X-DANCERS: GROUP CHOREOGRAPHY AS MASKED COMPLETION OF FORMATION-GROUNDED GROUP MOTION TENSOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Music-driven 3D group choreography requires jointly modeling individual motion, evolving formations, global trajectories and their synchronization with music. Existing methods model these factors separately and rely on dedicated mechanisms for different choreographic operations, limiting flexibility and generalization. We present **X-Dancers**, a unified framework that formulates group choreography as *bidirectional masked generation* over a continuous group motion tensor, where each cell couples a dancer's motion representation with its root state to jointly represent motion, formation and trajectory. To learn their joint distribution, we introduce the *ChoreoGraph Transformer*, which uses interleaved temporal and formation-grounded spatial attention to reason over observed group context, while masked conditional diffusion generates the unknown tensor cells. A spatial-temporal-channel masking strategy exposes the model to diverse conditioning patterns, enabling choreography synthesis, long-duration continuation, cast extension, and trajectory control via different conditioning masks without task-specific retraining. The experimental results demonstrate its state-of-the-art motion fidelity, music synchronization and formation quality, together with strong zero-shot generalization to long-duration choreography.

1 INTRODUCTION

Group dance is one of the most expressive forms of human movement art (Humphrey, 1959; Desmond, 1997), yet choreographing it is an iterative and laborious process (Foster, 2010). Choreographers continually refine formations[1], movements and transitions, while adapting the same choreography to new casts, group sizes, floor paths and performance durations (Butterworth & Wildschut, 2012). Automating this process thus requires more than generating a fixed performance, where a model must jointly capture continuous per-dancer motion, evolving formations and floor trajectories in synchronization with music, while allowing the resulting choreography to be extensible and revisable. Existing methods only partially support this flexibility. They either discretize group motion via vector quantization, limiting individual motion fidelity (Yang et al., 2024; Zhao et al., 2025), or model continuous motion for a fixed cast while handling global positioning through separate trajectory stages or dedicated objectives (Le et al., 2023a;b; Dai et al., 2025a;b). As a result, variable cast sizes, long-duration generation and editing typically require separate task-specific mechanisms (Yang et al., 2024; Le et al., 2024; Xu et al., 2026; Tseng et al., 2023).

In this work, we introduce **X-Dancers** from a different perspective: a group choreography should be modeled as a single joint process rather than assembled from separately generated components. Individual motion, formation, global trajectory and musical synchronization are inherently coupled, where changing the motion or position of one dancer changes the evolving formation, and thus constrains the rest of the group. Given music $\mathbf{m}$, we therefore model the motions and global placements of all dancers over the entire sequence as *one joint distribution*:

$$p\left(\left\{\mathbf{z}_{i,t},\ \mathbf{r}_{i,t}\right\}_{\substack{i=1,\dots,N\\ t=1,\dots,T}} \middle|\ \mathbf{m}\right), \tag{1}$$

[1]Formation of a dance team is the specification of positions of dancers or dance couples on the floor relative to each other and directions the dancers face or move with respect to others.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

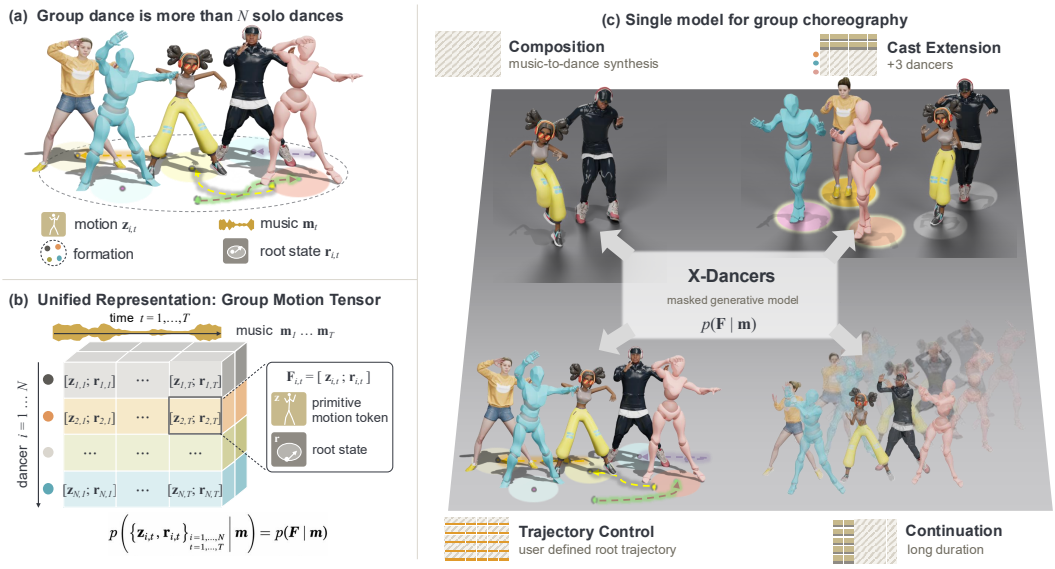


Figure 1: **X-Dancers as Unified Joint Process for Group Choreography.** (a) Group choreography requires jointly modeling individual motions and their coordination through evolving formations and trajectories. (b) We represent the choreography as a group motion tensor to enable joint modeling across dancers and time. (c) Spatial-temporal-channel masking turns different choreographic operations into conditioning patterns over the same joint process without re-training.

where $\mathbf{z}_{i,t}$ denotes the motion of dancer i at time t, and $\mathbf{r}_{i,t}$ its global root state. We instantiate this joint representation as a group motion tensor $\mathbf{F}$, with each cell $\mathbf{F}_{i,t} = [\mathbf{z}_{i,t}; \mathbf{r}_{i,t}]$, and formulate choreography generation as learning $p(\mathbf{F}|\mathbf{m})$ over *all dancers* $\times$ *all times* $\times$ *motion-and-placement*.

To learn this joint distribution, we formulate group choreography synthesis as bidirectional masked generation over the group motion tensor conditioned on music, where randomly masked cells are progressively recovered from the visible group context. We realize this with the *Choreo-Graph Transformer*, which propagates bidirectional information along the temporal and dancer axes through interleaved temporal and spatial attention, followed by a per-token diffusion head that recovers each masked cell. A key challenge arises from the structure of the dancer axis, where, unlike the fixed coordinates of the ordered frames of a video, dancers form a set with no intrinsic spatial coordinates. Their meaningful geometry is induced by the formation, which is itself an output of generation. This motivates our *formation-frame attention*, which constructs a formation coordinate frame from the root states visible at each stage of generation, and represents each visible dancer within this evolving frame. As a result, the spatial structure governing inter-dancer attention is not prescribed a priori, but is generated and progressively refined together with the choreography.

To enable generalization to diverse choreographic scenarios, we further propose a randomized spatial-temporal-channel masking strategy (STC-Mask) applied over timeframes, dancer poses and root channels of a cell to train the model. At inference, these masking patterns become conditioning interfaces for composition, arbitrary-length continuation, cast extension and trajectory control within the same generative process. Our contributions are:

- **Joint choreography modeling** couples continuous motion tokens and explicit root states in a group motion tensor for music-conditioned bidirectional masked generation.
- **Formation-grounded spatial reasoning** in the ChoreoGraph Transformer uses formation-frame attention to condition inter-dancer interactions on evolving frames derived from visible root states.
- **Unified choreographic operations** arise from STC-Mask training, enabling composition, continuation, cast extension and trajectory control through conditioning masks alone, without task-specific retraining.

2 Related Work

2.1 Multi-person Interaction Generative Modeling

Human motion generation has progressed rapidly with deep generative models, from diffusion-based single-person synthesis (Tevet et al., 2023; Zhang et al., 2024; Karunratanakul et al., 2023) to discrete variants built on VQ-VAEs (van den Oord et al., 2017) with autoregressive or masked transformers (Zhang et al., 2023; Jiang et al., 2023a; Guo et al., 2024; Pinyoanuntapong et al., 2024; 2025). Two-person interaction is addressed by joint-trajectory diffusion (Liang et al., 2024; Ruiz-Ponce et al., 2024; Shafir et al., 2024) and quantized masked modeling (Javed et al., 2025); number-free frameworks condition each person on the others (Fan et al., 2024) or compose two-person priors (Xu et al., 2025b; Ota et al., 2025), capturing little group-level structure beyond trajectories. Continuous masked autoregressive (MAR) modeling (Li et al., 2024c; Meng et al., 2025) has recently emerged as an expressive formulation for high-dimensional continuous distributions; we arrive at a related formulation from a different motivation, since group choreography is naturally a single joint distribution whose composition, restaging, extension, and revision modes are different mask queries over the same distribution, with the root state of every cell a maskable channel through which formation and trajectory enter the same query.

2.2 Music-driven Dance Generation

Single-dancer music-driven generation has matured from transformer-based models (Li et al., 2021; 2022) to diffusion-based ones (Tseng et al., 2023; Li et al., 2024a), with two-person extensions (Li et al., 2024b; Ghosh et al., 2025) adding synchronized partner interaction. Group-level frameworks (Wang et al., 2022; Le et al., 2023a;b; Dai et al., 2025a;b; Yang et al., 2025) coordinate dancers through group embeddings, contrastive objectives, or trajectory modules, and treat formation as a separately controlled quantity; variable cast sizes are reached by sequential dancer generation (Yang et al., 2024), discrete tokens over a free number of dancers (Zhao et al., 2025), or a phase manifold (Le et al., 2024), and long durations by dedicated sampling schemes (Dai et al., 2025a; Xu et al., 2025a; 2026). Each of these capabilities is delivered by its own mechanism. We instead recast composition, continuation, cast extension, trajectory control, and editing as different mask patterns over a single joint distribution, so that one trained model handles all five via mask-conditional sampling along the temporal, dancer, and channel axes.

2.3 Geometry-Grounded Attention and Controllable Generation

Attention over a set of agents needs a notion of position. Point cloud transformers (Zhao et al., 2021), multi-agent forecasters (Zhou et al., 2022; 2023), and topology-based models (Xu et al., 2022; Lee et al., 2024; Ray et al., 2025) derive it from observed coordinates; in generative models of molecules, protein structures, and multi-agent trajectories (Hoogeboom et al., 2022; Jumper et al., 2021; Jiang et al., 2023b), they are noisy states that are always present, and masked generative models (Chang et al., 2022; Li et al., 2024c; Guo et al., 2024) operate on grids whose positions are fixed in advance. Group choreography lies between these regimes: the dancer axis is a set whose positions are outputs of generation and are visible only once committed. Our formation-frame attention sits inside the masked generative backbone: a pre-shape normalization (Kendall, 1984; Dryden & Mardia, 2016) of the visible cast, recomputed from committed cells at every sampling step, encodes each dancer's place in the formation, so that spatial reasoning and generation are jointly supervised under a single objective. The same root channel provides control: whereas trajectory-conditioned generation adds guidance or dedicated controllers to a single dancer (Karunratanakul et al., 2023; Liu et al., 2024; Pinyoanuntapong et al., 2025) and group methods attach a trajectory module or positioning losses (Dai et al., 2025a;b), supplying the root channel of the tensor is a mask pattern seen during training, so control, composition, and editing share one sampler, one training objective, and one trained model.

3 Methodology

Given music $\mathbf{m}$ and N dancers, we model choreography as one joint distribution $p(\mathbf{F} \mid \mathbf{m})$ over a group motion tensor whose cells $\mathbf{F}_{i,t} = [\mathbf{z}_{i,t}; \mathbf{r}_{i,t}]$ couple continuous motion tokens with explicit

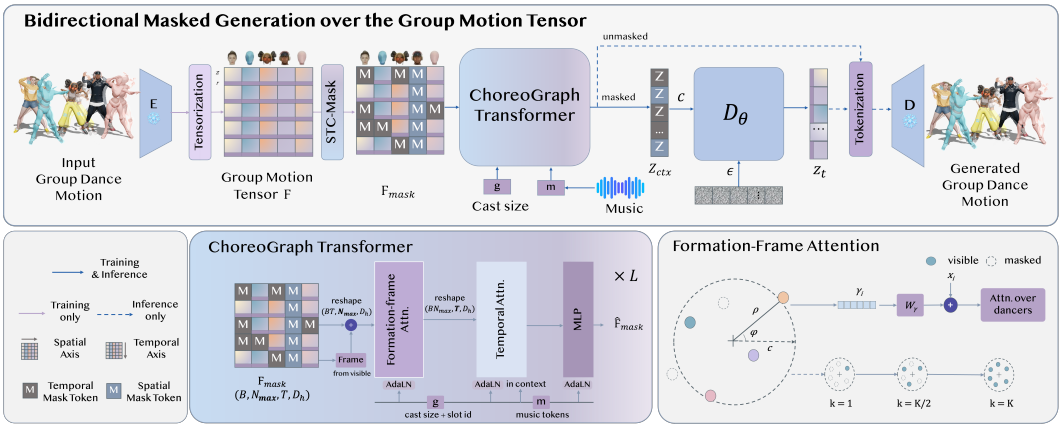


Figure 2: **Method overview of X-Dancers.** *Top*: X-Dancers performs bidirectional masked generation over the group motion tensor, jointly modeling dancer motion, formation and trajectory conditioned on music. The ChoreoGraph Transformer reasons over the observed group context, while token-wise diffusion generates the masked content. *Bottom*: ChoreoGraph interleaves formation-grounded spatial attention with temporal attention. Our formation-frame attention grounds inter-dancer reasoning in the evolving formation constructed from visible root states, allowing the spatial structure used by attention to evolve together with the generated choreography.

root states. Specifically, we learn $p(\mathbf{F}_{\mathcal{M}} \mid \mathbf{F}_{\bar{\mathcal{M}}}, \mathbf{m})$ through a bidirectional masked generative process, where $\mathcal{M}$ and $\bar{\mathcal{M}}$ denote masked and visible content varied by STC-Mask strategy across $\mathcal{M}$ across dancers, time and channels. Our ChoreoGraph Transformer integrates music with bidirectional context across dancers and time to condition per-token diffusion. To characterize dynamic spatial group formations, we introduced formation-frame attention, which constructs an evolving coordinate frame from visible root states so that inter-dancer interactions are grounded in the emerging formation. At inference, diverse choreographic operations are unified under this conditional completion paradigm through the choice of observed content, from composition, trajectory control, to continuation and cast extension.

3.1 Group Motion Tensor

To first obtain a shared representation of individual motion at each temporal step, we train a subject-agnostic autoencoder on single-dancer motion. The encoder E maps a motion sequence $\mathbf{x} \in \mathbb{R}^{T_m \times D_x}$ to tokens $\mathbf{z} \in \mathbb{R}^{T \times d}$, and the decoder $\mathcal{D}$ reconstructs $\hat{\mathbf{x}}$. Reconstruction losses on kinematic features and a KL penalty of small weight β characterize the token space as a manifold of plausible dance primitives shared across dancers. Both networks are frozen thereafter. While $\mathbf{z}_{i,t}$ captures how dancer i moves, it does not represent where the dancer stands, which is important for group formation. We therefore pair each motion token with the root state $\mathbf{r}_{i,t} = (\mathbf{p}_{i,t}, \sin\theta_{i,t}, \cos\theta_{i,t}) \in \mathbb{R}^4$, where $\mathbf{p}_{i,t} \in \mathbb{R}^2$ and $\theta_{i,t}$ denote the planar root position and heading of dancer i at token step t, with position expressed relative to the group centroid over the clip. We organize these states across all dancers and time steps into a group motion tensor

$$\mathbf{F} \in \mathbb{R}^{N \times T \times (d+4)}, \qquad \mathbf{F}_{i,t} = [\,\mathbf{z}_{i,t};\, \mathbf{r}_{i,t}\,], \tag{2}$$

indexed by the dancer set $\mathcal{V} = \{1, \ldots, N\}$ and temporal steps $\mathcal{T} = \{1, \ldots, T\}$. This representation jointly captures individual motion, global placement and their dependencies across dancers and time. Time has an intrinsic ordering aligned with the music, whereas the dancer ordering is canonicalized by polar angle about the group centroid at the first step, such that the slot index carries only this rank (Appendix G). Importantly, representing the entire choreography as $\mathbf{F}$ allows generation to be formulated as a masked completion problem.

3.2 ChoreoGraph Transformer

To parameterize this conditional completion process, we introduce the ChoreoGraph Transformer. Given a partially observed group motion tensor, it aggregates the visible motion, formation and

music context into $\mathbf{Z}_{\text{ctx}} \in \mathbb{R}^{N\times T\times D_h}$, which conditions a per-token diffusion head for recovering the masked cells. Each cell is embedded as $\mathbf{h}_{i,t} = W_{\text{in}}\mathbf{F}_{i,t} + W_\gamma \boldsymbol{\gamma}_{i,t} + \text{PE}(i,t)$, where a masked cell is represented by a learned mask token, PE is a sinusoidal encoding of dancer slot and temporal step, and $\boldsymbol{\gamma}_{i,t}$ denotes the formation descriptor introduced in Section 3.2.1. For channel-wise masking, the masked motion component is represented by the corresponding part of the mask token while the visible root state is retained (Appendix F).

3.2.1 Formation-Frame Attention

Unlike temporal attention, spatial attention has no predefined coordinate system. The dancer axis is a set whose geometry is determined by the evolving formation, which is itself only partially observed during masked generation. We thus introduce *formation-frame attention*, which grounds inter-dancer attention in the evolving formation characterized by formation frame descriptors.

Formation frame descriptors. At each step t, we construct the formation frame from the dancers whose root states are currently visible. Let $S_t \subseteq \mathcal{V}$ denote this visible set and $\mathbf{p}_{j,t}$ the root position of dancer j. We first summarize the current formation by its center $\bar{\mathbf{p}}_t$ and scale ℓ_t, and express each visible dancer relative to them: $\mathbf{q}_{i,t} = \frac{\mathbf{p}_{i,t} - \bar{\mathbf{p}}_t}{\ell_t}$. Thus, $\mathbf{q}_{i,t}$ describes *where dancer i lies within the current formation*, independent of its global translation and scale. This follows the normalization used in statistical shape analysis (Kendall, 1984; Dryden & Mardia, 2016), ensuring that the choreography is staged relative to the audience. Position alone cannot fully illustrate a dancer's role in an evolving formation. As illustrated in Figure 2, with $\mathbf{q}_{i,t} = \rho_{i,t}(\cos\varphi_{i,t}, \sin\varphi_{i,t})$, we represent its formation-relative position with its temporal change and local spacing:

$$\boldsymbol{\gamma}_{i,t} = \big(\rho_{i,t},\ \sin\varphi_{i,t},\ \cos\varphi_{i,t},\ \Delta_t\rho_{i,t},\ \rho_{i,t}\Delta_t\varphi_{i,t},\ d^{\text{nn}}_{i,t}\big), \tag{3}$$

where $d^{\text{nn}}_{i,t}$ is the distance to the nearest visible dancer. The descriptor captures three aspects of the evolving formation, including a dancer's relative position, its movement within the formation and its local spacing. For $i \notin S_t$, we set $\boldsymbol{\gamma}_{i,t} = \mathbf{0}$, leaving its spatial role unspecified until its root state becomes visible. The resulting descriptors are thus invariant to global translation and dilation, except for the metric spacing term $d^{\text{nn}}_{i,t}$, and are equivariant to dancer permutations.

Attention on the formation. During spatial attention, each dancer attends to the other dancers at the same time step using both its motion content and its position in the current formation. Let $\mathbf{a}_{i,t} = W_{\text{in}}\mathbf{F}_{i,t} + \text{PE}(i,t)$ denote the content embedding. We augment it with the formation descriptor before computing attention:

$$\alpha^h_{ij,t} = \underset{j\in\mathcal{V}}{\text{softmax}}\left[\frac{\big(W_{Q,h}\tilde{\mathbf{h}}_{i,t}\big)^\top\big(W_{K,h}\tilde{\mathbf{h}}_{j,t}\big)}{\sqrt{d_h}}\right], \qquad \tilde{\mathbf{h}}_{i,t} = \text{AdaLN}\big(\mathbf{a}_{i,t} + W_\gamma\boldsymbol{\gamma}_{i,t};\ \tilde{\mathbf{m}}_t, \mathbf{g}_i\big). \tag{4}$$

Thus, inter-dancer attention is determined not only by what each dancer is doing, but also by *where* and *how* it moves within the current formation. To be specific, the formation contribution to each attention head contains a learned bilinear kernel $\boldsymbol{\gamma}^\top_{i,t}\mathbf{B}_h\boldsymbol{\gamma}_{j,t}$, where $\mathbf{B}_h = W^\top_\gamma W^\top_{Q,h} W_{K,h} W_\gamma/\sqrt{d_h}$. This kernel allows a head to learn relations between formation roles, such as dancers at similar radii, opposite sides of the formation or dancers moving toward one another. Since $\boldsymbol{\gamma}$ is defined in the normalized formation frame, these relations remain consistent as the formation translates or scales. Moreover, since the frame is recomputed from currently visible root states, the same attention naturally follows the formation as it is progressively constrained by user-specified trajectories. The full decomposition is provided in Appendix E.3.

3.3 Training Procedure

STC-Masking Strategy. To support different choreographic operations within one model, STC-Mask varies the observed context during training to reflect their distinct conditioning requirements. Spatial and temporal masks hide selected dancers or time steps, while channel masks retain root states and hide motion tokens. The model thus learns to complete choreography from partial performances or prescribed trajectories. Appendix F specifies mask patterns and probabilities.

Training Objective. Regressing a cell from $\mathbf{Z}_{\text{ctx}}$ would average over the modes of dance. A diffusion head D_θ of AdaLN-Zero MLP blocks samples each masked cell $\mathbf{u} = \mathbf{F}_{i,t}$ independently given its context $\mathbf{c} = \mathbf{Z}_{\text{ctx}}[i,t]$, trained with the velocity objective of SiT (Ma et al., 2024) on the interpolant $\mathbf{u}_s = (1-s)\,\mathbf{u}_0 + s\,\boldsymbol{\epsilon}$,

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\mathbf{u}_0,\boldsymbol{\epsilon},s}\Big[\big\|\mathbf{w}\odot\big(D_\theta(\mathbf{u}_s,s,\mathbf{c}) - (\boldsymbol{\epsilon}-\mathbf{u}_0)\big)\big\|^2\Big], \qquad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), \quad s\sim\mathcal{U}[0,1], \tag{5}$$

where $\mathbf{w}$ weights the four root channels by λ_{r} and the d motion channels by one A one-step estimate $\hat{\mathbf{u}}_0$ of every masked cell is available at any noise level, which we regularize through the frozen decoder: a collision term $\mathcal{L}_{\text{coll}} = \text{mean}_{i\neq j,\,t}\,\big(\delta - \|\hat{\mathbf{p}}_{i,t} - \hat{\mathbf{p}}_{j,t}\|\big)_+$ pull root positions apart by a margin δ; a contact term $\mathcal{L}_{\text{con}}$ penalizes the speed of a decoded foot while its contact label marks it planted; a motion term $\mathcal{L}_{\text{mot}}$ applies an ℓ_1 loss to decoded joint positions against ground truth, down-weighted to zero beyond a cutoff on s. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{coll}}\mathcal{L}_{\text{coll}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{mot}}\mathcal{L}_{\text{mot}}, \tag{6}$$

with λ_{r}, δ, the cutoff and the weights given in Appendix D.

3.4 Group Choreography Synthesis

In the synthsis process, cells that are not pre-defined or conditioned by the user are masked and progressively recovered over K iterations in random order at a cosine-scheduled rate (Chang et al., 2022; Li et al., 2024c). Each choreographic operation is thus a choice of mask over the learned joint distribution. *(a) Composition* samples a fully masked tensor from music alone. *(b) Continuation* keeps the trailing k steps of the previous window as a prefix whose root states carry the formation across windows, choreographing music of any length window by window. *(c) Cast extension* keeps the cells of the existing dancers and completes appended rows up to $N_{\max}$, the new dancers placed by formation-frame attention relative to the cast, even mid-performance. *(d) Trajectory control* pre-defines values of the root channels given by users, so that only kinematics consistent with the prescribed paths are generated; uncontrolled dancers follow paths from a preceding composition pass (Appendix G). Figure 3 gives an gallery of results from each choreographic operations.

4 Evaluation Results

Implementation details. We train and evaluate our model and baseline methods on AIOZ-GDANCE (Le et al., 2023b), a large-scale 3D group dance corpus with ∼16.7 hours of paired music, RGB videos, and group motion annotations from in-the-wild videos. To optimize the ground truth data quality, we introduced a cleaning and refinement pipeline to refine the inaccurate annotations. Quantitative comparisons supply the same ground-truth root trajectories to all methods. The motion representation uses $D_x = 223$ features per frame, encoded by the frozen autoencoder into $d = 512$-dimensional tokens at one quarter of the frame rate. The ChoreoGraph Transformer uses $L = 4$ blocks of width 1024 and is paired with a 12-block MLP diffusion head. Further details are provided in Appendix D.

Evaluation metrics. We assess generated choreography at two complementary levels: *individual* and *group*. *Individual-level* metrics evaluate each dancer independently, using Single-Choreo FID (**SC-FID**) for motion fidelity, Beat Alignment Score (Li et al., 2021; Tseng et al., 2023; Dai et al., 2025a;b) (**BAS**) for music-motion alignment, per-dancer kinetic diversity (**Div**) for motion variety, and foot sliding distance (**FSD, mm**) for physical plausibility. *Group-level* metrics evaluate how dancers interact as a group, using Group-Choreo FID (**GC-FID**), our proposed distribution-level metric for group choreography, together with inter-dancer body collision rate (**Coll., %**) for spatial plausibility. Unlike pairwise kinetic correlation used in prior work (Le et al., 2023a; Yang et al., 2024), which primarily favors synchronized motion, GC-FID compares the distribution of generated group descriptors against that of ground truth, with feature contributions balanced between formation geometry and inter-dancer dynamics.

4.1 Results

Baseline comparison. Table 1 reports the comparison on the full AIOZ-GDANCE benchmark with the same dataset protocol as in Le et al. (2023a). Our method outperforms all baselines on

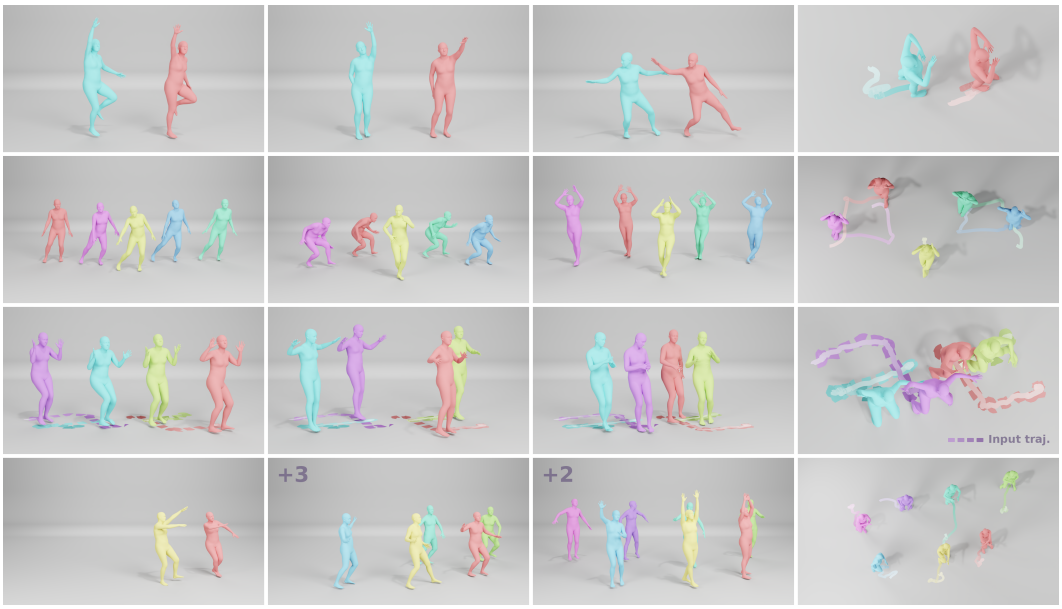


Figure 3: **Choreographic operations synthesis gallery.** The first two rows show generated performances with two and five dancers. The third row displays our method supports trajectory control task. The fourth row demonstrates cast extension from two to five and then seven dancers. The first three columns show successive moments, while the last column presents a top view with trajectory traces. Colors identify dancers within each sequence. Color gradients along trajectories indicate temporal order, from lighter (farther in time) to darker (closer in time).

every metric, from single-dancer motion plausibility to group dance choreography synthesis fidelity. Figure 6 visualizes representative results, which reveal the failure modes behind these numbers. GCD (Le et al., 2023a) ranks second in beat alignment but last in group quality: it couples dancers only through content similarity and reduces placement to three root coordinates within a 151-d pose, so the cast follows the beat but not a plausible formation. FreeDance (Zhao et al., 2025) quantizes spatial positions into each dancer's latent token, yielding largely homogeneous poses and drifting trajectories; despite its second-best collision rate, dancers thus collide severely in some frames and spread unreasonably far apart in others. TCDiff++ (Dai et al., 2025a) assumes dancers move alike and distinguishes them only by identity encodings and a swap-mode code, so formation never conditions pose: small casts collapse to homogeneous poses, and larger ones show a drifting dancer amid incoherent motion. In contrast, our group motion tensor pairs every motion token with an explicit root state, and formation-frame attention encodes each dancer by its coordinates in the committed cast, so formation conditions every pose, yielding coherent formations with differentiated yet coordinated motion. We conduct a user study, where our method again outweighs baseline methods in all aspects. Details and results are included in Appendix A.1.

Choreographic operations. Our group motion tensor representation supports composition and cast extension by varying which dancer rows are observed or generated. The first two rows of Figure 3 show two- and five-dancer compositions, with the latter also exhibiting a position exchange within the formation. The third row demonstrates dynamic cast extension, where a duet expands to five and then seven dancers during the performance. Existing dancer rows provide motion and placement as context for generating newly added rows, allowing the cast to expand within the same generation process without task-specific retraining. The same model further supports trajectory control through supplied root channels and continuation through a retained temporal prefix, as described in Section 3.4. Additional long-duration results are provided in demo videos (Appendix A).

Table 1: **Quantitative comparison on the full AIOZ-GDANCE dataset ($N_{\max} = 7$).** SC-FID, BAS, Div and FSD assess individual dancer quality; GC-FID and Coll. assess group coherence. **Best** / second-best are highlighted; ↑/↓ denote higher/lower is better.

	Individual-Level				Group-Level	
Method	SC-FID ↓	BAS ↑	Div ↑	FSD (mm) ↓	GC-FID ↓	Coll. (%) ↓
GCD (Le et al., 2023a)	37.27	0.361	10.52	21.40	34.63	3.261
CoDancers (Yang et al., 2024)	28.93	0.307	9.26	22.30	30.72	2.574
TCDiff (Dai et al., 2025b)	20.58	0.352	11.24	18.10	18.47	2.286
FreeDance (Zhao et al., 2025)	24.35	0.354	8.35	27.20	26.20	1.557
TCDiff++ (Dai et al., 2025a)	16.21	0.358	12.61	14.70	12.63	1.739
X-Dancers (Ours)	**7.25**	**0.365**	**13.05**	**10.58**	**7.64**	**1.008**
Ground Truth	–	0.367	9.48	4.70	–	0.956

Table 2: **Ablation on AIOZ-GDANCE.** Bold and underlined values mark the best and second-best updated results.

	Individual-Level				Group-Level	
Method	SC-FID ↓	BAS ↑	Div ↑	FSD (mm) ↓	GC-FID ↓	Coll. (%) ↓
A1: VQVAE-based masked transformer	31.74	0.348	5.93	13.70	105.83	1.762
A2: Diffusion transformer	22.91	0.367	10.59	15.30	74.62	2.138
B1: Spatial-only masking	18.83	0.358	7.24	13.10	25.94	1.396
B2: Temporal-only masking	14.29	0.362	6.15	11.32	43.20	1.865
C1: Graph attention	8.14	0.353	8.30	11.10	20.74	1.527
C2: w/o formation descriptors	9.74	**0.367**	12.16	10.81	13.72	1.011
Ours (full)	**7.25**	0.365	**13.05**	**10.58**	**7.64**	**1.008**

4.2 Ablation Studies

Effects of model components. Table 2 evaluates the variants defined in Appendix B. *(i) Generative paradigm.* **A1** degrades motion fidelity and group quality, supporting continuous representations. **A2** also loses group quality despite comparable beat alignment, favoring progressive completion with generated motion and placement as context. *(ii) Effect of STC-Mask.* **B1** reduces motion fidelity, group quality, and diversity. **B2** improves motion fidelity and foot stability over **B1** but worsens group quality, diversity, and collisions, showing that temporal completion alone does not ensure group coordination. STC-Mask learns from both spatial and temporal context while matching root-conditioned evaluation through channel masking. Channel-only training is excluded because it never trains root-state completion. *(iii) Formation-grounded spatial reasoning.* **C1** retains competitive motion fidelity but degrades group quality and diversity while increasing collisions. **C2** likewise reduces motion fidelity, group quality, and diversity despite comparable beat alignment. Since **C2** retains explicit root states, the degradation supports formation-relative geometry.

Analysis of formation-frame attention. Figure 4 visualizes first-layer attention during the same five-dancer position exchange shown in the second row of Figure 3, where the full trajectories are provided. Dancers 1 and 2 are fully visible, while the top-two keys of partially observed dancers 3, 4, and 5 define the illustrated neighborhoods. As dancers 1 and 3 move inward and dancers 4 and 5 move outward, the selected neighborhoods evolve with the formation. At $t = 1.07\,\text{s}$, query 3 favors dancers 5 and 1 on the opposite side. Dancers 1 and 2 are nearly collinear with query 3 and receive similar scores, with dancer 1 marginally higher. At $t = 4.27\,\text{s}$, as the exchange progresses, neighborhoods $\{3, 4, 2\}$ and $\{5, 1, 2\}$ each contain an approaching partner and the shared inner reference dancer 2. At $t = 6.40\,\text{s}$, after dancers 4 and 5 reach the outer positions, query 5 shifts its strongest attention toward dancer 4, which occupies a comparable outer radius on the opposite side while retaining inner dancer 1. These changes show that the attended context is not fixed across time, but reorganizes with the evolving formation. The bilinear formation kernel can

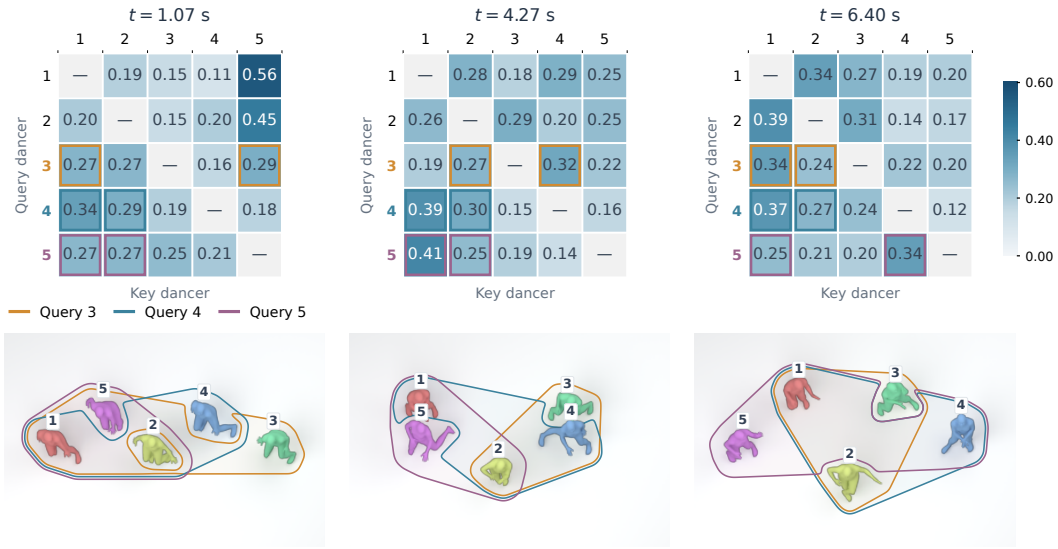


Figure 4: **Inter-dancer attention across a formation transition**. Columns show three timestamps of the five-dancer position-exchange example. (a) First-layer inter-dancer attention averaged over heads (self-attention omitted), probed on saved root conditions with motion channels masked; outlines mark the two strongest non-self keys for query dancers 3, 4, and 5. (b) The corresponding selections from (a) are shown in formation space, where each contour encloses a query dancer and its two selected keys, illustrating how the attention neighborhoods reorganize as the formation evolves.

express complementary relations through the formation descriptors, allowing attention to respond to changing relative radius, angular arrangement, and inter-dancer motion.

4.3 Limitations

While X-Dancers attains state-of-the-art group choreography synthesis with a single bidirectional masked generative model, a few limitations remain. First, the cast size is bounded by the training maximum $N_{\max}$. Zero-shot extending beyond it requires generating the cast in blocks, which causes severe collisions. Megacrew choreography scenes are thus out of reach of the current model. Second, the model suffers generalizing to some out-of-distribution music genres to the current dataset. Samples shown in demo videos (see Section. A). These limitations motivate future work on generalizing X-Dancers to larger casts and broader music genres.

5 Conclusion

We presented X-Dancers, a unified framework that models music-driven 3D group choreography as a single joint process over continuous individual motion, evolving formations, and global trajectories. By coupling motion tokens with explicit root states in a group motion tensor, bidirectional masked generation completes missing motion and placement from the observed group context. The ChoreoGraph Transformer interleaves temporal and formation-frame attention, so that the spatial structure governing inter-dancer interactions evolves with the choreography. Trained with STC-Mask, one model supports composition, trajectory control, long-duration continuation and cast extension through conditioning masks alone, without task-specific retraining. Experiments on AIOZ-GDANCE demonstrate state-of-the-art motion fidelity, music synchronization, and formation quality, while demonstrations exhibit differentiated yet coordinated motion under trajectory control and dynamic cast extension. X-Dancers thus combines high-quality synthesis with the flexibility to extend, recast, and revise performances as choreographic requirements change.

ETHICS STATEMENT

This work synthesizes 3D group choreography and does not involve the collection of new human-subject motion data. All motion and music data come from the publicly released AIOZ-GDANCE benchmark (Le et al., 2023b), which is built from in-the-wild videos; we use it under its original license and terms, and we release only the filtering script that defines the GDANCE-clean subset, not any redistributed media. Because our model generates full-body motion conditioned only on music and does not model identity, appearance, or face, it does not enable identity impersonation. We nevertheless note the general risk shared by all human-motion generators, that synthesized performances could be presented as authentic recordings of real dancers, and we encourage downstream users to label generated content as synthetic.

REPRODUCIBILITY STATEMENT

We have made the following efforts to support reproducibility. The full method, including the dance primitive manifold, the STC-Mask families, the ChoreoGraph Transformer with formation-frame attention, and the per-token diffusion head, is specified in Section 3 with all objectives written out. Section 4 states the dataset, hardware, latent dimension, number of blocks, head depth and batch size; Appendix D gives the remaining architecture and optimization hyperparameters, and Appendix I gives the exact definitions of the reported metrics. The GDANCE-clean subset is defined by a deterministic filter over the public AIOZ-GDANCE release; its thresholds and pseudocode are given in Appendix H, and the filter script is included in the supplementary material so that the subset can be regenerated exactly. All reported numbers are means over 5 fixed random seeds. Source code and the evaluation protocol are provided as anonymized supplementary material.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, specifically for grammar correction, wording, and copy-editing of author-written text, and as a coding assistant for routine implementation and visualization scripts. We have not used generative AI tools for research ideation, method design, experiment design, or analysis and interpretation of the results, and retrieval or generation of citations is not applicable to this work, as all references were collected and verified manually by the authors against their published venues. We have reviewed all AI-assisted work: AI-assisted code was read, tested, and verified against reference implementations by the authors, and every AI-edited passage of text was checked against the authors' intended technical claims. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Jo Butterworth and Liesbeth Wildschut. *Contemporary choreography*. Taylor & Francis, 2012.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022.

Yuqin Dai, Wanlu Zhu, Ronghui Li, Xiu Li, Zhenyu Zhang, Jun Li, and Jian Yang. TCDiff++: An end-to-end trajectory-controllable diffusion model for harmonious music-driven group choreography. *arXiv preprint arXiv:2506.18671*, 2025a.

Yuqin Dai, Wanlu Zhu, Ronghui Li, Zeping Ren, Xiangzheng Zhou, Jixuan Ying, Jun Li, and Jian Yang. Harmonious music-driven group choreography with trajectory-controllable diffusion. In *AAAI*, 2025b.

Jane Desmond. *Meaning in motion: New cultural studies of dance*. Duke University Press, 1997.

Ian L. Dryden and Kanti V. Mardia. *Statistical Shape Analysis: With Applications in R*. Wiley, 2nd edition, 2016.

Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *ECCV*, 2024.

Susan Foster. *Choreographing empathy: Kinesthesia in performance*. Routledge, 2010.

Anindita Ghosh, Bing Zhou, Rishabh Dabral, Jian Wang, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek, and Chuan Guo. DuetGen: Music driven two-person dance generation via hierarchical masked modeling. 2025.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *CVPR*, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *ICLR*, 2021.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *ICML*, 2022.

Doris Humphrey. *The art of making dances*. Rinehart, 1959.

Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. InterMask: 3D human interaction generation via collaborative masked modeling. In *ICLR*, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *NeurIPS*, 2023a.

Chiyu Max Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, and Dragomir Anguelov. MotionDiffuser: Controllable multi-agent motion prediction using diffusion. In *CVPR*, 2023b.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *ICCV*, 2023.

David G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 1984.

Nhat Le, Tuong Do, Khoa Do, Hien Nguyen, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Controllable group choreography using contrastive diffusion. *ACM TOG*, 2023a.

Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Music-driven group choreography. *CVPR*, 2023b.

Nhat Le, Khoa Do, Xuan Bui, Tuong Do, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Scalable group choreography via variational phase manifold learning. In *ECCV*, 2024.

Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyoobin Lee. MART: Multiscale relational transformer networks for multi-agent trajectory prediction. In *ECCV*, 2024.

Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *CVPR*, 2024a.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI Choreographer: Music conditioned 3d dance generation with AIST++. In *ICCV*, 2021.

Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation via actor-critic GPT with choreographic memory. In *CVPR*, 2022.

Siyao Li, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower GPT with off-policy reinforcement learning for dance accompaniment. In *ICLR*, 2024b.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024c.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. InterGen: Diffusion-based multihuman motion generation under complex interactions. *IJCV*, 2024.

Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *CVPR*, 2024.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *CVPR*, 2025.

Sakuya Ota, Qing Yu, Kent Fujiwara, Satoshi Ikehata, and Ikuro Sato. PINO: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups. In *ICCV*, 2025.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative masked motion model. In *CVPR*, 2024.

Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. MaskControl: Spatio-temporal control for masked motion synthesis. In *ICCV*, 2025.

Abhisek Ray, Ayush Raj, and Maheshkumar H Kolekar. Autoregressive adaptive hypergraph transformer for skeleton-based activity recognition. In *WACV*, 2025.

Pablo Ruiz-Ponce, German Barquero, Cristina Palmero, Sergio Escalera, and José García-Rodríguez. in2IN: Leveraging individual information to generate human interactions. In *CVPR Workshops*, 2024.

Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. EDGE: Editable dance generation from music. In *CVPR*, 2023.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *NeurIPS*, 2017.

Zixuan Wang, Jia Jia, Haozhe Wu, Junliang Xing, Jinghe Cai, Fanbo Meng, Guowen Chen, and Yanfeng Wang. GroupDancer: Music to multi-people dance synthesis with group dance dataset. In *ACM MM*, 2022.

Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. GroupNet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *CVPR*, 2022.

Jing Xu, Weiqiang Wang, Cunjian Chen, Jun Liu, and Qiuhong Ke. ST-GDance: Long-term and collision-free group choreography from music. In *BMVC*, 2025a.

Jing Xu et al. ST-GDance++: A scalable spatial-temporal diffusion for long-duration group choreography. *arXiv preprint arXiv:2603.22316*, 2026. VERIFY author list.

Wenning Xu, Shiyu Fan, Paul Henderson, and Edmond S. L. Ho. Multi-person interaction generation from two-person motion priors. 2025b.

Kaixing Yang, Xulong Tang, Ran Diao, Hongyan Liu, Jun He, and Zhaoxin Fan. CoDancers: Music-driven coherent group dance generation with choreographic unit. In *International Conference on Multimedia Retrieval*, 2024.

Kaixing Yang, Xulong Tang, Haoyu Wu, Qinliang Xue, Biao Qin, Hongyan Liu, and Zhaoxin Fan. CoheDancers: Enhancing interactive group dance generation through music-driven coherence decomposition. In *ACM MM*, 2025. VERIFY author list.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023.

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 TPAMI*, 2024.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

Yiwen Zhao, Yang Wang, Liting Wen, Hengyuan Zhang, and Xingqun Qi. FreeDance: Towards harmonic free-number group dance generation via a unified framework. In *ICCV*, 2025.

Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. HiVT: Hierarchical vector transformer for multi-agent motion prediction. In *CVPR*, 2022.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *CVPR*, 2023.

A Additional Results

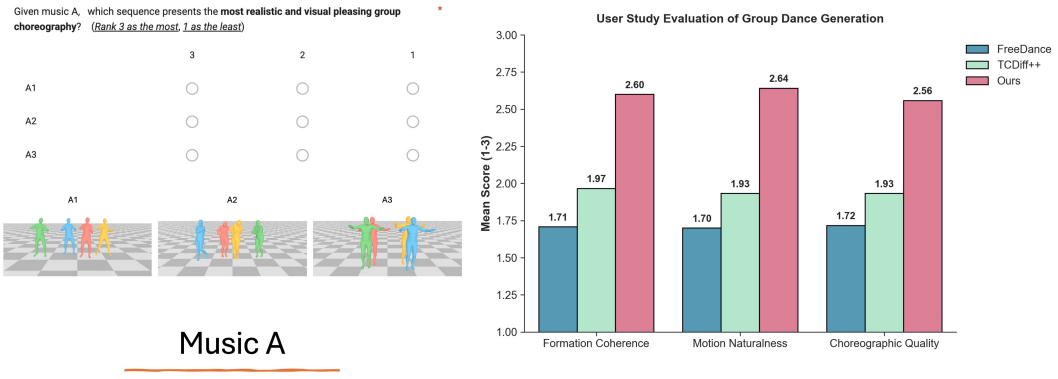
This section supplements Section 4.1 with visual comparisons and a user study of group choreography. Supplementary videos show the rendered examples: (1) video 01: presenting visual demos of the synthesis results of all the choreographic operation mentioned. (2) video 02: showcase the demo videos of baseline comparison, ablation studies, and limitations (Section 4.3). Figure 6 compares FreeDance, TCDiff++, and X-Dancers under the same music for casts of three and five dancers, illustrating differences in formation stability and pose coordination.

A.1 User study

This subsection details the user study reported in Section 4.1.

Protocol. We compared X-Dancers with FreeDance (Zhao et al., 2025) and TCDiff++ (Dai et al., 2025a) in a study with 20 participants. Each participant viewed 12 music sequences spanning different genres and group sizes. For each sequence, the three methods used the same music and were rendered with the same Blender pipeline, then displayed side by side in randomized order on a shared web interface. Participants ranked the samples from 1 to 3, with 3 indicating the best, on formation coherence, motion naturalness, and choreographic quality. These criteria assess spatial organization across dancers, physical plausibility and smoothness of individual motions, and overall visual appeal and musical synchronization, respectively.

Results. X-Dancers receives the highest mean score on all three criteria, followed by TCDiff++ and FreeDance, as shown in Figure 5. Participants favor both its formation coherence and motion naturalness, indicating that the perceived improvement spans group organization and individual motion. Its choreographic-quality score also supports the overall visual appeal.



(a) Question Template Overview

(b) User Study Evaluation Results

Figure 5: **User study.** The questionnaire appears on the left and mean scores across 20 participants and 12 music sequences on the right. Scores range from 1 for the worst to 3 for the best.

B Ablation Settings

This section defines the variants evaluated in Section 4.2 and Table 2, which examine the generative paradigm, masking strategy, and spatial representation. All use the same ground-truth root trajectories as described in Section 4. The spatial, temporal, and channel masks are defined in Appendix F.

A1. Vector-quantized motion tokens. Continuous motion tokens are replaced with vector-quantized codes, while explicit root states and masked generation over the group motion tensor are retained. This comparison evaluates motion discretization with explicit placement still available.

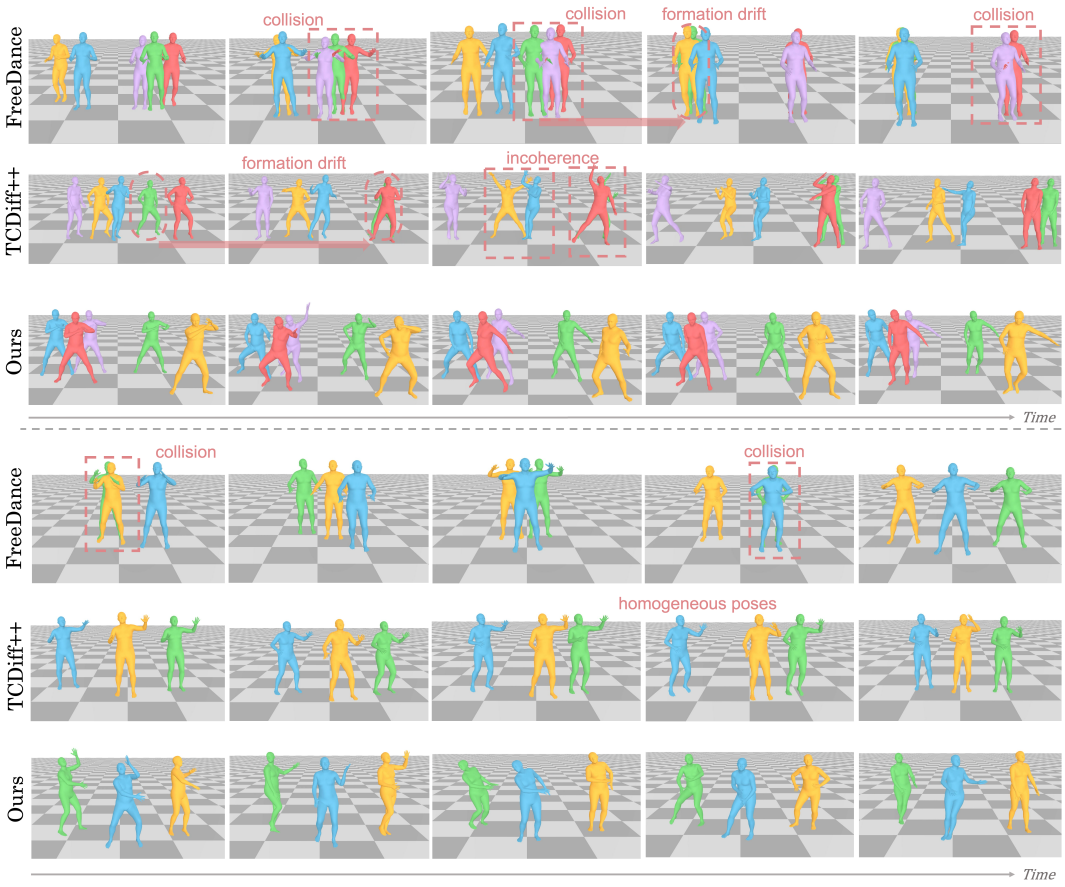


Figure 6: **Qualitative comparison of group choreography generation given the same music input.** *Top*: 5-dancer examples. *Bottom*: 3-dancer examples. Prior methods frequently exhibit degraded formation consistency and temporal coherence, leading to artifacts such as collisions, formation drift, and choreographic incoherence. Our method better preserves coherent formations and stable group dynamics over time.

A2. Joint diffusion. The continuous motion tokens and explicit root states are retained, but joint diffusion over the group motion tensor replaces bidirectional masked completion. The comparison tests progressive completion against joint denoising under the same representation.

B1. Spatial-only masking. Training uses only spatial masks, which hide each selected dancer's motion tokens and root states throughout the sequence. The remaining dancers provide context for completing the missing dancers. Temporal and channel masks are excluded.

B2. Temporal-only masking. Training uses only temporal masks, which hide motion tokens and root states at selected time steps while retaining unmasked steps as context. Spatial and channel masks are excluded, restricting training to completion along the temporal axis. Both **B1** and **B2** mask motion and root states together, so neither trains motion completion with root channels visible.

C1. Graph attention. Graph attention replaces formation-frame attention within a transformer with $L = 4$ blocks. This variant tests whether an alternative mechanism for inter-dancer interaction preserves group coordination as effectively as attention grounded in the visible formation.

C2. Without formation descriptors. The formation descriptors $\gamma_{i,t}$ and their projection are removed from the cell embeddings. Explicit root states, dancer-slot and temporal encodings, and the remaining transformer architecture are retained. This variant tests the contribution of formation-relative geometry beyond each dancer's root state and slot encoding.

Why channel-only masking is not included. Channel-only masking always reveals root trajectories and masks motion tokens. It therefore does not train completion of unknown root states, which composition, continuation, and cast extension require. We omit it as a standalone replacement for STC-Mask because it does not cover these generation tasks. It remains a relevant control for trajectory-conditioned evaluation, however. This comparison supports STC-Mask over spatial-only and temporal-only training without isolating the contribution of channel masking.

C NOTATION

This section summarizes the notation used in Section 3.

Symbol	Meaning
$\mathbf{x} \in \mathbb{R}^{T_m \times D_x}$	frame-rate motion of one dancer; D_x features per frame
E, $\mathcal{D}$	frozen encoder and decoder of the dance primitive manifold
$\mathbf{z}_{i,t} \in \mathbb{R}^d$, $\mathcal{Z}_{\text{prim}}$	motion token of dancer i at step t; the dance primitive manifold
$\mathbf{r}_{i,t} = (\mathbf{p}_{i,t}, \sin\theta_{i,t}, \cos\theta_{i,t})$	root state: planar position and heading
$\mathbf{F} \in \mathbb{R}^{N \times T \times (d+4)}$	group motion tensor; $\mathbf{F}_{i,t} = [\mathbf{z}_{i,t}; \mathbf{r}_{i,t}]$
$\mathcal{V}$, $\mathcal{T}$, Λ	dancer set, step set, channel-group index set $\mathcal{V} \times \mathcal{T} \times \{\mathbf{z}, \mathbf{r}\}$
$\mathcal{M} \subseteq \Lambda$	a mask (the cells to be generated)
$S_t \subseteq \mathcal{V}$	dancers whose root state is visible at step t
$\bar{\mathbf{p}}_t$, ℓ_t, $\mathbf{q}_{i,t}$	centroid, size and normalized coordinates of the formation frame
$\rho_{i,t}, \varphi_{i,t}, d^{\text{nn}}_{i,t}$	polar coordinates in the frame; nearest visible-neighbor distance
$\boldsymbol{\gamma}_{i,t} \in \mathbb{R}^6$, W_γ	formation descriptors and their projection
$\mathbf{h}_{i,t}$, $\mathbf{Z}_{\text{ctx}}$, D_h	cell embedding, transformer context, hidden width
$\mathbf{a}_i$, $\mathbf{B}$	content part of an embedding; bilinear formation kernel
$\mathbf{g}_i$, $\mathbf{m}_t$, $\tilde{\mathbf{m}}_t$	cast embedding; music token; music token with modality and position embedding
D_θ, $\mathbf{c}$, s, $\mathbf{u}_s$	diffusion head, its context vector, noise level, noised cell
λ_{r}, δ, λ_{coll}, λ_{con}, λ_{mot}	root-channel loss weight; collision margin; regularizer weights
K, k, $N_{\max}$	sampling iterations; overlap steps for continuation; training cast maximum

D IMPLEMENTATION DETAILS

This section details the settings in Section 4 and loss weights in Section 3.3.

Motion representation. Motions are represented at 30 fps with $D_x = 223$ features per frame: 24 joint rotations in the 6D representation (144), 24 local joint positions (72), the planar root velocity and height (3), and four foot-contact labels (4). Training clips are $T_m = 240$ frames (8 s).

Temporal motion tokenizer. The temporal individual motion tokenizer is an autoencoder based on 1D convolutional network with two downsampling stages of stride two, so tokens are produced at one quarter of the frame rate ($T = 60$ per clip) with $d = 512$ channels; the objective is $\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{recon}} + \beta\,\mathcal{L}_{\text{KL}}$ with a smooth-ℓ_1 reconstruction loss on the features, an ℓ_1 loss on forward-kinematics joint positions (weight 0.1), a foot-contact consistency loss (weight 0.05), and $\beta = 10^{-6}$. It is trained on every single-dancer sequence of the training split for 1000 epochs with batch size 128 and learning rate 2×10^{-4}, and frozen thereafter.

Root state. The root state is extracted from the ground-truth root joint at token rate: the planar position is centered by the mean of all dancers over the clip, and the heading is the direction of the displacement between consecutive token steps, the first step copying the second. At decoding, the position is read from the tensor and the heading is unused.

ChoreoGraph Transformer. $L = 4$ blocks with hidden width $D_h = 1024$, 16 attention heads, feed-forward width 4096 and dropout 0.2. Each block applies dancer-axis attention at every step, temporal attention for every dancer over $[\tilde{\mathbf{m}}_{1:T}; \mathbf{h}_{i,1:T}]$, and a feed-forward layer; every sublayer is modulated by adaptive layer normalization whose inputs are the music token of the current step and the cast embedding $\mathbf{g}_i$. Cast size and slot enter only through $\mathbf{g}_i$. The formation descriptors (6 channels) are projected to D_h by W_γ and added to the cell embedding; descriptors of masked cells are zero before projection.

Music encoder. Per-frame audio features are the 35-dimensional set of Li et al. (2021) (MFCC, chroma, onset and beat features), downsampled to the token rate, projected to width 256 and then to D_h, giving one music token $\mathbf{m}_t$ per step of the tensor. The tokens entering the temporal pass, $\tilde{\mathbf{m}}_t$, additionally carry a modality embedding and the sinusoidal encoding of step t, so that the concatenated sequence $[\tilde{\mathbf{m}}_{1:T}; \mathbf{h}_{i,1:T}]$ places each music token at the same temporal position as the cell it conditions. During training the music condition is dropped with probability 0.1 to enable classifier-free guidance.

Diffusion head and losses. The head is a 12-block AdaLN-Zero residual MLP of width 1024 conditioned on the context vector and the noise level, trained with the velocity objective of Ma et al. (2024). The root-channel loss weight is $\lambda_{\mathrm{r}} = 0.001$. The collision margin is $\delta = 0.8\,\mathrm{m}$ with $\lambda_{\mathrm{coll}} = 1.0$, applied after the first 100 epochs; the contact term uses the four foot joints and $\lambda_{\mathrm{con}} = 0.05$; the motion term uses $\lambda_{\mathrm{mot}} = 0.1$ with weight $(s_{\mathrm{c}} - s)_+ / s_{\mathrm{c}}$ and cutoff $s_{\mathrm{c}} = 0.2$ on the normalized noise level.

Training. Mask families are sampled with the probabilities of Appendix F. The training cast maximum is $N_{\max} = 7$. Optimization uses AdamW with learning rate 5×10^{-5}, weight decay 10^{-3}, 1000 warm-up iterations, batch size 128 and trained 1000 epochs. It takes 3 4 days on 4 $\times$ A100 GPUs.

Sampling. $K = 18$ iterations with a cosine schedule on the number of masked cells; the commitment order is a random permutation drawn once per sample. Classifier-free guidance on the music condition uses scale 3. The head integrates the flow with an adaptive a fixed-step Euler solver with 25 steps (T09). Continuation uses 8 s windows slided by stride of 4 s, i.e. $k = 30$ overlapping steps.

E Formation-Frame Attention: Properties and Details

This section expands the formation-frame attention formulation in Section 3.2.1. We establish the descriptor invariance properties, specify numerical safeguards, and derive the attention-logit decomposition before detailing the visible set during training and sampling.

E.1 Invariance properties of the descriptors

We verify the properties of the descriptors in Eq. (3) stated in Section 3.2.1. Fix a step t and the visible set S_t, and write $\mathbf{p}_j$ for $\mathbf{p}_{j,t}$.

Translation. Under $\mathbf{p}_j \mapsto \mathbf{p}_j + \mathbf{b}$ for all j, the centroid becomes $\bar{\mathbf{p}}_t + \mathbf{b}$, the size ℓ_t is unchanged, hence $\mathbf{q}_{i,t}$ is unchanged, and so are $\rho_{i,t}$, $\varphi_{i,t}$ and their step differences. Pairwise distances are unchanged, so $d^{\mathrm{nn}}_{i,t}$ is unchanged. The descriptors are therefore translation invariant.

Dilation. Under $\mathbf{p}_j \mapsto \lambda \mathbf{p}_j$ with $\lambda > 0$, the centroid becomes $\lambda \bar{\mathbf{p}}_t$ and the size $\lambda \ell_t$, hence $\mathbf{q}_{i,t}$ is unchanged, and so are the first five entries of $\boldsymbol{\gamma}_{i,t}$. Pairwise distances scale by λ, so $d^{\mathrm{nn}}_{i,t} \mapsto \lambda\, d^{\mathrm{nn}}_{i,t}$. The descriptors are dilation invariant except for the spacing entry, which is metric by design.

Permutation. For a permutation π of $\mathcal{V}$ applied to the positions and to S_t, the centroid and size are symmetric functions of the visible positions and are therefore unchanged, so $\mathbf{q}_{\pi(i),t}$ equals the former $\mathbf{q}_{i,t}$; the nearest visible neighbor distance is likewise a symmetric function of the other visible positions. Hence $\boldsymbol{\gamma}_{\pi(i),t}$ equals the former $\boldsymbol{\gamma}_{i,t}$, which is permutation equivariance. The velocity entries inherit each property because both endpoints transform identically.

E.2 Degenerate cases and numerical safeguards

The size ℓ_t is floored at $0.1\,\mathrm{m}$ and the distance to the centroid at $10^{-6}\,\mathrm{m}$, so that a lone visible dancer ($|S_t| = 1$) sits at $\mathbf{q}_{i,t} = \mathbf{0}$ with $\rho_{i,t} = 0$ and an arbitrary but finite $\varphi_{i,t}$; with no visible dancer ($|S_t| = 0$) every descriptor is zero. The spacing entry is capped at $10\,\mathrm{m}$. Angular differences are wrapped to $(-\pi, \pi]$ through $\mathrm{atan2}(\sin\Delta\varphi, \cos\Delta\varphi)$. Descriptors of dancers outside S_t are set to zero before projection, so masked positions contribute nothing to Eq. (3.2).

E.3 Decomposition of the attention logit

For the formation contribution in Section 3.2.1, let $\mathbf{a}_i$ denote the content part of $\mathbf{h}_i$ in Eq. (3.2), which consists of the projected cell and positional encoding at the first block. If this embedding were passed directly to Q/K projections, the unnormalized logit of head h would expand as

$$\begin{aligned} \widetilde{a}_{ij}^{(h)} &= \mathbf{a}_i^\top A_h \mathbf{a}_j + \mathbf{a}_i^\top A_h W_\gamma \boldsymbol{\gamma}_j \\ &\quad + \boldsymbol{\gamma}_i^\top W_\gamma^\top A_h \mathbf{a}_j + \boldsymbol{\gamma}_i^\top \mathbf{B}_h \boldsymbol{\gamma}_j, \\ A_h &= \frac{W_{Q,h}^\top W_{K,h}}{\sqrt{d_h}}, \qquad \mathbf{B}_h = W_\gamma^\top A_h W_\gamma. \end{aligned} \tag{7}$$

up to the linear and constant terms contributed by Q/K biases. The last term defines the 6×6 direct-path formation kernel visualized in Figure 7. Equation (7) is deliberately a weight-space diagnostic rather than a claim about the full forward pass: the actual first-block attention applies LayerNorm and music/cast-conditioned adaptive modulation before Q/K, and later blocks additionally mix descriptors with preceding sublayers. We therefore measure the realized descriptor effect by running the same root-only input twice, with the descriptors intact and zeroed, and reporting $\Delta A = A(\boldsymbol{\gamma}) - A(\mathbf{0})$. The diagnostic mirrors the content-to-position terms of disentangled attention (He et al., 2021) and relative geometric encodings for point clouds and multi-agent trajectories (Zhao et al., 2021; Zhou et al., 2022; 2023), while here the coordinates are themselves generated and progressively revealed.

E.4 The visible set during training and sampling

The visible set S_t of Section 3.2.1 is instantiated differently in the two regimes, but carries the same meaning. During training it consists of the unmasked cells at step t together with the masked cells whose root channels are kept visible by the channel family of Appendix F, which is why the channel family alone places every dancer in the frame. During sampling it consists of the cells supplied by the user together with the cells committed so far, and is recomputed at every iteration, so the frame is coarse at the first iterations and sharpens as the cast is placed; supplying root states for every cell, as trajectory control does, fixes the frame from the first iteration. In both regimes a zero descriptor means that the position of the dancer is not yet known.

F STC-Mask Families

This section specifies the mask patterns and probabilities used in Section 3.3 and the channel-wise embedding described in Section 3.2. Let $\mathcal{M} \subseteq \Lambda = \mathcal{V} \times \mathcal{T} \times \{\mathbf{z}, \mathbf{r}\}$ denote the set of channel groups to be generated. Each training sample draws one of the following five patterns, where ρ is a masking ratio drawn from a cosine schedule as in MaskGIT (Chang et al., 2022).

Family	Variant	Mask $\mathcal{M}$	Probability / operation
Spatial	random rows	$R \times \mathcal{T} \times \{\mathbf{z}, \mathbf{r}\}$, $R \subset \mathcal{V}$ random, $\lvert R\rvert = \lfloor \rho N \rfloor \geq 1$	0.15 / infer a dancer from the others, recast
Spatial	block rows	$\{m{+}1, \ldots, N\} \times \mathcal{T} \times \{\mathbf{z}, \mathbf{r}\}$, $m \sim \mathcal{U}\{1, \ldots, N{-}1\}$	0.10 / cast extension
Temporal	random steps	$\mathcal{V} \times C \times \{\mathbf{z}, \mathbf{r}\}$, $C \subset \mathcal{T}$ random per dancer, $\lvert C\rvert = \lfloor \rho T \rceil$	0.50 / composition, in-betweening
Temporal	block steps	$\mathcal{V} \times \{k{+}1, \ldots, T\} \times \{\mathbf{z}, \mathbf{r}\}$, $k \sim \mathcal{U}\{1, \ldots, T{-}1\}$	0.15 / continuation
Channel	motion only	$\mathcal{V} \times \mathcal{T} \times \{\mathbf{z}\}$	0.10 / trajectory control

Masked cells are replaced by the learned mask token. Under the channel family the root channels of every cell remain visible and are concatenated with the motion part of the mask token before W_{in}; the visible set S_t of Section 3.2.1 therefore contains every dancer, whereas under the other families it contains the dancers whose cells are unmasked at step t. The diffusion loss is applied to the cells in $\mathcal{M}$. At inference each operation supplies the complement of one of these patterns or of a union of them (Appendix G).

G Inference and Choreographic Operations

This section details the sampler and choreographic operations in Section 3.4, together with the dancer-slot ordering used by the group motion tensor in Section 3.1.

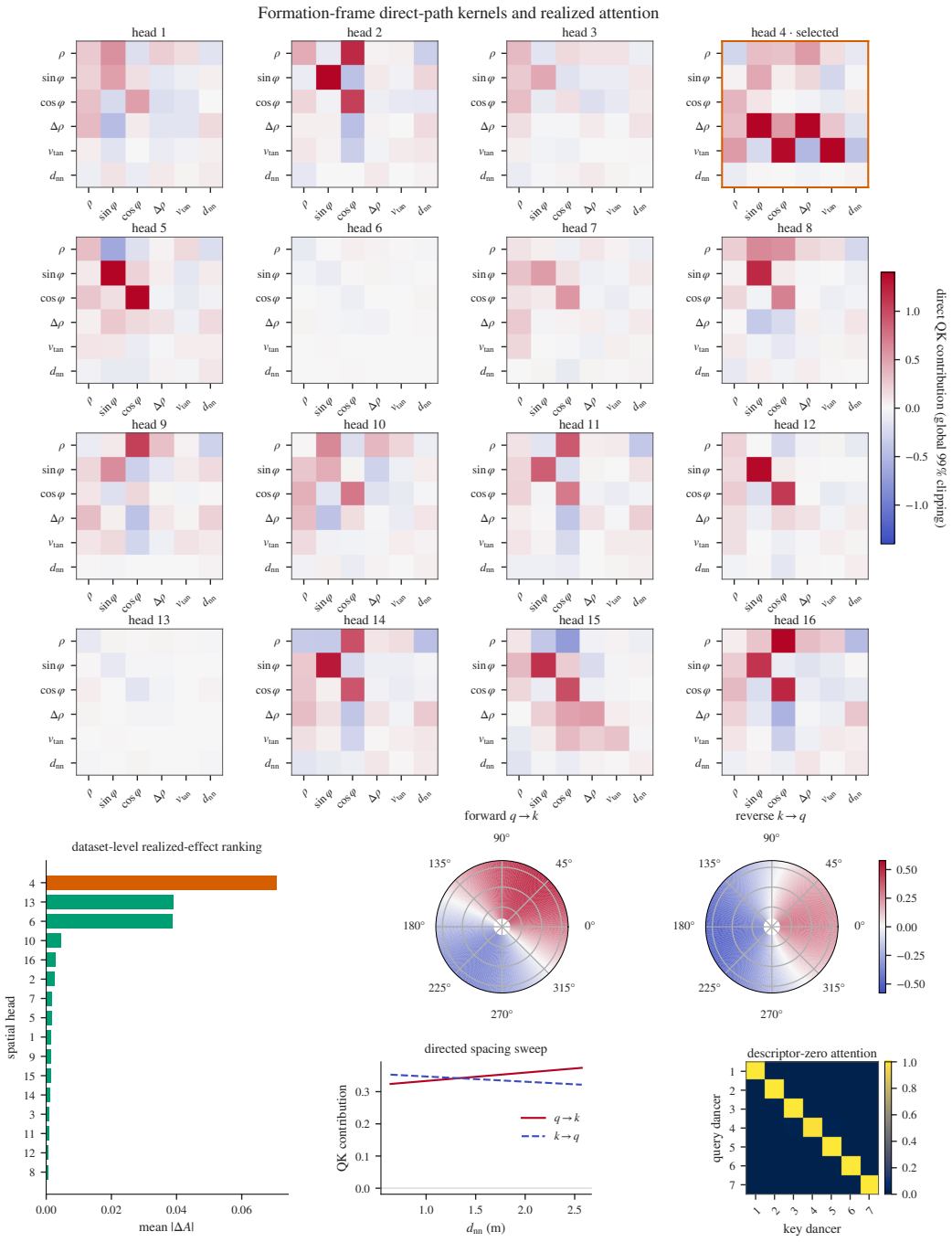


Figure 7: **Complete formation-kernel audit.** All 16 first-layer spatial heads use one shared, zero-centered color scale. The ranking is the mean absolute counterfactual attention change over 401 held-out clips: head 4 is globally first, while rank-2 head 13 is used in Figure **??** because it yields the selected persistent two-subgroup transition. The lower panels report both query–key directions, the nearest-neighbor sweep, and descriptor-zero attention for the globally top head on the median-response $N = 7$ frame.

G.1 MASKED SAMPLING

Given the user-defined cells, the sampler (i) draws a random commitment order over the masked cells, (ii) for iteration $\kappa = 1, \ldots, K$ sets the number of cells to remain masked after the iteration from the cosine schedule, (iii) recomputes the formation descriptors from the root states of all currently visible cells, (iv) runs the transformer once with and once without the music condition, (v) samples with the diffusion head exactly the cells that leave the masked set at this iteration, using classifier-free guidance on the head's velocity field, and (vi) writes them into the tensor. Supplied cells are never resampled.

G.2 CONTINUATION

Given music of arbitrary length, the model generates the first 8 second window from a fully masked tensor and then advances the window by 4 second; in each subsequent window the first $k = 30$ steps are supplied from the end of the previous window and the remaining steps are masked, the pattern of the temporal block variant. The supplied cells carry both the motion and the root states, so the formation is continued rather than re-planned, and the windows are stitched at the overlap.

G.3 CAST EXTENSION AND SLOT ORDERING

Cast extension supplies the rows of the existing N dancers and masks $N' - N$ appended rows, $N' \leq N_{\max}$, the pattern of the spatial block variant; the descriptors of the appended dancers are zero until their root states are committed, after which they enter the frame. Slots are canonicalized during training by the polar angle of each dancer about the centroid of the first step, and the block variant always masks the trailing slots, so the appended rows occupy the slots that carry the largest angular rank in the training distribution. Because the mask families factorize over the two axes, extension can be invoked inside a continuing performance by supplying the overlap steps of the existing rows and masking the appended rows of the new window.

H DATASET AND EVALUATION SPLITS

This section details the dataset preprocessing and evaluation splits for the experiments in Section 4. All experiments use AIOZ-GDANCE (Le et al., 2023b), a corpus of in-the-wild group dance with paired music and estimated group motion for casts of 2–7 dancers. We preserve the official split assignments for sequences retained after preprocessing.

Dataset preprocessing. The monocular SMPL reconstructions contain identity swaps, inaccurate depth, ground penetration, floating, and foot sliding. We recover source videos from motion identifiers and retain 1,556 of 1,624 clips, excluding 58 with unavailable videos and 10 that cannot be aligned with their motions. Camera calibration, 2D joints, dancer segmentation, and monocular metric depth guide four correction steps. **(1)** We repair identity swaps when mesh occlusion order disagrees with the video and reliable evidence supports relabeling. **(2)** We adjust each dancer's root depth along its viewing ray using contact-anchored depth estimates, while suppressing shared scale drift and limiting correction magnitude and temporal variation. **(3)** We robustly fit a shared ground plane to foot contacts across dancers, accepting it only when contact support and camera gravity agree. **(4)** We apply foot inverse kinematics to improve grounding, rigidly transform all dancers into a common floor-aligned frame, and lock planted feet to reduce sliding. Each step leaves the motion unchanged when evidence is insufficient. We will open-source the preprocessing code upon acceptance.

I EVALUATION METRICS

This section defines the individual-level and group-level metrics reported in Section 4. Each clip has 240 frames at 30 fps and 24 SMPL joints. Let $\mathbf{X}_{c,i,t,j}$ denote the world-space position of joint j of dancer i at frame t in clip c, measured in meters, and let $\Delta t = 1/30\,\mathrm{s}$.

Single-Choreo FID. SC-FID compares individual motion distributions using a 72-dimensional descriptor per dancer, comprising 36 kinetic and 36 pose statistics. Kinetic quantities are joint-position magnitudes, linear and angular velocity and acceleration magnitudes, and jerk magnitudes; pose quantities are torso orientation, pelvis height, two intra-limb angles, and two foot-contact ratios. We record the mean, standard deviation, minimum, maximum, and lower/upper quartiles of each quantity. Derivatives use finite differences at interval Δt. We fit empirical means $\boldsymbol{\mu}$ and covariances Σ to generated and reference descriptors and compute

$$\mathrm{FID} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|_2^2 + \mathrm{tr}\Big(\Sigma_g + \Sigma_r - 2(\Sigma_g\Sigma_r)^{1/2}\Big), \tag{8}$$

where g and r denote generated and reference motions.

Beat Alignment Score. Musical beats $\mathcal{B}$ come from the librosa onset envelope; motion beats $\mathcal{M}_i$ are local minima of average joint speed. BAS (Li et al., 2021; Tseng et al., 2023) averages the following score over N-dancer clips with tolerance Δt,

$$\mathrm{BAS} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{M}_i|}\sum_{m\in\mathcal{M}_i}\exp\!\left(-\frac{\min_{b\in\mathcal{B}}|m-b|^2}{2\Delta t^2}\right). \tag{9}$$

Group-Choreo FID. GC-FID applies Eq. (8) to a 75-dimensional descriptor per group clip. Its 60 kinetic features pool the componentwise mean, standard deviation, and maximum of 20 per-dancer quantities, including energy, speed, jerk, and inter-dancer kinetic correlations. Its 15 spatial features temporally summarize planar-root geometry through convex-hull area, centroid speed, nearest-neighbor distance statistics, principal-axis ratio, angular spread, and frame-to-frame changes. Nearest-neighbor distances contribute their mean, standard deviation, minimum, and maximum. The descriptor thus captures both coordinated movement and formation geometry.

Diversity. For each of C music conditions with N_c dancers, we generate R samples and extract 72-dimensional kinetic features $\mathbf{k}_{c,i}^{(r)}$ for dancer i in sample r. Div averages their pairwise distances,

$$\mathrm{Div} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c}\frac{2}{R(R-1)}\sum_{1\le r<s\le R}\|\mathbf{k}_{c,i}^{(r)} - \mathbf{k}_{c,i}^{(s)}\|_2. \tag{10}$$

Inter-dancer collision rate. Each dancer is approximated by a vertical capsule centered at its root, with radius 0.20 m and total height 1.75 m. Let $I_{c,t}$ indicate whether any pair intersects, determined by a central-segment distance of at most 0.40 m. Coll. reports the percentage of collision frames,

$$\mathrm{Coll.} = 100\,\frac{\sum_{c=1}^{C}\sum_{t=1}^{T_c} I_{c,t}}{\sum_{c=1}^{C} T_c}, \tag{11}$$

where T_c is the clip length and each frame is counted once.

Foot sliding distance. FSD averages planar heel/toe displacement during predicted ground contact, in millimeters. With binary contact labels $a_{c,i,t,j}$ and ground-plane projection P_{xy}, we sum over all clips, dancers, heel/toe joints, and adjacent frame pairs,

$$\mathrm{FSD} = 1000\,\frac{\sum_{c,i,j,t} a_{c,i,t,j}\,\big\|P_{xy}\big(\mathbf{X}_{c,i,t+1,j} - \mathbf{X}_{c,i,t,j}\big)\big\|_2}{\sum_{c,i,j,t} a_{c,i,t,j}}. \tag{12}$$

J Extended Related Work

This section extends the discussion in Section 2 with comparisons of group choreography capabilities, editing and control in single-dancer models, and attention over generated coordinates.

Group choreography, capability by capability. GroupDancer (Wang et al., 2022) and GDanceR (Le et al., 2023b) generate a fixed cast in one pass and coordinate dancers with group

embeddings; GCD (Le et al., 2023a) adds a contrastive objective that trades unison against diversity. TCDiff (Dai et al., 2025b) generates trajectories in a separate stage and conditions the motion stage on them; TCDiff++ (Dai et al., 2025a) folds both into one stage with positioning embeddings, a distance-consistency loss and a long-sequence sampling scheme. CoDancers (Yang et al., 2024) decomposes a group into choreographic units and generates dancers one after another conditioned on the previous ones, which admits any cast size and is the closest precedent to our cast extension; the difference is that X-Dancers generates or extends all dancers jointly under bidirectional context, keeps the existing rows fixed, and uses the same sampler for every other operation. FreeDance (Zhao et al., 2025) handles a free number of dancers with discrete tokens, and Le et al. (2024) generate unlimited casts through a phase manifold with constant memory; neither exposes trajectories or supports editing. ST-GDance and ST-GDance++ (Xu et al., 2025a; 2026) target long, collision-free ensembles with distance-aware spatial modeling and sparse temporal attention, and CoheDancers (Yang et al., 2025) decomposes coherence into individual and interaction terms. In X-Dancers, long duration, larger casts, trajectories and edits are all masks over one tensor.

Editing and control in single-dancer models. EDGE (Tseng et al., 2023) edits a single dancer by diffusion inpainting, Lodge (Li et al., 2024a) generates long solo dances coarse to fine, and MMM, MoMask and MaskControl (Pinyoanuntapong et al., 2024; Guo et al., 2024; Pinyoanuntapong et al., 2025) expose in-betweening, body-part editing and spatial control as mask patterns over discrete motion tokens. X-Dancers carries these patterns to a group: masks range over dancers, steps and channels, and the root channel makes trajectories a mask pattern rather than an external controller.

Attention over generated coordinates. Structure-module attention in AlphaFold2 (Jumper et al., 2021), equivariant diffusion for molecules (Hoogeboom et al., 2022) and joint multi-agent trajectory diffusion (Jiang et al., 2023b) all attend over coordinates that are themselves being generated, but in a noisy-state regime in which every coordinate is present at every step. Masked generation has a visible-or-absent regime: the frame must be built from the visible subset alone and mean the same thing under training masks and sampling commitments, which is what the formation frame of Section 3.2.1 provides.