Understanding Cinematography Styles with Generative Modeling of Camera Behaviors

Jinhe Na[1]*, Pengjie Wang[1]*, Rui He[3], Yingjie Xi[2], Boyuan Cheng[2], Gui Jing[1], Xiaosong Yang[2], Ying Cao[4]†

[1]Dalian Minzu University, China
[2]Bournemouth University, UK
[3]The Hong Kong Polytechnic University, China
[4]ShanghaiTech University, China

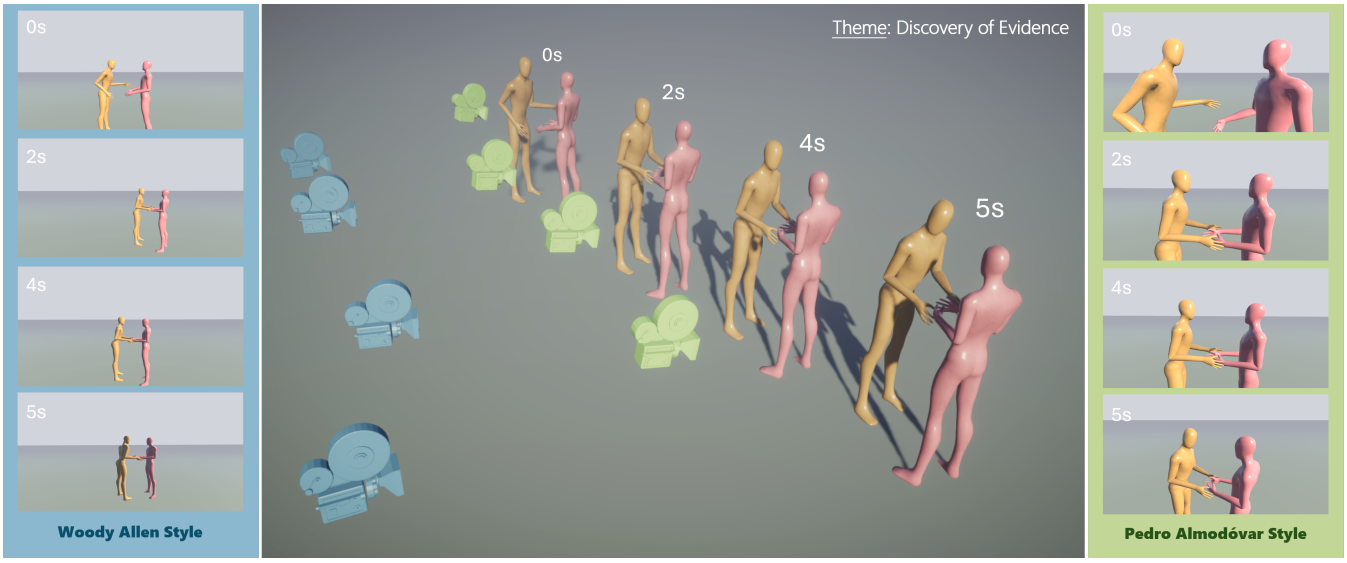


Figure 1: We propose a generative model that has the ability to mimic camera motion styles of different film directors. Given a 3D scene of two interacting characters as well as a scene theme (i.e., "discovery of evidence" in this example), our model generates two distinct camera trajectories—sequences of camera poses in 3D space—in the styles of Woody Allen and Pedro Almodóvar (highlighted in blue and green, respectively). 3D camera poses at four moments along each camera trajectory, alongside the corresponding 2D frames, are sampled and visualized. The model allows us to systematically and quantitatively compare and understand the cinematography styles of different directors in terms of shot type, shot composition and camera movement.

Abstract

In filmmaking, cinematography captures a story visually with shot sequences by placing and moving the camera. Professional filmmakers often develop their own unique shot design styles that deviate from conventional cinematographic rules to create a distinct look and feel for the films. However, understanding these styles can be challenging for film practitioners. In this paper, we represent a method to facilitate understanding of cinematography styles of different film directors by exploiting a generative model of stylized camera motions. Our generative model is able to synthesize camera motions in the style of a specified director given a 3D animation of characters and a scene theme. To train the model, we curate a new dataset of movie shots by various film directors, with extracted character and camera motions for each shot. Our evaluation demonstrates that our model outperforms existing alternatives in producing high-quality camera motions and replicating a director's style. Based on our trained model, we analyze the features that shape the unique styles of multiple film directors in placing and moving the camera across different film genres.

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

* Joint first authors. † Corresponding author.

1. Introduction

In filmmaking, cinematography is a challenging task that requires the filmmaker to portray the film's story events through carefully crafted visual representations, i.e., a sequence of shots. A shot is series of frames captured by a camera in a continuous, uninterrupted recording, encapsulating an ongoing event with a unique combination of camera placement and movement (e.g., pan, tilt, push in, pull out and tracking). To design the shots that effectively draw the audience's attention, evoke emotional reactions, and convey the story, filmmakers typically follow a set of conventional cinematography guidelines, such as "rule of third" for shot composition. In practice, filmmakers may extend the common cinematography rules with their stylistic variations, telling a story through unique approaches to composing a shot and moving the camera, which we refer to as *cinematography styles*. These styles contributes to the distinct look and feel of the filmmakers' movies. For example, Wes Anderson breaks the rule of thirds very often, and instead prefers to use center framing (i.e., put the main subject in the center). Furthermore, Wes Anderson frequently trucks the camera right to follow a character to create a storybook-like world of the film. These has been recognized as his trademark cinematographic techniques.

These cinematography styles can be well received by audiences and it is a common lesson for students to study them in a film school. However, it is rather difficult for film practitioners, either beginners or experts, to fully capture the essence of the styles and apply them effectively in practice. Common practice for investigating cinematographic characteristics of a particular filmmaker is to watch a large quantity of existing film clips and manually summarize observed patterns. This process, however, is quite tedious and time-consuming. More importantly, this approach prevents straightforward comparison of cinematographic features between filmmakers, since it is hard to factor out the effect of scene contents that vary greatly across the films of different filmmakers.

In view of the above problem, we aim to to computationally replicate camera motion styles of different filmmakers and quantitatively understand their cinematography styles in terms of how the camera is placed and moved. Studying what features a filmmaker's cinematography style offers several notable benefits: first, it provides useful insights for film practitioners, allowing them to learn and draw inspiration from these practically successful cinematography techniques; second, it can inform the development of computational models that generate stylistic camera motions, which will find a wide spectrum of applications in filmmaking, animation and gaming.

In this paper, we design a diffusion-based generative model of 3D camera motions and condition the model on a 3D animation of characters, a scene theme and a director label. As shown in Figure 1, our model can be controlled to generate camera pose sequences following different director styles, given the same input character animation and scene theme. Therefore, by examining synthetic samples from our model, it is possible to answer the question of how different directors make cinematographic design decisions under the same context. For training our model, we construct a new dataset of 3D character animations paired with camera motions, which are extracted from meticulously selected film clips by a range of well-known film directors. Our experiments show the superiority of our model over existing methods in sample quality and style adherence. We further show that our model can enable a novel *generative* paradigm for cinematography style study, by conducting a model-based study, where we analyze camera motions generated by our trained model, in terms of fundamental cinematography elements (shot type, shot composition and camera movement), to shed light on what distinguishes between the styles of different directors in placing and moving the camera. Our data and code are available at: https://github.com/NaJinhe/cam_style

Our paper makes the following contributions:

- **Generative model.** We propose a generative model with strong capabilities to replicate the camera motion styles of different film directors under various thematic contexts.
- **Dataset.** We introduce a new dataset, encompassing diverse 3D character animations along with stylistic camera motions, which are constructed from real shots taken from movies of various well-known film directors.
- **Cinematography style study methodology.** We introduce a novel method that harnesses the generative capabilities and controllability of our proposed model to analyze and understand the distinctive ways that different filmmakers place and move the camera.

2. Related Work

2.1. Cinematography in Filmmaking

Filmmaking follows a conventional workflow that includes scriptwriting, storyboarding, shot recording, and shot sequence editing [Ari91]. Within this workflow, the storyboarding stage involves planning a sequence of shots that compose the visual representation of each critical event in the film script [Gle12]. These shots are created through cinematography, which is essentially the art of capturing a story visually with a camera to create motion pictures we see on screen. Bowen's *Grammar of the Shot* [Bow13] summarizes some general rules for cinematography. However, these general rules can hardly encompass the extensive design space of cinematography. In practice, filmmakers would design many stylized variations within their specific working contexts to express intended emotions and convey the story event, which shape their own styles of cinematography.

To draw inspiration from well-established cinematography styles, film practitioners can refer to numerous style exemplars. This process, known as film analysis, involves a detailed shot-by-shot examination of a wide array of film clips, referred to as iconic analysis [AoTPB*92, SS87]. However, such film analysis process is usually expensive, and prevents an accurate comparison between the cinematographic techniques of different filmmakers, primarily due to variations in scene content across the filmmakers' movies. Some prior studies, such as [RS24], automate analysis on the cinematographic styles of film directors based on *existing* film examples, which avoids manual examination, but still suffers from the scene content variation issue. In contrast, our proposed generative model enables a new cinematography style study method that can address both issues: it compares statistics for important elements of cinematography that are *automatically* computed from *generated* camera motions of different styles on the *same* set of scenes.

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

2.2. Camera Motion Control

Automatic control of camera motion through real or virtual 3D environments is a widely studied area, with a rich history of investigation. Many of earlier works exploit path or motion planning techniques for creating camera trajectories in 3D space, while considering some important factors such as visibility, collision avoidance and smoothness [DZ94, OSTG09, LCL*10]. Optimization-based approaches have also been explored [GCLR15, HLH*16], which search the space of camera parameters (e.g., position, orientation, velocity) for solutions that enforce certain prior such as temporal coherence of camera poses and satisfy specific constraints such as projecting objects of interest at desired on-screen positions. Several works take data-driven approaches to extract camera motions from video examples and re-apply them to new scenes [CRE*14, SDM*14]. Cine-AI [EHG22] attempts to imitate a film director's cinematography style when placing and moving the camera in a virtual environment, but does not model the correlation between scene contents and camera motions.

Recent years have seen a rising interest in learning to control camera behaviors from real video footages, which captures the correlation between camera motions and scene contents from data directly. For example, several attempts have been made to imitate professional filming via camera drones by learning from single-subject videos captured by experts [HYK*19, HLY*19]. A deep reinforcement learning framework is proposed in [GCB*19], which drives the motion of a camera drone to select aesthetic viewpoints without human guidance. A recent work learns to transfer camera motion styles in reference 2D film clips to a 3D character animation [JWW*20] . A later approach extends such camera style transfer with additional low-level control over generated camera motions by introducing camera keyframes as constraints [JCW*21].

Motivated by the success of recent deep generative models in language and image generation, several method build generative models specialized in learning complex and diverse camera behaviors. A GAN-based framework is proposed in [WWK23], which synthesizes camera motion trajectories given input character motions and emotional states. A Transformer-based model is developed by [HWC24] to generate camera poses autoregressively given an input 3D scene of moving characters. A text-to-camera-motion diffusion model is proposed in [JWC*24], which translates camera shooting scripts, along with optional camera keyframes, into specific camera motions. DanceCamera3D [WJS*24] is a diffusion model trained to predict camera motion trajectories conditioned on dance motions and music audios. Recently, a diffusion model is proposed in [CXH*26] to jointly generate character and camera motions by explicitly learn pairwise interactions among the characters and camera.

Different from the prior methods, our camera motion generative model is scene-conditioned and style-aware, focusing on faithfully reproducing distinctive camera motion styles of various filmmakers under different scene contexts.

3. Our Model

We aim to build a generative models of camera motions $\mathbf{c}$, which is able to replicate the camera motion styles of different filmmakers, in order to facilitate the understanding of their cinematography styles. For this, our model is designed to process conditioning inputs including a 3D animation of characters $\mathbf{m}$, a scene theme y_h, and a director label y_d, so that it is possible to vary the director label while keeping the other inputs fixed to synthesize camera behaviors of different styles under the same context. Formally, our model learns a conditional distribution $p(\mathbf{c}|\mathbf{m}, y_h, y_d)$ from training data. We approach our learning problem with diffusion denoising probabilistic models [HJA20], in view of their strong performance in learning the distributions of complex and diverse data.

We assume that an input animation of N frames consists of one or two characters, and each character's motion is defined as a sequence of 3D poses, each being represented as a set of 3D joint locations. Then, the animation is formulated as $\mathbf{m} = (m^1, m^2, \ldots, m^N)$, where m^i is the concatenation of two distinct 3D poses at timestep i for two-character scenarios. When there is only a single character, m_i simply combines two duplicated, identical 3D poses. Following [JWW*20, HWC24, CXH*26], we define a camera motion as a sequence of camera poses $\mathbf{c} = (c^1, c^2, \ldots, c^N)$ in Toric space [LC15]. As shown in Figure 2, given two target characters a and b, a camera pose in Toric space is written as:

$$(x_a, y_a, x_b, y_b, \theta, \phi), \tag{1}$$

where (x_a, y_a) and (x_b, y_b) are the on-screen head positions of the two characters, and θ and ϕ represent the horizontal and vertical angles around the two characters, respectively. To apply the Toric representation to single-character scenarios, we set (x_a, y_a) and (x_b, y_b) to the on-screen positions of the character's left and right shoulders.

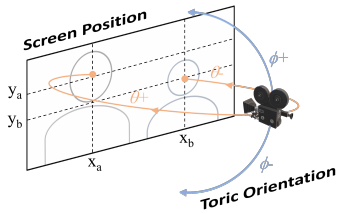


Figure 2: Toric camera pose representation.

3.1. Model Architecture and Training

Our conditional diffusion model defines a forward diffusion process that gradually add Gaussian noise to a camera motion sample $\mathbf{c}_0$ from the training dataset along a Markov chain of T steps, producing a sequence of noisy samples $(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T)$. The forward process is reversed by iteratively denoising pure Gaussian noise $\mathbf{c}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to recover the original sample $\mathbf{c}_0$ based on learned condition distributions $p_\varphi(\mathbf{c}_{t-1}|\mathbf{c}_t, \mathbf{u})$, where $\mathbf{u} = (\mathbf{m}, y_h, y_d)$ denotes the conditional inputs. $p_\varphi(\mathbf{c}_{t-1}|\mathbf{c}_t, \mathbf{u})$ is parameterized by optimizing a denoising network $f_\varphi(\mathbf{c}_t, t, \mathbf{u})$. f_φ takes as input a noisy camera motion sample $\mathbf{c}_t$, a timestep t, and the conditions $\mathbf{u}$, and predicts a clear sample $\hat{\mathbf{c}}_0$.

The architecture of the denoising network is build upon a stack of Diffusion Transformer (DiT) blocks [PX23], which is shown in Figure 3. The input noisy sample $\mathbf{c}_t = (c_t^1, c_t^2, \ldots, c_t^N)$ is first converted to a sequence of embeddings via a linear projection, and the

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

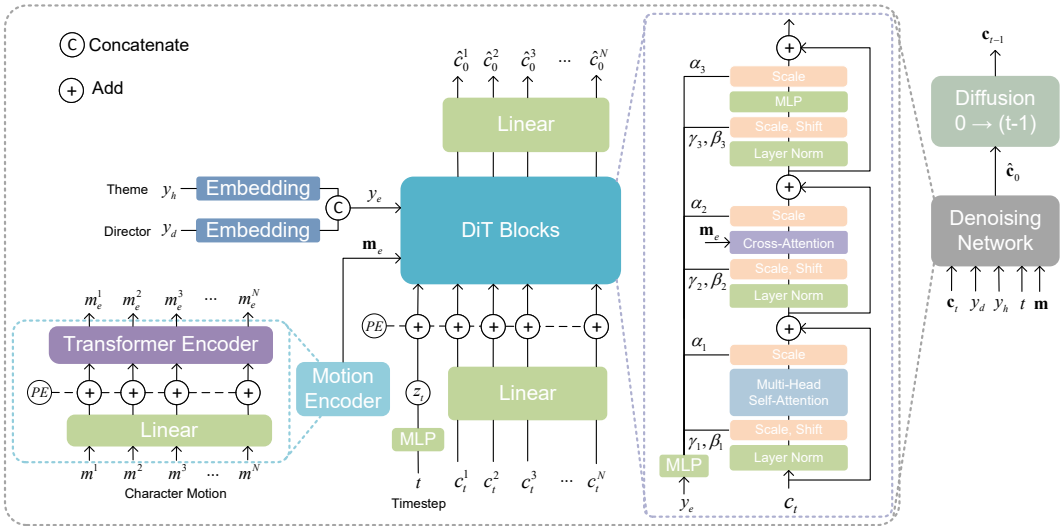


Figure 3: *Left*: Architecture of our denoising network, which takes as input a noisy camera motion $\mathbf{c}_t = (c_t^1, c_t^2, c_t^3, \cdots, c_t^N)$, a timestep t, and conditioning inputs (including a 3D character motion $\mathbf{m} = (m^1, m^2, m^3, \cdots, m^N)$, a scene theme y_h, a director label y_d), and predicts a clear camera motion $\hat{\mathbf{c}}_0 = (\hat{c}_0^1, \hat{c}_0^2, \hat{c}_0^3, \cdots, \hat{c}_0^N)$. *Right*: Iterative denoising process (generation process). Each reverse step goes from $\mathbf{c}_t$ to $\mathbf{c}_{t-1}$ by first estimating a clean sample $\hat{\mathbf{c}}_0$ from $\mathbf{c}_t$ using the denoising network, and then diffusing $\hat{\mathbf{c}}_0$ to timestep $t-1$ to obtain $\mathbf{c}_{t-1}$.

timestep embedding of t is prepended to the sequence. We then augment the resulting sequence with position embeddings before processing them by a sequence of DiT blocks. To incorporate the input animation $\mathbf{m} = (m^1, m^2, \ldots, m^N)$, we use a motion encoder that is a standard Transformer encoder to map $\mathbf{m}$ to a human motion representation $\mathbf{m} = (m_e^1, m_e^2, \ldots, m_e^N)$, which is then injected into each DiT block via a cross-attention layer. To condition on y_h and y_d, each of them is independently encoded into an embedding. The two embeddings are then concatenated together and integrated into each DiT block using the adaptive layer norm of DiT with zero initialization (i.e., AdaLN-Zero).

The denoising network is optimized by minimizing the following simplified objective [TRG*23, HJA20]:

$$L_{\text{simple}} = \mathbb{E}_{\mathbf{c}_0, \mathbf{u}=(y_h, y_d), t\sim[1,T]} \left[\|\mathbf{c}_0 - f_\varphi(\mathbf{c}_t, t, \mathbf{u})\|_2^2 \right], \qquad (2)$$

where, given $\mathbf{c}_0$, the noisy sample $\mathbf{c}_t$ at timestep t can be efficiently sampled from the marginal distribution $q(\mathbf{c}_t|\mathbf{c}_0)$ defined by the forward process. At inference time , our model generates camera motions given conditional inputs by running the learned reverse process that starts from $\mathbf{c}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At each reverse step, the denoising network produces the predicted clean sample $\hat{\mathbf{c}}_0$ from $\mathbf{c}_t$, which is then diffused to $\mathbf{c}_{t-1}$ using $q(\mathbf{c}_t|\mathbf{c}_0)$.

4. Our Dataset

To train our model, we collect a set of film clips that span a range of scene theme and film directors. Each clip correspond to a shot in the film, and is annotated with a theme label y_h and a director label y_d. Furthermore, for each shot clip, we estimate its character motion $\mathbf{m}$ and camera motion $\mathbf{c}$ in 3D space. This gives rise to our dataset $\mathcal{D} = \{(\mathbf{m}, \mathbf{c}, y_h, y_d)\}$ with a total of 3285 samples. We first describe our film clilp collection pipeline and the introduce our method for 3D character and camera motion estimation.

4.1. Film Clip Collection

We select a list of well-known film directors and, for each director, we consider one or several movie genres that the director is most known for: Wong Kar Wai for romance, Woody Allen for romance and comedy, Alfred Hitchcock and David Lynch for mystery, and Pedro Almodóvar for comedy. For each director, we download several film videos of the considered genres, e.g., the videos of the romance and comedy films by Woody Allen are downloaded; we also get the subtitle file for each downloaded video. For each film, we then apply the shot detection method of [PyS] to obtain all the shots. For each shot, we use its timestamps to extract the subtitles of the shot from the subtitle file, and leverage GPT-4 to determine the scene theme of the shot based on the extracted subtitles. Specifically, we first query GPT-4 to identify typical film scenes shared by film directors, using the prompt: "*Summarize a set of typical romance/comedy/mystery scenes of "two character" that are all shared by popular eastern and western directors.*" . This results in a list of scene themes for each genre. For example, five schemes including "chance meetings", "conflicted love", "intimate conversions", "moments of realization", "separation and reunion" are returned by GPT-4 for "romance". We then feed a shot's subtitles along the list of scene themes to GPT-4 and instruct it to choose one theme from the list based on the subtitles. We filter for shots whose subtitles are present to ensure that each shot in our dataset has a scene label obtained using the above method. We also only keep shots with one or two characters. For each shot, we sample a film clip consisting of a sequence of frames. Each film clip has a a maximum length of 120 frames (spanning 6 seconds with FPS 20); longer shots are cropped. After filtering out film clips where the human motion extraction (introduced in Section 4.2) fails to yield good results, e.g., due to severe occlusion, we end up with 3,285 film clips. In Table 1, we report the number of retained flim clips for different genres of each director.

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

Table 1: Number of film clips (shots) for each combination of director and genre in our dataset.

	Romance	Mystery	Comedy
Wong Kar Wai	190	-	-
Woody Allen	460	-	246
Alfred Hitchcock	-	1395	-
David Lynch	-	560	-
Pedro Almodóvar	-	-	434

4.2. Human and Camera Motion Extraction

For each film clip in our dataset, we extract the 3D motions of its characters, by applying Multi-HMR [BAG*24] to each frame. A character motion is a sequence of 3D character poses, each comprising a set of 3D joint positions. However, the per-frame 3D pose estimation of Multi-HMR may lead to temporal incoherency in the estimated character motions. To mitigate this issue, we propose to optimize the root positions of the characters in the estimated poses. Specifically, we first use a single-character motion estimation approach, GVHMR [SPX*24], to process the entire clip, obtaining a temporally coherent motion for each character in the clip individually. Suppose there are N frames in the clip with characters A and B. Let $\mathbf{x}_{\mathrm{A}}^0 \in \mathbb{R}^{N\times 3}$ and $\mathbf{x}_{\mathrm{B}}^0 \in \mathbb{R}^{N\times 3}$ be the root positions of A and B that are estimated by Multi-HMR, and $\bar{\mathbf{x}}_{\mathrm{A}} \in \mathbb{R}^{N\times 3}$ and $\bar{\mathbf{x}}_{\mathrm{B}} \in \mathbb{R}^{N\times 3}$ be the GVHMR predictions of the root positions of A and B. We then minimize the following objective to optimize the two characters' root positions $\mathbf{x}_{\mathrm{A}}, \mathbf{x}_{\mathrm{B}}$:

$$\mathcal{F}_{\text{total}} = w_{\mathrm{r}}\mathcal{F}_{\mathrm{r}} + w_{\mathrm{v}}\mathcal{F}_{\mathrm{v}} + w_{\mathrm{a}}\mathcal{F}_{\mathrm{a}}, \tag{3}$$

where w_{r}, w_{v} and w_{a} are weights to balance different terms. $\mathcal{F}_{\mathrm{r}}$ is a relative position term, which is defined as:

$$\mathcal{F}_{\mathrm{r}} = \sum_{i=1}^{N} \left\| (\mathbf{x}_{\mathrm{A}}[i] - \mathbf{x}_{\mathrm{B}}[i]) - (\mathbf{x}_{\mathrm{A}}^0[i] - \mathbf{x}_{\mathrm{B}}^0[i]) \right\|^2, \tag{4}$$

where $\mathbf{x}_{\mathrm{A}}[i] \in \mathbb{R}^3$ is the root position of character A at frame i. $\mathcal{F}_{\mathrm{r}}$ preserves the relative positions between A and B that are estimated by Multi-HMR. $\mathcal{F}_{\mathrm{v}}$ is a velocity term written as:

$$\mathcal{F}_{\mathrm{v}} = \sum_{i=2}^{N} \left\| \Delta\mathbf{x}_{\mathrm{A}}[i] - \Delta\bar{\mathbf{x}}_{\mathrm{A}}[i] \right\|^2 + \left\| \Delta\mathbf{x}_{\mathrm{B}}[i] - \Delta\bar{\mathbf{x}}_{\mathrm{B}}[i] \right\|^2, \tag{5}$$

where $\Delta\mathbf{x}_{\mathrm{A}}[i] = \mathbf{x}_{\mathrm{A}}[i] - \mathbf{x}_{\mathrm{A}}[i-1]$ denotes the velocity at frame i. $\mathcal{F}_{\mathrm{v}}$ matches the velocities of each character's root position to those estimated by GVHMR. $\mathcal{F}_{\mathrm{a}}$ is an acceleration term:

$$\mathcal{F}_{\mathrm{a}} = \sum_{i=2}^{N-1} \left\| \Delta\mathbf{x}_{\mathrm{A}}[i+1] - \Delta\mathbf{x}_{\mathrm{A}}[i] \right\|^2 + \left\| \Delta\mathbf{x}_{\mathrm{B}}[i+1] - \Delta\mathbf{x}_{\mathrm{B}}[i] \right\|^2, \tag{6}$$

which minimizes the accelerations of each character's root position to avoid motion jittering. We optimize $\mathcal{F}_{\text{total}}$ using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. Our implementation uses $w_{\mathrm{r}} = 10$, $w_{\mathrm{v}} = 500$ and $w_{\mathrm{a}} = 100$. We initialize $\mathbf{x}_{\mathrm{A}}$ as $\mathbf{x}_{\mathrm{A}}^0$ and $\mathbf{x}_{\mathrm{B}}$ as $\mathbf{x}_{\mathrm{B}}^0$, and run L-BFGS for 1000 iterations. This gives the optimized root positions of the two characters.

Based on the extracted 3D character motions for a film clip, we extract the camera motion by computing the Toric camera pose parameters for each frame, following the approach of [LC15]. See the supplementary material for full details.

5. Model Evaluation

5.1. Implementation Details

Our denoising network comprises 8 DiT blocks, each with 4 attention heads, an embedding dimensionality of 512, and an MLP using GELU activation. We use 80% and 20 % of our dataset as training and test sets, which contain 2629 and 656 samples, respectively. The number of diffusion steps is set to 1000. Our model is trained for 1400 epochs using the Adam optimizer with an initial learning rate of 1e-4 and a batch size of 64.

5.2. Evaluation Metrics

We evaluate the performance of our model using the following quantitative metrics:

- *FrameFID* computes FID between generated and real frames, measuring the quality of frame compositions. Following [WJS*24], to calculate FrameFID, we define the features of a frame as $\left(\frac{S3}{S1}, \frac{S3}{S2}\right)$, where S_1 denote the projection area of the full body on the image plane, S_3 is the area of the body parts inside the frame, and S_2 is the frame area.
- *CamSeqFID* evaluates the quality of generated camera motions by comparing them against camera motions extracted from film clips through FID. To compute CamSeqFID, we train a Transformer-based autoencoder on camera motion sequences in our dataset, and use its encoder to extract camera motion features.
- Character motion matching score (*CMM-Score*) is used to judge the consistency between the input character motion and the generated camera motion. It is effectively an indicator of how well the generated camera behavior portray the input animation. For computing CMM-Score, we train a character motion encoder and a camera motion encoder to project character and camera motions into a joint embedding space. Both encoders are learned jointly through contrastive learning as in CLIP, so that matched character and camera motions lie close in the embedding space while mismatched ones are far apart. CMM-Score between a character motion and a camera motion is defined as cosine similarity between their embeddings.
- Style Recognizability (*StyleReg*) evaluates how well generated camera motions emulates the target director's style. Our intuition is that, if a camera motion successfully exhibits the distinctive style of a director, a well-trained style recognition model should be able to correctly classify the camera motion as being created by the director. Motivated by this, we build a classifier that takes as input a character animation, a scene scheme label as well as a camera motion, and predicts a director label. We exploit the encoders used in CMM-Score to extract the features of the input character and camera motions and obtain the scheme features via a MLP. Then, all the features are concatenated and fed into a MLP to output class probabilities, one for each director in our dataset. To train the classifier, we use 20% of each director's samples for training and the rest for testing; the classifier yields

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

Table 2: Quantitative comparison of different models. The best results are in bold.

Method	FrameFID ↓	CamSeqFID ↓	CMM-Score ↑	StyleReg ↑
CineCamera	1.94	17.59	0.256	0.7881
DanceCamera3D	1.63	16.83	0.424	0.7363
Ours	**1.33**	**5.76**	**0.651**	**0.8796**

Table 3: Ablation study results. The best results are in bold.

Method	FrameFID ↓	CamSeqFID ↓	CMM-Score ↑	StyleReg ↑
w/o Character Motion	1.56	7.48	0.305	0.7651
w/o Scene Theme	3.36	7.29	0.562	0.8323
w/o Director Label	1.46	8.11	0.505	0.8049
Ours	**1.33**	**5.76**	**0.651**	**0.8796**

91.46% accuracy on the test split. We report the average accuracy of the classifier on generated camera motions as StyleReg.

5.3. Comparison to Baselines

Since there is no prior work specifically for generating camera motions conditioned on director and scheme labels, we use two recent camera motion generation models, CineCamera [JWC*24] (a text-driven model) and DanceCamera3D [WJS*24] (a music-driven model), as baselines, and evaluate our model against them. For CineCamera, we combine the input director and scheme labels as the input text; for DanceCamera3D, we replace its music pathway (from the music input to the music embedding) with a branch that linearly projects the input director and scheme labels, adds the resultant embeddings and feeds the obtained embedding into a MLP. All the models are trained on our training set, and tested on the test set.

The results are reported in Table 2. Our model outperforms CineCamera and DanceCamera3D across all the metrics, suggesting the strong performance of our model in generating high-fidelity camera motions that well match the input animations and faithfully reproduce the distinctive styles of various filmmakers. Figure 4 shows the visual comparison of camera motions generated by different methods. Our results better align with those directly extracted from film clips, compared to those from the other two models.

5.4. Ablation Study

We conduct an ablation study to evaluate the importance of the three conditional inputs in our model: character motion, scene theme, director label. We consider three variants of our model that remove the three conditions, respectively. The results are reported in Table 3. Removing the character motion results in a substantial reduction in CMM-Score and also degrades StlyeReg significantly, since the generated camera motion is irrelevant to the input character motion. The absence of the scene theme leads to a large increase in FrameFID, suggesting its importance for frame composition quality. Without the director label, significant degradations in CamSeqFID and StyleReg are observed, which implies that it is vital for generating the distinctive style of a specific director.

6. Model-based Study

The results in Section 5 has demonstrated the promising capabilities of our model in emulating camera motion styles of different filmmakers for a given scenario (specified by a 3D animation and a theme). With our model, we conduct a cinematography style study, where we perform quantitative analysis on samples from our model to gain an understanding of *what features contribute to the unique cinematography approaches of different filmmakers*.

For this study, we hypothesize that a cinematography style is characterized by three fundamental factors: *shot type* (e.g., close-up shot, medium shot), *shot composition* (i.e., the arrangement of visual elements within a frame), and *camera movement* (e.g, pan, push in and arc). Note that we do consider other cinematography factors such as lighting and color, which are not modeled by our model. Furthermore, we consider three common genres, including romance, comedy and mystery, which appear in our dataset. For each genre, we compare the styles of two directors who are famous for making the films of the genre. For this, we apply our model to each 3D character animation in the test split obtained in Section 5 to generate two camera motions, which are conditioned on the same, randomly selected scheme label under the genre but two different target director labels. This gives rise to a set of 656 camera motions for each of the two directors. Comparing the statistics of the two camera motion sets, in terms of the above cinematography factors, allows us to understand how the cinematographic decisions of the two directors differ when facing the same scenarios. Note that the aim of this study is to provide an demonstration of how our model can be applied to understand cinematography styles, rather than to draw rigorous conclusions about what defines a film director's style; imperfect model generations may result in the study's statistical results not being complete accurate. In the remainder of this section, we present the results and analysis for each of the aforementioned genres.

6.1. Romance

We first compare the styles of *Wong Kar Wai* and *Woody Allen* for the *romance* genre. Figure 5 visualizes the shot type distributions of the two directors. In the experiments, we consider 7 typical shot types, including extreme close-up (ECU), close-up (CU), medium close-up (MCU), medium shot (MS), medium full shot (MFS), full

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

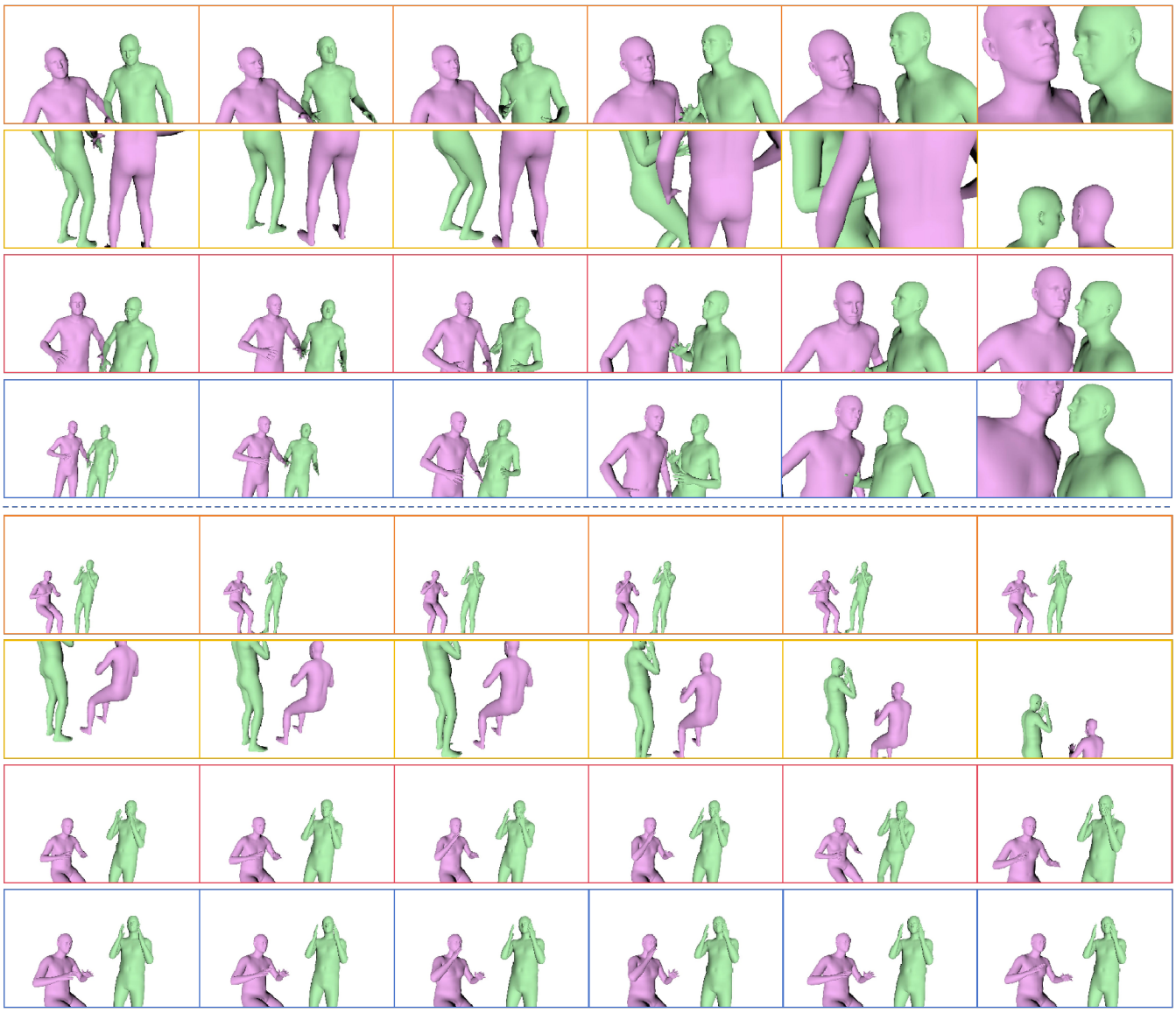


Figure 4: Comparison of generated camera motions by different models. For each camera motion, we display a sequence of 2D frames rendered using 6 camera poses that are sampled uniformly along the trajectory. We show two groups of results, each consisting of camera motions synthesized by CineCamera (orange), DanceCamera3D (yellow) and our model (red) as well as the camera motion extracted from the film clip (blue). The input scene themes for the top and bottom groups are "awkward situatio" and "intimate conversation", respectively.

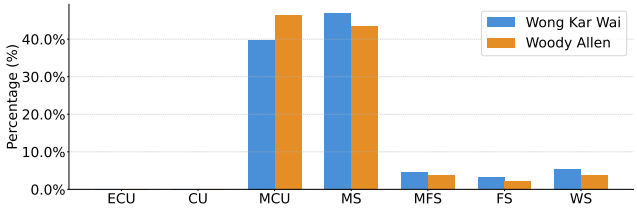


Figure 5: Shot type distributions of Wong Kar Wai and Woody Allen for the romance genre.

shot (FS), wide shot (WS). Wong Kar Wai uses wide shot more frequently than Woody Allen. The wide shot, which places the camera far from characters, is used to represent the relationship of characters to the environment. It has an effect of isolating characters from the world, making them lonely or overwhelmed. In contrast, Woody Allen relies more on medium close-up, which frames a character from the top of the head to the middle chest, to highlight character details and create an intimate feeling without losing physicality.

To investigate differences in shot composition between the two directors, we plot the spatial distributions of characters on the screen for Wong Kar Wai and Woody Allen in Figure 6. We nor-

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

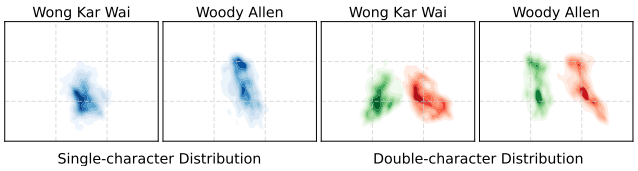


Figure 6: Distributions of on-screen character head positions for Wong Kar Wai and Woody Allen for the romance genre. In each subplot, the "rule of thirds" lines (gray dashed lines) are drawn for reference.

malize the 2D on-screen coordinates of character heads into $[-1, 1]$, and consider the distributions of character head positions for single-character and double-character scenarios separately. We observe that Wong Kar Wai prefers to frame characters at the lower part of the screen. Furthermore, the double-character composition of Wong Kar Wai adheres to the rule of thirds, positioning two characters at the two lower interaction points of the "rule of thirds" lines. Woody Allen, however, tends to position characters near the center of the frame. Such center composition, when combined with his frequent use of medium close-up, furthers his dialogue-driven film style. More specifically, for shooting a single character, the rule of thirds is broken, and the character is often placed at the center. When two characters are present in a frame, the composition is made symmetrical, where the two characters are placed along the two vertical grid lines of the rule of thirds.

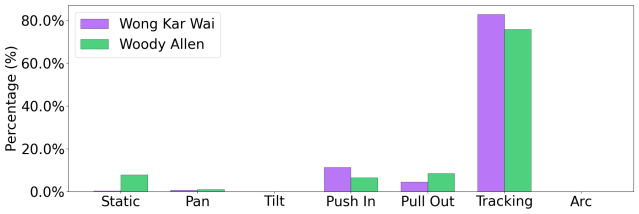


Figure 7: Camera movement distributions of Wong Kar Wai and Woody Allen for the romance genre.

Figure 7 compares the distributions of camera movements for Wong Kar Wai and Woody Allen. We consider several common camera movements, including static (the camera is still), pan (the camera rotates horizontally while being in a fixed location), tilt (the camera rotates vertically while being in a fixed location), push in (the camera moves closer to a subject), pull out (the camera moves further away from a subject), tracking (the camera follows a subject), and arc (the camera rotates around a subject). We observe pronounced preference distinctions between the two directors in camera movement. In particular, Wong Kar Wai opts to use push in and tracking more often to elevate tension in a shot and get audiences immersed into a scene. Woody Allen favors the static camera, which is great for dialogues; he also more frequently uses pull out, which can unveil the context of a scene or characters and emphasize some negative feelings such as isolation.

6.2. Comedy

We then analyze the style differences between *Pedro Almodóvar* and *Woody Allen* for the *comedy* genre. As shown in Figure 8, compared to Woody Allen, Pedro Almodóvar adopts medium close-up more extensively to highlight the exaggerated expressions and physical movements of the characters for a stronger comedic effect. In addition, full shot, which frames a character's entire body from the top to the bottom of the screen, is also more widely used by Pedro Almodóvar to advance the story by showing the entire bodies, poses and costumes of characters. Woody Allen relies more strongly on medium shot that starts just above the head and ends just above the waist, to capture the subtle movements of characters' upper bodies to express the humor through body language. Meanwhile, he uses more wide shots to present the strong contrast between characters' clumsy behaviors and realistic environments surrounding them, which makes the characters look more absurd.

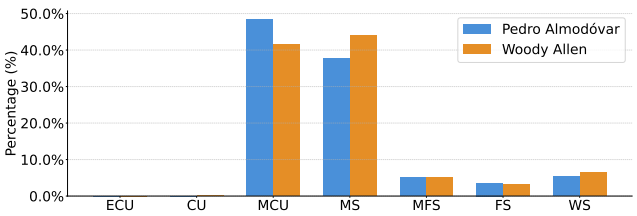


Figure 8: Shot type distributions of Pedro Almodóvar and Woody Allen for the comedy genre.

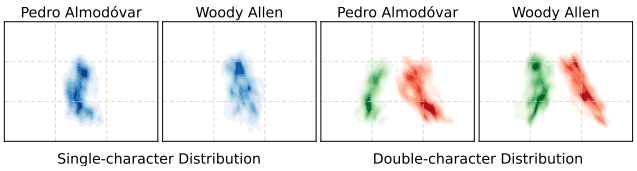


Figure 9: Distributions of on-screen character head positions for Pedro Almodóvar and Woody Allen for the comedy genre. In each subplot, the "rule of thirds" lines (gray dashed lines) are drawn for reference.

We show the spatial distributions of character heads on the screen in Figure 9. For the single-character case, Both directors embrace center framing, but Pedro Almodóvar's distribution spreads density more uniformly across the center region, while Woody Allen's distribution puts more density at a higher position near the top grid line of the rule of thirds. In case of two characters, both directors shoot symmetrical shots by positioning the two characters along the two vertical grid lines of the rule of thirds. However, Pedro Almodóvar tends to place the characters' heads at the two lower intersection points of the "rule of third" lines; Woody Allen makes use of all the four intersection points in his compositions.

Figure 10 shows the camera movement distributions of the two directors. Pedro Almodóvar utilizes push in more frequently to elevate tension in a scene, whereas Woody Allen relies more on tracking to make audiences move with a character in order to directly immerse them into the scene.

6.3. Mystery

We finally study differences between the cinematic decisions of *Alfred Hitchcock* and *David Lynch* for the *mystery* genre. As shown

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

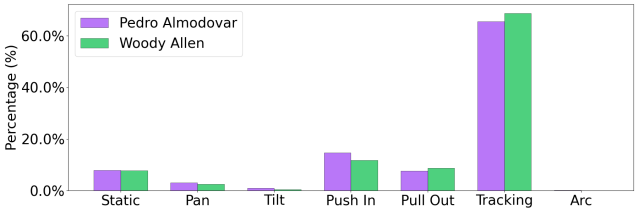


Figure 10: Camera movement distributions of Pedro Almodóvar and Woody Allen for the comedy genre.

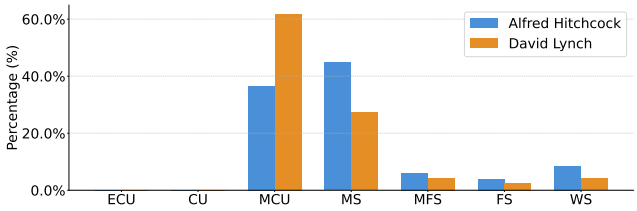


Figure 11: Shot type distributions of Alfred Hitchcock and David Lynch for the mystery genre.

in Figure 11, Alfred Hitchcock opts to include more environmental contexts on the screen by more frequently using medium shot, medium full shot, full shot and wide shot. The Alfred Hitchcock's shot type choices aim to build tension and suspense in a shot by emphasizing the relationship between characters and the environment surrounding them—e.g., a shot showing that a character is exploring an unfamiliar environment can keep audiences in suspense. This is in strong contrast to David Lynch's extensive use of medium close-up, which prioritizes a character's facial expressions and physical details to unveil the feelings (e.g., anxiety) and internal conflict of the characters.

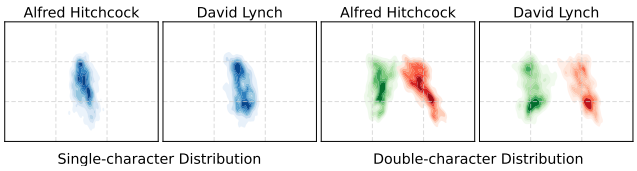


Figure 12: Distributions of on-screen character head positions for Alfred Hitchcock and David Lynch for the mystery genre. In each subplot, the "rule of thirds" lines (gray dashed lines) are drawn for reference.

According to the character head position distributions in Figure 12, Alfred Hitchcock likes to center characters in a frame in both single-character and double-character settings. In contrast, David Lynch, in general, tends to frame characters at a relatively lower position on the screen to weaken the spatial dominance of the characters, making them oppressive and overwhelmed, and suggesting their helplessness. It can also be observed that David Lynch's frame composition well follows the rule of thirds, with the higher density of the distributions near either the grid lines or intersection points of the rule of thirds.

Alfred Hitchcock and David Lynch exhibit distinct preference in planning camera movements, when shooting mystery scenes,

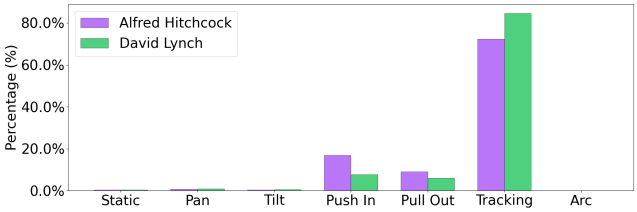


Figure 13: Camera movement distributions of Alfred Hitchcock and David Lynch for the mystery genre.

which is shown in Figure 13. Alfred Hitchcock's use of push in is much more frequent than that of David Lynch, drawing the viewer's attention to details crucial for the narrative. He also prefers to utilize pull out to reveal the context of characters in a shot. David Lynch relies more heavily on tracking to keep audiences thinking about what is going to happen next, thus building suspense and creating an immersive experience. Both directors rarely use static, pan and tilt for mystery scenes.

7. Conclusions and Discussions

In this paper, we present an attempt to model camera motions styles of film directors and understand their unique cinematography styles in terms of placing and moving the camera. To this end, we propose a diffusion-based generative model of camera motions and condition it on various conditional inputs including character motion, scene theme and director. Our model enables a controllable approach to synthesizing stylistic camera motion trajectories, and a novel methodology for cinematography style study. We hope our work can inspire future research in leveraging recent advances in deep generative models to study artist styles in other art or design domains.

Our method has several limitations. First, in our current trained model, the input director label and scene scheme are not fully decoupled; for example, our model can not faithfully generate the camera motion style of Alfred Hitchcock for a comedic scheme, since our training dataset lacks data from the movies of Alfred Hitchcock under the comedy genre (as shown in Table 1). This issue could be alleviated by further expanding the scale of our dataset so that it can cover as many genres as possible for each director. Further, our model can not handle director labels outside our training dataset. Second, the evaluation metrics used in our experiments rely on character and camera motions estimated from film clips. The estimation errors may adversely affect the reliability of the absolute metric scores. However, we evaluate the performance of our model based on its relative scores with respect to the baselines, which could provide a reasonable way to showcase the superiority of our model. In addition, these metrics may not be well aligned with human perception of camera motion quality and styles. Thus, we plan to conduct a large-scale human evaluation with cinematography experts to better gauge the abilities of our model. Third, our model generations may contain errors, especially for conditional inputs that are relatively rare in the training dataset. This could adversely affect the accuracy of the analysis results. To mitigate this issue, one way is to further improve the generation performance of our model, and another possibility to develop a mechanism for au-

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

tomated evaluation of the quality of individual samples from our model, so that low-quality samples can be excluded from the analysis.

Acknowledgments

This work was supported by the Research Funding of Liaoning Provincial Education Department under Reference LJ242412026O1.

References

[AoTPB*92] AUMONT J., OF TEXAS PRESS U., BERGALA A., MARIE M., NEUPERT R., VERNET M.: *Aesthetics of Film*. Film Studies. University of Texas Press, 1992. URL: https://books.google.com.hk/books?id=nmpZAAAAMAAJ. 2

[Ari91] ARIJON D.: *Grammar of the film language*. Silman-James Press, 1991. 2

[BAG*24] BARADEL* F., ARMANDO M., GALAAOUI S., BRÉGIER R., WEINZAEPFEL P., ROGEZ G., LUCAS* T.: Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV* (2024). 5

[Bow13] BOWEN C.: *Grammar of the Shot*. Routledge, 2013. 2

[CRE*14] CHRISTIAN K., RITSCHEL T., EISEMANN E., THORMÄHLEN T., SEIDEL H.-P.: Generating realistic camera shake for virtual scenes. *Journal of Virtual Reality and Broadcasting 10*, 7 (2014), 1–18. 3

[CXH*26] CHENG B., XI Y., HE R., NA J., CAO Y., WANG P., ZHANG J. J., YANG X.: Towards storytelling animations: Joint synthesis of human and camera motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2026), pp. 38376–38386. 3

[DZ94] DRUCKER M. S., ZELTZER D.: Intelligent camera control in a virtual environment. In *Proceedings of Graphics Interface* (1994), pp. 190–199. 3

[EHG22] EVIN I., HÄMÄLÄINEN P., GUCKELSBERGER C.: Cine-ai: Generating video game cutscenes in the style of human directors. *Proceedings of the ACM on Human-Computer Interaction 6*, CHI PLAY (2022), 1–23. 3

[GCB*19] GSCHWINDT M., CAMCI E., BONATTI R., WANG W., KAYACAN E., SCHERER S.: Can a robot become a movie director? learning artistic principles for aerial cinematography. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019), IEEE, pp. 1107–1114. 3

[GCLR15] GALVANE Q., CHRISTIE M., LINO C., RONFARD R.: Camera-on-rails: automated computation of constrained camera paths. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games* (2015), pp. 151–157. 3

[Gle12] GLEBAS F.: *Directing the story: professional storytelling and storyboarding techniques for live action and animation*. Routledge, 2012. 2

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems 33* (2020), 6840–6851. 3, 4

[HLH*16] HUANG H., LISCHINSKI D., HAO Z., GONG M., CHRISTIE M., COHEN-OR D.: Trip synopsis: 60km in 60sec. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 107–116. 3

[HLY*19] HUANG C., LIN C.-E., YANG Z., KONG Y., CHEN P., YANG X., CHENG K.-T.: Learning to film from professional human motion videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4244–4253. 3

[HWC24] HE R., WEI H., CAO Y.: An interactive system for supporting creative exploration of cinematic composition designs. In *Proceedings of the ACM Symposium on User Interface Software and Technology* (2024), pp. 78:1–78:15. 3

[HYK*19] HUANG C., YANG Z., KONG Y., CHEN P., YANG X., CHENG K.-T.: Learning to capture a film-look video with a camera drone. In *Proceedings of the International Conference on Robotics and Automation* (2019), pp. 1871–1877. 3

[JCW*21] JIANG H., CHRISTIE M., WANG X., LIU L., WANG B., CHEN B.: Camera keyframing with style and control. *ACM Transactions on Graphics (TOG) 40*, 6 (2021), 209:1–209:13. 3

[JWC*24] JIANG H., WANG X., CHRISTIE M., LIU L., CHEN B.: Cinematographic camera diffusion model. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e15055. 3, 6

[JWW*20] JIANG H., WANG B., WANG X., CHRISTIE M., CHEN B.: Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 45:1–45:14. 3

[LC15] LINO C., CHRISTIE M.: Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG) 34*, 4 (2015), 1–12. 3, 5

[LCL*10] LINO C., CHRISTIE M., LAMARCHE F., GUY S., OLIVIER P.: A real-time cinematography system for interactive 3d environments. In *SCA'10 Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2010), pp. 139–148. 3

[OSTG09] OSKAM T., SUMNER R. W., THUEREY N., GROSS M.: Visibility transition planning for dynamic camera control. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (2009), pp. 55–65. 3

[PX23] PEEBLES W., XIE S.: Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 4195–4205. 3

[PyS] Pyscenedetect. *https://github.com/Breakthrough/PySceneDetect*. 4

[RS24] RETAMOZO SANCHEZ M.: *Towards automatic analysis of cinematic style of film directors*. Phd thesis, The University of Queensland, 2024. URL: https://doi.org/10.14264/5c4fc7b. 2

[SDM*14] SANOKHO C. B., DESOCHE C., MERABTI B., LI T.-Y., CHRISTIE M.: Camera motion graphs. In *Symposium on Computer Animation* (2014), pp. 177–188. 3

[SPX*24] SHEN Z., PI H., XIA Y., CEN Z., PENG S., HU Z., BAO H., HU R., ZHOU X.: World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings* (2024). 5

[SS87] SOBCHACK T., SOBCHACK V.: *An Introduction to Film*. Little, Brown, 1987. URL: https://books.google.com.hk/books?id=SwMbAQAAIAAJ. 2

[TRG*23] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. In *The Eleventh International Conference on Learning Representations* (2023). URL: https://openreview.net/forum?id=SJ1kSyO2jwu. 4

[WJS*24] WANG Z., JIA J., SUN S., WU H., HAN R., LI Z., TANG D., ZHOU J., LUO J.: Dancecamera3d: 3d camera movement synthesis with music and dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 7892–7901. 3, 5, 6

[WWK23] WU X., WANG H., KATSAGGELOS A. K.: Automatic camera trajectory control with enhanced immersion for virtual cinematography. *arXiv preprint arXiv:2303.17041* (2023). 3

© 2026 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.