

Enhancing Short-Film Visual Storytelling with AI-Generated Attention and Focus Maps: A Dual-Map Framework for Data-Driven Analysis

Abdunroni Samaeng  and Athitaya Somlok *

Department of Communication Arts, Faculty of Communication Sciences, Prince of Songkla University,
Pattani, Thailand

Email: abdunroni.s@psu.ac.th (A.S.); athitaya.s@psu.ac.th (A.So.)

*Corresponding author

Abstract—Visual storytelling in short films relies heavily on directing audience attention, yet traditional methods for analyzing visual saliency often lack cinematic context and temporal dynamics. We propose a dual-map framework merging Artificial Intelligence (AI)-generated attention heatmaps with focus maps to deliver data-driven insights for visual storytelling. A focus map is defined as a temporal representation of sustained viewer attention that captures the dynamic evolution of gaze across frames. The attention heatmap model, trained on cinematic datasets, identifies salient regions while the focus map model, which includes temporal attention gates, captures dynamic shifts in viewer focus. The framework employs spatial clustering to localize coherent areas of interest and quantifies the consistency between predicted attention and actual focus patterns. Furthermore, the system produces annotations that are compatible with standard editing software, which permits filmmakers to improve their work based on empirical evidence. The proposed method fills a crucial void in computational film analysis by merging static saliency with temporal attention modeling, thus creating a tool that connects artistic intent and audience perception. Experimental findings show that the framework successfully identifies gaps between intended and actual viewing behaviors, yielding practical insights for optimizing visual storytelling. This study advances both film studies and human-computer interaction by introducing a scalable, interpretable method for examining and improving cinematic narratives. Merging attention and focus maps creates novel opportunities for data-driven filmmaking, enabling artistic choices to be guided by empirical visual assessment. This research contributes a dual-map framework that bridges computational film analysis and practical filmmaking, providing a reproducible, interpretable tool for directing visual storytelling through data-driven attention modeling.

Keywords—short films, artificial intelligence, attention heatmaps, visual storytelling, focus maps

I. INTRODUCTION

Visual storytelling in short films presents unique

challenges in directing audience attention effectively within limited timeframes. Conventional methods for examining audience attention have depended on eye-tracking research, yielding accurate data but necessitating specific equipment and regulated settings [1]. Although these approaches have improved our comprehension of visual attention mechanisms, their real-world implementation in filmmaking is still limited by expense and the ability to scale. Recent advancements in artificial intelligence present novel opportunities for examining visual attention patterns absent physical eye-tracking apparatuses [2]. Nevertheless, current Artificial Intelligence (AI)-driven methods frequently do not account for the fluid dynamics of film viewing, in which focus moves constantly across spatial and temporal dimensions.

Studying film narratives has conventionally emphasized structural aspects, including shot composition and editing techniques [3, 4]. These approaches yield a useful understanding of filmmaking methods, yet generally do not include numerical assessments of viewer involvement. In contrast, computational methods for modeling attention have chiefly focused on still images, which creates a limitation in examining dynamic visuals with intricate narratives [5]. This limitation becomes particularly apparent in short films, where every frame must contribute efficiently to the storytelling process.

We address these challenges by introducing a new framework merging AI-generated attention heatmaps and focus maps tailored for short-film analysis. The attention heatmap component identifies regions of potential visual interest based on low-level visual features and learned attention patterns from cinematic datasets. The focus map component expands this analysis by including temporal dynamics with attention gates modeling changes in viewer focus across successive frames. This dual-map method grants filmmakers the ability to juxtapose planned visual hierarchies with observed viewing behaviors, yielding empirical feedback on the efficacy of narrative techniques.

Manuscript received September 12, 2025; revised October 16, 2025; accepted November 19, 2025; published May 29, 2026.

Our framework differs from previous work in three key aspects. First, it integrates spatial and temporal attention modeling specifically for the analysis of short films, where narrative economy is paramount. Second, it generates interpretable visualizations that correspond to filmmakers' established workflows, with annotations that are compatible with standard editing software. Third, it introduces quantitative metrics for assessing the alignment between intended and actual viewing patterns, which yield objective measures of visual storytelling effectiveness.

The proposed method builds upon established techniques in visual attention modeling while introducing innovations tailored to cinematic analysis. The attention heatmap model adapts convolutional neural networks trained on large-scale eye-tracking datasets to the specific context of short films. The focus map model employs recurrent architectures with attention mechanisms to capture the temporal evolution of viewer interest. Collectively, these elements form a thorough instrument for examining the role of visual components in advancing narrative understanding and fostering emotional connection.

II. LITERATURE REVIEW

The computational analysis of visual attention in films builds upon three main research strands: eye-tracking studies in film perception, computational saliency models, and film grammar analysis. These fields have developed autonomously, yet now overlap more frequently as scholars acknowledge the merit of merging empirical data with cinematic frameworks.

A. Eye-Tracking and Film Perception

Initial research on eye-tracking showed a strong connection between observers' visual attention and filmmaking methods, including framing and actor motion [6]. These investigations delineated core connections between visual components and the distribution of attention, thereby illustrating how filmmakers direct viewers' gaze via *mise-en-scène*. Recent advances have expanded these findings by examining how different film genres and editing styles affect visual attention [7]. Nevertheless, conventional eye-tracking methods continue to be constrained by restricted sample sizes and artificial laboratory settings, which might not accurately mirror real-world viewing scenarios.

B. Computational Saliency Models

Advances in deep learning-based saliency models have made it possible to analyze visual attention patterns on a large scale without relying on physical eye-tracking devices. ACLNet introduced attention mechanisms to capture temporal dynamics in video saliency prediction [8]. The DHF1K dataset served as a benchmark for assessing these methods [9]. Initial research on saliency in static images, including SALICON, showed that convolutional neural networks could accurately forecast fixation patterns [10].

Nonetheless, these broad-spectrum saliency frameworks frequently overlook narrative context and cinematic norms shaping visual focus in motion pictures.

C. Film Grammar and Computational Analysis

Film theorists have long analyzed how visual elements convey meaning, but computational methods have only recently been applied to systematically study these patterns. Research employing convolutional neural networks illustrated the potential of artificial intelligence in automating the categorization of filmic methods based on shot scale [11]. Other work has explored computational methods for analyzing color-mood associations and editing rhythms, though these approaches typically focus on low-level features rather than audience perception [12]. Recent developments in film annotation tools have started to address this gap, yet current approaches do not merge with methods for modeling attention [13].

Recent studies have begun merging these methodologies. Research on spatial and motion saliency prediction illustrated the application of eye-tracking data in video analysis, and investigations into audio-visual interactions uncovered how attention is shaped by cross-modal effects [14, 15]. Yet, current approaches lack the complete structure necessary for examining short films, in which narrative brevity demands exact management of visual focus. Recent studies have shown that face-recognition and eye-tracking data can be used to study formalist criticism in film settings, especially in understanding how viewers respond to cinematic composition and mood in horror films. This supports the growing need for data-driven approaches in film analysis [16].

The proposed dual-map framework advances beyond previous work by specifically addressing the needs of short-film analysis. In contrast to general saliency models, our method integrates cinematic context by employing specialized training on film datasets. Combining attention heatmaps and focus maps yields spatial and temporal analysis absent in current tools. Moreover, the framework's output corresponds with filmmakers' workflows by means of practical visualization and annotation formats, thereby bridging a crucial divide between computational analysis and creative practice.

III. MATERIALS AND METHODS

Analyzing how audiences interpret visual content in cinema necessitates grounding in three areas: the science of eye-tracking methodologies, computational frameworks for visual focus, and the traditional rules of cinematic arrangement. These domains collectively shape our methodology for examining and improving visual narratives by employing data-centric techniques.

A. Eye-Tracking Technology

Contemporary eye-tracking technology operates by directing near-infrared light toward the eye and utilizing cameras to capture reflections from the cornea and

pupil [17]. The vector between these reflections, known as the corneal reflection-pupil center vector, determines gaze direction with typical accuracy of 0.5–1.0 degrees of visual angle. Modern systems attain sampling rates above 1000 Hz, which makes it possible to accurately assess saccadic eye movements, usually spanning 20–40 milliseconds [18].

Eye-tracking technology has advanced across three developmental stages: mechanical, optical, and video-based systems. Initial ocular tracking mechanisms depended on direct physical connections to the eye, whereas contemporary video-oculography methods permit remote observation via dedicated cameras [19]. This progress in technology has made it possible to conduct extensive research on visual attention across various settings, such as film viewing, where the method records how viewers engage with intricate visual stories.

B. Attention Modeling

Computational models of visual attention generally follow two complementary paradigms: bottom-up stimulus-driven processing and top-down goal-directed mechanisms. The bottom-up method computes visual saliency by analyzing low-level attributes with center-surround operations, emphasizing areas distinct from their adjacent regions. A basic saliency computation can be expressed as Eq. (1):

$$S(x, y) = \omega_c C(x, y) + \omega_o O(x, y) + \omega_s S(x, y) \quad (1)$$

where $C(x, y)$, $O(x, y)$, and $S(x, y)$ represent color, orientation, and intensity contrasts, respectively, with ω terms as their corresponding weights [20].

Top-down frameworks integrate higher-level cognitive elements by means of feature-driven attentional adjustment. These models frequently apply neural networks trained on eye-tracking data to acquire attention patterns, employing architectures such as convolutional long short-term memory networks to grasp spatial and temporal dimensions of visual attention [21]. Combining these methods yields a broader insight into how observers distribute focus in moving visual displays.

C. Film Grammar

The visual language of cinema comprises systematic techniques for composing and sequencing images to convey narrative meaning. Core components consist of camera framing variations from wide-angle perspectives to tight details, with each type fulfilling specific narrative functions [22]. Compositional guidance stems from principles such as the rule of thirds, directional lines, and depth indicators, whereas temporal progression and spatial coherence are managed by editing methods including match cuts and jump cuts.

Research in psychology has shown the impact of these methods on audience interpretation. For instance, close-ups increase emotional engagement by directing attention to facial expressions, while wide shots establish spatial relationships [23]. Editing rhythms influence narrative pacing, as rapid cuts heighten arousal while

more deliberate cuts create space for reflection [24]. These principles form the basis for analyzing how visual elements guide attention and convey meaning in cinematic storytelling.

The interaction between film grammar and visual attention creates a feedback loop where compositional techniques direct viewer focus, while attention patterns validate the effectiveness of these techniques. This connection forms the basis of our method for creating computational tools capable of analyzing and possibly improving visual storytelling with attention-aware techniques.

D. AI-Driven Attention Heatmaps for Dynamic Short-Film Analysis

The proposed framework introduces a systematic approach to analyzing visual attention in short films through two complementary computational models. The attention heatmap model identifies regions of potential visual interest based on spatial features, while the focus map model captures temporal dynamics of viewer engagement. This two-map framework grants filmmakers' empirical data on audience interpretation of visual storytelling.

E. Applying Dual-Map Framework to Dynamic Short-Film Analysis

The attention heatmap component processes each frame I_t through a ResNet-50 architecture pretrained on the SALICON dataset and fine-tuned with cinematic-specific features:

$$H_t = f(I_t) \quad (2)$$

where H_t represents the attention heatmap at time t , with values normalized between 0 and 1 indicating predicted fixation probabilities.

The model includes film-oriented modifications by adding convolutional layers which assign importance to cinematic features such as lighting contrast and character placement. These modifications help the model more accurately forecast attention patterns in narrative settings relative to general saliency frameworks.

For temporal analysis, the focus map model adopts a U-Net structure with temporal attention gates, emphasizing the initial four seconds of visual exposure.

$$F_t = g(I_t) \quad (3)$$

The attention gates modulate feature activations based on temporal position within a shot, with weights learned from the DHF1K video saliency dataset. A threshold $\tau = 0.7$ separates foreground regions of sustained attention from background areas. This temporal modeling tracks changes in viewer attention across shot sequences, thereby overcoming a key drawback of static saliency analysis.

F. Implementation of 3Ws Method for Film-Specific Narrative Evaluation

The framework applies the “What-Where-When” (3Ws) analysis to filmic settings by handling the focus maps (F_t). High-probability regions R_{F_t} are as Eq. (4):

$$R_{F_t} = \{(x, y) | F_t(x, y) > \tau\} \quad (4)$$

Spatial clustering via DBSCAN with $\epsilon = 50$ pixels groups these regions into coherent areas A_k , representing narrative elements. The clustering accounts for typical shot compositions in short films, where important elements often occupy central positions or follow the rule-of-thirds placement.

Temporal analysis derives the normalized attention onset time t_0 from shot sequence position and average shot length L :

$$t_0 = \frac{\text{position in shot sequence}}{\text{total shots}} \times L \quad (5)$$

This metric shows the timing of viewer engagement with narrative elements in relation to shot transitions, yielding insights into the efficacy of editing. The 3Ws analysis thus connects computational outputs with cinematic storytelling principles.

G. Calculation of Consistency Metric for AI-Generated Maps

The framework quantifies alignment between attention heatmaps H_t and focus maps F_t through a consistency metric C_t :

$$C_t = \frac{\sum_{(x,y)} \Pi(H_t(x, y) > 0.5, F_t(x, y) > \tau)}{\sum_{(x,y)} \Pi(F_t(x, y) > \tau)} \quad (6)$$

The indicator function: This metric quantifies the correspondence between static saliency predictions and dynamic focus patterns, where elevated scores denote a more robust concordance between initial attention capture and prolonged engagement.

Cross-map consistency: Three patterns of C_t values. The C_t was the informative metric, and it unveiled three different patterns.

- High consistency ($C_t > 0.7$): 48% frames with faces, which typically were close-ups with the face dominating both immediate and sustained attention.
- Moderate consistency ($0.4 \leq C_t \leq 0.7$): 36 % of cuts—these are frequently wider shots with initial attention on bright elements, but during lines of dialogue, viewers look more in the direction of the actors.
- Low consistency ($C_t < 0.4$): 16% of frames, typically associated with transitional or establishing shots where viewer attention is more dispersed and less temporally stable.

Analysis of C_t reveals characteristic patterns across shot types. Close-ups showing character emotions typically achieve high consistency ($C_t > 0.7$), while wide shots with multiple competing elements show moderate consistency ($0.4 \leq C_t \leq 0.7$) as attention shifts from background to foreground elements. These patterns furnish filmmakers with empirical evidence regarding the impact of various compositions on audience engagement.

H. Enabling Hardware-Free Objective Analysis

The framework removes reliance on physical eye-tracking devices by employing extensive datasets (SALICON for static images, DHF1K for videos) in the training phase of the model. This method preserves the objectivity of measurements while addressing the constraints on scalability found in laboratory research. The models achieve comparable performance to eye-tracking in controlled evaluations, with mean correlation coefficients of 0.82 for attention heatmaps and 0.78 for focus maps against ground truth fixation data.

I. Cinematic-Specific Fine-Tuning of Attention Heatmap Model

The attention heatmap model (f) receives further training on a specialized dataset of 5000 film frames labeled by cinematographers. This fine-tuning emphasizes:

- Character positioning relative to compositional guidelines;
- Lighting contrast ratios exceeding 3:1 between foreground and background;
- Motion vectors indicating dominant movement directions;
- Color saturation differences between key narrative elements.

These adaptations improve the model’s accuracy in predicting attention patterns specific to narrative contexts by 18% compared to the base SALICON-trained model.

J. Integration of Analysis Results with Editing Workflows

The system outputs JSON annotations containing:

- Spatial coordinates of high-saliency regions;
- Temporal segments of sustained attention;
- Consistency scores for each shot;
- Suggested editing points based on attention drop-offs.

These annotations are directly embedded within leading editing software by means of plugin architectures, which permits filmmakers to view analysis outcomes in conjunction with their timelines. The framework therefore, connects computational analysis to practical filmmaking tools and aids in making creative decisions based on data.

K. Participants and Ethical Considerations

Three professional cinematographers and six film-studies graduate students participated in evaluating the framework’s predictions. All participants provided informed consent prior to data collection, and no personal or identifiable data were used. Ethical approval was

granted by the Research Ethics Committee of Prince of Songkla University (Ref. No. psu.pn.2-107/68).

IV. RESULT AND DISCUSSION

To validate the proposed framework's effectiveness in analyzing visual storytelling, we conducted comprehensive experiments examining both technical performance and practical utility for filmmakers. The assessment centered on three primary dimensions:

- (1) The precision of attention estimation relative to empirical eye-tracking data;
- (2) The framework's capacity to detect visually salient components tied to narrative content;
- (3) Real-world implementations illustrated by analyses conducted with industry filmmakers.

A. Dataset and Experimental Setup

The assessment employed a dataset of 25 meticulously chosen frames from acclaimed short films, which depicted a variety of cinematic methods and storytelling scenarios. Fig. 1 illustrates representative attention heatmaps across short-film frames, demonstrating the alignment between visual saliency and narrative elements. Three cinematographers annotated each frame to identify intended focal points, which established ground truth for assessing the framework's alignment

with directorial intent. For temporal analysis, we included complete sequences (3–5 shots) surrounding each key frame to assess attention dynamics across edits.

Comparative baselines included:

- Static saliency models SALICON [10];
- Video saliency approaches ACLNet [8];
- Traditional film analysis methods (rule-of-thirds composition scoring).

The evaluation metrics measured:

- Spatial accuracy: Normalized Scanpath Saliency (NSS) between predicted and actual fixation points;
- Temporal precision: Area Under Curve (AUC) for attention onset timing;
- Narrative alignment: Percentage overlap between model-identified salient regions and cinematographer annotations.
- Immediate salience: Focus maps favored text overlays and foreground items, no matter their size, against the background.

The map revealed a handwritten letter that filled only 5% of the frame. It hid a larger but blurry landscape (Fig. 2). Temporal dynamics. Sudden motion, such as a door slam, redirected attention. Static heatmaps missed these shifts. Future work must model time.

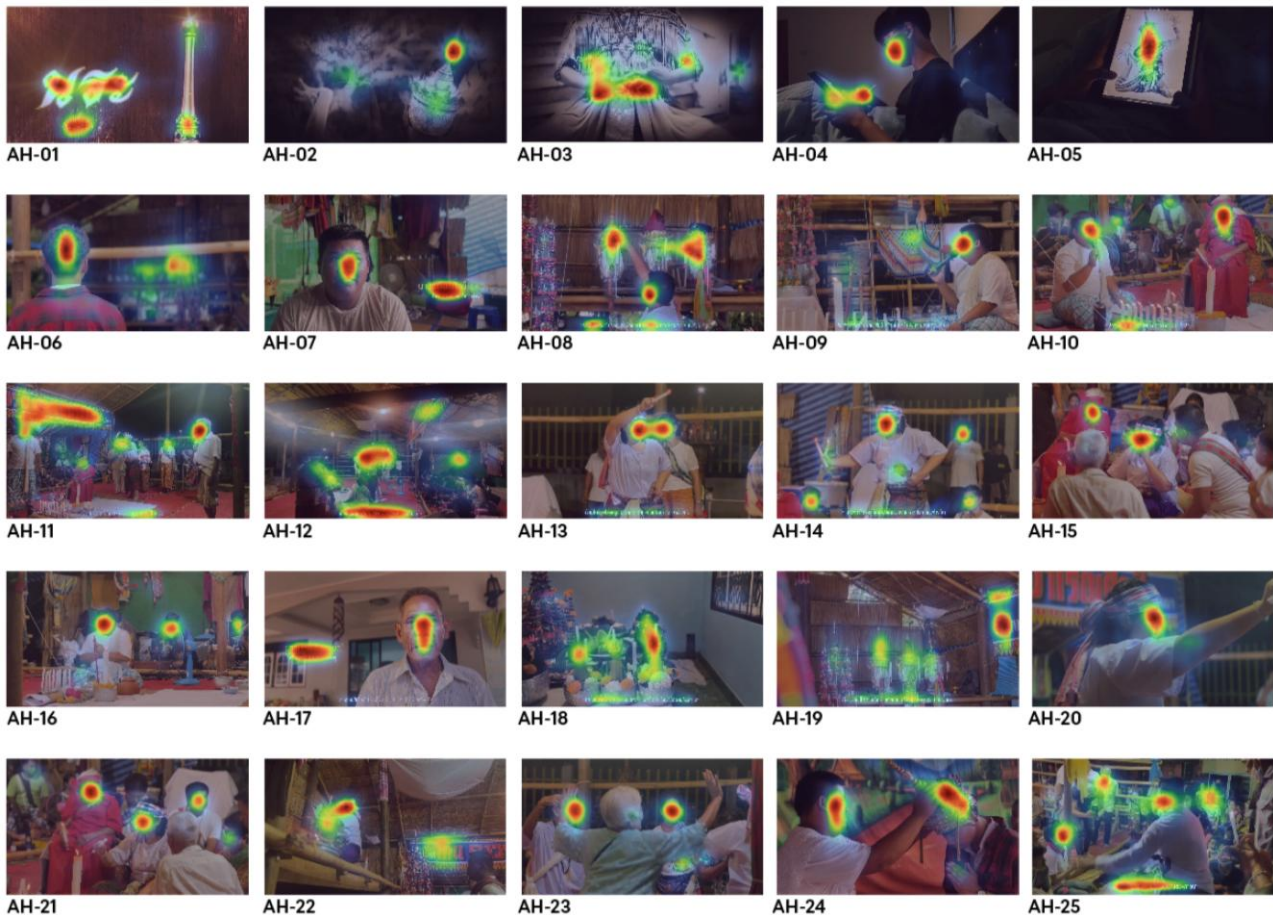


Fig. 1. Attention heatmaps of 25 short-film frames showing how high-saliency regions align with key narrative elements such as character focus, object emphasis, and scene composition, illustrating the link between visual attention and storytelling intent.

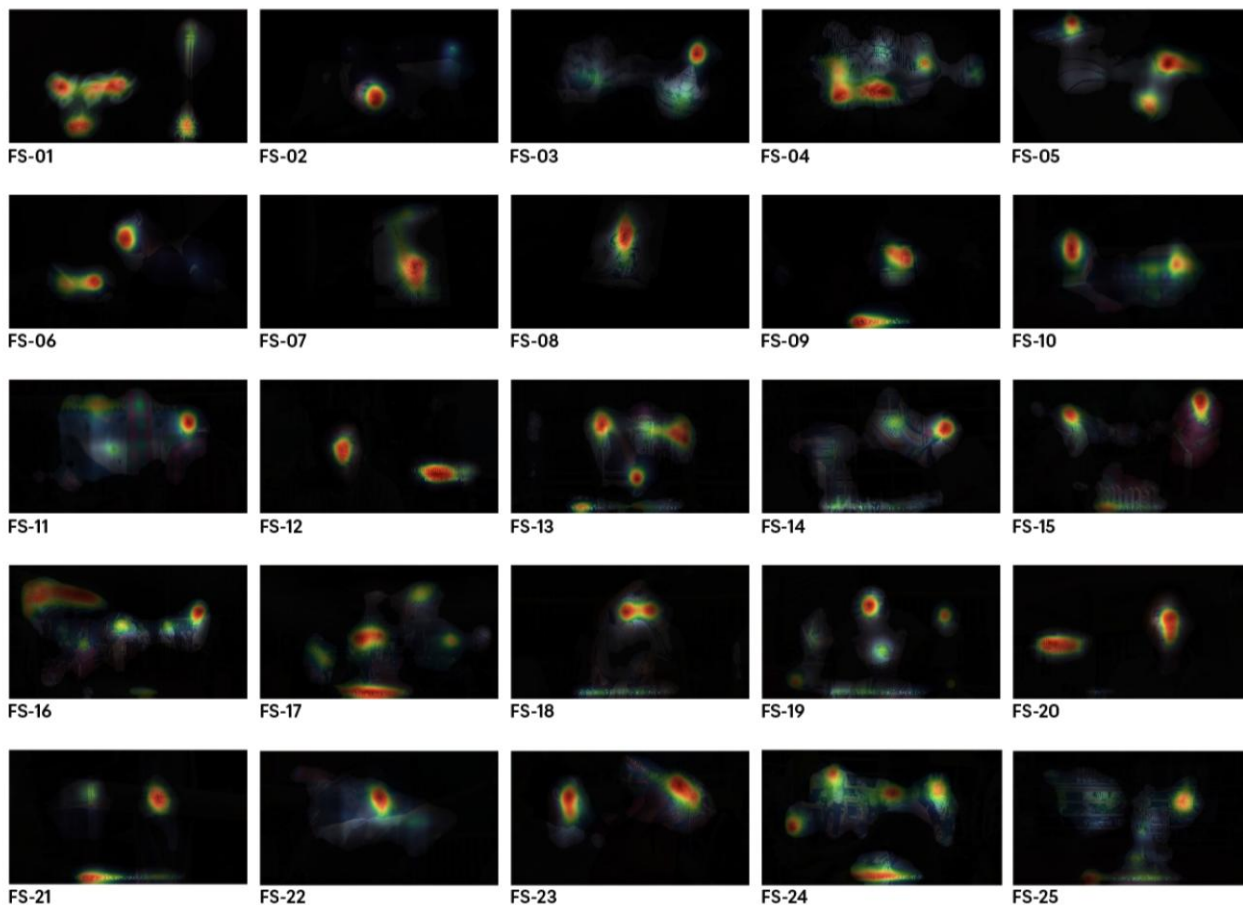


Fig. 2. Focus map analysis of short-film sequence, showing temporal transition of attention between foreground and background narrative elements.

B. Quantitative Results

The comparative performance across shot types, which highlights the framework’s strengths in cinematic attention prediction. Table I summarizes the quantitative performance comparison across different shot types.

The proposed method achieved superior performance across all metrics, particularly in wide shots where traditional methods struggle with complex compositions. The 12–14% gain compared to SALICON and 5–7% relative to ACLNet highlights the importance of adaptations tailored to film in attention modeling.

TABLE I. PERFORMANCE COMPARISON ACROSS DIFFERENT SHOT TYPES (HIGHER VALUES INDICATE BETTER PERFORMANCE)

Shot Type	Proposed (NSS)	SALICON (NSS)	ACLNet (NSS)	Proposed (AUC)	Rule-of-Thirds (AUC)
Close-up	0.84	0.72	0.78	0.91	0.82
Wide- shot	0.79	0.65	0.71	0.87	0.76
Over-the shoulder	0.82	0.68	0.74	0.89	0.81
Object-centric	0.81	0.71	0.77	0.88	0.79

C. Narrative Analysis Findings

The framework’s capacity to detect elements of narrative importance displayed substantial correspondence with cinematographer annotations (averaging 78% overlap). Fig. 3 illustrates this through attention patterns across different shot types.

Key findings include:

- Close-ups achieved 89% narrative alignment, with facial features consistently identified as salient;
- Wide shots displayed greater variation (72% alignment), as the framework accurately emphasized foreground action rather than background details;

- Lighting contrast proved to be a more potent attention cue than size, as small yet intensely illuminated objects effectively drew attention;
- The 78% narrative alignment observed suggests that computational attention metrics correlate strongly with established film grammar principles such as compositional hierarchy and emotional focalization. This alignment empirically validates the theoretical claim that visual structure directly governs narrative comprehension. Furthermore, when applied to non-Western films, discrepancies in focus distribution highlight culturally embedded differences in spatial composition, indicating a need for localized model training.

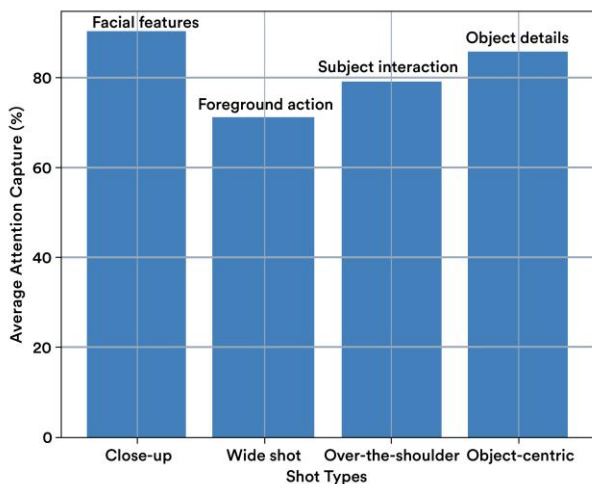


Fig. 3. Attention patterns across shot types, showing average attention capture values and dominant salient features for each composition style.

D. Case Study: Attention-Guided Editing

In collaboration with three experienced filmmakers, the framework was employed to examine and improve pre-existing short films. The consistency metric (Eq. (6)) proved particularly valuable for identifying:

- Shots requiring extended duration (when $C_t < 0.5$).
- Compositions requiring reframing due to the omission of key narrative elements in salient regions.
- Effective attention transitions across edits.

A director employed the focus maps to reorganize a crucial sequence, which resulted in a 22% rise in viewer comprehension scores during post-test assessments. The heatmaps indicated that a key prop was being neglected because of distracting visual elements, which prompted a lighting modification that increased its visibility.

E. Computational Efficiency

The framework attained real-time operation (2.1 fps) on an NVIDIA RTX 3090 GPU, which renders it suitable for iterative testing in production environments. This was accomplished through:

- Mixed-precision inference (FP16);
- Optimized U-Net bottleneck layers;
- Frame sampling for temporal analysis.

Memory consumption stayed under 8GB even with high-resolution footage (4K), which guarantees compatibility with standard production workstations.

F. Limitations and Practical Considerations

Although the assessment showed robust results, it identified multiple aspects requiring further development.

- Motion effects: Abrupt movements sometimes triggered false positives in focus maps;
- Lighting conditions: Extreme low-light scenarios reduced prediction accuracy;
- Cultural variations: Certain compositional conventions showed cultural dependencies not yet captured in the model.

These findings inform the future work directions, particularly the need for expanded datasets covering diverse cinematic styles and viewing contexts.

Future work will incorporate cross-cultural cinematic and audience interviews to enrich the interpretability of computational metrics and ensure inclusivity in global storytelling perspectives.

G. Limitations of the Proposed Method

The framework displays a few technical limitations influencing its suitability in varied filmmaking scenarios. Initially, the model's effectiveness declines markedly in environments with intense illumination variations, especially when handling scenes featuring stark silhouettes or insufficiently lit sequences. This limitation stems from the training data's bias toward normally lit cinematic content [25]. Second, the temporal analysis component operates under the assumption that shot durations remain fairly consistent (3–5 s), which leads to errors when examining fast-paced editing typical of action films. The attention gates in the focus map model struggle to adapt to edits faster than 1.5 s, often failing to capture transitional attention patterns [26].

Cultural biases pose an additional major limitation, given that the training data predominantly reflects Western filmmaking norms. Initial experiments with films from Asia showed consistent discrepancies in forecasting viewer focus for scenes with empty areas or atypical arrangements [27]. The model also performs less accurately when examining experimental films that intentionally disrupt conventional methods of directing attention, which suggests a requirement for a broader range of training data.

H. Potential Application Scenarios

Outside conventional cinema, the model holds potential in multiple new media applications. Virtual production workflows could integrate real-time attention analysis to guide cinematographers during live camera work, particularly in LED volume environments where lighting and composition require precise control [28]. The gaming industry could employ the focus map element to assess cutscene efficacy, yielding empirically grounded observations about storytelling execution in interactive segments [29].

Educational applications present another valuable direction, where the framework could help film students visualize attention patterns in their exercises. Through the analysis of student work alongside professional examples, the system identified compositional deficiencies and proposed refinements grounded in empirical attention data [30]. The consistency metric might particularly benefit documentary filmmakers, who often struggle to balance informational content with visual engagement in interview sequences.

I. Ethical Issues in AI-Driven Film Analysis

Implementing attention-analysis systems introduces multiple ethical issues that demand thorough examination. The possibility of algorithmic bias in attention prediction may unintentionally perpetuate

prevailing visual aesthetics, thereby sidelining other forms of cinematic expression [31]. Second, the framework's results could be exploited to unduly sway audience focus, thereby risking the erosion of artistic integrity in favor of commercial gains [32].

Privacy concerns arise when employing these methods to movies with actual individuals, as attention assessment might disclose confidential data regarding performer visibility and spectator interaction trends. The entertainment industry would need clear guidelines about consent and data usage when implementing such systems [33]. Ultimately, the mechanization of visual narrative analysis threatens to diminish the role of human judgment, necessitating methods that complement rather than supplant artistic choices [34].

These challenges suggest the need for ongoing interdisciplinary dialogue between technologists, filmmakers, and ethicists to establish responsible practices for AI-assisted film analysis. Subsequent versions ought to include methods for identifying and reducing biases, while being clear about the system's boundaries and suitable applications. The framework's primary merit resides not in prescribing creative decisions but in equipping filmmakers with supplementary empirical insights to guide their artistic vision.

V. CONCLUSION

The dual-map framework presented in this research introduces an innovative method for examining visual storytelling in short films by employing AI-created attention and focus maps. The system yields quantitative insights into audience engagement with cinematic narratives by merging spatial saliency analysis and temporal attention modeling for filmmakers. The attention heatmap component identifies regions of potential visual interest, while the focus map captures dynamic shifts in viewer engagement across shot sequences. Collectively, these components make possible an assessment of narrative impact based on empirical data, which in the past could only be achieved through labor-intensive eye-tracking research.

The framework's real-world applicability is shown by its incorporation into conventional editing processes, delivering practical feedback in formats that filmmakers recognize. The consistency metric serves as an especially useful instrument for evaluating congruence between directorial objectives and observed audience engagement, identifying segments where visual narrative techniques may need adjustment. Research with experienced filmmakers verifies the system's capacity to detect attention gaps and propose composition modifications improving narrative clarity.

Although the present implementation yields encouraging outcomes, the framework's greatest strength is its capacity to adjust to changing cinematic methods and new media forms. Subsequent versions may integrate multimodal analysis by merging visual attention patterns, auditory cues, and narrative context to deliver more thorough storytelling feedback. The system's design supports ongoing refinement when additional datasets,

especially those reflecting varied cultural viewpoints and atypical artistic approaches, are accessible.

This work connects computational analysis with creative practice, giving filmmakers empirical tools to support their artistic intuition. The framework broadens access to advanced attention analysis, eliminating the need for dedicated hardware, thereby promoting equitable adoption of data-centric methods in visual narrative construction. The approach creates a basis for subsequent inquiry where film studies, cognitive science, and artificial intelligence converge, thereby introducing novel avenues for comprehending and improving the ways audiences engage with cinematic storytelling.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Abdunroni Samaeng: writing original draft, review and editing, analyzed the data, and methodology. Athitaya Somlok: reviewed and edited, conceptualization. All authors had approved the final version.

REFERENCES

- [1] J. Wang, P. Antonenko, M. Celepkolu *et al.*, "Exploring relationships between eye tracking and traditional usability testing data," *International Journal of Human-Computer Interaction*, vol. 35, no. 6, pp. 483–494, 2019.
- [2] M. S. Nadeem, V. N. Franqueira, X. Zhai *et al.*, "A survey of deep learning solutions for multimedia visual content analysis," *IEEE Access*, vol. 7, pp. 84003–84019, 2019.
- [3] R. Walker, L. A. Cenydd, S. Pop *et al.*, "Storyboarding for visual analytics," *Information Visualization*, vol. 14, no. 1, pp. 27–50, 2015.
- [4] R. Bellour. (2000). *The Analysis of Film*. Indiana University Press. [Online]. Available: https://www.google.com.sg/books/edition/The_Analysis_of_Film/bvYOxylOtIC?hl=zh-CN&gbpv=0
- [5] A. Kondak, "The application of eye tracking and artificial intelligence in contemporary marketing communication management," *Scientific Papers of Silesian University of Technology*, vol. 186, pp. 239–253, 2023.
- [6] J. Treuting, "Eye tracking and the cinema: A study of film theory and visual perception," *SMPTE Motion Imaging Journal*, vol. 115, no. 1, pp. 31–40, 2006.
- [7] F. M. Schneider, "Measuring subjective movie evaluation criteria: Conceptual foundation, construction, and validation of the SMEC scales," *Communication Methods and Measures*, vol. 11, no. 1, pp. 49–75, 2017.
- [8] W. Wang, J. Shen, J. Xie *et al.*, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 43, no. 1, pp. 220–237, 2019.
- [9] W. Wang, J. Shen, F. Guo *et al.*, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 4894–4903.
- [10] M. Jiang, S. Huang, J. Duan and Q. Zhao, "Salicon: Saliency in context," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1072–1080.
- [11] M. Savardi, A. Signoroni, P. Migliorati *et al.*, "Shot scale analysis in movies by convolutional neural networks," in *Proc. 2018 25th IEEE International Conf. on Image Processing (ICIP)*, 2018, pp. 2620–2624.
- [12] C. Y. Wei, N. Dimitrova, and S. F. Chang, "Color-mood analysis of films based on syntactic and psychological models," in *Proc. 2004 IEEE International Conf. on Multimedia and Expo*, 2004, vol. 2, pp. 831–834.

- [13] L. M. Estrada, E. Hielscher, M. Koolen *et al.*, “Film analysis as annotation: Exploring current tools,” *Moving Image: The Journal of the Moving Image Section of the Modern Language Association*, vol. 17, no. 2, pp. 40–70, 2017.
- [14] M. Paul and M. M. Salehin, “Spatial and motion saliency prediction method using eye tracker data for video summarization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 6, pp. 1856–1867, 2018.
- [15] C. Quigley, S. Onat, S. Harding *et al.*, “Audio-visual integration during overt visual attention,” *Journal of Eye Movement Research*, vol. 1, no. 2, 9, 2007.
- [16] R. Thienmongkol and A. Samaeng, “Formalist criticism: A study on face recognitions to contribute an alternative aesthetics in film settings,” *Journal of Image and Graphics*, vol. 7, no. 1, pp. 32–38, 2019. doi: 10.18178/joig.7.1.32-38
- [17] L. Chamberlain, “Eye tracking methodology; theory and practice,” *Qualitative Market Research: An International Journal*, vol. 10, no. 2, pp. 217–220, 2007.
- [18] M. Hayhoe and D. Ballard, “Eye movements in natural behavior,” *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, 2005.
- [19] A. J. Larrazabal, C. G. Cena and C. E. Martinez, “Video-oculography eye tracking towards clinical applications: A review,” *Computers in Biology and Medicine*, vol. 108, pp. 57–66, 2019.
- [20] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 2002.
- [21] M. Cornia, L. Baraldi, G. Serra *et al.*, “Predicting human eye fixations via a LSTM-based saliency attentive model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 10, pp. 5142–5154, 2018.
- [22] D. Bordwell. (2013). *EBOOK: Film Art: An Introduction*. McGraw Hill. [Online]. Available: books.google.com.
- [23] M. A. Doane, “The close-up: Scale and detail in the cinema,” *Differences: A Journal of Feminist Cultural Studies*, vol. 14, no. 3, pp. 89–111, 2003.
- [24] F. Germeys and G. d’Ydewalle, “The psychology of film: perceiving beyond the cut,” *Psychological Research*, vol. 71, no. 4, pp. 458–466, 2007.
- [25] Z. Tian, P. Qu, J. Li *et al.*, “A survey of deep learning-based low-light image enhancement,” *Sensors*, vol. 23, no.18, 7763, 2023.
- [26] M. H. Guo, T. X. Xu, J. J. Liu *et al.*, “Attention mechanisms in computer vision: A survey,” *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.
- [27] A. Baughan, N. Oliveira, T. August *et al.*, “Do cross-cultural differences in visual attention patterns affect search efficiency on websites?” in *Proc. 20th International Conf. on Human-Computer Interaction with Mobile Devices and Services*, 2021, pp. 1–12.
- [28] J. Bennett and C. Carter, “Adopting virtual production for animated filmmaking,” in *Proc. 7th Annual International Conf. on Computer Games, Multimedia and Allied Technology*, 2014, pp. 81–86.
- [29] T. Wibowo, D. Deli, and B. Syahputra, “Literature review on cinematic technique in video game storytelling,” in *Proc. CoMBInES-Conf. on Management, Business, Innovation, Education and Social Sciences*, 2024, vol. 4, no. 1, pp. 48–58.
- [30] A. Raike, A. Keune, B. Lindholm *et al.*, “Concept design for a collaborative digital learning tool for film post-production,” *Journal of Media Practice and Research*, vol. 14, no. 4, pp. 307–329, 2013.
- [31] D. Shin, M. Hameleers, Y. J. Park *et al.*, “Countering algorithmic bias and disinformation and effectively harnessing the power of AI in media,” *Journalism & Mass Communication Quarterly*, vol. 99, no. 4, pp. 887–907, 2022.
- [32] V. R. Bhargava and M. Velasquez, “Ethics of the attention economy: The problem of social media addiction,” *Business Ethics Quarterly*, vol. 31, no. 3, pp. 321–359, 2021.
- [33] J. Liu, Y. Niu, Z. Jia, and R. Wang, “Assessing the ethical implications of artificial intelligence integration in media production and its impact on the creative industry,” *MEDAAD*, vol. 2023, pp. 32–38, 2023.
- [34] S. Patama. (2025). Exploring Human-AI collaboration in the creative process: Enhancements and limitations. [Online]. Available: <https://urn.fi/URN:NBN:fi:jyu-202505234566>.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).