DanceStyleCam: Style-Based 3D Multi-Style Dance Camera Movement Synthesis

Xiaoying Huang[1], **Sanyi Zhang**[2,5], **Xirui Wang**[3], **Qin Zhang**[5] and **Long Ye**[2,4]
[1]School of Information and Communication Engineering, Communication University of China, China
[2]School of Data Science and Media Intelligence, Communication University of China, China
[3]Hainan International College of Communication University of China, China
[4]State Key Laboratory of Media Convergence and Communication, Communication University of China
[5]Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education, China
{xiaoyinghuang, zhangsanyi, zhangqin, yelong}@cuc.edu.cn, 202329013018n@mails.cuc.edu.cn

Abstract

Fully automatic camera movement directly affects the art quality of dance expressiveness, especially in terms of visual expression, as well as choreography and music. Current studies mainly focus on synthesizing camera movements conditioned on dance and music, but they overlook the camera movement style, which is essential factor for artistic and visual coherence. In this paper, we introduce DanceStyleCam, a unified framework that incorporates the style-consistent characteristic into dance camera movement synthesis with diverse stylistic characteristics. Specifically, a style-aware feature learning module is proposed to map dance style information into compact embeddings, facilitating stable and discriminative style learning. To further guarantee that the generated camera movements remain faithful to the target style, we propose a style-consistent adversarial training scheme, leading and optimizing the model to learn better style-consistent representations. In addition, we also enrich the DCM dataset with diverse camera movement style annotations. Extensive experiments demonstrate that DanceStyleCam outperforms state-of-the-art methods in both generation quality and style consistency. Qualitative results further show that our method produces stylistically consistent camera movements while preserving smooth trajectories and natural shot transitions. Project page: https://github.com/YCCCCM/DanceStyleCam.

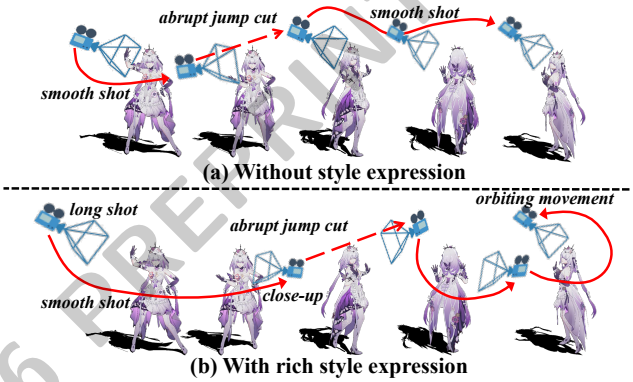

Figure 1: Comparison with and without style expression for camera movement synthesis. **(a)** Existing models without style expression tend to incorporate a limited set of camera movement patterns, **(b)** while our proposed method with rich style expression is good at mimicking a diverse range of camera movement patterns associated with specific style factors.

1 Introduction

With the advancement of digital media, dance performances driven by digital avatars have become a hot research area [Tseng *et al.*, 2023]. In such performances, camera work plays a crucial role in enhancing the expressive power of dance movements [Xie *et al.*, 2023]. However, carefully designing camera movement for a digital dance sequence re mains a time-consuming and skill-dependent task, which motivates us to explore fully automatic camera movement synthesis as a viable solution that inspires novel choreographic perspectives and enhances productivity for dance creators.

Earlier research on camera planning and control focused mainly on constraint-based methods [Christie *et al.*, 2008; Bares *et al.*, 2000], which was applied to gaming [Li and Cheng, 2008], film [Jiang *et al.*, 2024b], and drone filming [Huang *et al.*, 2021] applications. Recently, several studies have aimed to tackle the more challenging task of music- and dance-driven camera synthesis. Among these, DanceCamera3D [Wang *et al.*, 2024a] introduced the first 3D dance-camera-music dataset (DCM) and demonstrated the feasibility of synthesizing camera movements conditioned on music and dance. DanceCamAnimator [Wang *et al.*, 2024b] integrated prior knowledge from animation practice, mimicking animators' hierarchical camera-authoring process to generate camera movement. Cine-AI [Evin *et al.*, 2022] simplified the problem by reducing camera synthesis from 3D to 2D, removing roll and pitch orientations of the camera. This reduction significantly lowers the expressiveness of the generated camera movements while decreasing task complexity. TemMEGA [Xu *et al.*, 2025] further proposed a temporal masked generative modeling framework for offline and online

camera movement synthesis.

Though the aforementioned methods have made notable progress in camera movement synthesis, they focus on generating camera movement from the content of music and dance, overlooking the crucial relation between artistic expression and camera movement style. In professional cinematography, camera movement is inherently stylistic, while camera style can be defined as an explicit factor to convey specific narrative emotions or visual aesthetics, as demonstrated in cinematography [Jiang *et al.*, 2024a; Rao *et al.*, 2020]. However, in camera synthesis, this important stylistic factor remains underexplored and unmodeled, as shown in Figure 1. Therefore, how to achieve style-consistent dance camera movement synthesis automatically is a key but underexplored challenge. Besides, the newly synthesized camera movements should not only be synchronized with music and dance content but also exhibit distinct artistic stylistic expression.

To address the above limitations, we introduce DanceStyleCam ($\mathcal{DSC}$), a unified framework that incorporates the style-consistent characteristic into dance camera movement synthesis with diverse stylistic characteristics. The proposed DanceStyleCam consists of the following two core modules: (a) A style-conditioned camera movement synthesis network. We design a style-aware feature learning module that maps dance style information into compact embeddings. We combine the dance poses, music features, and the style embedding as joint input and feed it into the generator to produce style-conditioned camera movements. (b) A style-consistent adversarial training framework. To enhance style discriminability and generation quality, we design a training strategy based on a dual adversarial loss. In the meantime, the style consistency and movement constraints are enforced to ensure the synthesized movements are synchronized with music and dance while remaining visually faithful to the specified style. In addition, to support the style-consistent learning, we extend and annotate an existing dance-camera dataset. Building upon the DCM dataset, we introduce fine-grained style labels and align it with the FineDance dataset [Li *et al.*, 2023], resulting in a style-conditioned dance-camera dataset that supports style-driven synthesis. Extensive experiments demonstrate that our proposed method significantly outperforms existing methods in terms of both style consistency and content synchronization, validating the effectiveness of DanceStyleCam in the task of camera movement synthesis.

Our main contributions can be summarized as follows:

- We identify camera movement style as a previously underexplored yet critical factor in dance camera movement synthesis, and establish the first style-annotated benchmark for this task by enriching an existing 3D dance–camera–music dataset with explicit camera movement style labels.
- We propose **DanceStyleCam**, a unified framework for style-consistent dance camera movement synthesis, which explicitly disentangles content-driven movement synthesis from style conditioning via a dedicated style encoding mechanism and a style-consistent adversarial training strategy, enabling a single model to synthesize camera movements in multiple cinematographic styles.
- Extensive quantitative and qualitative evaluations demonstrate that **DanceStyleCam** consistently outperforms state-of-the-art methods, achieving superior generation quality while preserving stylistic fidelity and smooth, natural camera transitions.

2 Related Work

2.1 Camera Control and Planning

Automated camera control and planning has garnered significant attention in recent years due to the high level of expertise required for manually creating cinematic video, which is crucial for the media, entertainment, and gaming industries. In the fields of virtual environments and film production, a line of research focuses on the automated planning of camera behaviors, such as extracting reusable shot patterns from film clips or generating dynamic shots [Jiang *et al.*, 2020; Rao *et al.*, 2023; Wu *et al.*, 2023; Xie *et al.*, 2023; Pueyo *et al.*, 2024; Yang *et al.*, 2023]. For games and real-time rendering, related methods emphasize simulating director-level cinematographic techniques within third-person perspectives or cutscenes [Rucks and Katzakis, 2021; Evin *et al.*, 2022]. In aerial cinematography, a series of studies have also automated drone camera movements according to artistic principles [Gschwindt *et al.*, 2019; Huang *et al.*, 2021; Huang *et al.*, 2019a; Huang *et al.*, 2019b]. However, controlling cameras for dance is more complex, as it requires precise synchronization with both musical rhythm and motion.

Wang et al. [Wang *et al.*, 2024a] introduced DCM dataset and proposed DanceCamera3D, a Transformer-based diffusion model for synthesis. While capable of generating basic motion, this approach did not fully account for the mix of continuous camera movements and abrupt cuts characteristic of dance cinematography. DanceCamAnimator [Wang *et al.*, 2024b] achieved finer control over variable-length sequences by incorporating animator priors into a three-stage network, and Xu et al. [Xu *et al.*, 2025] further proposed TemMEGA, a temporal masking generative model for online synthesis.

Although the aforementioned works have advanced dance-driven camera movement synthesis, their focus remains primarily on generating motion from music and dance content, overlooking the artistic stylistic dimension inherent to camera movements itself. In professional cinematography, shot style is a key factor for conveying narrative emotion and visual aesthetics [Jiang *et al.*, 2024a; Rao *et al.*, 2020]. Yet, in camera synthesis, this important stylistic factor has not been systematically explored or modeled. Therefore, this paper addresses the critical and underexplored challenge of achieving style-consistent dance camera movement synthesis.

2.2 Music-Driven 3D Dance Generation

The music-driven 3D dance generation tasks share several core technical challenges with dance-driven camera synthesis, such as processing music audio and motion sequences and extracting spatio-temporal features. Early methods primarily utilized sequence-to-sequence models [Li *et al.*, 2021] and VQ-VAE [Siyao *et al.*, 2022]. Recent advances increasingly employ Transformer-based diffusion models for high-quality dance generation [Alexanderson *et al.*, 2023;

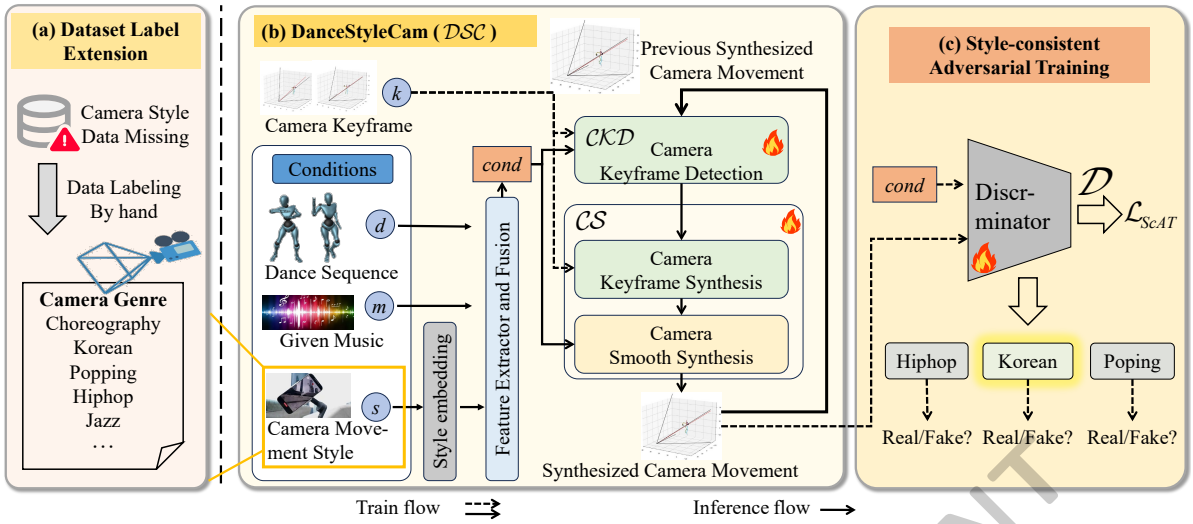

Figure 2: **Framework overview of the proposed DanceStyleCam.** Our method comprises three key components: the **Dataset Label Extension**, the two-stage $\mathcal{DSC}$ **camera movement synthesis models**, and the **style-consistent adversarial training (ScAT)**.

Li *et al.*, 2024b]. EDGE [Tseng *et al.*, 2023] integrates a Transformer-based diffusion model with Jukebox [Dhariwal *et al.*, 2020], enabling advanced editing capabilities. Kim et al. [Kim *et al.*, 2022] and Wu et al. [Wu *et al.*, 2022] leverage Generative Adversarial Networks (GAN) to investigate genre-specific control and dual learning between dance and music. LODGE [Li *et al.*, 2024a] further develops a coarse-to-fine, two-stage diffusion framework that incorporates adversarial training for style-diverse dance generation.

Notably, as a pivotal upstream task, contemporary dance generation research explicitly models dance style as a key factor [Li *et al.*, 2024a]. This provides a crucial precedent for its downstream counterpart: camera synthesis for dance. For a complete generative system to produce a cohesive artistic piece, the style expression of the camera must be consciously aligned with the dance motion. This inter-task stylistic consistency is essential for the final presentation but remains an unexplored dimension in existing camera synthesis.

3 Method

Given a **style label** s, a sequence of T frames of music $m = \{m_1, m_2, \ldots, m_T\}$ and dance motion $d = \{d_1, d_2, \ldots, d_T\}$, our goal is to synthesize the full camera movement sequence $c = \mathcal{DSC}(m, d, s) = \{c_1, c_2, \ldots, c_T\}$. Specifically, we introduce an explicit **style label** s as a novel conditioning variable—obtained through manual annotation—to enable the synthesis of style-consistent camera movements. We further propose a style-consistent adversarial training strategy that enhances the model's ability to understand and generate stylized dance cinematography. The overall pipeline of our method is illustrated in Figure 2.

3.1 Dataset Label Extension

To enable research on style-consistent dance camera synthesis, we augment the existing standard DCM dataset with fine-grained style annotations. While the original DCM dataset provides precisely aligned "music–dance–camera" triplets, it lacks precise descriptions of the artistic style of camera movements, containing only song-language tags. To bridge this gap, we construct the first dance–camera dataset with style annotations by extending the DCM dataset through systematic multimodal analysis and collaborative expert labeling.

Our annotation procedure is built upon a holistic analysis of the music, dance, and camera movement. Concretely, for each data sample, we simultaneously examine its musical characteristics (e.g., rhythm and melody), dance movements (e.g., movement vocabulary and intensity), and camera movement patterns (e.g., editing pace and shot transitions) to determine its dominant artistic style. The labeling process was conducted in collaboration with art professionals to ensure artistic credibility. Moreover, to maintain consistency with upstream tasks and improve the generality of the label taxonomy, the style categories we adopt are aligned with the mainstream music-to-dance generation dataset FineDance [Li *et al.*, 2023], thereby placing our "dance–camera" style labels in the same semantic space as the "music–dance" style labels.

The final annotations cover a variety of styles including 'Korean', 'Choreography', 'Traditional-Chinese', 'Hiphop', 'Popping', 'Locking', 'Urban' and 'Jazz'. The data distribution reflects the characteristics of the original DCM data, which originates from MikuMikuDance (MMD) community creations. This augmented DCM dataset lays the foundation for training and evaluating style-consistent camera movement synthesis models. (Dataset details in Appendix A)

3.2 Style-Based Dance Camera Synthesis Model

In professional animation production, animators follow a hierarchical procedure: they first identify keyframes on the timeline based on the music and dance, then design the camera movements at those keyframes, and finally smoothly interpolate the camera trajectory between keyframes. Thanks to the DCM dataset, which already provides annotated keyframe information, our model is designed based on this hierarchical

animator workflow. The $\mathcal{DSC}$ model consists of two parts: (a) **Camera Keyframe Detection** and (b) **Camera Synthesis**. The model architecture as shown in Figure 2 b.

Camera Keyframe Detection ($\mathcal{CKD}$). The objective of the camera keyframe detection stage is to predict a temporal keyframe position sequence $k = \{k_1, k_2, \ldots, k_T\}$ from the given music features m, dance sequence d, and the target style label s, where $k_t \in \{0, 1\}$ indicates whether the t-th frame is a keyframe.

Formally, we formulate this as a conditional probability prediction problem. At each time step t, the model predicts the probability of each frame in a historical–future window being a keyframe, conditioned on the context within that window. Specifically, we first map the discrete style label $\mathcal{S}$ to a style embedding vector $\mathbf{e}_s \in \mathbb{R}^{d_s}$ via a learnable style encoder. For the current frame t, we extract a sliding window of length $L = h + w$, comprising h frames of history and w frames of the future (including the current frame), obtaining subsequences of music features $\mathbf{m}_{t-h:t+w-1}$, dance motions $\mathbf{d}_{t-h:t+w-1}$, and the corresponding keyframe label subsequence $\mathbf{k}_{t-h:t+w-1}$.

The style label s is first encoded into a style embedding vector $\mathbf{e}_s$ via the style encoder. To condition the sequential model, $\mathbf{e}_s$ is broadcast (replicated) to match the sequence length L, denoted as $\mathbf{e}_s \oplus \mathbf{1}_L$. The model then encodes the concatenated features $[\mathbf{m}_{t-h:t+w-1}; \mathbf{d}_{t-h:t+w-1}; \mathbf{e}_s \oplus \mathbf{1}_L]$ through a Transformer encoder. The encoded features are fed into a classification head to predict the probability of each frame in the future w frames being a keyframe:

$$\hat{k}_{t+i} = P(k_{t+i} = 1 \mid \mathbf{m}_{t-h:t+w-1}, \mathbf{d}_{t-h:t+w-1}, \mathbf{e}_s \oplus \mathbf{1}_L), \tag{1}$$

where $i \in [0, w-1]$. A binary keyframe label is obtained by applying a threshold. The incorporation of the style embedding $\mathbf{e}_s$ enables the model to perceive and learn the shot-switching rhythms and patterns characteristic of different artistic styles. The entire stage can be summarized as a conditional mapping:

$$k = \mathcal{CKD}(m, d, s) = \{\arg\max P(k_t \mid m, d, s)\}_{t=1}^{T}. \tag{2}$$

The model is trained with a weighted binary cross-entropy loss to handle the class imbalance between keyframes and non-keyframes:

$$\mathcal{L}_{\text{kf}} = -\frac{1}{w}\sum_{i=0}^{w-1}\left[\lambda \cdot k_{t+i}\log(\hat{k}_{t+i}) + (1 - k_{t+i})\log(1 - \hat{k}_{t+i})\right], \tag{3}$$

where k_{t+i} is the ground-truth label, and $\lambda > 1$ is the weight assigned to the keyframe class from datasets.

Camera Synthesis ($\mathcal{CS}$). Given the keyframe sequence k, this stage synthesizes the complete camera movement sequence c that adheres to the target style s. The process is performed iteratively: for each pair of adjacent keyframe positions t_j and t_{j+1}, the model first synthesizes the camera parameters at these keyframes, $\hat{c}_{t_j}$ and $\hat{c}_{t_{j+1}}$, based on the history of camera movement, the current music–dance context, and the style embedding $\mathbf{e}_s$. Subsequently, instead of directly predicting the parameters of every frame between keyframes, the model predicts a style-dependent, monotonically increasing tween function $\rho(t; s) \in [0, 1]$. The camera movement for all non-keyframes is then obtained via linear interpolation:

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

This design ensures smooth motion within each shot while allowing different styles to exhibit distinct velocity profiles. The entire synthesis is performed by a unified, style-conditioned generative network. Formally, the stage can be summarized as a conditional mapping:

$$c = \mathcal{CS}(m, d, s, k), \tag{5}$$

where the keyframe sequence k is obtained from the preceding detection stage.

Synthesis Model Training. The model is optimized with a composite loss function that consists of reconstruction and kinematic terms. The overall synthesis loss is defined as:

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

where $\mathcal{L}_{\text{rec}}$ is the frame-wise reconstruction loss, $\mathcal{L}_{\text{vel}}$ and $\mathcal{L}_{\text{acc}}$ are velocity and acceleration smoothness losses, and $\mathcal{L}_{\text{ba}}$ is a body-attention loss that encourages the camera to focus on relevant body parts. In addition, the training of the $\mathcal{CS}$ model is also trained with the style-consistent adversarial loss $\mathcal{L}_{\text{ScAT}}$ (defined in Section 3.3). The full training objective is:

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{ScAT}}. \tag{7}$$

Inference. At inference time, the complete camera movement sequence for a given music clip m, dance sequence d, and style label s is generated in two steps:

$$c = \mathcal{DSC}(m, d, s) = \mathcal{CS}\big(m, d, s, \mathcal{CKD}(m, d, s)\big). \tag{8}$$

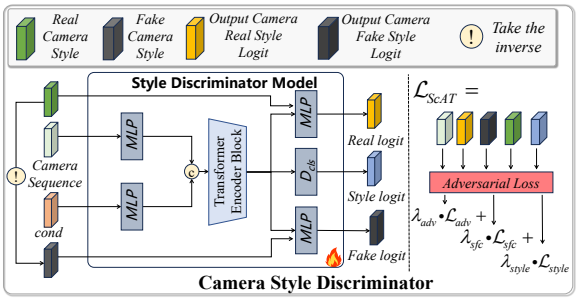

Figure 3: Overview of the style discriminator model in ScAT.

3.3 Style-Consistent Adversarial Training

In camera movement synthesis, style serves as an explicit, high-level artistic condition, and its effective incorporation is crucial for the visual expressiveness of the generated results. To further strengthen the model's understanding and generative capability of style features, we introduce a style-consistent adversarial training mechanism.

We design a **camera style discriminator model** $\mathcal{D}$. The model architecture is illustrated in Figure 3. The $\mathcal{D}$ takes the camera sequence, the fused music-dance conditional features,

and the style label as inputs. Its core consists of a lightweight Transformer encoder and multi-layer MLPs. The adversarial training first introduces the **Style-Focusing Contrastive (SFC) Loss** $\mathcal{L}_{sfc}$ to enhance the precision of style consistency and prevent confusion between different styles. Specifically, for the same sample synthesized by the generator $\mathcal{DSC}$, the discriminator simultaneously receives the correct style label s and a randomly sampled incorrect style label $\tilde{s}$. The goal of the generator is to maximize the authenticity score of the sample when conditioned on the correct label, while minimizing its score when conditioned on the incorrect label. Inspired by Wu et al. [Wu *et al.*, 2022], we design a loss function based on WGAN [Gulrajani *et al.*, 2017]:

$$\mathcal{L}_{sfc} = -\mathbb{E}[\mathcal{D}(\mathcal{DSC}(m, d, s) \mid s)] + \mathbb{E}[\mathcal{D}(\mathcal{DSC}(m, d, s) \mid \tilde{s})]. \tag{9}$$

Next is the **Style Classification Loss** $\mathcal{L}_{style}$. In parallel to the contrastive objective, the $\mathcal{D}$ also includes a style classification head $\mathcal{D}_{cls}$. We employ a cross-entropy loss with label smoothing to train this classification head, which in turn forces the generator to produce motions highly consistent with the target style. Its formula is as follows:

$$\mathcal{L}_{style} = \text{CE}(\mathcal{D}_{cls}(\mathcal{DSC}(m, d, s), s), \alpha), \tag{10}$$

where α is the label smoothing factor. Finally, we employ a **Basic Adversarial Loss** $\mathcal{L}_{adv}$ to drive the generator to output distributions that closely approximate the real data. Also based on the WGAN [Gulrajani *et al.*, 2017], this loss is:

$$\mathcal{L}_{adv} = -\mathbb{E}[\mathcal{D}(\mathcal{DSC}(m, d, s))]. \tag{11}$$

In summary, the total loss for style-consistent adversarial training is the weighted sum of the above components:

$$\mathcal{L}_{\text{ScAT}} = \lambda_{sfc}\mathcal{L}_{sfc} + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv}. \tag{12}$$

Discriminator Training. The discriminator $\mathcal{D}$ is optimized to accurately distinguish real camera sequences from generated ones while correctly classifying their style. For each training step, real camera sequences $\mathbf{c}_{\text{real}}$ and synthesized sequences $\mathbf{c}_{\text{fake}} = \mathcal{DSC}(m, d, s)$ are passed through $\mathcal{D}$ along with the corresponding fused condition (music and dance features) and the ground-truth style label s. The basic adversarial loss for the $\mathcal{D}$ is computed based on WGAN-GP and WGAN loss [Gulrajani *et al.*, 2017]:

$$\mathcal{L}_D^{\text{adv}} = \underbrace{\mathcal{D}(\mathbf{c}_{\text{real}}, \mathbf{h}_{\text{cond}}, s)}_{\text{real as positive}} + \underbrace{\max(0, 1 - \mathcal{D}(\mathbf{c}_{\text{fake}}, \mathbf{h}_{\text{cond}}, s))}_{\text{fake as negative}}, \tag{13}$$

where $\mathbf{h}_{\text{cond}}$ denotes the fused music-dance conditional features. To enforce Lipschitz continuity and stabilize training, a gradient penalty term is added. For each mini-batch, an interpolated sample is constructed as $\tilde{\mathbf{c}} = \epsilon\mathbf{c}_{\text{real}} + (1 - \epsilon)\mathbf{c}_{\text{fake}}$, with $\epsilon \sim U(0, 1)$. The gradient penalty is computed as:

$$\mathcal{L}_{\text{gp}} = \gamma \cdot \mathbb{E}_{\tilde{\mathbf{c}}}\left[(\|\nabla_{\tilde{\mathbf{c}}}\mathcal{D}(\tilde{\mathbf{c}}, \mathbf{h}_{\text{cond}}, s)\|_2 - 1)^2\right], \tag{14}$$

where γ is the penalty coefficient. The total loss for the $\mathcal{D}$ is thus:

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\mathcal{D}}^{\text{adv}} + \mathcal{L}_{\text{gp}}. \tag{15}$$

The $\mathcal{D}$ is updated by minimizing $\mathcal{L}_{\mathcal{D}}$, while the generator is optimized using the combined adversarial, style, and SFC losses as defined in $\mathcal{L}_{\text{ScAT}}$. This alternating training scheme enhances both the realism and style fidelity of the synthesized camera movements.

4 Experiments

4.1 Experiment Setup

Datasets. In this work, we employ the DCM dataset [Wang *et al.*, 2024a], which we have augmented with manual style annotations. The dataset comprises 108 pieces of animator-designed, paired data, including dance sequences, camera movements, music tracks, keyframes, and our newly added style labels. To ensure a fair comparison, we retain the original training and testing splits as defined in prior work.

Dataset Process. For music feature extraction, we employ Librosa [McFee *et al.*, 2015] to compute a 2D feature map $\mathbf{m} \in \mathbb{R}^{T\times 35}$, where T is the total number of frames (details in Appendix B). For dance and camera movements, we adopt the approach of DanceCamera3D [Wang *et al.*, 2024a]. The dance is represented by the global positions of 60 human body joints, denoted as $\mathbf{d}_i \in \mathbb{R}^{T\times 180}$. The camera is represented in the MMD format using polar coordinates, denoted as $\mathbf{c}_i \in \mathbb{R}^{T\times 8}$, which includes the global position of the reference point, the camera's rotation and distance relative to it, and the camera's Field of View. The style label is converted into a numerical encoding and is expanded to be temporally aligned with the other features, represented as $\mathbf{s} \in \mathbb{R}^{T\times 1}$.

Implementation Details. The $\mathcal{DSC}$ model comprises a total of 104M parameters. Its components include the Camera Keyframe Detection ($\mathcal{CKD}$) model with 39M, the Camera Synthesis ($\mathcal{CS}$) model with 64M parameters and the Discriminator ($\mathcal{D}$) with 1M parameters. All models were trained for one day using five NVIDIA GeForce RTX 3090 GPUs. For training the $\mathcal{CKD}$, we set a per-GPU batch size of 256, learning rate of 2×10^{-4}, and trained for 3000 epochs with the Adan optimizer [Xie *et al.*, 2024]. For the $\mathcal{CS}$ training, the per-GPU batch size was 256 and the learning rate was 2×10^{-4}. The loss weights were set as follows: $\lambda_{\text{rec}} = 2$, $\lambda_{\text{vel}} = 5$, $\lambda_{\text{acc}} = 5$, $\lambda_{\text{ba}} = 0.0075$, $\lambda_{\text{sfc}} = 4 \times 10^{-4}$, $\lambda_{\text{style}} = 5 \times 10^{-5}$, and $\lambda_{\text{adv}} = 5 \times 10^{-5}$. The label smoothing factor α was set to 0.015. The $\mathcal{CS}$ was trained for 1200 epochs using the Adan optimizer. To stabilize the adversarial training, we employed a dynamic training strategy for the $\mathcal{D}$. The training was divided at epoch 300. For the first phase, the $\mathcal{D}$'s learning rate was set to 1×10^{-5} and the gradient penalty weight γ was 0.5. In the second phase, these were adjusted to 5×10^{-6} and 0.8, respectively. The $\mathcal{D}$ was optimized using the Adam optimizer [Kingma and Ba, 2014].

Evaluation Metrics. We evaluate the generated results following prior work [Wang *et al.*, 2024a]. This includes **quantitative metrics** for quality, diversity, and dancer fidelity: Fréchet Inception Distance ($\text{FID}_{\text{k/s}}$), average feature distance ($\text{Dist}_{\text{k/s}}$), Dancer Missing Rate (DMR), and Limbs Capture Difference (LCD). For **qualitative assessment**, we provide visualization results. For **user study**, we randomly select 10 test dance sequences and generate corresponding camera movements using our method and all baselines, creating 30 comparison video pairs. These are presented in random order to 20 participants, comprising both professional art practitioners and individuals with minimal art experience, who select which camera movement better highlights the dance and music in each pair. To specifically evaluate the **Style Accu-**

Method	Quality		Diversity		Dancer Fidelity		User Study
	$FID_k \downarrow$	$FID_s \downarrow$	$Dist_k \uparrow$	$Dist_s \uparrow$	$DMR \downarrow$	$LCD \downarrow$	DSC WinRate ↑
GT	-	-	3.275	1.731	0.00142	-	42.29 ± 2.33 %
DanceRevolution [Huang *et al.*, 2020]	10.267	2.368	1.491	1.118	0.0062	0.154	90.12 ± 0.25 %
FACT [Li *et al.*, 2021]	5.205	0.960	1.505	1.007	0.0070	0.151	88.64 ± 1.44 %
DanceCamera3D [Wang *et al.*, 2024a]	3.749	0.280	1.631	1.326	0.0025	0.147	77.88 ± 2.23 %
DanceCamera3D* [Wang *et al.*, 2024a]	7.864	0.313	0.861	1.301	0.0047	0.151	-
DanceCamAnimator [Wang *et al.*, 2024b]	3.453	0.268	3.140	1.293	0.0022	0.152	69.14 ± 0.88 %
TemMEGA [Xu *et al.*, 2025]	3.237	0.255	1.961	1.347	0.0020	0.141	-
DanceStyleCam	**2.751**	**0.254**	**3.201**	**1.380**	**0.0016**	**0.138**	-

Table 1: Comparisons with SOTA methods on the DCM dataset.

Method	Style Accuracy			Style Consistency		
	Top-1 % ↑	Top-3 % ↑	Top-5 % ↑	Mean CosSim ↑	Max CosSim ↑	Min CosSim ↑
FACT [Li *et al.*, 2021]	27.34	64.18	80.22	0.3992	0.6923	-0.528
DanceCamera3D [Wang *et al.*, 2024a]	31.78	69.87	88.32	0.4927	0.8012	-0.382
DanceCameraAnimator [Wang *et al.*, 2024b]	35.66	76.74	90.7	0.5658	0.8711	-0.169
DanceStyleCam	**48.84**	**79.07**	**95.37**	**0.7565**	**0.9833**	**-0.0999**

Table 2: Comparisons of Style Consistency and Accuracy with SOTA methods on the DCM dataset.

racy and Consistency, we propose new consistency metrics. We construct a 30-dimensional prototype feature vector $\mathbf{s}_{\text{proto}}$ for each style from real data, comprising basic motion statistics (14-dim) and advanced descriptors (16-dim) (details in Appendix C). For each generated sample with feature $\mathbf{s}_{\text{gen}}$, we compute its cosine similarity to all style prototypes. The style consistency of the synthesized camera movements is assessed by the Top-K ranking accuracy of the target style and the distribution of target similarity scores.

4.2 Evaluation Results

Quantitative Results. We compare our model with state-of-the-art camera generation methods on the DCM dataset, reporting the quantitative results in Table 1. Our method achieves the best performance across all key metrics. The ScAT enables our model to significantly outperform existing approaches in terms of quality, diversity, and dancer fidelity. This demonstrates that explicitly modeling style through our proposed components not only enhances style consistency but also leads to comprehensive gains in synthesis quality.

To quantitatively evaluate the style accuracy of the synthesized camera movements, we extract the 30-dim style features from the samples synthesized by all compared methods and compute their Top-K classification accuracy and cosine similarity to the target style prototypes. The results presented in Table 2 verify the effectiveness of the introduced explicit style labels and ScAT. Our method achieves a significant advantage on all style accuracy metrics, indicating a high degree of conformity between the style features of the generated samples and the target prototypes. This demonstrates that the $\mathcal{D}$ is capable of performing both style classification and adversarial discrimination on both real and synthesized samples simultaneously. This dual objective collaboratively guides the generator not only to learn the real data distribution but also to precisely capture and replicate the unique motion patterns of different styles. This demonstrates that our method could transforms style from an implicit attribute into a learnable and consistently realizable dimension for synthesis.

Method	Quality		Diversity		Dancer Fidelity	
	$FID_k \downarrow$	$FID_s \downarrow$	$Dist_k \uparrow$	$Dist_s \uparrow$	DMR ↓	LCD ↓
w/o ScAT	3.457	0.301	3.221	1.291	0.0021	0.148
w/o $\mathcal{L}_{Adv}$	3.391	0.296	3.227	1.332	0.0020	0.145
w/o $\mathcal{L}_{style}$	2.819	0.256	3.199	1.357	**0.0012**	0.140
w/o $\mathcal{L}_{sfc}$	3.104	0.271	3.214	1.377	0.0016	0.144
Full	**2.751**	**0.254**	**3.201**	**1.380**	0.0016	**0.138**

Table 3: Ablation Study on ScAT Components for Synthesis Quality.

Qualitative Results. To comprehensively demonstrate the efficacy of our method, we visualize the samples alongside the rendered dance sequences for different style labels, as shown in Figure 4. Our method samples that distinctively embody their respective stylistic characteristics. For instance, the *Korean* is characterized by frequent shifts between long shots and close-ups, with relatively fewer abrupt jump cuts. In contrast, the *Choreography* exhibits a more dynamic and lively feel, featuring a higher frequency of jump cuts. The *Traditional-Chinese* predominantly features smooth, flowing camera movements with an increased emphasis on orbiting movements around the dancer. These visual results qualitatively validate that our method can accurately interpret diverse stylistic intents and realize style-consistent camera movement synthesis. More visual results and analysis of other styles can be found in Appendix D.

4.3 Ablation Studies

Synthesis Quality Perspective. From the perspective of synthesis quality, the ablation study clearly reveals the importance of the ScAT and its components, as shown in Table 3. When the entire ScAT is removed, all metrics show significant degradation. This confirms the fundamental role of ScAT in learning the real data distribution and generating di-

Method	Style Accuracy			Style Consistency		
	Top-1 % ↑	Top-3 % ↑	Top-5 % ↑	Mean CosSim ↑	Max CosSim ↑	Min CosSim ↑
w/o ScAT	38.26	76.74	91.21	0.5882	0.8739	-0.144
w/o $\mathcal{L}_{Adv}$	47.12	76.74	92.19	0.7125	0.9639	-0.147
w/o $\mathcal{L}_{style}$	44.22	72.11	95.37	0.6343	0.9282	-0.156
w/o $\mathcal{L}_{sfc}$	40.28	70.28	92.19	0.6121	0.8911	-0.167
Full	**48.84**	**79.07**	**95.37**	**0.7565**	**0.9833**	**-0.0999**

Table 4: Ablation Study on ScAT Components for Style Consistency and Accuracy

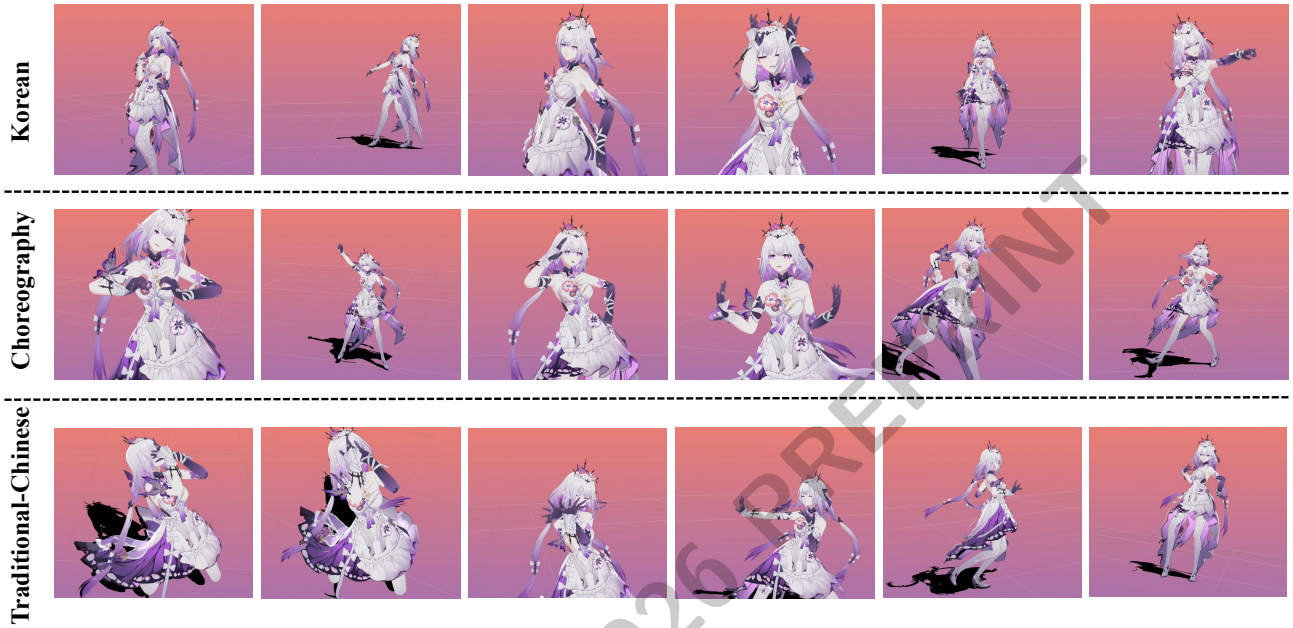

Figure 4: Visualization of the synthesized camera movements using DanceStyleCam. We showcase results for the three most prevalent style categories in dataset. The floor grid provides a spatial reference to validate the distinct camera movement patterns characteristic of each style.

verse and stable motions. Specifically, the absence of $\mathcal{L}_{Adv}$ leads to a marked deterioration in synthesis quality, which confirms the key contribution of adversarial training. In contrast, removing the style-related losses has a relatively limited impact on synthesis quality, suggesting that these losses primarily focus on style modeling without substantially interfering with the base motion generation pipeline. The full model achieves optimal performance across all metrics, demonstrating the effectiveness of the collaborative work of ScAT components in comprehensively improving synthesis quality.

Style Adherence Perspective. From the perspective of style accuracy, the ablation study further verifies the core role of the ScAT in achieving precise style aware, as shown in Table 4. Completely removing ScAT leads to a substantial decline in all style metrics, proving the effectiveness of the framework in achieving style alignment. Removing only the $\mathcal{L}_{Adv}$ has a relatively mild impact on style metrics, indicating that adversarial training itself contributes limitedly to style fidelity, with its primary role being to enhance generation realism. However, removing the style-related losses (without $\mathcal{L}_{style}$ and $\mathcal{L}_{sfc}$) significantly impairs style consistency, demonstrating the effectiveness of $\mathcal{L}_{style}$ and $\mathcal{L}_{sfc}$ in guiding the model to learn and replicate the specific motion patterns of each style. In particular, removing $\mathcal{L}_{sfc}$ causes a more pronounced drop in style classification metrics than removing $\mathcal{L}_{style}$, indicating that our dual adversarial loss mechanism is more effective at ensuring style specificity.

5 Conclusion

This paper presents DanceStyleCam (DSC), a novel framework for synthesizing style-consistent dance camera movement. We address a key gap in prior work by introducing camera movement style as an explicit conditioning signal, enabling style-aware synthesis beyond mere content generation. Our approach is based on a manually annotated paired dataset of styles and camera movement, and employs a two-stage camera movement generator guided by Style-consistent Adversarial Training. This mechanism enables the generator to precisely learn the unique motion patterns of different styles. Quantitative experiments and user studies on DCM dataset demonstrate that DSC achieves state-of-the-art performance in terms of synthesis quality, diversity, and dancer fidelity, while significantly outperforming existing methods in style consistency. These results validate the effectiveness of our framework and highlight the importance of explicit style modeling for advancing dance-driven camera synthesis.

Contribution Statement

Xiaoying Huang and Sanyi Zhang contributed equally to this work. Xiaoying Huang led the method design, experimental implementation, result analysis, and manuscript writing. Sanyi Zhang contributed to manuscript revision and research discussion. Xirui Wang contributed to method discussion and data preparation. Qin Zhang contributed to funding acquisition. Long Ye provided research direction, contributed to funding acquisition, and served as the corresponding author.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (Grant No. 62571500 and Grant No. 62271455), in part by the Beijing Natural Science Foundation (No. L252143), in part by the Fundamental Research Funds for the Central Universities (CUCZD2502), in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2025C03), and in part by Public Computing Cloud, CUC.

References

[Alexanderson *et al.*, 2023] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023.

[Bares *et al.*, 2000] William Bares, Scott McDermott, Christina Boudreaux, and Somying Thainimit. Virtual 3d camera composition from frame constraints. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 177–186, 2000.

[Christie *et al.*, 2008] Marc Christie, Patrick Olivier, and Jean-Marie Normand. Camera control in computer graphics. In *Computer graphics forum*, volume 27, 8, pages 2197–2218. Wiley Online Library, 2008.

[Dhariwal *et al.*, 2020] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[Evin *et al.*, 2022] Inan Evin, Perttu Hämäläinen, and Christian Guckelsberger. 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, 2022.

[Gschwindt *et al.*, 2019] Mirko Gschwindt, Efe Camci, Rogerio Bonatti, Wenshan Wang, Erdal Kayacan, and Sebastian Scherer. Can a robot become a movie director? learning artistic principles for aerial cinematography. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1107–1114. IEEE, 2019.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[Huang *et al.*, 2019a] Chong Huang, Chuan-En Lin, Zhenyu Yang, Yan Kong, Peng Chen, Xin Yang, and Kwang-Ting Cheng. Learning to film from professional human motion videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4244–4253, 2019.

[Huang *et al.*, 2019b] Chong Huang, Zhenyu Yang, Yan Kong, Peng Chen, Xin Yang, and Kwang-Ting Tim Cheng. Learning to capture a film-look video with a camera drone. In *Proceedings of the 2019 international conference on robotics and automation (ICRA)*, pages 1871–1877. IEEE, 2019.

[Huang *et al.*, 2020] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *International conference on learning representations*, 2020.

[Huang *et al.*, 2021] Chong Huang, Yuanjie Dang, Peng Chen, Xin Yang, and Kwang-Ting Cheng. One-shot imitation drone filming of human motion videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5335–5348, 2021.

[Jiang *et al.*, 2020] Hongda Jiang, Bin Wang, Xi Wang, Marc Christie, and Baoquan Chen. Example-driven virtual cinematography by learning camera behaviors. *ACM Trans. Graph.*, 39(4):45, 2020.

[Jiang *et al.*, 2024a] Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. In *Proceedings of the Computer Graphics Forum*, volume 43,2, page e15055. Wiley Online Library, 2024.

[Jiang *et al.*, 2024b] Xuekun Jiang, Anyi Rao, Jingbo Wang, Dahua Lin, and Bo Dai. Cinematic behavior transfer via nerf-based differentiable filming. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6723–6732, 2024.

[Kim *et al.*, 2022] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3490–3500, 2022.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

[Li and Cheng, 2008] Tsai-Yen Li and Chung-Chiang Cheng. Real-time camera planning for navigation in virtual environments. In *Proceedings of the International Symposium on Smart Graphics*, pages 118–129. Springer, 2008.

[Li *et al.*, 2021] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned

3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021.

[Li *et al.*, 2023] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023.

[Li *et al.*, 2024a] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1524–1534, 2024.

[Li *et al.*, 2024b] Sifei Li, Weiming Dong, Yuxin Zhang, Fan Tang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu. Dance-to-music generation with encoder-based textual inversion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

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

[Pueyo *et al.*, 2024] Pablo Pueyo, Juan Dendarieta, Eduardo Montijano, Ana Cristina Murillo, and Mac Schwager. Cinempc: A fully autonomous drone cinematography system incorporating zoom, focus, pose, and scene composition. *IEEE Transactions on Robotics*, 40:1740–1757, 2024.

[Rao *et al.*, 2020] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In *Proceedings of the European Conference on Computer Vision*, pages 17–34. Springer, 2020.

[Rao *et al.*, 2023] Anyi Rao, Xuekun Jiang, Yuwei Guo, Linning Xu, Lei Yang, Libiao Jin, Dahua Lin, and Bo Dai. Dynamic storyboard generation in an engine-based virtual environment for video production. In *ACM SIGGRAPH 2023 Posters*, pages 1–2. ACM, 2023.

[Rucks and Katzakis, 2021] James Rucks and Nikolaos Katzakis. Camerai: Chase camera in a dense environment using a proximal policy optimization-trained neural network. In *Proceedings of the 2021 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2021.

[Siyao *et al.*, 2022] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022.

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

[Wang *et al.*, 2024a] Zixuan Wang, Jia Jia, Shikun Sun, Haozhe Wu, Rong Han, Zhenyu Li, Di Tang, Jiaqing Zhou, and Jiebo Luo. Dancecamera3d: 3d camera movement synthesis with music and dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2024.

[Wang *et al.*, 2024b] Zixuan Wang, Jiayi Li, Xiaoyu Qin, Shikun Sun, Songtao Zhou, Jia Jia, and Jiebo Luo. Dancecamanimator: Keyframe-based controllable 3d dance camera synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10200–10209, 2024.

[Wu *et al.*, 2022] Shuang Wu, Shijian Lu, and Li Cheng. Music-to-dance generation with optimal transport. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4988–4994. International Joint Conferences on Artificial Intelligence Organization, 7 2022. AI and Arts.

[Wu *et al.*, 2023] Xinyi Wu, Haohong Wang, and Aggelos K Katsaggelos. The secret of immersion: actor driven camera movement generation for auto-cinematography. *arXiv preprint arXiv:2303.17041*, 4, 2023.

[Xie *et al.*, 2023] Chun Xie, Isao Hemmi, Hidehiko Shishido, and Itaru Kitahara. Camera motion generation method based on performer's position for performance filming. In *Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE)*, pages 957–960. IEEE, 2023.

[Xie *et al.*, 2024] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024.

[Xu *et al.*, 2025] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Smooth and flexible camera movement synthesis via temporal masked generative modeling. In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Yang *et al.*, 2023] Sizhe Yang, Yanjie Ze, and Huazhe Xu. Movie: Visual model-based policy adaptation for view generalization. *Advances in Neural Information Processing Systems*, 36:21507–21523, 2023.