Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SPH-DiffDance: Semantic Prototype Reconstruction and Hypergraph-Enhanced Diffusion Framework for Long-Sequence Music-Driven Dance Generation

Zhengping Hu

hzp@ysu.edu.cn

Yanshan University

Yifan Wu

Yanshan University

Xingyu Li

Yanshan University

Chuanxin Xu

Yanshan University

Research Article

Keywords: Music-driven dance generation, Long-sequence motion generation, Diffusion model, Semantic prototype reconstruction, Hypergraph-enhanced pose refinement

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10936613/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SPH-DiffDance: Semantic Prototype Reconstruction and Hypergraph-Enhanced Diffusion Framework for Long-Sequence Music-Driven Dance Generation

Zhengping Hu[1,2*], Yifan Wu[1†], Xingyu Li[1†], Chuanxin Xu[1†]

[1*]School of Information Science and Engineering, Yanshan University, Hebei 066004, China.

[2]Hebei Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Hebei 066004, China.

*Corresponding author(s). E-mail(s): hzp@ysu.edu.cn;
Contributing authors: yifanwu@stumail.ysu.edu.cn;
tutu7887@stumail.ysu.edu.cn; xuchuanxin0805@stumail.ysu.edu.cn;
†These authors contributed equally to this work.

Abstract

Long-sequence music-driven dance generation remains a challenging task, as existing methods still struggle with unstable global choreography semantics, insufficient local pose coordination, and difficulty in preserving complex motion details. Aiming to address these issues, we propose a Semantic Prototype Reconstruction and Hypergraph-Enhanced Diffusion Framework for Long-Sequence Music-Driven Dance Generation (SPH-DiffDance), refining motion from global semantic structure to local pose details. SPH-DiffDance follows a coarse-to fine diffusion framework, designing a Semantic Prototype Reconstruction (SPR) module at the coarse stage to provide semantic guidance for motion generation through constructing a learnable global dance prototype library. The motion latent features are queried and reconstructed, transforming continuous motion representations into prototype representations with semantic constraints. This process enhances the thematic consistency and long-range choreography stability of generated dances. Building upon established global semantics, a Hypergraph-Enhanced Pose Refinement (HEPR) module further improves local motion quality by modeling high-order motion correlations among the trunk, limbs, and feet through hyperedges. Compared with conventional skeleton graphs that

mainly describe pairwise joint connections, the hypergraph structure provides richer representations of complex human motion relationships and improves pose coordination and structural rationality. Foot-contact constraints are also incorporated to refine local motion details and reduce foot sliding and abnormal poses. Experiments on the AIST++ music-driven dance generation dataset demonstrate that SPH-DiffDance can generate long-sequence 3D dances with good musical synchronization, motion continuity, and structural consistency.

Keywords: Music-driven dance generation, Long-sequence motion generation, Diffusion model, Semantic prototype reconstruction, Hypergraph-enhanced pose refinement

1 Introduction

Dance is an artistic form that combines music, body movement, and emotional expression and has long served as an important means of cultural communication. With the increasing demand for virtual digital humans, game animation, immersive interaction, and intelligent choreography, the generation of realistic 3D dances from music has become an active research topic in computer vision, artificial intelligence, and multimodal generation [1, 2]. The objective of music-driven dance generation is not only to produce visually reasonable human motions, but also to maintain consistency between musical rhythm, energy variation, phrase structure, and movement style. Compared with general human motion generation, dance generation involves stronger temporal organization and artistic expression. Therefore, generating long-sequence dances that are synchronized with music while preserving natural human motion remains a challenging problem.

Early studies on music-driven dance generation focused mainly on learning the temporal mapping between music features and motion sequences. DanceRevolution introduced a curriculum learning strategy to capture long-term music-motion dependencies, which reduced the difficulty of training long dance sequences to some extent [3]. Later, the AIST++ dataset provided paired data containing various music styles and 3D human motions, becoming an important benchmark for music-conditioned dance generation [4]. Based on this dataset, Bailando used a choreography memory module to learn meaningful motion units and used actor-critic GPT for motion composition, improving the continuity and rhythmic alignment of generated dances [5]. These approaches established the foundation for structured dance generation. However, when generating longer sequences, they still tend to produce repeated motions and insufficient global choreography coherence due to limited modeling of long-term dance structures, while rhythm consistency and motion continuity across segments remain difficult to maintain.

The emergence of diffusion models has provided a new perspective for human motion and dance generation. Taking advantage of their ability to model complex data distributions, diffusion-based methods can gradually recover realistic motions from noisy representations [6]. In music-driven dance generation, EDGE incorporates music conditions into the diffusion process and enables editable dance synthesis [7]. DiffDance

further introduces a cascaded diffusion framework, where coarse motion generation and sequence refinement are combined to improve temporal details in long dances [8]. For longer choreography generation, LODGE adopts a coarse-to-fine strategy by learning feature dance primitives at the global level and generating detailed motions locally, achieving a better balance between overall arrangement and local quality [9]. Lodge++ extends this idea to ultra-long sequences and multi-dance scenarios, improving the ability to model complex choreography patterns [10]. PAMD further considers motion credibility and introduces foot-ground contact refinement to alleviate abnormal poses and foot sliding problems [11]. Despite these improvements, maintaining consistent global semantics while preserving detailed local structures in long-sequence dance generation remains unresolved.

Apart from the generation framework, human body representation and spatial relationship modeling also significantly influence dance quality. Although SMPL-based representation [12] and advanced pose estimation methods [13] provide effective human body modeling, they cannot fully capture the complex high-order coordination among the torso, hips, limbs, and feet during dance movements. Modeling foot-contact states has been shown to effectively reduce foot skating artifacts [14], while hypergraph neural networks offer a more flexible means of representing such high-order joint relationships by connecting multiple nodes simultaneously through hyperedges [15].

Although considerable progress has been achieved, long-sequence music-driven dance generation still faces several limitations. One major issue is the preservation of global choreographic semantics. A complete dance sequence is not simply a collection of independent movements; instead, it is organized by musical structure, movement themes, style transitions, and temporal coherence. Local music-motion alignment alone is insufficient to maintain consistency throughout the entire sequence. Another difficulty lies in local motion refinement. Since long dances are usually generated through multiple stages, discontinuities between segments, unnatural poses, and foot sliding may still occur. In addition, existing approaches rarely explicitly model high-order interactions between multiple joints , which limits their ability to generate complex dance movements with coordinated body structures.

To overcome these limitations, we proposes SPH-DiffDance, a long-sequence music-driven dance generation framework based on semantic prototype reconstruction and hypergraph-enhanced diffusion. Instead of directly generating complete long sequences, the proposed framework follows a coarse-to-fine diffusion strategy, decomposing choreography modeling into a global semantic learning stage and a local structural refinement stage. In the global stage, the Semantic Prototype Reconstruction (SPR) module builds a learnable prototype library to provide stable semantic representations for long-range choreography modeling. In the local stage, the Hypergraph-Enhanced Pose Refinement (HEPR) module captures high-order joint coordination through dynamic hypergraphs, with foot-contact constraints to reduce unnatural poses and foot sliding.

The main contributions of this work are summarized as follows:

1. In this paper, we propose SPH-DiffDance, a two-stage coarse-to-fine diffusion framework for long-sequence music-driven dance generation. The framework decomposes dance generation into global choreography modeling and local motion refinement

stages, reducing semantic drift and motion discontinuity during long-sequence synthesis.
2. The Semantic Prototype Reconstruction (SPR) module is introduced to improve global choreography representation. By constructing a library of learnable global dance prototypes , the proposed module performs semantic retrieval and reconstruction of latent motion features. Furthermore, a primitive feature dance extraction strategy is developed to convert global motion representations into hard and soft cues, providing semantic guidance for local diffusion generation.
3. We propose the Hypergraph-Enhanced Pose Refinement (HEPR) module to improve local motion coordination. By combining dynamic hypergraphs with SMPL skeleton topology, the module models high-order relationships among multiple body joints through hyperedges. Together with foot-contact constraints, it reduces foot sliding and abnormal poses, improving the structural quality of complex dance movements.
4. Experimental results on the AIST++ dataset demonstrate that SPH-DiffDance can generate long-sequence 3D dances with better rhythm synchronization, motion continuity, and structural consistency, showing improvements in both global semantic coherence and local motion quality.

2 Related Work

2.1 Music-driven dance generation

Music-driven dance generation aims to synthesize human dance sequences that are synchronized with input music while maintaining motion naturalness, style consistency, and expressive quality. Unlike general human motion generation, this task requires not only physically plausible body movements but also modeling the complex correspondence between musical rhythm, melody, energy variation, phrase structure, and dance patterns. Therefore, it involves challenges in temporal modeling, cross-modal alignment, and understanding of choreography.

Early approaches mainly treated dance generation as a mapping problem from music features to motion sequences. Li et al. [16] employed Transformer to capture long-term dependencies between music and motion through cross-modal attention. DanceNet (Music2Dance) [17] adopted a WaveNet-based temporal generation framework to predict continuous dance motions from music features. DanceRevolution [3] further introduced curriculum learning to gradually optimize long-sequence dance generation, reducing training difficulty and improving temporal stability.

However, these methods focused mainly on local temporal correlations and often suffered from accumulated errors and insufficient long-term consistency when generating extended dance sequences. To facilitate 3D dance generation, the AIST++ dataset and the FACT method [4] established a widely used benchmark with paired music and 3D motion capture data across different dance styles. FACT explored cross-modal Transformer learning between music and motion, enabling generated movements to better follow musical rhythms. The availability of AIST++ significantly improved the evaluation of motion quality, diversity, and beat synchronization.

As 3D dance generation developed, researchers gradually shifted their attention from direct music-motion mapping to the organization of choreography structures. Bailando [5] introduced a choreography memory module to represent continuous dance sequences using meaningful motion units and combined actor-critic GPT for motion composition. This work demonstrated that dance generation requires modeling of motion organization rather than only frame-level prediction. Nevertheless, existing motion-unit-based approaches mainly provide local generation guidance, while the global semantic evolution of long dance sequences remains insufficiently explored.

More recent work has explored controllable generation and multimodal conditions beyond conventional music-only settings. MoMask [18] introduced masked generative modeling for human motion synthesis and improved motion diversity through discrete motion representations. Although it mainly targets general motion generation, it provides a useful paradigm for dance synthesis. GCDance [19] incorporated dance categories as conditions to improve style control. Based on the OpenDance dataset, OpenDanceNet [20] explored large-scale multimodal dance generation by jointly utilizing music, text, keypoints, and character information.

Researchers have also started to consider richer human expressions and interactive scenarios instead of only single-person movements. SoulDance [21] jointly modeled body movements, hand gestures, and facial expressions through hierarchical motion representations. DuetGen [22] extended the task to two-person dance generation by learning the interaction and synchronization patterns between dancers.

Together, previous studies have substantially advanced dataset construction, cross-modal learning, and controllable synthesis. However, most approaches still emphasize local music-motion correspondence or external control signals, while long-term choreography semantics, such as motion theme continuity and phrase-level organization, remain inadequately modeled. Moreover, complex dance movements require coordinated changes among multiple body parts, which cannot be fully captured by temporal modeling alone. Addressing these issues requires better modeling of global choreography semantics, as well as local human body structures.

2.2 Long sequence dance generation based on diffusion models

Diffusion models have recently attracted considerable attention in human motion and dance generation due to their stable optimization and flexible sampling ability. Unlike autoregressive methods that generate motions sequentially, diffusion models recover motion sequences from noise through iterative denoising, providing better diversity for one-to-many generation problems [6]. Such characteristics are beneficial for music-driven dance generation, where a single musical sequence may correspond to multiple plausible choreographic choices.

EDGE [7] introduced diffusion models in music-driven dance generation by incorporating music conditions into the denoising process and enabling editable dance synthesis. DiffDance [8] further proposed a cascaded diffusion framework, where coarse motion sequences were refined through sequence super-resolution diffusion to improve temporal details and local motion quality.

However, directly applying diffusion models to long sequences introduces high computational costs and may weaken long-range structural consistency. LODGE [9]

addressed this issue through a coarse-to-fine diffusion framework. It first learned feature dance primitives at the global level and then generated detailed motions locally under music conditions. Lodge++ [10] further extended this idea to ultra-long sequences and multi-dance scenarios by enhancing choreography pattern learning and style constraints. These works suggest that intermediate representations can serve as a useful connection between global choreography planning and local motion synthesis.

In addition to long-term structure, physical plausibility and controllability have also attracted increasing attention. PAMD [11] introduced pose constraints and foot-ground contact refinement to alleviate abnormal poses and foot sliding. DanceLDM [23] explored latent diffusion for music- and text-conditioned dance generation and editing. Guo et al. [24] used the U-Net diffusion Transformer for keypoint-controlled 3D dance generation. FlowerDance [25] combined MeanFlow with physical consistency constraints to improve sampling efficiency while maintaining motion quality.

Despite the improvements brought about by diffusion-based approaches, several challenges have not been fully resolved. Existing coarse-to-fine frameworks usually employ dance primitives as guidance signals but lack explicit modeling of the semantic relationships behind these representations. Meanwhile, physical refinement methods focus mainly on foot contact and pose constraints, while high-order coordination among multiple joints is not explored sufficiently. Therefore, integrating semantic representation learning with human structure modeling is still necessary for high-quality long-sequence dance generation.

2.3 Semantic prototype learning and motion epresentation

Maintaining semantic consistency in long-sequence motion generation is challenging when relying only on continuous motion features. Similar motion patterns may correspond to different meanings depending on musical context, temporal position, and expressive intention. Therefore, extracting stable high-level semantic representations from motion sequences has become an important direction.

Prototype representations have also been explored in human motion analysis. ProtoFormer [29] extracted representative motion patterns through prototype attention, providing an intermediate representation between local motion features and high-level semantics. ProtoGCN [30] introduced prototypes in skeleton action recognition by learning core motion patterns and reconstructing action representations. MotionLCM [31] accelerated controllable motion generation through latent consistency modeling, while SemGeoMo [32] combined semantic and geometric guidance to improve motion generation.

For music-driven dance generation, intermediate motion representations have shown potential in improving long-sequence organization. Bailando [5] used choreography memory to learn representative motion units, while LODGE [9] introduced feature dance primitives as a bridge between global planning and local generation. Lodge++ [10] further extended this strategy to longer sequences and multiple dance styles. Despite these advances, existing motion primitives mainly serve as generation conditions rather than explicit semantic structures. The relationships among different motion patterns and their evolution throughout a dance sequence remain largely unexplored. In long-sequence dance generation, global choreography is dependent not

only on individual motion fragments but also on musical phrases, rhythm transitions, and thematic development.

Based on this observation, we develop SPR module to provide a more explicit semantic representation for long-sequence dance generation. Instead of treating prototypes as fixed motion fragments, SPR constructs a learnable semantic prototype space and performs feature querying, matching, and reconstruction in the latent motion space. These prototypes represent high-level choreography patterns, such as rhythmic movement tendencies and transitions between motion themes, providing a more stable global guidance for long-sequence generation.

2.4 Human Body Structural Modeling and Pose Refinement

High-quality human motion generation requires both temporal continuity and structural consistency with human anatomy. In music-driven dance generation, complex movements such as rotation, jumping, and weight transfer involve coordinated changes among the torso, limbs, and feet. Therefore, temporal modeling alone is insufficient to capture spatial dependencies within human motion.

Parametric human models provide effective structural priors for 3D motion generation. SMPL [12] represents human pose and shape using low-dimensional parameters and has been widely applied in reconstruction and motion synthesis. Recent human pose and mesh recovery approaches further demonstrate the importance of effective structural modeling for improving human body estimation [33]. SMPL-X [34] extends this representation by incorporating hand and facial expressions. DPoser-X [35] further learns human pose priors through diffusion models to improve complex pose modeling. However, these models mainly describe body structure and cannot fully capture dynamic cooperation among multiple body parts.

Foot contact modeling is another important factor affecting motion quality. Dance movements often contain rapid rotations, jumps, and weight shifts, where inaccurate contact estimation may cause foot sliding artifacts. Li et al. [13] alleviated this problem by explicitly modeling ground-contact events and constraining foot velocity during contact phases. PAMD [11] also demonstrated the importance of contact refinement for improving motion realism.

Graph neural networks have been widely applied in human pose and skeleton-based motion modeling by representing joints as nodes and skeletal connections as edges [36]. However, conventional graphs mainly describe pairwise relationships, while many human actions are generated through cooperation among multiple joints. Such high-order interactions cannot be fully represented by ordinary graph structures. Hypergraph representations offer another perspective by modeling relationships among multiple joints through hyperedges. HGNN [14] demonstrated the capability of hypergraphs in modeling high-order structural relationships. Hyper-GCN [37] captured multi-joint dependencies through adaptive hypergraph structures. Liu et al. [38] introduced semi-dynamic hypergraph learning for 3D pose estimation. AMHGCN [39] and H-DHGCN [40] further explored adaptive and directional hypergraph representations. HyperDiff [41] combined hypergraph structures with diffusion models, demonstrating their potential in generative human modeling.

While hypergraph-based methods have shown effectiveness in pose estimation and motion prediction, their application in music-driven dance generation remains limited. Existing approaches mainly focus on music conditions, motion primitives, or foot constraints, while explicit modeling of high-order body coordination is still lacking. During long-sequence generation, local structural errors may accumulate and result in unnatural limb relationships and unstable motion details.

Motivated by this limitation, we design HEPR for local motion optimization. By constructing dynamic hypergraphs, HEPR models high-order coordination among the torso, limbs, and feet beyond traditional skeleton connections. Combined with foot-contact constraints, it improves local structural consistency and motion naturalness. Together with SPR, the proposed framework enhances long-sequence dance generation from both global semantic organization and local structural refinement perspectives.

3 Methods

3.1 Overview of SPH-DiffDance

In this paper, we propose SPH-DiffDance, a long-sequence dance generation framework that incorporates semantic prototype reconstruction for global choreography coherence and hypergraph-enhanced pose refinement for local structural quality. The overall architecture of our framework is shown in Fig. 1.

Given a music feature sequence with L frames and a dance category label g, we first divide the complete sequence into k global segments, where each segment contains N frames. In the global generation stage, we employ a conditional diffusion model to gradually recover coarse-grained dance motions from noisy sequences under the guidance of music conditions and category labels. To alleviate semantic drift during long-sequence generation, we introduce a semantic prototype reconstruction (SPR) module, where we construct a learnable prototype memory bank and adaptively retrieve semantic prototypes to enhance the consistency of music rhythm, dance style, and motion arrangement. After obtaining the global coarse-grained sequence $d \in \mathbb{R}^{L\times 139}$, we further perform local refinement. Specifically, we represent each motion frame using the SMPL human body model and select representative dance primitives according to motion characteristics, prototype response intensity, and pose diversity. The selected primitives are further divided into hard cues d_h and soft cues d_s. The hard cues preserve structural continuity between local segments, whileas the soft cues provide semantic guidance for generating fine-grained motions. Based on these constraints, we generate multiple local dance fragments through the local diffusion process. To improve the spatial coordination and physical rationality of human motion, we further introduce a hypergraph-enhanced pose refinement (HEPR), which models high-order dependencies among body joints and incorporates foot-contact constraints to refine local poses. Finally, we fuse the generated fragments into a complete long dance sequence while maintaining music synchronization and motion continuity.

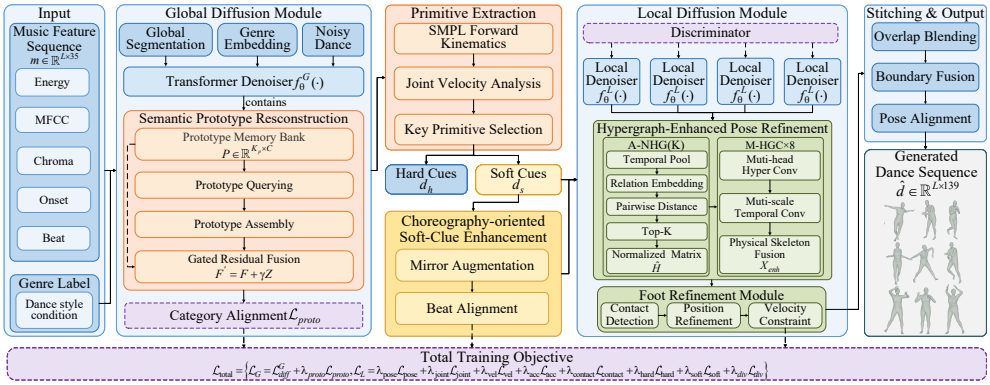


Fig. 1 Overall architecture of SPH-DiffDance.

3.2 Problem definition and symbolic representation

Given a music sequence with frames L, we first extract the corresponding acoustic features through Librosa and represent them as a music condition matrix $m \in \mathbb{R}^{L\times 35}$. The 35-dimensional feature consists of 1-dimensional energy envelope, 20-dimensional MFCC, 12-dimensional Chroma, 1-dimensional onset, and 1-dimensional beat. These features describe local variations, timbre structures, and rhythmic positions in music, providing multi-level conditions for subsequent dance generation.

The corresponding dance motion sequence is represented using the SMPL human body model. Each dance pose is represented as a 139-dimensional vector, where the first 4 dimensions describe foot-contact states, 3 dimensions represent root translation, and the remaining 132 dimensions represent human joint rotation parameters. In this work, we use a continuous 6D rotation representation to describe joint poses, including global root rotation and local rotations of 21 body joints. Meanwhile, the root translation parameters are retained to preserve spatial continuity and structural stability. Therefore, the complete dance sequence can be represented as $d \in \mathbb{R}^{L\times 139}$.

During long-sequence dance generation, directly modeling the complete sequence may cause semantic drift and local structural instability. Therefore, we divide the dance sequence into k global segments, where each segment contains N frames and satisfies $L = kN$. The i-th music segment and its corresponding motion segment are denoted as m^i and d^i, respectively. Each motion segment is further divided into local fragments d^i_j.

During diffusion-based generation, d_t represents the noisy dance sequence in diffusion step t. Our objective is to recover dance sequences that satisfy both music semantics and human motion constraints under a given music condition m, diffusion step t, and category label g. This formulation provides the basis for subsequent global semantic prototype diffusion and hypergraph-enhanced pose refinement.

3.3 Global semantic prototype diffusion

To capture the correspondence between long-term music structures and dance movements, we first construct a global semantic prototype diffusion module to generate

coarse-grained dance sequences. Since long dance sequences usually contain complex temporal dependencies, directly learning the entire motion distribution may lead to semantic inconsistency and discontinuous choreography. Therefore, we adopt a conditional diffusion framework to gradually recover dance motions from noisy states.

For a real dance sequence d_0, the forward diffusion process gradually adds Gaussian noise to destroy the original motion distribution. In the diffusion step t, the noisy sequence d_t is obtained as:

$$q(d_t \mid d_0) = \mathcal{N}\left(d_t; \sqrt{\overline{\alpha}_t} d_0, (1 - \overline{\alpha}_t) I\right) \tag{1}$$

where $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\alpha_t = 1 - \beta_t$, and β_t denotes the predefined noise scheduling coefficient. As t increases, the original motion structure is gradually destroyed, and d_t approaches a standard Gaussian distribution. During the reverse process, we learn a conditional denoising network to recover clean dance motions from noisy sequences.

Given the noisy dance sequence d_t, diffusion timestep t, music condition m, and category label g, the denoising network predicts the clean motion sequence:

$$\hat{d}_0 = f_0(d_t, t, m, g) \tag{2}$$

The basic training objective of the diffusion model is defined as follows:

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{d_0, t}\left[\|d_0 - f_0(d_t, t, m, g)\|_2^2\right] \tag{3}$$

To further enhance the controllability of music-driven dance generation, we introduce music features and category information as conditional inputs. The final global dance sequence is predicted as:

$$\hat{d}_0^g = f_0^G(d_t^g, t, e_g, g) \tag{4}$$

Where f_θ^G represents the global diffusion denoising network and e_g represents the semantic embedding of the category.

However, diffusion models mainly focus on motion distribution recovery and may ignore high-level semantic patterns shared among different dance categories. To address this limitation, we propose a SPR module that introduces a learnable prototype memory bank to enhance global motion semantics. As shown in Fig. 2, the Transformer output feature is represented as $F' \in \mathbb{R}^{B \times N \times C}$, where B, N, and C denote batch size, sequence length, and feature dimension, respectively. We construct a prototype memory bank $P \in \mathbb{R}^{K_p \times C}$, where each prototype vector represents a latent dance semantic pattern learned during training.

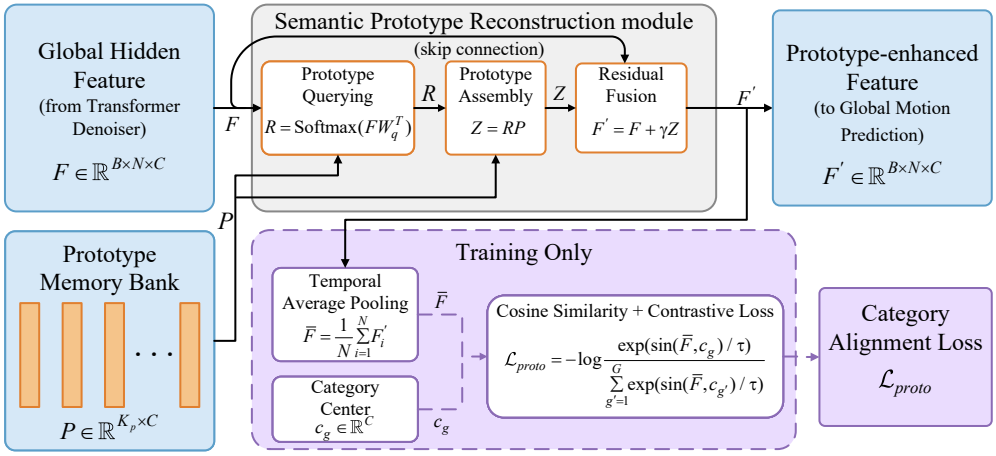


Fig. 2 Architecture of the Semantic Prototype Reconstruction Module

To establish the correspondence between motion features and semantic prototypes, we calculate the matching weights:

$$R = \text{Softmax}(FW_q^T), \quad R \in \mathbb{R}^{B\times N\times K_p} \tag{5}$$

where W_q is a learnable mapping matrix. The prototype features are then aggregated according to the matching weights:

$$Z = RP, \quad Z \in \mathbb{R}^{B\times N\times C} \tag{6}$$

This process enables the model to retrieve high-level semantic information from the prototype memory bank. We further fuse the original diffusion features and prototype features through residual fusion :

$$F^{'} = F + \gamma Z \tag{7}$$

where γ denotes the learnable fusion coefficient.

To improve semantic consistency among different dance categories, we introduce category semantic constraints. The enhanced feature is temporally averaged:

$$\bar{F} = \frac{1}{N}\sum_{i=1}^{N} F_i' \tag{8}$$

Each category in the training set maintains a semantic center $c_g \in \mathbb{R}^C$. Then, we optimize the category prototype alignment loss:

$$\mathcal{L}_{\text{proto}} = -\log \frac{\exp(\sin(\bar{F}, c_g)/\tau)}{\sum_{g'=1}^{G}\exp(\sin(\bar{F}, c_{g'})/\tau)} \tag{9}$$

where c_g represents the semantic center of category g, $sim(\cdot)$ denotes cosine similarity, and τ is the temperature coefficient.

After prototype reconstruction, we further extract representative dance primitives from the global coarse-grained motion sequence. These primitives serve as intermediate cues between global choreography planning and local motion generation. We first convert the generated SMPL parameters into 3D joint positions through forward kinematics. Let $p_{i,j}$ denote the position of the j-th joint in frame i. The average joint velocity of frame i is calculated as:

$$v_i = \frac{1}{J}\sum_{j=1}^{J}\|p_{i,j} - p_{i-1,j}\|_2 \tag{10}$$

where J denotes the number of human joints. Key poses usually occur near changes in movement direction, brief pauses, or transitions in motion energy. We therefore smooth the velocity sequence and detect its local extrema to obtain a candidate key-frame set:

$$\mathcal{D}_{\text{cand}} = \{d_i^{\text{cand}}\}_{i=1}^{M} \tag{11}$$

Selecting primitives only according to velocity extrema may retain redundant poses, especially when adjacent candidates exhibit similar body configurations. We therefore jointly consider motion saliency, prototype response intensity, and pose diversity. The final primitive set is selected by:

$$D_{\text{prim}} = \underset{\mathcal{S}\subset\mathcal{D}_{\text{cand}},|\mathcal{S}|=K}{\arg\max}\left(\sum_{d_i\in\mathcal{S}} E(d_i) + \lambda_p \sum_{d_i\in\mathcal{S}} R(d_i) - \lambda_r \sum_{d_i,d_j\in\mathcal{S}} \sin(d_i, d_j)\right) \tag{12}$$

where K is the number of selected primitives and is set to 8 in our experiments. $E(d_i)$ measures the motion saliency of the candidate frame d_i, while $R\,(d_i)$ denotes its response to the learned semantic prototypes. The last term suppresses highly similar candidates, preventing the selected primitive set from being dominated by repeated poses. In this way, the resulting primitives retain salient motion changes while covering different semantic patterns and pose configurations.

We further divide the selected feature dance primitives into hard cues d_h and soft cues d_s. Hard cues are placed near local segment boundaries to preserve structural continuity during fragment generation. Soft cues are used inside local segments to provide semantic and stylistic guidance, allowing the local diffusion stage to generate more expressive motions at rhythmically important positions. This division connects the global choreography representation learned in this section with the local conditional generation process introduced in the following section.

The global stage is optimized using the diffusion denoising loss together with the category semantic prototype constraint:

$$\mathcal{L}_G = \mathcal{L}_{diff}^{G} + \lambda_{proto}\mathcal{L}_{proto} \tag{13}$$

where $\mathcal{L}_{\text{diff}}^{G}$ denotes the denoising objective of the global diffusion network, and $\mathcal{L}_{\text{proto}}$ is the category semantic prototype loss defined in Eq.(9).

This objective encourages the model to recover a stable coarse-grained motion structure while preserving category-related dance semantics. The generated sequence

and the extracted cues then provide global motion priors for the subsequent local conditional generation and HEPR module.

3.4 Local conditional generation and pose refinement

The global semantic prototype diffusion stage produces a coarse-grained dance sequence with relatively stable choreography and category semantics. However, its long-range modeling objective makes it difficult to preserve all short-term motion details. Local errors may therefore appear in joint coordination, pose transitions, and limb continuity, especially during rotations, jumps, or rapid changes in body direction. Such movements involve coordinated variations among the torso, arms, and lower limbs, which cannot be fully represented by conventional pairwise skeleton connections. To refine these details, we introduce a local generation stage that progressively enhances short-term motion quality through cue-guided conditional diffusion and hypergraph-based pose refinement.

3.4.1 Soft-Cue Enhancement with Choreographic Priors

The global stage outputs hard cues d_h and soft cues d_s. Hard cues are mainly placed near fragment boundaries to maintain structural continuity, while soft cues describe representative poses inside the fragments. Although the original soft cues retain the general motion tendency of the global sequence, they do not explicitly account for left–right symmetry, repeated choreographic patterns, or alignment with salient musical events. We therefore enhance the soft cues before local diffusion:

$$d_s^* = \mathcal{A}(d_s, m) \tag{14}$$

Where $\mathcal{A}(\cdot)$ denotes the soft-cue enhancement function and m is the corresponding music condition. The enhancement consists of mirror augmentation and beat alignment:

$$d_s^* = \mathcal{A}_{beat}\left(\mathcal{A}_{mir}(d_s), m\right) \tag{15}$$

Where $\mathcal{A}_{mir}$ expands the range of choreographic expressions by exploiting the bilateral symmetry of the human body, while $\mathcal{A}_{beat}$ adjusts the temporal positions of the cues according to beat and onset information. Let the original soft-cue set be:

$$d_s = \{(s_i, t_i)\}_{i=1}^{r}, \quad s_i \in \mathbb{R}^{139} \tag{16}$$

where s_i is the i-th soft-cue pose and t_i is its temporal position. Based on the SMPL joint topology, we define a left–right joint permutation function $\pi(\cdot)$ and a mirror transformation R_{mir}. The mirrored pose is written as:

$$\tilde{s}_i = \mathcal{R}_{\min}\left(\pi(s_i)\right) \tag{17}$$

The original and mirrored cues are then combined into an augmented candidate set:

$$d_s^{\min} = d_s \cup \{(\tilde{s}_i, t_i)\}_{i=1}^{r} \tag{18}$$

This augmentation does not force the model to generate strictly symmetric motions. Instead, it provides additional choreographic alternatives, making it easier for the local diffusion model to respond to repeated phrases and direction changes while reducing unnatural left–right inconsistencies.

3.4.2 Beat-alignment constraint

Mirror augmentation enriches the spatial form of the soft cues, but their temporal positions may still be inconsistent with the musical rhythm. This mismatch is particularly noticeable around accents, impacts, and direction changes, where a visually meaningful pose may appear slightly before or after the corresponding musical event. We therefore align each soft cue with a nearby beat or onset. Let the beat positions in a local music segment be $\mathcal{B} = \{b_1, b_2, ..., b_q\}$ and let the onset positions be $\mathcal{O} = \{o_1, o_2, ..., o_u\}$. The two sets are merged into the set of musically salient positions :

$$\mathcal{T} = \mathcal{B} \cup \mathcal{O} \tag{19}$$

For the temporal position t_i of a soft cue, the nearest salient music position is determined by:

$$t_i^* = \arg\min_{\tau \in \mathcal{T}} |t_i - \tau| \tag{20}$$

Directly moving every cue to the nearest musical event may alter its original action context. We therefore limit the temporal displacement using a maximum offset δ:

$$t_i^* = t_i + \text{clip}(t_i^* - t_i, -\delta, \delta) \tag{21}$$

After alignment, the enhanced soft cue set is expressed as:

$$d_s^* = \{(s_i^*, t_i^*)\}_{i=1}^{r^*} \tag{22}$$

The aligned cues specify not only what type of motion should be emphasized, but also when it should occur. They therefore provide more explicit temporal guidance for the subsequent local generation stage without replacing the original motion context.

3.4.3 Local conditional diffusion generation

Given an enhanced soft-cue set, we use a local diffusion model to generate fine-grained dance fragments. For a global music segment m_g^i, it is divided into h local segments $m_g^i = \{m_g^{i,1}, m_g^{i,2}, \ldots, m_g^{i,h}\}, m_g^{i,j} \in \mathbb{R}^{n \times 35}$. For the j-th local segment m_j^l, the denoising network predicts the clean motion under the conditions of local music, hard cues, enhanced soft cues, and the category label:

$$\hat{d}_{0,j}^l = f_\theta^L(d_{t,j}^l, t, m_j^l, d_h^j, d_s^{j,^*}, g) \tag{23}$$

where $d_{t,j}^l$ represents the noisy action of the j-th local segment at diffusion step t, while d_h^j and $d_s^{j,*}$ are the hard and enhanced soft cues assigned to that fragment.

For each fragment, we construct a hard-cue condition matrix $C_h^j \in \mathbb{R}^{n \times 139}$ and a binary mask $M_h^j \in \mathbb{R}^{n \times 1}$. Hard cues are mainly placed near the beginning and end of the fragment:

$$C_h^j[\tau] = h_r, \quad M_h^j[\tau] = 1 \tag{24}$$

where h_r denotes the hard clue pose mapped to the local time position t. The remaining positions are not directly constrained, allowing the local diffusion model to generate transitional motions instead of simply copying the boundary poses. Soft cues are represented in the same form, using C_s^j and M_s^j, but are applied only around selected key moments.

The local forward diffusion process is defined as :

$$q(d_{t,j}^l \mid d_{0,j}^l) = \mathcal{N}\left(d_{t,j}^l; \sqrt{\overline{\alpha}_t} d_{0,j}^l, (1-\overline{\alpha}_t)I\right) \tag{25}$$

The corresponding reverse process predicts the clean fragment:

$$\hat{d}_{0,j}^l = f_\theta^L(d_{t,j}^l, t, \mathcal{C}^j) \tag{26}$$

Before temporal modeling, the noisy motion, music condition, hard cues, and soft cues are projected into the hidden space:

$$H_0 = \phi_d(d_{t,j}^l), \quad E_m = \phi_m(m_j^l), \quad E_h = \phi_h(C_h^j), \quad E_s = \phi_s(C_s^j) \tag{27}$$

The local Transformer captures the temporal dependencies within each fragment and injects the conditioning information through feature fusion:

$$H_{r+1} = \text{Block}\left(H_r, E_m, t, g\right) + \eta_h M_h^j \odot E_h + \eta_s(t) M_s^j \odot E_s \tag{28}$$

where η_h and η_s control the strengths of the hard and soft cues, respectively. The hard-cue weight remains relatively large to preserve fragment boundaries, whereas the soft-cue weight varies with the diffusion step. Stronger soft guidance is applied during the early denoising stage to establish the main action pattern, and the guidance is gradually weakened as the model begins to recover local details.

After several local Transformer blocks, the initial local pose sequence is obtained as:

$$X_{\text{pose}} = W_o H_R + b_o, \quad X_{\text{pose}} \in \mathbb{R}^{B\times n\times 139} \tag{29}$$

This sequence has already integrated local music and global choreography cues. Nevertheless, its spatial structure is still learned mainly through implicit feature interactions. To explicitly capture high-order cooperation among human joints while retaining the physical connections of the SMPL skeleton, we further process X_{pose} using the HEPR module described below.

3.4.4 Hypergraph-enhanced pose refinement

The overall architecture of the HEPR module is illustrated in Fig. 3. Given the initial local pose sequence $X_{pose} \in \mathbb{R}^{B\times n\times 139}$, we first project each frame into a joint-level feature space:

$$X_{\text{joint}} = \text{Reshape}(X_{\text{pose}} W_{\text{in}}), \quad X_{\text{joint}} \in \mathbb{R}^{B\times n\times J\times C_j} \tag{30}$$

where J represents the number of human joints, C_j represents the hidden feature dimension of each joint, and W_{in} is the mapping parameter from the pose space to the joint feature space. This operation converts the original 139-dimensional pose vector into a joint-level representation, so that the subsequent hypergraph module can model the spatial collaboration relationship at the human joint level.

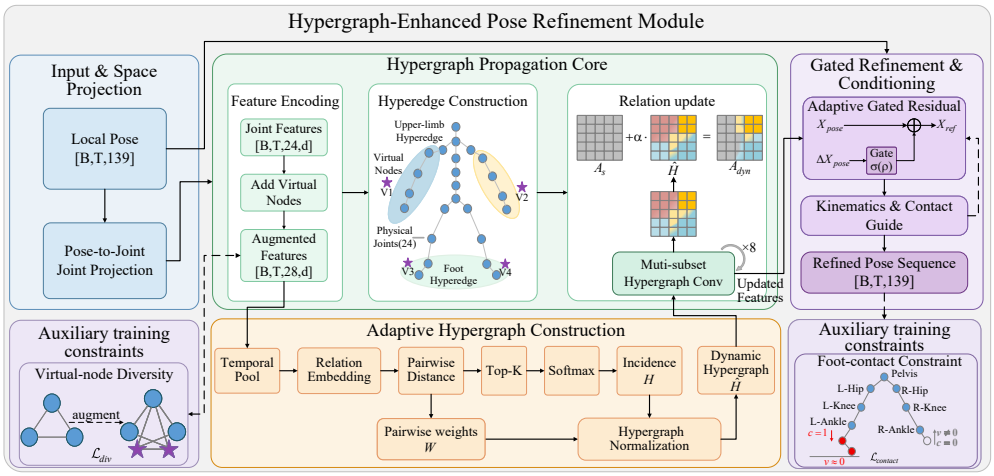


Fig. 3 Architecture of the Hypergraph-Enhanced Pose Refinement Module

In order to construct a hypergraph structure suitable for local dance movements, we regard each frame of human joint as a hypergraph node, and define the hypergraph as $\mathcal{G}_h = (\mathcal{V}_h, \mathcal{E}_h, W_h)$, where $\mathcal{V}_h$ represents the set of joint vertices, $\mathcal{E}_h$ represents the set of hyperedges, and W_h represents the hyperedge weight matrix. Different from the fixed human skeleton, the hypergraph structure allows a hyperedge to connect multiple related joints at the same time, which can describe the high-order synergy in dance movements. Specifically, for the joint features $X = \{x_1, x_2, \ldots, x_J\}, \quad x_i \in \mathbb{R}^{C_j}$, the feature distance $m_{i,j} = \|x_i - x_j\|_2$ between any two joints is first calculated, where $m_{i,j}$ represents the feature difference between the i-th joint and the j-th joint. In order to avoid each hyperedge containing too many irrelevant nodes, we select K joints that are most relevant to the current node according to the distance relationship to construct a hyperedge. Denote by set_i the nearest neighbor set corresponding to the i-th joint, then the hypergraph incidence matrix $H \in \mathbb{R}^{J \times J}$ is defined as:

$$h_{i,j} = \begin{cases} \frac{\exp(-m_{i,j})}{\sum_{k \in \text{set}_i} \exp(-m_{i,k})}, & j \in \text{set}_i \\ 0, & j \notin \text{set}_i \end{cases} \tag{31}$$

where $h_{i,j}$ denotes the degree to which the j-th joint belongs to the hyperedge centered on the i-th joint.

In this way, the model can dynamically adjust the connection relationship between joints according to the current action state. For example, the shoulder, torso and lower limbs may form a stronger association in rotation ; in the arm swing, the relationship

between the shoulder, elbow and wrist is more prominent. After obtaining the incidence matrix H, the vertex degree matrix and the hyperedge degree matrix are defined :

$$D_v(i,i) = \sum_{e \in \mathcal{E}_h} W_h(e,e) H(i,e) \tag{32}$$

$$D_e(e,e) = \sum_{\nu \in \mathcal{V}_h} H(v,e) \tag{33}$$

where the vertex degree matrix is used to describe the importance of nodes participating in the hyperedge connection, and the hyperedge degree matrix represents the number of nodes contained in the hyperedge.

Then, the hypergraph propagation matrix is normalized by $\hat{H} = D_v^{-\frac{1}{2}} H W_h D_e^{-1} H^T D_v^{-\frac{1}{2}}$, where $\hat{H} \in \mathbb{R}^{J \times J}$ represents the normalized hypergraph structure. Through the normalization operation, the information propagation between different joints and different super edges is more stable. In the process of hypergraph convolution, the joint features are propagated according to the normalized hypergraph structure :

$$X_{joint}^{(l+1)} = \sigma\left(\hat{H} X_{joint}^{(l)} W_g^{(l)}\right) \tag{34}$$

where $W_g^{(l)}$ is a learnable parameter and $\sigma(\cdot)$ is a nonlinear activation function. This process can aggregate the information of multiple joints in the same hyperedge and propagate the high-order cooperative relationship to each node.

Considering that human motion itself has a stable physical connection, relying solely on adaptive hypergraphs may ignore some basic motion constraints. Therefore, we further introduces the human skeleton prior matrix A_s to describe the natural joint connection relationship in the SMPL skeleton. The corresponding normalized skeleton adjacency matrix is $\hat{A}_s = D_s^{-\frac{1}{2}} A_s D_s^{-\frac{1}{2}}$, where D_s is the degree matrix of the physical skeleton graph. Subsequently, the physical skeleton graph and the adaptive hypergraph information are fused:

$$X_{enh} = \sigma\left((\hat{A}_s + \alpha \hat{H}) X_{joint} W_e\right) \tag{35}$$

where α balances the contributions of the fixed skeleton structure and the learned hypergraph, and W_e is a learnable feature transformation. This fusion preserves anatomical constraints while allowing the joint relationships to vary with the current motion. This fusion preserves anatomical constraints while allowing the joint relationships to vary with the current motion:

$$\Delta X_{pose} = \text{Flatten}(X_{enh}) W_{out} \tag{36}$$

where W_{out} is the mapping parameter from joint space to attitude space. In order to avoid the hypergraph branch causing too strong disturbance to the diffusion backbone generation result, we uses the gated residual method to refine the pose:

$$X_{ref} = X_{pose} + \sigma(\rho) \Delta X_{pose} \tag{37}$$

where ρ is the learnable gating parameter, and $\sigma(\rho) \in (0, 1)$ controls the injection strength of the hypergraph enhancement result.

Due to the introduction of learnable virtual nodes in the hypergraph enhancement process, different virtual nodes may learn highly similar motion patterns, thereby reducing the ability of hypergraphs to express different human collaborative relationships. Therefore, we further introduces the virtual node diversity constraint. Let the virtual node feature in the l-th layer hypergraph be $H^l = \{h_i^l\}_{i=1}^{V_l}$. First normalized:

$$\overline{h}_i^l = \frac{h_i^l}{\parallel h_i^l \parallel_2} \tag{38}$$

Then, define the virtual node diversity loss:

$$\mathcal{L}_v = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{V_l(V_l - 1)} \sum_{i \neq j} \frac{1}{\parallel \overline{h}_i^l - \overline{h}_j^l \parallel_2^2 + \epsilon} \tag{39}$$

where V_l denotes the number of virtual nodes and L denotes the number of hypergraph layers. This constraint can promote different virtual nodes to learn the motion patterns of different regions and improve the expression ability of hypergraph structure.

3.4.5 Foot refinement and local stability constraints

During dance generation, foot-contact consistency is important for maintaining realistic human motion. Since the diffusion model mainly focuses on overall motion generation, local foot sliding may still occur after pose refinement. Therefore, we introduce a foot-contact constraint after the HEPR module. Given the refined pose sequence X_{ref}, we obtain the 3D foot positions p_i^f through SMPL forward kinematics. The foot velocity is calculated as:

$$v_i^f = p_i^f - p_{i-1}^f \tag{40}$$

For frames predicted to be in contact with the ground, the foot velocity should remain close to zero. Therefore, we define the foot-contact loss as:

$$\mathcal{L}_{contact} = \frac{1}{n-1} \sum_{i=2}^{n} \left\| \hat{b}_i \odot v_{i,hor}^f \right\|_2^2 + \lambda_{\text{down}} \left\| \hat{b}_i \odot \min(v_{i,\text{ver}}^f, 0) \right\|_2^2 \tag{41}$$

where $\hat{b}_i$ represents the predicted foot-contact state, $\nu_{i,hor}^f$ and $\nu_{i,ver}^f$ denote the horizontal and vertical foot velocities, respectively. This constraint reduces foot sliding artifacts while preserving natural non-contact movements.

The overall local optimization objective combines pose reconstruction, joint motion constraints, foot-contact consistency, and cue-guided generation constraints:

$$\begin{aligned} \mathcal{L}_L = {} & \lambda_{pose}\mathcal{L}_{pose} + \lambda_{joint}\mathcal{L}_{joint} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{acc}\mathcal{L}_{acc} + \lambda_{contact}\mathcal{L}_{contact} + \\ & \lambda_{hard}\mathcal{L}_{hard} + \lambda_{soft}\mathcal{L}_{soft} + \lambda_{div}\mathcal{L}_{div} \end{aligned} \tag{42}$$

where each term constrains different aspects of local motion quality, including pose accuracy, temporal smoothness, contact stability, cue consistency, and hypergraph representation diversity.

3.4.6 Local fragment fusion and long-sequence recovery

After local diffusion and HEPR refinement, we obtain multiple generated motion fragments $\{\hat{d}_{0,l}^1, \hat{d}_{0,l}^2, ..., \hat{d}_{0,l}^h\}$. Since directly concatenating fragments may introduce discontinuities at boundaries, we adopt a linear blending strategy in overlapping regions:

$$\hat{d}_{blend}[\tau] = w(\tau)\hat{d}_{0,l}^j[\tau] + (1 - w(\tau))\hat{d}_{0,l}^{j+1}[\tau], \quad \tau = 1, \ldots, r_o \tag{43}$$

where $w(\tau)$ is a fusion weight that gradually decreases from 1 to 0. This smooth transition preserves the continuity between adjacent fragments.

Finally, all local fragments are arranged according to their temporal order to recover the complete dance sequence:

$$\hat{d} = \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_L\}, \quad \hat{d} \in R^{L \times 139} \tag{44}$$

where $\hat{d}$ denotes the final generated dance sequence.

In summary, the HEPR module achieves progressive optimization from global action semantics to local motion details through soft cue enhancement, beat constraints, local conditional diffusion, and hypergraph pose refinement. It improves the plausibility of human body structures and the continuity of dance sequences while ensuring musical synchronization.

3.5 Training objectives

SPH-DiffDance consists of two sequential stages: global semantic prototype diffusion and local pose refinement. The global stage focuses on learning the correspondence between music structures and coarse-grained dance patterns, while the local stage improves short-term motion details through cue guidance and pose constraints. Because the two stages optimize different objectives, we adopt a sequential training strategy. The overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \{\mathcal{L}_G, \mathcal{L}_L\} \tag{45}$$

where $\mathcal{L}_G$ represents the optimization objective of the global diffusion stage, and $\mathcal{L}_L$ represents the objective of the local refinement stage.

During training, we first optimize the global semantic prototype diffusion module to obtain stable global motion representations. Then, we fix the global-stage parameters and use the generated sequences to extract dance primitives, hard cues d_h, and enhanced soft cues d_s. These cues are subsequently used to train the local diffusion and HEPR modules. This decoupled strategy reduces the interference between global planning and local refinement objectives.

Through the coarse-to-fine generation strategy, SPH-DiffDance achieves a balance between long-term choreography consistency, music synchronization, and local motion realism.

4 Experiments

4.1 Datasets and evaluation metrics

4.1.1 Datasets

We use the AIST++ dataset [4] as the experimental benchmark for training and testing. AIST++ is a widely used dataset for music-driven 3D dance generation, constructed from the AIST Dance video database and contains 1,408 paired music-dance 3D motion sequences covering 10 different dance styles. It provides a unified benchmark for studying the cross-modal relationship between music features and human movements.Following previous work, the original 60 FPS motion sequences are uniformly downsampled to 30 FPS during training to reduce the computational cost of the long-sequence diffusion model while preserving the temporal correspondence between music and motion. During testing, the generated 30 FPS dance sequences are interpolated back to 60 FPS and evaluated following the requirements of existing methods.

4.1.2 Evaluation metrics

To evaluate the generated dances from different perspectives, we consider three important aspects of the generation process: motion quality, diversity, and music-motion synchronization. Accordingly, FID, DIV, and BAS are adopted as the evaluation metrics in our experiments.

Fréchet Inception Distance (FID) measures the difference between the distribution of generated motions and that of ground truth. Since the quality of a generated dance sequence is not only determined by its overall appearance but also closely related to its dynamic variations and spatial structures, we calculate FID in two feature spaces, including the motion dynamics feature space and the geometric feature space, which are denoted as FID_k and FID_g,respectively. The metric mainly reflects dynamic characteristics such as motion velocity and acceleration, while focuses on the rationality of human poses and spatial configurations. A lower FID value indicates that the generated motions have a distribution closer to ground truth.

Beyond motion quality, diversity is also an essential factor in evaluating dance generation models, since a model that produces only similar motion patterns may achieve reasonable quality scores but still fail to capture the richness of human dance. Therefore, Diversity (DIV) is employed to measure the variation among generated motions. In our experiments, the average Euclidean distance is calculated in both dynamic and geometric feature spaces, represented as DIV_k and DIV_g. A higher DIV value suggests that the model is capable of generating more diverse motion patterns while reducing the tendency toward repetitive output.

The relationship between music and motion is another key aspect of music-driven dance generation. To evaluate whether the generated movements can follow the rhythmic characteristics of the input music, the Beat Alignment Score (BAS) is introduced to measure the temporal consistency between motion beats and music beats. The music beats are extracted using Librosa, while motion beats are determined according to the

local minima of motion speed variations. A higher BAS value indicates stronger synchronization between the generated dance movements and the corresponding musical rhythm.

4.2 Implementation Details

The model in this paper is implemented based on the PyTorch framework, and training and testing are conducted in the NVIDIA GPU environment. In the experiments, the length of each AIST++ sequence is set to 180 frames, with 139-dimentional motion features, 24 human joints, 35-music features, and 10 categories of dance. The global diffusion is trained for 4000 epochs with a batch size of 128. The Adan optimizer is adopted with an initial learning rate of 4×10^{-4}. The hidden dimension of the Transformer encoder is set to 512, including an 8-layer network structure, 8 attention heads, and a dropout rate of 0.1. For the local diffusion stage, the model is trained for 800 epochs with a batch size of 64 and a learning rate of 3×10^{-4}. The HEPR module contains 4 virtual joint points, 8 hypergraph subsets, a joint feature dimension of 16, and a hypergraph loss weight of 0.01. During the local fine-tuning, pretrained diffusion model parameters are loaded for further optimization. The batch size is set to 32, and the learning rate is reduced to 3×10^{-5} to reduce structural deviation during the fine-tuning.

4.3 Comparison to existing methods

Under the same data split and evaluation protocol, we compare the proposed method with several mainstream music-driven dance generation methods. The comparison results are shown in Table 1. Overall, our method achieves a relatively balanced performance in terms of motion quality, action diversity and music-motion synchronization.

From the perspective of action quality, the proposed method achieves favorable FID scores. Specifically, the FID_g value of our method is 10.60, which is only slightly higher than that of ground truth (9.50) and significantly lower than FACT (22.11), EDGE (22.12), DiffDance (20.68), LODGE (18.79), and PAMD (17.46). This indicates that the generated motions have a spatial structure distribution closer to real dance sequences and can maintain more reasonable human poses. Meanwhile, the proposed method obtains an FID_k value of 35.82. Although this does not achieve the best result among all compared methods, it still outperforms several existing approaches, including FACT, DanceNet, and LODGE. This suggests that the model preserves the ability to generate dynamic motion variations.

It is worth noting that motion quality is influenced by multiple factors, including both motion realism and generation diversity. Therefore, a single metric cannot fully describe the overall performance of a generation model. By jointly optimizing semantic prototype reconstruction and hypergraph-based pose refinement, our method achieves a better balance between structural rationality, motion variation, and responsiveness to musical changes.

Table 1 Performance comparison of different methods on the AIST++ dataset

Method	Motion Quality		Motion Diversity		$BAS\uparrow$
	$FID_k\downarrow$	$FID_g\downarrow$	$DIV_k\rightarrow$	$DIV_g\rightarrow$	
Ground Truth	31.70	9.50	8.98	7.42	0.2649
Li et al. [16]	86.43	43.46	6.85	3.32	0.1607
DanceNet [17]	69.18	25.49	2.86	2.85	0.1430
DanceRevolution [3]	73.42	25.92	3.52	4.87	0.1950
FACT [4]	35.35	22.11	5.94	6.18	0.2209
Bailando [5]	28.16	**9.62**	7.83	6.34	0.2332
EDGE [7]	42.16	22.12	3.96	4.61	0.2334
DiffDance [8]	**24.09**	20.68	6.02	2.89	0.2418
MoMask [18]	44.92	26.20	4.67	1.06	0.2312
LODGE [9]	37.09	18.79	5.58	4.85	0.2423
GCDance [19]	30.93	18.25	5.22	<u>6.71</u>	0.2354
PAMD [11]	42.67	17.46	4.99	5.00	<u>0.2700</u>
OpenDanceNet [20]	<u>24.82</u>	12.54	5.24	5.29	0.2513
SPH-DiffDance (ours)	35.82	<u>10.60</u>	**9.36**	**6.90**	**0.2705**

Bold and underline indicate the best and second best results
$\downarrow$: lower is better, $\rightarrow$: closer to Ground Truth is better, $\uparrow$: higher is better

For motion diversity, the proposed method obtains better performance compared with most existing approaches. The DIV_k and DIV_g values reach 9.36 and 6.90, respectively, which are close to those of ground truth (8.98 and 7.42). Compared with Bailando, DiffDance, LODGE, and PAMD, our method shows clear advantages in diversity evaluation. This improvement can be attributed to the SPR module, which prevents the model from relying excessively on a limited number of motion patterns. As a result, the model can generate richer dance variations according to different music inputs. Meanwhile, the HEPR module further improves the coordination among local joints, allowing the generated motions to maintain structural consistency while providing more variation.

Regarding music-motion synchronization, our method achieves a BAS score of 0.2705, which is higher than all comparison methods and slightly exceeds PAMD (0.2700) and ground truth (0.2649). This result demonstrates that the generated motions can effectively respond to changes in musical rhythm. The improvement mainly comes from the combination of music-conditioned generation and semantic correspondence learning between music segments and motion patterns through prototype reconstruction. In addition, local pose refinement further strengthens the consistency between motion transitions and musical beats.

Overall, although the proposed method does not achieve the optimal result on every individual metric, it demonstrates strong competitiveness across several important indicators, including FID_g, DIV_k, DIV_g and BAS. These results verify that SPH-DiffDance can improve motion quality while maintaining diversity and music synchronization, making it suitable for long-sequence music-driven dance generation.

In addition to generation performance, we further evaluate the inference efficiency of SPH-DiffDance compared with LODGE under different sequence lengths. As shown in Table 2, our method contains 105.5M parameters, which is lower than LODGE (110.3M), indicating that the proposed SPR and HEPR modules achieve a lightweight model structure with limited parameter overhead. Since the model architecture and learnable parameters are independent of the generated sequence length, the number of parameters remains unchanged under different duration settings.

Table 2 Inference efficiency comparison between LODGE and SPH-DiffDance

Duration(s)	LODGE			SPH-DiffDance		
	Latency(ms)	Params(M)	FPS	Latency(ms)	Params(M)	FPS
5	1039.7	110.3	144.3	1139.3	105.5	131.7
10	1094.0	110.3	274.2	1095.0	105.5	274.0
15	1057.2	110.3	425.6	1120.2	105.5	401.7
20	1051.8	110.3	570.5	1139.4	105.5	526.6

Latency denotes the average inference time over 20 runs. Params represents the number of trainable parameters in millions (M). FPS denotes the generated motion frames per second

Both SPH-DiffDance and LODGE adopt parallel diffusion-based generation, where the whole motion sequence is processed simultaneously during each denoising step. Therefore, the inference latency remains relatively stable when increasing the sequence length. Although the number of generated frames increases with sequence duration, the parallel denoising mechanism and GPU acceleration reduce the impact on inference latency within the evaluated range. Meanwhile, the dynamic hypergraph construction in HEPR introduces additional computational overhead, resulting in a slightly higher latency than LODGE, but the overall cost remains acceptable (approximately 1.1s). Since longer sequences contain more generated frames under nearly unchanged inference time, the FPS of SPH-DiffDance increases from 131.7 to 526.6 when the sequence length increases from 5s to 20s, remaining comparable to LODGE (144.3 to 570.5) given the additional structural computation introduced by HEPR. These results demonstrate that SPH-DiffDance achieves a 4.3% reduction in model parameters over LODGE, with the marginal latency overhead attributable to dynamic graph construction in HEPR, confirming that the proposed structural modules maintain a lightweight and practical design for long-sequence dance generation.

4.4 Ablation studies

To further analyze the influence of the proposed modules on the music-driven dance generation task, ablation experiments were carried out on the AIST++ dataset. All experiments follow the same data division, training strategy, and evaluation protocol, with only the corresponding modules being modified.

The effectiveness of the SPR module and the HEPR module is evaluated by comparing different module combinations, with the results shown in Table 3. The baseline

model (vanilla diffusion without SPR and HEPR) achieves FID_k and FID_g values of 46.99 and 15.22, respectively. Meanwhile, its DIV_k, DIV_g and BAS values are 6.52, 5.69 and 0.2423. These results indicate that the basic diffusion framework can generate dance motions but still has limitations in motion quality, structural rationality, and music responsiveness. After introducing the SPR module, the model achieves significant improvements, where FID_k decreases from 46.99 to 36.06 and FID_g decreases from 15.22 to 10.75. Meanwhile, DIV_k, DIV_g and BAS increase to 8.67, 6.63 and 0.2585, respectively. This improvement demonstrates that by capturing global motion semantics through prototype-based retrieval and reconstruction, SPR provides more stable semantic guidance and thereby enhances the consistency of long-sequence dance generation. After adding HEPR, DIV_k increases to 9.05 and BAS improves to 0.2596. This improvement mainly benefits from the ability of hypergraph modeling to capture complex coordination relationships among multiple joints, indicating that hypergraph-based pose refinement enhances motion diversity and music responsiveness. When SPR and HEPR are integrated together, the model achieves the best performance. These results confirm the complementary roles of SPR and HEPR in improving both global choreographic coherence and local motion quality.

Table 3 Effect of different module combinations on dance generation performance

Ablation			Metrics				
Baseline	SPR	HEPR	$FID_k\downarrow$	$FID_g\downarrow$	$DIV_k\rightarrow$	$DIV_g\rightarrow$	$BAS\uparrow$
√			46.99	15.22	6.52	5.69	0.2423
√	√		36.06	10.75	8.67	6.63	0.2585
√		√	37.85	11.77	9.05	6.52	0.2596
√	√	√	**35.82**	**10.60**	**9.36**	**6.90**	**0.2705**

Bold indicates the best results

To analyze the influence of prototype loss weight λ_{proto} on generation performance, sensitivity experiments are conducted, and the results are shown in Table 4. For clarity, we report the three most representative metrics (FID_g, DIV_k, and BAS) in the sensitivity analysis. When $\lambda_{proto} = 0.05$, the model obtains FID_g, DIV_k and BAS values of 10.70, 8.83, and 0.2560, respectively. At this stage, the prototype constraint is relatively weak. Although the motion prototype information can provide certain semantic guidance, its ability to constrain the generation process is still insufficient, resulting in limited diversity and music-motion consistency. When λ_{proto} increases to 0.10, FID_g decreases to 10.60, indicating that stronger prototype constraints can improve motion quality. With $\lambda_{proto} = 0.2$, the model achieves the best overall performance. This indicates that an appropriate prototype constraint can achieve a better balance between motion stability and generation flexibility. When λ_{proto} is further increased to 0.30, the performance decreases slightly, possibly because overly strong semantic constraints limit the generation space of the diffusion model. Therefore, $\lambda_{proto} = 0.2$ is adopted as the prototype loss weight in the final model.

Table 4 Effect of semantic prototype reconstruction weight λ_{proto} on model performance

λ_{proto}	$FID_g\downarrow$	$DIV_k\rightarrow$	$BAS\uparrow$
0.05	10.70	8.83	0.2560
0.10	10.60	8.77	0.2543
0.20	**10.60**	**9.36**	**0.2705**
0.30	10.62	8.84	0.2569

Bold indicates the best results

The influence of the number of prototypes n is further analyzed, with the results presented in Table 5. When $n = 16$, the model obtains FID_g, DIV_k and BAS values of 10.61, 8.78, and 0.2543, respectively. This indicates that a small number of prototypes can provide basic semantic guidance but cannot fully represent the diverse dance styles in the AIST++ dataset. When n increases to 32, the performance remains comparable to $n = 16$ with only marginal variation. When $n = 64$, the model achieves the best overall performance across all metrics. These results suggest that 64 prototypes can better cover different dance categories and motion patterns, providing effective guidance for global generation. However, when n increases to 128, the performance decreases slightly. A possible explanation is that excessive prototypes lead to overly fine-grained semantic division, making some prototypes difficult to learn effectively. Therefore, 64 semantic prototypes are selected for the final model.

Table 5 Effect of different numbers of semantic prototypes on generation performance

n	$FID_g\downarrow$	$DIV_k\rightarrow$	$BAS\uparrow$
16	10.61	8.78	0.2543
32	10.64	8.68	0.2542
64	**10.60**	**9.36**	**0.2705**
128	10.66	8.70	0.2539

Bold indicates the best results

To further verify the effectiveness of hypergraph modeling and the diversity constrain in the HEPR module, different hypergraph configurations are compared, and the results are shown in Table 6. Compared with the baseline model that only uses the basic graph structure, introducing HyperGCN reduces FID_g from 10.75 to 10.67 and increases BAS from 0.2585 to 0.2703. This indicates that the hypergraph structure can improve human pose representation by modeling high-order relationships among multiple joints. After adding the diversity constraint, the complete model further improves across all metrics. These results suggest that the diversity constraint

helps prevent the model from relying excessively on similar local motions, enabling richer motion variations while maintaining structural rationality.

Table 6 Ablation study of the hypergraph-enhanced pose refinement module

Configuration	$FID_g\downarrow$	$DIV_k\rightarrow$	$BAS\uparrow$
Baseline (ProtoGCN only)	10.75	8.67	0.2585
+HyperGCN (w/o DivLoss)	10.67	8.68	0.2703
+HyperGCN + DivLoss (Ours)	**10.60**	**9.36**	**0.2705**

Bold indicates the best results

The influence of hypergraph structure parameters on generation performance is further investigated, with the results shown in Table 7. When the number of virtual key nodes is set to 4 and the neighbor connection number K is set to 9, the model achieves the best overall performance. When the number of virtual key nodes deviates from 4, performance degrades in both directions. A smaller number of virtual nodes limits the model's ability to represent high-order relationships among joints, while excessive nodes may introduce redundant relationships and increase optimization difficulty. Meanwhile, K determines the range of information propagation in the hypergraph. A smaller K makes it difficult to capture complete joint dependencies, whereas a larger K may introduce irrelevant connections. Therefore, appropriate hypergraph structure parameters can achieve a better balance between structural constraints and motion diversity.

Table 7 Performance comparison under different hypergraph structure configurations

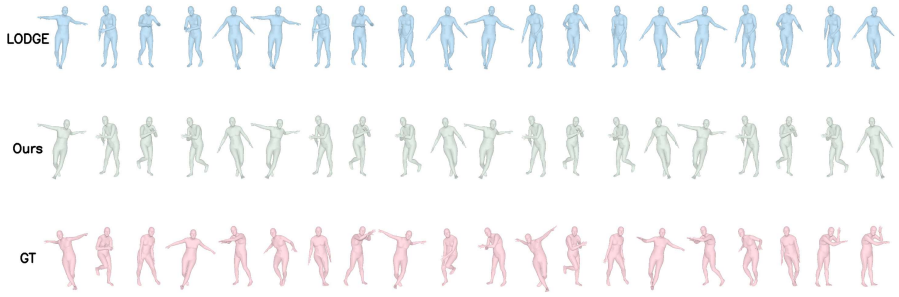
$VIRTUAL_NUM$	K	$FID_g\downarrow$	$DIV_k\rightarrow$	$BAS\uparrow$
2	9	10.70	9.13	0.2700
4	9	**10.60**	**9.36**	**0.2705**
8	9	10.93	8.92	0.2565
4	7	10.69	8.64	0.2560
4	11	11.28	8.69	0.2574

Bold indicates the best results

4.5 Visualization analysis

To visually demonstrate the generation quality, Fig. 4 presents a qualitative comparison among SPH-DiffDance, LODGE, and ground truth across three representative dance styles: Break, Hip-hop, and Waack. It can be observed that LODGE is able to generate generally continuous dance motions, but the movement range of some

sequences is limited and the posture variations are relatively simple, making it difficult to fully capture the characteristics of different dance styles. In contrast, the dance sequences generated by SPH-DiffDance are closer to ground truth in terms of overall motion trends, local pose variations and style characteristics. For the Break style, our method generates more obvious body weight shifts and lower-limb movements. In the Hip-hop style, the generated motions show more natural body swing and arm coordination. For the Waack style, the model maintains better rhythm variations and pose continuity. These qualitative results are consistent with the quantitative findings, further confirming that SPH-DiffDance generates more realistic and style-consistent dance sequences than LODGE.



(a) Break



(b) Hip-hop



(c) Waack

Fig. 4 Qualitative comparison of generated dance sequences across different dance styles. In each subfigure, the three rows correspond to LODGE (blue), SPH-DiffDance (green), and ground truth (pink), respectively.

5 Conclusion

In this paper, we present SPH-DiffDance, a coarse-to-fine diffusion framework for long-sequence music-driven 3D dance generation, incorporating the SPR and HEPR modules to address global semantic coherence and local structural refinement respectively. The SPR module learns representative dance patterns from motion latent representations and provides global semantic guidance for the generation process, which helps reduce semantic drift and preserve the continuity of dance themes in long sequences. Meanwhile, the HEPR module captures high-order relationships among human joints beyond conventional skeleton connections, enabling better coordination of local body structures and improving the spatial rationality of generated poses. By incorporating motion priors into the diffusion process, SPH-DiffDance further enhances the physical plausibility of generated movements. Experiments on the AIST++ dataset demonstrate that SPH-DiffDance can generate more natural and coherent dance sequences with improved motion quality, diversity, and music-motion synchronization compared with existing methods. Ablation studies further confirm the complementary contributions of SPR and HEPR across motion quality, diversity, and music synchronization metrics.

Although SPH-DiffDance improves the semantic consistency and structural quality of long-sequence dance generation, some limitations remain. The current framework mainly focuses on single-person dance generation, and the learned prototype library may not fully cover highly diverse dance styles. In future work, we will investigate more flexible semantic modeling strategies and extend the framework to multi-person dance generation and more diverse motion scenarios.

Acknowledgements. This work was supported by the Natural Science Foundation of Hebei Province of China (No. F2024203069), National Natural Science Foundation of China (No. 62001413).

Declarations

- Funding This study was funded by the Natural Science Foundation of Hebei Province of China (No. F2024203069), National Natural Science Foundation of China (No. 62001413).
- Conflict of interest The authors declare that they have no conflict of interest.
- Ethics approval Not applicable.
- Consent for publication Not applicable.
- Consent to participate Not applicable.
- Data availability This study uses the publicly available datasets AIST++ (can be access from https://google.github.io/aistplusplus_dataset). No other datasets were generated during the current study.
- Author contribution Yifan Wu, Xingyu Li, and Chuanxin Xu were responsible for design, data collection, writing, and overall investigation. Zhengping Hu was responsible for the supervision and editing of the manuscript.

References

1. Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., Wang, Y.: Human motion generation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(4), 2430–2449 (2024)

2. Sun, L., Li, H., Xu, F., Liu, Y.: Music-driven dance generation: A comprehensive review. Neurocomputing 682, 133466 (2026)

3. 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 (2020)

4. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: AI Choreographer: Music conditioned 3D dance generation with AIST++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13401–13412 (2021)

5. Siyao, L., Yu, W., Gu, T., Lin, C. Wang, Q., Qian, C., Loy, C.C., Liu,Z.: Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11050–11059 (2022)

6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. 33, 6840–6851 (2020)

7. Tseng, J., Castellon, R., Liu, K.: EDGE: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 448–458 (2023)

8. Qi, Q., Zhuo, L., Zhang, A., Liao, Y., Fang, F., Liu, S., Yan, S.: DiffDance: Cascaded human motion diffusion model for dance generation. In: Proceedings of the

31st ACM International Conference on Multimedia, pp. 1374–1382 (2023)

9. Li, R., Zhang, Y.X., 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 (CVPR), pp. 1524–1534 (2024)

10. Li, R., Zhang, H., Zhang, Y., Zhang, Y., Guo, J., Zhang, Y., Li, X., Liu, Y.: Lodge++: High-quality and long dance generation with vivid choreography patterns. arXiv preprint arXiv:2410.20389 (2024)

11. Wang H, Zhu Y, Lai Q, Zhang, Y., Xie, G.S., Geng, X. PAMD: Plausibility-Aware Motion Diffusion Model for Long Dance Generation. arXiv preprint arXiv:2505.20056 (2025)

12. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)

13. Li, J., Bai, Z., Kong, D., Chen, D., Li, Q., Yin, B.: 3D human pose estimation based on conditional dual-branch diffusion. Multimedia Syst. 31(1), 5 (2025)

14. Zou, Y., Yang, J., Ceylan, D., Zhang, J., Perazzi, F., Huang, J.B.: Reducing footskate in human motion reconstruction with ground contact constraints. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 459–468 (2020)

15. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(1), pp. 3558–3565 (2019)

16. Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., Li, H.: Learning to generate diverse dance motions with Transformer. arXiv preprint arXiv:2008.08171 (2020)

17. Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., Xia, S. Music2dance: Dancenet for music-driven dance generation. ACM Transactions on Multimedia Computing, Communications, and Applications 18(2), 1–21 (2022)

18. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: MoMask: Generative masked modeling of 3D human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1900–1910 (2024)

19. Liu, X., Dong, X., Qian, S., Kanojia, D., Wang, W., Feng, Z.: GCDance: Genre-controlled music-driven 3D full body dance generation. IEEE Transactions on Multimedia (2026)

20. Zhang, J., Kang, Z., Liu, L., Chang, J., Tian, Q., Gao, F., Wang, Y.: OpenDance: Multimodal controllable 3D dance generation with large-scale internet data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 28860–28870 (2026)

21. Li, X., Li, R., Fang, S., Xie, S., Guo, X., Zhou, J., Peng, J., Wang, Z.: Music-aligned holistic 3D dance generation via hierarchical motion modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14420–14430 (2025)

22. Ghosh, A., Zhou, B., Dabral, R., Wang, J., Golyanik, V., Theobalt, C., Slusallek, P., Guo, C.: DuetGen: Music driven two-person dance generation via hierarchical masked modeling. In: Proceedings of the ACM SIGGRAPH Conference Papers, pp. 1–11 (2025)

23. Su, M.C., Huang, W.L., Pan, T.Y.: DanceLDM: Latent-based diffusion model for dance generation and editing conditioned on music and text prompt. Multimedia Systems 32(4), 274 (2026)

24. Guo, P., Hao, T., Fu, W., Gao, Y., Li, Y.: Controllable 3D dance generation using diffusion-based Transformer U-Net. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39(3), pp. 3284–3292 (2025)

25. Yang, K., Tang, X., Peng, Z., Zhang, X., Wang, P., He, J., Liu, H.: FlowerDance: MeanFlow for efficient and refined 3D dance generation. arXiv preprint arXiv:2511.21029 (2025)

26. Pahde, F., Puscas, M., Klein, T., Nabi, M.: Multimodal prototypical networks for few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2644–2653 (2021)

27. Chen, S., Hou, W., Hong, Z., Ding, X., Song, Y., You, X., Liu, T., Zhang, K.: Evolving semantic prototype improves generative zero-shot learning. In: International Conference on Machine Learning, pp. 4611–4622. PMLR (2023)

28. Liu, D., Xia, Q., Meng, F., Ye, M., Zhang, J.: Multi-level semantic-assisted prototype learning for few-shot action recognition. Neurocomputing 636, 130022 (2025)

29. Han, C., Lu, Y., Sun, G., Liang, G.C., Cao, Z., Wang, Q., Guan, Q., Dianat, S.A., Rao, R.M., Geng, T., Tao, Z., Liu, D.: Prototypical Transformer as unified motion learners. arXiv preprint arXiv:2406.01559 (2024)

30. Liu, H., Liu, Y., Ren, M.,Wang, H., Wang, Y., Sun, Z.: Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 29248–29257 (2025)

31. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision, pp. 390–408. Springer (2024)

32. Cong, P., Wang, Z., Ma, Y., Yue, X.: SemGeoMo: Dynamic contextual human motion generation with semantic and geometric guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17561–17570 (2025)

33. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: PromptHMR: Promptable Human Mesh Recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1148–1159 (2025)

34. Pavlakos, G., Choutas, V., Ghorbani, N., et al.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10975–10985 (2019)

35. Lu, J., Lin, J., Dou, H., Zeng, A., Deng, Y., Liu, X., Cai, Z., Yang, L., Zhang, Y., Wang, H., Liu, Z.: DPoser-X: Diffusion model as robust 3D whole-body human pose prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9988–9997 (2025)

36. Li, B., Cao, Z., Li, H., Guo, Y. CIGPose: Causal Intervention Graph Neural Network for Whole-Body Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.23484–23494 (2026)

37. Zhou, Y., Xu, T., Wu, C., Wu, X., Kittler, J.: Adaptive hyper-graph convolution network for skeleton-based human action recognition with virtual connections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12648–12658 (2025)

38. Liu, S., Lv, P., Zhang, Y., Fu, J., Cheng, J., Li, W., Zhou, B., Xu, M.: Semi-dynamic hypergraph neural network for 3D pose estimation. In: International Joint Conference on Artificial Intelligence, pp. 782–788 (2020)

39. Li, J., Wang, J., Wu, L., Wang, X., Luo, X., Xu, Y.: AMHGCN: Adaptive multi-level hypergraph convolution network for human motion prediction. Neural Networks 172, 106153 (2024)

40. Cui, Q., Ding, Z., Chen, F.: Hybrid directed hypergraph learning and forecasting of skeleton-based human poses. Cyborg and Bionic Systems 5, 0093 (2024)

41. Han, B., Huang, Y., Gao, P.: HyperDiff: Hypergraph guided diffusion model for 3D human pose estimation. In: Proceedings of the IEEE International Workshop

on Multimedia Signal Processing (MMSP), pp. 72–77 (2025)