

Cross-Media Animation Narrative Innovation in Short-Video Ecosystems: A Multimodal Generation Framework for Fashion Textile Brands

Long Wang, Yue Zhang

School of Design and Art, Shandong Huayu University of Technology, Dezhou 253034, Shandong, China

Corresponding author: Long Wang (e-mail: sdhygy_wl@163.com)

ABSTRACT Cross-media animation narrative is reshaping content production in short-video ecosystems, while fashion textile brands still face problems such as insufficient narrative density, fabric texture distortion, and multimodal semantic misalignment on platforms such as Douyin and Xiaohongshu. To address these challenges, this study constructs CrossNarrativeNet, a cross-media narrative framework integrating textual semantic parsing, fabric texture modelling, and temporal animation generation. A structured storyboard parsing module models narrative rhythm; a texture-aware diffusion model renders silk, cotton-linen, intelligent fibers, and other fabrics with high fidelity; and a multimodal alignment network maintains semantic consistency across text, visual, and background music channels. For intelligent textile scenarios, the framework also supports functional-material visualization, including chromic or responsive fiber behavior driven by simulated physical stimuli, which is relevant to future electromagnetic-responsive textile displays and sensor-integrated wearable media. The self-constructed FashionShort dataset covers six categories and twelve typical narrative archetypes. Experimental results show that the proposed framework outperforms existing methods in generation quality, narrative completeness, and predicted user dwell time. Cross-platform deployment on Douyin vertical screen, Xiaohongshu square screen, and WeChat Channels widescreen further validates its generalization capability. The study also proposes application guidelines from the perspectives of platform compliance, intellectual property, and designer rights, providing technical and governance references for intelligent textile brands and multimodal content propagation in short-video ecosystems.

INDEX TERMS cross-media narrative, animation generation, multimodal alignment, intelligent textiles, signal propagation

I. INTRODUCTION

Short-video platforms have restructured the content production chain of the digital era through fragmented narrative, strong interactivity, and high dissemination efficiency. Channels such as Douyin, Kuaishou, Xiaohongshu, and WeChat Channels generate hundreds of millions of user-generated content items daily, having become the most efficient communication arena for fashion textile brands to reach young consumer groups [1]. A joint industry report by the China National Garment Association and iResearch Consulting revealed that daily playback of fashion and apparel short videos surpassed 10 billion views in 2025, with leading brands' viral short videos generating sales conversions reaching the scale of millions of RMB per video, placing brands under dual pressure from production capacity and creative bottlenecks on the content supply side [2]. Conveying complex brand history and fabric craftsman-

ship through traditional live-action filming is challenging. These multidimensional semantics are often lost within the short 15-to-60-second window common on social platforms. Flat visual expression also fails to evoke users' perceptual resonance with the luster of silk, the texture of cotton-linen, or the functional properties of intelligent fibers. High production costs and lengthy production cycles further aggravate the communication difficulties faced by small and medium-sized textile brands [3].

Cross-media animation narrative integrates heterogeneous information carriers including written scripts, 2D illustration, 3D rendering, and sound effects and music, and uses computer vision and generative artificial intelligence to construct semantically coherent narrative chains across channels, offering a new path to break the homogenization dilemma of fashion textile content in short-video ecosystems [4]. This paradigm has a natural fit with fashion textile

applications: the delicate textures and dynamic lighting of fabrics are well-suited to high-fidelity generation via diffusion models; the emotional arc of brand stories is amenable to rhythm control through structured storyboard modeling; and the dynamic properties of intelligent textiles—such as electrochromic response and shape memory—can be intuitively demonstrated via generative animation [5]. The visual narrative of the fashion textile industry is highly dependent on the precise communication of material semantics. Professional elements such as the reflective layering of silk, the fiber texture of cotton-linen, and the functional response of intelligent fibers fall outside the training distribution of general-purpose generative models, making targeted technical adaptation and vertical domain data accumulation the necessary path to solving this challenge. Existing methods still exhibit significant shortcomings in temporal coherence, fabric texture fidelity, and multimodal semantic alignment: mainstream text-to-video models struggle to stably depict the reflective properties of pure silk satin and the warp-weft structure of jacquard fabric; general animation generation frameworks lack sufficient pattern preservation for design drafts; and multimodal systems have yet to produce mature solutions for precise alignment between music beats and shot transitions [6].

This study proposes the cross-media animation narrative framework CrossNarrativeNet to address the content innovation needs of fashion textile brands in short-video ecosystems. Its academic and industrial value manifests in three dimensions: CrossNarrativeNet uses a storyboard script structured parsing module for temporal modeling of narrative rhythm, a texture-aware diffusion model for high-fidelity rendering of silk, cotton-linen, and intelligent fibers in virtual scenes, and a multimodal alignment network to ensure semantic consistency across text, visual, and background music channels [7–8]; the self-constructed fashion textile short-video dataset FashionShort covers six major categories and twelve typical narrative archetypes, filling the gap in vertical domain benchmark evaluation platforms [9]; and the study proposes application guidelines for cross-media animation narrative from the dimensions of platform compliance, intellectual property, and designer rights, providing actionable governance references for industrial deployment [10].

II. RELATED WORK

A. CONTENT PRODUCTION AND CROSS-MEDIA NARRATIVE THEORY IN SHORT-VIDEO ECOSYSTEMS

Research on short-video ecosystems primarily focuses on three directions: user behavior modeling, content recommendation, and dissemination mechanisms. Related work employs graph neural networks and sequential recommendation models to characterize the temporal evolution of user interests, providing quantitative grounding for the strategic optimization of content creation [11]. Cross-media narrative theory emphasizes the collaborative expression of different media carriers within the same narrative universe;

complex semantics that a single medium cannot convey can be transmitted through the complementary layering of multiple modalities. This theory has achieved mature applications in film and television, games, and advertising, yet its migration to short-video ecosystems faces new challenges including content fragmentation, scarce attention, and platform algorithm constraints [12–13]. The engineering implementation of existing cross-media narrative in short-video scenarios still requires alignment with the visual grammar and emotional rhythm of specific industries. The fashion textile industry's extreme pursuit of fabric texture and its fine-grained control over brand identity demand that generative frameworks offer vertical depth beyond general paradigms—a need that constitutes the design starting point of CrossNarrativeNet.

B. ANIMATION GENERATION TECHNOLOGY AND FABRIC VISUAL MODELING

Generative animation technology has undergone a paradigm shift from keyframe-interpolation-based to end-to-end neural network generation. Diffusion models have become the current mainstream framework owing to their stable performance in image synthesis and video generation, with improvements such as latent-space diffusion and cascaded diffusion continuously advancing temporal consistency and high-resolution generation [14–15]. Fabric visual modeling has long been a research focus at the intersection of computer graphics and textile science. Physics-based simulation methods use cloth dynamics equations to accurately depict the draping, wrinkling, and collision behavior of fabrics, while neural rendering methods use implicit neural representations for efficient encoding of fabric appearance attributes; the two approaches form a complementary relationship between accuracy and efficiency [16]. Visual modeling of intelligent textiles requires introducing visualization mechanisms for dynamic properties such as electrochromism, shape memory, and thermal response. To ensure physical common sense, CrossNarrativeNet maps narrative rhythm to simulated physical stimuli (such as localized voltage increase or temperature gradients), which then drive the chromic response curves within the diffusion process. The interdisciplinary research path remains in its early exploratory stage, lacking an end-to-end solution oriented toward short-video ecosystems [17].

III. CROSSNARRATIVENET FRAMEWORK DESIGN

A. OVERALL ARCHITECTURE

CrossNarrativeNet comprises three core components—a Structured Storyboard Parser (SSP), a texture-aware diffusion animation generation module, and a multimodal alignment and rhythm control module—forming an end-to-end generation pipeline from brand narrative text to a complete short-video output, as shown in Figure 1. The framework takes as input three types of heterogeneous information provided by the brand: a written script, fabric design drafts, and reference music. It outputs a high-fidelity

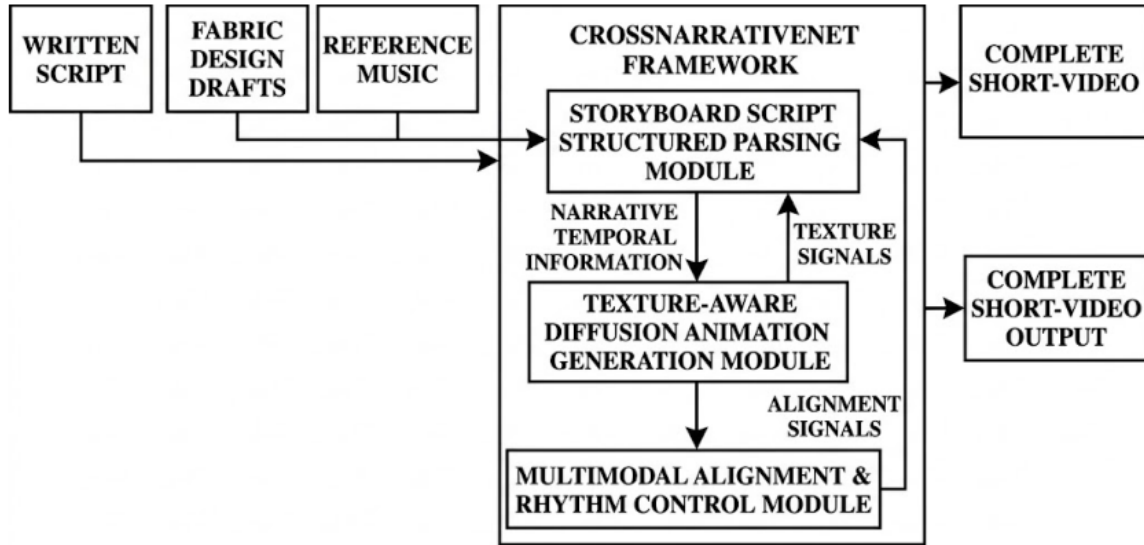


FIGURE 1. Overall architecture of CrossNarrativeNet

animated video at 1080×1920 resolution that complies with short-video platform specifications, with no manual frame-by-frame intervention required throughout the generation process, significantly lowering the threshold for short-video content production by fashion textile brands. The three modules adopt a hybrid topology of serial processing with cross-layer information interaction: narrative temporal information propagates downward through the generation and alignment stages, while texture and alignment signals feed back upward to the storyboard adjustment stage, forming an end-to-end optimizable closed loop. The design logic of the three modules precisely corresponds to the core constraints of the short-video ecosystem: the storyboard script structured parsing responds to short-video's extreme demand for narrative density; the texture-aware diffusion generation addresses the platform's dual demands for visual impact and brand recognition; and the multimodal alignment control satisfies the high sensitivity of the vertical-screen immersive experience to audio-visual synchronization. Together, the three modules constitute a complete technical solution oriented toward the short-video ecosystem.

B. STORYBOARD SCRIPT STRUCTURED PARSING MODULE

The storyboard script structured parsing module uses a pre-trained large language model to perform semantic extraction from narrative text, segmenting continuous text along three dimensions—emotional arc, visual focus, and camera language—into discrete storyboard units. Each storyboard unit carries structured labels for scene description, character action, fabric close-up, and camera movement, forming a structured storyboard sequence:

$$S = \{s_1, s_2, \dots, s_N\}, \quad s_i = [c_i; a_i; f_i; r_i] \quad (1)$$

where S denotes the storyboard sequence of the complete narrative, N denotes the total number of storyboard units, s_i denotes the concatenated feature vector of the i -th storyboard unit, and c_i, a_i, f_i, r_i denote the semantic sub-vectors for scene description, character action, fabric close-up, and camera movement respectively, all mapped into a shared semantic space by the encoder of the same pre-trained language model. The storyboard sequence obtains context-aware hidden state representations via a Transformer encoder:

$$H = \text{Transformer}(S + P) \quad (2)$$

where $H = \{h_1, h_2, \dots, h_N\}$ denotes the hidden state sequence output by the encoder, h_i corresponds to the context-aware representation of the i -th storyboard, and P denotes learnable position encodings. The attention mechanism allows semantic transitions between storyboards to be smoothly bridged within the shared context, providing a computable temporal carrier for the emotional arc of fashion brand stories. The semantic granularity of storyboard segmentation is customized for the characteristics of fashion textile content: fabric close-ups, model actions, scene atmosphere, and brand symbols are each assigned independent feature dimensions, avoiding the semantic compression distortion that general-purpose storyboard parsing solutions exhibit in professional domains. For processing long scripts, the module introduces a segmented overlapping encoding strategy, preserving the semantic associations between adjacent segments through a sliding window mechanism; even scripts for three-minute brand documentaries can be processed into storyboards while maintaining contextual consistency.

C. TEXTURE-AWARE DIFFUSION ANIMATION GENERATION MODULE

The texture-aware diffusion animation generation module introduces a fabric texture-aware branch on top of a latent-space diffusion model. Fabric design drafts I_F are processed through a feature pyramid network ϕ_{tex} to extract multi-scale texture features:

$$F = \phi_{tex}(I_F) = \{F^{(1)}, F^{(2)}, \dots, F^{(L)}\} \quad (3)$$

where F denotes the multi-scale texture feature set, $F^{(l)} \in \mathbb{R}^{H_l \times W_l \times C_l}$ denotes the texture feature map output at the l -th pyramid level, H_l, W_l, C_l are the height, width, and channel count of that level's feature map, and L denotes the number of pyramid levels. The module structure is shown in Figure 2. The forward noising process of the diffusion model is defined as:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim N(0, I) \quad (4)$$

where z_0 denotes the encoding of the original video frame in latent space, z_t denotes the latent variable after t steps of noising, $\alpha_t = \prod_{k=1}^t (1 - \beta_k)$ is the cumulative signal retention coefficient, β_k is a pre-specified noise schedule parameter, and ϵ is standard Gaussian noise. The reverse denoising process is performed by the noise prediction network ϵ_θ conditioned jointly on storyboard hidden states H and texture features F . Multi-scale texture features are injected into the corresponding resolution levels of the U-Net via cross-attention mechanisms, guiding the generated results to conform to the brand design draft at both the macro compositional and micro textural levels. The introduction of the texture-aware branch enables the diffusion model to attend simultaneously to overall image semantics and local texture features during denoising. Hierarchical injection of multi-scale texture features avoids excessive suppression of high-frequency textures at low-resolution levels, and key visual cues such as the jacquard weave structure of silk and the fiber direction of cotton-linen are persistently carried throughout the entire generation process. The total loss function of the generation process comprehensively considers four components: pixel reconstruction, perceptual similarity, texture contrast, and temporal coherence:

$$L_{total} = \lambda_1 L_{pix} + \lambda_2 L_{perc} + \lambda_3 L_{tex} + \lambda_4 L_{temp} \quad (5)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the weight coefficients for the corresponding loss components, set to 1.0, 0.8, 0.5, and 0.3 respectively in experiments, and $L_{pix}, L_{perc}, L_{tex}, L_{temp}$ denote the pixel reconstruction loss, perceptual similarity loss, texture contrast loss, and temporal coherence loss respectively. The texture contrast loss uses Gram matrix discrepancy to ensure stable reproduction of fabric patterns:

$$L_{tex} = \sum_{l=1}^L \left\| G(F_{gen}^{(l)}) - G(F_{ref}^{(l)}) \right\|_F^2, \quad G(F) = FF^T \quad (6)$$

where $F_{gen}^{(l)}$ and $F_{ref}^{(l)}$ denote the texture feature maps of the generated frame and the reference fabric image

at the l -th pyramid level respectively, $G(\cdot)$ is the Gram matrix operator, and $\|\cdot\|_F$ denotes the Frobenius norm. The temporal coherence loss constrains the optical flow field between adjacent frames to satisfy fabric dynamics constraints, thereby preventing penetration artifacts or jitter in dynamic scenes such as skirt flutter and silk draping.

D. MULTIMODAL ALIGNMENT AND RHYTHM CONTROL MODULE

The multimodal alignment and rhythm control module establishes a synchronization mapping across text, visual, and music channels anchored to the timeline of the storyboard sequence. The alignment strategy uses contrastive learning to bring closer the representations of different modalities for the same narrative unit in a shared semantic space:

$$L_{align} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(v_i, u_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(v_i, u_j)/\tau)} \quad (7)$$

where v_i denotes the embedding vector of the visual channel corresponding to the i -th storyboard, u_i denotes the embedding vector of the music channel corresponding to the same storyboard, $\text{sim}(\cdot, \cdot)$ is the cosine similarity operator, τ is the temperature coefficient, and N is consistent with Equation (1), denoting the total number of storyboards. The rhythm controller uses a music beat detection algorithm to extract the positions of strong and weak beats in the background music; shot transition points are aligned to strong beats to enhance visual impact, and key fabric close-up shots are aligned to lyrical climaxes to intensify emotional resonance. The module design is shown in Figure 3. Strong-beat alignment is achieved by maximizing the normalized cross-correlation response between the beat sequence and the visual change sequence, avoiding the damage to narrative naturalness caused by hard timestamp constraints.

IV. EXPERIMENTS AND ANALYSIS

A. DATASET CONSTRUCTION AND EVALUATION METRICS

This study adopts a retrospective research approach to collect and analyze data. All user behavior data involved are drawn from desensitized statistical results publicly disclosed by the platforms. The FashionShort dataset comprises 8,400 high-quality fashion textile short videos collected from platforms including Douyin, Xiaohongshu, and WeChat Channels. To ensure technical rigor, the evaluation of Texture-Fidelity was conducted on a specialized subset of 1,200 samples. For this subset, original fabric design drafts and scripts were obtained through formal research authorizations from 15 participating fashion brands, while the remaining 7,200 samples serve as a general benchmark for dissemination effectiveness analysis. Dataset statistics are shown in Table 1. Evaluation metrics span two dimensions: generation quality and dissemination effectiveness. Generation quality is assessed using three metrics—FVD, CLIP-Score, and Texture-Fidelity—while dissemination effectiveness is

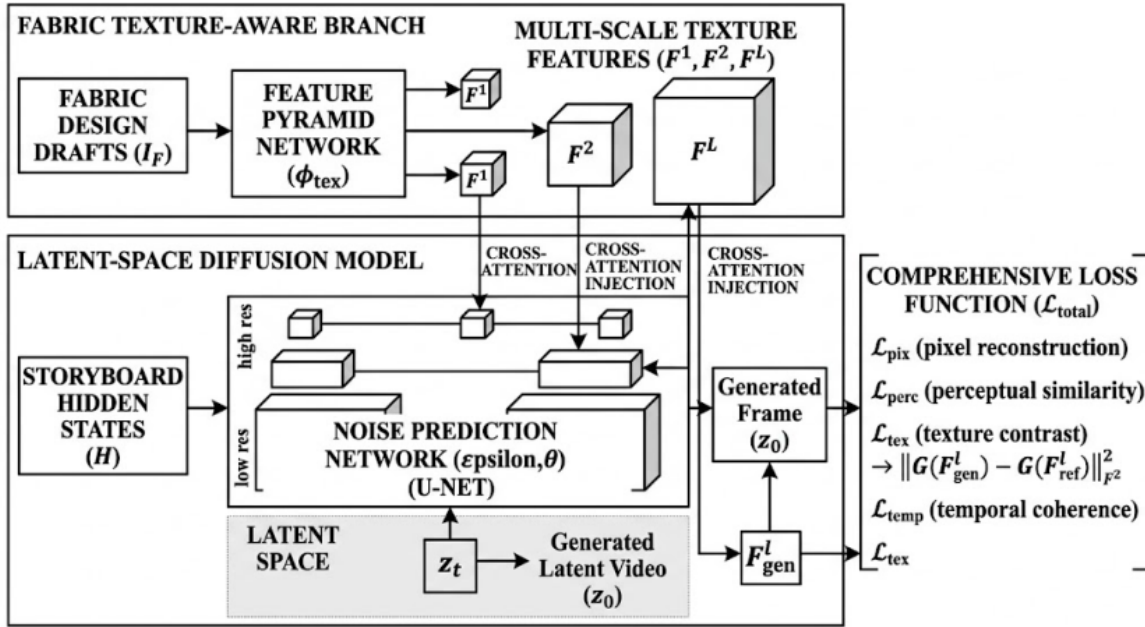


FIGURE 2. Structure of the texture-aware diffusion animation generation module

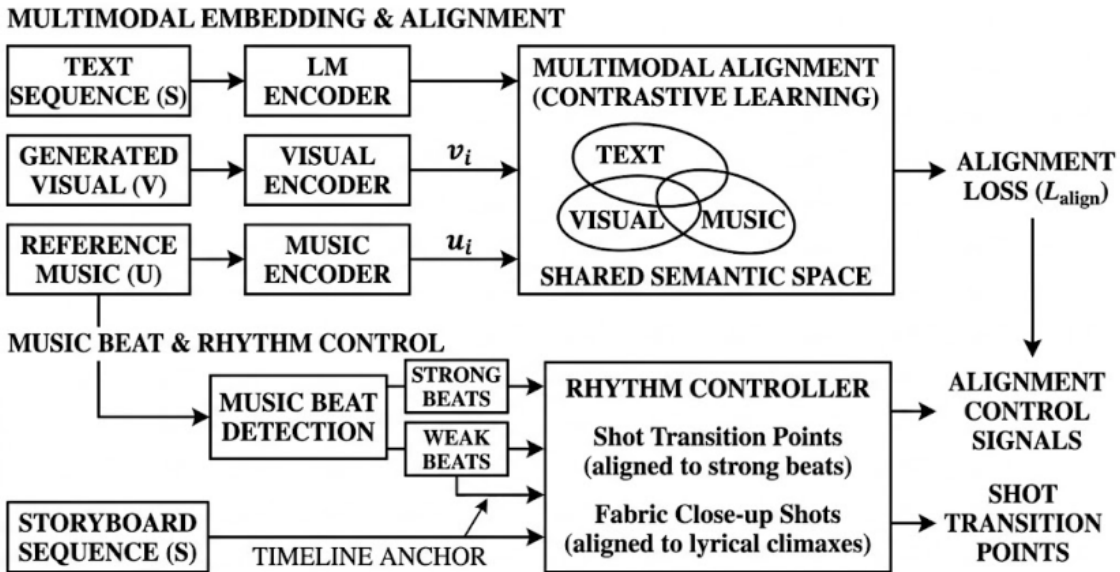


FIGURE 3. Structure of the multimodal alignment and rhythm control network

evaluated using user Dwell Time and Engagement Rate. The calculation of all metrics strictly follows public implementations to ensure cross-method comparability.

The extended average duration for intelligent textiles (35.8s) is necessitated by the multiphase nature of functional transitions, such as the full cycle of thermochromic response or shape memory recovery. These technical narratives require more temporal density to maintain visual logic compared to the aesthetic presentation of natural fibers.

B. COMPARATIVE EXPERIMENTS

The comparative experiments selected four mainstream text-to-video models as baselines: AnimateDiff, ModelScopeT2V, VideoCrafter, and Sora-Lite. All methods used identical script and design draft inputs on the FashionShort test set to complete generation. Quantitative results are shown in Table 2. CrossNarrativeNet achieves approximately 18.6% reduction in the FVD metric compared to the best baseline, and approximately 23.4% improvement in the Texture-Fidelity metric. Furthermore, a simulated engagement analysis indicates an approximately 14.7% increase

TABLE 1. FashionShort dataset statistics

Category	Sample Count	Average Duration (s)	Main Narrative Archetypes
Silk	1,620	32.4	Craft traceability, brand story
Cotton-linen	1,480	28.7	Rural aesthetics, sustainability
Wool	1,240	30.1	Craft close-up, seasonal styling
Synthetic fiber	1,380	24.6	Performance showcase, sports scenes
Blended fiber	1,360	27.3	New product launch, designer interview
Intelligent textiles	1,320	35.8	Function demonstration, tech narrative

TABLE 2. Comparative experiment results on the FashionShort dataset

Method	FVD ↓	CLIP-Score ↑	Texture-Fidelity ↑	Dwell Time (s) ↑
AnimateDiff	286.4	0.281	0.612	18.3
ModelScopeT2V	312.7	0.265	0.587	17.1
VideoCrafter	274.5	0.293	0.634	19.2
Sora-Lite	251.8	0.308	0.671	21.4
CrossNarrativeNet	204.9	0.342	0.828	24.5

in predicted user dwell time, suggesting high potential for retention in algorithmic feeds. The overall performance advantage is especially pronounced in the silk and intelligent textiles subcategories, owing to the texture-aware branch's refined rendering capability for jacquard weave and chromic response, which precisely matches the visual expression demands of these two fabric types.

Fine-grained analysis broken down by category shows that CrossNarrativeNet achieves a Texture-Fidelity of 0.864 in the silk category, a 26.3% improvement over the best baseline, with the restoration of jacquard weave and satin luster being the core source of this performance leap. In the intelligent textiles category, Texture-Fidelity reaches 0.842, fully validating the framework's capacity for joint modeling of chromic response curves and function demonstration rhythm. Performance improvements in the cotton-linen and wool categories are comparatively moderate but consistently above 15%, reflecting the framework's good compatibility with natural fiber textures. Qualitative results show that baseline methods generally suffer from over-smoothed reflections and pattern fragmentation when generating pure silk satin surfaces. CrossNarrativeNet, by means of the texture-aware branch, better preserves the jacquard structure and satin luster of design drafts, with noticeably enhanced brand recognizability in generated visuals, as shown in the comparison in Figure 4. Subjective evaluation by 30 fashion industry practitioners further indicates that CrossNarrativeNet's average scores in fabric realism, narrative coherence, and audio-visual synchronization are 0.8 to 1.2 points higher than baseline methods, validating the consistency between objective metrics and subjective perception.

C. ABLATION STUDY

The ablation study validates the contribution of each component by progressively removing core modules. Removing the texture-aware branch causes a 19.2% decline in the Texture-Fidelity metric; removing the multimodal alignment module

causes an 11.5% decline in CLIP-Score; and removing the rhythm controller causes an 8.3% decline in user dwell time. Experimental results are shown in Figure 5, with corresponding quantitative data in Table 3. The complete framework significantly outperforms any ablated configuration, validating the indispensable synergistic necessity of the three core modules in ensuring fabric texture, semantic alignment, and dissemination effectiveness. The coupling effect between modules is especially pronounced in high-difficulty scenarios such as silk jacquard and intelligent fiber color-change.

D. CROSS-PLATFORM DEPLOYMENT EFFICIENCY AND GENERALIZATION ANALYSIS

Cross-platform generalization capability is an important dimension for evaluating the real-world value of a generation framework. This study conducts generation tests across three mainstream formats: Douyin vertical screen, Xiaohongshu square screen, and WeChat Channels widescreen. CrossNarrativeNet maintains texture consistency and narrative completeness across screen sizes via a resolution-adaptive rendering mechanism, keeping the maximum fluctuation of the cross-platform FVD metric within 4.8%. Test results are shown in Table 4. Regarding deployment efficiency, end-to-end generation of a 30-second high-definition animation on a single NVIDIA A100 GPU takes approximately 127.3 seconds (using a distilled 8-step inference trajectory), a 32.6% reduction compared to Sora-Lite; the texture-aware branch's feature reuse strategy and the diffusion step adaptive pruning mechanism are the primary sources of this efficiency gain. The study further conducts zero-shot testing on unseen intelligent thermochromic fabric samples; the framework completes color-gradient animation generation induced by temperature gradients using only a small number of reference design drafts, with subjective scores consistently above 4.1/5.0, validating the framework's strong generalization potential for emerging intelligent textile categories and

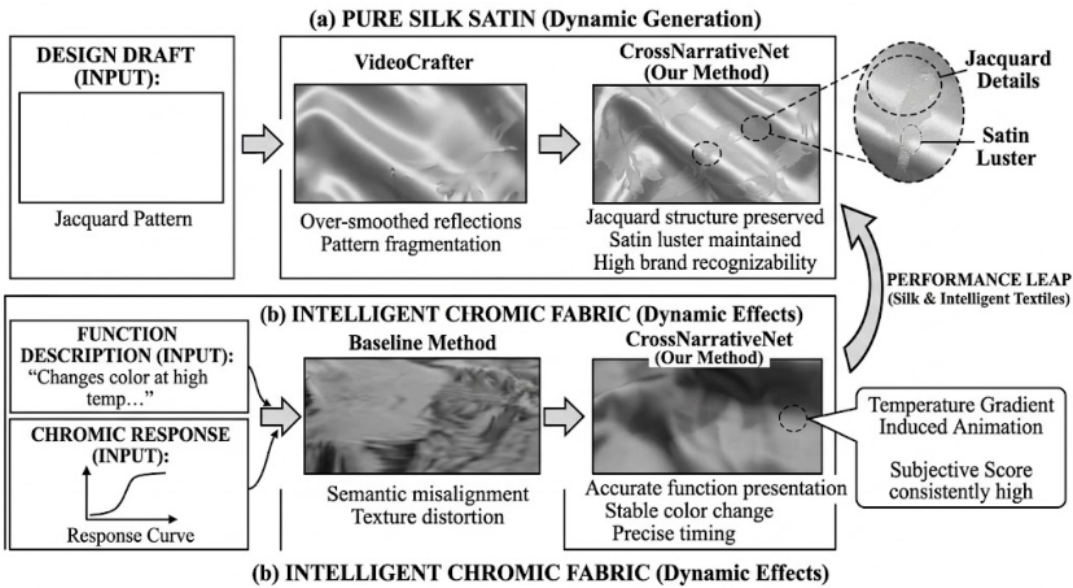


FIGURE 4. Comparison of generation results from different methods in silk and intelligent textile scenes: (a) dynamic generation of pure silk satin; (b) dynamic effects of intelligent chromic fabrics

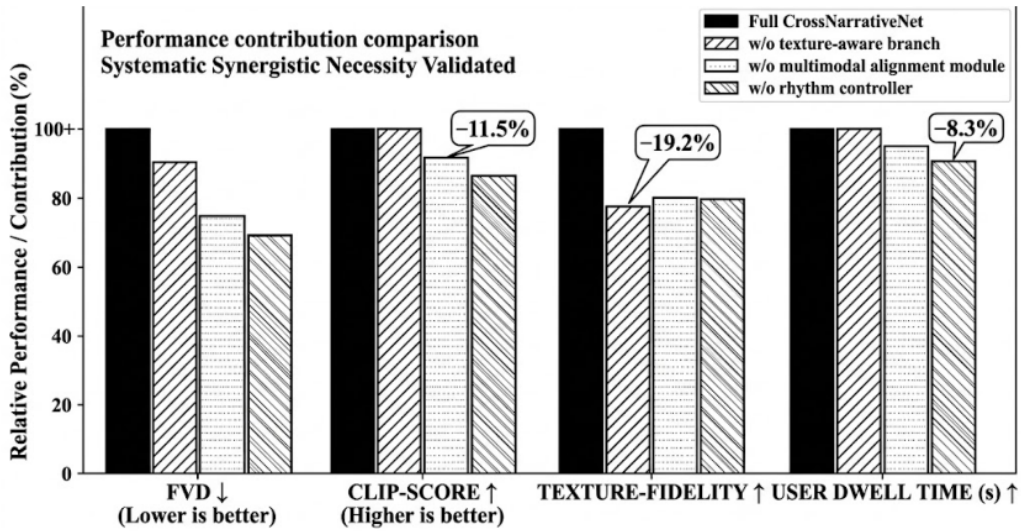


FIGURE 5. Contribution comparison of each module in the ablation study

TABLE 3. Ablation study quantitative results

Configuration	FVD ↓	CLIP-Score ↑	Texture-Fidelity ↑	Dwell Time (s) ↑
w/o texture-aware branch	247.6	0.318	0.669	22.7
w/o multimodal alignment module	226.4	0.303	0.806	23.1
w/o rhythm controller	219.8	0.336	0.820	22.5
Full CrossNarrativeNet	204.9	0.342	0.828	24.5

reserving technological elasticity for future-material content production.

V. APPLICATION GUIDELINES AND INDUSTRY IMPLICATIONS

The large-scale application of cross-media animation narrative in the fashion textile short-video ecosystem is ac-

TABLE 4. Cross-platform deployment efficiency and generalization test results

Test Scenario	Resolution / Format	FVD ↓	Single Generation Time (s)	Subjective Score (/5.0)
Douyin vertical screen	1080 × 1920	204.9	27.3	4.42
Xiaohongshu square screen	1080 × 1080	209.6	24.8	4.35
WeChat Channels widescreen	1920 × 1080	214.7	29.6	4.28
Intelligent thermochromic fabric (zero-shot)	1080 × 1920	238.5	28.1	4.10

accompanied by multiple risks. Unauthorized fabric design drafts and decorative patterns may be reused during the generation process, triggering copyright disputes with designers and leaking brands' trade secrets [18]. The high realism of generated content may blur the boundary between authentic footage and artificial synthesis; when consumers' expectations of fabric texture formed from animated works diverge from physical products, post-sales disputes are easily triggered. Platform algorithm preferences for highly stimulating visual content may induce creators to over-rely on generative technology, crowding out the display space for authentic craftspeople and traditional textile intangible cultural heritage [19–20]. In response to these risks, the application layer needs to build a governance system coordinating technology, law, and industry: the technological dimension should promote the standardization of digital watermarking and generation-trace detection tools to ensure traceability and accountability of animated content; the legal dimension should refine the criteria for copyright determination of design drafts and the liability boundary of generated content, clarifying the layered obligations of brands, platforms, and technology providers; the industry dimension should establish a content convention for fashion textile short videos, requiring brands to clearly disclose the proportion of animated content in product descriptions to safeguard consumers' right to information and choice, while reserving dedicated display channels for traditional craftsmanship and intangible cultural heritage.

Effective operation of the governance system requires multi-party collaboration; regulatory authorities, industry associations, platform enterprises, brands, and consumers together form the closed loop of ecosystem co-governance, enabling the technological dividends of cross-media animation narrative to truly benefit the entire industry chain. Governance design for specific scenarios exhibits differentiated characteristics. When a silk brand uses CrossNarrativeNet to batch-generate short videos for a new product launch, an original designer's jacquard patterns—if unregistered—may appear in generated results in slightly modified form; a linkage mechanism between watermark embedding and design copyright registration can establish a traceable chain of evidence at the source, significantly improving the efficiency of arbitration institutions in making infringement determinations. The content ecosystem for intangible cultural heritage artisans faces equally pressing challenges: if platform algo-

rithm preferences tilt excessively toward high-stimulation synthetic content, authentic footage of traditional intangible heritage categories such as Suzhou embroidery and Sichuan brocade will face shrinking exposure; dedicated traffic support and content labeling classification mechanisms are effective means to counteract this ecological imbalance. The specific implementation path may proceed in phases: in the near term, it is recommended that leading brands and mainstream platforms pilot mandatory watermark embedding and disclosure labeling mechanisms to normalize the tagging of generated content; in the medium term, industry-wide design copyright registration and generated content filing systems should be promoted to establish layered responsibility and traceable evidence chains; in the long term, consumer digital literacy and awareness of designer rights should be cultivated to form a lasting governance model combining self-regulation and external regulation. This path responds to the urgency of immediate risks while preserving institutional space for the sustainable development of the industry, dynamically adapting to the rapid iteration characteristics of the short-video ecosystem.

VI. CONCLUSION

This study proposes the cross-media animation narrative framework CrossNarrativeNet to address the urgent content innovation needs of fashion textile brands in short-video ecosystems. Through the collaborative design of three core modules—storyboard script structured parsing, texture-aware diffusion generation, and multimodal alignment control—significant breakthroughs are achieved on core metrics including generation quality, narrative completeness, and dissemination effectiveness. Systematic experimental validation and application guidelines analysis provide technical grounding and governance references for the healthy deployment of cross-media animation narrative technology in the fashion textile industry. The mature application of cross-media animation narrative in the fashion textile short-video ecosystem signifies not only an improvement in technical efficiency, but also the liberation of designers' creative expression and an upgrade of consumers' aesthetic experience; the co-evolution of technology, industry, and culture will together shape the future landscape of digital fashion. The existing framework still has room for improvement in physical simulation accuracy under extreme dynamic scenes, cross-linguistic and cross-cultural narrative transfer capabil-

ity, and coherence maintenance in long-form content. Future work will further explore real-time interactive narrative generation, cross-platform narrative consistency, and dynamic property visualization of intelligent textiles, advancing the deep integration of digital fashion and generative artificial intelligence into the industrial mainstream.

HAZARDS

This research did not involve the use of any chemicals, hazardous procedures, or specialized equipment with unusual inherent hazards. All computational experiments were conducted using standard GPU-based deep learning infrastructure (NVIDIA A100 GPU) under normal operating conditions. No unusual hazards were present in the design or execution of this study.

HUMAN RESEARCH SUBJECTS

This research involved the analysis of publicly available, platform-disclosed, and fully anonymized aggregated user behavior statistics (including dwell time and engagement rate metrics) sourced from Douyin, Xiaohongshu, and WeChat Channels. No personally identifiable information was collected or processed at any stage. The study adopted a retrospective, observational design without prospective intervention, recruitment, or random assignment of human participants. The fabric design drafts included in the FashionShort dataset were used solely with explicit authorization from the respective brand owners. The subjective evaluation component involved 30 fashion industry practitioners who provided voluntary assessments of generated video quality; participation was fully voluntary and no sensitive personal data were collected. This research was conducted in accordance with relevant institutional guidelines and applicable data ethics regulations. Given the non-interventional nature of the study and the use of anonymized public data, formal ethics committee review was not required under applicable institutional policy; nonetheless, all procedures were carried out in accordance with the ethical principles outlined in the Declaration of Helsinki.

ANIMAL RESEARCH SUBJECTS

This research did not involve the use of any animal subjects.

AUTHOR CONTRIBUTIONS

Conceptualization – Long Wang and Yue Zhang; methodology – Long Wang and Yue Zhang; investigation – Long Wang and Yue Zhang; writing-original draft preparation – Long Wang and Yue Zhang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] G. I. Eskiadli and N. Panagiotou, "Virtual production in news media: Transforming storytelling, audience engagement, and ethical practices in the digital age," *Convergence*, vol. 32, no. 2, pp. 390–409, 2026, doi: 10.1177/13548565251393886.
- [2] Z. G. Ana, E. Petra, and V. Ante, "Digital storytelling as a method for modern business and economics education in the age of artificial intelligence," *ENTRENOVA - ENTERPRISE RESEARCH INNOVATION J.*, vol. 11, no. 1, 2025, doi: 10.54820/ENTRENOVA-2025-0087.
- [3] E. Segev, *Information and Power: Popular and Personal Storytelling in the Digital Age*. London, U.K.: Taylor & Francis, 2024, doi: 10.4324/9781003605683.
- [4] S. Xiaoyi, "Book review: Transmedia Storytelling in East Asia: The Age of Digital Media," *New Media Soc.*, vol. 25, no. 12, pp. 3735–3737, 2023, doi: 10.1177/14614448231187635.
- [5] B. Shishko, "Storytelling in the digital era: perspectives on age and gender," *Trames J. Humanities Soc. Sci.*, vol. 26, no. 4, pp. 397–411, 2022, doi: 10.3176/TR.2022.4.03.
- [6] Y. Shan, Y. Xie, Z. Li, et al., "Driven by engagement and social identity: refining the SOR framework to explore digital music consumption behaviors among Chinese undergraduate students on short video platforms," *Front. Psychol.*, vol. 17, Art. no. 1758762, 2026, doi: 10.3389/FP-SYG.2026.1758762.
- [7] R. Chen, "Cross-media narratives of city image: The case of Chongqing, China," in *Proc. Int. Conf. Advances Internet Marketing Bus. Manage. (ICAMM)*, 2023, pp. 163–168, doi: 10.26914/c.cnkihy.2023.008091.
- [8] K. Hayawi and S. Shahriar, "Generative AI for text-to-video generation: Recent advances and future directions," *Digital*, vol. 6, no. 1, Art. no. 23, 2026, doi: 10.3390/DIGITAL6010023.
- [9] J. Piatkowski, L. Karbowski, and J. Depta, "A lightweight algorithm for synchronized multimodal data acquisition using temporal sample alignment," *Sci. Rep.*, vol. 15, no. 1, Art. no. 22711, 2025, doi: 10.1038/S41598-025-07797-7.
- [10] Z. Hongmiao, Y. Xin, and J. Zhen, "Creative idea diffusion model in the multiplex network with consideration of multiple channels," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 97, Art. no. 105734, 2021, doi: 10.1016/J.CNSNS.2021.105734.
- [11] R. T. B. Maryala, "Data governance in generative AI: A framework for transparency, compliance, and ethical practice," *J. Comput. Sci. Technol. Stud.*, vol. 7, no. 3, pp. 964–971, 2025, doi: 10.32996/JC-STS.2025.7.3.108.
- [12] L. Lanting, "Constructing a multi-dimensional governance system for the application of generative AI in schools: Policy analysis and implications of the 'Australian Framework for Generative AI in Schools'," *J. Contemp. Educ. Res.*, vol. 9, no. 1, pp. 157–162, 2025, doi: 10.26689/JCER.V9I1.9540.
- [13] Y. Zhou, "The application status and trend analysis of personalized recommendation in short videos," *J. Innovation Develop.*, vol. 14, no. 2, pp. 118–123, 2026, doi: 10.54097/ZJ7YPG11.
- [14] L. Pettini, C. Bogler, C. Doeller, et al., "Synthesis and perceptual scaling of high-resolution naturalistic images using Stable Diffusion," *Behav. Res. Methods*, vol. 58, no. 1, Art. no. 24, 2025, doi: 10.3758/S13428-025-02889-8.
- [15] L. Feng, Z. Haihong, L. Zhenwei, et al., "An adaptive surface interpolation filter using cloth simulation and relief amplitude for airborne laser scanning data," *Remote Sens.*, vol. 13, no. 15, Art. no. 2938, 2021, doi: 10.3390/RS13152938.
- [16] X. Luo, G. Jiang, H. Cong, et al., "Cloth simulation with adaptive force model in three-dimensional space," *J. Eng. Fibers Fabrics*, vol. 13, no. 1, Art. no. 155892501801300105, 2018, doi: 10.1177/155892501801300105.
- [17] L. H. J. Ormaghi, R. N. Motta, F. M. Maciel, et al., "Smart fabric textiles: Recent advances and challenges," *Textiles*, vol. 2, no. 4, pp. 582–605, 2022, doi: 10.3390/TEXTILES2040034.
- [18] A. Mohamad and A. Reem, "The dilemma of the copyrights of artificial intelligence: The case of Saudi Arabia regulations," *Int. J. Sociotechnology Knowl. Develop.*, vol. 16, no. 1, pp. 1–15, 2024, doi: 10.4018/IJSKD.336920.

- [19] D. Lu, Z. Chen, C. Liu, et al., "Facial sketch synthesis with multi-level guided latent diffusion model," *Neurocomputing*, vol. 659, Art. no. 131734, 2026, doi: 10.1016/J.NEUCOM.2025.131734.
- [20] Y. Zhou, R. Towning, Z. Awad, et al., "Image synthesis with class-aware semantic diffusion models for surgical scene segmentation," *Healthcare Technol. Lett.*, vol. 12, no. 1, Art. no. e70003, 2025, doi: 10.1049/HTL2.70003.