A Multimodal Generative Storytelling Framework for Sustainable Ecotourism Using Text and Image Fusion

Listra Frigia Missianes Horhoruw
Doctoral Program in Information Technology, Gunadarma University, Indonesia
listrajsc@gmail.com (corresponding author)

Lintang Yuniar Banowosari
Department of Informatics, Gunadarma University, Indonesia
lintang@staff.gunadarma.ac.id

Diana Ikasari
Faculty of Computer Science and Information Technology, Gunadarma University, Indonesia
d_ikasari@staff.gunadarma.ac.id

Received: 6 December 2025 | Revised: 13 February 2026, 28 February 2026, and 13 March 2026 | Accepted: 24 March 2026

 DOI: https://doi.org/10.48084/etasr.16742

ABSTRACT

This study presents a system-level multimodal generative storytelling approach designed to support sustainable ecotourism through narrative-oriented multimodal conditioning. The proposed architecture integrates Indonesian Text-to-Text Transfer Transformer (IndoT5) as the textual encoder–decoder and Bootstrapped Language-Image Pretraining (BLIP) as the visual encoder, employing an early fusion strategy to align semantic and visual representations. The model was trained on a curated dataset from Indonesian Super Priority Tourism Destinations (SPTDs) and optimized to generate coherent, natural, and emotionally expressive narratives. Performance evaluation was conducted using a combination of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Metric for Evaluation of Translation with Explicit Ordering (METEOR) metrics and semantic (BERTScore) similarity metrics. The results indicate that the proposed IndoT5-based approach achieves higher average performance, with ROUGE-L (0.66), METEOR (0.70), and BERTScore (0.89). In addition, expert-based qualitative evaluation demonstrated strong agreement within the defined narrative assessment scope, resulting in a Scale-Level Content Validity Index Averaged Across Items (S-CVI/Ave) of 1.00. The proposed approach effectively bridges visual perception and linguistic generation, offering a scalable solution for automated tourism storytelling that preserves contextual and emotional nuances.

Keywords-ecotourism storytelling; generative model; Indonesian language; text generation

I. INTRODUCTION

Tourism is a global economic driver, yet its rapid expansion poses significant threats to biodiversity and socio-cultural integrity, necessitating a transformative shift toward sustainable ecotourism [1]. However, effective governance in this domain is hindered by fragmented data ecosystems and a reliance on quantitative metrics such as visitor numbers and satisfaction scores, which fail to capture the rich emotional and narrative nuances of the tourist experience [2, 3]. Conventional frameworks often overlook the qualitative depth available in user-generated reviews and social stories, limiting their ability to induce sustainable behavioral change [4]. Consequently, integrating storytelling and emotion-aware analysis into tourism governance is crucial for enhancing destination identity and fostering deeper public engagement with sustainability principles [5, 6].

While advancements in Artificial Intelligence (AI) and multimodal learning have enabled the integration of text and visual data, existing research mainly focuses on passive tasks, such as sentiment analysis, image captioning, or emotion recognition, rather than active content creation [7, 8]. In [9], a combination of FastText and Bidirectional Long Short-Term Memory (Bi-LSTM) was used to enhance sentiment analysis on Indonesian tourism video commentary, demonstrating the relevance of deep learning methods for understanding user-generated tourism content. However, there is a significant gap in leveraging Large Language Models (LLMs) to go beyond mere classification and dynamically generate coherent, emotionally resonant narratives that synthesize multimodal

context [10, 11]. Several multimodal systems, such as the modified Contrastive Language-Image Pre-training (CLIP)-based architecture [12], demonstrate the potential of integrating image and text modalities for generating descriptive outputs. However, these approaches are domain-specific and do not address the narrative-level generation required for tourism storytelling. Addressing this limitation, where cultural and environmental issues are often overlooked in current generative models, the present study introduces a novel multimodal generative framework specialized for sustainable ecotourism. By synergizing visual and textual data from Indonesia's Super Priority Tourism Destinations (SPTDs), specifically Lake Toba [13] and Borobudur [14], this research proposes a generative model capable of producing context-aware storylines that not only describe attractions but also reinforce the values of ecological and cultural sustainability.

The contribution of this study lies in designing a system-level architecture that repositions multimodal storytelling as a reasoning mechanism for sustainable ecotourism. The novelty lies in how multimodal representations are structured, fused, and operationalized to support narrative coherence, contextual grounding, and decision-oriented interpretation, rather than isolated caption generation or descriptive tasks. Accordingly, this work presents a system-level engineering framework that integrates multimodal representation learning, generative language modeling, and agentic control into a unified storytelling framework.

II. MATERIALS AND METHODS

This study proposes a multimodal generative framework for adaptive storytelling to support sustainable ecotourism. The comprehensive workflow, as illustrated in Figure 1, consists of four stages: data acquisition, preprocessing, feature extraction and fusion, and model training. At the core of the system is a dual-encoder architecture that aligns linguistic and visual semantics to condition a generative LLM, ensuring that the generated narratives are both contextually accurate and emotionally resonant.

A. Data Collection

The dataset was collected from TripAdvisor in March 2024, focusing on two Indonesian SPTDs: Lake Toba and Borobudur Temple [13, 14]. Multimodal data were extracted using the Apify web-scraping platform, including review titles, textual content, user-uploaded images, and metadata (ratings and posting dates). A total of 5,735 text–image pairs were compiled. During preprocessing, duplicates, incomplete entries, and irrelevant content were removed. TripAdvisor is a significant source of user-generated tourism data. [15].

B. Text and Image Preprocessing

As depicted in Figure 1 and Table I, the raw data were standardized through two parallel streams (text pipeline and image pipeline) to improve model effectiveness.

C. Feature Extraction and Analysis

To capture high-quality multimodal representations, the framework employs specialized transformer-based encoders tailored to each modality. This stage ensures that raw inputs are transformed into dense, semantically rich vectors before fusion.

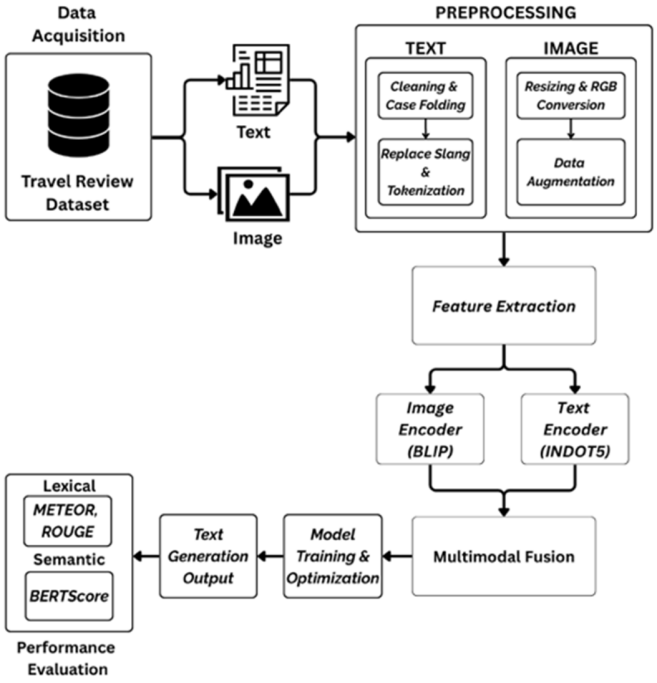

Fig. 1. Proposed multimodal generation pipeline.

TABLE I. TEXT AND IMAGE PREPROCESSING TASKS

Data modality	Preprocessing tasks
Text pipeline	Textual data underwent cleaning to remove noise and symbols, case folding for normalization, and slang replacement to handle informal linguistic patterns typical in user-generated content. Finally, tokenization was applied to prepare the sequences for the encoder.
Image pipeline	Visual data were resized to a standardized dimension of 150 × 112 pixels and converted to RGB to ensure channel consistency. No artificial data augmentation techniques were applied. The final dataset reflects the original collected samples after preprocessing.

1) Text Encoder

The encoder component of Indonesian Text-to-Text Transfer Transformer (IndoT5) was utilized to extract linguistic features from the preprocessed reviews [16]. IndoT5 was selected for its capability to understand local Indonesian contexts and produce a hidden state vector $h_i^{(t)}$ that captures the semantic depth of the textual input.

2) Image Encoder

For visual data, Bootstrapped Language-Image Pretraining (BLIP) was employed [17]. BLIP is designed to bridge the modality gap by generating visual embeddings $h_i^{(v)}$ that are semantically aligned with language understanding, making it highly effective for multimodal generation tasks.

To validate the quality and separability of these learned representations before fusion, the feature distributions were examined using t-Distributed Stochastic Neighbor Embedding (t-SNE). This visualization helps to assess whether the encoders have successfully mapped the inputs into a coherent latent space. As illustrated in Figure 2, the t-SNE projection provides a two-dimensional visualization of the high-dimensional IndoT5 textual embeddings (perplexity = 30, learning rate = 200, n_iter = 1000, random_state = 42). While the projection reveals globally compact distributions, substantial overlap between destination labels is observed. This indicates that the embeddings capture shared linguistic structures commonly present across tourism reviews rather than sharply separating geographic categories. The visualization is presented as structural insight and is complemented by quantitative validation metrics to avoid reliance solely on low-dimensional projection.

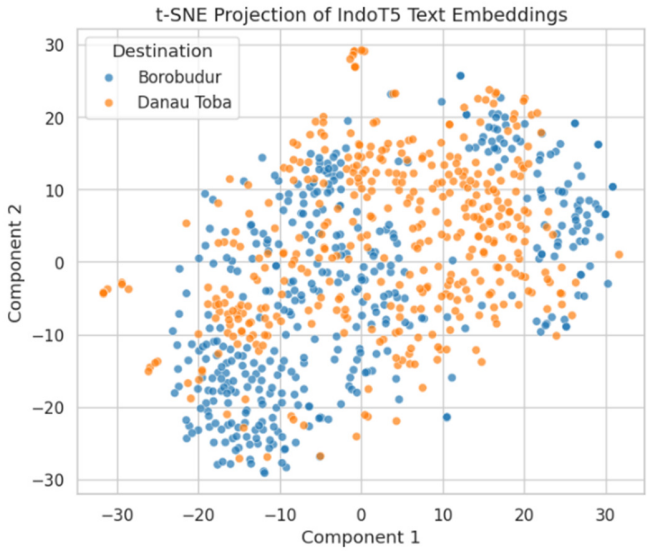

Fig. 2. t-SNE visualization of textual features extracted using IndoT5.

Similarly, Figure 3 presents the t-SNE projection of BLIP visual embeddings. Similar to the textual modality, the visual embeddings demonstrate global structural coherence with partial overlap between destination samples. This suggests that the encoder maintains modality-specific consistency while preserving shared visual characteristics common across tourism imagery.

The observed structural coherence across modalities suggests that the extracted representations maintain consistent distributional patterns. However, these projections are interpreted as descriptive structural illustrations and are complemented by quantitative clustering validation to ensure methodological rigor.

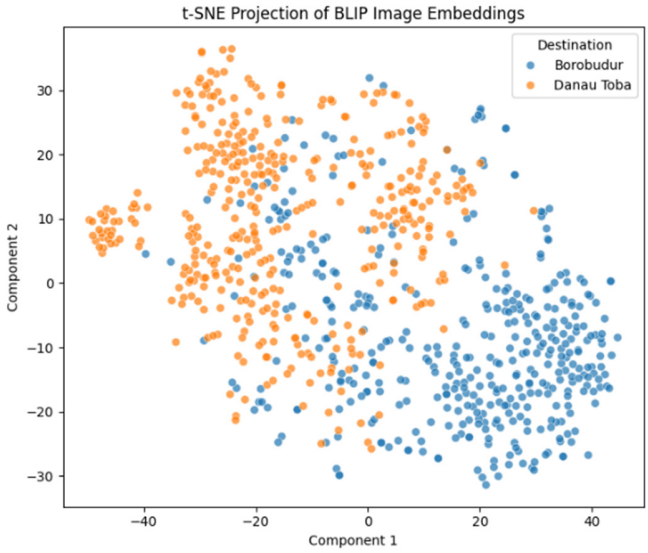

Fig. 3. t-SNE visualization of visual features extracted using BLIP.

D. Quantitative Embedding Space Validation

Quantitative clustering validation was conducted on a stratified balanced subset (n = 812; 406 samples per destination) to ensure fair comparison across modalities. This subset was used exclusively for embedding structure analysis and did not affect model training. The evaluation included silhouette score, Davies–Bouldin index, Calinski–Harabasz index (quantifying cluster variance through inter-/intra-cluster dispersion ratio), and cosine-based intra- and inter-cluster similarity analysis. Destination labels (Danau Toba and Borobudur) were used as reference clusters for both textual and visual embeddings. The quantitative embedding validation results are presented in Table II.

TABLE II. QUANTITATIVE EMBEDDING VALIDATION RESULTS

Metric	IndoT5 (text)	BLIP (image)
Silhouette score	0.0205	0.0585
Davies-Bouldin index	6.6648	4.0652
Calinski-Harabasz index	17.1544	47.9927
Intra-cluster cosine similarity	0.7536	0.6771
Inter-cluster cosine similarity	0.7442	0.6398

The textual embeddings exhibit limited destination-level separability, as reflected by a low silhouette score and relatively high Davies–Bouldin index. However, the high cosine similarity values across both intra- and inter-cluster comparisons indicate strong global semantic coherence across tourism narratives. This suggests that IndoT5 primarily captures shared linguistic structures rather than discriminative geographic distinctions. In contrast, the visual embeddings demonstrate relatively stronger cluster separation, as evidenced by higher silhouette and Calinski–Harabasz values and lower Davies–Bouldin index. The larger gap between intra- and inter-cluster cosine similarity further indicates clearer destination-level visual differentiation. These results stress the complementary characteristics of textual and visual representations: textual embeddings emphasize narrative

coherence, while visual embeddings preserve destination-specific structural variance. The proposed early fusion mechanism leverages this complementary structure to enhance multimodal conditioning in generative storytelling.

E. Multimodal Fusion and Model Development

IndoT5 is a pretrained encoder–decoder language model adapted from the Text-to-Text Transfer Transformer (T5) architecture for Indonesian language tasks. It is trained on large-scale Indonesian corpora and is designed to support diverse generative objectives through a unified text-to-text formulation, making it suitable as the generative backbone for narrative generation. BLIP is a vision–language model pretrained using image–text alignment objectives, enabling effective cross-modal representation learning for image captioning and visual–textual grounding tasks.

The core of the proposed framework lies in the effective integration of textual and visual representations to condition the generative process. As illustrated in Figure 1, this is achieved through an early fusion strategy, in which the linguistic features extracted by IndoT5 and the visual features extracted by BLIP are combined into a unified multimodal vector before decoding. In conventional multimodal learning, early fusion typically aims to enhance object recognition accuracy or descriptive completeness through low-level or token-level alignment between visual and textual features. In contrast, the early fusion strategy proposed in this study is narrative-oriented, designed to condition the generative process at the discourse level rather than at the lexical level. By aligning visual embeddings with semantic textual representations prior to decoding, the model enables visual context to influence narrative structure, emotional framing, and thematic continuity, which are critical for ecotourism storytelling.

1) Multimodal Fusion Strategy

To formalize the fusion process, the textual embeddings $h_i^{(t)}$ and visual embeddings $h_i^{(v)}$ are first projected into a shared latent space to align their dimensions. The fusion is then computed using a weighted concatenation mechanism, expressed as:

$$h_i^{(f)} = \alpha\left(W_t\, h_i^{(t)} + b_t\right) + (1-\alpha)\left(W_v h_i^{(v)} + b_v\right) \quad (1)$$

where $h_i^{(f)}$ is the resulting fused multimodal representation, W_t and W_v are the learnable projection matrices for text and image modalities, respectively, and b_t and b_v are the bias terms. $\alpha \in [0,1]$ is a gating parameter that controls the contribution weight of each modality, allowing the model to dynamically balance between visual cues and linguistic context.

2) Model Training and Optimization

The fused representation $h_i^{(f)}$ serves as the input context for the IndoT5 decoder. Unlike standard fine-tuning, the model is trained to generate the target sequence $Y = \{y_1, y_2, \ldots\ldots y_r\}$ by maximizing the conditional probability of the next token y_t. The probability distribution for the token y_t at time step t is calculated using a softmax function over the vocabulary, defined as:

$$P\left(y_t \middle| y_{<t}, h_i^{(f)}\right) = softmax\left(W_{vocab}\, .D\,(y_{<t}, h_i^{(f)})\right) \quad (2)$$

where $y_{<t}$ represents the sequence of previously generated tokens (history), $D(\cdot)$ denotes the decoder function of IndoT5, W_{vocab} is the weight matrix mapping the decoder output to the vocabulary size, and y_t is the predicted token at the current step.

The training objective is to minimize the negative log-likelihood loss (Cross-Entropy Loss) across the entire dataset:

$$\mathcal{L} = -\sum_{t=1}^{T} log\ P\left(y_t \middle| y_{<t}, h_i^{(f)}\right) \quad (3)$$

To ensure optimal convergence and generalization, the model was fine-tuned using the AdamW optimizer with the hyperparameters detailed in Table III.

TABLE III. TRAINING PARAMETERS

Parameter	Value
Learning rate	3×10^{-5}
Batch size	4
Epoch	10
Optimizer	AdamW

3) Inference and Refinement Strategy

Following the training phase, inference was executed using a controlled decoding strategy to balance fluency and creativity. Beam search with num_beams = 4 was employed to explore multiple probability paths simultaneously, reducing the risk of repetitive outputs. The specific parameters applied during inference are summarized in Table IV.

TABLE IV. INFERENCE PARAMETERS

Parameter	Value
Max length	200
Temperature	0.9
Top-p	0.95
Repetition penalty	1.1
Num beams	4

Subsequently, a Context-Aware Refinement Agent was deployed using the Meta-LLaMA-3.1-70B-Instruct model (via the OpenRouter API) [18, 19] to refine the generated draft to a professional storytelling standard. This process is managed by a custom-built Python orchestrator that goes beyond simple rewriting. Before refinement, the orchestrator computes an Emotion-Adaptive Representation (EAR) score using a lexicon-based density analysis and identifies dominant narrative themes utilizing K-Means clustering on Term Frequency–Inverse Document Frequency (TF-IDF) features. These semantic attributes are dynamically injected into the agent's system prompt, ensuring that the Large Language Model Meta AI (LlaMA) model refines the linguistic fluency and style while strictly preserving the specific emotional tone and thematic consistency of the original ecotourism experience. This refinement stage is not intended as a generic post-hoc editing step, but as an agentic control mechanism that governs narrative coherence and emotional fidelity while preserving

multimodal grounding. The refinement agent operates under explicit semantic and emotional constraints, ensuring that stylistic improvements do not distort the original multimodal representations.

F. Evaluation Metrics and Analysis

The evaluation was conducted using a dual-approach methodology combining quantitative metrics and expert-based qualitative validation. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and the Metric for Evaluation of Translation with Explicit Ordering (METEOR) were applied to measure lexical overlap and structural similarity, while BERTScore assessed semantic alignment beyond surface-level matching. Furthermore, expert evaluation was conducted using the Content Validity Index (CVI), in which Indonesian language and literature specialists assessed the produced narratives across five dimensions: relevance, coherence, fluency, naturalness, and emotional accuracy.

III. RESULTS

A. Quantitative Evaluation

Training and validation losses were analyzed to assess convergence stability and generalization. As shown in Figure 4, IndoBART exhibits fluctuations and a sharp decline in final-epoch training loss, indicating potential instability. In contrast, IndoT5 demonstrates smoother and more consistent convergence under multimodal conditioning. This behavior aligns with prior findings on T5-style encoder–decoder architectures, which report strong robustness in multilingual and low-resource generative tasks due to their unified text-to-text formulation and transfer learning framework [20]. The Indonesian adaptation IndoT5 further benefits from multilingual pretraining strategies [21], while related multilingual T5 variants have demonstrated stable cross-lingual and domain-specific narrative generation performance [22, 23].

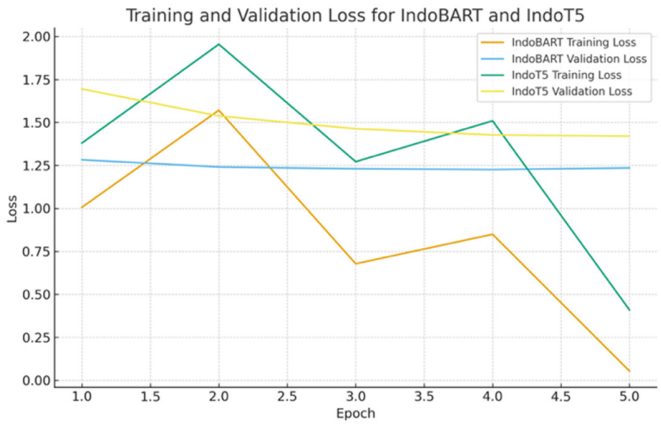

Fig. 4. Training and validation loss curves for IndoBART and IndoT5.

Following the convergence analysis, model performance was evaluated using ROUGE-L, METEOR, and BERTScore. As summarized in Table V, IndoT5 consistently outperforms IndoBART across all evaluation metrics, with significant gains in METEOR and BERTScore, indicating improved semantic alignment and contextual fidelity. This performance difference underscores the importance of the textual backbone in multimodal generative storytelling and aligns with prior findings on the robustness of T5-style encoder–decoder architectures in multilingual and low-resource generation settings.

TABLE V. QUANTITATIVE EVALUATION OF RAW GENERATED OUTPUTS

Evaluation metric	Model	
	IndoBART	**IndoT5 (proposed)**
ROUGE1	0.56	0.70
ROUGE2	0.50	0.64
ROUGE-L	0.53	0.66
ROUGE-LSum	0.53	0.66
METEOR	0.57	0.70
BERTScore	0.61	0.89

To further examine the robustness of these quantitative differences, a paired t-test was conducted on per-sample ROUGE-L, METEOR, and BERTScore values using a significance level of $\alpha = 0.05$. Inferential statistical analysis was performed on a randomly sampled subset ($n = 800$, random_state = 42) to ensure computational feasibility and reproducibility. The results, presented in Table VI, indicate statistically significant differences across all evaluation metrics. For ROUGE-L, IndoT5 ($M = 0.6504$, $SD = 0.3286$) significantly outperformed IndoBART ($M = 0.4202$, $SD = 0.2423$), $t(799) = 27.68$, $p < 0.001$, indicating a large effect size (Cohen's $d = 0.98$). Similarly, significant improvements were observed for METEOR, $t(799) = 16.51$, $p < 0.001$, $d = 0.58$, and BERTScore, $t(799) = 29.54$, $p < 0.001$, $d = 1.04$. These findings demonstrate that IndoT5 provides statistically and practically meaningful improvements over IndoBART in multimodal generative storytelling.

TABLE VI. STATISTICAL COMPARISON

Metric	IndoT5 mean (SD)	IndoBART mean (SD)	Mean diff	*t*	*p*	*d*
ROUGE-L	0.6504 (0.3286)	0.4202 (0.2423)	0.2301	27.68	<0.001	0.98
METEOR	0.6925 (0.3278)	0.5536 (0.2986)	0.1389	16.51	<0.001	0.58
BERTScore	0.8615 (0.1165)	0.7648 (0.0758)	0.0966	29.54	<0.001	1.04

The 95% confidence intervals for the mean differences were 0.2138-0.2465 for ROUGE-L, 0.1224-0.1554 for METEOR, and 0.0902-0.1031 for BERTScore. In all cases, the intervals exclude zero, confirming the robustness and practical significance of these improvements.

B. Multimodal Fusion and Context-Aware Refinement

The effectiveness of the proposed system stems from the early fusion strategy and the subsequent agentic refinement process. The early fusion effectively combines textual and visual features into a single multimodal representation, thereby improving contextual grounding, strengthening descriptive accuracy, and enhancing the relevance of generated storytelling to the depicted scenes. The performance is finalized by the context-aware refinement agent. After the IndoT5 model

generates an initial draft, the Meta-LLaMA-3.1-70B-Instruct model (via the OpenRouter API) is used as a stylistic editor.

This refinement process is managed by a custom-built Python orchestrator that dynamically computes the EAR score and identifies dominant K-Means narrative themes from the raw text, injecting these semantic attributes into the LLaMA agent's prompt. This ensures that the LLaMA model refines the linguistic fluency and flow while strictly preserving the specific emotional tone and thematic consistency of the original ecotourism experience.

C. Expert-Based Content Validity Evaluation

The qualitative evaluation was conducted by two independent experts with academic backgrounds in Indonesian language and literature. As the primary output of the proposed system is narrative storytelling, these experts were selected for their expertise in assessing narrative coherence, linguistic fluency, stylistic naturalness, and emotional expressiveness. The evaluation was performed independently using a five-point Likert scale, which was subsequently mapped to a CVI framework. The CVI framework comprises Item-level CVI (I-CVI), Scale-level CVI Averaged Across Items (S-CVI/Ave), and S-CVI based on Universal Agreement (S-CVI/UA). While S-CVI/Ave reflects the average relevance agreement across evaluation criteria, S-CVI/UA represents the proportion of items that achieved complete agreement among all evaluators.

The final refined narratives were then evaluated across five criteria: content relevance, coherence, linguistic fluency, naturalness, and emotional accuracy. Inter-rater agreement was assessed using Cohen's Kappa, indicating strong agreement between the evaluators. Table VII summarizes the CVI results for the proposed two-stage generative framework: the initial (1) multimodal grounding by IndoT5, followed by the final (2) context-aware refinement by the LLaMA agent. The refined outputs achieved an S-CVI/Ave of 1.00, reflecting complete agreement between the evaluators within the defined narrative evaluation scope. This result should be interpreted with caution, as it is influenced by the focused evaluation criteria and the limited number of experts. These findings indicate that the agentic refinement step enhances the perceived quality and readability of the storytelling outputs, supporting their suitability for professional tourism content.

TABLE VII. SUMMARY OF CVI RESULTS

INDEX	VALUE
Number of experts	2
Rating scale	Likert 1–5 (converted into relevant / not relevant)
I-CVI (per item)	1.00
S-CVI/Ave	1.00
S-CVI/UA	1.00

IV. DISCUSSION

The findings of this study contribute to the existing research on multimodal generative models by demonstrating that a narrative-oriented early fusion strategy combined with a T5-style encoder–decoder architecture enhances semantic alignment and contextual coherence in tourism storytelling tasks. The improvements observed across ROUGE-L, METEOR, and BERTScore indicate that IndoT5 generates narratives with stronger contextual representation when conditioned on multimodal inputs. These improvements should therefore be interpreted not merely as metric differences but as meaningful gains in narrative coherence and semantic alignment in multimodal storytelling.

These findings are consistent with [20–23], demonstrating the robustness of T5-style architectures and their multilingual variants in generative language modeling tasks. The unified text-to-text formulation enables stable contextual representation and flexible adaptation across diverse language generation tasks. In this study, the superior performance of IndoT5 compared to IndoBART suggests that task-agnostic pretraining better supports narrative-level generation in multimodal storytelling. Furthermore, the proposed narrative-oriented early fusion mechanism strengthens contextual grounding by integrating visual semantics at the representation level rather than at the token level, thereby improving discourse coherence beyond simple descriptive alignment.

Despite these contributions, this study has several limitations. First, a statistical evaluation was conducted on a reproducible sampled subset to ensure computational feasibility. Second, automatic similarity-based metrics may not fully capture creative expressiveness or stylistic richness in narrative storytelling. Third, expert-based validation involved a limited number of evaluators, which may constrain generalizability. Future research should explore larger-scale human-centered evaluation frameworks, cross-lingual extensions for international tourism contexts, and the integration of Explainable Artificial Intelligence (XAI) techniques to improve transparency in multimodal generative systems. Investigating adaptive fusion strategies and reinforcement learning-based narrative optimization may further enhance contextual sensitivity and sustainability-oriented storytelling performance.

V. CONCLUSION

This study presents a system-level multimodal generative storytelling framework that integrates early narrative-oriented fusion and agentic refinement to support sustainable ecotourism decision-making. By employing a caption-based bridging strategy utilizing Bootstrapped Language-Image Pretraining (BLIP) for visual interpretation and Indonesian Text-to-Text Transfer Transformer (IndoT5) as the text-to-text generative core, the proposed architecture synthesizes visual and linguistic data to produce context-aware narratives. The framework demonstrates robust quantitative performance, achieving a BERTScore of 0.89, substantially outperforming the IndoBART baseline (0.61) while indicating enhanced semantic alignment and contextual grounding. Furthermore, the two-stage inference process, featuring a context-aware refinement agent dynamically guided by Emotion-Adaptive Representation (EAR) and thematic context, resulted in narratives that achieved full agreement within the defined expert-based evaluation scope. The Scale-Level Content Validity Index Averaged Across Items (S-CVI/Ave) was 1.00, although the number of evaluators was limited. These findings suggest that the proposed framework offers a scalable and effective approach for automated storytelling generation that preserves contextual and emotional nuances while meeting

professional content standards. Future work will focus on integrating Explainable Artificial Intelligence (XAI) techniques, particularly Local Interpretable Model-agnostic Explanations (LIME), to enhance transparency and interpretability of the generative process. Additionally, future studies will explore cross-lingual storytelling extensions for broader destination promotion contexts.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests.

DATA AVAILABILITY

The data used in this study were collected from [13, 14]. Due to platform terms of service and ethical considerations, the raw textual reviews and images cannot be shared publicly. The final multimodal text–image pairs used in the experiments can be made available upon reasonable request from the corresponding author.

ACKNOWLEDGMENT

The authors would like to acknowledge the support and facilities provided by the Indonesian Education Scholarship, Center for Higher Education Funding and Assessment, and the Indonesian Endowment Fund for Education.

REFERENCES

[1] Q. B. Baloch *et al.*, "Impact of Tourism Development Upon Environmental Sustainability: A Suggested Framework for Sustainable Ecotourism," *Environmental Science and Pollution Research*, vol. 30, no. 3, pp. 5917–5930, Jan. 2023, https://doi.org/10.1007/s11356-022-22496-w.

[2] A. Stoffelen, "Disentangling the Tourism Sector's Fragmentation: A Hands-on Coding/Post-Coding Guide for Interview and Policy Document Analysis in Tourism," *Current Issues in Tourism*, vol. 22, no. 18, pp. 2197–2210, Nov. 2019, https://doi.org/10.1080/13683500.2018.1441268.

[3] W. Zhang and D. R. Fesenmaier, "Assessing Emotions in Online Stories: Comparing Self-Report and Text-based Approaches," *Information Technology & Tourism*, vol. 20, no. 1–4, pp. 83–95, Dec. 2018, https://doi.org/10.1007/s40558-018-0122-y.

[4] S. Sujatmiko, D. P. Ar, A. Hamdat, and K. N. Salam, "User-Generated Content (UGC) and Its Impact on Tourism Marketing: A Systematic Literature Review," *Golden Ratio of Mapping Idea and Literature Format*, vol. 5, no. 2, pp. 97–105, Jun. 2025, https://doi.org/10.52970/grmilf.v5i2.1491.

[5] H. Li, S. Zeng, and K. Tay, "Tourism Storytelling Research Progress and Trends: A Systematic Literature Review on SDGs," *Journal of Lifestyle and SDGs Review*, vol. 5, no. 1, Nov. 2024, Art. no. e02231, https://doi.org/10.47172/2965-730X.SDGsReview.v5.n01.pe02231.

[6] O. D. Rico Garcia, J. Fernandez Fernandez, R. A. Becerra Saldana, and O. Witkowski, "Emotion-Driven Interactive Storytelling: Let Me Tell You How to Feel," in *Artificial Intelligence in Music, Sound, Art and Design*, vol. 13221, T. Martins, N. Rodríguez-Fernández, and S. M. Rebelo, Eds. Cham, Switzerland: Springer International Publishing, 2022, pp. 259–274.

[7] W. Villalobos, Y. Kumar, and J. J. Li, "The Multilingual Eyes Multimodal Traveler's App," in *Proceedings of Ninth International Congress on Information and Communication Technology*, vol. 1004, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds. Singapore: Springer Nature Singapore, 2024, pp. 565–575.

[8] S. Sharma and N. Pandey, "Enhancing Sustainable Tourism with AI-Powered Cloud-Based Predictive Models for Intelligent Travel Destinations," in *2nd International Conference on Multidisciplinary Research and Innovations in Engineering*, Gurugram, India, Jul. 2025, pp. 529–534, https://doi.org/10.1109/MRIE66930.2025.11156218.

[9] D. Ariyus, D. Manongga, and I. Sembiring, "Enhancing Sentiment Analysis of Indonesian Tourism Video Content Commentary on TikTok: A FastText and Bi-LSTM Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18020–18028, Dec. 2024, https://doi.org/10.48084/etasr.8859.

[10] R. Zall, A. Kheyrkhah, E. Cambria, Z. Naseri, and M. R. Kangavari, "Intelligent Agents with Emotional Intelligence: Current Trends, Challenges, and Future Prospects." arXiv, 2025, https://doi.org/10.48550/ARXIV.2511.20657.

[11] X. Lin and X. Chen, "Improving Visual Storytelling with Multimodal Large Language Models." arXiv, 2024, https://doi.org/10.48550/ARXIV.2407.02586.

[12] D. Menaga and M. Sudha, "Leveraging a Modified Contrastive Language-Image Pre-training Model to Align Images and Text for Generating Remedy Text for Malus Pumila Lamina Images," *Engineering, Technology & Applied Science Research*, vol. 15, no. 2, pp. 21989–21997, Apr. 2025, https://doi.org/10.48084/etasr.9959.

[13] "Lake Toba - User Reviews and Photos," *Tripadvisor*, 2025. https://www.tripadvisor.in/Attraction_Review-g2301775-d338410-Reviews-Lake_Toba-North_Sumatra_Sumatra.html.

[14] "Borobudur Temple - Tripadvisor," *Tripadvisor*, 2025. https://www.tripadvisor.com/Attraction_Review-g790291-d320054-Reviews-Borobudur_Temple-Borobudur_Magelang_Central_Java_Java.html.

[15] R. P. Kusumawardani, R. A. Rahman, R. P. Wibowo, and A. Tjahjanto, "Understanding Fine-Grained Sentiments of Super-Priority Destination Visitors Using Multi-task Learning for Extraction of Aspect Terms and Polarity Classification on Reviews," *Procedia Computer Science*, vol. 234, pp. 602–613, 2024, https://doi.org/10.1016/j.procs.2024.03.045.

[16] Allenai, "Indonesian T5 Base." Hugging Face, Jan. 2024, [Online]. Available: https://huggingface.co/Wikidepia/IndoT5-base.

[17] E. A. Abed and T. Aguili, "Automated Medical Image Captioning Using the BLIP Model: Enhancing Diagnostic Support with AI-Driven Language Generation," *Diyala Journal of Engineering Sciences*, pp. 228–248, Jun. 2025, https://doi.org/10.24237/djes.2025.18215.

[18] Meta, "Meta-llama/Llama-3.1-70B-Instruct." Hugging Face, Dec. 2024, [Online]. Available: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

[19] "OpenRouter API: the Unified Interface for LLMs," *OpenRouter*, 2025. https://openrouter.ai/.

[20] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, Jun. 2020.

[21] M. Fuadi, A. D. Wibawa, and S. Sumpeno, "idT5: Indonesian Version of Multilingual T5 Transformer." arXiv, 2023, https://doi.org/10.48550/ARXIV.2302.00856.

[22] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498, https://doi.org/10.18653/v1/2021.naacl-main.41.

[23] M. Kale and A. Rastogi, "Text-to-Text Pre-Training for Data-to-Text Tasks," in *Proceedings of the 13th International Conference on Natural Language Generation*, 2020, pp. 97–102, https://doi.org/10.18653/v1/2020.inlg-1.14.