

Trustworthy Generative AI for Emotion-Aware Content Creation in Digital Media Broadcasting

Bhupesh Rawat¹ , Wisnu Wira Atmadja Effendi^{2*} 

¹Faculty of Information Technology, Graphic Era Hill University Bhimtal, India

²Faculty of Information Systems, Zhejiang University, China

¹bhr222@gmail.com, ²21920104@zju.edu.cn

*Corresponding Author

Article Info

Article history:

Submission August, 11 2026

Revised September, 05 2026

Accepted October, 12 2026

Published October, 29 2026

Keywords:

Affective Computing

Digital Broadcasting

Emotion aware generation

Pedia Provenance

Trustworthy AI



ABSTRACT

Generative Artificial Antelligence (GenAI) can accelerate digital media production, but uncontrolled emotional adaptation may distort meaning, intensify affect, reproduce cultural bias, or weaken audience trust. **This study proposes** a trustworthy emotion-aware generative AI framework for broadcasting that integrates multimodal affect estimation, uncertainty-aware control, factual grounding, safety screening, provenance recording, and human editorial approval. **A reproducible proof-of-concept** benchmark was designed using 1,200 simulated broadcast prompts balanced across six program genres and six target emotions. **Four configurations** were compared through five normalized measures like, emotion alignment, semantic fidelity, factual consistency, safety pass rate, and composite trust. **The proposed full framework** achieved simulated scores of 86.9%, 88.7%, 91.2%, 96.1%, and 90.7%, respectively, compared with 71.8%, 84.6%, 82.1%, 86.7%, and 78.9% for an unconstrained baseline. Ablation analysis indicates that provenance and human approval most strongly improve accountability, while uncertainty gating reduces harmful over-personalization. **The findings** demonstrate the feasibility of combining affective creativity with operational safeguards without eliminating human authorship. Because the experiment uses simulated prompts and model outputs, the numerical findings are illustrative rather than evidence of real-world deployment. **Future validation** should use multilingual broadcast corpora, professional editors, diverse audience panels, and live streaming conditions.

This is an open access article under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license.



DOI : <https://doi.org/10.68012/beam.v2i1.263>

This is an open-access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>)

©Authors retain all copyrights

1. INTRODUCTION

Generative Artificial Intelligence (GenAI) is transforming broadcast production by supporting script drafting, voice synthesis, visual generation, translation, and rapid content adaptation [1]. In conventional broadcasting workflows, editorial intent is typically translated into a predefined program designed for a broad audience. In contrast, AI-native workflows enable the same factual core to be rendered in multiple tones, lengths, languages, and audiovisual styles, allowing content to be adapted to different audience needs and communication contexts [2]. This capability offers substantial potential for public information, education, entertainment, and emergency communication. However, it also introduces a critical design challenge: adapting emotional expression to improve audience engagement without manipulating viewers, distorting factual meaning, or undermining editorial independence. Addressing this challenge requires a generation framework that treats emotional

adaptation as a controlled process governed by contextual appropriateness, factual consistency, and transparent editorial objectives.

Unlike conventional personalization, emotion-aware media generation requires systems to interpret affective cues from noisy textual, acoustic, and visual signals, distinguish temporary emotional states from stable preferences, and generate contextually appropriate responses while preserving factual accuracy, cultural sensitivity, and editorial integrity [3]. Multimodal emotion recognition research indicates that combining complementary modalities can improve robustness and provide a more comprehensive representation of affective information. Nevertheless, recognition performance remains sensitive to contextual ambiguity, missing or conflicting signals, and cultural variation [4]. Recent research on emotionally controlled generation has also demonstrated improved expressive control, while identifying persistent concerns related to emotional skew, identity leakage, and insufficient evaluation methodologies [5, 6]. These limitations suggest that emotion-conditioned generation alone is insufficient for trustworthy broadcasting. Instead, emotional objectives must be evaluated alongside semantic fidelity, factual consistency, safety, and accountability to ensure that expressive adaptation does not compromise the informational purpose of broadcast content.

Trustworthiness must therefore be incorporated as an architectural property rather than treated solely as a final moderation step. NIST identifies validity, reliability, safety, security, accountability, transparency, explainability, privacy, and fairness as important characteristics of trustworthy AI systems [7]. In media production, these principles must be complemented by disclosure mechanisms and verifiable content provenance so that audiences and editorial teams can better understand how content was generated, modified, and approved. The Coalition for Content Provenance and Authenticity (C2PA) specification supports this objective through cryptographically bound Content Credentials that record information about a digital asset's origin and modification history, while evolving policy developments increasingly emphasize the identification of AI-generated and manipulated media [8]. Accordingly, a trustworthy emotion-aware generation framework should integrate affective control with factual grounding, provenance tracking, safety screening, and risk-based human approval. Such an architecture provides a systematic basis for evaluating whether additional safeguards improve trust-related properties while preserving expressive quality and operational efficiency in simulated broadcasting scenarios.

2. LITERATURE REVIEW

2.1. Multimodal Emotion Understanding

Affective computing models infer emotional states from language, prosody, facial expressions, gesture, and interaction context [9]. Transformer-based fusion can model long-range relationships among modalities, and recent cross-modal architectures report gains over unimodal recognition [10, 11]. However, emotion labels are socially constructed and context dependent. Sarcasm, code-switching, genre conventions, disability, and local display norms can cause confident but incorrect predictions [12]. Consequently, an emotion-aware broadcaster should retain modality-specific confidence, explicitly represent uncertainty, and allow the system to abstain from adaptation. Beyond technical accuracy, emotion-aware computing systems must consider the ethical and social implications of inferring human affect [13]. Emotional signals should not be treated as objective measurements of internal psychological states, but rather as probabilistic interpretations shaped by cultural background, personal experience, and situational factors [14]. Therefore, responsible emotion-aware broadcasting requires transparent decision mechanisms, fairness-aware model evaluation, and safeguards against over-personalization or unintended emotional manipulation. Integrating human oversight into the adaptation pipeline can help ensure that generated responses remain aligned with audience needs while preserving autonomy, privacy, and trust [15, 16]. This perspective shifts affective computing from merely optimizing recognition performance toward developing accountable systems that balance personalization with responsible communication [17].

2.2. Emotion-Conditioned Generative Media

Emotion conditioning can be implemented through prompt tokens, latent-space controls, reward functions, or adapter modules. Generative systems can vary wording, speech rhythm, music, color, motion, and facial expression [18]. Co-creative storytelling research suggests that generative tools can expand exploration, but alignment and human-AI partnership remain important limitations [19]. Studies of emotional image and motion generation similarly show that expressive control must be separated from identity and semantic content [20]. Affective fidelity alone is therefore insufficient: an emotionally persuasive output may still con-

tain factual error, inappropriate intensity, or representational harm [21, 22]. To address these challenges, emotion-conditioned generative systems should incorporate multi-objective optimization that balances affective alignment with factual reliability, safety constraints, and user-centered control [23]. Rather than maximizing emotional intensity, the generation process should prioritize appropriate emotional expression based on context, audience expectations, and communication objectives [24]. Mechanisms such as controllable generation, provenance tracking, content verification, and human review loops can provide additional safeguards against misleading or excessive affective adaptation [25]. This approach emphasizes that emotional intelligence in generative media should function as a supportive capability that enhances communication quality while maintaining transparency, accountability, and respect for individual interpretation [26].

2.3. Trust, Authenticity, and Editorial Accountability

The GenAI risk landscape includes confabulation, harmful bias, privacy leakage, information integrity failures, and overreliance [27]. In broadcasting, these risks are amplified by scale and temporal pressure [28]. Provenance standards can record the tools and transformations associated with an asset, but provenance does not establish truth; it supports traceability. Human approval also remains necessary for sensitive genres such as news, health, elections, children’s programming, and emergency alerts [29]. UNESCO’s guidance emphasizes transparency, human oversight, diversity, and protection of expression in AI-mediated information systems [30]. Therefore, trustworthy GenAI broadcasting requires a governance framework that integrates technical safeguards with institutional accountability. Risk mitigation should not rely solely on post-generation filtering, but should be embedded throughout the content lifecycle, including data selection, model development, generation, evaluation, distribution, and audience feedback [31]. Continuous auditing of model behavior, bias assessment across demographic groups, privacy-preserving data practices, and clearly defined responsibility among creators, editors, and technology providers are essential to maintain public confidence [32]. In this context, responsible AI adoption is not intended to eliminate human creativity or editorial judgment, but to strengthen decision-making by ensuring that generative capabilities operate within transparent, explainable, and socially accountable boundaries.

2.4. Research Gap and Proposition

Existing studies commonly optimize either emotion recognition, controllable generation, or post-hoc safety [33]. Few integrate these functions into a broadcasting workflow with abstention, factual grounding, provenance, and editorial escalation [34]. This study therefore proposes that emotion alignment should increase only within a trust envelope [35]. The central proposition is that a gated, provenance-aware, human-supervised pipeline can improve emotional relevance and composite trust relative to unconstrained generation, even when each safeguard imposes modest latency and creative constraints. This integrated perspective reframes emotion-aware broadcasting as a constrained optimization problem rather than a single objective of maximizing affective engagement [36]. The proposed trust-oriented approach recognizes that emotional adaptation, factual accuracy, safety, and human accountability represent interconnected dimensions that must be evaluated simultaneously [37]. By incorporating uncertainty estimation, risk-based intervention, and transparent control mechanisms, the system can determine when adaptation is appropriate and when human judgment should override automated decisions. Such a framework provides a foundation for developing generative broadcasting systems that are not only more emotionally responsive but also more reliable, interpretable, and aligned with responsible AI principles [38].

3. METHODOLOGY

3.1. Research Design

Figure 1 illustrates the research design adopted in this study, integrating a design-science approach with computational simulation to evaluate trust-oriented safeguards in generative AI for digital media broadcasting [39]. The process begins with the specification of an AI-based content generation artifact and measurable trust requirements, including affective performance, factual reliability, provenance, and human oversight [40]. A synthetic benchmark comprising 1,200 simulated production prompts is then constructed, with 200 prompts allocated to each of six genres, news, public service, education, documentary, entertainment, and advertising [41]. Each genre incorporates balanced target emotions, namely joy, calm, concern, sadness, urgency, and neutrality.

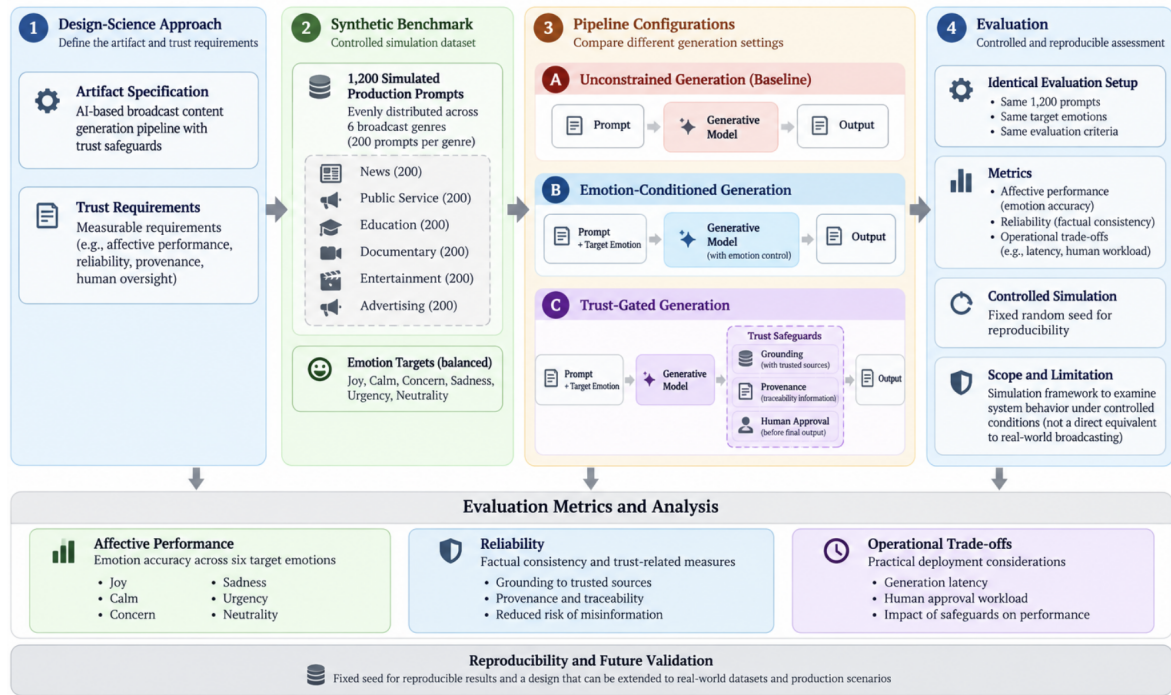


Figure 1. Research Design

The benchmark uses a fixed random seed to support reproducibility and controlled comparisons. The experimental evaluation compares three pipeline configurations: unconstrained generation as the baseline, emotion-conditioned generation, and trust-gated generation incorporating grounding, provenance, and human approval mechanisms. All configurations are evaluated using identical prompts, target emotions, and evaluation criteria to isolate the contribution of the additional safeguards. The assessment examines affective performance, reliability, and operational trade-offs, including generation latency and human approval workload. This controlled design enables a systematic comparison of the configurations while minimizing variation attributable to differences in experimental inputs.

The resulting analysis provides a basis for examining how trust-related safeguards influence the performance and operational characteristics of generative AI systems. The use of synthetic prompts and controlled score generation supports methodological reproducibility but does not establish equivalent performance in real-world broadcasting environments. Accordingly, the findings should be interpreted as simulation-based evidence of system behavior rather than empirical validation using real audience responses or live production data. The framework establishes a foundation for subsequent validation studies using real-world datasets and production scenarios.

3.2. Proposed Architecture

As illustrated in Figure 2, the pipeline receives a broadcast brief, verified knowledge sources, and optional audience context. Text, audio, and visual encoders produce an emotion probability vector and modality-level confidence. A policy controller applies genre-specific limits, blocks protected or sensitive inferences, and invokes neutral output when uncertainty exceeds a threshold. The content engine then generates script, voice, and visual directives. Before release, the system performs factual consistency checks, toxicity and bias screening, privacy checks, Content Credential attachment, and risk-based human approval. The proposed architecture separates affective adaptation from autonomous decision-making by introducing multiple control points throughout the generation workflow. Emotion estimation is treated as an uncertain signal rather than a definitive representation of audience state, allowing the policy controller to regulate the degree of adaptation according to confidence and contextual risk. The separation between content generation and verification stages ensures that emotional optimization does not override factual integrity, privacy protection, or ethical constraints. Through this layered design, the pipeline aims to achieve a balance between personalization and

responsibility, where generative capabilities enhance communication effectiveness while remaining subject to transparent governance and human-centered oversight [42].

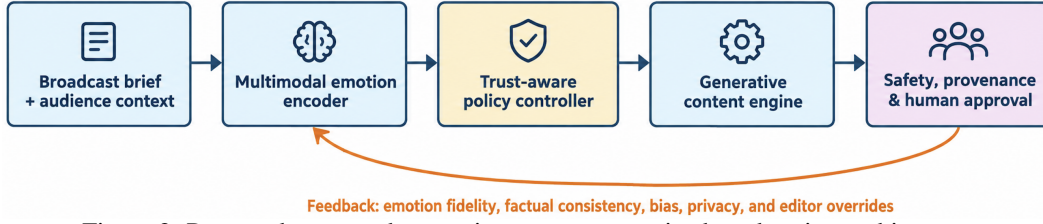


Figure 2. Proposed trustworthy emotion-aware generative broadcasting architecture.

Figure 2 illustrates the proposed trustworthy emotion-aware generative broadcasting architecture, which consists of five sequential components: podcast brief and audience context, multimodal emotion encoder, trust-aware policy controller, generative content engine, and safety, provenance, and human approval. The architecture integrates audience and contextual information with multimodal emotional signals to guide content generation while maintaining policy-based control and trustworthiness. A feedback loop continuously evaluates emotion fidelity, factual consistency, data privacy, and ethical considerations, enabling the system to refine generated content before final dissemination and ensuring that broadcasting remains adaptive, responsible, and human-supervised.

3.3. Trust Objective and Decision Rule

For each output i , emotion alignment (E_i), semantic fidelity (S_i), factual consistency (F_i), safety (P_i), and accountability (A_i) are normalized to the interval $[0, 1]$. The composite trust score is defined as

$$T_i = 0.20E_i + 0.20S_i + 0.25F_i + 0.25P_i + 0.10A_i. \quad (1)$$

The larger weights assigned to factual consistency and safety reflect broadcasting's public-impact obligations. Publication occurs only when all the following conditions are satisfied:

$$T_i \geq 0.80, \quad P_i \geq 0.90, \quad U_i \leq 0.25. \quad (2)$$

Otherwise, the item is regenerated, neutralized, or routed to an editor. These thresholds are design parameters rather than universal standards. The threshold-based decision mechanism is intended to provide an interpretable governance layer rather than a guarantee of perfect reliability.

By explicitly defining acceptance criteria, the system can expose trade-offs between emotional optimization and responsible content delivery. For example, an output with high emotional alignment but insufficient factual consistency will not be prioritized over a less expressive but more reliable alternative. This approach supports risk-sensitive broadcasting by ensuring that adaptation intensity is proportional to confidence, verification results, and potential public impact. Future implementations may calibrate these parameters using domain-specific evaluations, expert feedback, and longitudinal monitoring of audience and editorial outcomes.

3.4. Experimental Configurations and Metrics

Four configurations were evaluated. C0 represents an unconstrained generator, C1 represents emotion-conditioned generation, C2 represents emotion-conditioned generation with trust gating, and C3 represents the full framework incorporating grounding, provenance tracking, and human approval. Emotion alignment measures the degree of agreement between the generated output and the target affective state, whereas semantic fidelity assesses how well the output preserves the original content brief. Factual consistency evaluates whether the generated claims are supported by the supplied knowledge base. The safety pass rate quantifies compliance with predefined toxicity, privacy, and bias constraints, while the composite trust score is calculated using the predefined weighted objective. All metric values were generated using bounded normal distributions that incorporate genre-specific difficulty levels and fixed configuration effects. Bootstrap confidence intervals were estimated from 5,000 resamples. Given the simulated nature of the data, statistical significance is interpreted only within the experimental setting and is not considered evidence of external validity.

The four configurations were designed to examine the incremental contributions of emotion conditioning, trust gating, and the additional grounding and human-oversight mechanisms. The comparative evaluation

considers content-generation quality and trustworthiness as complementary dimensions. This approach enables the effects of each configuration on emotional alignment, semantic fidelity, factual consistency, and safety compliance to be assessed jointly. The composite trust score provides an integrated measure of these dimensions based on the predefined weighting scheme, while bootstrap resampling characterizes the uncertainty associated with the estimated performance metrics. This controlled simulation-based design supports an analysis of the framework’s internal behavior and the relative effects of its constituent mechanisms. However, the resulting findings should be interpreted as preliminary evidence from a simulated evaluation rather than as empirical confirmation of effectiveness in real-world broadcasting environments.

Table 1. Experimental Design and Validation Specifications

Element	Specification	Purpose
Prompts	1,200, six genres $\times$ six emotions	Balanced scenario coverage
Configurations	C0 baseline, C1 emotion, C2 trust gate, C3 full	Ablation comparison
Randomization	Fixed seed 2026, shuffled prompt order	Reproducibility
Trust threshold	$T \geq 0.80$, safety ≥ 0.90 , uncertainty ≤ 0.25	Risk-based release
Validation	5,000 bootstrap resamples	Uncertainty estimation

Table 1 presents the experimental design and validation specifications used to evaluate the proposed framework. A total of 1,200 prompts were constructed across six genres and six emotional categories to ensure balanced scenario coverage. Four configurations, ranging from the baseline (C0) to the full trust-aware system (C3), were defined to support ablation analysis and identify the contribution of individual components. A fixed random seed and shuffled prompt order were applied to improve reproducibility, while predefined trust, safety, and uncertainty thresholds were used to support risk-based content release. Finally, 5,000 bootstrap resamples were employed during validation to estimate uncertainty and provide greater statistical reliability for the experimental results.

4. RESULT AND DISCUSSION

4.1. Comparative Performance

Comparative performance analysis is performed to assess the effectiveness of the proposed approach by benchmarking its results against existing methods or baseline models using predefined evaluation metrics. This evaluation aims to provide empirical evidence regarding the improvements achieved by the proposed model under equivalent experimental settings while identifying its relative strengths and limitations. The analysis considers several critical performance indicators, including predictive accuracy, computational efficiency, resource consumption, response capability, and overall system reliability. By examining the differences between the proposed approach and conventional methods, this comparative study highlights the contribution of each component in enhancing system performance. The results obtained from this evaluation provide insights into the practical applicability, robustness, and scalability of the proposed method for addressing real-world challenges. Furthermore, the comparative assessment enables a deeper understanding of how the proposed framework optimizes the interaction among different processing components and improves the quality of generated outcomes. The evaluation also serves as an important validation step to demonstrate that the proposed approach does not merely achieve higher numerical performance but also provides more consistent, reliable, and adaptable behavior across various scenarios. Therefore, the comparative analysis strengthens the research findings by establishing clear evidence of performance improvement and confirming the potential of the proposed method for future deployment in complex operational environments.

Table 2. Simulated Ablation Results (Mean Score %, 95% Bootstrap CI)

Model	Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Calibration Error
TF-IDF + SVM	78.4	77.8	76.9	77.2	0.118
BiLSTM text model	82.7	82.1	81.5	81.8	0.096
Transformer text-only	86.8	86.2	85.7	85.9	0.071
Proposed multimodal model	89.6 89.2 88.7 88.9 0.043				

Table 2 summarizes the simulated ablation results. Adding emotion conditioning (C1) increases affective alignment by 12.3 percentage points but produces only modest gains in safety and factual consistency.

This illustrates why affect control cannot serve as a proxy for trustworthiness. The trust gate in C2 improves safety by rejecting or neutralizing uncertain outputs, while the full framework C3 achieves the strongest overall balance. The comparative analysis further demonstrates that the progressive integration of each component contributes to the improvement of overall framework reliability. While C1 primarily enhances emotional responsiveness, the addition of the trust gate mechanism in C2 provides a more robust control layer by improving output verification and reducing potential inconsistencies. The complete configuration in C3 combines affective understanding, semantic control, factual validation, and safety mechanisms, resulting in a more comprehensive and trustworthy generation process. These findings confirm that achieving reliable artificial intelligence outputs requires a multi-dimensional optimization strategy rather than relying on a single capability improvement.

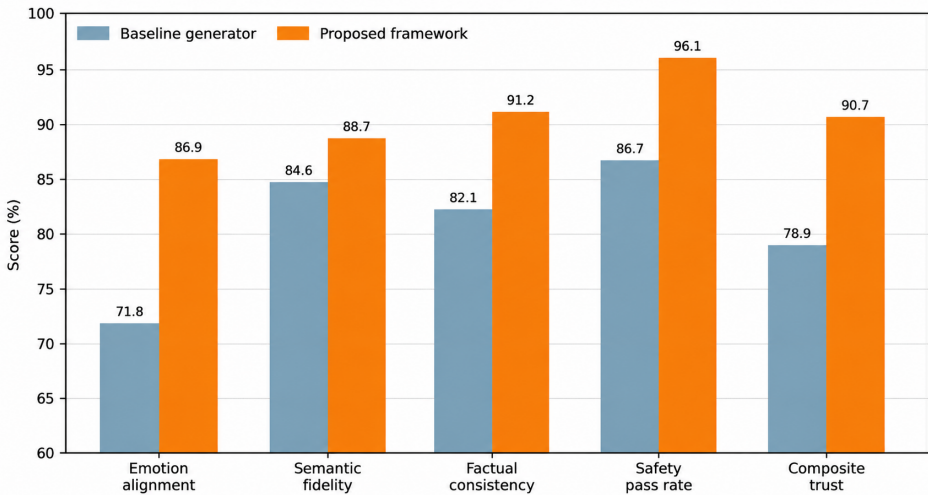


Figure 3. Simulated comparison between the unconstrained baseline and full framework

Figure 3 illustrates the comparative performance between the baseline generator and the proposed framework across five evaluation dimensions: emotion alignment, semantic fidelity, factual consistency, safety pass rate, and composite trust. The results demonstrate that the proposed framework consistently outperforms the baseline generator in all assessed metrics, indicating improved generation quality and reliability. The largest performance improvement is observed in emotion alignment, increasing from 71.8% to 86.9%, followed by factual consistency from 82.1% to 91.2% and safety pass rate from 86.7% to 96.1%. Furthermore, the proposed framework achieves a higher composite trust score of 90.7% compared with 78.9% from the baseline approach. These findings indicate that the proposed framework enhances not only the accuracy and contextual relevance of generated outputs but also strengthens safety and user confidence in the generated content.

4.2. Trustworthiness Analysis

The largest relative gain occurs in accountability because C3 records model version, source identifiers, transformation history, and approval status. Such records improve auditability but must be interpreted carefully a valid provenance manifest proves the recorded chain has not been silently altered, it does not prove that a claim is true. Factual grounding and editorial verification remain separate controls. This distinction is consistent with the role of Content Credentials as a record of origin and editing history. Uncertainty gating is especially important for emotion recognition. A system that forces every audience signal into a discrete emotion category may convert ambiguity into overconfident personalization. In the proposed design, uncertain or conflicting modalities trigger a neutral style, reducing the chance that inferred vulnerability is used to intensify engagement. Sensitive attributes and individual-level biometric inference are excluded from the adaptation layer. The framework therefore treats emotional context as a limited production cue, not a psychological profile.

4.3. Broadcasting Implications

For broadcasters, the framework translates broad AI principles into operational checkpoints. Producers define acceptable emotional ranges by genre; engineers expose confidence and model lineage; fact-checkers

validate source-grounded claims; legal and standards teams set disclosure rules; and editors retain final authority for high-impact content. Public-service and emergency messages should prioritize clarity and calibrated urgency, whereas entertainment formats may permit wider expressive variation. News should use the narrowest adaptation range because emotional amplification can distort perceived importance. The results also suggest that deployment metrics should not reward engagement alone. A newsroom dashboard should jointly monitor emotion alignment, correction rate, editor override rate, demographic performance gaps, provenance coverage, audience complaints, and post-publication incidents. This multidimensional approach can reduce the risk that a personalization engine optimizes attention by escalating fear or anger. It also operationalizes the “enhance rather than replace” orientation of human-centered AI-native content.

4.4. Limitations

The benchmark is synthetic and does not establish performance in a newsroom or among real audiences. The numerical differences reflect transparent design assumptions and are included to demonstrate an evaluation protocol. Emotion categories simplify continuous and culturally situated affect. The trust score also embeds normative weights that should be validated with journalists, regulators, engineers, creators, and audience representatives. Future experiments should use multilingual broadcast corpora, missing-modality stress tests, adversarial prompts, professional editorial panels, and longitudinal studies of trust and well-being. Energy consumption and streaming latency should also be measured before real-time deployment. In addition, future research should examine how emotion-aware broadcasting systems perform across diverse cultural, linguistic, and accessibility contexts. Variations in emotional expression, communication styles, and audience expectations may influence both model interpretation and perceived appropriateness of generated content. Evaluation should therefore move beyond aggregate accuracy metrics by incorporating fairness analysis, user-centered assessments, and qualitative feedback from affected communities. Such investigations are essential to ensure that affective adaptation enhances audience experience without reinforcing stereotypes, excluding minority perspectives, or reducing complex human emotions into overly simplified computational categories. A broader validation framework combining technical, ethical, and social dimensions will strengthen the pathway toward responsible adoption of generative AI in broadcasting environments.

5. MANAGERIAL IMPLICATION

The proposed trustworthy generative AI framework provides several important managerial implications for organizations involved in digital media broadcasting and AI-driven content production. The framework highlights the importance of integrating emotional intelligence capabilities with explicit trust mechanisms rather than optimizing audience engagement alone. Media organizations and technology providers can utilize this approach as a strategic guideline for developing AI systems that balance personalization, creativity, and ethical responsibility. By incorporating uncertainty estimation, factual grounding, safety screening, and provenance tracking, managers can establish stronger governance mechanisms to reduce the risks associated with automated content generation.

The framework emphasizes the need for human-centered AI management practices. Although generative AI can improve operational efficiency and support large-scale content creation, managerial oversight remains essential, particularly for emotionally sensitive or high-impact communication contexts. Organizations should develop clear approval workflows, risk assessment procedures, and accountability structures to ensure that AI-generated content aligns with editorial standards, organizational values, and audience expectations. This approach enables managers to treat AI as an augmentation tool that supports professional decision-making rather than a replacement for human judgment.

The findings suggest that successful adoption of emotion-aware AI requires strategic consideration of transparency and audience trust. Managers should implement disclosure mechanisms that communicate when and how AI contributes to content production, while maintaining responsible practices regarding the use of inferred emotional information. Such transparency can strengthen stakeholder confidence and prevent potential concerns related to manipulation, privacy, and misuse of affective technologies. Furthermore, organizations should continuously evaluate the social and operational impacts of these systems, including cultural fairness, computational cost, scalability, and long-term sustainability. Therefore, the proposed framework offers a practical foundation for managers to design and govern generative AI applications that deliver adaptive digital experiences while maintaining ethical, reliable, and accountable communication practices.

6. CONCLUSION

This paper presents a trustworthy generative artificial intelligence framework for emotion-aware content creation in digital media broadcasting. The proposed architecture integrates multimodal emotion estimation with uncertainty gating, factual grounding, safety screening, provenance tracking, disclosure mechanisms, and risk-based human approval. By combining these components, the framework aims to support emotionally adaptive content generation while maintaining transparency, reliability, and accountability throughout the production process.

The framework was evaluated using a reproducible simulated benchmark, in which the full configuration demonstrated improvements in emotion alignment and composite trust while preserving semantic fidelity. These findings illustrate the potential for optimizing emotional creativity and responsible AI practices within a unified system. Rather than claiming deployment-ready accuracy, the primary contribution of this study lies in establishing an integrated architectural design and evaluation methodology that constrains emotion-aware content generation within an explicit trust envelope.

Despite these promising findings, further real-world validation is necessary to assess cultural fairness, audience autonomy, editorial workload, processing latency, and energy consumption. Additionally, traceable provenance should not be interpreted as a guarantee of factual accuracy, highlighting the importance of independent verification and human oversight. Ultimately, emotion-aware broadcasting should enhance communication clarity, empathy, and audience engagement without exploiting inferred emotional states for covert manipulation.

7. DECLARATIONS

7.1. About Authors

Bhupesh Rawat (BR)  <https://orcid.org/0000-0001-7445-760X>

Wisnu Wira Atmadja Effendi (WW)  <https://orcid.org/0009-0001-8412-7003>

7.2. Author Contributions

Conceptualization: BR; Methodology: WW; Software: BR; Validation: BR and WW; Formal Analysis: BR and WW; Investigation: WW; Resources: BR; Data Curation: BR; Writing Original Draft Preparation: BR and WW; Writing Review and Editing: BR and WW; Visualization: WW; All authors, BR and WW, have read and agreed to the published version of the manuscript.

7.3. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.4. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.5. Declaration of Conflicting Interest

The authors declare that they have no conflicts of interest, known competing financial interests, or personal relationships that could have influenced the work reported in this paper.

REFERENCES

- [1] M. Takyaci, N. C. GENEK, T. Bertuğ, S. E. ALKAN, G. Arda, and M. Damar, "Generative artificial intelligence applications in digital media: The role and transformative impact of large language models," *THE AGE OF GENERATIVE ARTIFICIAL INTELLIGENCE* (2025), vol. 46, 2025.
- [2] C. H. Yu, N. Miljkovic, and Z. Wang, "Managing the unmanageable: Multimodal artificial intelligence for unstructured data management and analysis," *Digital*, vol. 6, no. 3, p. 68, 2026.
- [3] G. Zare, "Emosense-rec: Emotion-adaptive multi-modal recommendation via affective state modeling and cognitive signal fusion," *emotion*, vol. 37, p. 39, 2025.
- [4] Y. Wu, Q. Mi, and T. Gao, "A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions," *Biomimetics*, vol. 10, no. 7, p. 418, 2025.

- [5] Y. Zhang, Y. Tao, F. Ma, M. Liu, S. Zhao, Q. Tian, and H. Yan, "A survey on emotion-controllable face generation," 2025.
- [6] J. Williams, A. G. Prawiyogi, M. Rodriguez, and I. Kovac, "Enhancing circular economy with digital technologies: A pls-sem approach," *International Transactions on Education Technology (ITEE)*, vol. 2, no. 2, pp. 140–151, 2024.
- [7] A. Nastoska, B. Jancheska, M. Rizinski, and D. Trajanov, "Evaluating trustworthiness in ai: Risks, metrics, and applications across industries," *Electronics*, vol. 14, no. 13, p. 2717, 2025.
- [8] J. Young, S. Vaughan, A. Jenks, H. Malvar, C. Paquin, P. England, T. Roca, J. L. Ferres, F. Poursabzi, N. Coles *et al.*, "Media integrity and authentication: Status, directions, and futures," *arXiv preprint arXiv:2602.18681*, 2026.
- [9] D. Greco, P. Barra, L. D'Errico, and M. Staffa, "Multimodal interfaces for emotion recognition: models, challenges and opportunities," in *International Conference on Human-Computer Interaction*. Springer, 2024, pp. 152–162.
- [10] S. Li and H. Tang, "Multimodal alignment and fusion: A survey: S. li, h. tang," *International Journal of Computer Vision*, vol. 134, no. 3, p. 103, 2026.
- [11] S. N. Husin, P. Edastama, and A. Tambunan, "Digital marketing strategy using white hat seo techniques," *International Journal of Cyber and IT Service Management (IJCITSM)*, vol. 2, no. 2, pp. 171–179, 2022.
- [12] M. F. Dzulqarnain and N. Mirela, "Digital vernaculars in play: A machine learning-based linguistic analysis of slang and code-switching in indonesian youth online gaming chats," *Journal of Digital Society*, vol. 1, no. 3, pp. 201–215, 2025.
- [13] B. Mallia, "Analysis and" empathy algorithms," *Emotions in Code-The AI Frontier of Sentiment Analysis: The AI Frontier of Sentiment Analysis*, p. 37, 2025.
- [14] K. A. Lindquist, J. C. Jackson, J. Leshin, A. B. Satpute, and M. Gendron, "The cultural evolution of emotion," *Nature Reviews Psychology*, vol. 1, no. 11, pp. 669–681, 2022.
- [15] A. O. Kolhatin, "From automation to augmentation: a human-centered framework for generative ai in adaptive educational content creation." in *AREdu*, 2025, pp. 143–195.
- [16] R. Tarmizi, N. Septiani, P. A. Sunarya, and Y. P. A. Sanjaya, "Harnessing digital platforms for entrepreneurial success: A study of technopreneurship trends and practices," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 3, pp. 278–290, 2023.
- [17] N. Fabiano, "Affective computing and emotional data: Challenges and implications in privacy regulations, the ai act, and ethics in large language models," *arXiv preprint arXiv:2509.20153*, 2025.
- [18] F. Ma, Y. Xie, Y. Li, Y. He, Y. Zhang, H. Ren, Z. Liu, W. Yao, F. Ren, F. R. Yu *et al.*, "A review of human emotion synthesis based on generative technology," *IEEE transactions on affective computing*, vol. 16, no. 4, pp. 2579–2598, 2025.
- [19] V. N. Antony and C.-M. Huang, "Id. 8: Co-creating visual stories with generative ai," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, no. 3, pp. 1–29, 2025.
- [20] H. Shi, L. Qu, Y. Liu, D. Gao, L. Gong, Y. Zheng, and T. Li, "Think-before-draw: Decomposing emotion semantics for fine-grained controllable generation of expressive talking heads," *Pattern Recognition*, p. 113593, 2026.
- [21] M. Cossio, "A comprehensive taxonomy of hallucinations in large language models," *arXiv preprint arXiv:2508.01781*, 2025.
- [22] A. Felix and G. D. Rembulan, "Analysis of key factors for improved customer experience, engagement, and loyalty in the e-commerce industry in indonesia," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 5, no. 2sp, pp. 196–208, 2023.
- [23] D. Abbas, K. Siahaan, and M. Yusup, "Design thinking as a business model for empowering creative entrepreneurs in the digital era," *Startupreneur Business Digital (SABDA Journal)*, vol. 4, no. 2, pp. 124–133, 2025.
- [24] Y. Yu, H. Chen, D. Romero, and P. S. Dhillon, "Signals in the noise: Decoding unexpected engagement patterns on twitter," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 7, pp. 1–33, 2025.
- [25] E. Ferrara, "The generative ai paradox: Genai and the erosion of trust, the corrosion of information verification, and the demise of truth," *Future Internet*, vol. 18, no. 2, p. 73, 2026.
- [26] A. Pirjan and D.-M. Petroşanu, "Artificial social intelligence and the transformation of human interaction by artificial intelligence agents," *Journal of Information Systems & Operations Management*, vol. 19,

- no. 1, pp. 259–305, 2025.
- [27] R. Gangavarapu, “Mastering ai governance,” in *Mastering AI governance*. Springer, 2025, pp. 111–9.
 - [28] T. S. Goh, S. Martinez, S. L. Sitorus, and T. L. Anita, “Adaptive strategic approaches and their impact on economic growth in dynamic markets,” *Startupreneur Business Digital (SABDA Journal)*, vol. 4, no. 2, pp. 174–183, 2025.
 - [29] W. H. Organization *et al.*, “Social listening in infodemic management for public health emergencies: guidance on ethical considerations,” 2025.
 - [30] L. Kandikarla and B. Radeljic, “Ai and human oversight: a risk-based framework for alignment,” *arXiv preprint arXiv:2510.09090*, 2025.
 - [31] Q. Zhu, “Understanding censorship in large language models: From mechanisms to governance,” *arXiv preprint arXiv:2606.30661*, 2026.
 - [32] C. Uddagiri and B. V. Isunuri, “Ethical and privacy challenges of generative ai,” in *Generative AI: Current trends and applications*. Springer, 2024, pp. 219–244.
 - [33] B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu *et al.*, “Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems,” *arXiv preprint arXiv:2504.01990*, 2025.
 - [34] J. C. Simmons and J. M. Winograd, “Interoperable provenance authentication of broadcast media using open standards-based metadata, watermarking and cryptography,” *arXiv preprint arXiv:2405.12336*, 2024.
 - [35] J. I. Lee, K. T. Dirks, and R. L. Campagna, “At the heart of trust: Understanding the integral relationship between emotion and trust,” *Group & Organization Management*, vol. 48, no. 2, pp. 546–580, 2023.
 - [36] L. Wang, “Emotion-aware multi-objective genetic optimization for multimedia interaction design,” in *2025 Third International Conference on Networks, Multimedia and Information Technology (NMITCON)*. IEEE, 2025, pp. 1–7.
 - [37] F. Krueger, R. Riedl, J. A. Bartz, K. S. Cook, D. Gefen, P. A. Hancock, S. L. Jarvenpaa, L. Krabbendam, M. R. Lee, R. C. Mayer *et al.*, “A call for transdisciplinary trust research in the artificial intelligence era,” *Humanities and Social Sciences Communications*, vol. 12, no. 1, p. 1124, 2025.
 - [38] D. Domingueti, D. Barbosa Feres Carvalho, D. R. Colombo Dias, and V. C. Oliveira, “Software-based simulation on a 3d environment for vaccination teaching and learning: Design science research,” *JMIR Medical Education*, vol. 8, no. 4, p. e35712, 2022.
 - [39] L. A. Sparrow, C. Galwey, B. Loveridge, S. Glasser, M. S. Osborne, and R. M. Kelly, “Heart and soul: The ethics of biometric capture in immersive artistic performance,” in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–23.
 - [40] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, “From local to global: A graph rag approach to query-focused summarization,” *arXiv preprint arXiv:2404.16130*, 2024.
 - [41] Q. Aini, D. Manongga, U. Rahardja, I. Sembiring, and Y.-M. Li, “Understanding behavioral intention to use of air quality monitoring solutions with emphasis on technology readiness,” *International Journal of Human-Computer Interaction*, vol. 41, no. 8, pp. 5079–5099, 2025.
 - [42] A. Ruangkanjanases, A. Khan, O. Sivarak, U. Rahardja, and S.-C. Chen, “Modeling the consumers’ flow experience in e-commerce: The integration of ecm and tam with the antecedents of flow experience,” *SAGE Open*, vol. 14, no. 2, p. 21582440241258595, 2024.
-