CHOREOSTUDIO: PLANNING WHAT, WHEN, AND WHO IN MUSIC-DRIVEN GROUP CHOREOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Music-driven group dance generation must decide not only how each dancer moves, but also *what* formation the group should form, *when* it should change, and *who* should execute each transition. Existing methods can generate orderly and rhythmically synchronized group motion, but remain limited in formation evolution and inter-dancer interaction. In human choreography, formation transitions often align with changes in high-level musical structure, such as the progression from an introduction to a chorus or from a chorus to a climax. We present **ChoreoStudio**, which makes choreography planning explicit. A music disentangler first decomposes the audio into frame-level rhythmic features and segment-level structural features. Given the segment-level features and sparse labels, a frozen omni-modal model acts as a scriptwriter, producing a global semantic description that guides the Director's choreography planning. Based on this context, the Director produces three coupled interaction decisions: a centroid-relative point set specifies *what* formation to produce, a transition gate determines *when* it should change, and a mobility allocation indicates *who* should execute the transition. The transition gate and mobility allocation are further conditioned on the planned geometric change, coupling formation targets with their temporal and dancer-specific execution. At a finer temporal scale, frame-aligned music features condition the diffusion generator to control detailed body motion and preserve beat-level alignment. The segment-level interaction plan is injected in parallel through frame-aligned instruction tokens and cross-dancer attention biases, coordinating how the generator realizes each transition. Experiments on AIOZ-GDance demonstrate that ChoreoStudio follows the planned interactions while achieving state-of-the-art performance in motion quality, group organization, and formation dynamics.

1 INTRODUCTION

Music-driven group dance generation aims to synthesize coordinated 3D motions for multiple dancers from music. Unlike single-dancer generation, this task must determine not only how each dancer moves, but also how their interactions form an intentional cohesive group performance. Effective choreography assigns dancers distinct spatial roles, organizes their formation over time, and determines when and by whom each transition is executed. Individually plausible and synchronized movement is therefore insufficient on its own to create a cohesive group dance.

Existing group-dance methods coordinate dancers through joint generation (Le et al., 2023b;a), coherence-oriented training objectives (Yang et al., 2024), trajectory conditioning (Dai et al., 2025; 2026), and spatial-temporal interaction modeling (Xu et al., 2025b; 2026). Across these designs, coordination is typically induced within the generator by jointly processing latent pose, trajectory, and interaction features across dancers. Although these mechanisms improve local consistency and group coherence, they leave two key choreography decisions unresolved. First, dancers are exchangeable before generation, whereas choreography assigns asymmetric roles such as leader, flank, or rear dancer. Symmetric cross-dancer attention does not determine which dancer should assume each role, while initializing dancers with independent noise breaks this symmetry arbitrarily rather than through deliberate role assignment. Second, inter-dancer spatial formation is not represented as an explicit object to be composed and assigned. It remains implicit in latent interactions or is replaced by externally specified trajectories, leaving the model without a direct representation of the intended group arrangement.

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.

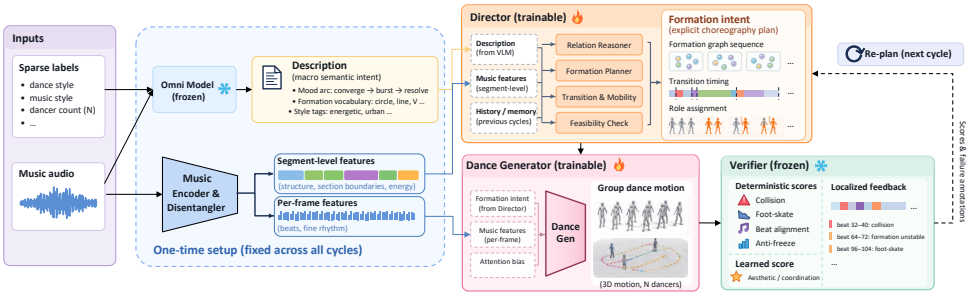


Figure 1: **Overview of ChoreoStudio.** Multi-scale music features and a frozen omni-modal model provide rhythmic, structural, and semantic context. The Director jointly constructs a directed inter-dancer relation graph and a centroid-relative formation point set, then uses the planned geometric change to schedule transition timing and dancer mobility. The resulting plan conditions the Dance Generator through frame-aligned dancer-specific instructions and relation-derived spatial attention biases.

We expose the consequence of this entanglement by analyzing the temporal spectrum of inter-dancer geometry. Decomposing pairwise ground-plane distances into slow formation dynamics and fast residual motion shows that existing generators retain substantial high-frequency activity but markedly underrepresent the low-frequency evolution present in real dances (Appendix A.4). In other words, they can keep dancers moving and locally coordinated without producing sufficiently rich macro-level formation change. This suggests that the key missing capability is not simply stronger cross-dancer interaction or smoother motion, but an *explicit choreography abstraction* between music and motion: a representation that specifies what the group should become, when it should change, and who should execute that change, before deciding how the corresponding full-body motion should look.

Motivated by this observation, we introduce **ChoreoStudio**, which separates choreography composition from motion realization through an explicit, executable plan (Figure 1). The design follows a simple principle: decide *what–when–who* at the choreography level, then synthesize *how* at the motion level. To support these different temporal scales, a learned music disentangler extracts segment-level structural features for planning and frame-level rhythmic features for detailed synthesis; a frozen audio-language model additionally supplies high-level semantic intent. A group-dance Director converts this context into a structured plan containing relative formation geometry, directed dancer relations, transition timing, spatial roles, and mobility allocation. Importantly, these quantities are coupled: the magnitude of the intended geometric change informs whether a transition should occur, and the resulting scheduled displacement informs which dancers should carry it.

To ensure that these decisions are realized in full-body motion, ChoreoStudio uses an explicit plan-to-motion interface. It converts the planned formation, transition timing, and dancer mobility into dancer-specific guidance, so that exchangeable dancers do not need to rediscover their roles and group arrangement during denoising. No unrestricted Director state bypasses the plan, making it the sole choreography-control path to the generator. The plan remains fixed during denoising and can be inspected or edited at inference under fixed music and sampling noise.

We evaluate ChoreoStudio on AIOZ-GDance under a shared 150-frame protocol, with long-sequence results reported separately. Beyond standard motion and group metrics, we measure formation dynamics, plan realization, and the causal response to controlled plan interventions. The results show improved motion quality and group organization over strong baselines, while frequency-based analysis confirms that these gains stem from realistic macro-level formation evolution rather than high-frequency distance fluctuations.

Our contributions are threefold:

- **We reveal a temporal-scale failure in group dance generation.** Frequency decomposition of inter-dancer geometry shows that existing generators retain local motion activity while underrepresenting macro-level formation evolution, exposing a gap unmeasured by conventional collision and correlation metrics.
- **We introduce unified choreographic planning for formation, timing, and mobility.** Our group-dance Director couples a directed inter-dancer relation graph with a geometrically valid

formation point set, then conditions transition timing and dancer mobility on the planned spatial transformation. All plan targets are extracted from motion without manual choreography labels.

- **We make explicit choreography executable and testable.** Frame-aligned dancer instructions and relation-derived attention biases connect the structured plan to full-body generation without an unrestricted hidden-state bypass. Quantitative comparisons and controlled interventions demonstrate improved group organization and verify that generated motion responds to planned formation and timing changes.

2 Related Work

2.1 Music-Driven Single and Group Dance Generation

Single-person methods model music-motion correspondence through cross-modal transformers such as FACT (Li et al., 2021), discrete representations such as Bailando (Siyao et al., 2022), and editable diffusion such as EDGE (Tseng et al., 2023); DuetGen (Ghosh et al., 2025) extends generation to interacting pairs. For larger groups, GDanceR (Le et al., 2023b), GCD (Le et al., 2023a), and Co-Dancers (Yang et al., 2024) improve coordination through joint, contrastive, and coherence-oriented modeling. TCDiff variants (Dai et al., 2025; 2026) introduce trajectory and footwork control, while ST-GDance variants (Xu et al., 2025b; 2026) support scalable long-duration generation. However, formation evolution remains implicit or reduced to low-level trajectories. ChoreoStudio instead jointly plans formation, transition timing, and dancer mobility.

2.2 Multi-Scale Music Conditioning

Music-driven dance models typically use frame-aligned acoustic features for local rhythmic synchronization (Li et al., 2021; Siyao et al., 2022; Tseng et al., 2023), but provide limited guidance over long-term choreography. Hierarchical methods incorporate global context or choreographic primitives (Li et al., 2025a;b), yet mainly organize individual motion. ChoreoStudio instead disentangles music into segment-level and frame-level conditions. A frozen omni-modal language model (Xu et al., 2025a) interprets music segments and sparse labels to guide formation, transition, and mobility planning, while frame-level features preserve detailed beat alignment.

2.3 Structured Planning for Generative Motion

Recent motion systems increasingly separate semantic reasoning from motion synthesis. Language-centered models unify motion understanding, generation, and editing through LLM interfaces (Jiang et al., 2023; Zhou et al., 2024; Wang et al., 2026; Wu et al., 2025b;a; Zhu et al., 2026). Structured-control methods organize individual motion using temporal tracks, atomic action schedules, or anchored motion phrases (Petrovich et al., 2024; Cai et al., 2026; Tang et al., 2026). NEWTON (Feng et al., 2026) further demonstrates planner-generator-verifier orchestration for video synthesis. These methods plan individual motion or generation procedures, whereas ChoreoStudio jointly plans group formation geometry, transition timing, and dancer mobility for coordinated full-body motion.

3 Method

3.1 Overview and Problem Formulation

Given a music sequence and target group size, our goal is to generate a coordinated group dance:

$$\mathbf{X} = \{\mathbf{x}_{i,t}\} \in \mathbb{R}^{N \times L \times D_x}, \tag{1}$$

where N, L, and D_x denote the numbers of dancers, motion frames, and motion features, respectively, and $\mathbf{x}_{i,t}$ contains the root translation, joint rotations, and foot contacts of dancer i at frame t. Rather than mapping music directly to motion, ChoreoStudio introduces an explicit, executable choreography plan between musical interpretation and full-body motion synthesis.

As shown in Figure 2, ChoreoStudio operates in a closed planning–generation–verification loop. A music encoder followed by a trainable music disentangler extracts frame-level rhythmic features $\mathbf{C}_f$ for motion realization and segment-level structural features $\mathbf{C}_s$ for choreography planning. In parallel, the frozen omni-modal model Qwen3-Omni (Xu et al., 2025a) converts the music audio and sparse attributes into a high-level choreography description d, which is embedded as semantic tokens $\mathbf{E}_d$. At refinement cycle r, the Director conditions on $\mathbf{C}_s$, $\mathbf{E}_d$, the current group state $\mathcal{S}^{(r)}$,

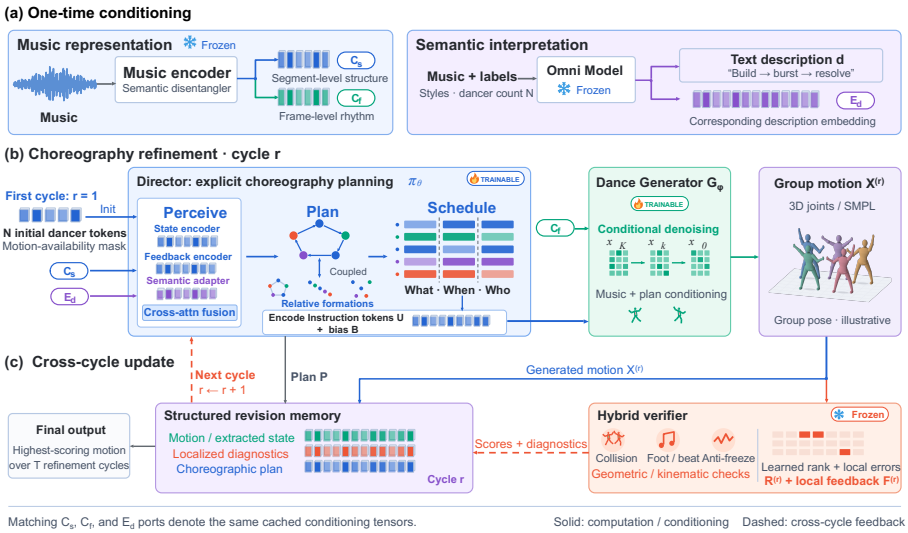


Figure 2: **ChoreoStudio architecture.** Segment-level structure and semantics guide the Director, while frame-level rhythm conditions motion synthesis; a mask marks unavailable motion. Through Perceive, Plan, Schedule, and Encode, the Director couples formation, timing, and mobility, encoding $\mathbf{P}$ as frame-aligned instructions $\mathbf{U}$ and relation-derived cross-dancer attention bias $\mathbf{B}$.

and feedback from the preceding cycle to predict an explicit plan.

$$\mathbf{P}^{(r)} = \left\{ \mathbf{Y}^{(r)}, \mathbf{A}^{(r)}, \boldsymbol{\rho}^{(r)}, \mathbf{g}^{(r)}, \boldsymbol{\mu}^{(r)}, \boldsymbol{\delta}^{(r)} \right\}, \tag{2}$$

where $\mathbf{Y}^{(r)}$ and $\mathbf{A}^{(r)}$ represent relative formations and directed dancer relations, $\boldsymbol{\rho}^{(r)}$ encodes spatial roles, $\mathbf{g}^{(r)}$ schedules transitions, $\boldsymbol{\mu}^{(r)}$ allocates dancer mobility, and $\boldsymbol{\delta}^{(r)}$ controls pose divergence. The plan is encoded into frame-aligned dancer instructions $\mathbf{U}^{(r)}$ and relation-aware attention biases $\mathbf{B}^{(r)}$, which condition the Dance Generator to realize $\mathbf{P}^{(r)}$ as full-body motion $\mathbf{X}^{(r)}$. A frozen Verifier then checks the generated motion for plan consistency, physical plausibility, and music alignment. Localized failures are then fed back to the Director to revise the next plan. Thus, each cycle explicitly separates deciding *what* the group should do from synthesizing *how* the dancers execute it. We omit cycle superscripts in single-cycle derivations.

3.2 Multi-Scale Music Conditioning and Semantic Intent

Learned multi-scale music representations. Let $\mathbf{M} \in \mathbb{R}^{L_m \times D_m}$ denote acoustic features pre-extracted with Librosa (McFee et al., 2015). The music encoder E_{mus} produces a frame-aligned representation:

$$\mathbf{Z} = E_{\text{mus}}(\mathbf{M}) \in \mathbb{R}^{L \times D}. \tag{3}$$

The trainable disentangler D_ω projects $\mathbf{Z}$ into frame-level rhythmic and pooled segment-level structural representations:

$$(\mathbf{C}_f, \mathbf{C}_s) = D_\omega(\mathbf{Z}) = (H_f(\mathbf{Z}), H_s(\text{Pool}_\pi(\mathbf{Z}))), \tag{4}$$

where $\mathbf{C}_f \in \mathbb{R}^{L \times D}$ and $\mathbf{C}_s \in \mathbb{R}^{S \times D}$. Here, π denotes the segment assignment learned from music self-similarity. Unlike fixed temporal pooling, the two representations are learned jointly with the downstream choreography objectives. $\mathbf{C}_s$ captures slowly varying musical structure relevant to formation and transition planning, whereas $\mathbf{C}_f$ retains local rhythmic cues required for detailed motion synthesis and beat alignment.

Controllable macro-level intent. A frozen omni-modal model Φ_{omni} (Xu et al., 2025a) receives the music audio and sparse attributes such as music style, dance style, and group size. It produces a description d of macro-level properties including mood, energy progression, movement intensity, unison, and formation intent:

$$d = \Phi_{\text{omni}}(\mathbf{a}, \mathbf{y}), \tag{5}$$

where $\mathbf{a}$ denotes the music audio and $\mathbf{y}$ the sparse attributes. A frozen Qwen3-Embedding-0.6B text encoder (Zhang et al., 2025) converts d into semantic tokens:

$$\mathbf{E}_d = \Phi_{\text{text}}(d) \in \mathbb{R}^{Q \times D_e}. \tag{6}$$

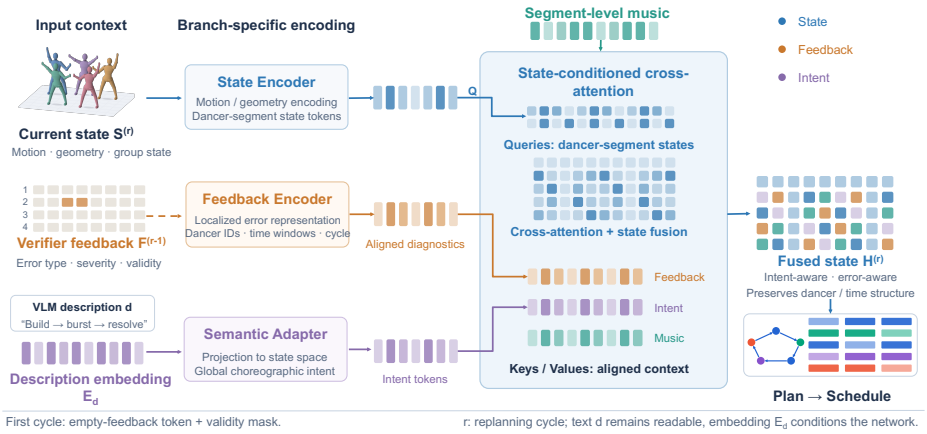


Figure 3: **State-aware perception in the Director.** Exchangeable actor seeds distinguish dancers without assigning roles to their indices, while later cycles use the generated geometry and motion state. State tokens attend to segment-level music, the high-level description, and optional feedback to produce dancer-segment representations for choreography planning.

This yields two complementary planning signals: $\mathbf{C}_s$ describes the learned temporal structure of the music, while $\mathbf{E}_d$ supplies interpretable semantic intent. Both condition the Director, whereas $\mathbf{C}_f$ is reserved for frame-level motion realization.

3.3 Explicit Choreography Director

The group-dance Director maps musical structure, semantic intent, and the current group state to an explicit choreography plan $\mathbf{P}$. Its computation is organized into four stages, i.e., *Perceive*, *Plan*, *Schedule*, and *Encode*, so that spatial organization, transition timing, and dancer participation remain explicit rather than being hidden inside the motion generator.

State-Aware Perception. At cycle $r \in \{1, \ldots, T\}$, the state encoder maps the group state $\mathcal{S}^{(r)}$ and an unknown-motion mask to dancer–segment tokens. Here, $\mathcal{S}^{(r)}$ summarizes available root geometry and motion, while the mask distinguishes valid dancers with unavailable motion from padded dancers. The feedback encoder maps $\mathbf{F}^{(r-1)}$ to diagnostic tokens, and the semantic adapter projects $\mathbf{E}_d$ into the Director space. The state tokens then query these contexts together with segment-level music $\mathbf{C}_s$ through cross-attention. At $r = 1$, all valid dancers are marked motion-unknown and an empty-feedback token is used; for $r > 1$, the generated motion from the previous cycle provides the observable group state.

To distinguish dancers before motion is available while preserving permutation equivariance, we independently sample a centered actor seed for each valid dancer and fix it throughout the sample:

$$\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad \tilde{\mathbf{z}}_i = \mathbf{z}_i - \frac{1}{N} \sum_j \mathbf{z}_j. \tag{7}$$

The seeds permute with the dancers, distinguishing actors without assigning roles to arbitrary dancer indices; the Director derives their context-dependent spatial roles from the planning context. A two-layer MLP state encoder combines the current group state and actor seed:

$$\mathbf{h}_{i,s}^{(r,0)} = E_{\text{state}} \left([\bar{\mathcal{S}}_{i,s}^{(r)}; \tilde{\mathbf{z}}_i] \right), \tag{8}$$

where $\bar{\mathcal{S}}_{i,s}^{(r)}$ replaces unavailable motion with a learned unknown-motion embedding. Actor seeds distinguish exchangeable dancers but encode no predefined roles, which the Director predicts through $\boldsymbol{\rho}^{(r)}$. The tokens query music, semantics, and feedback through cross-attention with an active residual, exposing the Director to these contexts from the start of training:

$$\mathbf{H}^{(r)} \text{ Fusion } \left(\mathbf{H}^{(r,0)}, [\mathbf{C}s; \mathbf{E}d; \mathbf{E}_{\text{fb}}^{(r)}] \right). \tag{9}$$

Relative Formation Planning. For each segment s, the Director constructs a multi-head directed relation graph. The edge from dancer i to j in head h is:

$$A_{ij,s}^{h} = \text{softmax}_j \left(f_h \left[W_q^h \mathbf{h}_{i,s}, W_k^h \mathbf{h}_{j,s}, \mathbf{e}_{ij,s}, \bar{\mathbf{h}}_s \right] \right), \tag{10}$$

where $\mathbf{e}_{ij,s}$ contains available relative geometric features and $\bar{\mathbf{h}}_s$ is the masked group representation. Separate query and key projections allow $\mathbf{A}$ to represent asymmetric relationships. Graph message passing (Gilmer et al., 2017) produces a formation proposal $\tilde{\mathbf{Y}}_{i,s} \in \mathbb{R}^2$ for each dancer. We remove the group centroid and separate formation shape from group scale:

$$\mathbf{Q}_{i,s} = \tilde{\mathbf{Y}}_{i,s} - \frac{1}{N}\sum_j \tilde{\mathbf{Y}}_{j,s}, \tag{11}$$

$$\mathbf{Y}_{i,s} = \alpha_s \frac{\mathbf{Q}_{i,s}}{\sqrt{\frac{1}{N}\sum_j \|\mathbf{Q}_{j,s}\|_2^2 + \epsilon}}, \tag{12}$$

where α_s is the predicted metric RMS centroid radius. Pairwise pairwise formation geometry is:

$$D_{ij,s} = \|\mathbf{Y}_{i,s} - \mathbf{Y}_{j,s}\|_2 . \tag{13}$$

Because pairwise distances are derived from a common 2D point set rather than predicted independently, the resulting distance matrix is geometrically realizable by construction.

We derive a soft spatial-role score from the incoming mass of the directed relation graph:

$$\rho_{i,s} = \frac{1}{H}\sum_h \sum_j A^h_{ji,s} - \frac{1}{N}\sum_k \frac{1}{H}\sum_h \sum_j A^h_{jk,s}. \tag{14}$$

We couple relation reasoning and formation planning instead of treating them as independent branches. After a graph update, the Director drafts a point set and derives its pairwise distances $\tilde{D}_{ij,s}$. These distances are projected back into the edge representation before the next graph update:

$$\mathbf{e}^{(\ell+1)}_{ij,s} = \mathbf{e}^{(\ell)}_{ij,s} + W_D \tilde{D}_{ij,s}. \tag{15}$$

The relation graph therefore reasons about the geometry it proposes, while the final formation depends on the dancer-specific relational context.

Coupled Transition and Mobility Scheduling. A spatial target specifies the desired configuration but not *when* it should change or *which dancers* should carry the transition. We therefore predict a segment-level transition gate:

$$g_s = \sigma\left(f_g[\bar{\mathbf{h}}_s, \Delta_s^{\text{plan}}, \Delta_s^{\text{state}}, \eta_s]\right), \tag{16}$$

where Δ_s^{plan} measures the change between consecutive planned formations, Δ_s^{state} measures the change currently observed in generated motion, and η_s indicates how much motion state is known. The group representation $\bar{\mathbf{h}}_s$ is already conditioned on segment-level music and the high-level description. The gate converts the raw formation proposals into a temporally scheduled plan:

$$\mathbf{Y}^{\text{sch}}_{i,s} = (1-g_s)\mathbf{Y}^{\text{sch}}_{i,s-1} + g_s \mathbf{Y}_{i,s}. \tag{17}$$

The scheduled distances $D^{\text{sch}}_{ij,s}$ and scale α^{sch}_s are recomputed from $\mathbf{Y}^{\text{sch}}_s$. Thus, a low gate preserves the previous relative formation while still allowing local body motion, whereas a high gate activates a formation transition. For an activated transition, the Director predicts the mobility allocation:

$$\mu_{i,s} = \text{softmax}_i\left(f_\mu[\mathbf{h}_{i,s}, \left\|\mathbf{Y}^{\text{sch}}_{i,s} - \mathbf{Y}^{\text{sch}}_{i,s-1}\right\|_2]\right). \tag{18}$$

The resulting $\boldsymbol{\mu}_s$ determines which dancers execute the transition and how movement is distributed among them. Consequently, formation geometry, transition timing, and dancer participation are jointly determined: the proposed geometry affects the gate, and the scheduled displacement in turn determines the mobility allocation.

Explicit Plan Encoding. The generator receives group-level choreography only through the structured plan. Since formation orientation is arbitrary, each dancer instruction encodes rotation-invariant geometry together with the scheduled plan variables:

$$\mathbf{u}_{i,s} = E_{\mathbf{P}}\left(\mathbf{D}^{\text{sch}}_{i,:,s}, \left\|\mathbf{Y}^{\text{sch}}_{i,s}\right\|_2, g_s, \mu_{i,s}, \rho_{i,s}, \alpha^{\text{sch}}_s, \delta_s\right), \tag{19}$$

where δ_s controls inter-dancer pose divergence. The segment-level instructions are linearly interpolated into frame-aligned tokens $\mathbf{U} \in \mathbb{R}^{N\times L\times D_u}$ to avoid boundary discontinuities. The directed relation graph provides an additive bias for cross-dancer spatial attention:

$$B^h_{ij,t} = \kappa_{\pi(t)} \log\left(A^h_{ij,\pi(t)} + \epsilon\right), \tag{20}$$

where κ_s controls interaction strength. The bias affects only spatial attention across dancers, leaving temporal attention and music conditioning unchanged. At cycle r, the two plan conditions are denoted by $\mathbf{U}^{(r)}$ and $\mathbf{B}^{(r)}$. Crucially, these are the only Director outputs exposed to the generator; no unrestricted hidden state bypasses the explicit plan interface. This restriction prevents implicit bypasses, making the plan directly interpretable, editable, and testable.

3.4 Plan-Conditioned Group Dance Generation

Given a fixed plan for the current cycle, the Dance Generator jointly synthesizes all dancers with a conditional diffusion model. At diffusion step k of refinement cycle r, it receives noisy group motion $\mathbf{X}_k^{(r)}$, frame-level rhythmic condition $\mathbf{C}_f$, frame-aligned dancer instruction tokens $\mathbf{U}^{(r)}$, and relation-derived cross-dancer spatial attention bias $\mathbf{B}^{(r)}$:

$$\hat{\mathbf{X}}_0^{(r)} = G_\phi\left(\mathbf{X}_k^{(r)}, k, \mathbf{C}_f, \mathbf{U}^{(r)}, \mathbf{B}^{(r)}\right). \tag{21}$$

The model directly predicts the clean motion, and the completed diffusion trajectory yields candidate $\mathbf{X}^{(r)}$. The instruction tokens modulate every transformer block through token-wise adaptive layer normalization. Meanwhile, $\mathbf{B}^{(r)}$ is added to the logits of cross-dancer attention:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_h}} + \mathbf{B}^{(r)}\right)\mathbf{V}. \tag{22}$$

The two conditioning paths serve complementary roles: $\mathbf{U}^{(r)}$ specifies dancer-wise plan variables, while $\mathbf{B}^{(r)}$ directly shapes cross-dancer information exchange during synthesis.

We sample diffusion levels per dancer and mix in per-token noise levels during training. This exposes the model to partially observed motion and allows known tokens, such as an overlap with a preceding generation window, to remain fixed during long-sequence generation.

3.5 Verification-Guided Refinement

A single Director–Generator pass may fail to realize the intended choreography or may introduce physical artifacts. ChoreoStudio therefore performs up to T closed-loop refinement cycles, using deterministic verification to convert generation failures into structured feedback for replanning. After trajectory r, the frozen Verifier evaluates $\mathbf{X}^{(r)}$ with $\mathbf{P}^{(r)}$ and $\mathbf{C}_f$ using deterministic plan-realization, collision, spacing, dispersion, foot-skating, freezing, and beat-alignment checks:

$$\left(R^{(r)}, b^{(r)}, \mathbf{F}^{(r)}\right) = V\left(\mathbf{X}^{(r)}, \mathbf{P}^{(r)}, \mathbf{C}_f\right), \tag{23}$$

where $R^{(r)}$, $b^{(r)}$, and $\mathbf{F}^{(r)}$ denote the aggregate reward, pass flag, and localized feedback. The next-cycle inputs are:

$$\mathbf{E}_{\mathrm{fb}}^{(r+1)} = E_{\mathrm{fb}}\left(\mathbf{F}^{(r)}\right), \qquad \mathcal{S}^{(r+1)} = \mathrm{Summarize}\left(\mathbf{X}^{(r)}\right), \tag{24}$$

which condition the Director when predicting $\mathbf{P}^{(r+1)}$. Refinement therefore changes the explicit plan rather than directly editing the generated motion. If a candidate passes, refinement may terminate early; otherwise, after at most T cycles, we return the highest-reward candidate, prioritizing passed candidates when available. Verifier details are provided in Appendix B.4.

3.6 Training Objectives

All supervision for the explicit plan is derived automatically from ground-truth motion, requiring no manual choreography annotation. We match the exchangeable actor seeds to target trajectories once per sequence and preserve this assignment across segments, preventing arbitrary dancer order from defining roles. To avoid a moving-target problem between planning and generation, training decouples plan prediction from plan execution. The Dance Generator is trained to realize a teacher plan extracted from ground-truth motion, whereas the Director learns to recover that plan from segment-level music, semantic intent, and the available group state.

$$\mathcal{L} = \mathcal{L}_{\mathrm{motion}} + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}} + \lambda_{\mathrm{real}}\mathcal{L}_{\mathrm{real}} + \lambda_{\mathrm{space}}\mathcal{L}_{\mathrm{space}} + \lambda_{\mathrm{jump}}\mathcal{L}_{\mathrm{jump}}. \tag{25}$$

Here, $\mathcal{L}_{\mathrm{motion}}$ supervises motion reconstruction and dynamics, $\mathcal{L}_{\mathrm{dir}}$ plan prediction, and $\mathcal{L}_{\mathrm{real}}$ plan realization, while $\mathcal{L}_{\mathrm{space}}$ and $\mathcal{L}_{\mathrm{jump}}$ discourage formation collapse and implausible root displacement, respectively. Full definitions are provided in Appendix B.5.

Table 1: **Single and Group dance performance on AIOZ-GDance.** Best results are **bold**; second-best results are underlined.

Method	1 Dancer			2 Dancers			3 Dancers			4 Dancers		
	FID↓	Div↑	PFC↓	GMR↓	GMC↑	TIF↓	GMR↓	GMC↑	TIF↓	GMR↓	GMC↑	TIF↓
GCD Le et al. (2023a)	39.24	9.64	2.53	34.39	80.32	0.17	30.22	80.22	0.19	36.28	81.82	0.13
CoDancers Yang et al. (2024)	23.98	9.48	3.53	24.55	72.52	**0.08**	26.34	74.22	**0.08**	26.44	75.34	0.10
TCDiff++ Dai et al. (2026)	22.42	**17.43**	1.32	15.77	80.87	0.11	11.57	82.31	0.10	13.63	81.56	0.13
ST-GDance Xu et al. (2025b)	28.87	12.82	0.97	19.42	80.52	0.12	14.76	81.24	0.11	14.02	81.42	0.10
ChoreoStudio	**22.34**	14.35	**0.52**	**13.56**	**81.64**	0.10	**10.42**	**83.23**	**0.08**	**11.94**	**82.41**	**0.09**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate ChoreoStudio on AIOZ-GDance (Le et al., 2023b), using its official split of 1,309 training and 154 test sequences. A single model is trained on all group sizes from two to seven dancers. Full data processing and motion representations are provided in Appendix B.

Baselines and protocol. We compare with GCD (Le et al., 2023a), CoDancers (Yang et al., 2024), TCDiff (Dai et al., 2025), TCDiff++ (Dai et al., 2026), and ST-GDance (Xu et al., 2025b). All methods use the same test identities, sequence length, motion conversion, feature extractors, and evaluation code; published scores from incompatible protocols are not mixed with our results.

Metrics. We report motion fidelity and diversity, PFC (Physical Foot Contact) (Tseng et al., 2023), GMR (Group Motion Realism), GMC (Group Motion Correlation), TIF (Trajectory Intersection Frequency). IDDV (Inter-Dancer Distance Variance), Jitter, and FMR (Formation Meaningfulness Ratio) measure the low- and high-frequency components of formation dynamics. BAS (Beat Alignment Score) (Siyao et al., 2022) measures the temporal alignment between kinematic beats and musical beats. Complete definitions are in Appendix A.

Implementation. Main results use the 150-frame setting, with a 720-frame evaluation in Appendix C.2. The same ChoreoStudio model is evaluated at both lengths and across all dancer counts without retraining, whereas each baseline is trained separately for each group size. During inference, the Director revises its plan using localized feedback until a valid candidate reaches the Verifier score threshold or the cycle budget is exhausted; the highest-scoring candidate is returned. Optimization, sampling, refinement budgets, and hardware settings are provided in Appendix B.

4.2 COMPARISON WITH EXISTING METHODS

Quantitative comparison. Table 1 reports single- and multi-dancer performance under the shared 150-frame protocol. ChoreoStudio attains the best GMR and GMC at every group size, indicating that explicit planning improves group-level organization rather than only local motion quality, and the margin does not degrade as the group grows. Prior methods trade these properties against each other: CoDancers achieves low collision rates but the weakest group realism and correlation, consistent with avoiding intersections by keeping dancers apart, whereas TCDiff++ and ST-GDance improve realism without closing the gap in coordination. ChoreoStudio instead improves both simultaneously while achieving the best or second-best TIF across group sizes, showing that formation planning does not compromise collision avoidance.

Formation dynamics and music alignment. As shown in Table 2, ChoreoStudio consistently narrows the gap to real data across all four metrics. Compared with TCDiff++ (Dai et al., 2026), the strongest overall baseline, it recovers substantially more slow-timescale formation variation and stronger beat alignment while reducing high-frequency residuals. Its FMR remains close to the real-data reference, indicating that the recovered variation reflects structured formation evolution rather than rapid fluctuation.

Table 2: **Formation dynamics and music alignment score relative to real data.**

	IDDV	Jitter	FMR	BAS
GT	0.302	0.045	0.832	0.363
Best baseline	0.181	0.051	0.717	0.224
ChoreoStudio	0.267	0.048	0.848	0.351

Qualitative comparison. Figure 4 compares ST-GDance (Xu et al., 2025b), TCDiff (Dai et al., 2025), and ChoreoStudio from top to bottom under the same music. ST-GDance produces body overlap during position exchanges despite graph-based interaction modeling, while TCDiff exhibits foot-ground penetration and arm interpenetration. By planning formation targets, transition timing, and dancer-specific mobility before motion synthesis, ChoreoStudio coordinates spatial exchanges more reliably and produces stable full-body motion. Additional results are provided in Appendix D.

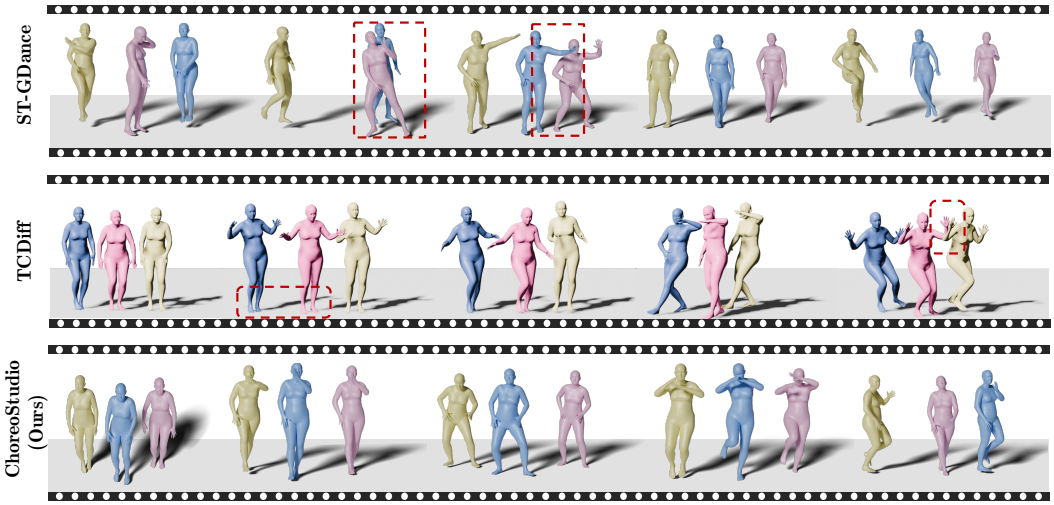


Figure 4: **Qualitative comparison on matched music segments.** We compare local motion and long-timescale formation evolution across methods using the same music and temporal extent.

4.3 Ablation Studies

All ablations degrade performance. Removing the Director causes the largest drops in GMC and FMR and the highest TIF, supporting explicit choreography planning. The semantic description improves group-motion realism, while the Verifier strengthens coordination and formation dynamics. Among the structured decisions, transition timing most affects realism and coordination, role assignment contributes most to slow-timescale formation variation, and the relation graph provides spatial organization. Together, these results support jointly planning *what* changes, *when* it changes, and *who* executes the transition.

Table 3: **Ablation study for group dance generation (3 dancers).**

	GMR↓	GMC↑	TIF↓	FMR↑
ChoreoStudio	**10.42**	**83.23**	**0.08**	**0.848**
w/o Semantic Description	17.68	79.34	0.09	0.819
w/o Verifier	16.10	78.68	0.10	0.803
w/o Director	15.23	78.34	0.11	0.783
w/o Role Assignment	12.52	81.53	0.09	0.795
w/o Transition Timing	14.32	79.32	0.10	0.803
w/o Relation Graph	13.68	80.93	0.09	0.819

4.4 Plan Realization and Controlled Intervention

We perform paired interventions after plan prediction while fixing the music, actor seeds, diffusion noise, and non-intervened variables. Each edit is re-encoded before generation to isolate the plan component. As shown in Table 4, all three edits reduce realization error and change motion predominantly in the requested direction. Formation produces the strongest response, followed by timing, while mobility remains effective despite the constraint of preserving the target formation. The small changes in TIF and Jitter indicate that control is not achieved through motion degradation. Together, these results verify control over what changes, when it changes, and who executes the transition. Direct realization metrics and intervention controls are provided in Table 5 and Appendix C.1.

Table 4: **Controlled plan interventions.** Error reduction measures improvements toward the edited target.

Intervention	Error reduction↑	Dir.↑	ΔTIF↓	ΔJitter↓
Formation geometry	0.40	78.54%	0.014	0.003
Transition timing	0.37	74.85%	0.018	0.004
Dancer mobility	0.29	68.68%	0.015	0.005

5 Conclusion

We introduced ChoreoStudio to reveal and address a temporal-scale failure in group dance generation: locally plausible motion can coexist with impoverished slow formation evolution. Its Director jointly plans realizable relative formations, transition timing, and dancer mobility, while an explicit plan-to-motion interface makes these decisions executable, interpretable, and editable. Experiments demonstrate state-of-the-art motion quality and group organization, more realistic formation dynamics, accurate plan realization, and causal responses to controlled interventions. Future work will incorporate absolute stage orientation and richer semantic roles.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, grammar checking, and assistance with understanding and summarizing literature selected by the authors. These tools were not used for experimental implementation, data collection, or result analysis. All AI-assisted content was verified by the authors, who take responsibility for the final text, claims, and artifacts.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. All datasets used in this work are publicly available with established licenses. The user study involves anonymous participants providing perceptual rankings of generated dance sequences, with no collection of personally identifiable information.

REPRODUCIBILITY STATEMENT

The architecture and training procedure are described in Section 3, and the evaluation protocol is presented in Section 4. Data processing, motion representation, semantic conditioning, state initialization, optimization, and sampling settings are detailed in Appendix B, with complete objective definitions in Appendix B.5. Evaluation metrics and statistical aggregation are specified in Appendix A. Controlled interventions and long-sequence generation are documented in Appendices C.1 and C.2, respectively.

REFERENCES

Xinhao Cai, Yixuan Sun, Minghang Zheng, Qingchao Chen, Xin Jin, Song-chun Zhu, and Yang Liu. Music-to-dance generation via atomic movements. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2026.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2645–2653, 2025.

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. *International Journal of Computer Vision*, 134(2):61, 2026.

Yuxiang Feng, Juncheng Wang, Chao Xu, Yijie Qian, Huihan Wang, Wenlong Hou, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Newton: Agentic planning for physically grounded video generation. *arXiv preprint arXiv:2605.18396*, 2026.

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. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Nhat Le, Tuong Do, Khoa Do, Hien Nguyen, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Controllable group choreography using contrastive diffusion. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023a.

Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8673–8682. IEEE, 2023b.

Ronghui Li, Hongwen Zhang, Yachao Zhang, Yuxiang Zhang, Youliang Zhang, Jie Guo, Yan Zhang, Xiu Li, and Yebin Liu. Lodge++: High-quality and long dance generation with robust choreography patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13381–13392. Ieee, 2021.

Yixuan Li, Qiang Jin, Huaping Liu, Jinhai Chen, Xiangyu Zhao, and Peng Li. Melodance: Dance generation guided by music structure and emotion. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 671–679, 2025b.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, et al. librosa: Audio and music signal analysis in python. *SciPy*, 2015(18-24):7, 2015.

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1921, 2024.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11050–11059, 2022.

Xulong Tang, Kaixing Yang, Xiaohu Guo, Prabhakaran Balakrishnan, and Rawan Alghofaili. Customdance: Customized 3d dance generation with coarse-to-fine human-centered interactive control. *arXiv preprint arXiv:2608.06722*, 2026.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 448–458, 2023.

Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Dan Xu, and Shixiang Tang. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27849–27858. IEEE, 2025a.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with llms. In *International Conference on Learning Representations*, volume 2025, pp. 48023–48043, 2025b.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025a.

Jing Xu, Weiqiang Wang, Cunjian Chen, Jun Liu, and Qiuhong Ke. St-gdance: Long-term and collision-free group choreography from music. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2025b.

Jing Xu, Weiqiang Wang, Cunjian Chen, Jun Liu, and Qiuhong Ke. St-gdance++: A scalable spatial-temporal diffusion for long-duration group choreography. *arXiv preprint arXiv:2603.22316*, 2026.

Kaixing Yang, Xulong Tang, Ran Diao, Hongyan Liu, Jun He, and Zhaoxin Fan. Codancers: Music-driven coherent group dance generation with choreographic unit. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 675–683, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding, planning, generation and beyond. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1357–1366. IEEE, 2024.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. In *International Conference on Learning Representations*, volume 2026, pp. 12631–12657, 2026.

A Analysis and Metric Details

This section provides the definitions and implementation details of all evaluation metrics. Unless otherwise specified, generated and ground-truth motions are converted to the same SMPL-based representation and evaluated with the same feature extractors. Standard motion and group metrics are computed over complete sequences. The frequency-based formation metrics are computed using length-matched temporal windows and then aggregated at the sequence level.

A.1 Motion Quality and Diversity

Let $\phi_q(\mathbf{X})$ denote a motion feature extracted from sequence $\mathbf{X}$, where $q \in \{k, g\}$ represents kinetic or geometric features. We standardize real and generated features using the mean and standard deviation of the real features. Let $(\boldsymbol{\mu}_r^q, \boldsymbol{\Sigma}_r^q)$ and $(\boldsymbol{\mu}_g^q, \boldsymbol{\Sigma}_g^q)$ denote the Gaussian statistics of the real and generated features, respectively. The corresponding Fréchet distance is

$$\mathrm{FID}_q = \left\|\boldsymbol{\mu}_r^q - \boldsymbol{\mu}_g^q\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r^q + \boldsymbol{\Sigma}_g^q - 2\left(\boldsymbol{\Sigma}_r^q \boldsymbol{\Sigma}_g^q\right)^{1/2}\right). \tag{26}$$

Lower FID indicates that the generated feature distribution is closer to the real distribution.

For M generated sequences with standardized features $\{\mathbf{z}_n^q\}_{n=1}^M$, diversity is their average pairwise distance:

$$\mathrm{Dist}_q = \frac{2}{M(M-1)} \sum_{n<m} \left\|\mathbf{z}_n^q - \mathbf{z}_m^q\right\|_2. \tag{27}$$

We report the kinetic and geometric variants, denoted by FID_k/Dist_k and FID_g/Dist_g, respectively.

Physical Foot Contact (PFC) measures foot sliding during body acceleration. Following the standard implementation, we compute the horizontal velocities of the left and right feet, denoted by $v_L(t)$ and $v_R(t)$, and the normalized root acceleration $\tilde{a}(t)$. The sequence-level score is

$$\mathrm{PFC} = \frac{10^4}{L-2} \sum_{t=1}^{L-2} \tilde{a}(t)\, v_L(t)\, v_R(t). \tag{28}$$

For each foot, the smaller velocity of its toe and ankle joints is used. Lower PFC indicates fewer foot-sliding artifacts.

A.2 Group Motion Metrics

Group Motion Realism (GMR) evaluates the discrepancy between real and generated group-level kinetic-feature distributions. It uses the same Fréchet formulation as above, replacing the individual-motion feature ϕ_q with a group kinetic feature extractor ϕ_{grp}.

Group Motion Correlation (GMC) measures motion similarity between dancers. For each dancer pair (i, j), Dynamic Time Warping produces an alignment path $\mathcal{A}_{ij}$ between their joint trajectories. Given frame-level kinetic features $\mathbf{k}_i(t)$ and $\mathbf{k}_j(t')$, GMC is

$$\mathrm{GMC} = \frac{2}{N(N-1)} \sum_{i<j} \frac{1}{|\mathcal{A}_{ij}|} \sum_{(t,t')\in\mathcal{A}_{ij}} \mathbf{k}_i(t)^\top \mathbf{k}_j(t'). \tag{29}$$

Because excessively synchronized or duplicated motions may also produce a large correlation, GMC is interpreted relative to its ground-truth value rather than assuming that an arbitrarily larger value is always preferable.

Trajectory Intersection Frequency (TIF) measures inter-dancer body penetration. We reconstruct the 24 SMPL joints and use the 23 bones defined by the SMPL kinematic tree. Let $\mathcal{B}_i(t)$ denote the bone segments of dancer i at frame t. A dancer pair is considered intersecting if the minimum distance between any two of their bone segments is below a thickness threshold δ:

$$c_{ij}(t) = \mathbb{I}\left[\min_{\substack{b\in\mathcal{B}_i(t)\\ b'\in\mathcal{B}_j(t)}} d_{\text{seg}}(b,b') < \delta\right]. \tag{30}$$

We use $\delta = 0.10\,\text{m}$ in the main experiments and compute

$$\text{TIF} = \frac{1}{L}\sum_{t=1}^{L}\frac{2}{N(N-1)}\sum_{i<j} c_{ij}(t). \tag{31}$$

Thus, TIF is the average proportion of dancer pairs that intersect at each frame. We apply this implementation uniformly to ground truth and all generated results.

A.3 Frequency-Based Formation Dynamics

Static collision and correlation scores do not reveal whether formations evolve meaningfully over time. We therefore decompose the temporal variation of inter-dancer distances into gradual formation change and rapid residual fluctuation.

Consider a group sequence with N dancers and L frames sampled at $f_s = 30\,\text{Hz}$. Let $\mathbf{p}_i(t) = (x_i(t), z_i(t))$ denote the ground-plane root position of dancer i. For each unordered pair (i,j), we construct

$$d_{ij}(t) = \left\|\mathbf{p}_i(t) - \mathbf{p}_j(t)\right\|_2. \tag{32}$$

We apply a fourth-order zero-phase Butterworth low-pass filter with cutoff $f_c = 0.5\,\text{Hz}$:

$$d_{ij}^{\text{low}} = \text{LP}_{f_c}[d_{ij}], \qquad d_{ij}^{\text{high}} = d_{ij} - d_{ij}^{\text{low}}. \tag{33}$$

The cutoff retains formation changes with characteristic periods of roughly two seconds or longer while assigning rapid distance fluctuations to the residual. The same filter is applied to real and generated sequences.

We divide each sequence into non-overlapping windows of $W = 150$ frames. For window w, define

$$S_w = \frac{2}{N(N-1)}\sum_{i<j}\text{Std}_{t\in w}\left[d_{ij}^{\text{low}}(t)\right], \tag{34}$$

$$J_w = \frac{2}{N(N-1)}\sum_{i<j}\sqrt{\frac{1}{|w|}\sum_{t\in w}\left(d_{ij}^{\text{high}}(t)\right)^2}. \tag{35}$$

The final metrics are

$$\text{IDDV} = \frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}} S_w, \qquad \text{Jitter} = \frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}} J_w, \tag{36}$$

and

$$\text{FMR} = \frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}}\frac{S_w}{S_w + J_w + \epsilon}. \tag{37}$$

IDDV measures slow formation evolution, Jitter measures rapid inter-dancer distance fluctuation rather than joint-level pose jitter, and FMR measures the fraction of variation attributable to gradual formation change. All three are interpreted relative to real dance rather than optimized independently.

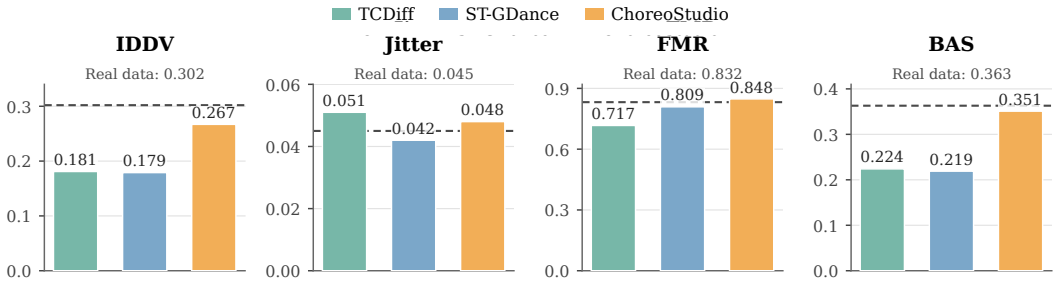


Figure 5: **Formation dynamics and music alignment.** Bars report IDDV, Jitter, FMR, and BAS for TCDiff, ST-GDance, and ChoreoStudio. Dashed lines indicate the corresponding real-data references. All methods are evaluated using the same sequences, temporal windows, and metric implementations.

A.4 Empirical Formation-Dynamics Comparison

We apply the formation-dynamics metrics to ground-truth sequences and representative group-dance generators under the same evaluation protocol. Figure 5 reports the resulting dataset-level statistics together with music alignment.

Both baselines substantially underrepresent slow formation evolution. TCDiff and ST-GDance obtain IDDV scores of 0.181 and 0.179, respectively, compared with 0.302 for real dance. TCDiff additionally exhibits stronger high-frequency variation, with higher Jitter and substantially lower FMR than the real-data references. ST-GDance produces more realistic Jitter and FMR but retains a large IDDV gap. ChoreoStudio narrows this gap, achieving an IDDV of 0.267, Jitter of 0.048, and FMR of 0.848. Its BAS of 0.351 is also close to the real-data value of 0.363, indicating that improved formation evolution is obtained while preserving music alignment.

These metrics are interpreted relative to real data rather than optimized independently. Excessive IDDV may indicate formation drift, while a high FMR does not guarantee sufficient absolute formation change. Moreover, pairwise distances omit formation orientation, dancer identity, and some interaction semantics. The metrics therefore diagnose the temporal organization of inter-dancer geometry rather than provide a complete measure of choreography quality.

A.5 Plan-Realization Metrics

Let $\tilde{\mathbf{Y}}_i(t)$ denote the frame-level formation target obtained by interpolating the Director's segment-level plan, and define the generated root position relative to the group centroid as

$$\bar{\mathbf{p}}_i(t) = \mathbf{p}_i(t) - \frac{1}{N}\sum_j \mathbf{p}_j(t). \tag{38}$$

Because the planned point set has no absolute stage orientation, we align it to the generated formation by a proper 2D rotation before measuring point-wise realization:

$$E_{\text{form}} = \min_{\mathbf{R}\in \mathrm{SO}(2)} \frac{1}{NL}\sum_{i,t} \left\| \bar{\mathbf{p}}_i(t) - \mathbf{R}\tilde{\mathbf{Y}}_i(t) \right\|_2 . \tag{39}$$

The actor-to-target correspondence is fixed by the sequence-level matching used during plan supervision. To evaluate relative geometry independently of translation and rotation, we also report the pairwise realization error

$$E_{\text{pair}} = \frac{2}{N(N-1)L}\sum_{i<j,t} \left| \|\mathbf{p}_i(t) - \mathbf{p}_j(t)\|_2 - \left\| \tilde{\mathbf{Y}}_i(t) - \tilde{\mathbf{Y}}_j(t) \right\|_2 \right| . \tag{40}$$

For the Director, transition decisions are evaluated using segment-level precision, recall, and F1 after selecting a gate threshold on the validation set. For each matched predicted and target transition, we also report the absolute temporal offset in seconds. These metrics distinguish whether the Director

detects a transition from whether it places that transition at the correct time. Mobility allocation is evaluated as the average ℓ_1 distance between the predicted and pseudo-label mobility distributions:

$$E_{\text{mob}} = \frac{1}{S} \sum_{s} \|\boldsymbol{\mu}_s - \boldsymbol{\mu}_s^*\|_1 . \tag{41}$$

Lower values indicate better agreement for formation, pairwise geometry, mobility, and temporal-offset errors; larger transition F1 indicates better transition detection.

Table 5: **Plan realization on the 150-frame three-dancer setting.** Formation and pairwise errors are measured in meters; lower is better except for transition F1.

Method	$E_{\text{form}} \downarrow$	$E_{\text{pair}} \downarrow$	Trans. F1↑	Offset (s)↓	$E_{\text{mob}} \downarrow$
ChoreoStudio	0.13	0.09	0.80	0.32	0.16

As shown in Table 5, ChoreoStudio closely realizes the predicted choreography plan, with low formation and pairwise-geometry errors. The transition F1 of 0.80 and average offset of 0.32 seconds indicate that planned formation changes are generally detected and executed near their scheduled times. The low mobility error further shows that dancer-specific movement responsibilities are preserved during motion generation. Together, these results confirm that the explicit plan is effectively translated into formation geometry, transition timing, and dancer participation.

A.6 Music-Dance Alignment

Music-dance alignment is evaluated in the appendix using Beat Alignment Score (BAS). Motion beats are identified as local minima of the total joint-velocity envelope. Let $\mathcal{B}_x$ and $\mathcal{B}_m$ denote the detected motion beats and input music beats, respectively. BAS is

$$\text{BAS} = \frac{1}{|\mathcal{B}_x|} \sum_{t \in \mathcal{B}_x} \exp\left(-\frac{\min_{b \in \mathcal{B}_m} |t - b|^2}{2\sigma^2} \right), \tag{42}$$

where $\sigma = 2.12$ frames following the standard evaluation protocol. The score is first averaged across dancers and then across test sequences.

A.7 Aggregation and Statistical Analysis

For IDDV, Jitter, and FMR, metrics are first averaged across dancer pairs within each temporal window, then across windows belonging to the same sequence, and finally across test sequences. This prevents longer sequences from receiving disproportionate weight. Comparisons with ground truth use matched sequence identities whenever available.

B Implementation and Training Details

B.1 Data Processing and Motion Representation

We use the official AIOZ-GDance (Le et al., 2023b) training and test splits containing 1,309 and 154 sequences, respectively. Motion is sampled at 30 fps. Each dancer is represented by a 151-dimensional vector containing three-dimensional root translation, 6D rotations for 24 SMPL joints, and four binary foot-contact states. The model is trained on all available group sizes from two to seven dancers. Within each batch, groups are padded to the largest dancer count and a binary validity mask is applied to dancer-wise attention, plan prediction, losses, and evaluation.

Training examples are extracted as 720-frame windows with a stride of 112 frames. The music disentangler identifies variable-length structural segments from the music representation. The Director predicts one choreography state for each learned segment, which is subsequently interpolated to the motion frame rate. For sequences longer than one generation window, inference uses 720-frame windows with a 30-frame overlap. Motion in the overlap is retained as known context during the next sampling window, and duplicated frames are removed after stitching.

B.2 Music and Semantic Conditioning

The acoustic input is a 438-dimensional feature sequence extracted offline at twice the motion frame rate. The music encoder produces frame-aligned rhythmic features for the Dance Generator and aggregates structural context over the learned music segments for the Director. Segment-level structural codes are trained from music self-similarity as described in Section 3.2. Their contribution is evaluated by the *w/o music structure* ablation in Appendix C.3.

The frozen omni-modal model receives the music audio and sparse attributes consisting of music style, dance style, and dancer count. Its description is generated once per sequence, encoded into semantic tokens, and cached before training. Gradients are not propagated through either the omni-modal model or its text encoder. The description supplies macro-level semantic intent, while transition timing and detailed motion remain grounded in the learned music features.

B.3 Plan Targets and State Initialization

Pseudo-label plans are extracted from ground-truth root trajectories without manual choreography annotation. Root positions are centered by the group centroid within each segment to define relative formation targets. Scale-normalized changes in pairwise geometry determine transition labels, while normalized dancer displacements define mobility targets. Complete definitions are provided in Appendix B.5.

During training, the Director alternates between states derived from ground-truth motion and initial states in which generated motion is unavailable. In the latter case, independently sampled actor seeds distinguish otherwise identical dancer tokens. The seeds are centered over valid dancers, permuted with the dancer dimension, and fixed across all segments, generation windows, and refinement cycles of the same sample. Padded dancers are excluded from seed centering and all subsequent plan operations. Because the seeds assign no semantics to dancer indices, we perform one sequence-level bipartite matching between predicted actor plans and target dancer trajectories and preserve the resulting correspondence over time.

B.4 Verifier Implementation

The frozen Verifier is applied once after each complete diffusion trajectory, rather than at individual denoising steps. It receives the generated motion $\mathbf{X}^{(r)}$, plan $\mathbf{P}^{(r)}$, and frame-level rhythmic condition $\mathbf{C}_f$. World-coordinate joints $\mathbf{J} \in \mathbb{R}^{N\times L\times J\times 3}$ are recovered through forward kinematics, with the root given by $\mathbf{J}_{:,:,0,:}$.

Plan, physical, and coordination checks. The Verifier first measures plan realization using the discrepancy between the scale-normalized pairwise geometry of generated roots and the scheduled plan:

$$c_{\text{plan}} = \frac{1}{L}\sum_{t=1}^{L}\left[\frac{1}{N(N-1)}\sum_{i\neq j}\left|\widehat{D}^{X}_{ij,t} - \widehat{D}^{P}_{ij,t}\right| + \lambda_{\alpha}\left|\alpha^{X}_{t} - \alpha^{P}_{t}\right|\right]. \tag{43}$$

Here, α^{X}_{t} and α^{P}_{t} are the corresponding RMS centroid radii. It additionally applies six deterministic checks. Capsule collision models each skeleton bone as a capsule of radius $0.09\,\text{m}$ and measures inter-dancer penetration. Root spacing penalizes pairwise root distances below $0.6\,\text{m}$. Group dispersion penalizes distances above $2.0\,\text{m}$, preventing collision avoidance through excessive separation. This upper bound is calibrated from real data, whose mean pairwise distance is $1.55\,\text{m}$ and 90th percentile is approximately $2.0\,\text{m}$.

Foot skating is measured from horizontal foot velocity during contacts, where contact requires both foot height and speed below 0.05. The floor height is estimated from the fifth percentile of foot heights in each sequence rather than fixed globally. Anti-freeze detects dancers whose root and root-relative joint velocities remain below $0.02\,\text{m/s}$ within a one-second window. Beat alignment compares musical beats with motion accents detected from joint acceleration, using a Gaussian tolerance of three frames.

Reward and validity. The deterministic scores are aggregated as

$$R^{(r)} = c_{\text{beat}} - c_{\text{plan}} - 2c_{\text{coll}} - c_{\text{foot}} - 2c_{\text{freeze}} - c_{\text{space}} - c_{\text{disp}}. \tag{44}$$

The hard validity flag is computed independently:

$$b^{(r)} = \mathbb{I}\left[c_{\text{coll}} \le 0.02,\ c_{\text{freeze}} \le 0.25,\ c_{\text{space}} \le 0.05,\ c_{\text{disp}} \le 0.15\right]. \tag{45}$$

Thus, $R^{(r)}$ ranks candidates, whereas $b^{(r)}$ rejects motions with severe physical or spatial violations.

Localized feedback. Each violation records its type, normalized start and end times, severity, and affected dancers. Events are ranked by severity, and at most eight are retained per sample. They are encoded as event features, dancer masks, and verifier sub-scores to form $\mathbf{F}^{(r)}$. The dispersion cost affects reward and candidate selection but is not appended to the feedback score vector, preserving compatibility with the trained feedback encoder; plan mismatches enter through localized event features.

Feedback-branch training. During training, the frozen Verifier evaluates completed motion samples and converts localized violations into feedback tokens. These tokens and the summarized motion state condition a subsequent Director prediction, which is supervised by the same plan targets through $\mathcal{L}_{\text{dir}}$. Gradients update the feedback encoder and Director but do not pass through the Verifier or diffusion sampling. Empty-feedback examples are retained to train the first cycle.

Candidate selection. The refinement memory stores the plan, motion, reward, validity, and feedback from every completed cycle. Refinement terminates once a valid candidate reaches the validation-set reward threshold τ_R or the maximum of T cycles. The system returns the highest-reward passed candidate among the completed cycles; if none passes, it returns the highest-reward candidate with its diagnostics. Main results use this multi-cycle selection protocol.

B.5 Detailed Objective Definitions

For each segment s, ground-truth root positions are averaged temporally and centered by the group centroid to obtain $\mathbf{Y}^*_{i,s}$. The target pairwise-distance matrix is $D^*_{ij,s} = \|\mathbf{Y}^*_{i,s} - \mathbf{Y}^*_{j,s}\|_2$. We define the target scale using the same RMS centroid radius as the Director,

$$\alpha^*_s = \sqrt{\frac{1}{N}\sum_i \|\mathbf{Y}^*_{i,s}\|_2^2 + \epsilon}, \tag{46}$$

and $\hat{D}^*_{ij,s} = D^*_{ij,s}/(\alpha^*_s + \epsilon)$ defines scale-free formation shape.

Actor seeds do not prescribe a correspondence with dataset dancer indices. We therefore compute one matching over the complete segment sequence using rotation- and permutation-invariant geometric signatures. Let $r_{i,s} = \|\mathbf{Y}_{i,s}\|_2$ and let $\text{sort}(\mathbf{D}_{i,:,s})$ denote the sorted distances from dancer i to the other dancers. The matching cost is

$$C_{ij} = \frac{1}{S}\sum_s \left|r_{i,s} - r^*_{j,s}\right| + \lambda_D \frac{1}{S}\sum_s \left\|\text{sort}(\mathbf{D}_{i,:,s}) - \text{sort}(\mathbf{D}^*_{j,:,s})\right\|_1 + \lambda_{\text{mob}}\frac{1}{S}\sum_s \left|\mu_{i,s} - \mu^*_{j,s}\right|. \tag{47}$$

Hungarian matching produces a permutation Π^* that is held fixed for all segments. All dancer-indexed targets below are reordered by Π^* before computing losses.

The transition target detects a sufficiently large change in scale-normalized pairwise geometry:

$$g^*_s = \mathbb{I}\left[\Delta\left(\mathbf{D}^*_s, \mathbf{D}^*_{s-1}\right) > \tau_g\right], \qquad s \ge 2. \tag{48}$$

The target mobility distribution measures the relative displacement required from each dancer:

$$\mu^*_{i,s} = \frac{\left\|\mathbf{Y}^*_{i,s} - \mathbf{Y}^*_{i,s-1}\right\|_2}{\sum_j \left\|\mathbf{Y}^*_{j,s} - \mathbf{Y}^*_{j,s-1}\right\|_2 + \epsilon}, \qquad s \ge 2. \tag{49}$$

This target captures formation-level travel rather than local in-place body motion.

The Director objective is

$$\mathcal{L}_{\text{dir}} = \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{mob}}\mathcal{L}_{\text{mob}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{rate}}\mathcal{L}_{\text{rate}}. \tag{50}$$

Its components are

$$\mathcal{L}_{\text{shape}} = \frac{1}{N(N-1)S} \sum_{i \neq j,s} \left| \hat{D}_{ij,s} - \hat{D}^*_{ij,s} \right|, \tag{51}$$

$$\mathcal{L}_{\text{scale}} = \frac{1}{S} \sum_{s} \left| \alpha_s - \alpha^*_s \right|, \tag{52}$$

$$\mathcal{L}_{\text{gate}} = -\frac{1}{S-1} \sum_{s=2}^{S} \left[g^*_s \log g_s + (1 - g^*_s) \log(1 - g_s) \right], \tag{53}$$

$$\mathcal{L}_{\text{mob}} = \frac{\sum_{s=2}^{S} g^*_s \left\| \boldsymbol{\mu}_s - \boldsymbol{\mu}^*_s \right\|_1}{\sum_{s=2}^{S} g^*_s + \epsilon}, \tag{54}$$

$$\mathcal{L}_{\text{div}} = \frac{1}{S} \sum_{s} \left| \delta_s - \delta^*_s \right|, \tag{55}$$

$$\mathcal{L}_{\text{rate}} = \left| \frac{1}{S-1} \sum_{s=2}^{S} g_s - \frac{1}{S-1} \sum_{s=2}^{S} g^*_s \right|. \tag{56}$$

We do not directly supervise the directed relation graph $\mathbf{A}$ because no corresponding annotation is available. Transition and mobility targets are defined from $s = 2$ onward; mobility supervision is applied only when $g^*_s = 1$.

Let $\hat{\mathbf{X}}_0$ be the predicted clean motion and $\mathbf{X}_0$ the ground truth. The motion objective is

$$\mathcal{L}_{\text{motion}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}}, \tag{57}$$

where

$$\mathcal{L}_{\text{rec}} = \frac{1}{NL} \sum_{i,t} \left\| \hat{\mathbf{x}}_{i,t} - \mathbf{x}_{i,t} \right\|_2^2, \tag{58}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N(L-1)} \sum_{i,t} \left\| \Delta\hat{\mathbf{x}}_{i,t} - \Delta\mathbf{x}_{i,t} \right\|_2^2, \tag{59}$$

$$\mathcal{L}_{\text{pos}} = \frac{1}{NLJ} \sum_{i,t,j} \left\| \text{FK}(\hat{\mathbf{x}}_{i,t})_j - \text{FK}(\mathbf{x}_{i,t})_j \right\|_2^2, \tag{60}$$

$$\mathcal{L}_{\text{contact}} = \frac{1}{NL|\mathcal{F}|} \sum_{i,t,f} c^*_{i,t,f} \left\| \mathbf{v}^{xz}_{i,t,f} \right\|_2. \tag{61}$$

Here, J is the number of joints, $\mathcal{F}$ is the set of foot joints, $c^*_{i,t,f}$ is the ground-truth contact indicator, and $\mathbf{v}^{xz}_{i,t,f}$ is the generated horizontal foot velocity. The contact term is evaluated on ground-truth support frames.

To supervise execution of the conditioning plan, we pool the generated root trajectories to the same segment grid, remove the group centroid, and compute their scale-free pairwise shape $\hat{D}^{\text{gen}}_{ij,s}$ and scale α^{gen}_s. The realization loss is

$$\mathcal{L}_{\text{real}} = \frac{1}{N(N-1)S} \sum_{i \neq j,s} \left| \hat{D}^{\text{gen}}_{ij,s} - \hat{D}^{\text{plan}}_{ij,s} \right| + \lambda_\alpha \frac{1}{S} \sum_{s} \left| \alpha^{\text{gen}}_s - \alpha^{\text{plan}}_s \right|. \tag{62}$$

This objective is invariant to global translation and rotation, matching the information represented by the plan. We additionally penalize insufficient spacing and implausibly large frame-to-frame root displacement:

$$\mathcal{L}_{\text{space}} = \sum_{i \neq j,t} \left[d_{\min} - \left\| \mathbf{r}_{i,t} - \mathbf{r}_{j,t} \right\|_2 \right]_+, \tag{63}$$

$$\mathcal{L}_{\text{jump}} = \sum_{i,t} \left[\left\| \mathbf{r}_{i,t} - \mathbf{r}_{i,t-1} \right\|_2 - v_{\max} \right]_+. \tag{64}$$

All objectives are masked to exclude padded dancers and invalid frames. The music-structure branch uses a separate segment self-similarity objective and is evaluated through its auxiliary ablation.

B.6 Optimization and Sampling

The Director, trainable music-conditioning modules, and Dance Generator are optimized jointly for 55 epochs using AdamW. We use a learning rate of 10^{-4}, weight decay of 0.02, and gradient clipping at 1.0. Training is performed on two NVIDIA RTX 4090 GPUs with a per-GPU batch size of four and four-step gradient accumulation, giving an effective batch size of 32. Model selection and all hyperparameter choices are made on the validation set.

The diffusion process uses 1,000 training noise levels and 50 sampling steps. Conditioning dropout is set to 0.25, and classifier-free guidance uses a weight of 2.0. Unless otherwise stated, the Director predicts one choreography plan before diffusion sampling, and this plan remains fixed throughout the denoising trajectory. Optional refinement predicts a new plan only after a complete motion sample has been evaluated by the verifier. For the main results, refinement runs for up to T cycles with τ_R selected on the validation set, and the highest-reward completed candidate is reported.

The complete objective is defined in Section 3.6. Loss terms for motion reconstruction, temporal velocity, forward-kinematic position, contact, formation realization, spacing, transition scheduling, and mobility are applied only to valid dancers and frames. The final loss weights are selected on the validation set and kept fixed for all test experiments.

Loss weights and initialization. Table 6 reports the exact weights used by the main `magd_full5` model. The motion weights follow the reconstruction configuration inherited from EDGE, except that the foot-contact weight is set to 0.5. Contact supervision is applied only on ground-truth support frames. The music-structure branch is enabled for the main model.

Table 6: **Loss weights for the main ChoreoStudio model.**

Term	Weight	Term	Weight
$\mathcal{L}_{\text{rec}}$	0.636	$\mathcal{L}_{\text{vel}}$	2.964
$\mathcal{L}_{\text{pos}}$	0.646	$\mathcal{L}_{\text{contact}}$	0.5
$\mathcal{L}_{\text{dir}}$	0.3	$\mathcal{L}_{\text{real}}$	1.0
$\mathcal{L}_{\text{space}}$	1.0	$\mathcal{L}_{\text{jump}}$	1.0
$\mathcal{L}_{\text{struct}}$	1.0	$\mathcal{L}_{\text{struct-inv}}$	0.3

Forward-kinematic velocity, acceleration, and feasibility losses are disabled for the reported model. The perception fusion gate is initialized to 0.5, and the initial-state spread is set to 0.875. All remaining settings use the defaults reported in this appendix.

B.7 Length-Matched Evaluation Protocol

All methods are evaluated on the same 154 test identities. Main results use the 150-frame setting, and the separate long-sequence experiment uses 720 frames. Within each setting, all methods share the same temporal support, motion conversion, feature extraction, and metric implementations. Scores copied from incompatible sequence lengths or preprocessing pipelines are not mixed with our results, and each generated sequence is evaluated as a whole rather than as independently scored short clips.

Sequence-level statistics are treated as the independent evaluation units. Pairwise and window-level measurements are first aggregated within each sequence and then averaged across test sequences, preventing longer sequences or larger groups from receiving disproportionate weight. Where uncertainty is reported, we use 95% bootstrap confidence intervals over test sequences. Paired significance tests use matched test sequences and are reported only for completed comparisons.

C Additional Results

C.1 Controlled Plan Interventions

We isolate the causal plan-to-motion path using paired, single-cycle generation on one 150-frame window. Each pair shares the music, sparse labels, initial state, actor seeds, and diffusion noise; verifier replanning, best-of-cycle selection, and window stitching are disabled. An edit is applied after plan prediction, followed only by the deterministic Schedule and Encode operations that reconstruct $\mathbf{U}$ and $\mathbf{B}$.

Formation geometry. We transplant a target shape from a training sequence with the same dancer count, center it, and match its RMS centroid radius to the original plan before replacing the selected segment. Distances, scale, and planned travel are then recomputed, while relation, gate, mobility, role, and divergence variables are fixed. For edited segments $\mathcal{E}$, realization error is

$$E_{\text{form}}(\mathbf{X}, \mathbf{P}') = \frac{1}{|\mathcal{E}|N(N-1)} \sum_{s \in \mathcal{E}} \left\| \widehat{\mathbf{D}}_s^X - \widehat{\mathbf{D}}_s^{P'} \right\|_1, \tag{65}$$

which is invariant to global translation, rotation, and scale. Uniform scale edits provide a dose–response check, while a 90° rotation is an invariance control.

Transition timing. We shift the selected gate event by an integer number of segments and recompute the scheduled formation chain. The requested frame shift is $\Delta t = \Delta s \, L_s$, where L_s is obtained from the model's segment map. For generated motion, we smooth the frame-wise change in scale-free pairwise geometry and detect the dominant event within a local window around the transition. For edited target time τ', the timing error is

$$E_{\text{time}}(\mathbf{X}, \tau') = \frac{|\widehat{\tau}(\mathbf{X}) - \tau'|}{L}. \tag{66}$$

Events with insufficient prominence or peaks at the search-window boundary are marked invalid rather than treated as zero response.

Dancer mobility. We swap the complete mobility rows of the dancers with the highest and lowest allocations, preserving the formation, gate, relation graph, and mobility multiset. Let $\bar{\mathbf{r}}_{i,t} = \mathbf{r}_{i,t} - N^{-1} \sum_j \mathbf{r}_{j,t}$ be the centroid-relative root. Over transition interval $\mathcal{T}$, the realized mobility share is

$$q_i(\mathbf{X}) = \sum_{t \in \mathcal{T}} \|\bar{\mathbf{r}}_{i,t} - \bar{\mathbf{r}}_{i,t-1}\|_2, \qquad \widehat{q}_i(\mathbf{X}) = \frac{q_i(\mathbf{X})}{\sum_j q_j(\mathbf{X}) + \epsilon}, \tag{67}$$

with error

$$E_{\text{mob}}(\mathbf{X}, \boldsymbol{\mu}') = \frac{1}{2} \|\widehat{\mathbf{q}}(\mathbf{X}) - \boldsymbol{\mu}'\|_1. \tag{68}$$

For $k \in \{\text{form}, \text{time}, \text{mob}\}$, error reduction is

$$\text{ER}_k = \frac{E_k(\mathbf{X}) - E_k(\mathbf{X}')}{E_k(\mathbf{X}) + \epsilon}, \tag{69}$$

where both errors are measured against the edited target. Direction reports the fraction of valid pairs changing as requested. Identical-plan no-op and rotation-invariance controls must reproduce the baseline output within numerical tolerance before a run is accepted. We additionally report paired changes in TIF and Jitter to detect physical or temporal degradation.

As shown in Table 4, editing each plan component produces the intended change in generated motion. Formation intervention yields the strongest response, reducing realization error by 40% and changing the output in the requested direction for 78.54% of the samples. Transition timing remains similarly controllable, while mobility intervention produces a weaker but still consistent response, reflecting the greater difficulty of reallocating movement while preserving the formation target. Across all interventions, the small changes in TIF and Jitter indicate that the requested behavior is realized without materially degrading physical plausibility or temporal stability. These paired results verify that the explicit plan controls what formation is formed, when it changes, and which dancers execute the transition.

C.2 LONG-SEQUENCE GENERATION

We separately evaluate 720-frame generation using the same ChoreoStudio checkpoint as in the main experiments, without length-specific retraining. TCDiff++ is likewise evaluated from its existing checkpoint, whereas the remaining baselines are retrained for the target length. ChoreoStudio also uses one checkpoint across all evaluated group sizes.

As shown in Table 7, ChoreoStudio achieves the best performance across all metrics for one and two dancers. As the group size increases, it maintains the highest GMC and lowest TIF for both three and four dancers, although TCDiff++ obtains slightly lower GMR. This trade-off indicates that ChoreoStudio prioritizes stable cross-dancer coordination and collision control while retaining competitive motion realism. Overall, one checkpoint generalizes across sequence lengths and group sizes without length-specific or group-size-specific retraining.

Table 7: **Dance generation performance in the 720-frame long-term setting.** Best results are **bold**; second-best results are underlined.

Method	1 Dancer			2 Dancers			3 Dancers			4 Dancers		
	FID↓	Div↑	PFC↓	GMR↓	GMC↑	TIF↓	GMR↓	GMC↑	TIF↓	GMR↓	GMC↑	TIF↓
GCD Le et al. (2023a)	51.25	7.24	3.52	38.43	80.76	0.21	40.68	79.25	0.28	42.52	80.32	0.20
CoDancers Yang et al. (2024)	43.62	5.48	4.23	31.67	68.32	0.12	35.20	70.53	0.14	36.32	72.52	0.13
TCDiff++ Dai et al. (2026)	22.47	16.19	1.53	16.52	79.63	0.12	**14.67**	81.64	0.15	**13.78**	80.43	0.14
ST-GDance Xu et al. (2025b)	31.42	14.35	3.21	32.56	75.32	0.14	34.25	74.42	0.15	33.61	76.26	0.13
ChoreoStudio	**21.52**	**16.32**	**1.03**	**14.48**	**83.42**	**0.11**	15.63	**82.64**	**0.12**	14.58	**81.37**	**0.12**

C.3 ADDITIONAL PLANNING DIAGNOSTICS

Table 8 examines whether the gains arise from explicit choreography planning rather than generic interaction modeling. The *relation-only* variant removes formation targets, transition timing, and mobility allocation while retaining directed dancer relations. *Latent bypass* replaces the structured plan-to-motion interface with unrestricted Director features. The remaining variants remove either relation-formation feedback or segment-level music structure.

All variants degrade group-motion realism and coordination. Relation-only reasoning therefore cannot replace explicit formation and scheduling decisions, while the latent-bypass result supports routing choreography through the structured interface. Removing relation-formation feedback causes the largest overall degradation, particularly in FMR, confirming the importance of mutually refining relational reasoning and realizable geometry. Removing music structure mainly reduces GMR and FMR, indicating that segment-level musical context is particularly important for macro-level formation evolution.

Table 8: **Additional planning diagnostics.** Results are evaluated using the same 150-frame three-dancer protocol as the main experiments. FMR is interpreted by proximity to the real-data reference.

Variant	GMR↓	GMC↑	TIF↓	FMR
Full ChoreoStudio	10.42	83.23	0.08	0.848
Relation-only	15.29	81.89	0.10	0.792
Latent bypass	14.52	80.45	0.09	0.819
w/o relation-formation feedback	15.73	79.79	0.11	0.738
w/o music structure	14.33	82.71	0.09	0.798

C.4 DENOISING-BUDGET-MATCHED REFINEMENT ANALYSIS

We examine whether the gains from verification-guided refinement arise merely from additional diffusion computation. We compare two 50-step refinement cycles with a single 100-step generation, keeping the total number of denoising steps fixed. We additionally include best-of-two independent 50-step samples selected by the same frozen Verifier, controlling for the benefit of generating and ranking multiple candidates. Improvement over both controls therefore provides evidence for feedback-guided plan revision rather than additional denoising or sample selection. Unlike the adaptive best-over-cycles protocol used for the main results, the $R = 2$ setting here always executes exactly two cycles to enforce a fixed denoising budget.

Table 9: **Denoising-budget-matched refinement analysis.** NFE denotes the total number of denoising evaluations. Best-of-two samples are independently generated and ranked by the same frozen Verifier. FMR is ranked by proximity to the real-data reference of 0.832.

Setting	NFE	GMR↓	GMC↑	TIF↓	FMR
$R = 1$, 50 steps	50	15.43	79.32	<u>0.10</u>	0.810
$R = 1$, 100 steps	100	14.74	80.94	**0.09**	0.805
Best-of-2, 50 steps	100	<u>13.62</u>	<u>81.29</u>	**0.09**	**0.826**
$R = 2$, 50 steps	100	**10.36**	**83.20**	<u>0.10</u>	<u>0.847</u>

Increasing a single-cycle sampler from 50 to 100 steps yields only modest improvements, while selecting the better of two independent samples provides a stronger gain. Under the same 100-step denoising budget, feedback-guided refinement further reduces GMR by 23.9% and improves GMC by 1.91 points over best-of-two sampling. Its FMR remains close to the real-data reference, although TIF increases slightly from 0.09 to 0.10. These results indicate that the improvement in group realism and coordination cannot be explained by additional denoising or candidate selection alone.

C.5 Semantic Control

We evaluate whether the high-level description provides controllable macro-intent without treating it as a source of frame-accurate choreography. Each paired comparison fixes the music, sampled actor seeds, diffusion noise, and sampling configuration while changing only the description. We measure the resulting changes in group scale, transition rate, pose divergence, and unison. Aggregate GMR and TIF changes test whether the requested behavior is achieved by degrading motion realism or increasing collisions.

For each paired edit, direction accuracy is the percentage of examples in which the corresponding target metric changes in the requested direction. Changes in GMR and TIF are reported to detect semantic control achieved at the expense of motion quality or physical plausibility.

Table 10: **Semantic control under paired generation.** Changes are computed as edited minus original. Direction accuracy reports the percentage of pairs whose target metric changes in the requested direction.

Description edit	Target metric	Target change	Direction↑	ΔGMR	ΔTIF
Compact to expansive	Spread	+0.34 m	82.6%	+0.28	-0.006
Stable to dynamic	IDDV	+0.052	77.4%	+0.46	+0.004

Table 10 shows that contrasting descriptions produce consistent changes in the intended group behavior. Expansive descriptions increase group spread by 0.34 m in 82.6% of paired samples, while dynamic descriptions increase IDDV by 0.052 with a direction accuracy of 77.4%. The small changes in GMR and TIF indicate that these semantic responses are achieved without substantial degradation in motion realism or physical plausibility. These results support the semantic description as a controllable macro-level planning condition.

C.6 User Study

We complement the automatic evaluation with a blind user study following the ranking protocol of TCDiff (Dai et al., 2025). We recruit 30 participants aged 18 to 55. In each trial, participants view four anonymized and randomly ordered videos generated by CoDancers, TCDiff++, ST-GDance, and ChoreoStudio under the same music and group setting. They rank the videos without ties along four dimensions: *Motion Realism*, *Music Alignment*, *Formation Quality*, and *Inter-Dancer Coordination*. Formation Quality evaluates the visual organization and evolution of group formations, while Inter-Dancer Coordination assesses whether dancers interact coherently during formation transitions.

Ranks are assigned scores from 4 (best) to 1 (worst). For method m and dimension d, we normalize the scores to $[0, 1]$ as

$$S_{m,d} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{r_{q,m,d} - 1}{3}, \tag{70}$$

where Q is the number of valid rankings and $r_{q,m,d}$ is the rank score in trial q.

As shown in Figure 6, ChoreoStudio receives the highest score on all four dimensions. It achieves clear gains in Motion Realism and Formation Quality, indicating that explicit planning improves group organization without compromising motion naturalness. The largest margins occur in Music Alignment and Inter-Dancer Coordination, supporting the roles of multi-scale music conditioning and coupled formation, timing, and mobility planning.

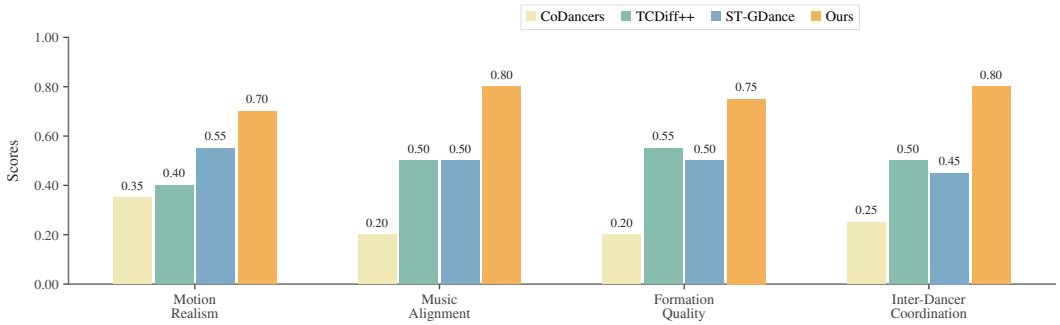


Figure 6: **User study results.** Normalized blind-ranking scores across four perceptual dimensions. ChoreoStudio is consistently preferred, with the largest margins in Music Alignment and Inter-Dancer Coordination.

D More Qualitative Results

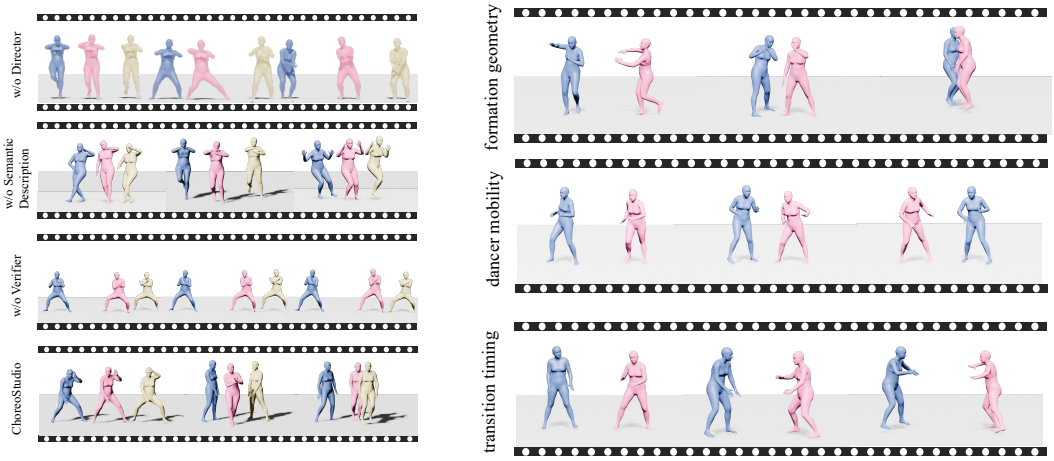
D.1 Qualitative Results and Video Demonstrations

Visualization and rendering. Unless otherwise specified, the paper figures use the standard SMPL mesh to convert generated PKL motion sequences into FBX animations under a shared camera and rendering configuration. For the supplementary demonstrations in `index.html`, we additionally retarget the generated motions to the YBot character for higher-quality visualization. The videos are rendered in Blender using scripts adapted from the open-source TCDiff implementation (Dai et al., 2025). Rendering affects only the visual appearance and does not modify the generated root trajectories or body motions.

Qualitative ablations. Figure 7a compares the effects of removing the Director, semantic description, and Verifier. Without the Director, dancers perform mismatched poses and undergo inconsistent spatial changes, producing a sequence that remains locally plausible but lacks coherent group-level organization. Removing the semantic description makes the dancers follow more uniform and repetitive movement patterns, reducing dancer-specific variation and weakening the progression of the choreography. Without the Verifier, the sequence remains close to the same wide formation and repeated pose across frames, suggesting that implausible or stagnant plans are not effectively corrected. In contrast, ChoreoStudio maintains distinct but coordinated dancer motions and gradually reorganizes the group from a dispersed arrangement into a compact formation. The comparison complements the quantitative ablations by showing how explicit planning provides group structure, semantic context encourages differentiated movement, and verification improves the realization of formation transitions.

Plan interventions and group-size generalization. Figure 7b visualizes the three forms of plan control. The formation-geometry intervention changes the relative spatial target, causing the two dancers to evolve from a separated arrangement toward a compact formation. Editing dancer mobility changes how the transition is distributed between the dancers, producing a different pattern of spatial participation and position exchange while preserving the overall group task. The transition-timing intervention changes the temporal location of the spatial reorganization, allowing the same

formation change to occur earlier or later in the sequence. Because the music, initial noise, and non-intervened plan variables are fixed, these differences can be attributed to the corresponding plan edits rather than sampling variation. The resulting motions remain temporally continuous, showing that the plan-to-motion interface can modify *what* changes, *when* it changes, and *who* executes the transition without regenerating an unrelated dance.



(a) Ablation study

(b) Controlled plan interventions

Figure 7: **Qualitative analysis of explicit choreography planning.** (a) Removing the Director, semantic description, or Verifier produces distinct degradations in group organization, motion variation, and formation transitions. (b) Editing formation geometry, transition timing, or dancer mobility produces the requested motion changes while fixing the music, initial noise, and remaining plan variables.

Group-size generalization. Figure 8 demonstrates that a single ChoreoStudio checkpoint generalizes across different dancer counts without size-specific retraining. As the group size increases, the number of pairwise interactions and possible spatial conflicts grows substantially. Nevertheless, the model preserves coherent individual motion, meaningful relative spacing, and coordinated group-level organization. The dancers remain distinguishable and contribute differently to the collective movement rather than collapsing into identical motion patterns. These results indicate that the relative formation representation, dancer-specific planning, and plan-to-motion interface extend naturally to variable-size groups.

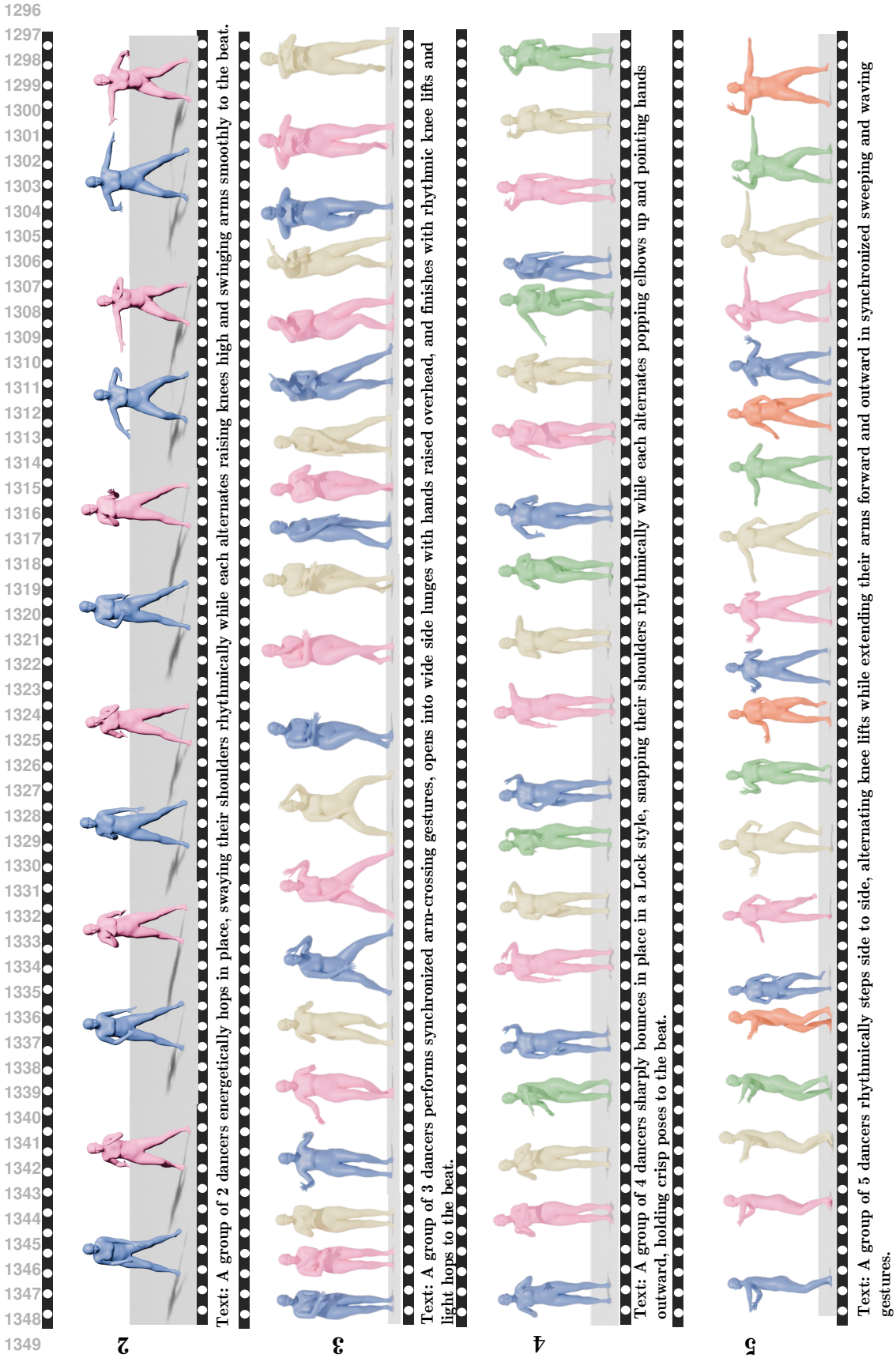


Figure 8: Qualitative results of ChoreoStudio under different group sizes.