

中图法分类号: 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-15

论文引用格式: Huang Xiaoying, Zhang Sanyi, Wang Xirui, Guo Xiaoxuan, Zhang Qin, Ye Long. Trajectory- and style-controllable dance camera movement synthesis[J/OL]. Journal of Image and Graphics, XXXX: 1-15. DOI: 10.11834/jig.260441. (黄霄莹, 张三义, 王玺瑞, 郭晓璇, 张勤, 叶龙. 3D 数字人舞蹈运镜的轨迹与风格可控合成[J/OL]. 中国图象图形学报, XXXX: 1-15. DOI: 10.11834/jig.260441.) [DOI: 10.11834/jig.260441]

3D 数字人舞蹈运镜的轨迹与风格可控合成

黄霄莹¹, 张三义², 王玺瑞³, 郭晓璇¹, 张勤⁵, 叶龙^{2,4}

1. 中国传媒大学信息与通信工程学院, 北京 100024; 2. 中国传媒大学数据科学与智能媒体学院, 北京 100024; 3. 中国传媒大学海南国际学院, 海南 572400; 4. 媒体融合与传播国家重点实验室(中国传媒大学), 北京 100024; 5. 媒介音视频教育部重点实验室(中国传媒大学), 北京, 100024

摘要: 目的 现有音乐与舞蹈驱动自动运镜合成方法主要关注生成与节奏和动作匹配的相机轨迹, 缺乏对运镜风格与空间轨迹的显式控制, 难以满足专业创作中的精细化编辑需求。方法 针对该问题, 本文提出多风格舞蹈运镜可控合成框架(controllable dance camera movement synthesis framework, CtrlDanceCam), 将风格标签和关键帧相机参数作为用户控制条件。框架包含关键帧预测模块(keyframe prediction module, KPM)和可控轨迹合成模块(controllable trajectory synthesis module, CTSM): KPM 基于音乐、舞蹈与风格条件预测符合目标风格的关键帧分布; CTSM 以关键帧为锚点, 通过风格条件化的可学习插值函数生成完整相机运动轨迹, 并支持用户指定关键帧参数进行轨迹控制。进一步设计风格感知对抗训练策略(style-aware adversarial training, SAAT), 增强不同运镜风格的区分能力与生成一致性。结果 在公开舞蹈-运镜配对数据集上的实验表明, CtrlDanceCam 的基于运动学特征的弗雷歇起始距离(Fréchet inception distance, FID_k)为 3.022, 平均欧氏距离(average Euclidean distance, Dist_k)为 3.101, 在提升生成质量的同时保持了良好的轨迹多样性。消融实验和可视化结果进一步验证了各模块及双重控制机制的有效性。结论 CtrlDanceCam 实现了舞蹈运镜的风格与轨迹双重可控合成, 在保证生成质量与轨迹多样性的同时增强了用户编辑能力, 可为虚拟制作和数字表演中的自动运镜提供技术支撑。代码开源地址为: <https://www.scidb.cn/s/vUr2Av>。

关键词: 舞蹈相机轨迹合成; 可控合成; 风格感知; 对抗训练; 关键帧预测; 数字人动画; 跨模态生成

Trajectory- and style-controllable dance camera movement synthesis

Huang Xiaoying¹, Zhang Sanyi², Wang Xirui³, Guo Xiaoxuan¹, Zhang Qin⁵, Ye Long^{2,4}

1. School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; 2. School of Data Science and Media Intelligence, Communication University of China, Beijing 100024, China; 3. Hainan International College of Communication University of China, Hainan 572400, China; 4. State Key Laboratory of Media Convergence and Communication, Communication

收稿日期: 2026-07-20; 修回日期: 2026-09-18

* 通信作者: 叶龙, 通信作者, 男, 教授, 主要研究方向为多媒体视音频处理。E-mail: yelong@cuc.edu.cn; 张三义, 男, 副教授, 主要研究方向为图像分割, 虚拟数字人。E-mail: zhangsanyi@cuc.edu.cn; 王玺瑞, 男, 学生, 主要研究方向为虚拟数字人生成。E-mail: 202329013018n@mails.cuc.edu.cn; 郭晓璇, 女, 学生, 主要研究方向为音频鉴伪, 数字人生成。E-mail: xiaoxuanguo@mails.cuc.edu.cn; 张勤, 男, 教授, 主要研究方向为多媒体视音频处理。E-mail: zhangqin@cuc.edu.cn

基金项目: 国家自然科学基金(62571500, 62271455); 北京市自然科学基金(L252143); 虚拟现实技术与系统全国重点实验室(北京航空航天大学)开放课题基金(VRLAB2025C03); 中国传媒大学中央高校基本科研业务费专项资金资助(CUC26TD09); 中传云计算平台。

Supported by: National Natural Science Foundation of China (62571500, 62271455); Natural Science Foundation of Beijing, China (L252143); Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2025C03); Fundamental Research Funds for the Central Universities (CUC26TD09); Public Computing Cloud, CUC.

University of China, Beijing; 5. Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, China

Abstract: Objective In digital human dance performance, camera movement quality directly impacts the artistic expression of choreography and the audience's viewing experience. Camera trajectory design serves multiple functions: guiding viewer attention toward key performers, conveying emotional atmosphere through framing and motion, and enhancing the visual impact of dance movements through coordinated camera dynamics. However, manually designing camera movements for a given dance sequence is labor-intensive and demands both cinematic expertise and deep understanding of choreographic structure. Consequently, automatic dance camera movement synthesis has emerged as an important research direction in cross-modal content generation. Existing music- and dance-driven camera synthesis methods, including DanceCamera3D, DanceCamAnimator, and TemMEGA, focus on generating camera trajectories aligned with basic rhythm and motion patterns of the input performance. While these methods have advanced automation, they share a critical limitation: they do not support explicit user control over the generated results. In professional filmmaking workflows, practitioners require fine-grained editorial control along two independent dimensions. The first is style controllability—the capacity to specify the overall artistic style of camera work, such as K-pop, Choreography, or Jazz. The second is trajectory controllability—the ability to input camera parameters at specific keyframes, including position, orientation, and field of view, to guide the generated trajectory toward a desired creative intent. No existing dance camera synthesis method has explicitly modeled either form of user control, leaving a significant gap between research prototypes and practical production tools. This work addresses the problem of controllable multi-style dance camera movement synthesis, targeting dual-dimensional user control over both the artistic style and the spatial trajectory of synthesized camera motion. **Method** We propose CtrlDanceCam, a framework that, for the first time in dance camera synthesis, enables users to interactively specify style labels and keyframe constraints as explicit conditioning signals. The framework is built upon a controllable camera synthesis network backbone, ControlCamNet, comprising two tightly integrated modules: a keyframe prediction module (KPM) and a controllable trajectory synthesis module (CTSM). Given an input music sequence and a dance motion sequence with a user-specified style label, the KPM predicts a binary keyframe sequence identifying shot boundaries. KPM employs a Transformer encoder augmented with a learnable style encoder that maps discrete style labels into continuous embeddings. These style-conditioned embeddings, fused with music and dance features, enable prediction of keyframe distributions characteristic of the target style—for instance, sparse keyframes for jazz-style long takes versus dense keyframes for locking-style rapid cuts. The module is trained with a weighted binary cross-entropy loss to address class imbalance. The CTSM synthesizes the complete camera motion sequence using predicted or user-specified keyframes as anchor points. For each shot defined by consecutive keyframes, CTSM predicts camera parameters at boundary frames and learns a style-conditioned, monotonically increasing interpolation function $\rho(t; s)$ governing the intra-shot motion profile. Non-keyframe camera parameters are obtained through $\rho(t; s)$ -weighted interpolation between boundary predictions. This design ensures smooth intra-shot transitions while allowing style-dependent intra-shot motion profiles. When users provide hard constraints by specifying camera parameters at particular keyframe indices, CTSM substitutes these values for model predictions, enabling precise trajectory-level editing. To enhance style-conditioning responsiveness and synthesis quality, we design a style-aware adversarial training (SAAT) strategy with three complementary loss components: a controllable style contrastive loss that compels the generator to produce outputs judged as realistic under the correct style label while penalized under incorrect labels, a style classification loss with label smoothing for style-consistent generation, and a base Wasserstein adversarial loss with gradient penalty. These losses synergistically prevent inter-style confusion, ensure intra-style consistency, and promote overall realism. To support style-controllable training, we augment the dance-camera-music (DCM) dataset with fine-grained style annotations. Through multiple rounds of discussion with professional animators, and referencing the FineDance style taxonomy for semantic alignment with upstream music-to-dance generation, each sample is annotated with one of eight styles: K-pop, Choreography, Traditional Chinese Dance, Hip-hop, Popping, Locking, Urban, and Jazz. Statistical analysis across these styles reveals quantifiable differences in keyframe density (ranging from 2.640% for Jazz to 31.484% for Locking), shot length, field of view (FOV) variation, and camera velocity, confirming

the validity of the labeling scheme and motivating the need for style-adaptive keyframe prediction. **Result** Comprehensive experiments on the DCM dataset—the only publicly available dance-camera paired benchmark—demonstrate the effectiveness of CtrlDanceCam. In terms of generation quality, the full model achieves Fréchet inception distances of 3.022 (FID_k) and 0.251 (FID_s). The average Euclidean distances are 3.101 ($Dist_k$) and 1.355 ($Dist_s$). For dancer tracking fidelity, the dancer missing rate (DMR) is 0.0019 and the limbs capture difference (LCD) is 0.140. These results outperform DanceCamera3D, DanceCamAnimator, and TemMEGA on five of the six metrics, while remaining competitive on $Dist_k$. Ablation studies confirm the contribution of each architectural component. Removing SAAT degrades FID_k to 3.512 and FID_s to 0.345. Removing the controllable style contrastive loss raises FID_s to 0.317, confirming its contribution to style discrimination. Replacing the KPM-predicted keyframes with uniform sampling drives FID_k to 3.742 and FID_s to 0.366, while stripping style conditioning from KPM disproportionately harms FID_s to 0.331, confirming that style-conditioned keyframe prediction forms the upstream basis of the framework's style controllability. Substituting the learnable interpolation function $\rho(t; s)$ with fixed linear interpolation causes FID_s to rise sharply from 0.251 to 0.346—a substantially larger degradation than the corresponding FID_k increase from 3.022 to 3.351—revealing that $\rho(t; s)$ contributes an independent, intra-shot layer of style expression beyond KPM's inter-shot style differentiation. Keyframe density sensitivity analysis further shows that the KPM adaptive strategy achieves the optimal balance across all metrics, outperforming every fixed-density configuration from 3% to 30%. Qualitative visualizations confirm that distinct style labels yield clearly differentiated camera trajectories, and that sparse user-specified keyframe constraints effectively adjust the generated camera motion toward the desired result, jointly validating the dual-dimensional controllable synthesis. **Conclusion** CtrlDanceCam introduces a novel paradigm for dance camera synthesis by enabling simultaneous trajectory-level and style-level user control. Its modular architecture decomposes controllable synthesis into interpretable sub-problems of style-conditioned keyframe prediction and style-conditioned intra-shot interpolation. The design facilitates future extensions to richer control modalities including path constraints, velocity constraints, and fine-grained shot composition parameters. The style labeling scheme established for DCM provides a foundation for future style-aware camera synthesis research. CtrlDanceCam demonstrates the feasibility of integrating interactive user control into cross-modal generative frameworks, offering a practical technical pathway for deploying automatic camera synthesis in professional creative workflows such as virtual production, digital performance, and immersive media creation. Code is available at ScienceDB: <https://www.scidb.cn/s/vUr2Av>

Key words: dance camera movement synthesis; controllable synthesis; style-aware; adversarial training; keyframe prediction; digital human animation; cross-modal generation

论文引用格式: Huang Xiaoying, Zhang Sanyi, Wang Xirui, Guo Xiaoxuan, Zhang Qin, Ye Long. Trajectory- and style-controllable dance camera movement synthesis [J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260441. (黄霄莹, 张三义, 王玺瑞, 郭晓璇, 张勤, 叶龙. 3D 数字人舞蹈运镜的轨迹与风格可控合成[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260441.)

0 引言

随着数字媒体与虚拟现实技术的快速发展,多模态数字人建模与驱动技术受到广泛关注(高玄等, 2024),数字人舞蹈表演已成为计算机图形学与视觉领域的重要研究方向(Tseng等, 2023)。在舞蹈表演

中,运镜设计对于引导观众注意力、烘托情感氛围和增强动作表现力具有重要作用(Xie等, 2023)。然而,为一段舞蹈精心设计运镜,是一项耗时且高度依赖专业技能的任务。因此,全自动的舞蹈运镜合成技术受到广泛关注,其目标是根据输入的音乐和舞蹈动作自动生成符合艺术规律的相机运动轨迹。

相机运动轨迹的自动化规划与控制是计算机图形学中的经典问题。早期研究主要采用基于规则、约束或参考轨迹的方法(Christie等, 2008; Bares等, 2000),通过预定义的电影语法或用户指定的帧约束生成相机路径。这类方法在游戏过场动画(Rao等, 2023)、电影制作(Rucks和Katzakis, 2021)及无人机航拍(Gschwindt等, 2019)等场景中得到了应用。在无人机摄影领域, Huang等(2019)根据艺术原理实现了相机运动的自动化规划。近年来,基于深度学

习的相机控制方法取得了明显进展:Jiang等(2024)提出了电影摄影相机扩散模型,将扩散生成范式引入相机轨迹规划;Pueyo等(2024)提出了全自主无人机摄影系统 CineMPC,实现了包含变焦、对焦、位姿与场景构图的完整自主控制。在国内,吴晓雨等(2023)针对云演艺场景设计了一套基于机位计算的智能虚拟拍摄系统。然而,上述方法主要面向通用拍摄场景,未针对舞蹈表演中音乐节奏响应、舞蹈动作跟随等特定需求进行专门设计。

针对舞蹈表演场景,研究者开始探索音乐与舞蹈联合驱动的运镜合成。Wang等(2024a)首次构建了三维舞蹈-运镜-音乐(dance-camera-music, DCM)配对数据集,并提出一种基于Transformer扩散模型的运镜生成方法 DanceCamera3D,验证了从音乐和舞蹈联合条件生成三维运镜的可行性。随后Wang等(2024b)提出 DanceCamAnimator (DCA),其融入动画师先验知识,模仿专业动画制作者的分层级创作流程,采用“关键帧预测—镜头合成—参数细化”的三阶段范式,实现了对可变长度序列的精细生成。Xu等(2025)提出了 TemMEGA,一种基于时序掩码生成模型的运镜合成方法,支持在线与离线合成。在相邻领域,虚拟摄影与电影摄影的自动化也取得了进展:Jiang等(2020)提出了基于示例驱动的虚拟电影摄影方法,从已有电影片段中学习相机行为;Evin等(2022)设计了 Cine-AI 系统,通过模仿人类导演风格生成游戏过场动画的相机镜头序列。这些工作在提升运镜生成自动化程度方面取得了进步。

然而,上述舞蹈运镜合成方法均关注于从音乐

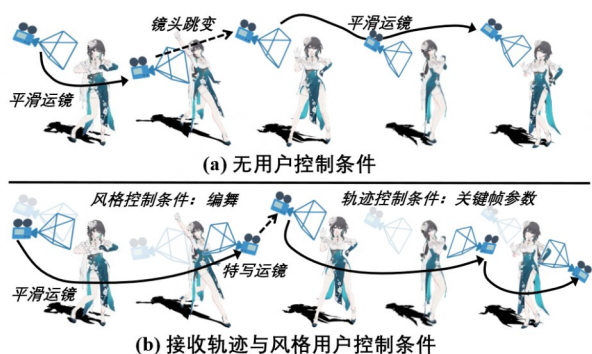


图1 与没有可控合成的方法对比

Fig. 1 Comparison with methods lacking controllable synthesis. (a) Existing methods without user control; (b) proposed method with user control conditions

与舞蹈内容中自动生成符合基本节奏与动作的运镜轨迹,缺乏对用户控制需求的支持。值得注意的是,DCA虽然采用了关键帧预测与插值合成的高层流程,但其关键帧预测主要隐式学习动画师先验知识,并未接收用户显式指定的控制条件;此外,上述方法均未对运镜的艺术风格进行显式建模。在专业影视与内容创作实践中,用户往往需要对生成结果进行精细化编辑:一是风格的可控性,即允许指定目标运镜风格(如韩流、编舞、爵士等),以满足不同艺术表达需求;二是轨迹的可控性,即允许输入特定关键帧的相机参数(如位置、朝向、视场角),以引导生成符合创作意图的完整运镜轨迹。现有方法尚未对这两类控制需求进行显式建模。

在多模态三维数字人运动生成领域,文本、音频和动作等条件已被广泛用于人体运动生成与编辑(赵宝全等,2024)。其中,在音乐驱动的舞蹈生成领域,风格的显式建模已得到广泛探索,为运镜合成中的风格可控性研究提供了重要参照。舞蹈生成方法已从基于跨模态Transformer的连续序列建模(Li等,2021)、基于离散动作表示的自回归生成(Li等,2022),逐步发展到扩散模型(Tseng等,2023; Alexanderson等,2023)。在风格化生成方面, Kim等(2022)开创了基于生成对抗网络的多风格舞蹈生成范式; LODGE(Li等,2024)通过从粗到细的两阶段扩散框架结合对抗训练,实现了风格多样的舞蹈生成,其扩展版本 LODGE++(Li等,2025)进一步提升了长序列舞蹈的编排丰富度; EDMG(Zhang等,2025)从舞蹈流派的基本动作出发探索了高效长序列舞蹈生成。李晨光等(2025)提出了体型感知的音乐驱动舞蹈动作风格化生成方法;石敏等(2025)通过舞蹈动作基元符号和音乐语义驱动舞蹈动画生成; FODGE(Huang等,2026)通过全身体优化实现了高保真舞蹈生成。上述研究表明,将风格明确建模为影响生成结果的关键因素已成为舞蹈生成领域的基本共识。然而,这一共识尚未在舞蹈运镜合成任务中得到体现,风格感知的运镜合成仍是一个亟待解决的问题。

在可控生成与风格条件建模的方法论层面,条件生成对抗网络(Mirza和Osindero,2014)和条件扩散模型(Dhariwal和Nichol,2021)是两种主流范式,通过条件注入实现细粒度的生成控制。在运动风格化方面, LILAC(Ren等,2025)提出了基于流式变分

自编码器-扩散模型的长序列任意运动风格化方法,展示了风格条件在运动生成中的灵活应用。然而,舞蹈运镜的可控合成面临独特挑战:需要同时处理连续条件(轨迹参数)和离散条件(风格标签)的联合建模,且需保证相机运动的物理平滑性。

针对上述问题,本文提出了一种多风格舞蹈运镜可控合成框架(**controllable dance camera movement synthesis framework, CtrlDanceCam**)。该框架将关键帧参数和风格标签作为显式用户控制条件引入运镜合成过程。框架基于可控运镜合成骨干网络(**controllable camera synthesis network, ControlCamNet**),包含关键帧预测模块(**keyframe prediction module, KPM**)和可控轨迹合成模块(**controllable trajectory synthesis module, CTSM**)。KPM 基于 Transformer 编码器将风格标签编码为条件嵌入,预测符合目标风格特征的关键帧分布;CTSM 以关键帧为锚点,通过风格条件化的可学习插值函数合成完整的相机运动序列,同时支持用户输入关键帧相机参数作为用户控制条件。此外,本文设计了风格感知对抗训练策略(**style-aware adversarial training, SAAT**),包含可控风格对比损失、风格分类损失和基础对抗损失,迫使生成器在正确风格标签下产生更强的真实性信号,有效防止多风格间的混淆,同时提升生成质量与风格一致性。DCA(Wang 等, 2024b)虽然同样采用了关键帧预测与插值合成的高层流程,但与本文 CtrlDanceCam 框架在三个层面存在本质区别。在用户控制条件来源上,DCA 隐式学习动画师先验知识以实现自动化生成,而 CtrlDanceCam 显式接收用户控制条件(关键帧参数、风格标签),将生成范式由"自动生成"转向"交互可控"。在风格建模维度上,DCA 未对运镜的艺术风格进行显式建模,CtrlDanceCam 通过八类运镜风格标注和 SAAT 训练策略,首次在运镜合成中引入了风格感知机制。在插值策略上,DCA 采用固定的插值方式,CtrlDanceCam 设计了依赖于风格标签的可学习插值函数,使得不同风格呈现差异化的镜头内运动剖面。

本文的主要贡献包括以下三个方面:1)首次在舞蹈运镜合成任务中提出轨迹与风格可控合成机制,将用户控制条件融入生成流程;2)设计了含 KPM 和 CTSM 的 ControlCamNet 骨干网络和 SAAT 风格感知对抗训练策略,分别实现结构化的条件响应和多风格的精细化区分;3)对 DCM 数据集进行了八

类运镜风格的细粒度标注,为该领域的风格化研究提供了数据基础。

1 方法

1.1 框架概述

本文提出一种多风格舞蹈运镜可控合成 CtrlDanceCam 框架,如图 2 所示,实现音乐与舞蹈动作驱动的舞蹈运镜的可控合成。该框架允许通过指定风格标签或提供关键帧运镜参数,灵活控制合成运镜的艺术风格与空间轨迹。给定目标风格标签 s 、音乐序列 $m = \{m_1, \dots, m_T\}$ 及舞蹈动作序列 $d = \{d_1, \dots, d_T\}$,框架的目标是生成完整的相机运动序列 $c = \text{CtrlDanceCam}(m, d, s) = \{c_1, \dots, c_T\}$ 。

1.2 数据集标签扩展

为实现风格可控的舞蹈运镜合成,需要具备细粒度风格标注的训练数据。然而,现有的舞蹈-相机运动数据集,如 DCM,仅提供对齐的“音乐-舞蹈-相机”三元组,缺乏对相机运动艺术风格的描述。为此,本文对 DCM 数据集进行了风格标签扩展。

在扩展过程中,团队与动画及舞蹈领域的专业人士进行了多轮讨论,确保了标签准确性。同时,为使风格语义与上游的音乐-舞蹈生成任务对齐,本文参考了主流音乐到舞蹈生成数据集 FineDance(Li 等, 2023)所采用的风格分类体系。最终,本文为数据集中的每条样本标注了八种主要的舞蹈运镜风格,包括“韩流”、“编舞”、“中国传统舞”、“嘻哈”、“机械舞”、“锁舞”、“都市”和“爵士”。扩展后数据集中各风格的数据分布如图 3 所示。

为量化不同运镜风格的差异,对八种风格对应的真实数据(**ground truth, GT**)进行了分析。表 1 展示了八类风格在关键帧密度、视场角(**field of view, FOV**)变化幅度和运动速度等维度的对比。

由表 1 可知,八类运镜风格在关键帧密度、镜头长度、FOV 变化和相机速度等维度上呈现可量化的差异。爵士风格具有最低的关键帧密度(2.640%)和最长的平均镜头长度(38.023 帧),运镜更偏向长镜头和低频切换;锁舞风格则具有最高的关键帧密度(31.484%)和最短的平均镜头长度(3.158 帧),体现出频繁切换和强节奏响应的特点。韩流与编舞风格的关键帧密度较高(约 8%),FOV 变化也较明显,

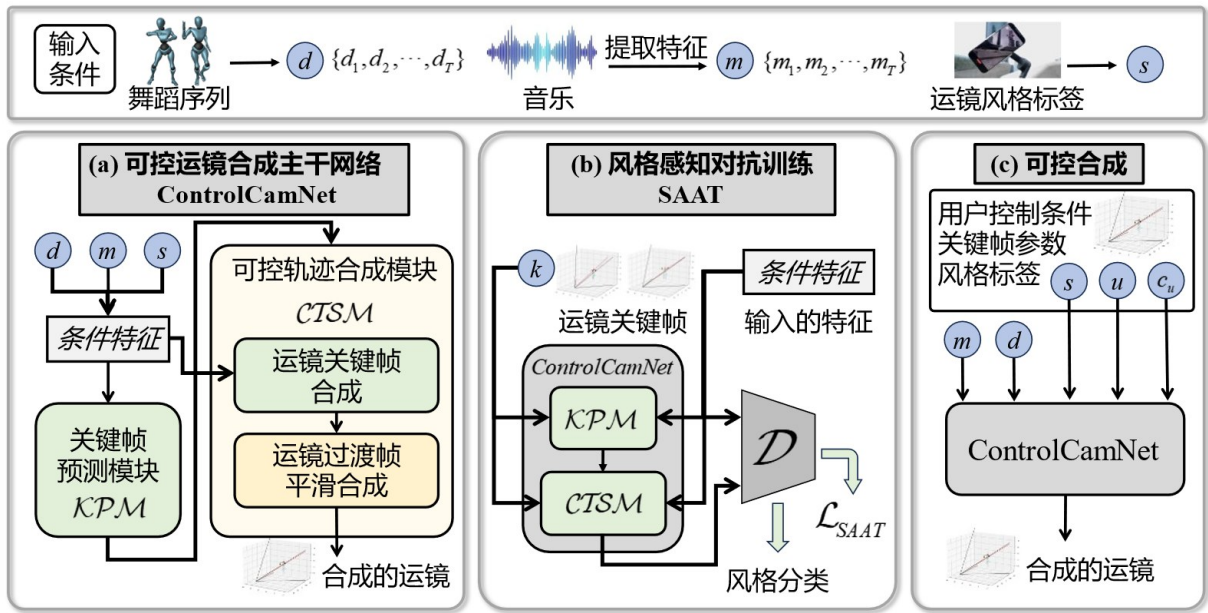


图2 CtrlDanceCam 框架图,包含3个主要部分:(a)可控运镜合成主干网络;(b)风格感知对抗训练;(c)可控合成
Fig. 2 Framework of CtrlDanceCam. It includes (a) ControlCamNet; (b) style-aware adversarial training; (c) controllable synthesis

表1 八类运镜风格的特征统计对比

Table 1 Statistical comparison of characteristics of eight types of camera movement styles

指标	韩流	编舞	传统中国舞	嘻哈	机械舞	锁舞	都市	爵士
关键帧密度	8.085	7.728	4.978	5.186	12.124	31.484	3.389	2.640
平均镜头长度	12.303	12.885	19.815	18.957	8.229	3.158	29.657	38.023
FOV 均值	29.908	27.153	28.718	33.935	23.714	28.162	30.000	17.117
FOV 标准差	7.359	8.514	10.822	12.104	9.001	4.996	0.000	3.525
平均相机速度	1.115	1.094	1.182	1.405	1.619	1.436	0.632	0.442

注:关键帧密度单位为%,平均镜头长度单位为帧,FOV 均值和标准差单位为°,平均相机速度为归一化值。

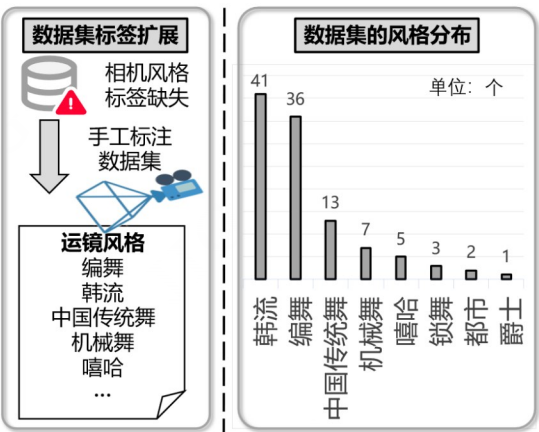


图3 数据集风格标签手工扩展与风格分布图
Fig. 3 Manual style-label annotation and style distribution of the dataset

符合其活跃的镜头调度模式。嘻哈和机械舞具有较高的平均相机速度(1.405和1.619),反映出更强的动态运动特征。都市风格的FOV标准差为0.000,其视场角基本固定,更偏向稳定松弛的运镜模式。传统中国舞的FOV标准差最高(10.822°),反映出其运镜中存在较大幅度的视场角变化。这些统计差异表明,不同风格标签对应着可量化的差异化运镜模式,固定关键帧密度难以适配所有风格,进一步支撑了风格条件化KPM和自适应关键帧预测的必要性。

1.3 可控合成模型设计

为实现轨迹与风格的双重可控合成,本文构建了可控运镜合成主干网络(controllable camera synthesis network, ControlCamNet),其核心由两个可灵活响应用户控制条件的子模块构成:关键帧预测模块(keyframe prediction module, KPM)和可控轨迹合

成模块 (controllable trajectory synthesis module, CTSM)。整体网络旨在接收并融合用户的轨迹控制条件(关键帧参数)与风格控制条件。

CtrlDanceCam 的高层流程(关键帧预测→镜头合成)与 DCA 共享了直观的创作管线,但在三个关键层面进行了本质扩展:(a) 用户控制条件性质:DCA 隐式学习动画师先验, CtrlDanceCam 显式响应用户控制条件,用户控制条件的来源和性质本质不同;(b) 风格建模:CtrlDanceCam 首次引入八类运镜风格标签和 SAAT 训练策略,实现了 DCA 不具备的风格感知能力;(c) 插值策略:CtrlDanceCam 的 CTSM 模块设计了依赖于风格标签的可学习插值函数 $\rho(t; s)$,使得不同风格呈现差异化的运动剖面,而 DCA 采用固定的插值策略。

关键帧预测模块(KPM)。KPM 负责预测时序关

键帧位置序列 $k = \{k_1, \dots, k_T\}$, 其中 $k_t \in \{0, 1\}$ 表示第 t 帧是否为关键帧。该模块将用户指定的风格标签作为条件,使关键帧分布与目标风格相符,同时其输出为用户提供了轨迹控制的基础接口。

本文将关键帧预测建模为条件概率预测问题。首先,通过可学习的风格编码器将离散风格标签 s 映射为风格嵌入向量 e_s 。对于当前帧 t ,提取一个长度为 $L = h + w$ 的滑动窗口,包含 h 帧历史与 w 帧未来,并提取窗口内的音乐特征 $m_{t-h:t+w-1}$ 、舞蹈动作 $d_{t-h:t+w-1}$ 及对应的关键帧标签子序列 $k_{t-h:t+w-1}$ 。风格嵌入 e_s 被广播至序列长度 L ,记为 $e_s \oplus I_L$ 。模型将拼接后的特征 $[m_{t-h:t+w-1}, d_{t-h:t+w-1}, e_s \oplus I_L]$ 输入 Transformer 编码器,最终通过一个分类头预测未来 w 帧中每一帧为关键帧的概率:

$$\hat{k}_{t+i} = P(k_{t+i} = 1 | m_{t-h:t+w-1}, d_{t-h:t+w-1}, e_s \oplus I_L), (1)$$

KPM 采用加权二元交叉熵损失进行训练,以处理关键帧和非关键帧之间的类别不平衡问题:

$$\mathcal{L}_{kf} = -\frac{1}{w} \sum_{i=0}^{w-1} [\lambda \cdot k_{t+i} \log(\hat{k}_{t+i}) + (1 - k_{t+i}) \log(1 - \hat{k}_{t+i})], (2)$$

式中 k_{t+i} 为真实值标签, $\lambda > 1$ 是从数据集中分配给关键帧类别的权重。通过阈值处理即可得到二值关键帧序列。风格嵌入的引入使得模型能够感知并学习不同艺术风格所特有的镜头切换模式。该阶段可

表示为:

$$k = \text{KPM}(m, d, s) = \left\{ \arg \max P(k_t | m, d, s) \right\}_{t=1}^T. (3)$$

可控轨迹合成模块(CTSM)。CTSM 是 ControlCamNet 的核心,负责合成最终的运镜序列 c 。该模块设计为能够同时响应两种用户控制条件:1) 由 KPM 预测或用户直接指定的关键帧序列 k ; 2) 风格标签 s 。在轨迹可控模式下,用户可提供关键帧索引 $U = \{u_1, u_2, \dots\}$ 及其期望的相机参数 $\{c_{u_1}^*, c_{u_2}^*, \dots\}$, CTSM 将以此作为用户控制条件进行合成。

合成过程以相邻关键帧之间的片段为单位迭代进行。对于每对相邻关键帧位置 t_j 和 t_{j+1} ,模型首先基于相机运动历史、当前音乐、舞蹈上下文及风格嵌入 e_s ,并融合用户可能提供的该关键帧处的相机参数,预测这两个关键帧处的相机参数 $\hat{c}_{t_j}$ 和 $\hat{c}_{t_{j+1}}$ 。随后,模型预测一个依赖于风格标签、单调递增的插值函数 $\rho(t; s) \in [0, 1]$ 。最终,所有非关键帧的相机运动通过风格条件化的加权插值获得:

$$c_t = \hat{c}_{t_j} + \rho(t; s)(\hat{c}_{t_{j+1}} - \hat{c}_{t_j}), t \in [t_j, t_{j+1}], (4)$$

该设计确保了每个镜头内部运动的平滑过渡,同时允许不同风格呈现差异化的运动剖面。整个合成过程可形式化为:

$$c = \text{CTSM}(m, d, s, k, \{(u_j, c_{u_j}^*)\}), (5)$$

当用户未提供轨迹控制条件(即 $U = \emptyset$)时, $\{(u_j, c_{u_j}^*)\}$ 为空集,模型退化为基于预测关键帧的纯风格条件合成。

ControlCamNet 的合成主干采用由重构项和运动学项组成的多任务损失函数进行训练:

$$\mathcal{L}_{\text{syn}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}} + \lambda_{\text{acc}} \mathcal{L}_{\text{acc}} + \lambda_{\text{ba}} \mathcal{L}_{\text{ba}}, (6)$$

式中, $\mathcal{L}_{\text{rec}}$ 为逐帧重建损失, $\mathcal{L}_{\text{vel}}$ 和 $\mathcal{L}_{\text{acc}}$ 分别为速度和加速度平滑度损失, $\mathcal{L}_{\text{ba}}$ 为身体注意力损失,鼓励相机聚焦于舞蹈相关的身体部位。各权重在实验中配置为 $\lambda_{\text{rec}} = 2, \lambda_{\text{vel}} = 5, \lambda_{\text{acc}} = 5, \lambda_{\text{ba}} = 0.0075$ 。

1.4 风格感知对抗训练

为增强模型对风格控制条件的响应能力与合成质量,本文设计了风格感知对抗训练(style-aware adversarial training, SAAT),其核心是一个同时承担真实性判别与风格分类的判别器 $\mathcal{D}$,如图4所示。该判别器以相机序列和融合的条件特征为输入,由两个 Transformer 编码器和多层感知器(multilayer

perceptron, MLP) 构成, 旨在同时判别序列的真实性及其风格类别的准确性。

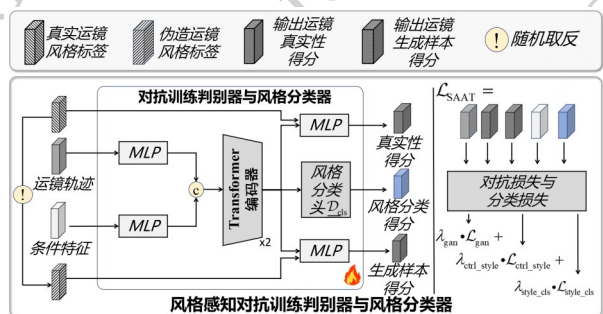


图4 风格感知对抗训练(SAAT)框架图, 包含对抗判别器、风格分类器和三个损失函数

Fig. 4 Framework of style-aware adversarial training (SAAT), including an adversarial discriminator, a style classifier, and three loss functions

SAAT 基于带梯度惩罚的 Wasserstein 生成对抗网络 (Wasserstein generative adversarial network with gradient penalty, WGAN-GP) 框架 (Gulrajani 等, 2017) 设计, 在其基础上引入了风格相关的双重判别机制。风格可控合成的核心难点在于: 不同风格的运动模式, 如编舞与都市, 可能高度相似, 单纯的条件注入难以实现精细化区分。SAAT 通过以下三个损失函数协同解决该问题。

可控风格对比损失 $\mathcal{L}_{ctrl_style}$ 。该损失旨在增强模型对风格控制条件的精确服从, 防止不同风格间的混淆。对于 CtrlDanceCam 合成的样本, 判别器同时接收正确的风格标签 s 和一个随机采样的错误风格标签 $\tilde{s}$ 。生成器的目标是最大化样本在正确标签下的真实性分数, 同时最小化其在错误标签下的分数。该损失基于 WGAN 框架设计:

$$\mathcal{L}_{ctrl_style} = -E[D(CtrlDanceCam(m, d, s)|s)] + E[D(CtrlDanceCam(m, d, s)|\tilde{s})], \quad (7)$$

该设计迫使生成器只有在接收正确风格标签时才能产生高真实性信号, 从而建立风格标签与生成结果之间更强的条件关联。

风格分类损失 $\mathcal{L}_{style_cls}$ 。判别器 $\mathcal{D}$ 包含一个风格分类头 $\mathcal{D}_{cls}$, 通过风格分类损失进行训练。本文采用带标签平滑的交叉熵损失训练此分类头, 以迫使生成器产生与目标风格高度一致的运动:

$$\mathcal{L}_{style_cls} = CE(\mathcal{D}_{cls}(CtrlDanceCam(m, d, s)), s; \alpha), \quad (8)$$

式中 CE 表示交叉熵 (cross-entropy, CE) 损失, α 为标

签平滑因子。

基础对抗损失 $\mathcal{L}_{gan}$ 。采用基础对抗损失驱动生成器输出尽可能接近真实数据分布的样本:

$$\mathcal{L}_{gan} = -E[D(CtrlDanceCam(m, d, s))]. \quad (9)$$

风格感知对抗训练的总损失为上述三个损失函数的加权和:

$$\mathcal{L}_{SAAT} = \lambda_{ctrl_style} \mathcal{L}_{ctrl_style} + \lambda_{style_cls} \mathcal{L}_{style_cls} + \lambda_{gan} \mathcal{L}_{gan}, \quad (10)$$

式中 $\lambda_{ctrl_style} = 4 \times 10^{-4}$, $\lambda_{style_cls} = 5 \times 10^{-5}$, $\lambda_{gan} = 5 \times 10^{-5}$ 。框架总训练目标为合成损失与 SAAT 损失之和:

$$\mathcal{L}_{gen} = \mathcal{L}_{syn} + \mathcal{L}_{SAAT}, \quad (11)$$

判别器与生成器通过交替优化上述损失, 共同提升模型在风格可控合成方面的性能——可控风格对比损失防止风格混淆, 风格分类损失确保风格内一致性, 基础对抗损失提升整体真实性。

1.5 可控合成推理

在推理阶段, CtrlDanceCam 支持两种可控合成模式: 1) 根据输入的音乐 m 、舞蹈 d 及用户指定的风格标签 s , KPM 预测生成关键帧序列 $k = KPM(m, d, s)$; 2) 基于关键帧序列 k 、风格标签 s 以及可能的用户轨迹控制条件, CTSM 合成完整相机序列 c 。

在轨迹可控合成中, 用户输入一组关键帧索引 $U = \{u_1, u_2, \dots\}$ 及其对应的期望相机参数 $\{c_{u_1}^*, c_{u_2}^*, \dots\}$ 作为用户控制条件。在合成阶段, 用户指定的参数将替换模型原本的预测值。具体而言, 对于 $\forall u_j \in U$, 令 $\hat{c}_{u_j} = c_{u_j}^*$ 。合成过程可形式化为:

$$c = CTSM(m, d, s, k, \{(u_j, c_{u_j}^*)\}), \quad (12)$$

该模式允许用户通过稀疏的关键帧输入直观地引导整个运镜轨迹的走向, 实现了轨迹层面的可控性。

在风格可控合成中, 用户通过指定风格标签 s 来控制生成运镜的艺术风格。 s 可以是单一标签, 也可以是一组时序标签 $\{s_i\}_{i=1}^T$ 。风格标签被编码为风格嵌入并注入生成过程, 实现了艺术风格层面的可控性。

结合相机轨迹可控与风格可控合成模式, 本框架统一的推理过程可概括为:

$$\begin{aligned} \mathbf{c} &= \text{CtrlDanceCam}(\mathbf{m}, \mathbf{d}, s, U, \{c_u^*\}) \\ &= \text{CTSM}(\mathbf{m}, \mathbf{d}, s, \mathbf{k}, \{(u_j, c_{u_j}^*)\}) \end{aligned} \quad (13)$$

当 $U = \emptyset$ 时, 退化为无轨迹控制的风格可控合成; 当 s 为数据集已知风格时, 即为本文实验验证的标准风格可控合成; 二者结合则实现了轨迹与风格的双重控制, 充分体现了 CtrlDanceCam 框架在舞蹈运镜跨模态合成中的灵活可控性。

2 实验结果和分析

2.1 数据集与数据处理

实验基于目前该领域唯一公开的三维舞蹈-运镜-音乐 (dance-camera-music, DCM) 配对数据集 (Wang 等, 2024a)。为支撑风格可控合成的研究目标, 对数据集中所有运镜序列进行了细粒度的风格标签标注。模型训练与测试严格遵循原始数据集的既定划分, 以确保实验的公平性与可复现性。在特征提取阶段, 音乐信号通过 Librosa 处理为 $T \times 35$ 维的时序特征 $\mathbf{m} \in \mathbf{R}^{T \times 35}$; 舞蹈动作表征为 60 个人体关节的全局位置序列 $\mathbf{d} \in \mathbf{R}^{T \times 180}$; 相机运镜采用极

坐标下的 MikuMikuDance (MMD) 相机参数表示格式, 表示为 $\mathbf{c} \in \mathbf{R}^{T \times 8}$, 包含参考点位置、相机旋转、距离及视场角等信息; 风格标签 s 经数值编码后广播至时间维度, 形成 $T \times 1$ 的时序对齐条件。

2.2 实现细节

CtrlDanceCam 框架总参数量为 103 M, 其中 KPM 包含 39 M, CTSM 包含 64 M。所有实验均在一台配备两张 NVIDIA RTX PRO 5000 GPU 的工作站上完成。KPM 训练时批次大小为 1024, 学习率为 2×10^{-4} , 采用加权二元交叉熵损失处理类别不平衡问题, 其中正样本损失权重设置为 2.0, 负样本损失权重为 0.5。CTSM 模块批次大小为 512, 学习率同为 2×10^{-4} 。各损失函数的权重配置如下: 重建损失 $\lambda_{\text{rec}} = 2$, 速度平滑损失 $\lambda_{\text{vel}} = 5$, 加速度平滑损失 $\lambda_{\text{acc}} = 5$, 身体注意力损失 $\lambda_{\text{ba}} = 0.0075$ 。对抗训练采用两阶段动态策略: 前 250 个 epoch 判别器学习率为 1×10^{-5} , 梯度惩罚权重 $\gamma = 0.5$; 后续训练中学习率调整为 5×10^{-6} , $\gamma = 0.8$ 。可控风格对比损失权重 $\lambda_{\text{ctrl_style}} = 4 \times 10^{-4}$, 风格分类损失权重 $\lambda_{\text{style_cls}} = 5 \times 10^{-5}$, 对抗损失权重 $\lambda_{\text{gan}} = 5 \times 10^{-5}$ 。总训练轮数为 1200 轮, 使用 Adan 优化器进行训练。

表 2 在 DCM 数据集上与现有方法的对比

Table 2 Comparison with existing methods on the DCM dataset

方法	FID _k ↓	FID _s ↓	Dist _k ↑	Dist _s ↑	DMR ↓	LCD ↓
GT	—	—	3.275	1.731	0.0014	—
吴等	14.295	2.621	3.836	1.914	0.0194	0.188
DanceRevolution	10.267	2.368	1.491	1.491	0.0062	0.154
FACT	5.205	0.960	1.505	1.007	0.0070	0.151
DanceCamera3D	3.749	0.280	1.631	1.326	0.0025	0.147
DanceCamAnimator	3.453	0.268	3.140	1.293	0.0022	0.152
TemMEGA	3.237	0.255	1.961	1.347	0.0020	0.141
CtrlDanceCam	3.022	0.251	3.101	1.355	0.0019	0.140

注: ↑ 表示数值越大性能越优, ↓ 表示数值越小性能越优, 加粗表示对应指标的最优结果。

2.3 评价指标

为全面评估生成运镜轨迹的质量, 遵循 DCM (Wang 等, 2024a) 的评价指标, 从生成质量、多样性和舞蹈演员跟随保真度三个方面进行评价, 包括: 用于衡量生成质量的弗雷歇起始距离 (Fréchet inception distance, FID_k/FID_s), 用于衡量多样性的平均欧氏距离 (average Euclidean distance, Dist_k/Dist_s), 以及

用于评估舞蹈演员跟随保真度的舞者丢失率 (dancer missing rate, DMR) 与肢体捕捉差异 (limbs capture difference, LCD)。其中, 下标 k 和 s 分别表示运动学特征 (kinetic features) 空间和镜头特征 (shot features) 空间。

2.4 与现有方法的对比

表 2 展示了本文方法与多种现有方法在 DCM
© 中国图象图形学报版权所有

数据集上的对比结果,对比方法包括吴等(2022)、DanceRevolution (Huang 等, 2021)、FACT (Li 等, 2021)、DanceCamera3D (Wang 等, 2024a)、DanceCamAnimator (Wang 等, 2024b) 和 TemMEGA (Xu 等, 2025)。CtrlDanceCam 在生成质量和舞蹈演员跟随保真度方面表现突出,其中 FID_t 和 FID_s 分别为 3.022 和 0.251,DMR 和 LCD 分别为 0.0019 和 0.140,优于其他对比方法;在轨迹多样性方面, $Dist_t$ 和 $Dist_s$ 分别

达到 3.101 和 1.355,保持了较高的生成多样性。总体结果表明,本文方法在提升运镜生成质量与舞者跟随保真度的同时,能够兼顾轨迹多样性。KPM 与 CTSM 通过风格条件化建模分别学习不同风格的关键帧分布和镜头内运动模式,SAAT 进一步增强了生成结果的真实性与风格一致性。



图5 四种风格运镜可视化。歌曲1~4分别对应韩流、爵士、传统中国舞、都市风格。地面网格提供空间参考
Fig. 5 Qualitative results for four styles. Song 1~4: Korean, Jazz, Traditional-Chinese, Urban. Ground grid provides spatial reference



图6 可控合成可视化。(a)基线(同图5(d));(b)仅风格控制(编舞);(c)轨迹+风格双控制
Fig. 6 Controllable synthesis. (a) baseline, same as Fig. 5(d); (b) style control only (Choreography); (c) trajectory + style control
© 中国图象图形学报版权所有

2.5 可控合成分析

图5展示了四种风格下的运镜可视化结果,并以歌曲编号(歌曲1—4)区分不同样例。图5(a)对应歌曲1(韩流风格),呈现出明显的景别变化,近景与全身构图交替出现,体现出较活跃的镜头调度;图5(b)对应歌曲2(爵士风格),整体取景距离较近,并表现出较明显的低角度拍摄特征;图5(c)对应歌曲3(传统中国舞风格),更多采用完整人体构图和较开阔的取景方式,镜头变化相对平稳;图5(d)对应歌曲4(都市风格),整体构图变化较为流畅,人物在画面中的尺度和位置变化相对平缓。上述可视化结果表明,不同风格标签能够使生成运镜在景别、拍摄视角和构图方式等方面呈现出差异化特征。

图6展示了用户交互控制下的生成结果,三个子图构成“自动生成→风格切换→轨迹修正”的完整可视化过程。图6(a)为图5(d)中都市风格的自动生成结果,作为可控合成的参照基线;图6(b)在保持音乐与舞蹈输入不变的情况下,将风格标签由都市切换为编舞,生成结果在景别变化和构图方式上表现出明显差异,说明风格条件能够有效调节整体

表3 风格感知对抗训练(SAAT)消融实验

Table 3 Results of ablation experiment on style-aware adversarial training (SAAT)

方法	FID_k ↓	FID_s ↓	$Dist_k$ ↑	$Dist_s$ ↑	DMR ↓	LCD ↓
w/o SAAT	3.512	0.345	3.241	1.304	0.0030	0.155
w/o $\mathcal{L}_{gan}$	3.423	0.322	3.244	1.347	0.0028	0.154
w/o $\mathcal{L}_{style_cls}$	3.093	0.278	3.201	1.366	0.0017	0.148
w/o $\mathcal{L}_{ctrl_style}$	3.193	0.317	3.239	1.361	0.0021	0.151
Full CtrlDance-Cam	3.022	0.251	3.101	1.355	0.0019	0.140

注: ↑表示数值越大越优, ↓表示数值越小越优;加粗表示对应指标的最优结果;w/o表示移除相应模块或损失项。

运镜模式;图6(c)进一步在编舞风格基础上对部分关键帧相机参数进行人工指定,使相关时刻附近的构图和拍摄视角发生定向调整。对比图6(b)与图6(c)可见,用户通过少量关键帧约束即可对生成结果进行局部修正,从而实现轨迹层面的可控编辑。

表4 关键帧预测模块(KPM)消融实验结果

Table 4 Results of ablation experiment on keyframe prediction module (KPM)

方法	FID_k ↓	FID_s ↓	$Dist_k$ ↑	$Dist_s$ ↑	DMR ↓	LCD ↓
Uniform-KF	3.742	0.366	2.834	1.196	0.0030	0.156
KPM w/o style cond.	3.382	0.331	2.981	1.274	0.0024	0.149
Full CtrlDanceCam	3.022	0.251	3.101	1.355	0.0019	0.140

注: ↑表示数值越大越优, ↓表示数值越小越优;加粗表示对应指标的最优结果;w/o表示移除相应条件。

3 讨论

为进一步验证 CtrlDanceCam 各核心组件的贡献以及方法在不同配置下的性能表现,本节从消融实验和关键帧密度敏感性两个维度展开讨论。消融实验涵盖 SAAT 风格感知对抗训练及其子损失项、KPM 关键帧预测模块和 CTSM 可控轨迹合成模块三个层面;关键帧密度敏感性分析考察固定密度设置与自适应策略对合成质量的影响。

3.1 SAAT 风格感知对抗训练消融

表3展示了 SAAT 及其各子损失项的消融结果。完整模型在生成质量和整体综合性能上表现最佳,同时保持了良好的轨迹多样性与舞者跟随保真度。移除整个 SAAT 模块导致 FID_k 升至 3.512、 FID_s 升至 0.345,验证了对抗训练在提升生成真实性与风格一致性方面的关键作用。在各子损失中, $\mathcal{L}_{gan}$ 的移除对生成质量影响较大(FID_k 升至 3.423),证明基础对抗损失是保证输出真实性的主要部分;移除可控风格对比损失 $\mathcal{L}_{ctrl_style}$ 后, FID_s 升至 0.317,表明该损失有助于减少不同风格间的混淆并增强风格区分能力。移除 $\mathcal{L}_{style_cls}$ 后, FID_s 由 0.251 升至 0.278,虽然 $Dist_s$ 小幅升至 1.366,但 FID_s 和 LCD 等指标出现退化,表明风格分类损失主要用于增强风格一致性,而非单纯提升多样性。消融结果表明各子损失协同工作,使框架同时实现高保真轨迹生成与精准风格控制。

3.2 KPM 关键帧预测模块消融

KPM 与 CTSM 在框架中构成管道式协作关系——CTSM 以 KPM 预测的关键帧为合成锚点,直接移除任一模块将导致推理流程断裂。为验证 KPM 中关键设计要素的贡献,进行替代性消融:(a)

Uniform-KF:将 KPM 预测的关键帧替换为均匀间隔采样,送入 CTSM 正常合成;(b) KPM w/o style cond.:保留 KPM 结构但移除风格嵌入条件 e_s ,仅基于音乐和舞蹈特征预测关键帧。实验结果如表 4 所示。

由表 4 可知,Uniform-KF 在各项指标上均不及完整模型, FID_k 从 3.022 升至 3.742, FID_s 从 0.251 升至 0.366。均匀采样无法捕捉风格相关的镜头切换模式:不同风格具有差异化的关键帧密度偏好和时序分布特征——爵士风格关键帧稀疏(密度 2.640%)、镜头长度较长,而锁舞风格关键帧密集(密度 31.484%)、切换频繁。KPM 通过风格嵌入学习到了这些差异化的关键帧分布。KPM w/o style cond. 变体中风格相关指标退步幅度更大(FID_s 从 0.251 升至 0.331),说明风格嵌入主要影响关键帧的时序分布模式。移除风格条件后,KPM 退化为风格无关的通用关键帧检测器,无法为不同风格提供差异化的镜头切换策略,导致合成运镜的风格区分度下降。综上,KPM 通过风格条件化的关键帧预测,为下游 CTSM 提供了风格感知的合成锚点,是框架风格可控能力的上游基础。

3.3 CTSM 可控轨迹合成模块消融

CTSM 内部包含三个步骤:1)预测关键帧处相机参数;2)预测风格条件插值函数 $\rho(t; s)$;3)按公式插值合成运镜。其中步骤 1)是核心生成能力,步骤 3)是纯算术操作,唯一可独立消融的是步骤 2)—— $\rho(t; s)$ 。 $\rho(t; s)$ 决定了镜头内从关键帧 A 到关键帧 B 的过渡节奏。将 $\rho(t; s)$ 替换为固定线性插值(Linear)后,实验结果如表 5 所示。

表 5 可控轨迹合成模块(CTSM)消融实验
Table 5 Results of ablation experiment on controllable trajectory synthesis module (CTSM)

方法	$FID_k \downarrow$	$FID_s \downarrow$	DMR $\downarrow$	LCD $\downarrow$
Linear	3.351	0.346	0.0023	0.148
Full	3.022	0.251	0.0019	0.140

注:↓表示数值越小越优;加粗表示对应指标的最优结果。

由表 5 可知,将 $\rho(t; s)$ 替换为线性插值后, FID_k 和 FID_s 均出现退化,但 FID_s 的退步(从 0.251 升至 0.346)显著大于 FID_k 的退步(从 3.022 升至 3.351)。 FID_k 的小幅上升说明即使使用线性插值,CTSM 仍能

预测合理的关键帧相机参数,轨迹整体空间走向保持基本正常。 FID_s 的明显恶化揭示了 $\rho(t; s)$ 的核心作用:风格可控合成不仅需要关键帧分布的风格化(由 KPM 负责),也需要镜头内运动节奏的风格化(由 $\rho(t; s)$ 负责)。不同风格的 FOV 标准差和相机速度差异显著(如表 1 所示), $\rho(t; s)$ 赋予不同风格以差异化的镜头内过渡曲线——线性插值消解了这些差异,使所有风格退化为统一的匀速过渡。该实验表明,CTSM 中的 $\rho(t; s)$ 通过风格条件化的镜头内运动剖面,对运镜风格表达产生了独立于 KPM 关键帧预测的贡献。结合表 4 的 KPM 消融,框架的风格可控性由关键帧分布的风格和镜头内过渡的风格两个层面协同实现。

表 6 关键帧密度敏感性分析
Table 6 Sensitivity analysis under keyframe density

方法	$FID_k \downarrow$	$FID_s \downarrow$	$Dist_k \uparrow$	$Dist_s \uparrow$	DMR $\downarrow$	LCD $\downarrow$
3%(稀疏)	3.754	0.365	2.761	1.176	0.0031	0.157
5%	3.436	0.318	2.938	1.263	0.0025	0.149
10%	3.218	0.286	3.047	1.327	0.0022	0.145
20%	3.509	0.334	2.872	1.232	0.0026	0.151
30%(密集)	3.872	0.383	2.603	1.115	0.0031	0.159
Full CtrlDance-Cam	3.022	0.251	3.101	1.355	0.0019	0.140

注:↑表示数值越大越优,↓表示数值越小越优;加粗表示对应指标的最优结果。

3.4 关键帧密度敏感性分析

为探究关键帧密度对合成质量的影响,在不同关键帧占比下评估模型性能。关键帧位置由 KPM 按置信度排序选取 top-K 个,所有变体使用完整的 CTSM。实验结果如表 6 所示。

由表 6 可知,固定关键帧密度难以同时适配不同运镜风格。在稀疏端(3%),关键帧数量不足导致 CTSM 缺乏足够的合成锚点,运镜轨迹在长镜头内出现漂移和抖动, FID_k 升至 3.754,DMR 升至 0.0031。在密集端(30%),过多关键帧过度约束了生成空间,CTSM 退化为简单内插而丧失轨迹的生成多样性, FID_k 升至 3.872, $Dist_k$ 降至 2.603。10%接近全局平均密度,因此在固定密度设置中表现相对较好($FID_k=3.218$)。3%和 5%的低密度设置更接近

爵士(2.640%)、都市(3.389%)等低切换频率风格,但对于锁舞(31.484%)、机械舞(12.124%)等高密度风格缺少足够合成锚点;20%和30%的高密度设置虽能覆盖高频切换风格,但对多数中低密度风格形成过强约束。相比之下,KPM自适应预测能够根据风格标签和音乐-舞蹈内容动态调整关键帧分布,在全部指标上取得最优结果($FID_k=3.022$),避免了全局统一密度带来的问题。

4 结 论

针对舞蹈运镜自动合成中可控性不足的问题,本文提出CtrlDanceCam多风格舞蹈运镜可控合成框架。该框架通过以下三个层面的设计实现轨迹与风格的可控合成。

在方法层面,本文构建了ControlCamNet骨干网络。KPM基于Transformer编码器将风格标签编码为条件嵌入,预测符合目标风格特征的关键帧分布;CTSM以关键帧为锚点,通过风格条件化的可学习插值函数 $\rho(t; s)$ 合成完整相机运动序列。模块化的管道设计将可控合成任务分解为可解释的子问题:风格条件化关键帧预测负责镜头间的风格区分,风格条件化插值负责镜头内的运动剖面差异化。消融实验验证了这一分工的有效性——KPM风格嵌入的移除使风格指标 FID_s 从0.251升至0.331, $\rho(t; s)$ 替换为线性插值使 FID_s 升至0.346,两者对风格表达产生独立且互补的贡献。

在训练策略层面,本文设计了SAAT风格感知对抗训练。三个协同优化的损失函数——可控风格对比损失、风格分类损失和基础对抗损失——共同防止多风格间混淆并提升生成质量。SAAT消融表明,移除可控风格对比损失后 FID_s 由0.251升至0.317,说明该损失有助于减少不同风格间的混淆并增强风格区分能力;同时,正确风格标签下更高的真实性响应进一步强化了风格标签与生成结果之间的条件关联。

在数据层面,本文对DCM数据集进行了八类运镜风格的细粒度标注。统计特征分析揭示了不同风格在关键帧密度、镜头长度、FOV变化和相机速度等维度上的可量化差异(如爵士关键帧密度为2.640%,锁舞为31.484%),为该领域的风格化研究

提供了数据基础与分析依据。

在DCM数据集上,CtrlDanceCam在生成质量($FID_k=3.022$, $FID_s=0.251$)和舞蹈演员跟随保真度($DMR=0.0019$, $LCD=0.140$)方面优于现有方法,同时在多样性指标($Dist_k=3.101$, $Dist_s=1.355$)上保持了较高的轨迹多样性。可视化实验表明,不同风格标签能够产生差异化的运镜轨迹,稀疏的用户关键帧约束能够精准引导轨迹走向。

本文方法存在以下局限。第一,轨迹控制灵活性受限于DCM数据集的标注内容,目前仅支持关键帧相机参数指定,尚未覆盖路径约束、速度约束和避障等更精细的编辑方式。第二,风格标签体系基于人工标注,八类风格对运镜艺术表达的覆盖度有限,且风格边界的模糊性(如编舞与都市在统计特征上存在部分重叠)可能影响风格控制的精度。第三,SAAT虽通过可控风格对比损失实现了隐式风格解耦,但未引入显式风格解耦或跨模态对比学习等更先进的技术,风格控制的细粒度仍有提升空间。

当前公开的DCM数据集仅提供相机位姿及关键帧相关标注,缺乏路径约束、速度约束和避障等精细化用户控制条件的对应标注,成为进一步扩展轨迹控制能力的主要瓶颈。未来将围绕精细化控制数据构建展开研究,通过补充路径、速度及避障等用户控制条件标注,并利用CTSM的模块化接口将其引入生成过程,以提升轨迹控制的灵活性与精细度。在风格建模方面,将进一步探索显式风格解耦与跨模态对比学习,提升风格控制的细粒度与鲁棒性;同时扩充和细化风格标签体系,以覆盖更丰富的运镜艺术表达。在应用方面,将进一步探索实时交互式编辑以及上游舞蹈生成模型的联合优化。

参考文献(References)

- Alexanderson S, Nagy R, Beskow J and Henter G E. 2023. Listen, denoise, action! Audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics*, 42 (4): 44: 1-44: 20 [DOI: 10.1145/3592458]
- Bares W H, McDermott S, Boudreaux C and Thainimit S. 2000. Virtual 3D camera composition from frame constraints//*Proceedings of the Eighth ACM International Conference on Multimedia*. Los Angeles: ACM: 177-186 [DOI: 10.1145/354384.354463]
- Christie M, Olivier P and Normand J M. 2008. Camera control in computer graphics. *Computer Graphics Forum*, 27 (8): 2197-2218

- [DOI: 10.1111/j.1467-8659.2008.01181.x]
- Dhariwal P and Nichol A. 2021. Diffusion models beat GANs on image synthesis//Advances in Neural Information Processing Systems 34 (NeurIPS). Virtual: Curran Associates: 8780-8794.
- Evin I, Hamalainen P and Guckelsberger C. 2022. Cine-AI: Generating video game cutscenes in the style of human directors. Proceedings of the ACM on Human-Computer Interaction, 6(CHI PLAY): 1-23 [DOI: 10.1145/3549486]
- Gao X, Liu D Y and Zhang J Y. 2024. Multi-modal digital human modeling, synthesis, and driving: a survey. Journal of Image and Graphics, 29(9): 2494-2512 (高玄, 刘东宇, 张举勇. 2024. 多模态数字人建模、合成与驱动综述. 中国图象图形学报, 29(9): 2494-2512) [DOI: 10.11834/jig.230649]
- Gschwindt M, Camei E, Bonatti R, Wang W, Kayacan E and Scherer S A. 2019. Can a robot become a movie director? Learning artistic principles for aerial cinematography//Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China: IEEE: 1107-1114 [DOI: 10.1109/IROS40897.2019.8967592]
- Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C. 2017. Improved training of Wasserstein GANs//Advances in Neural Information Processing Systems 30 (NeurIPS). Long Beach: Curran Associates: 5767-5777 [DOI: 10.5555/3295222.3295327]
- Huang C, Lin C E, Yang Z, Kong Y, Chen P, Yang X, et al. 2019. Learning to film from professional human motion videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE: 4244-4253 [DOI: 10.1109/CVPR.2019.00437]
- Huang R, Hu H, Wu W, Sawada K, Zhang M and Jiang D X. 2021. Dance revolution: Long-term dance generation with music via curriculum learning//Proceedings of the 9th International Conference on Learning Representations (ICLR). Virtual: OpenReview.net
- Huang X Y, Zhang S Y, Zhang Q, Guo X X and Ye L. 2026. FODGE: High-fidelity dance generation via full-body optimization//Proceedings of the 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE: 12552-12556 [DOI: 10.1109/ICASSP55912.2026.11463867]
- Jiang H, Wang B, Wang X, Christie M and Chen B. 2020. Example-driven virtual cinematography by learning camera behaviors. ACM Transactions on Graphics, 39(4): 45: 1-45: 14 [DOI: 10.1145/3386569.3392427]
- Jiang H, Wang X, Christie M, Liu L and Chen B. 2024. Cinematographic camera diffusion model. Computer Graphics Forum, 43(2): e15055 [DOI: 10.1111/cgf.15055]
- Kim J, Oh H, Kim S, Tong H and Lee S. 2022. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE: 3480-3490 [DOI: 10.1109/CVPR52688.2022.00348]
- Li C G, Wen Y H, Jing Y C, Wang Y, Zhang Y Y, Zhou Z Z, et al. 2025. Shape-aware stylized dance motion generation driven by music. Journal of Computer-Aided Design and Computer Graphics, [Online first]: 1-13 (李晨光, 温玉辉, 景宇宸, 汪意, 张亚宇, 周正忠, 等. 2025. 体型感知的音乐驱动舞蹈动作风格化生成. 计算机辅助设计与图形学学报, [在线首发]: 1-13 [DOI: 10.3724/SP.J.1089.2024-00492])
- Li R, Yang S, Ross D A and Kanazawa A. 2021. AI choreographer: Music conditioned 3D dance generation with AIST++//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE: 13381-13392 [DOI: 10.1109/ICCV48922.2021.01315]
- Li R, Zhao J, Zhang Y, Su M, Ren Z, Zhang H, et al. 2023. Fine-Dance: A fine-grained choreography dataset for 3D full body dance generation//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE: 10200-10209 [DOI: 10.1109/ICCV51070.2023.00939]
- Li R, Zhang Y, Zhang Y, Zhang H, Guo J, Zhang Y, et al. 2024. LODGE: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 1524-1534 [DOI: 10.1109/CVPR52733.2024.00151]
- Li R H, Zhang H W, Zhang Y C, Zhang Y X, Zhang Y L, Guo J, et al. 2025. LODGE++: High-quality and long dance generation with robust choreography patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, [Online first]: 1-15 [DOI: 10.1109/TPAMI.2025.3618016]
- Li S Y, Yu W J, Gu T P, Lin C Z, Wang Q, Qian C, et al. 2022. Bailando: 3D dance generation by actor-critic GPT with choreographic memory//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE: 11040-11049 [DOI: 10.1109/CVPR52688.2022.01077]
- Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL].[2026-08-12]. <https://arxiv.org/abs/1411.1784> [DOI: 10.48550/arXiv.1411.1784]
- Pueyo P, Dendarieta J, Montijano E, Murillo A C and Schwager M. 2024. CineMPC: A fully autonomous drone cinematography system incorporating zoom, focus, pose, and scene composition. IEEE Transactions on Robotics, 40: 1740-1757 [DOI: 10.1109/TRO.2024.3353550]
- Rao A, Jiang X, Guo Y, Xu L, Yang L, Jin L, et al. 2023. Dynamic storyboard generation in an engine-based virtual environment for video production//ACM SIGGRAPH 2023 Posters. Los Angeles: ACM: 40: 1-40: 2 [DOI: 10.1145/3588028.3603647]
- Ren P and Yang H. 2025. LILAC: Long-sequence incremental low-latency arbitrary motion stylization via streaming VAE-diffusion with causal decoding[EB/OL].[2026-06-15]. <https://arxiv.org/abs/2510.15392>

- Rucks J and Katzakis N. 2021. CamerAI: Chase camera in a dense environment using a proximal policy optimization-trained neural network//Proceedings of the 2021 IEEE Conference on Games (CoG). Copenhagen: IEEE: 4-8 [DOI: 10.1109/CoG52621.2021.9619120]
- Shi M, Sun B L, Zhu D M and Zhuo X R. 2025. KM2D: Method for generating dance animation driven by dance movement primitives and musical semantics. Journal of Computer-Aided Design and Computer Graphics, [Online first]: 1-12 (石敏, 孙碧莲, 朱登明, 嵯心如. 2025. KM2D: 舞蹈动作基元符号和音乐语义驱动的舞蹈动画生成方法. 计算机辅助设计与图形学学报, [在线首发]: 1-12) [DOI: 10.3724/SP.J.1089.2024-00325]
- Tseng J, Castellon R and Liu C K. 2023. EDGE: Editable dance generation from music//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE: 448-458 [DOI: 10.1109/CVPR52729.2023.00051]
- Wang Z X, Jia J, Sun S K, Wu H, Han R, Li Z, et al. 2024a. DanceCamera3D: 3D camera movement synthesis with music and dance//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 7892-7901 [DOI: 10.1109/CVPR52733.2024.00754]
- Wang Z X, Li J Y, Qin X Y, Sun S K, Zhou S, Jia J, et al. 2024b. DanceCamAnimator: Keyframe-based controllable 3D dance camera synthesis//Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM). Melbourne: ACM: 10200-10209 [DOI: 10.1145/3664647.3680980]
- Wu S, Lu S and Cheng L. 2022. Music-to-dance generation with optimal transport//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI). Vienna: IJCAI: 4988-4994 [DOI: 10.24963/ijcai.2022/691]
- Wu X Y, Min J X, Qiu J C and Wu J Q. 2023. Intelligent virtual filming system for cloud performances based on camera position calculation. Journal of Communication University of China (Science and Technology), 30(4): 17-26 (吴晓雨, 闵静萱, 邱驹成, 吴建琴. 2023. 基于机位计算的云演艺智能虚拟拍摄系统. 中国传媒大学学报(自然科学版), 30(04): 17-26)
- Xie C, Hemmi I, Shishido H and Kitahara I. 2023. Camera motion generation method based on performer's position for performance filming//Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE). Nara: IEEE: 957-960 [DOI: 10.1109/GCCE59613.2023.10315539]
- Xu C H, Lyu G T, Yan J X, Yang M L and Deng C. 2025. Smooth and flexible camera movement synthesis via temporal masked generative modeling//Advances in Neural Information Processing Systems 38 (NeurIPS). San Diego: Curran Associates: 13217-13235 [DOI: 10.52202/085713-0446]
- Zhang J M, Sun Y L, Zhang H W and Tang J H. 2025. EDMG: Towards efficient long dance motion generation with fundamental movements from dance genres//Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM). Dublin: ACM: 10447-10456 [DOI: 10.1145/3746027.3755677]
- Zhao B Q, Fu Y Y, Su Z, Wang R M, Lyu C L and Luo X N. 2024. A survey on multimodal information-guided 3D human motion generation. Journal of Image and Graphics, 29(9): 2541-2565 (赵宝全, 付一榆, 苏卓, 王若梅, 吕辰雷, 罗笑南. 2024. 多模态信息引导的三维数字人运动生成综述. 中国图象图形学报, 29(9): 2541-2565) [DOI: 10.11834/jig.230626]