Echoes of Imagination

1st Deivanai Thiyagarajan
Department of Engineering Education
University of Florida
Gainesville, United States
dthiyagarajan@ufl.edu

***Abstract*—This project, Echoes of Imagination, aims to generate illustrative visuals from narrative text through progressive multimodal modeling. The initial stage implemented a text-conditioned Variational Autoencoder (VAE) [1] in which only the decoder was conditioned on textual embeddings derived from BERT [5] and CLIP encoders. To improve text–image alignment, the latent dimension was increased, and both the encoder and decoder were subsequently conditioned on text features. Although this enhanced reconstruction quality, generation from text alone remained unstable. Building on these findings, the next phase transitioned to diffusion-based architectures, employing text-conditioned U-Nets with encoder–decoder cross-attention to strengthen semantic grounding and improve image fidelity. This iterative development establishes a foundation for robust multimodal storytelling by linking textual semantics and visual synthesis within a unified generative framework.**

***Index Terms*—Diffusion Models, Text-to-Image Generation, Multimodal AI, Story Visualization, Responsible AI, Variational AutoEncoders**

I. Introduction

Storytelling is one of humanity's oldest and most expressive traditions, yet it remains largely confined to static text. Echoes of Imagination explores how multimodal AI can bridge this gap by transforming written narratives into coherent and expressive visuals, creating more immersive storytelling experiences.

The project began with a text-conditioned Variational Autoencoder (VAE) [1], where text embeddings from BERT [5] and CLIP [4] guided the decoder during image reconstruction. While the model could reproduce visual features, text-only inputs yielded vague outputs, showing weak text–image alignment. Conditioning both the encoder and decoder on text and increasing the latent dimension improved quality but retained the blurriness typical of VAEs [1].

To overcome these limits, the project transitioned to diffusion-based architectures, using text-conditioned U-Nets that refine images through iterative denoising. Integrating BERT [5] and CLIP [4] embeddings via cross-attention enhanced semantic coherence and image fidelity. The final phase involved fine-tuning a Stable Diffusion model with DDPM [2] scheduling, resulting in more realistic and text-aligned outputs. This work presents the system's design, fine-tuning process, interface, and preliminary results toward AI-driven visual storytelling.

II. Related Work

Text-to-image synthesis has seen transformative advances with the emergence of diffusion models. Earlier GAN-based systems such as AttnGAN and DM-GAN achieved limited fidelity for complex prompts. Latent diffusion models (LDMs) like Stable Diffusion [3] compress high-dimensional images into manageable latent spaces, enabling efficient training and high-quality outputs. Multimodal embedding models such as CLIP [4] and BERT [5] further allow semantic conditioning on text. The *Echoes of Imagination* framework leverages these advances while emphasizing continuity across narrative fragments—a direction less explored in conventional text-to-image pipelines.

III. System Architecture and Pipeline

The proposed system operates iteratively to generate contextually coherent visuals from narrative text. Each iteration produces an image that reflects the current story segment while maintaining visual continuity with previous scenes. The overall architecture and data flow are shown in Fig. 1. The process consists of four key stages:

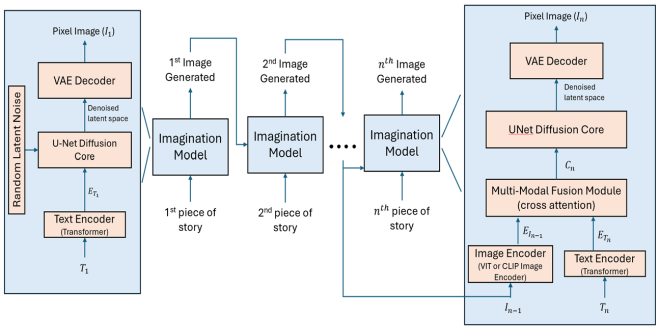

Fig. 1. Architecture Diagram.

A. Data Input

The system receives the story text, which is divided into smaller narrative units or storylets T_i.

B. Model Encoding (Iteration n)

The current storylet T_n and the previously generated image I_{n-1} are independently encoded. The text encoder (BERT or CLIP) produces the textual embedding $E(T_n)$, while the image encoder extracts the visual embedding $E(I_{n-1})$.

C. Diffusion Model and Conditioning

The diffusion backbone is a latent-space UNet denoiser trained with a Denoising Diffusion Probabilistic Model (DDPM) [2]. The model operates over 1000 diffusion steps, predicting noise residuals and refining latent representations iteratively. A pretrained VAE [1] encodes images into a $4 \times 32 \times 32$ latent space, reducing computational load by 16× compared to pixel-space diffusion. Conditioning on BERT [5] embeddings aligns latent noise predictions with narrative meaning.

D. Training Strategy

Training was performed on the HiPerGator cluster using an NVIDIA B200 GPU with 72GB VRAM. The model employed mixed-precision training, gradient accumulation, and distributed data parallelism to efficiently manage memory. The learning rate was set to 1×10^{-4} with cosine annealing and batch size of four. Loss stabilized around 0.38 after five epochs, down from an initial 0.82.

The dataset combined COCO and Flickr30k captions to provide both object-centric and descriptive contexts. When unavailable, dummy data generators ensured consistent training across subsets.

E. Inference and Fusion

The Multi-Modal Fusion Module, implemented via cross-attention, combines the text and image embeddings to form a joint conditioning vector $C_n = f(E(T_n), E(I_{n-1}))$. This conditioning vector guides the U-Net diffusion core to denoise a random latent variable z, generating a new latent image that preserves stylistic and character features from I_{n-1}.

F. User Interface and Reconstruction

The VAE [1] decoder transforms the latent output into a pixel-space image I_n, which is displayed to the user (see Figure 2). The generated image I_n is then reused as I_{n-1} for the next iteration, enabling consistent narrative visualization across the story sequence.

This iterative process establishes a feedback loop between text and image modalities, ensuring temporal coherence and visual consistency throughout the generated sequence.

IV. Model Implementation Details

A. Frameworks and Libraries

The project was implemented in `Python` using `PyTorch` for deep learning. Key libraries and tools include:

- **Core:** `numpy`, `pandas`, `matplotlib`, `seaborn`, `scikit-learn`, `tqdm`
- **Deep Learning:** `torch`, `torchvision`, `torchaudio`, `transformers`, `accelerate`, `sentencepiece`, `einops`
- **Text-to-Image / VAE / CLIP:** `diffusers`, `open_clip_torch`, `Pillow`
- **UI and Visualization:** `gradio`, `streamlit`
- **Utility:** `requests`, `jsonlines`, `pyyaml`

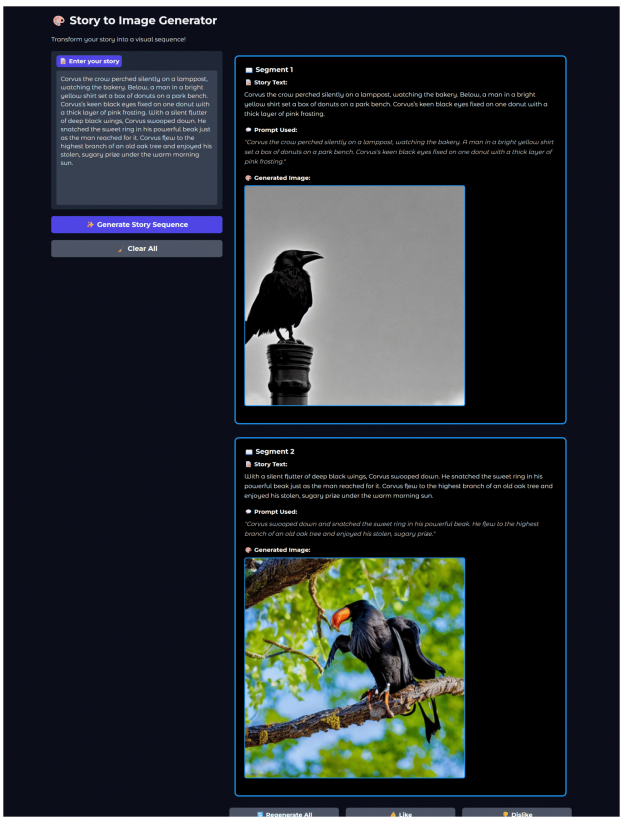

Fig. 2. User Interface Diagram.

- **Optional (diagrams or notebooks):** `jupyter`, `graphviz`
- **Dataset Handling:** `kagglehub`

B. Hardware and Training Setup

Training was conducted on an HPC node (`hpg-b200`) with the following specifications:

- GPU: 1
- CPU: 4 cores
- Memory: 30 GB

C. Hyperparameters

The VAE-based model was trained using the following settings:

- Optimizer: `AdamW`
- Learning rate: 2×10^{-4}
- Batch size: 16
- Number of epochs: 5
- Latent dimension: initially 4, later increased to 8
- Loss: Mean Squared Error (MSE) reconstruction loss + Kullback–Leibler (KL) divergence, optionally combined with L1 loss for better image reconstruction

D. Reproducibility

To ensure reproducibility:

- All random seeds for `PyTorch`, `numpy`, and `Python random` modules were fixed.
- Pretrained text encoders (`BERT`, `CLIP`) were frozen during initial training.
- Training scripts, dataset splits, and preprocessing routines are version-controlled.
- Core model components (`VAEEncoder`, `TextConditionedDecoder`, reparameterization, KL-loss functions) are modular and fully documented.
- Outputs and checkpoints are saved for each epoch to allow exact reconstruction of training experiments.

V. Inference Prototype

The interface prototype for Echoes of Imagination provides a user-friendly environment for transforming narrative text into illustrated sequences. It is implemented using Gradio, a lightweight Python library for building interactive web interfaces.

A. Input and Output

- Input: Users provide a story in free-form text via a multi-line Textbox. The system expects complete sentences but is robust to longer paragraphs.
- Processing: The system automatically: 1.Splits the input story into smaller storylets (4–5 sentences each) using NLTK's sentence tokenizer.
  2.Summarizes each storylet via a BART-based summarization pipeline to reduce token count (¡77 tokens), ensuring compatibility with the diffusion model. 3. Generates an image for each summarized storylet using a Stable Diffusion 2.1 pipeline.
- Output: For each storylet, the user sees: The original story segment (Markdown block), The generated illustration (Image block), Optional summarized text for context

B. Interface Layout

The interface features a column-based output display for generated storylets and associated images. Users can interact with the system via the following controls:

- Generate Story Sequence: Runs the pipeline on the provided story.
- Regenerate All: Re-generates images for previously processed storylets.
- Clear All: Clears the current output.
- Feedback Buttons (Like / Dislike): Collects user preferences for evaluation purposes.

The layout is designed for clarity and minimal user friction. Each storylet is displayed sequentially with its corresponding illustration, enabling a coherent storytelling experience.

Evaluation and Results

The initial evaluation focused on understanding the model's ability to reconstruct images and generate new content based on textual prompts. Early experiments were conducted using a Variational Autoencoder (VAE) [1] as the core generative backbone.

C. VAE Experiments

- Latent Dimension = 4, Text Conditioning Only in Decoder: Reconstruction quality of training images was poor. While the VAE could partially capture coarse structure, fine details were largely lost. A low-dimensional latent space constrained the model's ability to encode sufficient information for detailed image reconstruction. Conditioning only in the decoder was insufficient to influence meaningful image generation.
- Image reconstruction quality improved noticeably. Structural and color fidelity were better preserved compared to the previous configuration. Despite improved reconstruction, generating novel images using only textual prompts resulted in pure noise, indicating that the latent space alone was not sufficient for robust text-to-image generation.
- Conditioning the encoder in addition to the decoder did not result in measurable improvement for prompt-based image generation. Reconstructions remained comparable, but new generations continued to be noisy. The current VAE setup struggles to capture the correlation between latent representations and textual prompts when generating unseen images. This highlights the need for more advanced architectures or training strategies (e.g., full diffusion pipelines) for effective story-to-image synthesis.

D. Early Visual Evaluation

- Reconstruction Comparison: Sample reconstructions showed progressive improvement from latent dimension 4 to 8, with better preservation of shape and color details (see Figure 3).

Fig. 3. Image Reconstruction.

- Prompt-Based Generation: Images generated purely from text prompts were noisy in all configurations, reinforcing the limitation of the current VAE conditioning approach (see Figure 4).

E. Insights

Increasing the latent dimension and applying L1 regularization improved reconstruction but was insufficient for text-driven generation. Conditioning both encoder and decoder did

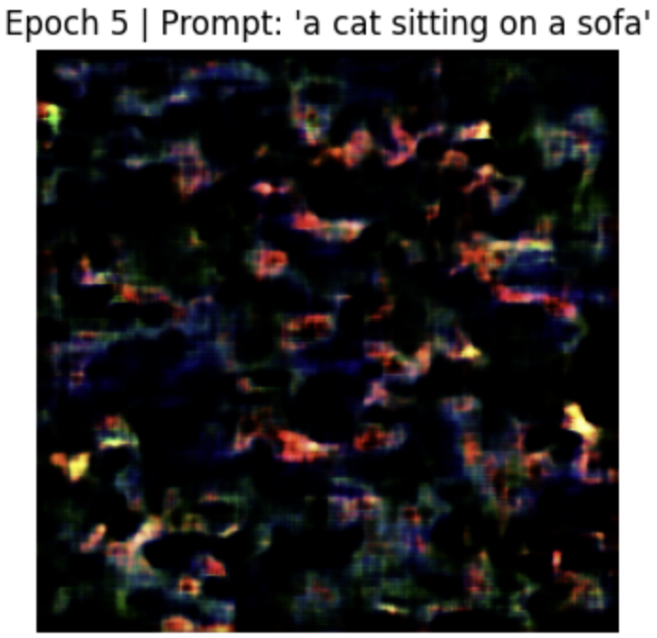

Fig. 4. Inference Diagram.

not meaningfully enhance generation quality. These results motivated the integration of a pretrained diffusion pipeline (e.g., Stable Diffusion) for robust text-to-image generation in the interface prototype.

Challenges and Next Steps

Several technical and data-related challenges were encountered during the development of the story-to-image generation system:

- Limited Latent Space in VAE: Early experiments with low-dimensional VAEs resulted in poor reconstruction quality and noisy image generation, highlighting the difficulty of encoding sufficient visual detail while conditioning on text prompts.
- Conditioning the VAE decoder or both encoder and decoder on textual prompts was insufficient for generating coherent images from new prompts, indicating that the model architecture was not robust enough for story-based image synthesis.
- Training more complex models, such as diffusion pipelines or large GANs from scratch, proved computationally intensive. On the current setup, training a diffusion model for just one epoch takes over 2 hours, limiting the ability to experiment with architectural or hyperparameter changes.

F. Next Steps

To address these challenges and improve model performance before Deliverable 3, the following steps are planned:

- Leverage Pretrained Diffusion Models: Transition from training VAEs or GANs from scratch to using pretrained Stable Diffusion pipelines, which provide high-quality image generation out-of-the-box.
- Fine-Tuning via Lightweight Architectures: Implement fine-tuning techniques such as LoRA (Low-Rank Adaptation) or similar methods on top of the pretrained model, specifically targeting story-driven image caption datasets. This approach reduces training time while adapting the model to the target task.
- Exploration of Alternative Generative Models: Evaluate deep convolutional GANs and text-conditioned diffusion models as alternatives for improved text-to-image fidelity, especially for long or complex story inputs.
- Interface Enhancements: Refine the Gradio interface to improve usability, feedback collection, and sequential storylet visualization, preparing the system for end-user evaluation.

VI. Responsible AI Reflection

During the development of the *Echoes of Imagination* story-to-image system, several ethical and fairness considerations were identified:

- **Bias in Generated Images:** Pretrained diffusion models can inherit biases present in their training data, potentially generating images that reinforce stereotypes or omit representation of certain groups. We plan to mitigate this by curating diverse and balanced prompts during evaluation and by exploring fine-tuning on datasets with equitable representation.
- **Content Safety:** Generated images may unintentionally produce unsafe or inappropriate content. To address this, we will integrate content moderation filters provided by the model's safety pipeline and allow user feedback to flag undesirable outputs.
- **Privacy of User Data:** Since the interface collects user-provided stories, there is potential for exposure of sensitive or personal information. We plan to implement local inference options and ensure that input data is not stored permanently, thereby minimizing privacy risks.
- **Transparency and User Awareness:** Users should understand that the images are AI-generated and may not faithfully reflect reality. The interface will include disclaimers and guidance on responsible use of generated content.

In the refinement phase, we aim to systematically incorporate these measures, ensuring that the system aligns with responsible AI principles while maintaining creativity and usability.

Acknowledgment

The author thanks the University of Florida HiPerGator team for computational resources. The implementation and source code for *Echoes of Imagination* are available at: https://github.com/yourusername/echoes-of-imagination.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[6] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pp. 38–45, 2020.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[8] A. Abid, M. Farooqi, and J. Zou, "Gradio: Hassle-Free Sharing and Testing of Machine Learning Models in Python," *arXiv preprint arXiv:1906.02569*, 2021.

[9] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.