An Interactive System for Supporting Creative Exploration of Cinematic Composition Designs

Rui He
The Hong Kong Polytechnic University
Kowloon, Hong Kong
henry.he@connect.polyu.hk

Huaxin Wei*
The Hong Kong Polytechnic University
Kowloon, Hong Kong
huaxin.wei@polyu.edu.hk

Ying Cao
ShanghaiTech University
Shanghai, China
caoying59@gmail.com

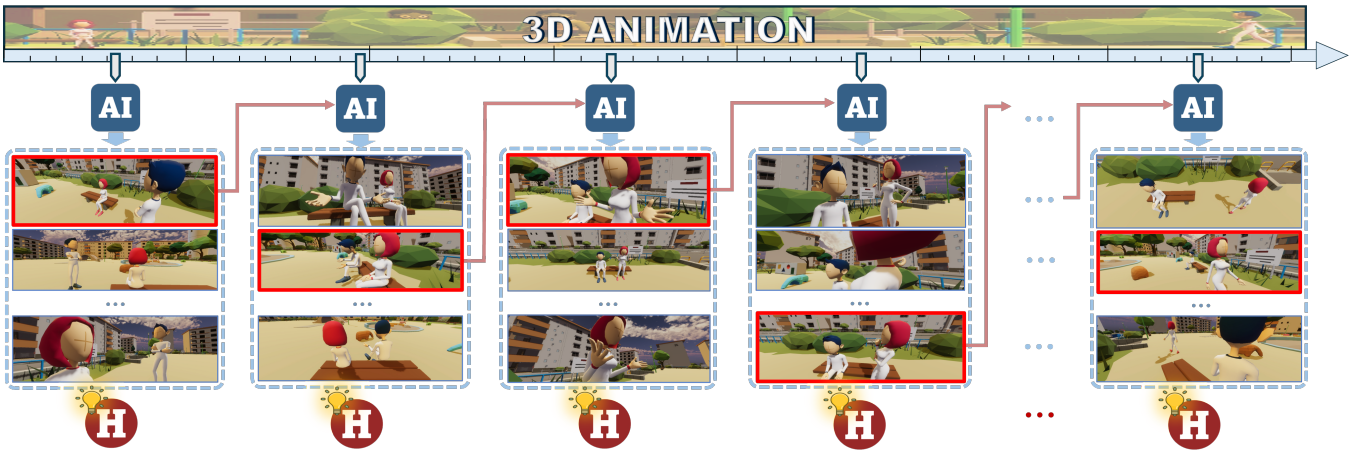

Figure 1: Cinemassist is an interactive system that helps users design cinematic compositions for 3D animations by providing suggestions based on user-selected keyframes and input semantics (such as movie genre and intended emotion state). It combines human decision-making with AI-generated suggestions, allowing users to have creative control while benefiting from AI assistance. For each keyframe (indicated by a marker on the animation timeline), our model ("AI") suggests a list of composition design options (outlined by a blue dashed rectangle) and the user ("H") explores the options to select a preferred one (outlined in red). The user's decision will influence the model's suggestions in the next keyframe.

ABSTRACT

Designing cinematic compositions, which involves moving cameras through a scene, is essential yet challenging in filmmaking. Machinima filmmaking provides real-time virtual environments for exploring different compositions flexibly and efficiently. However, producing high-quality cinematic compositions in such environments still requires significant cinematography skills and creativity. This paper presents *Cinemassist*, a tool designed to support and enhance this creative process by generating a variety of cinematic composition proposals at both keyframe and scene levels, which users can incorporate into their workflows and achieve more creative results. At the crux of our system is a deep generative model trained on real movie data, which can generate plausible, diverse camera poses conditioned on 3D animations and additional input semantics. Our model enables an interactive cinematic composition design workflow where users can co-design with the model by being inspired by model-generated suggestions while having control over the generation process. Our user study and expert rating find Cinemassist can facilitate the design process for users of different backgrounds and enhance the design quality especially for users with animation expertise, demonstrating its potential as an invaluable tool in the context of digital filmmaking.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**.

KEYWORDS

Creativity Support Tool, Digital Filmmaking, Machine Learning, Intelligent Cinematography, Machinima

ACM Reference Format:
Rui He, Huaxin Wei, and Ying Cao. 2024. An Interactive System for Supporting Creative Exploration of Cinematic Composition Designs. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST*

*Huaxin Wei is the corresponding author.

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0628-8/24/10
https://doi.org/10.1145/3654777.3676393

'24), October 13–16, 2024, Pittsburgh, PA, USA. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3654777.3676393

1 INTRODUCTION

Planning camera positions and movements through a scene of an ongoing event for storytelling is a fundamental task in filmmaking, which we refer to as *cinematic composition design*. Setting a camera pose – i.e., the camera angle and distance relative to actors in the scene – determines how the actors and the set are composed on the screen, and composing in different ways can convey different meanings and messages to audiences, as shown in Figure 2. Designing effective cinematic compositions is a daunting task since it requires extensive cinematographic knowledge, experience and creativity. In digital filmmaking, machinima tools that enable more efficient exploration and evaluation of design alternatives within virtual environments have become increasingly popular. Unfortunately, cinematic composition design in such environments remains very challenging, as one needs to navigate through huge amounts of options to find an ideal composition for a moment at frame level and ensure the compositions of different moments are connected coherently [13]. Although many cinematographic rules and referential examples are available, machinima-based filmmakers still struggle to draw inspiration from them in practice.

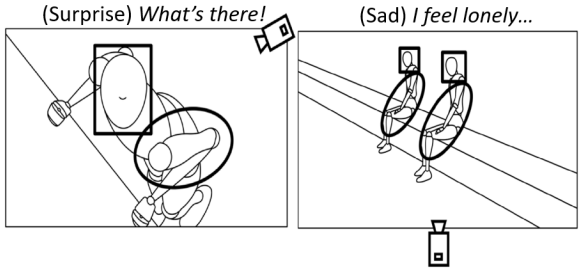

Figure 2: Composing a scene from different camera viewpoints can convey different emotions.

Although there are a few studies on automating cinematography in virtual environments, none of them can serve as effective solutions to supporting creativity in the cinematic composition design process. In view of this, we aim to create an intelligent creativity support tool to aid the ideation and exploration of cinematic compositions for not only machinima filmmakers but also those traditional filmmakers who use the machinima platform during pre-production to develop storyboards. To this end, we conducted a field interview with three professional machinima filmmakers, gaining a better understanding of the machinima filmmaking workflow, the composition design process and design difficulties. Informed by the interview findings and our literature review, we develop a novel system, called Cinemassist, to support the cinematic composition design process. Given a 3D animation along with additional conditioning semantics (e.g., movie genre and intended emotion state), our system is able to support a user in designing a trajectory of 3D camera poses (i.e., cinematic compositions) through several iterations. At each iteration, both our system (computer) and the user (human) can contribute to, and thus, influence the design. *On the computer side*, our system first suggests a set of diverse composition alternatives that fit the content of the user-chosen keyframe and are coherent with the previous composition designs. *On the human side*, the user then turns to the suggested designs for inspirational ideas and create a final composition, which will be fed into the system to generate the suggestions for the next iteration. This formulates an interactive workflow that alternates human decisions and computer suggestions throughout the design process so that the user can get inspired by diverse suggested options while enjoying considerable amounts of freedom to customize the design.

We conduct a user study where representative users of different backgrounds design cinematic composition sequences with and without our system, as well as an expert rating study assessing the users' design outcomes. The results suggest that Cinemassist can effectively facilitate the users' creative process in different ways, helping them yield better design outcomes, and boost the performance of the users with animation expertise. Furthermore, based on the study results, we identified several design implications for further iterating the design of the system and informing the design of relevant creativity support tools for tasks within such a 3D real-time environment. In summary, the main contributions of this paper include:

- **Artefact contribution.** We introduce the design rationale and system implementation of an intelligent creativity support tool, Cinemassist to facilitate the creative process for cinematic composition design in machinima environment. Our intended audiences are both machinima filmmakers, including digital game producers who craft game cut-scenes using existing animations, and traditional filmmakers who use machinima tools for storyboard ideation during pre-production.
- **Technical methodology contribution.** We propose a deep generative model that is trained on existing movie data to generate different camera pose trajectories to depict the story of a given 3D animation in diverse ways. The model is at the core of Cinemassist, enabling its distinctive interactive workflow, its ability to recommend diverse and coherent cinematic composition solutions, and its flexibility in gaining the knowledge necessary to guide users without the need to hard-code rules.
- **Design implications.** According to the user study and expert rating results, we draw four design implications that can inform the design of related creativity support tools for design tasks in a 3D real-time environment.

2 BACKGROUND AND RELATED WORK

2.1 Machinima Filmmaking

Visual storytelling involves the recreation of a narrative using static or motion imagery materials [16]. However, visual storytelling in a machinima environment differs significantly from that of traditional forms [12]. In traditional filmmaking, a pivotal stage in the production process is the shot design phase [5]. During this stage, the cinematic composition of every ongoing "beat" [41] of a story scene is meticulously planned using "storyboards" as references for the final imagery implementation [23]. Creative cinematic composition translates a scene from a unique viewing perspective for

the intended emotion expression (referred to as "focalization" in narratology) [1, 10]. To acquire this skill, novice designers often turn to the principles of cinematography, which encapsulate the fundamental "grammar of the shot" [10]. Conversely, experienced designers tend to expand their design repertoire through film analysis and drawing inspiration from master exemplars[6, 40, 49]. This convention coincides with many design creativity studies stating that referencing "design examples" from similar or even different domains can foster the creativity of designers, even those already trained to be creative [9, 27, 31, 50].

Machinima is defined as "animated filmmaking within a real-time virtual 3D environment" [39]. Popular machinima environments include commercial digital game engines like Unity and Unreal, as well as 3D filmmaking platforms such as Autodesk 3ds Max and Maya. In contrast to traditional filmmaking, which often relies on fixed storyboards, the "real-time" advantage of machinima allows for a more flexible and cost-effective exploration of creative cinematic composition. Machinima filmmakers can conveniently edit the placement, movement, and animation of virtual characters, enabling them to instantly experiment with and iterate upon design alternatives from different viewing directions before settling on the most suitable one [36]. Given this strength, the 3D digital filmmaking and animation industry has increasingly employed advanced machinima techniques for script rehearsal and the exploration of innovative cinematic composition design ideas before committing to storyboarding [33]. A prominent example is the production of the animation film "EVANGELION:3.0+1.0 THRICE UPON A TIME," which has gained significant commercial success and professional acclaim [3]. Nevertheless, despite the advantages offered by machinima, crafting coherent storytelling through composition sequences within such a dynamic environment remains a challenge [32], which requires machinima filmmakers to explore and evaluate an overwhelming amount of design options. In this work, we focus on designing a system to facilitate this process.

2.2 Creativity Support Tool

The third wave of HCI-oriented creativity research introduced Creativity Support Tools (CSTs), advocating for their potential to enhance human creativity [19–21, 48]. Despite extensive studies on CSTs, tools for designers and artists working in storytelling and content generation contexts are still scarce. Frich et al. highlighted CSTs' roles in enhancing design outcomes, improving design knowledge, and providing new execution methods [20, 44]. Yet, in machinima filmmaking, the focus has largely been on novice needs with systems like the one proposed by Nicolas et al., which assesses shot design "correctness" using a rule-based approach [14]. However, requests of experienced filmmakers who look for real-time feedback on potential design alternatives were neglected. Davis et al.'s "Distributed Exploratory Visualization" model highlighted this request, emphasizing the need for CSTs to facilitate a fluid and low-cost creative exploration and evaluation process [13]. Building on this, our paper presents a new CST system designed to offer real-time, high-quality, and diverse shot suggestions for advanced machinima filmmakers, emphasizing the reciprocal exchange between the user and the system in the creative process – a previously overlooked gap in the field. Consequently, this system plays a "skis" role of CST in Kumiyo Nakakoji's theory [44] and enables an execution method for cinematic composition design, augmenting creative practitioners's creative workflow.

2.3 Intelligent Cinematography

The automation of cinematography tasks in both physical and virtual settings has been the focus of numerous autonomous or interactive cinematography tools. For physical scenes in live-action films, numerous studies have proposed methods for suggesting coherent cinematic composition sequences based on provided video clips. For instance, Moorthy et al. [43] developed a tool that generated shot sequences at appropriate timings by leveraging a master shot that encapsulated the entire scene. Arev et al. [4] recommended shot sequences by identifying the narrative focus within a scene through the analysis and editing of multiple footage sources. Although many of these methods are transferable to the context of animation or digitally constructed films, composition for virtual scenes has its distinct characteristics and challenges.

For virtual scenes, the pioneering study [25] proposed a set of machinima cinematography guidelines in addition to classical rules summarized by Arijon [5]. Subsequent works [15, 28, 35] have built upon those guidelines to enable autonomous cinematography in the machinima environment. However, formalizing all possible rules into computational models can be challenging and there are also great variations on how the guidelines can be applied in practice. Consequently, such rule-based methods are prone to generating results with limited variety. Recent works started to investigate data-driven approaches, training machine learning models from data to predict cinematic compositions directly. For example, Edirlei et al. [15] trained a SVM model to select among several pre-determined, fixed camera poses in the scene. Jiang et al. [29] trained a deep learning model to extract camera behaviours from a reference film clip and re-apply them to a given 3D animation. Unfortunately, all of these methods are fully automatic. They do not allow users to specify their preferences and inject their thoughts into the results, a critical need we identified earlier in the previous section.

Evin et al. [17] introduced Cine-AI, an interactive cinematography tool to generate machinima shot sequences resembling those crafted by a human director. It allows for manual adjustments to the results. However, its generation is agnostic of the input scene content and purely relies on a predefined set of rules. As a result, the generated shot sequences are often of low quality, suffering from the problems of being less relevant to the input scene and lacking coherency. Therefore, a lot of manual refinement is still needed to yield reasonable results, making it ineffective to support machinima filmmaking, which requires immediate quality feedback at the narrative layer [13].

3 UNDERSTANDING MACHINIMA FILMMAKING WORKFLOW

3.1 Field Interview

To better understand machinima filmmakers' design process and design thinking, we conducted field interviews with three professional machinima filmmakers. The three interviewees — P1 (female, junior game designer), P2 (female, junior game designer), and P3 (male, lead game artist) — come from a world-leading digital game

company. P1 and P2 have 3~5 years of professional narrative design experience using machinima tools for MMO game cut-scenes prototyping. P3 has more than 10 years of professional experience using machinima tool for not only cut-scenes but also 3D digital film production. The interview lasted for approximately 60 minutes and was conducted in a face-to-face way. Our interview questions were structured around three topics with no specific limitations on the answers: machinima filmmaking workflow, cinematic composition design, and difficulties encountered in the design process. Consequently, we summarize their feedback as follows.

Machinima filmmaking workflow. Based on our feedback from the three interviewees, we have summarized a common four-stage workflow in digital game cut-scene and 3D digital film production. In the first stage, designers are provided with textual scripts describing the scenes. The second stage involves translating these scripts into tangible scenes, which includes constructing the scene environment and planning the 3D animation within the scene by specifying characters' positions, movements, and behaviours using character models and animations. Notably, the planning process is often based on keyframes that represent pivotal moments within the animation. Subsequently, the third stage focuses on designing a sequence of cinematic compositions to tell the animation which is based on keyframes as well. Notably, P3 commented: "*The 3D scene and animation production and cinematic composition design can be decoupled not only in cut-scene production but also in traditional filmmaking workflows, as the latter stage can be conducted using interim artworks.*" Finally, in the fourth stage, the scene is rendered, incorporating the cinematic sequences into video or descriptive script formats.

Cinematic composition design. During the design of cinematic composition sequences (the third stage of the workflow described above), all three interviewees emphasized their inclination to validate and refine their initial designs for each keyframe. Specifically, they control the virtual camera to composite three principal aspects: the positions, orientations, and scales of the narrative focus on the screen. When time permits, our interviewees might explore alternative design possibilities in pursuit of more creative outcomes. Evaluation criteria for each design encompass alignment with the scene content at the current keyframe, coherency with the preceding keyframes, and effective expression of the intended emotion within the animation. Furthermore, all interviewees noted their propensity to draw inspiration from "master examples" that transcend conventional film cinematography norms. P1 shared: "*I meticulously archive film clips that I find intriguing for reference in my working context.*" Nevertheless, P1 also acknowledged the challenge of comprehending the underlying principles behind these master examples and applying them to enhance her own cinematography work.

Difficulties. All interviewees expressed challenges associated with the cinematic composition stage, notably the time constraints imposed by limited production time. In particular, P1 and P2 concurred that placing and orienting the camera to compose each intended keyframe is time-consuming, and demands precise and intensive manual control through the use of a mouse and keyboard. P3 complained: "*Trying out different cinematic composition alternatives is essential, yet this exploration is significantly constrained by the production agenda.*"

3.2 Design Requirements

Based on the feedback we collected in the field interview, we summarize our design requirements for building a tool that intervenes in the cinematic composition design process, i.e., the third stage of machinima filmmaking workflow when the 3D scene and animation is assumed to be already prepared. First, Cinemassist should significantly enhance the exploration of creative cinematic composition designs within the typical machinima filmmaking workflow (**R1**). Secondly, Cinemassist should provide real-time recommendations for cinematic composition design, at both keyframe and scene levels [13] (**R2**). Thirdly, Cinemassist should recommend a diverse range of plausible design alternatives that serve as inspirational examples, expanding upon conventional design paradigms rooted in cinematography rules (**R3**). Fourthly, Cinemassist should recommend coherent cinematic compositions that seamlessly align with the specific scene context, minimizing the need for manual refinement. (**R4**). Finally, the cinematic composition designs recommended by Cinemassist can be displayed in real-time to facilitate users' quick evaluation (**R5**). In the following section, we flesh out our design features and connect them to the above requirements using labels R1, R2, and so on.

4 OUR SYSTEM

Based on the design requirements distilled from the field interview, we have developed Cinemassist, a creativity support tool designed to enhance the exploration of creative cinematic composition designs. Particularly, Cinemassist implements an interactive paradigm that alternates between human decision and intelligent suggestion throughout the design process of cinematic compositions. This tool is intended for both machinima filmmakers and traditional filmmakers who ideate on and pre-visualize their storyboards within a machinima environment. In this section, we provide a functional overview of Cinemassist with discussion of our design rationale. The Cinemassist interface is shown in Figure 3, comprising three components: (a) a control panel for configuring the input 3D animation and high-level semantics; (b) a design panel for designing cinematic compositions at frame level; (c) a storyboard panel for visualizing recommended composition sequences at scene level. Furthermore, the interface provides a scene view to display the 3D animation in real time, along with suggested camera poses in 3D space, a cinema view to preview the composition from a particular camera pose, as well as an animation timeline to facilitate keyframe selection. In the remainder of this section, we elaborate on each of the three components.

Input configuration. A user can load a 3D animation into the system. In our implementation, we prepared animations for experimental purposes using Unity Timeline which plans and records the position, movement and interaction between virtual characters using pre-given 3D assets. The animation will be displayed in the scene view, under which the timeline is presented. Then, on the

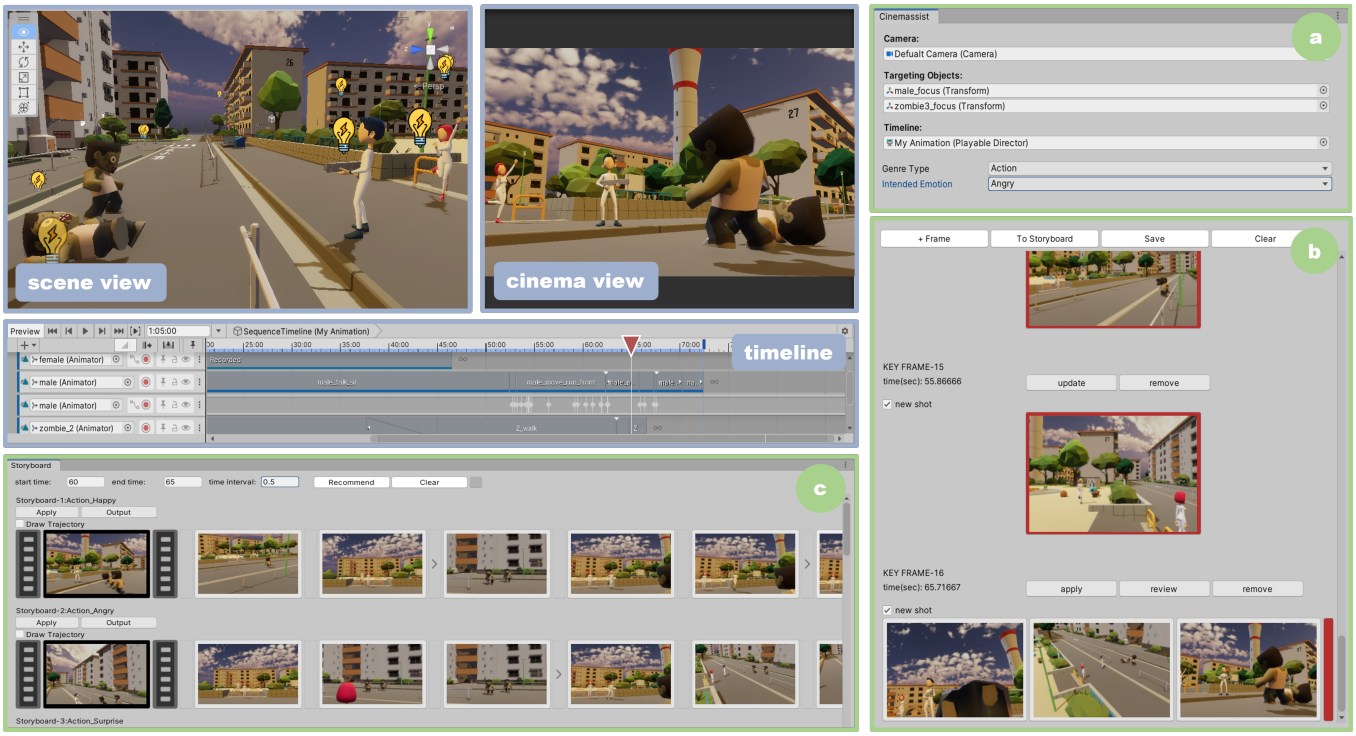

Figure 3: Cinemassist interface consists of three panels: (a) control panel, (b) design panel, and (c) storyboard panel, alongside a scene view, a cinema view, and an animation timeline.

control panel (Figure 3 (a)), the user can select two scene objects as the characters of interest and a camera object in the scene as the default control camera (**R1**). Notably, our design decision of focusing on two characters is primarily due to expressing camera poses with the Toric coordinate system which assumes there are two target characters on the screen (Section 5.2). Additionally, the user can choose one of the movie genre categories (including "action", "romance" and "thriller") that the story of the animation is expected to belong to, and one intended emotion state that the user intends to express through cinematic compositions to the audience. We consider five common intended emotion states including "happy", "angry", "surprise", "sad" and "fear" according to Plutchik's wheel of emotions [42].

Frame-level design exploration. Before initiating the design of cinematic compositions for the specified animation, the user can drag the arrow along the timeline to preview the entire sequence. They can then position the arrow at a critical moment in the animation, designated as a "keyframe". At this juncture, the user clicks the "+Frame" button on the frame-level suggestion panel (Figure 3 (b)) to capture this keyframe. To design the cinematic composition for this keyframe, besides manually looking for a plausible camera pose using the default control camera, our system automatically proposes a set of potential camera poses (**R2**, **R3**). These suggestions are visually represented as light bulbs within the scene view, facilitating *in-situ* exploration in 3D space. When a user selects one of these suggestions, the corresponding 2D composition is immediately displayed in the cinema view, allowing for real-time evaluation (**R5**). Alternatively, users can press the "review" button, which expands to show a row of 2D compositions for all suggested camera poses, ranked by quality scores predicted by our model. This feature provides users with quick visual references to assess the suggestions (**R3**, **R5**). Once a desirable camera pose is found, the user can click the "record" button to confirm the composition of the current keyframe using this pose.

Then, the user moves on to identify the next keyframe, on which the aforementioned system suggestion and user decision procedures are iterated. Notably, at the end of each iteration, the user-chosen option is fed into our model so that our model makes per-iteration predictions based on both the current keyframe content and the previously designed compositions, resulting in relevant and coherent composition suggestions. Additionally, for each keyframe, our system can suggest if it is a shot boundary after which a new shot should be started, to facilitate better camera planning in practice. At last, the user can export the resulting composition sequence and review the animation rendered under it in the cinema view by clicking the "play" button, to get a quick sense of its overall quality (**R5**).

Scene-level design exploration. In addition to the frame-level design exploration above, our system also provides a functionality

to allow the user to explore across sequences of cinematic compositions at scene level. To do this, on the storyboard panel, the user first chooses a range on the timeline by specifying the start and end time points, and set a time interval. Then, after the "recommend" button is clicked, our system automatically samples a sequence of uniformly spaced keyframes using the interval over the specified time range and suggests multiple storyboards (Figure 3 (c)) (**R2**, **R3**, **R4**). Each storyboard is a sequence of cinematic compositions, one for each keyframe. Note that the chosen range can vary to cover the entire animation or only a small part of it. The user can select a satisfying sequence to export into the design panel and continue iterating on it via the frame-level design exploration. Further, the user can also export a cinematic composition sequence in progress from the design panel into the storyboard panel, and our system is able to automatically extend the partial sequence into multiple complete sequences by generating the remaining compositions (**R1**).

5 OUR MODEL

5.1 Problem Formulation

Given a 3D animation as well as some high-level semantics including a movie genre and an intended emotion state, our model aims to generate a camera trajectory through the 3D scene so that the resulting cinematic compositions can well portray the event of the animation while reflecting the input semantics. Specifically, let $\mathbf{s}_{1:T} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T)$ be a sequence of keyframes sampled from the animation, and $\mathbf{h} = \{\mathbf{h}_g, \mathbf{h}_e\}$ be the input semantics where $\mathbf{h}_g$ and $\mathbf{h}_e$ denote the genre and intended emotion, respectively. Our model outputs a sequence of 3D camera poses $\mathbf{c}_{1:T} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T)$, one for each keyframe. To enable our Cinemassist system, we design our model with three major considerations in mind:

- Our model should be *multimodal*, generating diverse alternative options from a single input so that the user can get inspired by exploring different options;
- Our model should be highly *controllable* so that, instead of directly producing end results, it should provide a mechanism to continuously incorporate user decisions into the generation process to offer the user great freedom to control the design procedure;
- Our model should be *learnable*, so that it can acquire enormous amounts of cinematic composition knowledge from data directly, rather than relying on fixed rules.

We address these considerations by proposing an auto-regressive and probabilistic generative model. Compared to other types of deep generative models such as GAN and VAE, auto-regressive models are unique in that they generate a sequence of values by conditioning the prediction of the next value on the previous values. This makes them the ideal choice for implementing our desired controllable workflow where user decisions (i.e., the past values) can be continuously integrated into the generation process to influence future predictions. Formally, our model aims to learn a conditional distribution $p(\mathbf{c}_{1:T}|\mathbf{s}_{1:T}, \mathbf{h})$ of camera pose sequences $\mathbf{c}_{1:T}$ given input 3D animations $\mathbf{s}_{1:T}$ and semantics $\mathbf{h}$, which can be sampled to produce diverse camera pose sequences that are consistent with the inputs (multimodal), as shown in Figure 5. Further, we assume the conditional distribution can be decomposed into the form: $p(\mathbf{c}_{1:T}|\mathbf{s}_{1:T}, \mathbf{h}) = \prod_{t=1}^{T} p(\mathbf{c}_t|\mathbf{c}_{1:t-1}, \mathbf{s}_{1:t}, \mathbf{h})$.

In other words, the output camera poses are generated auto-regressively. At each iteration t, the prediction of the current camera poses $\mathbf{c}_t$ is conditioned on the previously predicted ones $\mathbf{c}_{1:t-1}$ and will be used as an input into the next iteration. With such a mechanism, the user can decide a camera pose for the current iteration based on the model predictions, and the user's decision will influence the model prediction at the next iteration. This allows for great user control over the generation (controllable). Finally, we learn the model from existing movie data, without hard-coding any rules (learnable). In the remainder of this section, we first introduce our frame and camera representation and then present the architecture, training details and generation process of our model.

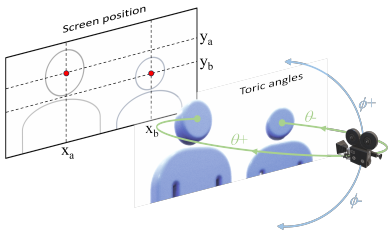

Figure 4: Our camera pose representation is based on the Toric space.

5.2 Frame and Camera Presentation

Each frame in the animation is represented by the 3D poses of one or more characters, each of which is described by a set of 3D body joint positions. Following [27], we express a camera pose in the Toric coordinate system. The Toric representation is defined in the local reference frame of two given target characters, making it easier to capture the correlation between the 3D character poses and the relative camera poses in Toric coordinates. Concretely, the representation $\mathbf{c}_t$ of a camera pose, given two target characters, is written as $\mathbf{c}_t = \{(x_a, y_b), (x_b, y_b), \theta, \phi\} \in \mathbb{R}^6$, where (x_a, y_a) and (x_b, y_b) are the 2D screen positions of the two character heads respectively, and θ and ϕ are two parametric angles in the Toric representation [34], as shown in Figure 4. Notably, given the 3D head positions of the two characters and our camera pose representation, we can reconstruct the camera pose in the 3D Cartesian space [8], and use it to composite the two characters on the screen. Moreover, the Toric camera representation is defined based on two targets, but real 3D animations can contain an arbitrary number of characters. Therefore, for a scene with more than two characters, the user is asked to select two of them as the targets, and in case of a single-character scene, we circumvent it by introducing a dummy "observer" character and treating its head as the camera pointing to the actual character in the scene. This method aligns with the "shot and reverse shot" technique [38], commonly employed in dialogue scenes between two characters, which shows each speaker from the perspective of the other. We find this works well in practice.

We further quantize our continuous camera pose representation uniformly into K discrete bins. A camera pose sequence is now represented as a sequence of discrete tokens: $(\mathbf{c}_{<BOS>}, \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_T)$, where $\mathbf{c}_{<BOS>}$ is a special token denoting the start of the sequence. $\mathbf{c}_t \in [0,1]^K$ is now the one-hot vector of t-th camera pose. We train our model to predict the sequence auto-regressively based on the inputs for conditional camera pose generation.

5.3 Network Architecture and Training

As shown in Figure 5, our model consists of two main components: 1) a keyframe feature extractor that learns a feature vector for each keyframe, which captures its local context information; 2) a camera pose generator that synthesizes camera pose sequences based on the extracted keyframe features and input semantics.

Keyframe feature extractor. As shown in Figure 5 (a), this component first uses a character feature extractor $f_{character}$ to compute a feature vector for each input animation frame, which characterizes the 3D spatial relation between the two characters in the frame. Similar to the cinematic character features in [27], the vector is composed of three quantities $(d_{a,b}, \theta_a, \theta_b)$, where $d_{a,b}$ is the 3D distance between the character heads, and θ_a (θ_b) is the angle between the line linking character heads and the line linking the shoulders of character a (character b). Then, for each keyframe (outlined in red in Figure 5 (a)), we refine its feature vector by training a temporal feature extractor $f_{temporal}$. The input to $f_{temporal}$ includes character feature vectors for a local window of 5 frames centered at the keyframe. It uses a stack of 1D convolutions along the temporal axis to process the inputs, producing a 1123-dimensional refined feature vector for the keyframe, which augments its original character features with its local context information. The final output of this component is a sequence of refined keyframe feature vectors $(\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_T)$, which will be used to condition the camera pose generator.

Camera pose generator. As shown in Figure 5 (b), this component is implemented with a decoder-only Transformer to generate a sequence of camera poses in an auto-regressive manner. We chose the Transformer architecture due to its great success in various sequential modelling tasks, which is attributed to the strong performance of its underlying self-attention mechanism on modelling long-term dependencies. In our context, it can learn to capture temporal dependencies among camera poses at different steps, which is crucial to ensuring the coherency of generated sequences. To predict a camera pose at step t, it generates an output embedding $\hat{\mathbf{e}}_t$ conditioned on the sequence of previous embeddings $(\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_{t-1})$ as well as the embeddings of the input semantics $\mathbf{e}_g$, $\mathbf{e}_s$. To obtain $\mathbf{e}_i, i \in \{0, 1, \ldots, t-1\}$, we encode the generated camera pose at step i with an encoder E_c, concatenate the resulting vector with the keyframe feature vector $\mathbf{\Phi}_{t+1}$, and feed it into a fusion module g_{fuse}. Both E_c and g_{fuse} are implemented with MLPs. $\mathbf{e}_g$ and $\mathbf{e}_s$ are obtained by passing the one-hot encodings of the input genre and intended emotion through two semantic encoders E_g and E_e, which are also MLPs. Then, the output embedding $\hat{\mathbf{e}}_i$ is fed into a camera pose decoder D_c, giving a probability distribution $\mathbf{c}_t$ over K quantized classes. D_c is a MLP ending with a softmax layer to predict class probabilities. We additionally predict the probability $\hat{b}_t$ of the current step being a shot boundary from $\hat{\mathbf{e}}_i$ using another decoder D_b, which is a MLP with a sigmoid layer at the end.

During training, the keyframe feature extractor and camera pose generator are trained jointly on a dataset $\mathcal{D} = (\mathbf{v}_{1:N}, \mathbf{c}_{1:N}, \mathbf{h}, \mathbf{m})$, which will be detailed in the next section. $\mathbf{v}_{1:N}$ is a 3D animation with N frames and $\mathbf{c}_{1:N}$ are the camera poses for each frame. $\mathbf{h}$ are the corresponding semantic labels, and $\mathbf{m}$ is a binary mask vector indicating if a frame is a shot boundary. We employ a cross-entropy loss function for both camera pose predictions and shot boundary predictions. Our ultimate objective is to optimize a combination of these two losses using the Adam optimizer with a learning rate of 0.001. We train our model for 98 epochs. For each training animation sequence, we randomly select keyframes at random time points while ensuring the interval between two consecutive keyframes lies between 1 and 5 seconds. This allows our model to better generalize to handle different keyframe sampling intervals that may be encountered at test time.

5.4 Generation

Once trained, our model can be used to generate camera poses for a given animation and input semantics in different ways to support the functionalities of Cinemassist. First, the auto-regressive nature of our model makes it naturally support interactive generation through several iterations. In particular, at each iteration, our model predicts a distribution over all possible camera poses, and a set of samples can be drawn from the distribution as suggested options ranked by their likelihood. The user then can select a desired option, and our model takes as input the selected option to make a prediction for the next iteration. This enables the frame-level design exploration functionality of Cinemassist on the design panel. Second, we also sample high-confidence sequences of camera poses (shown in Figure 6 with annotations of distinct camera trajectories) from our model via beam search without any user intervention, which supports our scene-level design exploration on the storyboard panel. Notably, we adopted a beam size of 15 that yields the best results compared to other beam size options. Third, our model can also accept a partial sequence as input and auto-complete the remaining part. This makes it possible to export the sequence being designed on the design panel into the storyboard panel and automatically generate suggestions on the complete design sequence, as introduced in Section 4. To improve the "creativity" of our model, we divide the inputs of the softmax function at the final layer of the camera pose decoder by a temperature parameter (1.0 in our implementation), to make the sampling more stochastic.

6 OUR DATASET

To train our model, we construct a dataset from 18 movies belonging to three different genre categories including romance, action and thriller. The movies are carefully selected so that they receive high IMDb ratings and span a wide range of cinematography styles. For each movie, we divide it into scenes using the scene boundary annotations provided by MovieNet [26], and further partition each scene into a sequence of shots using MovieNet's shot-detector tool. As shown in Figure 7, for each scene with N frames, we extract the 3D poses of characters in each frame using Metrabs [47]. Notably,

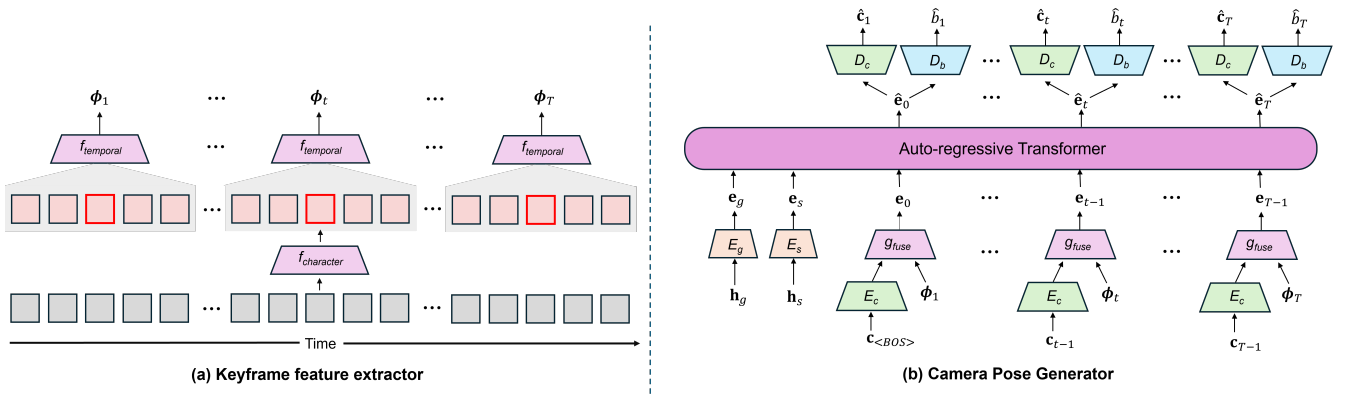

Figure 5: Overview of our model.

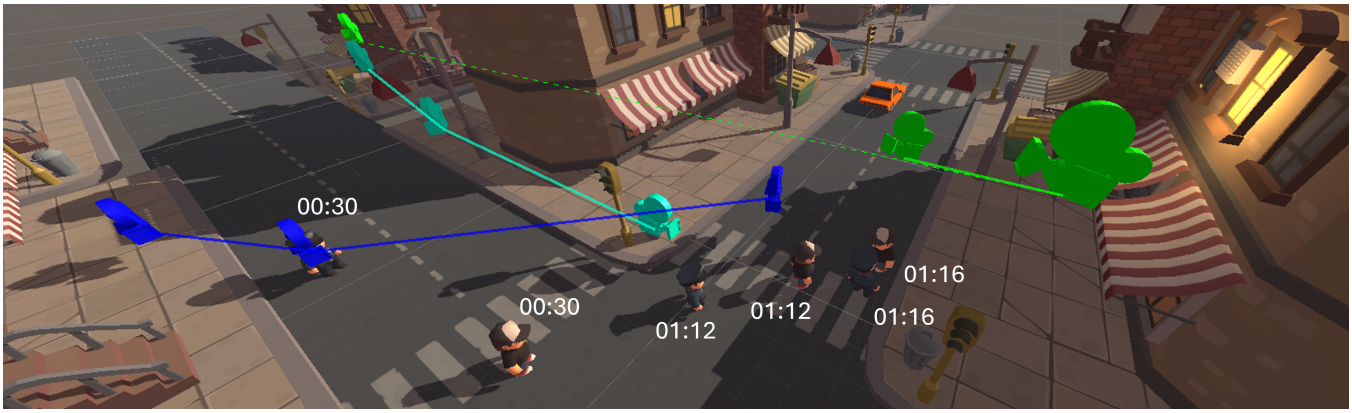

Figure 6: Our deep generative model can generate diverse sequences of camera poses to convey the story of an animation. The three sequences generated for an animation of two characters in this example are visualized as distinct camera trajectories coloured in blue, cyan and green in the 3D scene. Each trajectory consists of three camera poses generated for three keyframes where the two characters belonging to the same keyframe are labelled with their corresponding timestamps. The solid lines on a trajectory indicate the transition of the same camera within a shot and the dashed lines represent the jump between the cameras of different shots.

Metrabs has the capability of estimating the full-body 3D poses for all the people in a frame even if they are partially occluded and clipped on the screen. This gives rise to a 3D animation $\mathbf{v}_{1:N}$ of the characters in the scene. Then, we utilize the 3D poses of the characters in each frame to calculate its Toric camera parameters [34], resulting in a sequence of camera pose annotations $\mathbf{c}_{1:N}$. It is worth noting that the Toric camera representation assumes two targets are presented in the scene. When a frame contains more than two characters, we select the largest two characters on the screen as targets according to Hitchcock's cinematography principles [22], and for a frame with a single character, we introduce a dummy observer as discussed in Section 5.2. To obtain the genre label of a scene, we use its movie genre on IMDb. We also estimate the scene's emotion label, by employing a text-to-emotion model [2] to predict a probability distribution over the five basic emotion categories and assign the most likely category to the scene. This leads to the semantic labels $\mathbf{h}$. Further, we assign each frame with a label indicating if it is a shot boundary, producing a shot boundary mask $\mathbf{m}$. In this way, for each scene, we can obtain a sample of the form: $(\mathbf{v}_{1:N}, \mathbf{c}_{1:N}, \mathbf{h}, \mathbf{m})$. By filtering out scenes with no characters or more than 10 characters, we end up with our dataset of 977 samples and split it into 70% training and 30% test sets. Please refer to the supplementary materials for a detailed summary of our dataset.

7 EVALUATION

In this section, we conduct a user study with representative end-users of different backgrounds, followed by another expert rating study assessing the users' design outcomes in the initial study. Our main aim is to evaluate whether and how our system can facilitate the user's design process and influence the design outcome.

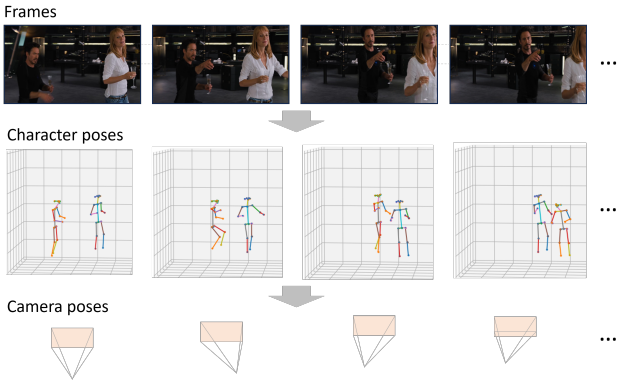

Figure 7: 3D camera pose estimation on our dataset. For each scene frame (top), the full-body 3D poses of the characters of interest are predicted (middle), and the Toric camera parameters are computed from the character poses, yielding the camera pose estimated from the scene (bottom).

7.1 Evaluation with Users of Different Backgrounds

Our system's usefulness may vary with the end-users' heterogeneous backgrounds. Particularly, novice users — who have limited 3D digital filmmaking experience — could find it difficult to adopt machinima tools to produce cinematic content. In contrast, non-novice users who own 3D digital filmmaking experience often use machinima filmmaking tools in their practical work or study purposes but could face distinct design challenges. Recognizing these potential disparities among our target end-users, we recruited 18 participants of diverse backgrounds for this study, all of whom have experience in using prominent 3D digital filmmaking tools such as Unity, Unreal Engine, Autodesk 3ds Max and Maya. 9 of the 18 participants have 3+ years of experience in digital filmmaking, while the relevant experience of the rest ranges from 1 to 3 years. Moreover, it should be noted that the participants have diverse backgrounds —almost half of them are animation professionals (majoring in animation or working on animations in digital gaming or filming industry) and the remaining ones are studying in or graduated from design schools, without expertise on animation. The detailed background and experience of all the participants are provided in the supplementary materials.

The participants were tasked with designing cinematic compositions on Unity Timeline for a 3D animation, which lasted for 72 seconds and was planned within a virtual scene in Unity by a digital game expert using open-source assets from the Unity Asset Store. The story content within the animation follows the three phases of "Hero's Journey" [11], a classical story pattern widely adopted in digital game design and filmmaking.

7.1.1 Procedure.

Prior to the study, we read the study orientation to each participant, introducing the walk-through of the study. Further, we conducted a tutorial session to familiarize each participant with cinematic composition design on Unity Timeline by having them watch a short tutorial video of using Unity for cinematic composition design and finish a warm-up exercise using a demonstration animation.

Task session. The participants were given an animation and asked to perform two tasks of cinematic composition design on Unity Timeline for the animation, one without our system and one with our system. The order of the two tasks was randomized for each participant. No time limitation was set for either task to allow participants to complete the tasks to the best of their abilities. Moreover, we did not pre-specify genre types and intended emotion states that the participants are expected to express with composition designs, and instead offer them sufficient freedom to determine what specific message they wish to impart based on their own understanding of the storyline in the animation. This helps leave more room for the participants to unleash their creativity in the tasks. During the design process with our system, they are free to experiment with different settings of the input semantics to explore distinct camera behaviours. As the deliverable for each task, the participant was required to save the outcome composition sequences as "storyboards", each comprising a sequence of distinct camera viewpoints on the animation, and export the corresponding videos. To generate the video of a composition sequence, we interpolate the camera movement between two consecutive keyframes using linear interpolation and render the video at FPS 24 using Unity Recorder. Moreover, the participants' design workflow for the two tasks was all screen-recorded. We also counted the completion time of a task for each individual.

Post-task questionnaire session. After finishing a task, the participant was required to give self-confidence scores towards the task's design outcome using a 5-point Likert score and leave their scoring reason. Specifically, the confidence questions here not only assess the overall quality of the design outcome but also its novelty and appropriateness, which are widely agreed as the nature of "creativity" [46]. In addition, we invited each participant to evaluate the functionality and usability of Cinemassist using a 5-point Likert scale (1: strongly agree, 5: strongly disagree) and leave the scoring rationale as well. Significantly, questions 1 to 4 evaluate Cinemassist's two functionalities in facilitating the participants' divergent thinking and convergent thinking which are known as two integral aspects of the creative process [18, 24, 45]. The participants further report their satisfaction towards Cinemassist's usability through questions 5 to 7 that follow the "System Usability Scale (SUS) standard" [7].

Reflective interview session. Upon completing both tasks and their corresponding questionnaires, we conducted semi-structured reflective interviews with the participants. The interview questions were formulated based on our observations of the participants' design processes and aimed at eliciting feedback for improving the functionality and usability of our system. With the consent of each participant, the interview session was recorded in audio for our later analysis.

7.1.2 Results.

In Figure 8, we illustrate the self-assessed confidence levels of participants in the questionnaire session concerning their design quality for the design tasks. The data uncovers a pronounced trend in the participants' perceptions of "overall quality". While over 75% of participants were confident in their designs from the task without

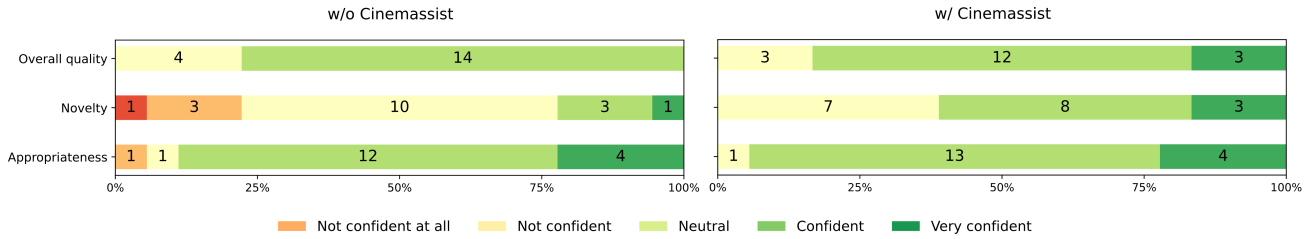

Figure 8: Self-confidence scores that the participants report after designing cinematic compositions without our Cinemassist system (left) and with our system (right) in the user study.

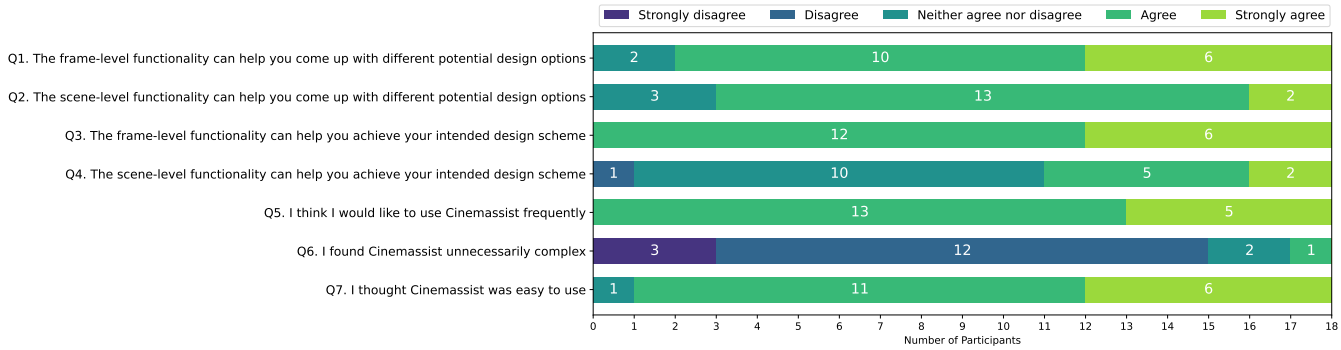

Figure 9: Self-satisfactory score distributions of the participants in the user study for each question regarding the usefulness and usability of our system.

our system, none reported being "very confident". Post the task with our system, a similar proportion of participants retained their confidence, and three individuals upgraded their self-evaluation to "very confident". P11 explained: "*When compositing the shots, I employed more professional angles, enhancing the narrative.*" P13 agreed: "*The quality seems to be guaranteed by AI, making my design outcome resembles real film.*" Conversely, self-confidence scores regarding the "novelty" of the design outcomes showed a significant shift. After the task without our system, less than 25% felt "confident" or "very confident," with the majority indicating only "neutral" confidence. Additionally, 25% expressed low confidence levels. In contrast, following the task with our system, the proportion of those feeling "confident" or "very confident" rose to over 50%, with no instances reported low confidence. P13 highlighted: "*I achieved some breakthrough to express the scene content.*" P16 commented: "*Some unexpected shots make the outcome more creative.*" The "appropriateness" of the designs showed minimal variation in confidence levels between the tasks, indicating a subtle effect on participants' self-perception in this aspect.

Figure 9 summarises the participants' assessment towards the usefulness of the Cinemassist's two functionalities in addition to its usability. Specifically, more than 80% of the participants either agreed or strongly agreed that both the frame-level and scene-level functionalities can help them come up with different potential design options. Particularly, in terms of the frame-level function, P2 supplemented: "*When informed with frame-level cinematic composition design alternatives, my original "serial" ideation workflow shifted to a "parallel" workflow.*" P4 added: "*The recommendations freed me from the constraints of traditional exemplars.*" Concerning the scene-level function, P5 assessed: "*The recommended sequences are of above-average quality, with no glaring errors, and any that may exist can be easily corrected.*" Similarly, P7 commented: "*The continuity of the sequences appears to be good already, requiring little to no manual adjustment.*" In contrast, although all participants agreed or strongly agreed that the frame-level function can help them achieve their intended design scheme, less than 40% of them still held their agreement when it comes to the scene-level function. For instance, regarding the frame-level function, Participant P10 remarked: "*There are consistently recommended alternatives close to my envisioned concept, allowing me to apply them directly or fine-tune them to achieve the desired result immediately.*" However, when turning to the scene-level function, P10 observed: "*None of the recommended sequences closely matched my original concept in an overall sense.*" P5 also acknowledged: "*While portions of the sequence may be close to what I envisioned, I still found myself dedicating time to revise the remaining frames to align with my initial concept.*" As for the usability aspects, all participants felt they would like to use Cinemassist frequently in their working context. More than 80% of them disagree that Cinemassist is unnecessarily complex and more than 90% of them corroborated that our system is easy to use.

Notably, we observed some participants had spent a longer completion time for the task with our system than for the task without our system. We asked them to give their reasons to explain this phenomenon in the reflective interview session. Significantly, P11 explained: "*As the two functions of Cinemassist had saved my time manually placing the camera to potential options, I turned to spend more time exploring and evaluating whose composition looks best and how the story can be best featured by these options in different ways.*" When asked about any suggestions to iterate the current system and improve its usability, the participants commented on three aspects. Firstly, the scene-level function of our system should recommend more diverse and easy-to-edit composition sequences. P10 addressed this aspect and suggested: "*Can the scene-level function enable the user to directly edit the camera trajectory of the recommended composition sequence as like how they adjust the "inspirational bulbs" at frame-level?*" Secondly, the semantic input needs to be enhanced beyond the existing "genre" and "intended emotion" parameters to include additional multimodal semantic specifications such as intended directorial style, film script, and accompanying background music or voiceover. Finally, the system should iterate to facilitate or automate the task of adjusting "target objects" over time to match the narrative's development within the 3D animation. P8 complained: "*I tend to forget to shift the target objects as the story unfolds, making the system still focus on the wrong objects when giving recommendations. Can the system remind me to do this or help detect the right targets?*" P8 also suggested: "*Sometime, I just want to focus on capturing only one object or other time taking care of multiple ones at the same time. Can the system expand its capabilities in this aspect beyond the fixed "two-targets" settings?*"

7.2 Expert Rating Study

We further evaluate the quality of the resulting composition designs from the previous user study as well as automatic composition outcomes generated by our system, by inviting two cinematography experts to rate the designs. The two experts are professors of a digital filmmaking school who have more than 10 years of experience in digital animation film production and instruction. We generate the automatic results by selecting keyframes at uniform time intervals and randomizing the selection of targeting objects and the input semantics to our model.

7.2.1 Procedure.

The experts were first provided with the story script of the animation and then required to assess the 18 participants' design works. When assessing a participant's performance, three design works were shown to an expert in randomized order: the participant's two cinematic composition sequences designed without and with using our system, in addition to an automatic sequence automatically generated by our system. The assessment process required experts to rank the sequences as "1st," "2nd," and "3rd" based on their merit, corresponding to the order presented on the slides across two criteria: continuity and novelty which are known as two key aspects of storyboarding [30]. Additionally, we also invited the experts to rank the three sequences based on their overall quality.

7.2.2 Results.

Based on the assessment results provided by the two experts, it was evident that their evaluations varied considerably between participants with a background in animation and those without such a background. Particularly, we categorized participants as having an animation background if they were professionals in the field of animation filmmaking or if they were students majoring in this discipline. As a result, participants P1 through P9 were classified as having an animation background, while participants P10 through P18 were classified as not having such background.

The rating results are shown in Figure 10. We can see using Cinemassist can lead to remarkably improved performance for the participants with animation backgrounds. The two experts ranked the design outcomes of the task with our system (human + Cinemassist) in the first place nearly 75% of the time across the three criteria. In contrast, the design outcomes from the task without our system (human) were predominantly placed in the second. Furthermore, the automatic results from Cinemassist were generally ranked in the last place. The two experts consistently identified the main flaw in the automatically generated sequences as "focusing on the wrong characters" during the progression of the story in the animation. This is mainly because of the content-agnostic and random setting of the inputs to our system. We expect better results can be achieved with the automatic method by using more sophisticated keyframe selection algorithms and powerful emotion predictors from input 3D animations. Figure 11 shows a visual comparison of two outcomes created by P7 without and with our system, respectively, and an automatically generated work. In this comparison, both experts ranked the outcome produced with our system (human + Cinemassist) in the first place for its continuity, novelty, and overall quality. Notably, one expert commented: "*The composition of this work (human + Cinemassist) appears novel.*" Additionally, the second expert remarked: "The other work (human) seems conical and lacks creativity." Furthermore, the two experts ranked the automatically generated work (Cinemassist) in the last place concerning continuity and overall quality. One expert elaborated: "I did not observe the appearance of the two zombies when expected" indicating that the storyboard failed to focus on the intended targets.

We also observe that Cinemassist does not bring performance gain to the non-animation background participants. In particular, most of their designs of the task without our system are ranked first place while those of the task with our system are frequently ranked second place and Cinemassist's automatic results are almost always third place. One expert commented on a sequence designed by P13 in the task with our system: "*This work is not clear on what audio-visual language in film creation is, therefore there is a misuse of shot types.*" We posit that this phenomenon may occur because users with non-animation background, who lack essential knowledge of film editing, might struggle to correctly integrate the suggested compositions into a coherent sequence for a novel yet accurate expression.

Consequently, we find although most participants in our initial user study felt more confident regarding their design outcomes for the task supported by our system, the following expert rating study offered two distinct types of assessment results. The first type supported that our system advanced animation background users' design outcomes while the second didn't provide comparable support for our system's usefulness to non-animation users.

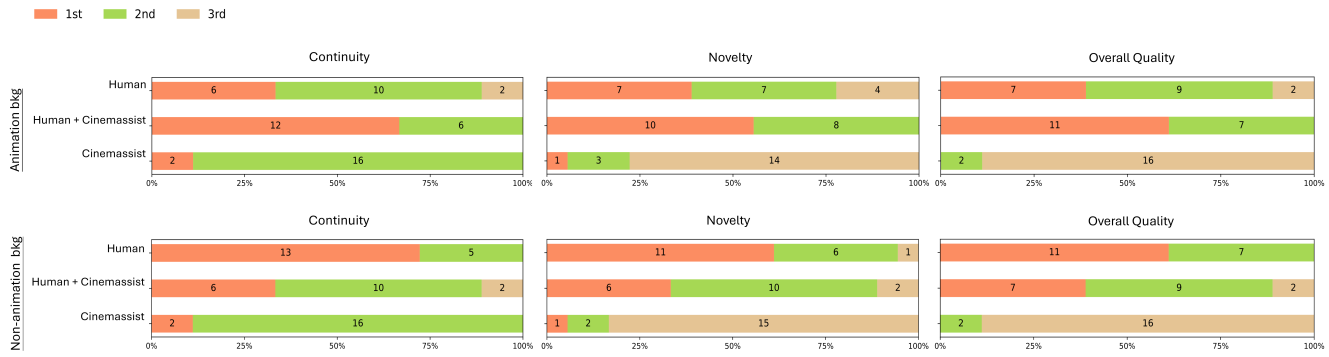

Figure 10: Results of the expert rating study: the preference ranking of the composition sequences created by the users in Section 7.1 using the standard Unity functionality (Human) and our system (Human + Cinemassist) along with those generated by our system automatically (Cinemassist). The rating is performed on three criteria, i.e., the continuity, novelty and overall quality of the composition sequences. The ranking results are categorized based on the animation expertise of the composition sequence creators, distinguishing between users with an animation background and those without, presented as top and bottom rankings respectively.

Interestingly, among all the rating results, two exceptional automatic results still achieved first place concerning their composition novelty, as is shown in Figure 10.

8 DISCUSSION

8.1 Data-driven System vs Rule-based System

Cinemassist is designed as a ready-to-use creativity support tool for enhancing cinematic composition in machinima environments. Our user study results indicate that this tool, which utilizes a data-driven approach, offers a variety of design options at both keyframe and scene levels, significantly improving the novelty and overall quality of user-generated designs. By contrast, autonomous or interactive cinematography systems that rely on predefined rules for machinima filmmaking tend to provide only standard design proposals. These systems lack awareness of the design context, which greatly limits their ability to enhance the creativity or efficiency of the design process. While some systems attempt to introduce variability and novelty by incorporating randomized noise into their rule-based recommendations, this often results in a loss of initial appropriateness, rendering them less suitable as design alternatives. The data-driven approach adopted by our system addresses these limitations effectively by exploiting a large number of creative examples to learn a deep generative model that considers both spatial and semantic context and can synthesize novel options beyond training examples. This methodology can produce more contextually appropriate and creatively stimulating design options. However, the effectiveness of the data-driven approach can be limited by the size of the training dataset. Introducing additional rules may potentially enhance the precision and relevance of the model's recommendations, ensuring robust and applicable designs across various scenarios.

8.2 Conventional design process vs AI-facilitated design process

Our user study results highlight that, compared to traditional methods, our system significantly enhances the design process by presenting users with various potential camera pose schemes at both keyframe and scene levels. User feedback suggests that this AI-assisted process not only reduces the time needed to realize design options but also provides more opportunities to explore and select the most optimal designs. As such, it would be interesting to investigate if and how AI-suggested design options may influence the original cognitive models of machinima filmmakers [13], specifically in terms of ideation, realization, exploration, and evaluation, which is left as future work.

Our findings also reveal that the system's effectiveness depends on the user's background. Particularly, there is a significant improvement in the design outcomes of users with animation background with the help of our system. However, for those non-animation background users, our system can not help improve their performance according to the expert ratings, although they became more confident about their results when facilitated by our system. This may be related to the confidence-correctness mismatch in AI-assisted decision-making scenarios [37]. One possible reason is that to better support creativity, we diversify outputs from our model via random sampling, leading to some less likely suggestions that are inspiring but may not be properly used by novices with limited understanding of cinematic composition. To aid novices more effectively, we could customize our system to offer only a few high-probability options, to increase their chance of achieving high-quality results.

8.3 Creativity Support Tools Design Implications

Based on the results of our study, we also draw four design implications as references for the design of related creativity support

Figure 11: Example cinematic composition sequences created by a participant without the support of our system (Human) and with the assistance of our system (Human + Cinemassist), along with the composition sequence generated automatically by our system (Cinemassist). The right arrows displayed in-between two compositions indicate shot boundary.

tools for design tasks within 3D real-time environment. First, displaying the recommended design scheme outcomes could facilitate the designer's "divergent thinking" by enabling instant evaluation (**I1**). Secondly, visualizing the recommended designs as "reminders" in the design space of machinima filmmaking — the 3D scene — could facilitate the designer's "convergent thinking" by fostering further "in-situ" exploration and adjustment (**I2**). Thirdly, enabling the recommended design to transit between the above two presenting modes instantly could jointly facilitate the designer's creative process by facilitating the transition between divergent thinking and convergent thinking as "two ends of the cognitive continuum" [18] (**I3**). Lastly, given the two types of user needs we observed in our user study, we emphasize the importance of considering "levels of control" (**I4**).

9 LIMITATIONS & IMPROVEMENT AREAS

We acknowledge several limitations in the design and implementation of our system. Primarily, our system is tailored to a particular filmmaking scenario, wherein the 3D scene and animation are pre-arranged in a real-time environment. In practice, however, within the conventional 2D/3D animation filmmaking framework, the crafting of cinematic composition might occur before or concurrently with the creation of the animation's visual and auditory elements at the story-boarding stage. As such, our future work will extend our system to support the more conventional filmmaking workflow. Secondly, we also realize that a scene's genre type and intended emotion type might be too simplistic to fully encapsulate the scene's semantic context. In our future iteration of the generative model, we plan to try extending the semantic input with story script, voice-over, or background music. Besides, we will also try training our model conditioned on different director styles, enabling the user to acquire cinematic composition proposals that follow an intended director style. Thirdly, we observe that the appropriateness of our system's cinematic composition suggestions at the frame and scene level could be significantly influenced by two appropriately assigned target objects at a given time which frequently changes as the animation's narrative progresses. Given this issue,

our system's future iteration will consider automatically recognizing appropriate targets at a specific time point besides keeping the user's discretion in assigning them. In addition, our current system assumes an animation at one time only focuses on the interaction between two target objects, which is not always true for "solo" and "multiple" target cases. As such, future advancements in our system should also fit one or multiple target cases to our current camera pose representation and enlarge our current training dataset with such scenes.

10 CONCLUSION & FUTURE WORK

In summary, this work proposes a creativity support tool for facilitating cinematic composition in 3D digital filmmaking within a real-time environment. Based on a Transformer-based auto-regressive model trained on movie data, our model can recommend various potential cinematic composition design options at both keyframe and scene levels that users can refer to and incorporate into their design workflows to achieve more creative design outcomes. We conducted a user study followed by an expert rating study, whose results suggest that our tool can not only facilitate the users' design workflow but also advance their design outcomes in various ways. We also draw several design implications from building the system, which we believe can inform future trends in developing such creativity support tools for design tasks within 3D real-time environments. In the immediate future, we plan to further analyse the user study's qualitative data as well as the task session video recordings, with the goal of understanding more deeply the interaction between users and the system and users' experiential views. Our ultimate goal is to draw a design framework for building designer-centred CST that harnesses AI to augment designers' creativity in an interactive, iterative workflow.

ACKNOWLEDGMENTS

We would like to express our gratitude to Dr. Junnan Yu and Dr. Jae-Eun Oh for their invaluable suggestions on the evaluation of our system. We are also grateful to the eighteen participants and two experts who contributed their time and insights to our two evaluative studies. Additionally, we appreciate the reviewers for their thoughtful and constructive feedback on an earlier version of this paper. Lastly, our thanks extend to The Hong Kong Polytechnic University and its School of Design for their generous support, as well as the financial assistance provided by the Research Studentship funded by the University Grants Committee of Hong Kong.

REFERENCES

[1] H Porter Abbott. 2002. *The Cambridge introduction to narrative.* Cambridge University Press.
[2] Shivam Sharma Karan Bilakhiya Aman Gupta, Amey Band. 2020. *text2emotion.* https://pypi.org/project/text2emotion/
[3] Hideaki Anno. 2021. Evangelion: 3.0+1.01 Thrice Upon a Time [Film]. *IMDb* (2021).
[4] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic Editing of Footage from Multiple Social Cameras. *ACM Trans. Graph.* 33, 4, Article 81 (jul 2014), 11 pages. https://doi.org/10.1145/2601097.2601198
[5] Daniel Arijon. 1991. *Grammar of the film language.* Silman-James Press.
[6] J. Aumont, University of Texas Press, A. Bergala, M. Marie, R. Neupert, and M. Vernet. 1992. *Aesthetics of Film.* University of Texas Press. https://books.google.com.hk/books?id=nmpZAAAAMAAJ
[7] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.
[8] Marcel Berger. 2009. *Geometry i.* Springer Science & Business Media.
[9] Nathalie Bonnardel. 2000. Towards understanding and supporting creativity in design: analogies in a constrained cognitive environment. *Knowledge-Based Systems* 13, 7-8 (2000), 505–513.
[10] Christopher Bowen. 2013. *Grammar of the Shot.* Routledge.
[11] Joseph Campbell. 2008. *The hero with a thousand faces.* Vol. 17. New World Library.
[12] Marc Christie, Patrick Olivier, and Jean-Marie Normand. 2008. Camera control in computer graphics. In *Computer Graphics Forum*, Vol. 27. Wiley Online Library, 2197–2218.
[13] Nicholas Davis, Boyang Li, Brian O'Neill, Mark Riedl, and Michael Nitsche. 2011. Distributed Creative Cognition in Digital Filmmaking. In *Proceedings of the 8th ACM Conference on Creativity and Cognition* (Atlanta, Georgia, USA) *(C&C '11)*. Association for Computing Machinery, New York, NY, USA, 207–216. https://doi.org/10.1145/2069618.2069654
[14] Nicholas Davis, Alexander Zook, Brian O'Neill, Brandon Headrick, Mark Riedl, Ashton Grosz, and Michael Nitsche. 2013. Creativity support for novice digital filmmaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 651–660.
[15] Edirlei E. S. de Lima, Cesar T. Pozzer, Marcos C. d'Ornellas, Angelo E. M. Ciarlini, Bruno Feijó, and Antonio L. Furtado. 2009. Virtual Cinematography Director for Interactive Storytelling. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology* (Athens, Greece) *(ACE '09)*. Association for Computing Machinery, New York, NY, USA, 263–270. https://doi.org/10.1145/1690388.1690432
[16] Will Eisner. 2008. Graphic storytelling and visual narrative: Principles and practices from the legendary cartoonist (Rev. ed). *NY: Norton.(Original work published 1996)* (2008).
[17] Inan Evin, Perttu Hämäläinen, and Christian Guckelsberger. 2022. 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.
[18] HJ Eysenck. 2003. Creativity, personality and the convergent-divergent continuum. (2003).
[19] Gerhard Fischer. 2004. Social creativity: turning barriers into opportunities for collaborative design. In *Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices-Volume 1*. 152–161.
[20] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–18.
[21] Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2018. Twenty years of creativity research in human-computer interaction: Current state and future directions. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 1235–1257.
[22] Quentin Galvane and Rémi Ronfard. 2017. Implementing hitchcock-the role of focalization and viewpoint. In *Eurographics Workshop on Intelligent Cinematography and Editing*. The Eurographics Association.
[23] Francis Glebas. 2012. *Directing the story: professional storytelling and storyboarding techniques for live action and animation.* Routledge.
[24] Joy Paul Guilford. 1967. The nature of human intelligence. (1967).
[25] Li-wei He, Michael F Cohen, and David H Salesin. 1996. The virtual cinematographer: A paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 217–224.
[26] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 709–727.
[27] David G Jansson and Steven M Smith. 1991. Design fixation. *Design studies* 12, 1 (1991), 3–11.
[28] Arnav Jhala and R Michael Young. 2011. Intelligent machinima generation for visual storytelling. In *Artificial Intelligence for Computer Games*. Springer, 151–170.
[29] Hongda Jiang, Bin Wang, Xi Wang, Marc Christie, and Baoquan Chen. 2020. Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 45–1.
[30] Steven Douglas Katz. 1991. *Film directing shot by shot: visualizing from concept to screen.* Gulf Professional Publishing.
[31] Arthur Koestler. 1964. The act of creation. (1964).
[32] Aki Kubota. 2021. Hideaki Anno: The Final Challenge of Evangelion [Film]. *NHK* (2021).
[33] Nicole E Lemon. 2012. *Previsualization in Computer Animated Filmmaking.* Ph. D. Dissertation. The Ohio State University.
[34] Christophe Lino and Marc Christie. 2015. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–12.
[35] Christophe Lino, Marc Christie, Fabrice Lamarche, Schofield Guy, and Patrick Olivier. 2010. A real-time cinematography system for interactive 3d environments.

In *SCA'10 Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 139–148.
[36] Henry Lowood and Michael Nitsche. 2011. *The machinima reader*. MIT Press.
[37] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 840, 20 pages. https://doi.org/10.1145/3613904.3642671
[38] James Mairata, Mairata, and Aboujieb. 2018. *Steven Spielberg's Style by Stealth*. Springer.
[39] Paul Marino. 2004. *3D game-based filmmaking: The art of machinima*. Paraglyph Press.
[40] Marcos Mateu-Mestre and Jeffrey Katzenberg. 2010. *Framed ink: Drawing and composition for visual storytellers*. Design Studio Press.
[41] R. McKee. 1997. *Story: Substance, Structure, Style and the Principles of Screenwriting*. HarperCollins Publishers.
[42] Manshad Abbasi Mohsin and Anatoly Beltiukov. 2019. Summarizing emotions from text using Plutchik's wheel of emotions. In *7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)*. Atlantis Press, 291–294.
[43] K. L. Bhanu Moorthy, Moneish Kumar, Ramanathan Subramanian, and Vineet Gandhi. 2020. GAZED– Gaze-Guided Cinematic Editing of Wide-Angle Monocular Video Recordings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376544
[44] Kumiyo Nakakoji. 2006. Meanings of tools, support, and uses for creative design processes. In *International design research symposium*, Vol. 6. 156–165.
[45] Mark A Runco. 2014. Creativity theories and themes: research, development, and practice. (2014).
[46] Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity research journal* 24, 1 (2012), 92–96.
[47] István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. 2020. Metrabs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 1 (2020), 16–30.
[48] Ben Shneiderman. 2002. Creativity support tools. *Commun. ACM* 45, 10 (2002), 116–120.
[49] T. Sobchack and V.C. Sobchack. 1987. *An Introduction to Film*. Little, Brown. https://books.google.com.hk/books?id=SwMbAQAAIAAJ
[50] EDP Symons. 1986. Edison's electric light. Biography of an invention.