JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Application of Comics Generation Techniques in the Visualization of Vietnamese Folk Tales

Minh Tan Le*, Dinh Tan Loc Nguyen
Ho Chi Minh City University of Technology and Engineering, Vietnam

*Corresponding author. Email: tanlm@hcmute.edu.vn

ARTICLE INFO

Received: 26/03/2026
Revised: 02/06/2026
Accepted: 04/08/2026
Online First: 25/08/2026
Published:

KEYWORDS

Story-Iter;
Diffusion;
Generative AI;
Comics generation;
Folk comics.

ABSTRACT

In recent years, the integration of Artificial Intelligence into tasks requiring creative cognition has become increasingly prevalent. Although the nature of machine creativity remains a subject of debate, it is undeniable that these models have exerted a positive influence, ranging from the reduced lead times and significant cost savings for enterprises to versatile support in day-to-day activities. This study focuses on the generation of Vietnamese folk comics utilizing the Story-Iter architecture, which possesses an inherent advantage in maintaining contextual consistency via historical image references. Furthermore, this study proposes an enhancement involving the selective curation of reference image sets rather than incorporating the entire dataset or a fixed sequence of recent frames, thereby mitigating the integration of noise during the generation of novel imagery. This strategy effectively minimizes the inclusion of noisy data during the synthesis of new images. Experimental results performed on our self-built small-sized story dataset demonstrate that the approach improves upon the original Story-Iter framework, while simultaneously highlighting persistent challenges.

DOI: https://doi.org/10.54644/jte.2026.2148

1. Introduction

The rapid expansion of Artificial Intelligence (AI) has significantly transformed various aspects of daily life. As AI applications become broadly accessible to the public via mobile and computing devices, their presence in specialized creative fields has been pervasive. Key events in 2023, including the release of the SUNO music generation platform [1] and the integration of the DALL-E 3 image synthesis model into ChatGPT [2], highlight this trend. The proliferation of research problems arising from this period serves as a clear testament to the substantial progress made in generative AI.

Generative image synthesis for comic illustration has become a focal point of recent research, beginning with the advent of Generative Adversarial Networks (GANs) in 2014 [3]. A notable contribution to the field is StoryGAN (2019) [4], which employs recurrent neural networks to ensure contextual continuity. Research efforts have since sought to refine semantic correspondence and sequential stability in generated narratives [5], [6]. Nevertheless, GAN-derived models often produce distorted imagery that lacks the required visual fidelity. Recent advancements in diffusion-based models have largely overcome these spatial inconsistencies, establishing a new benchmark for image quality in this domain.

The limitations of GANs in maintaining image integrity and long-term context have shifted the research focus toward diffusion techniques for comic generation. By iteratively denoising latent representations, diffusion models achieve superior detail retention and stylistic continuity across frames. While these methods improve text-reference integration and consistency, long-form narratives with multiple interacting entities often suffer from error propagation. This has prompted the emergence of multiple advanced architectures, including StoryDiffusion (2024) [7], DreamStory (2025) [8], and Story-Iter (2026) [9] which leverage diffusion processes to enhance the overall quality of generated image sequences.

Existing research has largely overlooked Asian and Vietnamese folk literature in its experimental evaluations. Prominent benchmarks, including the Pororo-SV dataset employed by StoryGAN [4] and VLC-StoryGAN [6], and the StorySalon dataset used in the Intelligent Grimm research [10], lack representation from these regions. Despite the substantial volume of the StorySalon corpus (11,280 narratives), its composition is biased toward European and Grimm's folk traditions. Acknowledging that cultural divergence presents a substantial challenge for generative models, the present study constructed the VNFairyTales dataset. This modest custom dataset was engineered to align with the specific requirements of the task and was compiled from web-based digital archives.

This research introduces an approach to dynamic contextual image selection as an optimization for Story-Iter. Experimental results demonstrate that while a fixed-window approach (restricting inputs to the most recent frames) is suboptimal compared to global inclusion, a selective sampling mechanism yields superior performance across all quantitative benchmarks. Additionally, this study proposes domain-specific negative prompts tailored to the nuances of traditional narratives, addressing the limitations of existing prompt engineering techniques that have been predominantly validated on Western or mainstream pop-culture datasets.

2. Related work

2.1. Story-Iter architecture

Story-Iter, introduced by Mao et al., addresses the challenge of maintaining contextual coherence in long-sequence narrative synthesis [9]. Utilizing the Stable Diffusion backbone, the framework optimizes computational efficiency and accelerates convergence. The researchers highlighted the prevalence of error propagation, characterized by a loss of contextual alignment in subsequent frames, and the complexity of modeling inter-entity interactions. The system is designed for the dynamic integration of reference frames to guide the generative process. To manage the computational overhead associated with large reference sets, a Global Reference Cross-Attention (GRCA) mechanism was implemented [9]. This module generates a global embedding matrix that serves as a conditioning signal for the synthesis of each individual frame.

Story-Iter implements a dynamic referencing scheme where the comprehensive historical sequence (or a part of it) informs the generation process. For the initial frame, the model is conditioned solely on input prompts. In subsequent iterations, the architecture performs a denoising trajectory, mapping a latent noise distribution to a coherent output image according to the flow below:

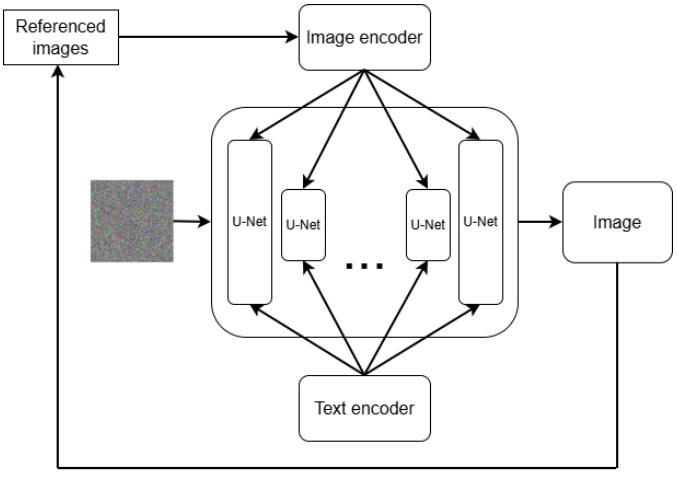

Figure 1. *Story-Iter architecture overview.*

In Figure 1, attention operations are performed across the U-Net framework. Reference image embeddings are derived using a frozen CLIP encoder. The combination of GRCA and prompt attention yields the following expression [9]:

$$I_k^i = \text{Attention}\left(I_k^i, T_k, T_k\right) + \lambda \text{GRCA}\left(I_k^i, x_{1,\ldots,B}^{i-1}\right) \quad (1)$$

Where λ functions as a weighting factor to calibrate the contribution of GRCA against textual attention. Within the U-Net architecture, the model alternates between GRCA and standard self-attention modules.

GRCA serves as an adapter within this architecture, processing the transformation of inputs for pre-trained models such as Stable Diffusion. However, this configuration inherits the baseline model's limitations, including difficulties in accurately rendering complex inter-object interactions [9]. Utilizing a comprehensive set or a fixed subset of previous frames $x_{1,\ldots B}^{i-1}$ implies that noisy data, which may lack direct relevance to the current context, can degrade output quality. Consequently, instances of character facial identity confusion occur frequently.

Story-Iter was preferred over solutions such as StoryDiffusion and DreamStory. While addressing long-form storytelling, they lack the flexible context usage of other architectures. StoryDiffusion is constrained by a fixed "ID length" for initial context, while DreamStory utilizes "multimodal anchors" [8] referencing self-generated portraits. DreamStory also requires a Large Language Model for descriptive metadata, leading to significantly higher computational demands than earlier approaches. While Story-Iter successfully utilizes GRCA, a critical research gap remains: the model blindly ingests sequential historical frames. This rigid inclusion inevitably incorporates irrelevant noisy data (e.g., past characters no longer in the current scene), leading to severe identity confusion in long narratives. This work directly addresses the gap by proposing a novel contextual filter mechanism that ensures the model only accounts for historical data to improve output consistency.

2.2. SigLIP

Proposed by Xiaohua Zhai et al., Sigmoid Loss for Language–Image Pretraining (SigLIP) [11] is a multimodal representation learning model designed to enhance the training methodologies of image-text alignment frameworks such as CLIP [12]. The primary objective of this model is to learn a joint embedding space for visual and textual data, ensuring that semantically related pairs exhibit high degrees of similarity within the shared latent space. Similarity is determined as:

$$P(y = 1 \mid x, t) = \sigma(\langle f_I(x), f_T(t) \rangle) \quad (2)$$

Defined by image x, text t, embedding functions $f_I(x)$, $f_T(t)$, and the sigmoid operator σ, SigLIP optimizes image-text pairs via binary loss, contrasting with CLIP's batch-wide softmax approach [12]. This study adopts SigLIP to quantify the similarity between prompts and historical frames, enabling the selective acquisition of reference data for the next generation step. Figure 2 illustrates the architecture consisting of two encoders that handle image or text inputs, while the split losses represent the separated training process in which the gradients are backpropagated independently.

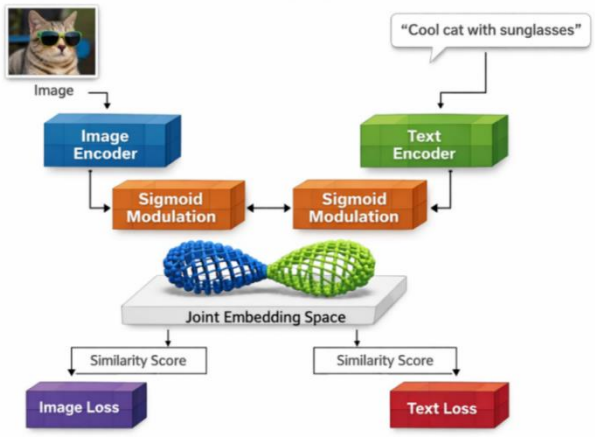

Figure 2. *SigLIP architecture overview.*

3. Data and Methodology

3.1. Dataset

The study establishes a dataset based on 23 Vietnamese folktales via a five-stage workflow: (1) multi-source online collection, (2) scene-based narrative segmentation, (3) English translation and cultural interpretation, (4) LLM-assisted (OpenAI's ChatGPT-4o mini, Google's Gemini-3 thinking, Claude's Sonnet 4.6) visual prompt engineering, and (5) data cleaning and standardization. It was named VNFairyTales for convenience.

VNFairyTales comprises 447 complete story segments. Depending on the complexity of each story, the narratives were partitioned into approximately 11 to 34 scenes. To optimize compatibility with the model, the prompts average 32 words in length, providing sufficient descriptive detail for the architecture to interpret the context. A significant characteristic of this dataset is the inclusion of distinct cultural elements, such as the "áo tứ thân" (four-part dress) and thatched houses, which present a substantial challenge for models predominantly trained on synthetic or generalized data. Over 92% of the prompts incorporate keywords oriented toward traditional artistic styles, ensuring the synthesis of illustrative artwork rather than irrelevant realistic imagery. Furthermore, the dataset exhibits high story diversity, improving the experimental objectivity.

While VNFairyTales is currently smaller than its counterparts, the limitations remain regarding the translation of subtle cultural nuances into English-language prompts; however, this work is hoped to establish a foundation for future scholarly inquiry. The VNFairyTales dataset, along with references, is hosted on GitHub for public access [13]. The dataset was validated across a diverse array of quantitative metrics, and the results are publicly accessible at https://lab.hcmute.fit/2MG6NZ16G.

3.2. Proposed Methodology

In this study, the base generative Story-Iter and the SigLIP evaluator are inherited components. Our primary novel contribution is the design and integration of the correlation-based, dynamic contextual filter block (Figure 3), which fundamentally alters how reference images are curated prior to generation to address the limitations of Story-Iter. Rather than incorporating a fixed set of the most recent images into the model, this filtering process is governed by a rule that identifies a quantity of images whose textual prompts exhibit the highest degree of similarity to the current prompt. Consider the following illustrative sequence of prompts:

Prompt 1: The stepmother brushes a beautiful girl's hair near a thatched hut. Another girl observes from afar.

Prompt 2: While the beautiful girl is bathing in a pond, another girl steals shrimps by putting them into her basket.

Prompt 3: A deity manifests before the beautiful girl to offer help.

Prompt 4: The beautiful girl feeds fish in a well with rice. Another girl watches behind the bush nearby.

Using a fixed reference image count of 2, Prompt 4 would typically reference Prompts 2 and 3. However, Prompts 1 and 2 are more contextually appropriate due to the presence of both primary characters and relevant environmental descriptors.

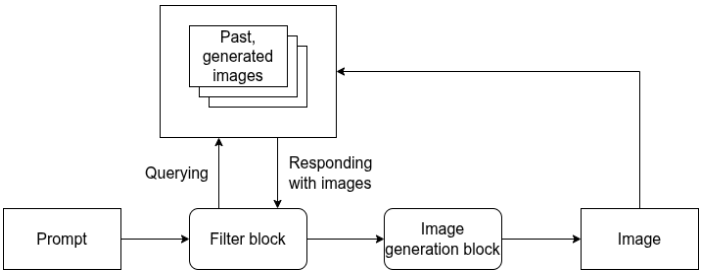

Figure 3. *Proposed architecture, with a contextual filter block.*

This proposed architecture includes a contextual filter block. The filter block operates to select high-ranking past generated images for context supplementation, thereby removing the semantic noise, and experiments confirm that the proposed filter method produces a visible enhancement. This research also leverages the Google SigLIP model [14] to assess the correlation between prompts and previously generated images. The algorithm selects a number (IdLen) of highest-scoring images to provide preferable contextual reference.

4. Results and Discussion

4.1. Experimental Setup

Table 1. *Experiment parameters.*

Parameter	Value	Description
Seed	42	To fix the random seed for number generation
numInferenceSteps	50	Number of times the diffusion model improves image
negPrompts	photorealistic, 3D render, glitch, distorted, abstract, anime, ugly, mutated, disfigured	Negative prompt, which describes which features should be removed
Width, height	640 × 640	Size of output image
IdLen	-1 (all), 2, 4, 6, 8, 10	Number of referenced images

Table 1 presents the parameter values used in the experiment. Negative prompts were adopted from prior studies [7]–[9] to ensure comic-style outputs. To prevent image distortion while remaining within the constraints of GPU memory, hyperparameters were optimized for dimensions and refinement iterations, ensuring efficient processing times while producing high-quality images.

Experimental validation was conducted using an NVIDIA RTX 3090 (24GB). Empirical observation suggests that while minor morphological distortions occur, principally within complex textures, the model maintains high fidelity regarding salient character attributes such as facial consistency and postural articulation.

4.2. Evaluation Metrics

In this study, quantitative evaluation is performed based on three metrics: CLIP-T [12], CLIP-I [12], and Character-wise Contextual Similarity (aCCS) [15]. CLIP-T assesses the semantic alignment between text and imagery using outputs from the CLIP model (ViT-L/14). Although it does not directly measure contextual maintenance, contextual frames nonetheless influence new image synthesis, resulting in variations in the alignment between textual content and visual output. CLIP-I compares two consecutive frames to analyze contextual consistency. For both CLIP-T and CLIP-I, higher values indicate stronger performance.

The aCCS metric [15] evaluates character consistency by using Grounding DINO SwinB v0.1.0-alpha2 to isolate characters from frames and comparing them via CLIP. It computes the average correlation coefficient for each character, bridging the gap between CLIP-T's text focus and CLIP-I's frame focus. While aCCS ignores background elements, it provides a precise measure of character identity. Greater values signify more effective character maintenance.

Finally, a user study evaluation was conducted via an online survey platform. Four representative frames were randomly selected from each architecture across three specific narrative tracks: *Mai An Tiêm*, *Chim Năm Trâu Sáu Cột và Chim Bắt Cô Trói Cột*, and *Trái Sầu Riêng*. Both models can use 4 past frames for context. Participants were instructed to evaluate the generated samples on a 5-point Likert scale, assessing criteria of narrative alignment, fidelity to traditional Vietnamese folktale culture, and overall aesthetic preference. Each respondent completed a total of 15 questions.

4.3. Experimental Results

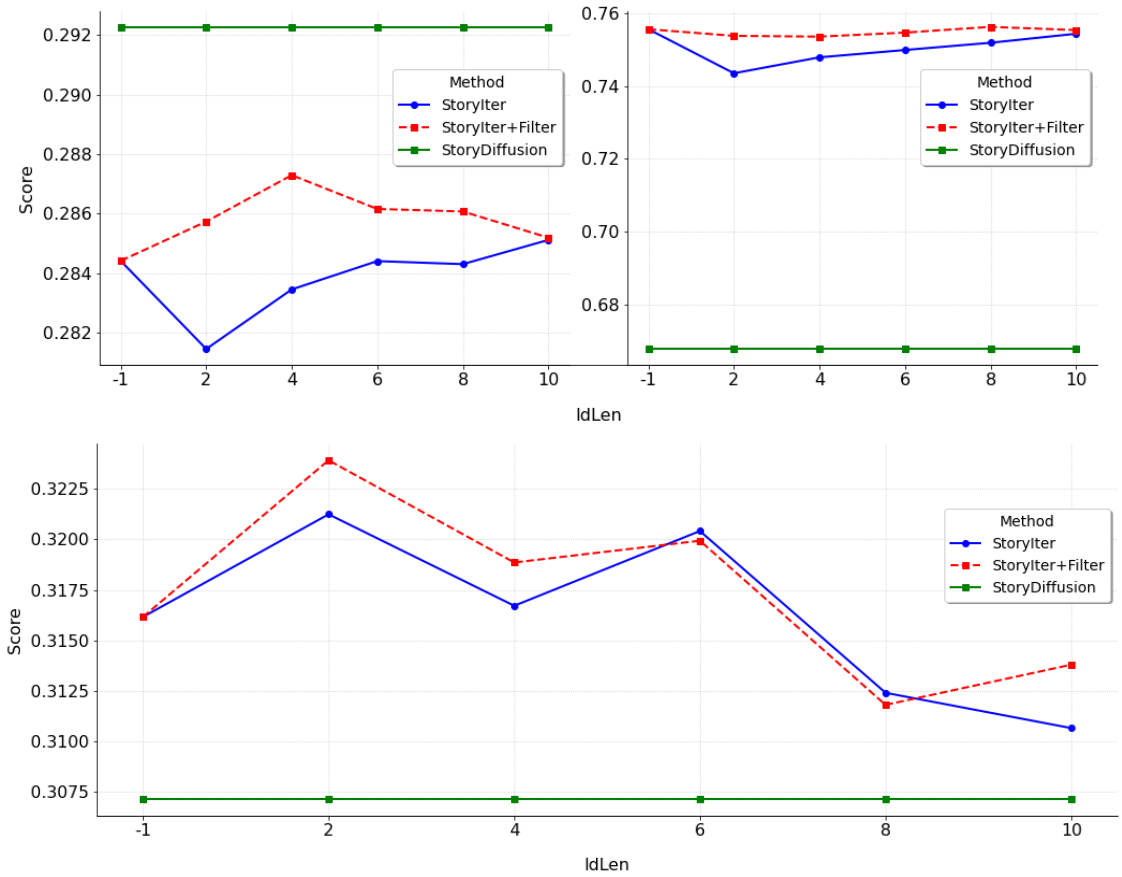

Figure 4. *Evaluation results for CLIP-T (top-left), CLIP-I (top-right), and aCCS (below) of the original Story-Iter (blue lines), StoryDiffusion (green line) and the enhancement (red, dash line).*

This study evaluates the baseline Story-Iter framework, its enhanced variant, and StoryDiffusion. Unlike Story-Iter, the context length of StoryDiffusion cannot be easily adjusted. Instead, the architecture employs a specific sampling rate to determine the context window size. Prior ablation studies indicate that rather than implementing a dynamic IdLen parameter, a constant value of 50% yields optimal performance across most scenarios [7]. Consequently, the performance curve corresponding to this 2024 model remains as a horizontal line throughout the evaluation.

Evaluation results of both CLIP-T and CLIP-I demonstrated in Figure 4 indicate a general trend: the proposed enhancement yields no significant difference when all images are selected (IdLen = -1). However, the efficacy increases sharply starting from a reference count of two (IdLen = 2) and gradually converges until no substantial divergence remains. This greatly demonstrates the weakness of standard Story-Iter in case of high context limitations. The visuals furthermore illustrate the steady performance of proposed method in both metrics.

The aCCS metric demonstrates greater consistency across evaluations. The proposed method yields discernible improvements in most cases, though marginal performance fluctuations were observed in others. Interestingly, increased IdLen values correlated with lower scores regardless of the architectural version, highlighting a diminishing return of dense context that will be discussed in subsequent sections.

Although both models fell short of matching StoryDiffusion in the CLIP-T evaluation, the performance deficits were minimal, remaining under 2.70%. Conversely, they substantially outperformed the StoryDiffusion in all remaining tests. The proposed method achieved a minimum improvement of 12.84% in the CLIP-I metric, and 1.5%–5.46% increase in aCCS.

Table 2. *Results and improvement margins (%) relative to context length: Upgrade vs. standard Story-Iter (bold) and StoryDiffusion (italic).*

IdLen	-1	2	4	6	8	10
CLIP-T	0.284 **0.00%** *-2.69%*	0.286 **+1.52%** *-2.24%*	0.287 **+1.35%** *-1.71%*	0.286 **+0.62%** *-2.09%*	0.286 **+0.62%** *-2.12%*	0.285 **+0.03%** *-2.42%*
CLIP-I	0.756 **0.00%** *+13.15%*	0.754 **+1.38%** *+12.87%*	0.754 **+0.76%** *+12.84%*	0.755 **+0.64%** *+13.01%*	0.756 **+0.58%** *+13.24%*	0.755 **+0.13%** *+13.11%*
aCCS	0.316 **0.00%** *+2.94%*	0.324 **+0.83%** *+5.46%*	0.319 **+0.68%** *+3.81%*	0.320 **-0.15%** *+4.16%*	0.312 **-0.19%** *+1.52%*	0.314 **+1.01%** *+2.17%*

Excluding the full-set baseline, experimental results for the improved Story-Iter in Table 2 show a CLIP-T score increase of 0.03% to 1.52% over the original solution. Similarly, CLIP-I yields favorable results, with a maximum performance gain of 1.38% in the two-image experiment, eventually decreasing to 0.13%. As IdLen increases, the metrics remain marginally higher than the baseline but exhibit a diminishing trend toward 0% as IdLen reaches its maximum value.

Prompt 7: ...the old mandarin **pointing towards** the distant sea on a map held by servants, Legendary Vietnamese king listening thoughtfully...

Prompt 9: ...**Mai's wife** carrying a child and a small bundle, standing firmly beside **Young man** who is in chains, guards and crying relatives in background...

Prompt 18: ...**Mai's wife** hugged **their child** and murmured... the couple planted even more melons... merchant ships and fishing boats ... bringing rice, clothing, chickens, pigs, tools, and even other types of seeds to trade for the melons.

Prompt 20: In those days, people clamored to buy the melons for their seeds, so within just a few years, the melon variety had spread everywhere. The names of **Mai An Tiem and his wife** became widely known...

Figure 5. *A sample of Mai An Tiêm tale generated by vanilla Story-Iter (above) and our method. The bold, blue parts of the prompt are details that were successfully captured by the newer approach but were ignored by the initial model.*

The aCCS metric converges with the baseline average when the full historical context is retained (IdLen = -1). Reducing IdLen to 2 results in a performance gain of 2.53%, though the score subsequently declines by up to 3.70% at an IdLen of 8. The most significant net improvement occurs at IdLen = 10 with a 1.01% increase, effectively rebounding from the marginal performance dips observed at IdLen 6 and 8.

Experimental evaluations also demonstrate that the enhanced architecture exhibits high efficacy in generating the *Mai An Tiêm* folktale narrative. This paper includes four generated frames where all models were restricted to a 4-frame context window. In the seventh frame embedded in Figure 5, the action of pointing toward the map was successfully captured by the newer model. In the ninth frame, the baseline model inadvertently inverted the gender of *Mai*'s wife. Conversely, our model with the integrated filtering mechanism preserved the character details accurately. This discrepancy may be attributed to the robust influence of the context history buffer, as the baseline model exhibited a severe bias resulting from the first eight reference frames containing exclusively male characters. Furthermore, while both models failed to synthesize the hugging posture, our proposed method successfully generated the child specified in the text prompt while emphasizing the protagonists better. The vanilla Story-Iter framework also exhibits identity confusion, repeating this error in the twentieth frame. In contrast, the enhanced methodology reliably synthesized a scene featuring *Mai An Tiêm* and his wife as the primary subjects. This visually demonstrates the underlying mechanism of our contribution: by filtering out temporally recent but semantically irrelevant male-dominated frames, our context filter prevented the GRCA attention weights from being biased, thus preserving the female character's identity.

Table 3. *User study average scores of our method from 40 respondents.*

Stories	***1*** ***Mai An Tiêm***	***2*** ***Chim Năm Trâu Sáu Cột và Chim Bắt Cô Trói Cột***	***3*** ***Trái Sầu Riêng***
Narrative alignment (max 5)	3.525 (+0.125)	3.300 (+0.200)	3.275 (-0.150)
Vietnamese folktale Culture alignment (max 5)	3.475 (+0.400)	3.275 (+1.175)	3.250 (-0.150)
Preference for new method	70.732%	65.854%	48.780%

The proposed solution generated comic sequences that were overall more favorably received by the majority of the 40 respondents as presented in Table 3. Specifically, *Mai An Tiêm* emerged as one of the most effectively synthesized narratives across all evaluation categories, securing a preference rate exceeding 70%. Furthermore, over 65% of the participants favored the version of *Chim Năm Trâu Sáu Cột và Chim Bắt Cô Trói Cột* produced by the enhanced method compared to the baseline model. Conversely, the results for *Trái Sầu Riêng* indicated a highly competitive distribution. Although the vanilla architecture was preferred, the performance margin remained narrow. Given that this specific narrative presented persistent difficulties in preceding evaluations, these subjective user preferences are predictable.

The evaluation notebook was made available to the public via https://lab.hcmute.fit/2MHR9S1QV. All the results were also uploaded, including story frames on https://github.com/RoggerTan/StoryVisual-Experiment-Results.

4.4. Discussion

The convergence of performance metrics at IdLen = -1 and the diminishing marginal utility as the reference window expands are logically sound, as the selective filtering algorithm becomes identical to the original approach when IdLen reaches the total number of frames. Nevertheless, the aCCS scores reveal a critical challenge concerning the volume of historical context utilized during synthesis. Given that folk narratives often involve multiple characters per scene, the corresponding frames frequently

contain several distinct identities. Both the baseline and enhanced architectures retrieve prior frames based solely on visual data, lacking the inherent ability to disambiguate specific character identities within those references. Consequently, dense contextual data can inadvertently degrade subsequent generation quality, leading to identity confusion or morphological distortions in the rendered subjects.

The model demonstrated less competitive performance regarding CLIP-T scores when compared to StoryDiffusion. Nonetheless, this metric accounts less for contextual maintenance. The observed success in aCCS and particularly within the CLIP-I benchmarks validates the efficacy of the proposed approach in enhancing overall consistency by reducing noise. Maintaining context is critical for the culturally specific task of generating Vietnamese folktale comics. While our proposed filtering stage improved performance, two theoretical limitations remain. First, SigLIP selects references based on prompt relevance rather than image quality. In instances where the chosen frames exhibit distorted details, such as facial deformation, the filter block inadvertently introduces noise into the diffusion process. For example, in the IdLen = 4 experiment, our model generated a *Mai An Tiêm* frame sequence displaying a deformed female face with a flattened nose at the fifteenth frame, a structural defect that persisted into the next sequential output. The second obstacle is the model's lack of domain-specific training on comics data, relying instead on Google's general WebLI dataset [11]. These challenges are likely to intensify in long-form storytelling scenarios (over 100 frames). To address both limitations, a theoretical evaluation block could be integrated to perform assessments on critical regions, such as character portraits. This architectural addition aims to establish a definitive, more robust scoring framework that accounts for both macro-level and detail-oriented semantic visual quality. To achieve this, commonly deployed image encoder models, such as InsightFace's ArcFace [16] and FaRL [17], can be leveraged.

The current experimental design presents several opportunities for refinement. Most notably, the simplicity of the user study survey limits the qualitative validation of the results compared to related literature [7]–[9]. Additionally, the inherent characteristics of Vietnamese folktales result in relatively short sequence lengths per data sample. Lastly, the reliance on manual grid searching rather than automated hyperparameter optimization suggests that the model's performance may not yet have reached its theoretical peak, potentially affecting the overall results.

5. Conclusion

The results yield modest yet consistent improvements over the previous architecture. While Story-Iter leverages GRCA to effectively process exhaustive global context, empirical data indicates a substantial performance degradation when the context window is constrained. In real-world deployment scenarios where computational costs and VRAM limitations are of high concern, our proposed selection strategy provides a critical enhancement for maintaining generative quality. Furthermore, to minimize the subjectivity and errors in prompts self-written by real-world users, a large language model can be integrated as a processing input layer following our guidelines established for the construction of VNFairyTales.

In future work, we intend to mitigate these constraints by performing an in-depth structural analysis of current model backbones to develop advanced optimization techniques. Additionally, to provide a more convincing experiment, we plan to incorporate human evaluation benchmarks to supplement our quantitative metrics.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Google's Gemini for English language editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] H. King, "Exclusive: Gen AI music app Suno comes out of stealth," *Axios*, Dec. 20, 2023. [Online]. Available: https://www.axios.com/2023/12/20/suno-gen-ai-music-microsoft. [Accessed: Mar. 25, 2026].
[2] OpenAI, "DALL·E 3 is now available in ChatGPT Plus and Enterprise," *OpenAI*, Oct. 19, 2023. [Online]. Available: https://openai.com/index/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise/. [Accessed: Mar. 25, 2026].
[3] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Nov. 2020, doi: 10.1145/3422622.
[4] Y. Li *et al.*, "StoryGAN: A Sequential Conditional GAN for Story Visualization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6329–6338.
[5] A. Maharana, D. Hannan, and M. Bansal, "Improving Generation and Evaluation of Visual Stories via Semantic Consistency," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 2021, pp. 2432–2443, doi: 10.18653/v1/2021.naacl-main.194.
[6] A. Maharana and M. Bansal, "Integrating visuospatial, linguistic, and commonsense structure into story visualization," in *Proc. 2021 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, Nov. 2021, pp. 6772–6786, doi: 10.18653/v1/2021.emnlp-main.543.
[7] Y. Zhou, D. Zhou, M. M. Cheng, J. Feng, and Q. Hou, "StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 110315–110340, doi: 10.52202/079017-3501.
[8] H. He *et al.*, "DreamStory: Open-domain story visualization by LLM-guided multi-subject consistent diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 12, pp. 11874–11891, 2025, doi: 10.1109/TPAMI.2025.3600149.
[9] J. Mao *et al.*, "Story-Iter: A Training-free Iterative Paradigm for Long Story Visualization," Jan. 2026. [Online]. Available: https://openreview.net/forum?id=puBVb9vTah. [Accessed: Mar. 25, 2026].
[10] C. Liu, H. Wu, Y. Zhong, X. Zhang, Y. Wang, and W. Xie, "Intelligent Grimm-open-ended visual storytelling via latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6190–6200.
[11] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11975–11986.
[12] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.
[13] Minh Tan Le *et al.*, "VNFairyTales," *GitHub*, 2026. [Online]. Available: https://github.com/RoggerTan/VNFairyTales. [Accessed: Mar. 25, 2026].
[14] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "google/siglip-base-patch16-224," *Hugging Face*, 2023. [Online]. Available: https://huggingface.co/google/siglip-base-patch16-224. [Accessed: Mar. 25, 2026].
[15] J. Cheng *et al.*, "Theatergen: Character management with LLM for consistent multi-turn image generation," 2024, *arXiv:2404.18919*. [Online]. Available: https://arxiv.org/abs/2404.18919.
[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.
[17] Y. Zheng et al., "General facial representation learning in a visual-linguistic manner," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18697–18709.

Minh Tan Le received his Bachelor's degree in Information Technology from Ho Chi Minh City University of Technology and Engineering in 2019, and completed his Master's degree in Computer Science in 2021. He is currently a Lecturer at the Faculty of Information Technology, Ho Chi Minh City University of Technology and Engineering, Vietnam. His academic and research interests center on Generative Artificial Intelligence and Computer Vision, with a focus on developing innovative solutions that push the boundaries of intelligent computing.
Email: tanlm@hcmute.edu.vn. ORCID: https://orcid.org/0009-0004-4912-6795.

Dinh Tan Loc Nguyen has been studying Data Engineer at Ho Chi Minh City University of Technology and Engineering since 2023. His academic and research interests center on Generative Artificial Intelligence.

Email: 23133041@student,hcmute.edu.vn. ORCID: https://orcid.org/0009-0007-8294-9007