

Research on Immersive Narrative Cognitive Mechanism of Animated Films Under Generative Media Environment

Meiling Zhou

*Nanjing University of Information Science and Technology, Nanjing, China
zhoumeiling02420@outlook.com*

Abstract. Generative artificial intelligence, virtual production, and intelligent image generation are reshaping the creation process and viewing experience of animated films. Animated narrative is gradually shifting from traditional linear expression to a new form that integrates algorithmic generation, audiovisual immersion, and audience cognitive participation. Focusing on the cognitive mechanism of immersive narrative in animated films under the generative media environment, this study constructs a technical pathway covering sample collection, multimodal feature extraction, audience cognition experiments, and predictive model analysis. Public animation clips, AI-assisted animated shorts, and traditional animated clips from 2018 to 2025 are selected as research samples. NLP, computer vision, audio feature analysis, eye-tracking, and deep learning models are used to measure visual attention, spatial presence, emotional arousal, character empathy, and narrative comprehension. The results show that the multimodal fusion model outperforms single text-based or visual models in immersion prediction. Visual density, shot rhythm, spatial depth, dialogue emotion, and character credibility are key factors influencing immersive narrative effects. The study provides an operational methodological framework for generative animation evaluation, intelligent visual analysis, and immersive media cognition research.

Keywords: Generative Media, Animated Film, Immersive Narrative, Cognitive Mechanism, Deep Learning

1. Introduction

Generative artificial intelligence, virtual production, and intelligent image generation are changing the production logic of animated films. Scene construction, character movement, visual style, and emotional atmosphere are increasingly shaped by algorithmic generation [1]. In this media environment, immersive narrative in animated films no longer depends only on plot development. It is also closely related to visual complexity, spatial hierarchy, audiovisual rhythm, and audience perception [2]. Existing studies have examined AI-based image generation, digital animation aesthetics, and immersive experience evaluation. However, the mechanism through which generative media affects visual attention, character empathy, emotional arousal, and narrative comprehension still requires a computable explanation [3]. This research combines media narrative analysis with computational methods. It uses natural language processing, computer vision, eye-

tracking, and deep learning models to extract multimodal features from animated films and examine the relationship between generative visual features and audience immersive cognitive responses.

2. Literature review

2.1. Generative media in animated film production

The research on generative media concentrates on the application of artificial intelligence in the field of text creation, image creation, motion extension, and scene extension. The essence of it is the transition from a manual process of creating animations into a joint effort made by people and artificial intelligence. Artists are able to utilize algorithm-based solutions for quick generation of storyboards, background designs, facial expressions, and style outlines [4]. The focus of this line of research lies in technical performance and visual generation potential, together with the connection between the created elements and the logic of the story. Animated movies do not consist of only static images. Motives of characters, transitions in the plot, and emotional structure of the movie will determine whether or not these images are suitable for continuous viewing [5].

2.2. Immersive narrative in audience cognition

Research on immersive storytelling tends to describe the psychological process of the immersion experience using the approaches of narrative transportation, spatial presence, and character identification. What matters to the researchers is the persistent condition of attention engagement, emotion engagement, and meaning construction in the course of watching [6]. The animated movies possess certain cognitive peculiarities different from those of live-action movies. Stylized characters, exaggerated movements, and fabricated spaces free one from realism and make the spectator depend more on visual and emotional symbols [7]. As such, in this type of media, immersive experience depends not only on storytelling but also on the perception of forms.

2.3. Computational analysis of narrative cognition

The application of computational narrative study to media analysis changes the texts, images, and responses from the audience into data items that can be acquired, modeled, and predicted. Natural Language Processing will help to identify the dialogue emotion, theme shift, and relations between characters, and the computer vision technology will be able to analyze color, camera movements, facial expressions, and space structures [8]. The multimodal deep learning technology takes the study of narrative cognition from comparison of one variable to feature fusion, and questionnaires, eye tracking, and emotions ratings become part of the same explanation framework [9]. This method in animated film studies is used to understand how generative visual features shape the immersive experience of the viewer.

3. Experimental methods

3.1. Animated film sample construction

Data collection is based on publicly available animated film clips, official trailers, open-licensed animated shorts, and public video subtitles from 2018 to 2025. This period covers changes before and after the wider entry of generative media into animation production. The samples are divided into a generative-media-feature group and a traditional animation control group. The generative-

media-feature group may include public AI animated shorts or AI-assisted visual clips, such as *The Frost*, *Critterz*, and selected experimental animation clips produced with AI assistance. The traditional animation control group may include clips from *Spider-Man: Across the Spider-Verse*, *Elemental*, *The Mitchells vs. the Machines*, and *Suzume*, all of which have clear narrative structures and mature audiovisual styles. Each clip is controlled within 90 to 120 seconds to reduce the influence of duration differences on fixation time and emotional ratings. Public sources include IMDb non-commercial datasets, TMDb film metadata, official YouTube trailer subtitles, open animation platforms, and public promotional clips. Finally, 96 clips are constructed, including 48 generative-media-feature samples and 48 traditional animation samples. Film title, release year, genre, subtitle text, shot sequence, color distribution, character movement, facial expression change, and background audio features are extracted to support multimodal modeling.

3.2. Multimodal feature extraction model

Multimodal feature extraction is conducted on the clip samples established in Section 3.1. Each animated clip is divided into textual, visual, and audio inputs. The textual layer uses BERT-based models to extract dialogue emotion, keyword density, and character relationships. The visual layer uses CNN and ViT models to extract camera movement, color saturation, spatial hierarchy, and character expressions [10]. The audio layer extracts loudness, rhythm variation, and emotional intensity. To reduce scale differences among modalities, attention weights are first calculated to estimate the contribution of each modality to immersive narrative reception. A unified multimodal feature representation is then generated, as shown in Formula (1).

$$\alpha_{m,i} = \frac{\exp(q^T \tanh(W_m x_{m,i}))}{\sum_{m \in \{v,t,a\}} \exp(q^T \tanh(W_m x_{m,i}))}, \quad z_i = \sum_{m \in \{v,t,a\}} \alpha_{m,i} W_m x_{m,i} \quad (1)$$

Here, $x_{v,i}$, $x_{t,i}$, and $x_{a,i}$ denote the visual, textual, and audio features of the i -th clip. $\alpha_{m,i}$ denotes the attention weight of the corresponding modality, and z_i represents the fused narrative feature. Then, z_i is input into the prediction model to calculate the error between predicted immersion scores and actual audience ratings. The training objective is shown in Formula (2).

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \lambda \|\theta\|_2^2 \quad (2)$$

Here, y_i denotes the actual audience immersion score, $\hat{y}_i$ denotes the predicted value, N denotes the number of clip-level samples, and $\lambda \|\theta\|_2^2$ controls model complexity. Formula (1) completes the fusion of textual, visual, and audio features, while Formula (2) constrains prediction error, enabling the model to analyze the relationship between generative visual features and audience cognitive responses.

3.3. Audience cognitive response experiment

The audience experiment will select 24 clips from the sample space created in Section 3.1. The two samples are made up of 12 clips in the generative-media-feature group and 12 clips in the traditional animation control group respectively. There are similar conditions between the two samples in terms of clip duration, theme types, number of characters, and completeness of the narratives. 120 young audiences aged between 18 and 30 are recruited for this experiment. All subjects are required to have normal or corrected-to-normal vision and hearing ability. Prior to the experiment, a

questionnaire is filled out concerning their viewing frequencies of animations, familiarity with AI-generated pictures, and the selected movies previously viewed. Nine-point eye-tracker calibration is conducted prior to the formal experiment. Those with fixations exceeding 0.5° need to be calibrated again. Clip presentation adopts Latin square design in order to alleviate potential order effect and fatigue effect. In clip viewing stage, average fixation durations, revisit frequencies, areas-of-interest ratio, and pupil changes are logged. The main areas of interests include characters' faces, action areas, background spatial layers, and important narrative elements. A seven-point scale survey on spatial presence, character empathy, emotional arousal, and narrative comprehension [11] is conducted immediately after viewing each clip. An interview with open-ended questions on visual complexity, character credibility, and narrative coherence is conducted after the experiment. The records with eye-tracking loss rate higher than 25% are excluded. Subjective scores, eye-tracking behaviors, and the multimodal features collected in Section 3.2 are aligned by the ID of clips to create a three-layer data set of film clip, multimodal feature, and audience cognitive responses.

4. Results

4.1. Multimodal feature performance in immersion prediction

In the pilot stage, data were collected from 120 participants. After removing eight records with excessive eye-tracking loss or abnormal questionnaire responses, 112 valid samples remained, producing 2,688 clip-level observations across 24 animated clips. Five-fold cross-validation is used to evaluate model performance, with Accuracy, Macro-F1, AUC, and RMSE as evaluation metrics. The results show that the text-only model has a certain explanatory capacity for narrative comprehension, but performs weakly in predicting spatial presence and visual attention. The vision-only model captures attention changes more effectively, but explains differences in emotional acceptance less well. The text-visual bimodal model reaches an Accuracy of 0.76 and an AUC of 0.82, indicating that cross-modal features improve predictive stability. After adding audio rhythm and emotional intensity, the multimodal fusion model reaches an Accuracy of 0.81, a Macro-F1 of 0.78, and an RMSE of 0.094. This indicates that joint modeling of visual, textual, and audio features predicts cognitive responses to immersive animated narrative more effectively. As shown in table 1.

Table 1. Performance comparison of multimodal immersion prediction models

Model Input	Accuracy	Macro-F1	AUC	RMSE
Text features only	0.67	0.63	0.71	0.143
Visual features only	0.70	0.66	0.75	0.132
Text + visual features	0.76	0.72	0.82	0.108
Text + visual + audio features	0.81	0.78	0.87	0.094

4.2. Cognitive path of immersive narrative reception

Cognitive path analysis shows that the influence of generative media features on immersive experience is concentrated in four stages: visual attention, emotional arousal, narrative comprehension, and character empathy. From the standardized coefficient matrix, visual density has a coefficient of 0.78 in visual attention, and shot rhythm has a coefficient of 0.72 in generative animation, which means that dense visuals and quick rhythm tend to lead to immediate perceptual entry. Spatial depth also has a coefficient of 0.73 in narrative comprehension. Credibility of

characters has a coefficient of 0.80 and 0.82 in narrative comprehension and character empathy respectively, implying that judgment on character logic and spatial relationship influences story comprehension. Linguistic emotion also has a coefficient of 0.71 and 0.74 in emotional arousal and character empathy, meaning that it is still a significant variable linking visual perception to character empathy. The average score of immersion of the generative-media-feature group is 5.62, which is higher than 5.18 in the traditional animation group. If the visual complexity score surpasses 6.20, the mean score of narrative comprehension reduces from 5.54 to 5.07. This implies that visual complexity increases cognitive load, as shown in Figure 1.

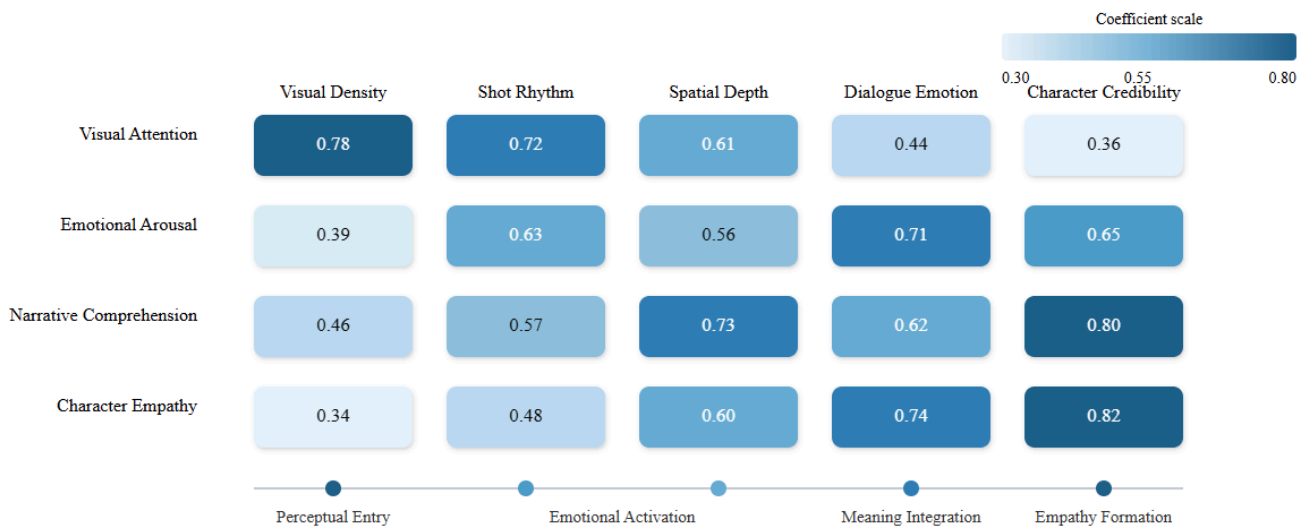


Figure 1. Standardized coefficient matrix of multimodal narrative features and cognitive responses

5. Discussion

The influence of generative media on immersive narrative in animated films is reflected not only in image generation efficiency and visual novelty, but also in the dynamic relationship among visual features, narrative structure, and audience cognitive responses. The experimental results show that visual density and shot rhythm can rapidly activate attention, spatial depth helps audiences understand event relations, while dialogue emotion and character credibility further affect empathy formation. Immersive experience depends on the continuous connection among visual attraction, emotional arousal, and narrative comprehension. When generated images become overly complex, audiences may experience higher cognitive load, leading to weaker narrative understanding. Therefore, generative animation creation needs to maintain a balance between technological generation capacity and narrative comprehensibility.

6. Conclusion

Focusing on the cognitive mechanism of immersive narrative in animated films under the generative media environment, this study establishes an analytical framework composed of public sample collection, multimodal feature extraction, audience experiments, and prediction modeling. The results show that the fusion of textual, visual, and audio features can more accurately predict audience immersion. Character credibility, spatial depth, and dialogue emotion are important variables connecting generative visual features with audience cognitive responses. Generative media provides animated films with new possibilities for visual expression and narrative organization, but

its effectiveness still depends on the coordination among visual style, story logic, and audience cognitive processing. The proposed method can support AI animation evaluation, immersive content design, and intelligent media analysis.

References

- [1] Epstein, Z., et al. (2023). Art and the science of generative AI. *Science*, 380(6650), 1110–1111.
- [2] Amankwah-Amoah, J., et al. (2024). The impending disruption of creative industries by generative AI: Opportunities, challenges, and research agenda. *International Journal of Information Management*, 79, 102759.
- [3] Bender, S. (2025). Generative-AI, the media industries, and the disappearance of human creative labour. *Media Practice and Education*, 26(2), 200–217.
- [4] Zhao, X., & Zhao, X. (2024). Application of generative artificial intelligence in film image production. *Computer-Aided Design & Applications*, 21(S27), 15–28.
- [5] Kim, J.-G. (2024). Current use and issues of generative AI in the film industry. *Journal of Information Technology Applications and Management*, 31(3), 181–192.
- [6] Peng, J., et al. (2025). Optimizing the conceptual design process of animation scenes using generative AI. *Frontiers in Computer Science*, 7, 1741768.
- [7] Rosati, E. (2025). The future of the movie industry in the wake of generative AI: A perspective under EU and UK copyright law. *Computer Law & Security Review*, 59, 106207.
- [8] Gupta, P., et al. (2024). Generative AI: A systematic review using topic modelling techniques. *Data and Information Management*, 8(2), 100066.
- [9] Nikulina, O., et al. (2024). Narrate, act, and resonate to tell a visual story: A systematic review of how images transport viewers. *Journal of Advertising*, 53(4), 605–625.
- [10] Ma, Z., et al. (2024). Present, empathetic, and persuaded: A meta-analytic comparison of storytelling in high versus low immersive mediated environments. *Human Communication Research*, 50(1), 27–38.
- [11] Bowman, N. D., et al. (2024). Immersive storytelling as a technology, a practice, and an experience. *Media and Communication*, 12.