Music-to-Dance Generation via Atomic Movements
Supplementary materials

Xinhao Cai[1], Yixuan Sun[2], Minghang Zheng[1], Qingchao Chen[3], Xin Jin[5], Song-chun Zhu[5], and Yang Liu[1,4]

[1] Wangxuan Institute of Computer Technology, Peking University
[2] School of Electronics Engineering and Computer Science, Peking University
[3] National Institute of Health Data Science, Peking University
[4] State Key Laboratory of General Artificial Intelligence, Peking University
[5] Beijing Institute for General Artificial Intelligence
{xinhao.cai, ilyssa_syx}@stu.pku.edu.cn
{minghang,qingchao.chen,yangliu}@pku.edu.cn
jinxinbesti@foxmail.com

Abstract. Music-driven dance generation aims to produce human motion that is both rhythmically synchronized and semantically consistent with music. While recent neural approaches have achieved impressive visual realism, they typically model motion as a continuous signal and neglect its compositional nature, making generated dances structurally incoherent and difficult to control. In this work, we introduce a structure-aware framework that models choreography as a sequence of atomic movements—semantically interpretable motion events that serve as the building blocks of dance. To construct this atomic movement vocabulary, we first segment large-scale dance data and cluster them into atomic movement groups. We then employ a large language model to semantically relabel and refine the clusters, yielding a set of interpretable and reusable atomic movements. Based on these atomic movement annotations, we design a two-stage generation framework that mirrors the human choreography process. In the atomic movement planning stage, the model predicts the type, duration, and timing of atomic movements conditioned on the input music, forming a symbolic dance allocation. In the completion stage, a transition-aware generator synthesizes smooth and stylistically coherent motion conditioned on the planned structure. Extensive experiments demonstrate that our method produces dances with significantly improved structural coherence, rhythmic alignment, and perceptual naturalness compared to existing baselines, while providing enhanced interpretability and controllable editing through explicit structural representation.

1 Introduction

The task of generating dance conditioned on music has become an important research topic at the intersection of digital entertainment, embodied AI, and

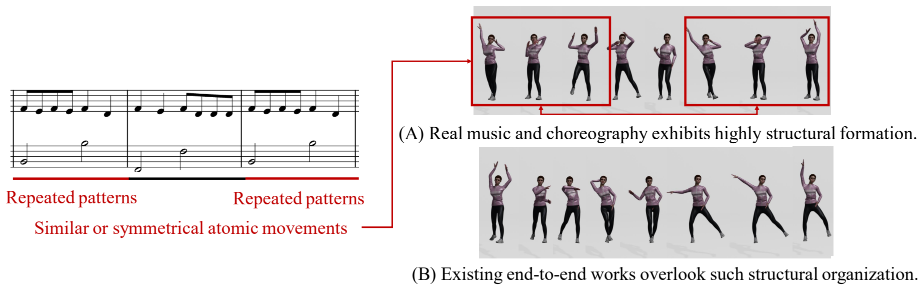


Fig. 1: (A)There exists explicit repeated patterns in music and choreography. (B)Existing methods, usually in an end-to-end form, treat music-to-dance generation as a simple sequence-to-sequence task, neglecting the inner structure.

computational creativity. It aims to synthesize motions both physically natural and musically expressive, enabling applications in virtual performance, interactive art, and intelligent choreography.

Both choreography and music theory [2, 11, 14, 16, 19] emphasize that a performance is organized around discrete, meaningful, and reusable units. A choreographer first plans the sequence of such fundamental atomic movements and their temporal arrangement, while the dancer subsequently refines transitions and expressive variations.

However, most existing neural approaches [23] treat dance generation as a direct sequence-to-sequence translation. These models learn frame-level or short-term correspondences, such as beat alignment or local rhythm consistency, but overlook the **structural organization** that underlies real choreography. As a result, while they can produce visually smooth motion clips, the generated dances often lack segment-level coherence across sections.

Thus, we revisit dance generation from a **structural and compositional** perspective. We hypothesize that complex dances can be represented as sequences of **atomic movements** that carry interpretable semantics and repeat or recombine to form longer compositions.

To realize this idea, we construct a high-quality set of **atomic movements**. In this work, we define an atomic movement as a basic motion process that is repeatable, with variation, and can be freely combined. Specifically, an atomic movement must satisfy three criteria: 1)It must involve **a clear process** rather than a single frame. A static pose is comprehensive for a single moment but ultimately frozen in time. In contrast, an atomic movement unfolds through time, reveals shifts in emotion or energy, and allows more dynamic with the rhythm and phrasing of the music, thus creating a deeper artistic atmosphere. 2)It should have **repetition and variation**. The movements should reappear throughout the dance sequences and evolve in rhythm, dynamics, range, and direction. It may be sped up or slowed down, enlarged or minimized, reversed, fragmented, and reassembled. The coexistence of repetition and variation allows the movement vocabulary of the entire dance to remain both rich and diverse, yet unified. 3)It should be **semantically interpretable**, meaning that each atomic movement can be described naturally in language, which facilitates clustering, retrieval, and human-understandable labeling.

We then design a pipeline for constructing atomic motion groups. First, we perform **segmentation** to divide continuous motion into procedural segments distinct from their surrounding context so that the movements involve clear processes. Next, we apply **clustering** over motion features to identify frequently repeating motions across the dataset. The resulting clusters correspond to high-level and repetitive motion patterns - such as kicking or spinning - that capture the dominant movement tendencies of each motion, while preserving diverse variations within each group. Finally, we perform **in-group re-clustering**: a tagging LLM annotates each motion segment with fine-grained natural-language descriptions, aligning segment semantics with human interpretation. A reasoning LLM then analyzes the annotations within each cluster to identify, summarize, and separate consistent intra-cluster patterns, re-clustering segments accordingly, yielding finer-grained, semantically interpretable motion categories.

Based on these learned atomic movements, we propose a two-stage structure-aware generation framework. In the **atomic movement planning** stage, given the input music, a planning network predicts which atomic movements should appear, along with their temporal locations and durations. This stage effectively produces a symbolic *dance score*, describing the high-level choreography structure. In the **motion completion** stage, a diffusion-based generator synthesizes transitions for the planned atomic movements and re-create the chosen atomic movements for better diversity, ensuring smooth transitions and rhythmic synchronization. This two-stage design offers two major advantages over conventional end-to-end approaches. First, it mirrors the real-world creative process of choreography: the first stage acts as the choreographer, planning the dance structure, while the second stage serves as the dancer, realizing the plan through expressive transitions. Second, because the atomic movements are discrete, meaningful, and temporally grounded, the generated dances become highly controllable—users can easily modify or replace specific movements, adjust durations, or reassemble dance sections to produce new variations without retraining. These properties make our framework not only effective for generation but also practical for editing and interactive design.

Comprehensive experiments demonstrate that our method produces more structurally coherent, rhythmically aligned, and perceptually natural dances than previous models. Furthermore, the explicit atomic representation and two-stage pipeline enable unprecedented controllability and interpretability in music-conditioned choreography generation.

Our main contributions are summarized as follows:

- We propose a three-stage pipeline for atomic movement discovery, which segments continuous motion into process-complete units, clusters them into repeating movement patterns while preserving inside-cluster variation, and clarifies finer-grained patterns via LLM-based re-clustering to produce semantically interpretable atomic movements.

– We propose a two-stage structure-aware framework that separates atomic movement planning from transition-based motion completion, explicitly modeling both choreography design and performance realization.
– Our approach achieves superior structural coherence, rhythmic consistency, and controllability compared to existing methods, offering a new perspective on interpretable and editable dance generation.

2 Related Works

2.1 Human Motion Generation

The task of synthesizing plausible human motion sequences has been extensively studied in computer vision, graphics, and robotics. Early methods [8] rely on retrieval or matching-based strategies over motion-capture databases, where appropriate sequences are retrieved or selected based on similarity. While such methods ensured physical realism by using real motion data, they are constrained by limited diversity and poor generalization beyond the database.

With the advent of deep generative models, condition-controlled motion generation became a major trend. TEMOS [17] and T2M-GPT [24] introduced VAE- or Transformer-based architectures, mapping free-form textual descriptions to 3D motions using large-scale datasets such as HumanML3D [6]. Diffusion-based models, such as MDM [22] and MotionDiffuse [26], further improved sequential coherence and sample diversity. However, text-to-motion tasks provide concrete textual guidance and demand precise motion. In contrast, dance is an inherently artistic and compositional form that requires not only kinematic realism but also aesthetic creation, involving structural planning and artistic variation, where similar movements are creatively diversified. Therefore, directly applying text-conditioned motion frameworks to dance fails to capture these artistic and hierarchical characteristics. Our work instead focuses on uncovering and utilizing discrete atomic movements that serve as repeatable choreographic building blocks aligned with musical structure.

2.2 Music-to-Dance Generation

Music-conditioned dance generation aims to produce motion sequences that synchronize rhythmically and stylistically with music. With larger music-motion datasets, generative models became dominant. FACT [13] used a cross-modal Transformer to autoregressively predict joint trajectories from musical features. Bailando [20] and TM2D [5] adopted VQ-VAE or GPT-based quantized representations to enhance music-conditioned controllability. Each discrete code in the learned vocabulary typically corresponds to a very short motion segment, resulting in decoded outputs that are essentially static poses and lack process information and segment-level structural information. EDGE [23] introduced diffusion models for dance generation with local editing and continuation capabilities. But these diffusion-based methods are usually trained within a short temporal window, which constrains their receptive field and prevents them from capturing global choreographic structures that unfold over longer timescales. In contrast, our pipeline handles the entire music at the first stage, therefore extracting more structural information of the music. Moreover, diffusion-based inpainting can edit motion at the signal level but lacks symbolic interpretability. EDMG [25] further works on long dance generation efficiency. DanceBA [4] and Beat-it [10] optimize the alignment of rhythm. But they neglect the inner structure of the dance choreography. Beat-it [10]. LODGE [12] proposed a coarse-to-fine diffusion pipeline for long dance sequences. However, it generates only keyframes as coarse guidance, hindering the generation of a repeatable structure conditioned on the music.

Our approach complements these paradigms by introducing a structural layer of atomic actions, planned and synthesized hierarchically, bridging symbolic-level control and continuous motion fidelity.

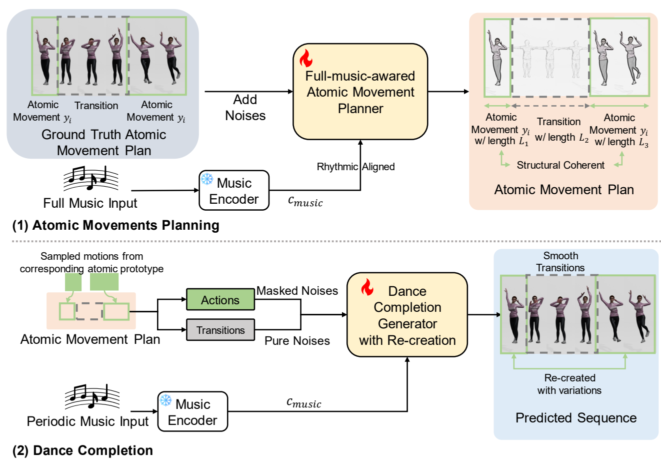


Fig. 2: The framework of our method. We propose a two-stage pipeline. 1)Atomic movement planning: the model allocates the temporal position, the duration, and the type of the atomic movements. 2)Dance Completion: The model optimizes the atomic movements and generates the transitions between them.

3 Method

3.1 Problem Formulation and Overview

Given a music condition $M^{1:L}$, where L is the length of the music, our goal is to generate a dance sequence $X^{1:F}$ aligned with the music, where $F = L \times R_F$ and R_F is a fixed frame rate. To enable the atomic-movement-based dance generation, we first collect and annotate the atomic movements included in the dance dataset, which we will introduce in Sec. 3.2.Then, as we demonstrate in Fig. 2 and Sec. 3.3, our generation consists of two stages: (1) atomic movement planning and (2) dance completion. We first allocate the type, temporal position, and duration of each atomic movement conditioned on the full music. Subsequently, we optimize each movement and generate the transitions between them to make the dance continuous and fluent.

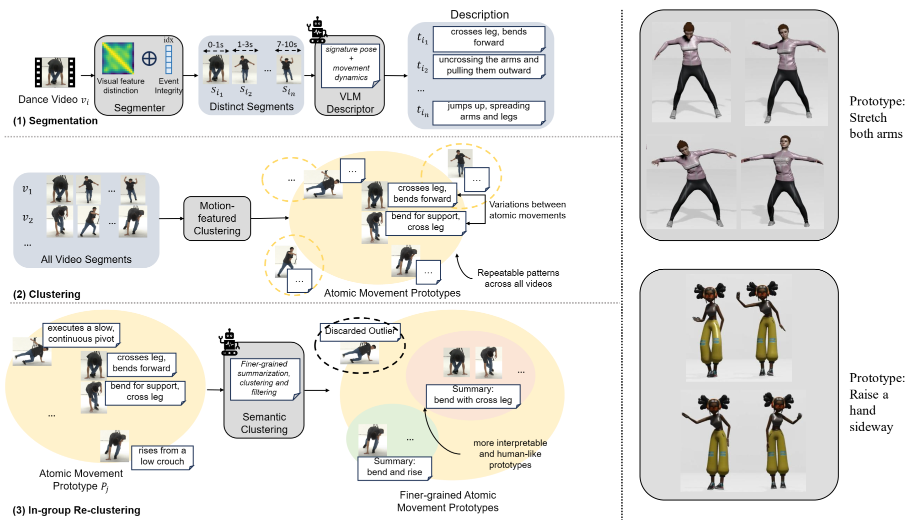


Fig. 3: The framework of our atomic movement discovery. 1)Segmentation exhibits distinct segments; 2)Clustering obtains atomic movement prototypes. 3)In-group Re-clustering re-splits each prototype by genres and leverages an LLM to further split them into finer-grained atomic movement prototypes. We show samples on the right.

3.2 Atomic Movement Discovery

Pipeline Overview. To find the atomic movements that 1)involve clear processes, 2)are repetitive and various, and 3)semantically interpretable, we design a three-stage pipeline to construct such atomic movement annotations(Fig. 3) on AIST++ [13] dataset. **(1)Segmentation** : to obtain clear motion processes that are distinct from their temporal context, we employ a temporal-event-proposal method following [15] to segment continuous motion into complete, self-contained motion events. **(2)Clustering**: to identify basic motion units that repeats with variation, we encode each segment using TMR's [18] motion encoder and apply K-Means clustering across the entire dataset, organizing the motion segments into repeatable prototypes while preserving variation. **(3)In-group Re-clustering**: We use Gemini to provide fine-grained textual labels that align segment semantics with human interpretation, and a reasoning LLM abstracts and clarifies variation patterns within each cluster, producing finer-grained and semantically interpretable atomic movement categories.

Segmentation. We aim to extract complete motion processes that are clearly distinct from their temporal context. For example, a kick - comprising the preparatory weight shift, leg extension, and recovery - forms a complete motion segment;

Algorithm 1 Adaptive segmentation via visual similarity

Input: Paired T-frame dance video $V = \{v_1, \ldots, v_T\}$ and T-frame 3D dance motion $M = \{m_1, \ldots, m_T\}$, N, minimum motion segment length $L_{\min}$

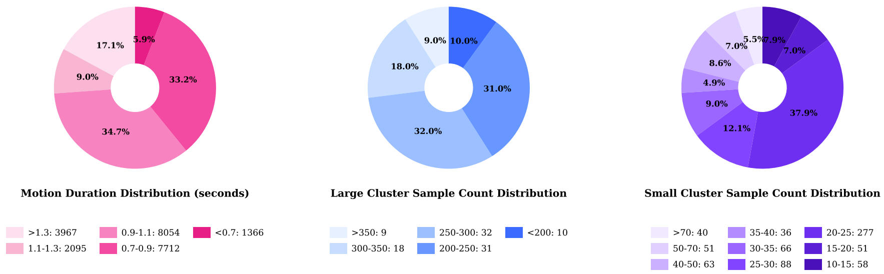
Output: motion segment count K, cut points $B = \{b_k\}_{k=1}^{K}$ and motion units $S = \{s_i\}_{i=1}^{K+1}$.

1: Extract per-frame visual features $F = \{f_1, \ldots, f_T\}$ using an I3D encoder.
2: Compute a frame-wise similarity matrix $A \in \mathbb{R}^{T \times T}$ where $A_{pq} = \cos(f_p, f_q)$.
3: $c_t \leftarrow A_{[t,:]} \in \mathbb{R}^T$ // Each row t of A encodes the similarity between frame t and all other frames.
4: $\tilde{c}_t \leftarrow [c_t; t/T] \in \mathbb{R}^{T+1}$ // Append normalized frame index to c_t to encourage temporally local grouping.
5: Perform N-means clustering on $\{\tilde{c}_t\}_{t=1}^T$ to obtain frame labels $\{l_t\}_{t=1}^T$.
6: Set cut points at label transitions: $\tilde{B} = \{\, t \mid \ell_t \neq \ell_{t-1},\ 2 \leq t \leq T \,\}$.
7: Iteratively merge any segment shorter than $L_{\min}$ with the shorter of its two temporal neighbors until all segments satisfy the minimum length, yielding $B = \{b_k\}, K = |B|$.
8: Segment M into motion units $S = \{s_i\}_{i=1}^{K+1}$, where $s_i = M_{[b_{i-1},\, b_i)}$, with $b_0 = 1$ and $b_{K+1} = T + 1$.

any single frame of the extended leg or a very brief leg movement would fail to convey the full event. To obtain such complete motion processes, we follow the temporal event proposal method in [15], performing segmentation by clustering by-frame similarity matrix and placing cut points at cluster boundaries. The idea is that frames are similar to frames within the same motion segment, and less similar to frames from other motion segments; therefore, frames within the same motion segment have the same similarity pattern, and tend to be clustered into a single motion segment. As detailed in Algorithm 1, our segmentation algorithm partitions the dance sequence by clustering temporally-augmented visual similarity features. By identifying transitions in the cluster labels and iteratively merging segments shorter than L_{min}, the algorithm adaptively extracts motion units that preserve the structural integrity of the dance while responding to changes in visual patterns.

Furthermore, we employ Gemini-2.5-Pro [21] to describe each dance video segment in terms of a *signature pose* and *movement dynamics*. A signature pose is a relatively stable, sculptural posture that serves as a visual symbol of an atomic movement. Movement dynamics capture the transitions between static poses, encompassing the path, rhythm, intensity, and fluidity of motion, and thereby defining the diversity and specificity of each atomic movement. Specifically, we use PoseScript [3] to describe selected keyframes identified as *motion beats* (local minima of segment-wise joint velocities). These PoseScript annotations are provided as auxiliary cues to the VLM descriptor, which then generates fine-grained natural language descriptions that capture both the signature pose and movement dynamics.

Clustering. To find the repetition and variation in atomic movements, we cluster motion segments across the entire dataset to extract atomic movement pro-



(a) **(b)** **(c)**

Fig. 4: Statistics of atomic motion lengths and clustering.

totypes $\mathcal{P} = \{P_j\}$ that exhibit repeatability across pieces and preserve their variety.

Concretely, each segment is encoded by a pretrained TMR [18] motion encoder and projected into the joint motion-text embedding space; we then run K-Means to group segments into recurring atomic movement prototypes. TMR was trained with text-motion contrastive learning on HumanML3D [6], whose captions emphasize high-level action descriptions rather than fine-grained dance-specific details; as a result, the encoder naturally groups high-level similar segments while retaining intra-class variation. Because not every segment exhibits a clear pattern, we keep only segments near cluster centers and discard ambiguous edge points, yielding a set of high-confidence atomic movement prototypes. We yield 100 atomic movement prototypes from the complete AIST++ dataset; each prototype contains 268.57 segments on average (Fig. 4b). This provides sufficient data density for the model to learn robust class representations. Also, these large clusters ensure that the atomic movements are repetitive, and the difference in a specific cluster guarantees the diversity.

In-group Re-clustering. To increase the interpretability of the atomic movements,we conduct semantical re-clustering for each prototype P_j. We further decompose each prototype to identify and label common patterns within a motion class. To better capture such intra-group patterns, we first pre-split each atomic movement prototype P_j by dance genre, as motions from different genres tend to be semantically distinct.

To obtain finer-grained, more interpretable clusters, we use a summarizing LLM to iteratively: (i) identify a subset of mutually similar captions as a sub-prototype, and (ii) distill it into a concise semantic tag. We repeat this procedure until the number of ungrouped segments falls below a threshold. On average, each group yields 7.3 sub-prototypes with 31.8 samples each (Fig. 4c), preserving fine-grained diversity without severe data sparsity. This also makes the atomic movements more interpretable.

3.3 Dance Generation

Atomic Movement Planning. Once a set of atomic movements has been defined, the central challenge is to determine how they should be temporally organized to form a complete dance that aligns with the given music and allocate these atomic movements to critical and appropriate positions. Traditional music-to-dance methods attempt to learn a direct mapping from music to dense motion frames, which captures local rhythm but ignores long-term compositional structure. In contrast, we explicitly model the planning process as predicting a symbolic sequence of atomic movements—each with a specific category, temporal position, and duration conditioned on full music instead of clips—thereby bridging musical phrasing with choreographic structure.

Given a music sequence, the planner aims to produce a symbolic *dance score* that specifies when and how each atomic movement should occur. Our goal is to allocate every motion frame an atomic movement label or a token indicating that this frame should belong to a transition. We adopt a Discrete Denoising Diffusion Probabilistic Model(D3PM) [1].

Assume that there are K types of atomic movements, we use numbers 0 to K to denote these atomic movements and the transition. We have the atomic movement annotation for ground truth dance $Y_{gt} = [\hat{y_1}, \hat{y_2}, ..., \hat{y_F}], y_i \in \{0, 1, 2, ..., K\}$ denotes the atomic category of i-th frame or no atomic movement at i-th frame if $\hat{y_i} = 0$.Let $q(y_t|y_{t-1})$ denote the forward diffusion process that gradually perturbs the true atomic labels into random noise according to a transition matrix $Q_t \in \mathbb{R}^{(K+1)\times(K+1)}$:

$$q(y^t|y^{t-1}) = \mathrm{Cat}(y^t; Q_t[y^{t-1}, :]), \tag{1}$$

where y^t is the movement type at $t-$th frame. At each step, a token is retained with probability α_t or replaced with a uniformly random category with probability $(1 - \alpha_t)$. After T steps, the sequence approaches a uniform distribution.

We first use a pretrained music encoder to encode the music sequence M into music features $c_{music} = \mathrm{Enc}(M)$. The reverse process $p_\theta(y_{t-1}|y_t, \mathbf{c}_{music}, t)$ is parameterized by a Transformer-based diffusion model conditioned on the music features:

$$p_\theta(y_{t-1}|y_t, \mathbf{c}_{music}, t) = \mathrm{Cat}(\pi_\theta(y_t, \mathbf{c}_{music}, t)). \tag{2}$$

The model is trained by minimizing the variational objective:

$$\mathcal{L}_{\mathrm{D3PM}} = -\mathbb{E}_q\big[\log p_\theta(y_{t-1}|y_t, \mathbf{c}_{music}, t)\big]. \tag{3}$$

During inference, we start from a uniformly random sequence y_T and iteratively sample from $p_\theta(y_{t-1}|y_t, \mathbf{c}_{music}, t)$ until $t = 0$. The resulting sequence $\hat{\mathbf{Y}} = [\hat{y}_1, \ldots, \hat{y}_F]$ represents the planned atomic movements, specifying both the category and temporal arrangement of each motion event. Also, we observed that any mislabeled frame within a continuous atomic movement segment could severely disrupt the planning results. Thus, we apply a lightweight post-processing module after the first-stage atomic movement planning. Specifically, we first perform a sliding-window majority vote to remove isolated frame-level errors and smooth local inconsistencies. We then apply a minimum-duration merging heuristic that detects abnormally short segments and reassigns them to the most semantically compatible neighboring segment. This two-step refinement effectively corrects over-segmentation errors while preserving genuine action boundaries, leading to more stable and coherent atomic-level plans.

This stage of atomic movement planning allows the model to allocate atomic movements to proper temporal positions and plan where to put the transitions between them, therefore improving the structural coherence and rhythmic alignment as the full music is used here. This explicit planning procedure also enhances the structural interpretability and segment-level controllability.

Dance Completion. After obtaining the planned atomic movement sequence, the goal of the dance completion stage is to generate a continuous motion trajectory that connects these atomic segments smoothly while maintaining rhythmic and stylistic consistency with the given music. Although the atomic movement sequence provides a structural outline of the choreography, it only specifies the categorical composition of movements. A realistic dance requires not only accurate realization of each atomic movement but also natural transitions between neighboring motions. Also, as choreographic theory [19] indicates, structured movements should exhibit variation when they recur. Accordingly, we apply appropriately scaled masked noise to the movement primitives to increase choreographic diversity and to improve their coherence with the preceding and succeeding transitional movements. To this end, we introduce a diffusion-based completion model that refines and connects the planned motion sequence, ensuring continuity, musical alignment, and expressive diversity.

Given the predicted atomic label sequence $\hat{\mathbf{Y}} = [\hat{y}_1, \ldots, \hat{y}_F]$, we first retrieve representative motion samples from the pre-clustered atomic movement database. For each atomic label $\hat{y}_i$, we select a motion segment whose duration most closely matches the planned temporal length, rescale it to fit the target duration, and place it into the corresponding frame range. For these chosen atomic movements, we will apply a masked noise, allowing them to exhibit better diversity and coherence with adjacent motions. This yields a coarse motion sequence $\mathbf{M}_0$, in which atomic segments are filled while transitional frames remain uninitialized. We employ a Denoising Diffusion Probabilistic Model(DDPM) [7] to refine $\mathbf{M}_0$ and generate smooth transitions. It consist of a forward perturbation process and a reverse denoising process. The former maps data m_0 to m_t via $q(m_t|m) = \mathcal{N}(\sqrt{\alpha_t}m, (1-\alpha_t)\mathbf{I})$, where α_t governs the noise schedule. During denoising, we optimize a network $f_\theta(m_t, t, \mathbf{x}_{music}, \mathbf{M}_0, \mathbf{w})$ to predict the clean signal m_0, with music condition $\mathbf{x}_{music}$, draft motion sequence $\mathbf{M}_0$ generated by stage1 and mask $\mathbf{w}$ controlling the noise ratio on the atomic movements chosen.

This masked strategy enables the model to preserve the semantics of known atomic movements while synthesizing diverse and musically aligned transitions. The training objective follows the standard denoising loss:

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\mathbf{m}_0}\left[\left\|f_\theta(m_t, t, \mathbf{x}_{music}, \mathbf{M}_0, \mathbf{w}) - \mathbf{m}_0\right\|_2^2\right]. \tag{4}$$

In addition, we apply a transition loss to ensure temporal smoothness across atomic boundaries: $\mathcal{L}_{trans} = \sum_{b\in\text{boundaries}} \left\|\mathbf{M}_{b^-} - \mathbf{M}_{b^+}\right\|_1$, which penalizes discontinuities between neighboring atomic movements, where $\mathbf{M}_{b^-}$ and $\mathbf{M}_{b^+}$ denote the pose frame before and after the atomic movement boundaries. The total training loss is thus: $\mathcal{L} = \mathcal{L}_{diff} + \lambda_{trans}\mathcal{L}_{trans}$.

Consequently, $\hat{\mathbf{M}}$ exhibits coherent structure and smooth temporal dynamics.

To conclude, this dance completion stage generates transitions to make the dance more natural and smooth, while simultaneously allowing the atomic movement to be re-created to improve diversity and aesthetics.

4 Experiments

Method	$FID_k\downarrow$	$FID_g\downarrow$	$Div_k \rightarrow$	$Div_g \rightarrow$	BAS↑	$R\uparrow$
Ground Truth	17.1	10.6	8.19	7.45	0.2374	42.1
DanceNet [27]	69.18	25.49	2.8	2.85	0.143	14.1
DanceRevolution [9]	73.42	25.92	3.52	4.87	0.195	13.7
Bailando [20]	28.16	9.62	7.83	6.34	0.2332	17.9
EDGE [23]	42.16	22.12	3.96	4.61	0.2334	16.3
Lodge [12]	37.09	18.79	5.58	4.85	0.2423	18.2
Ours	**25.26**	**9.03**	**8.01**	**6.69**	**0.2470**	**26.6**

Table 1: Quantitative results on AIST++ [13]. An up-arrow↑ indicates the performance is better if the value is higher. We use **bold** to represent the best result in the table.

Method	$FID_k\downarrow$	$FID_g\downarrow$	$Div_k \rightarrow$	$Div_g \rightarrow$	BAS↑	$R\uparrow$
w/o Post-process	25.26	9.03	8.01	6.69	0.2470	26.6
w/ Post-process	24.02	8.78	8.03	6.68	0.2472	27.5
Using GT	21.59	8.82	8.13	6.94	0.2455	29.8

Table 2: Ablation studies on the planning results.

4.1 Metrics

Following the existing music-to-dance generation works, we employ the following metrics to evaluate the quality of the dance sequences generated.

1)Frechet Inception Distance (FID). We compute the FID between the motion features of generated dances and those of the corresponding ground-truth sequences, qualifying the overall distribution distance in kinematic or geometry spaces. 2)Motion Diversity (Div). To assess the variety of generated dance motions, we follow [12, 20] and compute the mean pairwise Euclidean distance within the kinematic or geometry feature space. 3)Beat Alignment Score (BAS). To evaluate rhythmic synchronization between the generated dance and the input music, we adopt the Beat Alignment Score (BAS) as proposed in [12, 23]. Also, to evaluate the structural consistency of the dance generated, we additionally propose a new metric: 4)R-precision. We split the music into segments and calculate the music features of each clip. We examine whether the motion feature of the dances in every pair of the most-similar-music-clip is among the three highest most-similar-dance-clip. The precision is reported as the R. We also report the variance between 5 samples generated by the same music as 5)MultiModality to test the diversity of the model.

4.2 Quantitative Comparisons

Table 1 presents a quantitative comparison of our method against several representative baselines. Across all metrics, our approach demonstrates clear improvements in both structural quality and rhythmic consistency.

Method	$FID_k\downarrow$	$FID_g\downarrow$	$Div_k \rightarrow$	$Div_g \rightarrow$	BAS↑	$R\uparrow$	$MultiModality\uparrow$
Ground Truth	17.1	10.6	8.19	7.45	0.2374	42.1	-
Fixed Atomic Movements	24.13	8.44	6.99	4.61	0.2356	27.2	1.43
Flexible Atomic Movements	25.26	9.03	8.01	6.69	0.2470	26.6	2.25

Table 3: Ablation studies on the flexibility of atomic movements.

Method	$FID_k\downarrow$	$FID_g\downarrow$	$Div_k \rightarrow$	$Div_g \rightarrow$	BAS↑	$R\uparrow$	$MultiModality\uparrow$
Ground Truth	17.1	10.6	8.19	7.45	0.2374	42.1	-
Single Candidate	25.30	9.01	6.42	4.11	0.2458	28.9	2.18
Random Choice	27.94	9.98	8.07	6.73	0.2466	24.5	2.62
Duration Nearest	25.26	9.03	8.01	6.69	0.2470	26.6	2.43

Table 4: Ablation studies on the retrieval of the atomic movements.

1)In terms of motion fidelity, our method achieves the lowest $\mathrm{FID_k}$ and $\mathrm{FID_g}$ scores (25.26 and 9.03, respectively), outperforming all diffusion- and VQ-VAE-based baselines. This indicates that the generated dances are not only physically realistic in terms of kinematic smoothness but also compositionally coherent in spatial structure. Such results prove that our method better captures the atomic movements, which are the common patterns shown in the real dance, and thus produce results that are closer to ground truth. **2)For motion diversity**, our approach achieves high scores in both $\mathrm{Div_k}$ and $\mathrm{Div_g}$, approaching the diversity level of real human dances. This demonstrates that the proposed two-stage framework does not sacrifice diversity for structural regularity—on the contrary, by introducing partial noise during motion completion, our model enriches local variations within each atomic movement and generates a wider range of expressive dynamics. **3)In terms of rhythmic alignment**, our model attains the highest Beat Alignment Score, indicating superior synchronization with musical beats. This benefit arises from the explicit temporal planning in the first stage, which allocates atomic movements according to musical structure, and from the rhythm-conditioned diffusion refinement that further enhances temporal coherence. **4)R-precision** evaluates the structural consistency of dance sequences under similar musical contexts. Our method achieves the highest R score, substantially outperforming previous methods that lack explicit compositional modeling. This result confirms that our framework successfully captures the relationship between musical and choreographic structure—when two music segments share similar phrasing or rhythm, the generated dances exhibit correspondingly similar motion patterns.

Overall, these results quantitatively validate that our structure-aware two-stage framework produces more coherent, musically aligned, and controllable dances than existing end-to-end generative baselines.

4.3 Ablation Studies and Discussions

Planning Results. Table 2 compares three settings: using our raw planning results, applying the proposed post-processing module, and directly using ground-truth atomic labels. Without post-processing, frame-level mispredictions in the

Cluster Base Num	ReCluster?	$FID_k\downarrow$	$FID_g\downarrow$	$Div_k\rightarrow$	$Div_g\rightarrow$	BAS↑	$R\uparrow$
75	✗	33.24	13.31	7.81	6.12	0.2413	20.2
100	✗	32.68	12.09	7.78	6.05	0.2420	21.3
125	✗	34.57	14.28	7.75	6.01	0.2415	21.1
100	w/o LLM	30.11	11.27	8.25	6.05	0.2431	23.3
100	w/ LLM	25.26	9.03	8.01	6.69	0.2470	26.6

Table 5: Ablation studies on the methodology of clustering.

planner occasionally fragment continuous atomic movements, which propagate inconsistencies to stage 2. Introducing the majority-vote smoothing and minimum-duration merging noticeably improves both fidelity and alignment, indicating that corrected action boundaries help the model generate more stable and musically aligned motions. Using ground-truth atomic action labels further boosts all metrics, demonstrating that a more precise atomic movement planning benefits dance quality. This also proves that with human editing before stage 2, the model is capable of generating even better dances with a more appropriate atomic movement allocation.

Atomic Movement Flexibility. We further examine whether allowing the re-creation of retrieved atomic movements in the second stage contributes to overall motion quality. In the Fixed Atomic Movements setting, the selected atomic instances are directly concatenated and temporally rescaled, without further modification. In contrast, the Flexible Atomic Movements variant enables the diffusion-based generator to adjust and refine each atomic motion during denoising, allowing smoother transitions and better adaptation to the accompanying music. As reported in Tab. 3, the fixed version exhibits lower motion diversity and weaker beat alignment, indicating limited expressiveness and poor synchronization with music. By enabling refinement, the flexible variant achieves a better balance between realism, rhythmic coherence, and diversity, suggesting that moderate adaptability in atomic movements allows the generator to better capture high-level musical variations while preserving natural continuity. It also matches the real-world dance performance: the dancer will add variation to the repeated atomic movements. We also use MultiModality to test the diversity among the results generated by the same music. With flexible atomic movements, the results yield a better variety. This proves that rather than simply copying the movements in the datasets, our flexible atomic movements adjust the movement chosen and perform them in diverse ways.

Retrieval of atomic movements. To investigate the effect of instance selection strategies when retrieving atomic movements from their corresponding groups, we compare three variants: (1) *Single Candidate*, where each atomic category is always represented by a fixed motion instance; (2) *Random Choice*, where an instance is randomly sampled from the corresponding group; and (3) *Duration Nearest*, where we select the instance whose length is closest to the predicted duration. As shown in Tab. 4, *Single Candidate* yields the lowest motion diversity, since using a single fixed instance suppresses the natural variability of motion patterns. Although *Random Choice* improves diversity, it leads to higher FID scores due to the scaling of the movements, which makes them a little more unnatural. Also, such randomness makes the R lower because, under

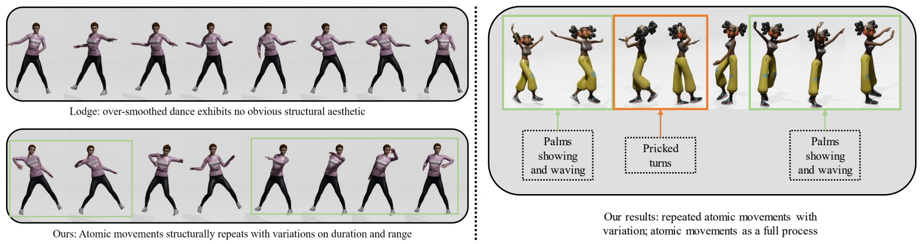


Fig. 5: Qualitative results and comparisons with Lodge [12].

similar musical structures, the atomic movements chosen with a larger randomness might differ more. Our proposed *Duration Nearest* selection achieves the best overall balance—maintaining high motion diversity while preserving structural and rhythmic coherence. This result validates that aligning the atomic instance duration with the predicted temporal structure effectively enhances motion smoothness and musical synchronization. This strategy also performs better in MultiModality than Single Candidate, proving that this design helps the model generate more diverse dance than using a single candidate.

Cluster Method. We conduct experiments to investigate how the clustering method affect the results as shown in Tab. 5. When varying the base number of clusters from 75 to 125 without re-clustering, an excessively small number of clusters slightly harms geometric fidelity and choreography quality, as over-merged clusters blur fine-grained motion distinctions. Conversely, too large a number of clusters (e.g., 125) leads to over-fragmentation and decreased motion coherence. More importantly, introducing our LLM-assisted re-clustering significantly improves all metrics. Furthermore, compared with the naïve k-means re-grouping, the LLM-based semantic refinement improves FID scores and enhances rhythmic and structural consistency. This demonstrates that leveraging language priors helps to merge semantically similar yet dynamically diverse movements, leading to more interpretable and musically aligned atomic motion categories.

4.4 Qualitative Results

On the left of Fig. 5, we qualitative compare our method with Lodge [12]. We can find that our results exhibit typical atomic movements repeatedly with variations on duration and range. However, methods like Lodge suffers from an over-smooth dance with no apparent dance primitive. On the right, we show more results where the atomic movements show as a full process, exhibit repetition and variation, and are interpretable. We show more visualizations in the supplementary materials.

5 Conclusions

In this work, we present a structure-aware framework for music-conditioned dance generation that explicitly models dance as a sequence of semantically interpretable atomic movements. By decoupling the generation process into atomic movement planning and motion completion, our method effectively mimics the two-stage choreography workflow of human creation — planning and performance — achieving both structural coherence and expressive flexibility. Comprehensive experiments demonstrate our approach produces more musically aligned and structurally consistent dances and enables fine-grained, interpretable control over composition.

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
2. Conklin, D.: Pattern in music. Journal of Mathematics and Music **15**(2), 95–98 (2021). https://doi.org/10.1080/17459737.2021.1947404, https://doi.org/10.1080/17459737.2021.1947404
3. Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory: PoseScript: 3D Human Poses from Natural Language. In: ECCV (2022)
4. Fan, C., Guan, J., Zhao, X., Xu, D., Lin, Y., Ye, T., Feng, P., Pan, H.: Align your rhythm: Generating highly aligned dance poses with gating-enhanced rhythm-aware feature representation. arXiv preprint arXiv:2503.17340 (2025)
5. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration (2023), https://arxiv.org/abs/2304.02419
6. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239 (2020)
8. Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. ACM Transactions on Graphics (ToG) **39**(4), 53–1 (2020)
9. Huang, R., Hu, H., Wu, W., Sawada, K., Zhang, M., Jiang, D.: Dance revolution: Long-term dance generation with music via curriculum learning. In: International Conference on Learning Representations (2021)
10. Huang, Z., Xu, X., Xu, C., Zhang, H., Zheng, C., Qin, J., He, S.: Beat-it: Beat-synchronized multi-condition 3d dance generation (2024), https://arxiv.org/abs/2407.07554
11. Leh, A., Endres, D., Hegele, M.: A primitive-based representation of dance: modulations by experience and perceptual validity. J Neurophysiol **130**(5), 1214–1225 (Oct 2023)
12. Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1524–1534 (2024)
13. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++ (2021), https://arxiv.org/abs/2101.08779

14. McDonough, J., Herczyński, A.: Fractal patterns in music. Chaos, Solitons & Fractals **170**, 113315 (2023). https://doi.org/https://doi.org/10.1016/j.chaos.2023.113315, https://www.sciencedirect.com/science/article/pii/S0960077923002163
15. Nam, J., Ahn, D., Kang, D., Ha, S.J., Choi, J.: Zero-shot natural language video localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1470–1479 (2021)
16. Okada, N., Iwamoto, N., Fukusato, T., Morishima, S.: Dance motion segmentation method based on choreographic primitives. pp. 332–339 (03 2015). https://doi.org/10.5220/0005304303320339
17. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022)
18. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: International Conference on Computer Vision (ICCV) (2023)
19. Reese, S.: Music in motion: Interpreting musical structure through choreography in gershwin's "an american in paris" (04 2019)
20. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation via actor-critic gpt with choreographic memory. In: CVPR (2022)
21. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., et al.: Gemini: A family of highly capable multimodal models (2024), https://arxiv.org/abs/2312.11805
22. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu
23. Tseng, J., Castellon, R., Liu, C.K.: Edge: Editable dance generation from music. arXiv preprint arXiv:2211.10658 (2022)
24. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
25. Zhang, J., Sun, Y., Zhang, H., Tang, J.: Edmg: Towards efficient long dance motion generation with fundamental movements from dance genres. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 10447–10456. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755677, https://doi.org/10.1145/3746027.3755677
26. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
27. Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., Xia, S.: Music2dance: Dancenet for music-driven dance generation. ACM Trans. Multimedia Comput. Commun. Appl. **18**(2) (Feb 2022). https://doi.org/10.1145/3485664, https://doi.org/10.1145/3485664