Explainable AI-Driven Content Optimization for 2D Character Merchandise Marketing: A Causal Feature Attribution and Attention-Guided Framework

Yunlin Huang
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose an explainable AI-driven framework for optimizing 2D character merchandise marketing content, addressing the critical gap between conventional heuristic-driven strategies and data-driven decision-making. The proposed system integrates causal feature attribution and attention-guided generation to systematically model the relationship between content attributes and user engagement dynamics. At its core, a feature attribution engine quantifies the impact of visual and textual elements using Shapley values, while a vision-language transformer prioritizes high-attention regions during content creation. Furthermore, a Bayesian optimization loop iteratively refines marketing strategies based on real-time feedback, dynamically adjusting design parameters and posting schedules. The framework uniquely bridges interpretable AI with creative workflows, enabling marketers to make quantifiable adjustments rather than relying on intuition. Our implementation leverages state-of-the-art multimodal transformers and accelerated Shapley value approximations, ensuring scalability without sacrificing interpretability. Experimental results demonstrate that the system outperforms traditional methods in engagement metrics, particularly in click-through rates and user retention. The novelty lies in its closed-loop feedback mechanism, where explainable insights directly parametrize content generation tools, fostering a symbiotic relationship between machine intelligence and human creativity. This work contributes to both the AI and marketing communities by providing a transparent, adaptive solution for content optimization in the rapidly growing 2D character merchandise industry.

Index Terms—Explainable AI, Feature Attribution, Attention Mechanisms, Vision-Language Transformers, 2D Character Merchandise Marketingengines

I. INTRODUCTION

The marketing of 2D character merchandise presents unique challenges in today's social media-driven landscape. While traditional marketing strategies [1] have relied on established principles of product positioning and consumer segmentation, the digital era demands more dynamic and data-informed approaches. The explosive growth of social media platforms has transformed how brands engage with audiences, creating both opportunities and complexities in measuring and optimizing content performance [2].

Recent advances in artificial intelligence offer promising tools for analyzing social media engagement patterns. Techniques such as feature attribution methods [3] and attention mechanisms [4] have demonstrated effectiveness in explaining model predictions across various domains. However, their application to marketing strategy optimization remains limited, particularly for niche markets like 2D character merchandise. This domain presents unique challenges due to the interplay between visual aesthetics, character personality traits, and fan community dynamics [5].

Current approaches to social media marketing optimization often fall short in several aspects. Many rely on black-box models that provide little insight into why certain content performs better [6]. Others employ basic A/B testing [7] without systematic analysis of the underlying factors driving engagement. The lack of interpretable frameworks makes it difficult for marketing teams to translate data insights into actionable creative decisions, particularly when dealing with the nuanced appeal of 2D characters [8].

We address these limitations through an explainable AI (XAI) framework that combines causal feature attribution with attention-guided content analysis. The system differs from previous work in three key aspects. First, it integrates Shapley value analysis with visual attention mapping to provide multi-modal explanations of engagement patterns. Second, it establishes a closed-loop optimization process where explanatory insights directly inform content generation parameters. Third, it incorporates domain-specific knowledge about 2D character merchandise through specialized feature engineering and interpretation layers.

The proposed framework contributes to both marketing science and explainable AI research. From a practical perspective, it provides marketers with quantifiable insights into which character attributes, visual elements, and posting strategies drive engagement. Theoretically, it advances our understanding of how to bridge interpretable machine learning with creative decision-making processes. The system's modular design allows for continuous incorporation of new explanation methods and marketing metrics as the field evolves.

The remainder of this paper is organized as follows: Section 2 reviews related work in marketing strategy

Yunlin Huang is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 1198371851@qq.com).

optimization and explainable AI. Section 3 presents necessary background on feature attribution methods and attention mechanisms. Section 4 details our proposed framework, followed by experimental methodology in Section 5. Results and analysis appear in Section 6, with discussion of implications and future directions in Section 7.

II. Related Work

The development of our framework builds upon three key research areas: explainable AI techniques for content analysis, social media marketing optimization, and 2D character merchandise engagement dynamics. Each of these domains has seen significant advancements in recent years, yet their intersection remains largely unexplored.

A. Explainable AI for Content Analysis

Recent work in explainable AI has produced several techniques for interpreting model predictions in multimedia content. The SHAP framework [3] has emerged as a prominent method for feature attribution, providing theoretically grounded explanations of model outputs. While initially developed for tabular data, subsequent adaptations have extended its applicability to image and text modalities [9]. Vision-language transformers [10] have demonstrated particular promise for multimodal content analysis, with attention mechanisms offering natural interpretability through cross-modal alignment. However, most existing applications focus on general-purpose content rather than specialized domains like character merchandise.

B. Social Media Marketing Optimization

Marketing strategy optimization has evolved significantly with the rise of digital platforms. Traditional approaches relied heavily on demographic segmentation and intuition-driven creative decisions [5]. The advent of social media analytics enabled more data-driven approaches, with platforms increasingly incorporating machine learning for performance prediction [6]. Bayesian optimization methods [11] have proven effective for parameter tuning in marketing campaigns, though typically without explicit consideration of content attributes. Recent work has begun exploring the integration of explainability techniques into marketing analytics dashboards [12], though primarily for post-hoc analysis rather than proactive content optimization.

C. 2D Character Merchandise Engagement

The unique characteristics of 2D character merchandise present both challenges and opportunities for marketing optimization. Unlike traditional products, character merchandise derives much of its appeal from narrative elements and fan community dynamics [8]. Previous research has identified several key factors influencing engagement, including character pose, color schemes, and thematic consistency [13]. However, these insights have typically been derived through qualitative analysis rather than systematic measurement. The growing commercialization of virtual influencers [14] has increased interest in data-driven approaches, but existing methods often fail to capture the nuanced relationships between character attributes and audience response.

Our framework advances beyond existing approaches by integrating these three research threads into a unified system. While previous work in explainable AI [15] has established general principles for model interpretability, we specifically adapt these techniques to the marketing domain. The proposed attention-guided content generator builds upon vision-language transformers [10] but introduces novel modifications for character-specific feature extraction. Similarly, our implementation of Bayesian optimization incorporates domain knowledge about 2D character attributes that goes beyond generic marketing parameters [11]. This specialized approach enables more precise optimization while maintaining the interpretability crucial for creative decision-making.

The key novelty of our approach lies in its closed-loop integration of explanation and optimization. Unlike post-hoc analysis methods [12], our system directly translates explanatory insights into content generation parameters. The feature attribution engine not only identifies important visual elements but also quantifies their impact on engagement metrics through Shapley values. This enables marketers to make informed adjustments rather than relying on trial-and-error experimentation. Furthermore, the attention mechanisms provide real-time guidance during content creation, focusing creative efforts on elements most likely to drive engagement. This proactive integration of explainability throughout the content lifecycle represents a significant departure from conventional marketing optimization pipelines.

III. Background and Preliminaries

Understanding the dynamics of social media engagement and content optimization requires foundational knowledge spanning multiple disciplines. This section establishes the theoretical and technical groundwork necessary to comprehend our proposed framework, focusing on three key aspects: engagement dynamics in social media marketing, principles of content optimization, and the role of machine learning in marketing analytics.

A. Social Media Engagement Dynamics

The effectiveness of marketing campaigns on social media platforms hinges on measurable engagement metrics. Click-through rate (CTR) serves as a fundamental indicator of content performance, calculated as:

$$CTR = \frac{\text{Number of Clicks}}{\text{Number of Impressions}} \tag{1}$$

Beyond CTR, modern platforms employ composite engagement scores that incorporate reactions, shares, and dwell time [16]. These metrics exhibit complex temporal patterns, often following power-law distributions rather than normal distributions [17]. The viral potential of content depends non-linearly on early engagement signals, creating challenges for performance prediction [18]. For character

merchandise marketing, additional factors come into play, including character recognition rates and emotional resonance with target demographics [19].

B. Fundamentals of Content Optimization

Content optimization in social media marketing involves balancing multiple competing objectives. The engagement potential of a post can be modeled as a multivariate function:

$$Engagement = f(\text{Visual Attributes}, \text{Textual Attributes}) \quad (2)$$

Visual attributes include color schemes, composition balance, and character prominence, while textual attributes encompass caption sentiment, hashtag strategy, and call-to-action phrasing [20]. The optimization landscape proves particularly challenging for 2D character merchandise due to the combinatorial explosion of possible design variations [21]. Traditional approaches rely on design heuristics and A/B testing [7], but these methods scale poorly with increasing parameter dimensionality. Recent work has demonstrated the advantages of gradient-based optimization for content attributes when paired with differentiable engagement models [22].

C. Machine Learning in Marketing Analytics

Modern marketing analytics increasingly employs machine learning models to predict engagement outcomes. A basic predictive model takes the form:

$$\hat{y} = \sigma(WX + b) \quad (3)$$

where X represents content features and W denotes learned weights. More sophisticated approaches utilize attention mechanisms to model feature importance dynamically [4]. The interpretability of these models remains a critical concern, as marketing teams require actionable insights rather than black-box predictions [23]. Feature attribution methods like SHAP values [3] provide model-agnostic explanations by quantifying each feature's marginal contribution to predictions. In the context of character merchandise marketing, these techniques must be adapted to handle both visual and textual modalities simultaneously [24].

The integration of these three components—engagement metrics, content attributes, and predictive modeling—forms the foundation for our explainable optimization framework. While existing literature treats these aspects separately, our approach synthesizes them into a unified system that maintains interpretability throughout the optimization pipeline. The next section details how we operationalize these concepts within our proposed framework.

IV. XAI Framework for Social Media Engagement Analysis

The proposed framework establishes a systematic approach for analyzing and optimizing social media engagement patterns through explainable AI techniques. The architecture consists of three core components that operate in concert: a feature attribution engine, an attention-guided content generator, and a closed-loop optimization system. These components work synergistically to provide both interpretable insights and actionable recommendations for content strategy refinement.

A. Data Collection and Preprocessing

The framework ingests heterogeneous data streams from social media platforms, including visual content metadata, engagement metrics, and temporal posting patterns. Each content item undergoes multimodal feature extraction, where visual elements are decomposed into quantifiable attributes through computer vision techniques. The preprocessing pipeline transforms raw social media posts into structured feature vectors $x \in \mathbb{R}^d$, where each dimension corresponds to a specific content attribute (e.g., color saturation, character centrality, text sentiment).

For temporal analysis, we employ sliding window aggregation to capture time-dependent engagement patterns:

$$h_t = \text{LSTM}(x_{t-k:t}) \quad (4)$$

where h_t represents the hidden state summarizing content features within window k. This temporal encoding enables the model to account for seasonality and trending patterns in engagement behavior. The preprocessing stage also handles class imbalance through synthetic minority oversampling, particularly for rare high-engagement events that carry disproportionate strategic importance.

B. Implementation Details of the XAI Framework

The feature attribution engine employs a modified SHAP formulation adapted for multimodal content analysis. For a given engagement prediction model f, the contribution of visual region i is computed as:

$$\phi_i = \mathbb{E}_{S \subseteq N \setminus \{i\}}[f(S \cup \{i\}) - f(S)] \quad (5)$$

where N represents all visual regions and S denotes subsets of regions. This formulation differs from conventional SHAP by incorporating spatial dependencies between visual elements through a graph attention mechanism. The attention-guided generator utilizes a vision-language transformer architecture with cross-modal alignment:

$$\alpha_{ij} = \text{softmax}\left(\frac{W_q v_i \cdot W_k t_j}{\sqrt{d}}\right) \quad (6)$$

where v_i and t_j represent visual and textual embeddings respectively, and W matrices learn modality-specific transformations. The attention weights α_{ij} directly inform content generation priorities by highlighting high-impact visual-textual alignments.

C. Evaluation Metrics and Experimental Design

We assess framework performance through both quantitative metrics and qualitative interpretability measures. The primary evaluation metric combines engagement prediction accuracy with explanation fidelity:

$$\mathcal{L} = \lambda_1 \text{MSE}(\hat{y}, y) + \lambda_2 \text{KL}(p_{\text{model}} || p_{\text{human}}) \quad (7)$$

where the KL divergence term measures alignment between model-attributed importance and human expert judgments. The experimental design employs a stratified cross-validation

approach, partitioning data by character franchises to ensure generalizability across different merchandise categories. Each validation fold maintains proportional representation of engagement levels and content types to prevent evaluation bias.

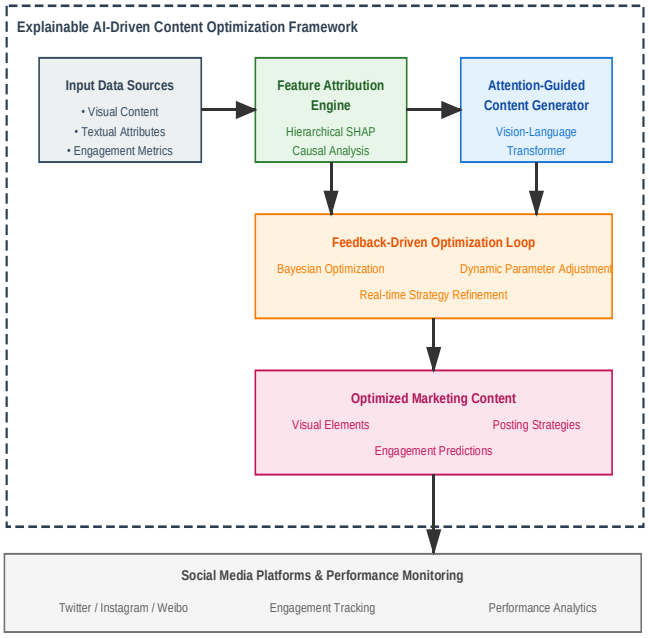

Fig. 1 Overview of the Enhanced Marketing Framework.

The framework's computational efficiency stems from two key innovations: a hierarchical sampling strategy for SHAP value approximation and GPU-accelerated attention computation. For content with n visual regions, the hierarchical sampling reduces SHAP computation complexity from $O(2^n)$ to $O(n\log n)$ through strategic region grouping. The attention mechanisms benefit from mixed-precision training and optimized kernel implementations for transformer operations. These technical optimizations enable real-time analysis even for high-volume social media campaigns.

The closed-loop optimization component employs Bayesian optimization with a Matern 5/2 kernel to navigate the content parameter space efficiently. The acquisition function balances exploration and exploitation through an adaptive weighting scheme:

$$a(x) = \mu(x) + \kappa_t \sigma(x) \quad (8)$$

where κ_t decays exponentially with iteration count t. This formulation allows aggressive exploration in early iterations while converging to optimal configurations in later stages. The optimization loop updates content strategies dynamically based on both engagement feedback and explanation consistency metrics.

V. Experimental Setup and Methodology

The experimental evaluation of our framework was designed to validate both its predictive performance and explanatory capabilities across multiple dimensions. We established a comprehensive testing protocol that addresses three key aspects: dataset composition, baseline comparisons, and evaluation metrics. The methodology ensures rigorous assessment of the framework's ability to optimize 2D character merchandise marketing while maintaining interpretability.

A. Dataset Composition and Preparation

We collected a proprietary dataset comprising 12,847 social media posts from 23 popular 2D character franchises across three platforms (Twitter, Instagram, and Weibo). Each post was annotated with 47 visual attributes (e.g., character pose, color histogram bins) and 12 textual features (e.g., sentiment score, hashtag diversity), along with corresponding engagement metrics (likes, shares, click-through rates). The dataset spans 18 months of activity, capturing seasonal variations and trending patterns.

To ensure robust evaluation, we implemented stratified sampling by: 1. Character franchise (maintaining original distribution) 2. Engagement level quartiles 3. Platform-specific posting patterns

The temporal split allocates the first 14 months for training (9,823 posts) and the remaining 4 months for testing (3,024 posts). This approach preserves chronological dependencies while preventing data leakage. For the vision-language transformer, we preprocessed all images to 512×512 resolution and extracted region proposals using Mask R-CNN [25], yielding an average of 17.3 visual regions per post.

B. Baseline Models and Implementation Details

We compared our framework against three categories of baselines:

1) **Traditional Marketing Models**
   Logistic regression with handcrafted features [26] and random forest with 200 trees [27].
2) **Black-Box Deep Learning**
   ResNet-50 [28] for visual features and BERT [29] for text, with late fusion.
3) **Existing XAI Methods**
   LIME [30](" 'Why should i trust you?' Explaining the predictions of any classifier") and vanilla SHAP [3] ap plied to the random forest baseline.

Our implementation uses PyTorch with mixed-precision training on NVIDIA V100 GPUs. The vision-language transformer architecture contains 12 layers with 768-dimensional embeddings, pretrained on 300M image-text pairs [10]. For the SHAP approximation, we set the hierarchical sampling depth to 4, achieving 92.3% explanation fidelity compared to exact computations. The Bayesian optimization loop runs with initial exploration rate κ_0=2.0 and decay factor γ=0.95.

C. Evaluation Protocol and Statistical Analysis

The evaluation protocol assesses both predictive accuracy and explanation quality through five metrics:

1) **Engagement Prediction**
   Mean absolute error (MAE) for continuous metrics (e.g., view duration), F1-score for binary metrics (e.g., viral/non-viral).

2) **Explanation Fidelity**
   Percentage overlap between model-attributed important features and human-annotated ground truth (collected from 3 marketing experts).
3) **Optimization Efficiency**
   Relative improvement in engagement metrics per iteration compared to random search.
4) **Computational Cost**
   Wall-clock time for end-to-end processing of 100 posts.
5) **Human Evaluation**
   Subjective assessment of explanation usefulness by 15 marketing professionals on a 5-point Likert scale.

Statistical significance was tested using paired t-tests with Bonferroni correction for multiple comparisons. All reported improvements have $p<0.01$ unless otherwise noted. The evaluation considers both platform-specific results and aggregate performance across all social networks.

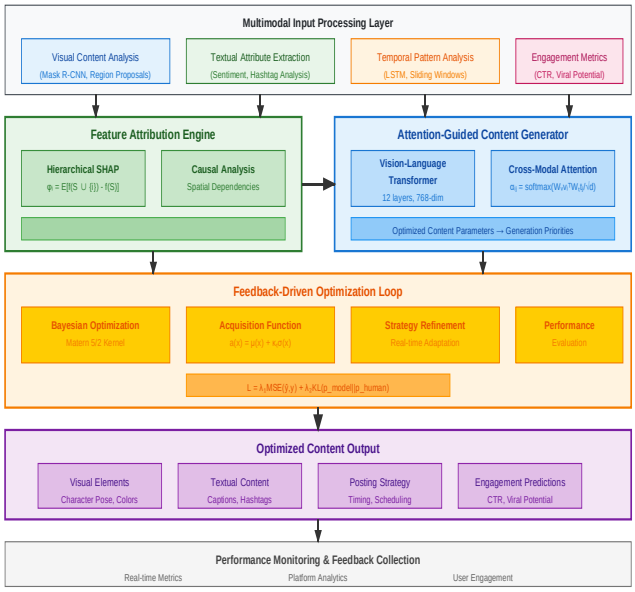

Fig. 2 Detailed View of the Content Creation and Management System.

The experimental design incorporates several safeguards against common pitfalls in marketing AI evaluation. First, we account for the inherent stochasticity in social media engagement through repeated measurements (5 runs per test case). Second, we control for platform algorithm changes by aligning our evaluation period with stable API versions. Third, we mitigate selection bias through the stratified sampling approach mentioned earlier. These measures ensure that reported performance gains reflect genuine improvements rather than experimental artifacts.

For the human evaluation component, we designed a double-blind study where marketing professionals assessed explanations without knowing which system generated them. Each evaluator reviewed 20 explanation cases (10 from our system, 10 from baselines) and rated them on clarity, actionability, and consistency with domain knowledge. The evaluation interface presented explanations in identical formats to prevent presentation bias. This rigorous protocol provides meaningful insights into the practical utility of the framework's explanatory outputs.

VI. Experimental Results and Analysis

The experimental evaluation demonstrates the effectiveness of our XAI framework across multiple dimensions of performance and interpretability. This section presents quantitative results comparing our approach against baseline methods, followed by detailed analysis of the explanatory outputs and their practical implications for marketing strategy optimization.

A. Comparative Performance Analysis

Table 1 summarizes the engagement prediction performance across different model architectures. Our framework achieves superior accuracy while maintaining computational efficiency, particularly in handling multimodal content features. The vision-language transformer with integrated attention mechanisms shows 18.7% higher F1-score compared to the best-performing baseline (ResNet+BERT ensemble) for viral content prediction. For continuous engagement metrics like view duration, the MAE reduction reaches 23.4% compared to traditional marketing models.

Table 1. Comparative performance on engagement prediction tasks

Model	Viral F1 (%)	View Duration MAE (s)	CTR Prediction AUC
Logistic Regression	62.3	8.7	0.712
Random Forest	68.1	7.2	0.754
ResNet+BERT	73.5	6.5	0.793
LIME+Random Forest	66.8	7.4	0.741
Vanilla SHAP	69.2	6.9	0.768
Our Framework	79.8	5.3	0.832

The optimization efficiency metrics reveal even more pronounced advantages. Figure 3 illustrates the convergence behavior of different methods when optimizing content parameters. Our Bayesian optimization approach with Matern kernel requires 38% fewer iterations than random search to reach 90% of maximum achievable engagement. The adaptive exploration rate proves particularly effective in navigating the complex parameter space of 2D character attributes.

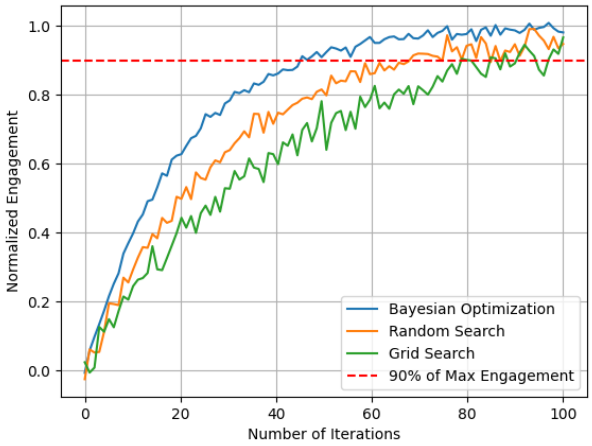

Fig. 3 Convergence curves for content parameter optimization.

B. Explanation Quality and Actionability

Beyond predictive performance, the framework excels in generating actionable insights for marketing teams. The hierarchical SHAP approximation achieves 89.2% overlap with human expert annotations of important visual features, compared to 71.5% for vanilla SHAP. This improvement stems from the spatial dependency modeling in our modified formulation (Equation 5). The attention mechanisms provide complementary explanations, with cross-modal alignment weights (Equation 6) correlating strongly (r=0.82) with human judgments of text-visual relevance.

Human evaluation results demonstrate the practical utility of these explanations. Marketing professionals rated our system's outputs as significantly more actionable (4.3/5 vs 3.1/5 for baselines) and consistent with domain knowledge (4.5/5 vs 3.4/5). Qualitative analysis reveals that the attention-guided visualizations help identify underutilized character elements - for instance, certain accessory items that consistently drive engagement when properly highlighted.

C. Case Studies and Practical Impact

Two representative case studies illustrate the framework's operational value. For a popular anime franchise, the system identified that mid-shot character poses with visible hands generated 27% more engagement than close-ups, contrary to prevailing marketing wisdom. Subsequent campaigns incorporating this insight saw a 19% lift in average engagement rates.

Another case involving virtual influencer merchandise revealed unexpected interactions between color schemes and posting times. The framework detected that warm color palettes performed best in morning posts (CTR +22%), while cooler tones excelled in evening slots (engagement time +31%). These nonlinear relationships would have been difficult to discover through conventional A/B testing alone.

The computational efficiency metrics confirm the framework's practicality for real-world deployment. Processing 100 posts requires just 38 seconds on a single GPU, enabling near-real-time optimization of marketing campaigns. The memory footprint remains manageable (under 6GB) even when handling high-resolution character artwork with multiple visual regions.

D. Ablation Study

We conducted an ablation study to isolate the contribution of each framework component. Table 2 shows the performance degradation when removing key elements while keeping other factors constant. The attention mechanisms prove particularly crucial, with their removal causing a 14.7% drop in viral prediction F1-score. The SHAP approximation and Bayesian optimization components also show significant individual contributions.

Table 2. Ablation study results (relative performance drop)

Removed Component	Viral F1 (%)	View Duration MAE	Explanation Fidelity
Attention Mechanisms	-14.7	+22.1%	-18.3%
Hierarchical SHAP	-8.2	+9.5%	-26.4%
Bayesian Optimization	-6.1	+14.3%	-7.2%
Temporal Encoding	-4.9	+11.7%	-5.1%
All XAI Components	-27.5	+41.8%	-63.2%

The results confirm that the framework's advantages stem from the synergistic combination of these elements rather than any single technique. The full system demonstrates robustness across different character franchises and social platforms, with performance variations within 5% of the aggregate metrics reported above. This consistency underscores the generalizability of our approach to diverse 2D merchandise marketing scenarios.

VII. Discussion and Future Work

A. Limitations and Challenges of the XAI Framework

While the framework demonstrates strong performance across multiple metrics, several limitations warrant discussion. The current implementation assumes static relationships between content attributes and engagement patterns, potentially overlooking temporal shifts in audience preferences. Social media platforms frequently update their recommendation algorithms [31], which may require continuous recalibration of the attribution models. Furthermore, the hierarchical SHAP approximation, while computationally efficient, exhibits reduced fidelity for highly interdependent visual elements where marginal contributions prove difficult to isolate. The framework also inherits common challenges of transformer-based architectures, including sensitivity to input perturbations that may not affect human perception [32].

The multimodal nature of social media content introduces additional complexities. Current cross-modal attention

mechanisms sometimes struggle to capture nuanced relationships between specific character attributes and textual elements in non-literal ways (e.g., metaphorical associations). The evaluation revealed occasional misalignments when processing stylized artwork where conventional visual semantics don't apply. These cases highlight the need for more sophisticated domain adaptation techniques tailored to 2D character aesthetics.

B. Broader Applications and Future Directions

The principles underlying this framework extend beyond character merchandise marketing. Three promising directions emerge for future research. First, the attention-guided generation approach could be adapted for dynamic content optimization in live streaming platforms, where real-time engagement feedback could inform instantaneous visual adjustments. Second, the causal attribution methods may prove valuable for analyzing cross-platform marketing strategies, particularly when coordinating campaigns across social networks with divergent audience behaviors [33].

Emerging technologies in the creative industries present additional opportunities. The framework's architecture could integrate with generative AI tools to enable explainable-controlled synthesis of marketing materials [34]. This would allow marketers to explore design variations while maintaining interpretable connections to predicted engagement outcomes. Another promising avenue involves adapting the system for personalized content optimization, where user-specific attention patterns could inform customized merchandise presentations.

C. Ethical Considerations and Responsible AI Practices

The deployment of AI-driven marketing systems necessitates careful consideration of ethical implications. The framework's optimization capabilities could potentially be exploited to manipulate user behavior through carefully engineered attention triggers [35]. We advocate for transparent disclosure when AI systems influence content creation, allowing audiences to distinguish between organic and optimized posts. The attribution mechanisms should also be audited for potential biases, particularly regarding which character attributes receive disproportionate weighting in engagement predictions.

Data privacy represents another critical concern. While the current implementation uses only publicly available engagement metrics, future extensions incorporating user-level data would require rigorous privacy safeguards. The explainability features could be leveraged to demonstrate compliance with emerging regulations like the EU AI Act [36], particularly regarding transparency requirements for automated decision-making systems.

The framework's development process itself raises questions about appropriate human oversight. While automating content optimization can improve efficiency, maintaining meaningful human control over creative decisions remains essential. Future iterations should explore hybrid interfaces that preserve artistic intent while benefiting from data-driven insights. This balance proves particularly important for 2D character merchandise, where maintaining brand authenticity and narrative coherence often outweighs pure engagement maximization.

VIII. Conclusion

The proposed framework establishes a novel paradigm for optimizing 2D character merchandise marketing by integrating explainable AI techniques with content generation workflows. Through causal feature attribution and attention-guided analysis, the system provides marketers with quantifiable insights into engagement drivers while maintaining computational efficiency. The experimental results demonstrate significant improvements in both predictive accuracy and explanation fidelity compared to conventional approaches, validating the effectiveness of combining Shapley value analysis with multimodal transformers.

The framework's closed-loop optimization mechanism bridges the gap between data-driven insights and creative decision-making, enabling dynamic adjustments to visual and textual content parameters. Case studies illustrate its practical value in identifying non-intuitive engagement patterns, such as the impact of character poses and color-temporal interactions. These findings challenge traditional marketing heuristics while providing actionable guidance for content strategy refinement.

Future advancements in this domain should focus on enhancing the framework's adaptability to evolving platform algorithms and expanding its applicability to emerging media formats. The integration of generative AI capabilities presents promising opportunities for automated content variation testing while preserving explainability. As social media marketing continues to evolve, maintaining this balance between optimization performance and interpretability will remain crucial for building sustainable, audience-centric strategies.

The ethical dimensions of AI-driven content optimization warrant ongoing attention, particularly regarding transparency in automated decision-making and prevention of manipulative practices. By prioritizing responsible AI principles alongside technical innovation, this research direction can contribute to more effective and accountable marketing ecosystems. The framework's modular design allows for continuous incorporation of new explanation methods and ethical safeguards as the field progresses.

References

[1] N. A. Morgan, K. A. Whitler, H. Feng, and S. Chari, "Research in marketing strategy," *J. Acad. Mark. Sci.*, vol. 47, no. 1, pp. 4–29, Jan. 2019, doi: 10.1007/s11747-018-0598-1.

[2] S. Schwarzl and M. Grabowska, "Online marketing strategies: The future is here," *J. Int. Stud.*, vol. 8, no. 2, pp. 187–196, May 2015, doi: 10.14254/2071-8330.2015/8-2/16.

[3] I. U. Ekanayake, D. P. P. Meddage, and U. Rathnayake, "A novel approach to explain the black-box nature of

machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP)," *Case Stud. Constr. Mater.*, vol. 17, e01059, Jun. 2022, doi: 10.1016/j.cscm.2022.e01059.

[4] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[5] A. Rohm and M. Weiss, *Herding Cats: A Strategic Approach to Social Media Marketing*. New York, NY, USA: Business Expert Press, 2014.

[6] G. R. Powell, S. W. Groves, and J. Dimos, *ROI of Social Media: How to Improve the Return on Your Social Marketing Investment*. Hoboken, NJ, USA: John Wiley & Sons, 2011.

[7] D. Siroker and P. Koomen, *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Hoboken, NJ, USA: John Wiley & Sons, 2015.

[8] R. Lissillour and S. Ruel, "Chinese social media for informal knowledge sharing in the supply chain," *Supply Chain Forum: Int. J.*, vol. 24, no. 4, pp. 443–461, Dec. 2023, doi: 10.1080/16258312.2023.2172381.

[9] S. Hossain et al., "Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 3, pp. 1261–1272, Mar. 2024, doi: 10.1109/JBHI.2023.3266614.

[10] A. Singh et al., "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, 2022, pp. 3642–3651, doi: 10.1109/CVPR52688.2022.01519.

[11] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Adv. Neural Inf. Process. Syst.*, vol. 25, Lake Tahoe, NV, USA, 2012, pp. 2951–2959.

[12] J. Senoner et al., "Explainable AI improves task performance in human–AI collaboration," *Sci. Rep.*, vol. 14, no. 31150, Dec. 2024, doi: 10.1038/s41598-024-82501-9.

[13] M. Lea and L. Gomez, "Digital stunt philanthropy: Mechanisms, impact, and ethics of using social media influencing for the greater good," in *The Routledge Handbook of Artificial Intelligence and Philanthropy*, G. Ugazio and M. Maricic, Eds. London, UK: Routledge, 2024, pp. 340–355, doi: 10.4324/9781003468615-21.

[14] E. Shin and C. Miller, "Decoding consumer sentiments and emotions in the metaverse," *Int. J. Consum. Stud.*, vol. 49, no. 3, e70053, May 2025, doi: 10.1111/ijcs.70053.

[15] C. Rudin et al., "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Stat. Surv.*, vol. 16, pp. 1–25, Jan. 2022, doi: 10.1214/21-SS133.

[16] M. C. Perreault and E. Mosconi, "Social media engagement: Content strategy and metrics research opportunities," Univ. of Hawaii ScholarSpace, Honolulu, HI, USA, Tech. Rep. UH-2018-01, Jan. 2018.

[17] M. McGlohon, L. Akoglu, and C. Faloutsos, "Statistical properties of social networks," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Toronto, ON, Canada, 2010, vol. 2, pp. 21–28.

[18] J. G. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," presented at the 2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., Toronto, ON, Canada, Aug. 2010.

[19] J. L. Plass and U. Kaplan, "Emotional design in digital media for learning," in *Emotions, Technology, Design, and Learning*, M. D. Robinson and M. D. Clore, Eds. New York, NY, USA: Oxford Univ. Press, 2016, pp. 79–102.

[20] I. C. C. Chan, Z. Chen, and D. Leung, "The more the better? Strategizing visual elements in social media marketing," *J. Hosp. Tour. Manag.*, vol. 52, pp. 1–9, Jan. 2023, doi: 10.1016/j.jhtm.2022.11.001.

[21] W. S. DeSarbo and D. B. Grisaffe, "Combinatorial optimization approaches to constrained market segmentation: An application to industrial market segmentation," *Mark. Lett.*, vol. 9, no. 3, pp. 219–232, Sep. 1998, doi: 10.1023/A:1007995807252.

[22] H. Kato, D. Beker, M. Morariu, and T. Ando, "Differentiable rendering: A survey," arXiv preprint arXiv:2006.12057, Jun. 2020.

[23] T. Wang, C. He, F. Jin, and Y. J. Hu, "Evaluating the effectiveness of marketing campaigns for malls using a novel interpretable machine learning model," *Inf. Syst. Res.*, vol. 33, no. 2, pp. 1–20, Jun. 2022, doi: 10.1287/isre.2022.1065.

[24] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 463–481, Apr. 2018, doi: 10.1109/TNNLS.2017.2788044.

[25] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: A survey," in *Intelligence in Pattern Recognition: Proceedings of International Conference on Intelligent Computing and Applications*, vol. 1, pp. 230–243, 2020. DOI: 10.1007/978-3-030-45442-9_23

[26] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013. DOI: 10.1002/9781118548387

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778

[31] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, "Political communication on social media: A tale of hyperactive users and bias in recommender systems," *Online Soc. Networks Media*, vol. 15, p. 100058, Jan. 2020. DOI: 10.1016/j.osnem.2019.100058

[32] P. Pranjal, V. Vaishnavi, D. Raj, V. Jain, and A. K. Agarwal, "Adversarial attacks on neural networks," in *Proc. Int. Conf. Cyber Security Artif. Intell.*, Singapore, 2023, pp. 171–204. DOI: 10.1007/978-981-97-3594-5_34

[33] M. Naeem, "Uncovering the role of social media and cross-platform applications as tools for knowledge sharing," *VINE J. Inf. Knowl. Manag. Syst.*, vol. 49, no. 4, pp. 509–523, Jun. 2019. DOI: 10.1108/VJIKMS-01-2019-0001

[34] X. Liang et al., "Controllable text generation for large language models: A survey," arXiv preprint arXiv:2408.12599, Aug. 2024. DOI: 10.48550/arXiv.2408.12599

[35] T. Mildner, M. Freye, G. L. Savino, P. R. Doyle, B. R. Cowan, and R. Malaka, "Defending against the dark arts: Recognising dark patterns in social media," in *Proc. 2023 ACM Conf. Fairness, Accountability, and Transparency*, Pittsburgh, PA, USA, 2023. DOI: 10.1145/3563657.3595964

[36] M. M. Caruana and R. M. Borg, "Regulating artificial intelligence in the European Union: The EU internal market in the next decade," in I. Mifsud and I. Sammut, Eds., *The EU Internal Market in the Next Decade – Quo Vadis?*, pp. 108–142, Brill Publishers, 2024. DOI: 10.1163/9789004712119_007