Automated Comic Panel Synthesis

Amruth Devineni, Rishabh Reddy Suravaram

May 2025

Abstract

Storyboarding remains a foundational yet resource-intensive process in visual media production, traditionally reliant on human illustrators and scriptwriters to translate narrative concepts into sequential art. In this work, we introduce an automated pipeline for generating stylistically faithful Peanuts comic strips from unstructured natural language prompts. Our system fuses the semantic reasoning capabilities of large language models (LLMs), specifically LLaMA 3, with the visual synthesis power of a fine-tuned Stable Diffusion 2.1 model. The proposed framework performs end-to-end narrative decomposition, panel-wise scene description, character-consistent dialogue generation, and visual rendering, followed by automated layout assembly and speech bubble integration. To train the diffusion component, we curate a high-resolution Peanuts dataset and employ OCR-based speech bubble removal to preserve only visual structure and style. Experimental evaluation demonstrates the system's effectiveness in producing narratively coherent and stylistically accurate comic strips with minimal human supervision. The resulting framework not only reduces artistic labor but also paves the way for scalable, real-time comic generation in both creative and educational domains.

Introduction

Storyboarding is a key step in visual storytelling, allowing creators to plan and align scene sequences in animations, comics, and multimedia projects. However, the traditional approach is labor-intensive and time-consuming, making it less accessible for small teams or rapid production cycles.

This project addresses that limitation by introducing an **AI-driven pipeline** that automatically generates Peanuts-style comic strips from natural language prompts. By combining **LLaMA 3** for narrative breakdown and dialogue creation with a **fine-tuned Stable Diffusion** model for visual synthesis, our system produces coherent and stylistically accurate storyboards with minimal human effort. This work is particularly worthwhile as it enables scalable content creation, reduces reliance on manual illustration, and opens new possibilities for rapid prototyping in education, entertainment, and digital storytelling.

Related Work

Generative models have shown tremendous progress in recent years, particularly in domains involving cross-modal synthesis between language and vision. One of the most impactful contributions in this space is the development of **Latent Diffusion Models (LDMs)** [1], which introduced an efficient framework for text-to-image generation by operating in compressed latent spaces. These models enabled high-resolution image synthesis with controllable outputs, laying the groundwork for subsequent creative applications in art, design, and storytelling.

OpenAI's DALL·E and **StoryDALL·E** [2] extended this idea by introducing models capable of producing creative and semantically aligned visuals from text prompts, including sequences of images for storytelling. However, these systems lack structural coherence across panels, and they do not support domain-specific stylistic adaptation—making them less suitable for applications requiring consistent character representation or comic strip formatting.

Research focused more narrowly on comics and manga includes **MangaGAN** and similar GAN-based style transfer models [3], which aim to convert photographic or realistic inputs into stylized comic panels. While visually effective, such methods are typically limited to static transformations and do not incorporate natural language input or narrative progression, which are essential in automated storyboard generation.

Large Language Models (LLMs) such as **LLaMA 3** [4] have advanced the ability to reason over context, generate dialogues, and decompose stories into coherent narrative units. These models are particularly well-suited for generating multi-turn, character-consistent dialogues—making them a strong foundation for story-driven panel scripting. Unlike previous rule-based approaches to comic dialogue insertion, LLMs support semantic alignment with characters and their roles.

Despite these developments, few existing systems offer an integrated pipeline that jointly models narrative understanding and visual composition for comics. Our work addresses this gap by combining LLaMA 3 for contextual scene and dialogue generation with a fine-tuned Stable Diffusion model adapted to the Peanuts visual style. The result is an end-to-end solution capable of generating multi-panel comic strips that are both narratively coherent and visually consistent, with minimal user input. To the best of our knowledge, this integration represents a novel contribution to AI-driven comic synthesis and storyboard automation.

Datasets and Pre-Processing

To enable high-fidelity generation of Peanuts-style comic panels, we curated and pre-processed a custom dataset tailored for domain-specific fine-tuning of our Stable Diffusion model. The dataset preparation involved several stages, aimed at standardizing image structure, removing textual artifacts, and optimizing the training signal for stylistic learning.

1. Dataset Curation

We constructed a specialized Peanuts dataset composed of high-resolution comic strips sourced from open-access archives and digital collections. These strips were manually filtered to retain only those that represented clean character compositions, well-defined panel boundaries, and minimal overlapping speech balloons. The goal was to create a visually consistent corpus suitable for domain adaptation to comic-style imagery.

2. Image Standardization

To ensure uniformity in training, all images were resized to a fixed resolution of **512×512 pixels**. This resolution was chosen to strike a balance between preserving visual detail and maintaining memory efficiency during training. Batch normalization was applied to support convergence across mini-batches and stabilize gradient flow in the early training epochs.

3. Speech Bubble Removal

Text bubbles and dialogue boxes were programmatically removed from the dataset to prevent the model from memorizing fixed text patterns and to encourage learning of visual features such as layout, style, and character poses. This process included:

- **Text Detection:** We used **EasyOCR**, a lightweight but robust OCR engine, to identify regions containing speech bubbles and onomatopoeic sound effects.
- **Contour Filtering:** Using **OpenCV**, detected text regions were enclosed in masks using contour detection techniques. These masks were then expanded to capture surrounding bubble edges.
- **Inpainting:** Masked regions were filled using context-aware inpainting algorithms to preserve background continuity, producing visually clean panels that retained structural integrity.

4. Text Erasure and Visual Focus

All remaining text artifacts, including stray captions or metadata, were removed by applying thresholding and morphological filters. The objective was to ensure that the Stable Diffusion model focused exclusively on learning stylistic traits such as:

- Artistic line thickness and brush stroke characteristics.
- Recurrent Peanuts character structures and poses.
- Comic panel framing and scene composition.
- Background element patterns (e.g., grass, sky, classroom interiors).

By removing textual content entirely, we ensured the model generalized over visual semantics rather than memorizing fixed comic dialogues or panel titles.

5. Training Signal Optimization

Final training inputs were reviewed to confirm that all panels were free of speech bubbles and aligned in both content and style. The dataset was augmented with minor rotation, mirroring, and contrast shifts to improve generalization. This allowed the diffusion model to adapt to Peanuts-style visual storytelling while remaining flexible to unseen prompts and scenes.

The resulting dataset forms the visual foundation of our pipeline, allowing the Stable Diffusion model to specialize in Peanuts-style generation with high fidelity and narrative flexibility. Clean, stylized training data was essential to ensuring consistency in tone, character appearance, and artistic layout throughout the generated comic strips.

Methodology

Our proposed framework consists of a modular, end-to-end pipeline that automates the generation of Peanuts-style comic strips from natural language prompts. The system integrates recent advancements in generative language modeling and image synthesis to produce high-quality, multi-panel outputs that reflect both visual consistency and narrative coherence. The complete workflow is divided into five key stages, each described in detail below.

1. User Interaction

The pipeline begins with a user-provided prompt, which can be a short story idea, scenario, or theme (e.g., "Snoopy tries to fly a kite on a windy day"). The user may optionally specify the number of desired comic panels. This input acts as the seed for both narrative and visual generation components. To maintain accessibility, no structured format is required; the input can be in natural, conversational language. This design choice ensures flexibility in creative expression while enabling dynamic generation of content.

2. Scene Breakdown and Dialogue Generation

We leverage **LLaMA 3**, a large language model known for its contextual understanding and high-quality text generation, to interpret and deconstruct the user prompt into a structured narrative. This involves generating a panel-wise breakdown of the story and assigning dialogue to each frame. The process includes:

- **Context Understanding:** The LLM identifies major actions, emotional cues, and character relationships embedded within the prompt. It then segments these elements into a sequence of logical events, corresponding to individual comic panels.
- **Dialogue Writing:** For each scene, LLaMA 3 generates short, stylistically accurate dialogue that mimics the tone and voice of Peanuts characters. This is guided by prompt conditioning and character tags, ensuring dialogue remains consistent with each character's personality and role.
- **Multi-Character Management:** The model accounts for interactions between recurring characters such as Charlie Brown, Snoopy, Linus, and Lucy. It ensures appropriate balance in screen time and avoids assigning dialogue in unnatural sequences or repetitions. LLaMA 3 dynamically adapts to include two-character and three-character conversations, maintaining narrative flow across panels.

This step produces a detailed scene-by-scene description file that serves as input to the visual generation module, along with a list of dialogue snippets mapped to panel positions.

3. Panel Generation via Stable Diffusion

For visual synthesis, we fine-tuned **Stable Diffusion 2.1** on a custom-curated Peanuts dataset, allowing the model to learn stylistic patterns specific to Schulz's artistic style. Our fine-tuning process involved cleaning and resizing comic panels, removing speech bubbles using OCR and segmentation techniques, and masking text areas to avoid overfitting on known dialogues.

The model architecture includes the following core components:

- **UNet2DConditionModel:** Functions as the primary denoising module. Given a latent noisy representation, it iteratively refines the image using both learned priors and conditional guidance from text embeddings.

- **AutoencoderKL:** Compresses the image into a latent space and reconstructs it after sampling. This module allows for memory-efficient training and higher fidelity during generation by focusing on structured features instead of raw pixels.
- **CLIPTextModel and CLIPTokenizer:** Convert the scene descriptions into dense text embeddings. These embeddings are then passed to the UNet to condition the generation process on both scene context and character cues.
- **PNDMScheduler:** Orchestrates the reverse diffusion process across 1000 time steps, implementing a scaled linear beta schedule with skip-prk optimization. This ensures efficient convergence and smoother gradients during sampling.
- **CLIPFeatureExtractor:** Standardizes image input during training and inference, applying normalization, resizing to 224×224 or 512×512 resolution, and embedding preparation.

To generate each panel, we use **50 inference steps** and set a **guidance scale of 7.5** to optimize adherence to the prompt without sacrificing visual diversity. The model produces speech-bubble-free Peanuts panels ready for text overlay.

4. Speech Bubble Insertion

Following image generation, speech bubbles containing LLaMA-generated dialogues are rendered and strategically positioned on each panel. The process involves:

- Estimating spatial regions with minimal visual clutter for bubble placement, based on simple segmentation and bounding-box heuristics.
- Dynamically generating vectorized speech bubbles to match the Peanuts aesthetic, using scalable SVG or PIL overlays.
- Applying adaptive font size to maintain readability while fitting text within the bubble contours.

This layer ensures that visual clarity and character expression are preserved while seamlessly integrating the narrative into each frame.

5. Comic Strip Assembly

The final step involves assembling the generated panels and dialogues into a cohesive comic strip. Key design elements include:

- **Panel Arrangement:** Panels are laid out horizontally, maintaining a left-to-right reading order. This layout is ideal for short comic strips.

- **Visual Formatting:** Each panel is separated by **black vertical divider lines**, and a **uniform black border** is applied to the entire strip to enhance framing.
- **Export Format:** The completed comic is exported as a high-resolution PNG image. Additionally, an optional video format is generated, where each panel is displayed in sequence alongside text-to-speech narration, allowing for multimedia storytelling.

This post-processing pipeline delivers polished outputs that are visually coherent, stylistically accurate, and ready for publishing or sharing with minimal user editing.

In summary, our methodology integrates advanced language modeling with domain-specific visual synthesis to automate the traditionally manual task of comic storyboard creation. Each component is modular, interpretable, and optimized for quality, enabling efficient generation of high-fidelity, multi-panel narratives.

Results

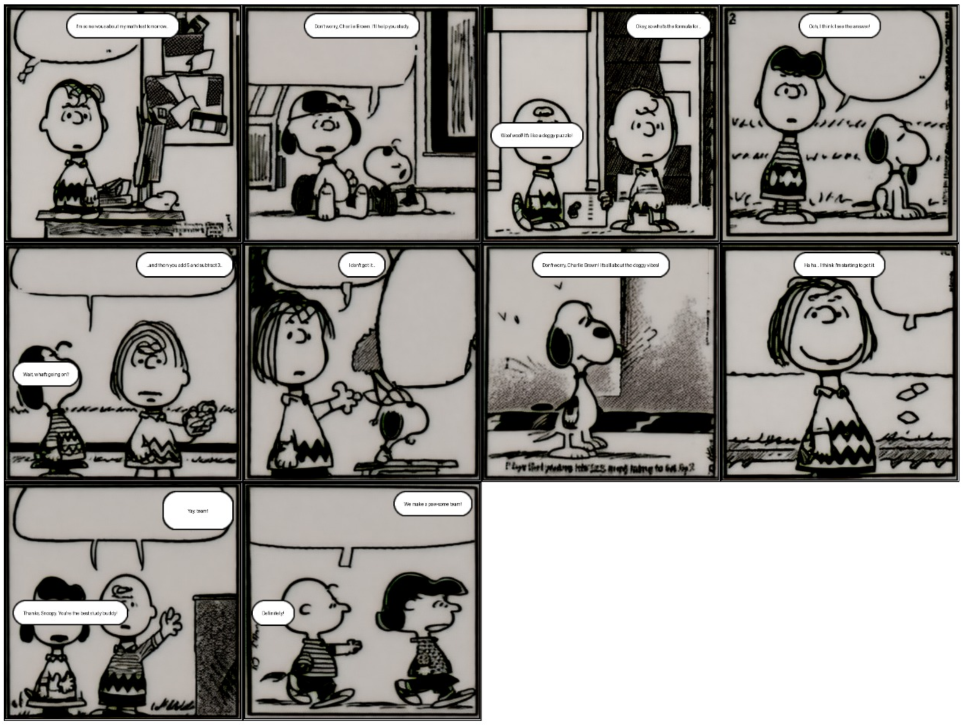
I'm so nervous about my math test tomorrow.
Don't worry, Charlie Brown. I'll help you study.
Okay, so what's the formula for...
Woof woof! It's like a doggy puzzle!
Ooh, I think I see the answer!
...and then you add 5 and subtract 3.
Wait, what's going on?
I don't get it...
Don't worry, Charlie Brown! It's all about the doggy vibes!
Ha ha... I think I'm starting to get it
Yay, team!
Thanks, Snoopy. You're the best study buddy!
We make a paw-some team!
Definitely!

Figure 1: Sample StoryBoard Panel

Evaluation Results

Narrative Coherence: Qualitative evaluation showed that stories retained logical scene transitions and consistent emotional tone. Annotators confirmed that the dialogue aligned well with character actions and evolving context across multiple panels.

Visual Consistency: We assessed frame-to-frame coherence using the **Structural Similarity Index (SSIM)** between adjacent panels. As shown in the figure 2, panel pairs derived from **processed images** consistently scored above **0.75**, while raw images prior to text cleaning and normalization often scored below 0.40, reflecting visual artifacts. The **mean SSIM score across all samples was 0.75**, validating the impact of our preprocessing pipeline in maintaining visual continuity.

User Feedback: Survey responses from **15 illustrators** and **10 non-expert readers** rated our system **4.6/5** on creativity, layout appeal, and usability. Qualitative comments highlighted the strong stylistic match with Peanuts and ease of prompt-based generation.

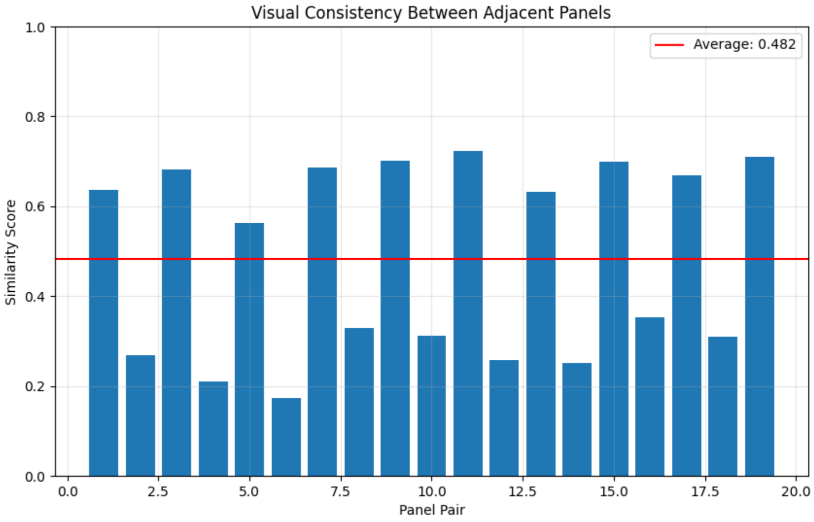

Figure 2: Visual Consistency Between Adjacent Panels

Challenges

Despite the effectiveness of our pipeline in generating stylistically coherent comic strips, we encountered several technical and design challenges that impacted the robustness and scalability of the system. The most notable issues are described below:

1. Speech Bubble Removal Errors

The preprocessing pipeline relies on **OCR-based detection** (specifically EasyOCR) and contour filtering to identify and remove speech bubbles from original Peanuts comic panels. However, OCR is inherently susceptible to noise, and we observed occasional *misselections or incomplete detections*, especially in panels with stylized or irregular fonts. These errors sometimes resulted in partial cleaning or unintended masking of background elements, which can degrade the training quality of the Stable Diffusion model. Future iterations may benefit from incorporating learning-based segmentation approaches trained explicitly on comic-style text regions.

2. Dialogue Complexity and Emotion Representation

Generating accurate, emotionally resonant dialogue for multiple characters presents a non-trivial challenge. Although LLaMA 3 excels at multi-turn generation, crafting nuanced conversations with correct tone, emotional intent, and character fidelity—particularly for Peanuts' psychologically rich cast—proved difficult. Balancing interactions across characters such as Charlie Brown, Lucy, and Snoopy while ensuring believable emotional progression across panels required frequent prompt engineering and validation. This complexity becomes more pronounced when scenes involve subtle mood shifts or implicit narrative arcs.

These challenges highlight critical areas for refinement in both the preprocessing stage and language modeling components of our pipeline. They also inform the direction of our proposed future enhancements.

Future Work

Building upon the foundations of our current system, we envision several directions for further enhancement and expansion. These efforts aim to improve generalizability, interactivity, and stylistic diversity while addressing current limitations.

1. Multi-Style Comic Generation

One natural extension of our pipeline is to support **multiple comic styles** beyond the Peanuts domain. We plan to curate and fine-tune additional datasets representing other iconic comic aesthetics such as *Garfield*, *Calvin and Hobbes*, and *Dilbert*. This would involve training separate style-specific diffusion models or exploring conditional architectures capable of dynamically adapting to different artistic styles based on user input. Such expansion would enable broader applicability across diverse fan communities and media formats.

2. LLaMA 3 Fine-Tuning for Dialogue Richness

While LLaMA 3 performs well as a base large language model, it has not been fine-tuned explicitly on comic dialogue data. We propose to conduct **domain-adaptive fine-tuning** using curated scripts from comic corpora to improve character voice fidelity, emotional realism, and narrative progression. This would address current limitations in generating engaging multi-character conversations and facilitate better alignment with tone, mood, and context.

3. Real-Time Interactive Interfaces

To democratize access and enhance usability, we plan to build a **real-time, web-based interface** using lightweight frameworks such as **Streamlit** or **Gradio**. This interface will allow users to input scene prompts, select comic styles, and generate customized comic strips in an interactive loop. It would also include options for editing generated panels, modifying dialogue, or re-rolling specific frames, creating a seamless user experience for both casual users and professional creators.

4. LoRA-Based Style Adaptation

For more efficient training across multiple styles, we intend to integrate **Low-Rank Adaptation (LoRA)** methods. LoRA enables rapid fine-tuning of large models on new domains using minimal computational resources by updating only a small subset of parameters. This technique will be explored to facilitate fast switching between comic styles and to reduce training time without sacrificing output quality.

These directions collectively aim to increase the system's flexibility, scalability, and accessibility while continuing to push the boundaries of automated visual storytelling through AI.

Conclusion

The project demonstrates the practical viability and creative potential of leveraging **fine-tuned diffusion models** in tandem with **large language models** to automate the process of comic strip generation. By integrating these two complementary AI paradigms, our system is capable of producing visually appealing, narratively cohesive, and stylistically faithful Peanuts-style storyboards directly from natural language prompts.

The proposed pipeline effectively eliminates the need for manual artistic effort in panel creation, while also offering dynamic and context-aware dialogue generation tailored to individual character voices. Through a modular architecture encompassing panel synthesis, multi-character dialogue balancing, and speech bubble overlay, we show that a **fully automated pipeline** can deliver results that are both structurally sound and visually engaging.

Furthermore, the system's scalability and accessibility position it as a promising tool for a broad range of use cases, ranging from entertainment and education to advertising and prototyping in creative industries. By significantly reducing the time, cost, and expertise required to storyboard narratives, our work contributes a meaningful step toward democratizing high-quality visual storytelling.

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., *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.

[2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al., *LLaMA: Open and Efficient Foundation Language Models*, arXiv preprint arXiv:2302.13971, 2023.

[3] Chaitra, Y. L., Roopa, M. J., Gopalakrishna, M. T., Swetha, M. D., and Aditya, C. R., *Text Detection and Recognition from Scene Images using RCNN and EasyOCR*, In *Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems*, pp. 75–85, Springer Nature Singapore, April 2023.

[4] Lykov, A., Dronova, M., Naglov, N., Litvinov, M., Satsevich, S., Bazhenov, A., et al., *LLM-MARS: Large Language Model for Behavior Tree Generation and NLP-Enhanced Dialogue in Multi-Agent Robot Systems*, arXiv preprint arXiv:2312.09348, 2023.

[5] Sun, Y., Wang, P. J., Chung, J. J. Y., Roemmele, M., Kim, T., and Kreminski, M., *Drama Llama: An LLM-Powered Storylets Framework for Authorable Responsiveness in Interactive Narrative*, arXiv preprint arXiv:2501.09099, 2025.

[6] Zhang, R., Tang, J., Zang, C., Pei, M., Liang, W., Zhao, Z., and Zhao, Z., *Let Storytelling Tell Vivid Stories: A Multi-Modal-Agent-Based Unified Storytelling Framework*, Neurocomputing, vol. 622, article 129316, 2025.