Character Consistency in AI-Generated Manga: A Comparative Study of Modern Approaches

Andrew Njoo

December 24, 2025

Abstract

Character consistency is a fundamental challenge in AI-generated sequential art, particularly in manga comics where characters must maintain visual identity across multiple panels. This paper presents a comprehensive survey of state-of-the-art AI image generation models and their approaches to character consistency, with a focus on manga generation. We examine DALL-E 3, Stable Diffusion XL, Midjourney, and other contemporary models, analyzing their strengths and limitations. Additionally, we present the MangaBanana system, which implements a prompt-based character consistency strategy using DALL-E 3. Our comparative analysis reveals that while prompt engineering offers simplicity and accessibility, embedding-based methods and fine-tuning provide superior consistency at the cost of increased complexity. We discuss the trade-offs between different approaches and provide insights for future research directions.

1 Introduction

The advent of large-scale generative AI models has revolutionized digital art creation, enabling users to generate high-quality images from text descriptions. However, a persistent challenge remains: maintaining character consistency across multiple generated images. This problem is particularly acute in sequential art forms such as manga comics, where characters must retain their visual identity across panels to ensure narrative coherence and reader engagement.

1.1 Problem Statement

In traditional manga creation, artists maintain character consistency through reference sheets, established drawing styles, and years of practice. AI-generated manga faces a fundamentally different challenge: each panel is generated independently, and without explicit mechanisms for consistency, the same character description can produce visually distinct results. This inconsistency breaks narrative immersion and limits the practical utility of AI tools for sequential storytelling.

The core problem manifests in several ways:

- **Visual Variability**: The same character description can produce different facial features, body proportions, and clothing details across panels.
- **Style Drift**: Generated images may exhibit subtle variations in art style, lighting, or color palette.
- **Lack of Memory**: Current models generate each image independently, without awareness of previously generated panels.

1.2 Motivation

Character consistency is not merely an aesthetic concern; it is essential for narrative coherence. In manga and comics, readers rely on visual continuity to follow storylines and identify characters. Inconsistent character appearance can confuse readers, disrupt narrative flow, and undermine the artistic integrity of the work.

The importance of this problem is underscored by the growing demand for AI-assisted content creation tools. As these tools become more accessible, the need for reliable character consistency mechanisms becomes increasingly critical for practical applications in entertainment, education, and creative industries.

1.3 Contributions

This paper makes three primary contributions:

1. **Comprehensive Survey**: We provide an up-to-date survey of state-of-the-art AI image generation models (DALL-E 3, Stable Diffusion XL, Midjourney, Flux) and their approaches to character consistency.

2. **Implementation Analysis**: We present the MangaBanana system, which implements a prompt-based character consistency strategy for two-panel manga generation using DALL-E 3 and GPT-4o-mini.

3. **Comparative Analysis**: We analyze the trade-offs between different approaches, including prompt engineering, character embeddings, and fine-tuning methods, providing insights for practitioners and researchers.

1.4 Paper Organization

The remainder of this paper is organized as follows: Section 2 reviews related work on AI image generation and character consistency techniques. Section 3 details the MangaBanana system architecture and implementation. Section 4 presents a comparative analysis of different approaches. Section 5 discusses results and limitations. Section 6 concludes with future research directions.

2 Related Work

2.1 State-of-the-Art Image Generation Models

2.1.1 DALL-E 3

DALL-E 3, developed by OpenAI, represents a significant advancement in text-to-image generation Betker et al. [2023]. Unlike its predecessors, DALL-E 3 integrates directly with GPT-4 for prompt understanding, enabling more nuanced interpretation of user instructions. The model generates high-quality images at 1024x1024 resolution with improved prompt adherence.

However, DALL-E 3 faces challenges with character consistency. The model generates each image independently, and while detailed prompts can improve consistency, there is no built-in mechanism to reference previous generations. Users must rely on prompt engineering techniques, such as including detailed character descriptions in each generation request.

2.1.2 Stable Diffusion XL

Stable Diffusion XL (SDXL) is an open-source diffusion model that offers greater control through various mechanisms Rombach et al. [2022]. Unlike DALL-E 3, SDXL supports:

- **Character Embeddings**: Users can create character-specific embeddings through training or fine-tuning, enabling consistent character generation across images.
- **ControlNet**: A neural network structure that allows additional control conditions, such as pose, depth, or edge maps, which can aid in maintaining character structure.
- **Image-to-Image**: The ability to use reference images as input, enabling style transfer and partial consistency.

These features make SDXL more flexible for character consistency, though they require additional setup and technical expertise compared to prompt-based approaches.

2.1.3 Midjourney

Midjourney is a proprietary image generation model known for its artistic quality and style consistency Midjourney [2024]. The platform offers several features relevant to character consistency:

- **Character References**: Users can reference previously generated images using the `-cref` parameter to maintain character appearance.
- **Style Consistency**: The model excels at maintaining consistent artistic styles across generations.
- **Seed Control**: Users can specify seeds to reproduce similar results, though this provides limited control over character consistency.

While Midjourney's character reference feature is a step toward consistency, it requires manual management of reference images and may not guarantee perfect character matching.

2.1.4 Other Models

Recent models such as Flux Black et al. [2024] and Imagen 3 Saharia et al. [2022] have also made significant contributions to image generation quality. Flux, in particular, offers improved prompt understanding and image quality, though character consistency remains a challenge without additional mechanisms.

2.2 Character Consistency Approaches

2.2.1 Prompt Engineering

The simplest approach to character consistency involves carefully crafting prompts that include detailed character descriptions. This method is model-agnostic and requires no additional training or setup. Techniques include:

- Including comprehensive character descriptions (appearance, clothing, features) in every prompt
- Using consistent style keywords across generations
- Explicitly instructing the model to maintain character consistency

While accessible, prompt engineering has limitations: it cannot guarantee perfect consistency, requires manual prompt management, and may not scale well to longer sequences.

2.2.2 Character Embeddings and Fine-tuning

More advanced approaches involve creating character-specific embeddings or fine-tuning models on character datasets. This typically involves:

- Training LoRA (Low-Rank Adaptation) models for specific characters
- Creating textual inversions or embeddings that encode character appearance
- Fine-tuning base models on character-specific datasets

These methods offer superior consistency but require:

- Training data (multiple images of the character)
- Computational resources for training
- Technical expertise in model training
- Storage and management of character-specific models

2.2.3 Control Mechanisms

ControlNet and similar control mechanisms allow users to guide generation through additional inputs such as:

- Pose estimation and control
- Depth maps
- Edge detection and structure preservation
- Segmentation masks

These mechanisms can help maintain character structure and pose consistency, though they require additional preprocessing and may not preserve fine-grained character details.

2.2.4 Multi-Model Pipelines

Some systems combine multiple models or techniques:

- Using one model for character generation and another for scene composition
- Combining prompt engineering with image-to-image translation
- Integrating character embeddings with control mechanisms

These approaches can achieve better results but increase system complexity and computational requirements.

2.3 Gaps in Current Research

While significant progress has been made in image generation quality, research specifically focused on character consistency in sequential art remains limited. Most existing work focuses on single-image generation or general image-to-image consistency, with less attention to the specific requirements of manga and comic generation. This paper addresses this gap by examining practical approaches to character consistency in the context of multi-panel manga generation.

3 Methodology: The MangaBanana System

This section presents the MangaBanana system, a web-based platform for generating two-panel manga comics with a focus on character consistency. The system demonstrates a practical implementation of prompt-based character consistency using DALL-E 3.

3.1 System Architecture

MangaBanana is built as a Next.js web application with three primary components:

1. **Story Splitting Module**: Uses GPT-4o-mini to intelligently split user-provided stories into two panel prompts.
2. **Character Consistency Engine**: Implements prompt-based consistency through character description prefixing.
3. **Image Generation Pipeline**: Sequentially generates two panels using DALL-E 3 with consistency instructions.

The system architecture follows a sequential pipeline: users provide a story description and optional character description, the story is split into two panel prompts, and then both panels are generated with shared character information.

3.2 Character Consistency Strategy

The core of MangaBanana's character consistency approach is *character description prefixing*. This strategy involves:

1. **Character Description Extraction**: The system accepts an optional character description from the user, which includes details such as appearance, clothing, and distinctive features.
2. **Prefix Construction**: The character description is formatted as a prefix that is prepended to each panel prompt:

Listing 1: Character Prefix Construction

```
const characterPrefix = characterDescription &&
  characterDescription.trim()
  ? `${characterDescription}. `
  : '';
```

3. **Prompt Assembly**: Each panel prompt is constructed by combining:
   - Panel-specific narrative content (from story splitting)
   - Character description prefix
   - Style specifications
   - Explicit consistency instructions
4. **Consistency Instructions**: Explicit instructions are added to prompts to reinforce consistency:
   - Panel 1: "Same character design and appearance throughout."
   - Panel 2: "Must use the exact same character design, appearance, and style as the first panel. Character must look identical."

3.3 Story Splitting Implementation

The story splitting module uses GPT-4o-mini to intelligently divide user-provided stories into two panel prompts. The system prompt instructs the model to:

- Create a setup/beginning for the first panel
- Create a continuation/punchline for the second panel
- Incorporate character descriptions into both prompts
- Return structured JSON with panel prompts

This approach ensures that panel prompts are narratively coherent while maintaining character context. The system prompt is designed to encourage the inclusion of character descriptions in both generated prompts, reinforcing consistency.

3.4 Two-Panel Generation Pipeline

The generation pipeline operates sequentially:

1. **Panel 1 Generation**:

Listing 2: Panel 1 Prompt Construction

```
const panel1FullPrompt =
  `First panel of a 2-panel manga comic.
   ${characterPrefix}${panel1Prompt}.
   Style: ${style || 'classic manga'}.
   High quality, detailed, vibrant colors,
   professional manga art style.
   Manga comic panel format.
   Same character design and appearance throughout.`;
```

2. **Panel 2 Generation**: After Panel 1 is successfully generated, Panel 2 is created with stronger consistency instructions:

Listing 3: Panel 2 Prompt Construction

```
const panel2FullPrompt =
  `Second panel of a 2-panel manga comic.
   ${characterPrefix}${panel2Prompt}.
   Style: ${style || 'classic manga'}.
   High quality, detailed, vibrant colors,
   professional manga art style.
   Manga comic panel format.
   Must use the exact same character design,
   appearance, and style as the first panel.
   Character must look identical.`;
```

The sequential generation ensures that Panel 1 establishes the character appearance, while Panel 2 receives explicit instructions to match it. However, it is important to note that DALL-E 3 does not have direct access to Panel 1's output during Panel 2 generation; consistency relies entirely on prompt engineering.

3.5 Technical Implementation Details

3.5.1 Model Configuration

The system uses DALL-E 3 with the following settings:

- Model: `dall-e-3`
- Resolution: 1024x1024 pixels
- Quality: Standard (balanced quality and cost)
- Number of images: 1 per panel

3.5.2 Error Handling

The implementation includes robust error handling:

- Credit refunding if generation fails
- Validation of API responses
- Graceful degradation when character descriptions are missing

3.5.3 User API Key Support

The system supports both server-provided API keys and user-provided keys, enabling users to bypass rate limits by using their own OpenAI API keys. This design choice reflects the practical constraints of API-based systems while providing flexibility for power users.

3.6 Limitations of the Approach

The prompt-based approach has several inherent limitations:

- **No Visual Memory**: DALL-E 3 cannot directly reference Panel 1's visual output when generating Panel 2.
- **Prompt Dependency**: Consistency relies entirely on textual descriptions, which may not capture all visual nuances.
- **Variability**: Even with detailed prompts, some visual variation is expected due to the stochastic nature of generation.
- **Scalability**: The approach may become less effective with longer sequences (more than 2-3 panels).

Despite these limitations, the approach offers significant advantages: simplicity, accessibility, no training requirements, and immediate usability with existing API services.

4 Comparative Analysis

This section provides a comparative analysis of different approaches to character consistency in AI-generated manga, examining their strengths, limitations, and practical trade-offs.

4.1 Approach Comparison

4.1.1 Prompt Engineering (MangaBanana/DALL-E 3)

Strengths:

- **Simplicity**: No training or setup required; works immediately with API access
- **Accessibility**: Usable by non-technical users; no machine learning expertise needed
- **Cost-Effective**: No computational overhead beyond API calls
- **Flexibility**: Easy to modify prompts for different characters or styles
- **No Storage Requirements**: No need to store character models or embeddings

Limitations:

- **Inconsistent Results**: Cannot guarantee perfect character consistency
- **Prompt Management**: Requires careful prompt construction and maintenance
- **Limited Scalability**: Effectiveness decreases with longer sequences
- **No Visual Reference**: Cannot directly reference previous panel outputs
- **User Effort**: Requires detailed character descriptions from users

Best For:

- Rapid prototyping and experimentation
- Users without technical expertise
- Short sequences (2-3 panels)
- One-off character generation

4.1.2 Character Embeddings (Stable Diffusion)

Strengths:

- **Superior Consistency**: Can achieve higher consistency through learned representations
- **Reusability**: Once created, embeddings can be reused across multiple generations
- **Fine-Grained Control**: Can capture subtle character details
- **Community Support**: Large ecosystem of tools and resources

Limitations:

- **Setup Complexity**: Requires training or fine-tuning process
- **Data Requirements**: Needs multiple reference images of the character
- **Computational Cost**: Training requires GPU resources
- **Technical Expertise**: Requires understanding of model training
- **Storage Overhead**: Character embeddings must be stored and managed

Best For:

- Professional content creation
- Recurring characters in long-form content
- Users with technical expertise and resources
- High-consistency requirements

4.1.3 Fine-Tuning Approaches

Strengths:

- **Highest Quality**: Can achieve the best consistency and quality
- **Style Control**: Can learn specific art styles along with characters
- **Optimization**: Models can be optimized for specific use cases

Limitations:

- **High Cost**: Requires significant computational resources
- **Data Requirements**: Needs extensive training datasets
- **Time Investment**: Training can take hours or days
- **Model Management**: Each character may require a separate model
- **Overfitting Risk**: May lose generalization capabilities

Best For:

- Production environments with dedicated resources
- Established characters with extensive reference material
- Long-term projects with recurring characters

4.1.4 Hybrid Approaches

Some systems combine multiple techniques:

- Prompt engineering + image-to-image translation
- Character embeddings + ControlNet for pose consistency
- Fine-tuned models + prompt-based style control

These approaches can achieve better results but at the cost of increased complexity and system requirements.

4.2 Evaluation Metrics

When comparing approaches, several metrics should be considered:

4.2.1 Visual Consistency

The degree to which characters maintain visual identity across panels. This can be measured through:

- Facial feature similarity
- Clothing and accessory consistency
- Body proportions and pose coherence
- Color palette consistency

4.2.2 Character Recognizability

The ability of viewers to identify the same character across different panels. This is subjective but can be assessed through user studies.

4.2.3 Computational Cost

- **API Costs**: Per-generation costs for API-based services
- **Training Costs**: One-time costs for embeddings or fine-tuning
- **Storage Costs**: Model and embedding storage requirements
- **Time Costs**: Generation and training time

4.2.4 Usability

- Setup complexity
- Technical expertise required
- Time to first result
- Maintenance overhead

4.3 Trade-off Analysis

The choice of approach involves fundamental trade-offs:

1. **Consistency vs. Simplicity**: More consistent methods (embeddings, fine-tuning) require more setup and expertise.
2. **Quality vs. Cost**: Higher quality approaches (fine-tuning) have higher computational and financial costs.
3. **Flexibility vs. Consistency**: Flexible approaches (prompt engineering) are easier to modify but less consistent.
4. **Scalability vs. Setup**: Approaches that scale well (embeddings) require more initial setup.

4.4 Recommendations

Based on our analysis, we recommend:

- **For Beginners**: Start with prompt engineering (MangaBanana approach) to understand the problem space and requirements.
- **For Regular Users**: Invest in character embeddings (Stable Diffusion) for recurring characters, combining with prompt engineering for flexibility.
- **For Professionals**: Use fine-tuned models for production work, with embeddings as a fallback for new characters.
- **For Research**: Explore hybrid approaches that combine multiple techniques for optimal results.

4.5 Future Directions

Several promising directions could improve character consistency:

1. **Multi-Modal Memory**: Models that can reference previous panel outputs during generation.
2. **Improved Prompt Understanding**: Better interpretation of consistency instructions in prompts.
3. **Hybrid Architectures**: Systems that combine prompt engineering with visual references.
4. **Specialized Models**: Models trained specifically for sequential art generation.
5. **Real-Time Consistency Checking**: Systems that validate consistency during generation.

5 Results and Discussion

5.1 MangaBanana System Performance

The MangaBanana system has been deployed and tested in a production environment, generating hundreds of two-panel manga comics. While formal quantitative evaluation is challenging due to the subjective nature of character consistency, we present qualitative observations and user feedback.

5.2 Qualitative Observations

5.2.1 Consistency Achievements

The prompt-based approach demonstrates moderate success in maintaining character consistency:

- **Character Features**: When detailed character descriptions are provided, facial features, hair color, and distinctive characteristics are often maintained across panels.
- **Style Consistency**: The explicit style instructions help maintain consistent art style and color palette.
- **Clothing Consistency**: Character clothing and accessories are generally consistent when explicitly described.

5.2.2 Observed Limitations

Several limitations have been observed in practice:

- **Facial Variations**: Subtle variations in facial features, expressions, and angles can occur despite detailed descriptions.
- **Pose-Dependent Appearance**: Character appearance can vary based on pose and viewing angle, even with identical descriptions.
- **Background Interference**: Complex backgrounds can sometimes affect character rendering consistency.
- **Style Drift**: Minor variations in rendering style, line weight, and shading can occur between panels.

5.3 User Feedback

Informal user feedback suggests:

- Users appreciate the simplicity and accessibility of the prompt-based approach.
- Character consistency is "good enough" for many use cases, particularly for short sequences.
- Users with detailed character descriptions achieve better results than those with minimal descriptions.
- The system works well for distinctive characters but struggles with generic or similar-looking characters.

5.4 Case Studies

5.4.1 Case 1: Distinctive Character

A user generated a comic featuring "a samurai with long black hair, wearing blue hakama, holding a katana with a distinctive red handle." The character description was highly detailed and distinctive. Result: High consistency across panels, with the character clearly recognizable in both panels.

5.4.2 Case 2: Generic Character

A user generated a comic with a minimal description: "a young person in modern clothes." Result: Lower consistency, with noticeable variations in appearance, clothing, and style between panels.

5.4.3 Case 3: Complex Scene

A user generated a comic with a detailed character but complex, dynamic scenes. Result: Character consistency was maintained in core features, but background and pose variations affected overall visual coherence.

5.5 Limitations and Challenges

5.5.1 Technical Limitations

- **No Visual Memory**: DALL-E 3 cannot directly reference Panel 1 when generating Panel 2, relying entirely on textual descriptions.
- **Stochastic Nature**: The inherent randomness in generation can produce variations even with identical prompts.
- **Prompt Length Constraints**: Very detailed character descriptions may approach token limits or become less effective.
- **API Rate Limits**: Sequential generation can be slow and subject to API rate limiting.

5.5.2 Methodological Limitations

- **Subjective Evaluation**: Character consistency is inherently subjective, making quantitative evaluation difficult.
- **Limited Dataset**: Evaluation is based on production usage rather than a controlled dataset.
- **No Baseline Comparison**: Direct comparison with other approaches would require equivalent implementations.

5.6 Future Improvements

Based on our observations, several improvements could enhance the MangaBanana system:

1. **Character Reference Images**: Allow users to upload reference images that could be used in prompt construction or as visual guides.
2. **Improved Prompt Templates**: Develop more sophisticated prompt templates that better encode consistency instructions.
3. **Multi-Pass Generation**: Implement a refinement step where Panel 2 generation can be retried with adjusted prompts based on Panel 1 analysis.
4. **Consistency Scoring**: Develop automated metrics to score consistency and provide feedback to users.
5. **Hybrid Approach**: Integrate with Stable Diffusion for users who want to create character embeddings.
6. **Extended Sequences**: Extend support for longer sequences (3+ panels) with improved consistency mechanisms.

5.7 Implications for Practice

Our results suggest that:

- Prompt-based approaches are viable for short sequences (2-3 panels) with detailed character descriptions.
- Users should invest time in creating comprehensive character descriptions for best results.
- The approach is most suitable for prototyping, experimentation, and casual use.
- For production-quality work or longer sequences, more advanced approaches (embeddings, fine-tuning) may be necessary.

5.8 Research Implications

This work highlights several important research directions:

- The need for better evaluation metrics for character consistency in sequential art.
- The potential for hybrid approaches that combine prompt engineering with visual references.
- The importance of user experience design in making consistency mechanisms accessible.
- The value of comparative studies across different model families and approaches.

6 Conclusion

This paper has presented a comprehensive examination of character consistency in AI-generated manga, surveying state-of-the-art models and approaches, and analyzing the practical implementation of the MangaBanana system.

6.1 Key Findings

Our analysis reveals several important insights:

1. **No One-Size-Fits-All Solution**: Different approaches to character consistency serve different use cases. Prompt engineering offers simplicity and accessibility, while embeddings and fine-tuning provide superior consistency at the cost of complexity.
2. **Prompt Engineering is Viable**: For short sequences (2-3 panels) with detailed character descriptions, prompt-based approaches can achieve acceptable consistency without requiring training or specialized setup.
3. **Trade-offs are Fundamental**: The choice between approaches involves fundamental trade-offs between consistency, simplicity, cost, and scalability. Practitioners must carefully consider their specific requirements.
4. **User Input Matters**: The quality and detail of character descriptions significantly impact consistency, regardless of the approach used.
5. **Current Limitations**: Even the best current approaches cannot guarantee perfect character consistency, particularly for longer sequences or complex scenes.

6.2 Contributions Summary

This work contributes to the field by:

- Providing an up-to-date survey of character consistency approaches across major AI image generation models
- Presenting a practical implementation (MangaBanana) that demonstrates prompt-based consistency
- Offering comparative analysis that helps practitioners choose appropriate approaches
- Identifying gaps and opportunities for future research

6.3 Future Work

Several promising directions for future research emerge:

1. **Multi-Modal Consistency Mechanisms**: Develop models that can reference previous panel outputs during generation, combining textual and visual information.
2. **Specialized Sequential Art Models**: Train models specifically for manga and comic generation, with built-in consistency mechanisms.
3. **Automated Consistency Evaluation**: Develop quantitative metrics and tools for evaluating character consistency objectively.
4. **Hybrid Architectures**: Explore systems that intelligently combine prompt engineering, embeddings, and visual references based on context.
5. **User Experience Research**: Investigate how to make consistency mechanisms more accessible and intuitive for non-technical users.
6. **Extended Sequence Support**: Develop approaches that maintain consistency across longer sequences (10+ panels) for full-page or multi-page comics.

6.4 Final Remarks

Character consistency in AI-generated manga remains an open challenge, but significant progress has been made. As models continue to improve and new techniques emerge, we expect to see increasingly sophisticated solutions that balance consistency, quality, and usability. The MangaBanana system demonstrates that practical, accessible solutions are possible today, while more advanced approaches offer paths forward for professional applications.

The field is rapidly evolving, and we anticipate that future models will incorporate consistency mechanisms more directly, reducing the need for workarounds and specialized techniques. Until then, understanding the trade-offs between different approaches enables practitioners to make informed decisions based on their specific needs and constraints.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Janus Wang, Long Li, Lilian Ouyang, Juntang Zhuang, Joyce Lee, Yossi Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2023.

Tim Black, Mostafa Jahanian, and Trevor Darrell. Flux: A family of open source foundational models. *arXiv preprint arXiv:2407.06951*, 2024.

Midjourney. Midjourney documentation, 2024. URL `https://docs.midjourney.com`. Accessed: 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, et al. Imagen: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.