Addressing Character Consistency Challenges in AI Filmmaking

Marcin Kalinski[*1], Michal Podstawski[*1], Krzysztof Ostrowski[1], Malgorzata Kudelska[1], Patryk Bartkowiak[1], Piotr Lyczko[1], Michal Piasecki[1], Thomas Visscher[2], Haohong Wang[2]
[1]*TCL Research Europe, Warsaw, Poland*
[2]*TCL Research America, San Jose, USA*
E-mails: {name.surname}@tcl.com

***Abstract*—Generative AI is revolutionizing filmmaking workflows, but achieving visual consistency across characters remains a significant challenge. This paper explores the practical application of state-of-the-art techniques to address this challenge, leveraging insights from extensive experiments and real-world use cases in film productions. Key difficulties include preserving character identities across varied poses and expressions (spanning both realistic and animated styles). By sharing lessons learned, best practices, and practical insights, this paper emphasizes the transformative potential of generative AI techniques in crafting cohesive cinematic experiences and engaging product-centric storytelling, setting the stage for future advancements in AI-driven filmmaking.**

***Index Terms*—Generative AI, AI filmmaking, Low-Rank Adaptation (LoRA), visual consistency, character consistency**

I. Introduction

Maintaining consistency in character representation is one of the most significant challenges in AI-generated content, particularly in filmmaking. Generative models such as Mid-Journey [1] and Stable Diffusion (based on [2]) process each image independently, without an understanding of characters as cohesive entities. This often results in discrepancies in a character's appearance, clothing, or expressions across frames, disrupting the narrative flow. The inherent randomness of these generative processes further exacerbates such inconsistencies, making character coherence a critical barrier to the practical adoption of AI-generated films.

In traditional filmmaking, visual consistency is guaranteed by physical or geometric constraints [3]–[5]. Optical projections naturally preserve the physical form of actors across scenes. Stop-motion animation achieves the same through the constancy of puppets. Animated films rely on conceptual representations: in 2D animation, detailed reference sketches define how characters should appear across angles and poses, ensuring uniformity. In 3D animation, this consistency is extended through geometric models (meshes) enriched with textures, enabling precise representation of characters in spatial and dynamic contexts. These traditional methods inherently achieve consistency through physical presence or well-defined visual concepts, which are absent in generative AI models.

*Equal contribution. Authors are listed in alphabetical order.

To address these challenges, we propose a robust framework that leverages the synergy of state-of-the-art generative models and tools. This framework introduces a structured approach covering the entire production pipeline, including data collection, model fine-tuning, and post-processing. By combining curated datasets with targeted adjustments to model configurations, our solution enables consistent character representation across scenes. Its effectiveness has been validated through real-world applications, demonstrating superior quality and narrative coherence compared to existing approaches. This framework provides a significant step forward in making generative AI a viable tool for filmmaking and other creative domains.

II. Related work

A. Diffusion models

Diffusion models have emerged as a powerful class of generative models [2], [6], [9], excelling in high-quality image, audio, and multimodal generation. They model data distribution $p(x)$ by learning to reverse a stochastic forward process that corrupts data into noise. The forward process adds Gaussian noise over T timesteps:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where α_t determines the noise schedule. The reverse process, parameterized by a neural network $\epsilon_\theta(x_t, t)$, denoises step-by-step:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{2}$$

Stable Diffusion (SD): Stable Diffusion [2], including Stable Diffuson XL (SDXL) [7], extends diffusion models by operating in a latent space instead of pixel space, significantly reducing computational requirements. A pre-trained Variational Autoencoder (VAE) encodes the input image into a compressed latent representation z, where the forward diffusion process is defined as:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t}z_{t-1}, (1-\alpha_t)\mathbf{I}). \tag{3}$$

The reverse process generates latent codes, which are decoded back into the image space via the VAE decoder. Stable Diffusion excels in generating high-quality images

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

with reduced computational overhead compared to pixel-space diffusion models.

Flux: Flux [10] builds on prior work, such as transformer-based architectures optimized for diffusion tasks [8], [9], but emphasizes computational efficiency and applicability to high-quality outputs under constrained resources. It enhances training stability and convergence by introducing a gradient-guided denoising objective:

$$L_{\text{flux}} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \nabla_\theta \epsilon_\theta(x_t, t)\|^2\right]. \quad (4)$$

This approach improves sample fidelity and reduces the number of timesteps required for generation.

MidJourney and Closed Diffusion Models: MidJourney [1], DALL·E [11], and RunwayML [12] are proprietary diffusion models designed for user-friendly generation of stylized and artistic images. However, their closed nature limits transparency and prevents fine-tuning, making them less suitable for tasks requiring domain-specific adaptations, such as consistent character generation.

In contrast, open models like Stable Diffusion and Flux offer greater flexibility and customization through techniques like LoRA [13], enabling more precise control and adaptation.

B. Low-Rank Adaptation (LoRA)

Diffusion models have revolutionized image synthesis, producing high-quality outputs, but struggle with consistent character representations across images. Efficient fine-tuning for specific tasks is challenging, as full model fine-tuning demands prohibitive computational resources, limiting its practicality for real-world applications.

LoRA addresses the challenge of fine-tuning large diffusion models by approximating weight updates with low-rank matrices:

$$\Delta W = AB, A \in \mathbb{R}^{d\times r}, B \in \mathbb{R}^{r\times k}, r \ll \min(d,k). \quad (5)$$

In diffusion models, LoRA is applied to attention and feed-forward layers, where query and key projections are updated as:

$$Q = Q_0 + A_Q B_Q, \quad K = K_0 + A_K B_K, \quad (6)$$

allowing efficient fine-tuning while freezing pre-trained weights. LoRA enables consistent character representation across images by associating unique tokens with specific character features, which is crucial for creative workflows like filmmaking.

There are many well-adopted tools that enable LoRA training, both for SD and for Flux. It includes Kohya [14] and SimpleTuner [15].

C. Video generation and limitations of existing approaches

In filmmaking, the final output must be a video, yet achieving this with generative AI introduces unique challenges. While video models like VideoGPT [16], Imagen Video [17], and Phenaki [18] introduce mechanisms for spatio-temporal alignment, they often struggle with maintaining character consistency across longer sequences or between different videos. This limitation is particularly problematic in filmmaking, where character fidelity across diverse scenes and narrative coherence are essential.

To address these challenges, current generative video workflows leverage a hybrid approach. Image-based diffusion models, such as Stable Diffusion and Flux, are employed to produce keyframes, which serve as anchors for the sequence. To bridge the gaps between these keyframes and ensure temporal coherence, tools like Runway [12] and Kling [19] are used to interpolate intermediate frames. This method effectively combines the precision of image-based diffusion for maintaining visual consistency with the temporal alignment capabilities of video synthesis tools, mitigating key limitations of existing generative video models.

By prioritizing character consistency in image generation, techniques like LoRA fine-tuning establish a solid foundation for maintaining fidelity in keyframes, enabling visually cohesive and temporally aligned video narratives for filmmaking.

III. Proposed solution

Building towards the described hybrid approach, we propose a framework that enables generating consistent characters across multiple images. It leverages well-adopted state-of-the-art techniques, such as diffusion model tuning and LoRA training. The achieved high level of consistency enables the usage of the approach in AIGC filmmaking, which has been proven by real-world adaptation.

The main contributions of the project include:

1) The framework that enables overcoming the main challenges related to generating consistent characters.
2) A set of well-defined guidelines for building both realistic and synthetic datasets that provide the best overall quality.
3) A detailed multi-step procedure with iterative refinement for generating synthetic datasets.
4) Ready-to-use tools for inpainting that improve the final outcome of the procedure.

A. Framework design

The proposed framework introduces a systematic set of stages to facilitate the generation process. The first stage involves constructing an appropriate dataset. The nature of this stage varies depending on whether the dataset is derived from real-world data, such as images of an actual person, or if it involves creating a fictional character from scratch. In the latter case, we propose a meticulously designed procedure to ensure the generated dataset aligns with the desired characteristics and consistency requirements. This includes curating diverse representations of the character, ensuring variations in poses, expressions, and contexts to enhance the robustness of the model during training.

The next stage focuses on fine-tuning LoRA models using the curated dataset. During this phase, the LoRA methodology is applied to adapt large diffusion models, such as Stable Diffusion, for task-specific generation. This stage outputs the initial set of generated images, which capture the essence of

the character based on the training data. Special attention is given to the consistency of generated features, ensuring uniformity across different poses or scenarios.

However, when working with fictional characters, especially where the training dataset is synthesized using tools like MidJourney, the consistency of the character's face in the input dataset is often not very high, particularly in full-body shots where the face appears small. This can result in the initially trained LoRA struggling to generate consistent and accurate representations of the character in such cases.

To overcome these limitations, the framework incorporates a post-processing inpainting phase. In this phase, the pre-trained LoRA model is reapplied to refine and enhance the quality of the initial outputs. This post-processing step improves visual coherence, corrects artifacts, and ensures that the generated images meet high-quality standards for the intended application. For synthetic datasets, this process extends further by establishing a feedback loop. Specifically, the outputs refined through inpainting are reintroduced into the synthetic dataset, enabling iterative improvements in training. By using the inpainted faces to improve the training dataset, subsequent LoRA models are trained on more consistent and higher-quality data, leading to better results, particularly in challenging scenarios like full-body shots.

This loop allows the synthetic dataset to progressively enhance its consistency, fidelity, and adaptability, ultimately leading to superior results in subsequent iterations. By leveraging this targeted refinement cycle for synthetic datasets, the framework achieves a level of quality and uniformity that surpasses traditional methods while maintaining robust performance across both realistic and synthetic applications.

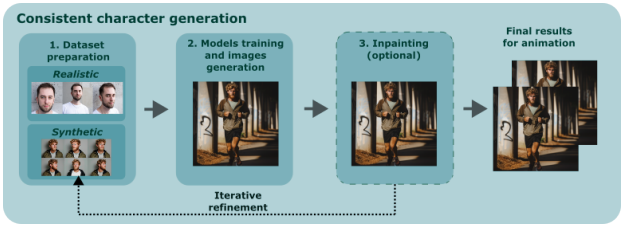

Fig. 1. Main stages of the process.

B. Datasets preparation

The preparation of datasets is a vital part of the process. Well-structured datasets serve as the foundation for model training, enabling fine-tuning to achieve fidelity, coherence, and adaptability across various applications. For tasks like AI filmmaking, where maintaining visual consistency is paramount, datasets must be carefully designed to reflect real-world or synthetic scenarios. This involves capturing or generating data with diverse poses, lighting conditions, expressions, and contexts while adhering to standardized practices for robust model training. The goal is to provide the variability needed for generalization while ensuring uniformity in critical attributes to achieve reliable and consistent outputs.

1) Realistic Datasets: Realistic dataset photographs require controlled conditions to ensure high fidelity and uniformity, capturing diverse angles, lighting, and emotional expressions.

To ensure robust training the following guidelines for dataset preparation are proposed:

1) **Image Specifications:** Use square images with a minimum resolution of 1024 × 1024 pixels for compatibility with diffusion-based models.
2) **Close-Up Images:** Capture images covering:
   - *Head Positions:* Horizontal (left profile, frontal, right profile) and vertical (upward, downward).
   - *Emotional Expressions:* Joy, sadness, anger, surprise, and fear.
3) **Portraits:**
   - *Close-Up Portraits:* Shoulders visible.
   - *Medium Portraits:* Waist visible.
   - *Full-Body Portraits:* Entire figure visible.
4) **Environmental Diversity:** Acquire data under diverse conditions:
   - Controlled studio lighting.
   - Indoor settings with diffused or harsh artificial light.
   - Outdoor environments, including urban and natural settings.

These guidelines provide a robust foundation for training LoRAs models, ensuring high-quality and consistent outputs across varied applications.

Fig. 2. Example of LoRA results. Left: real photo from the training dataset. Rest: images generated using SDXL with LoRA.

2) Synthetic Datasets: The synthetic dataset creation process is entirely AI-driven and follows an iterative, hierarchical structure. Each layer builds upon the outputs of the previous layer, refining and expanding the dataset to ensure consistent and high-quality results. The following guidelines outline the step-by-step procedure for generating a comprehensive synthetic dataset:

1) **Initial Image Generation:**
   - Using tools like MidJourney or Stable Diffusion (enhanced with tools such as IP-Adapter [22] or PuLID [23]) create a composite image with 6-8 portraits of one character presented from different angles. Such an approach ensures that the character remains consistent across the sub-images and provides a solid foundation for creating further coherent images.
   - Cut the composite image into individual portraits.
2) **Close-Up Layer:**

- Using an image-to-image pipeline to generate new close-up portraits.
- Introduce variations in:
- *Images with head rotations*: frontal, left, and right profiles.
- *Images with different expressions*: joy, sadness, anger, fear, surprise.
- Maintain consistency in character features while diversifying expressions and angles.

3) **Close-Up Portrait Layer :**
   - Expand close-ups to include shoulders and chest, ensuring continuity in facial features.
   - Add subtle variations like head tilt, lighting, and background scenarios for a broader context.
4) **Portrait Layer:**
   - Extend to upper-body portraits, including the waist.
   - Focus on proportionality, coherence, and introduce diverse postures and gestures to enhance adaptability.
5) **3/4 Shot Layer :**
   - Generate 3/4 shots using the previous layer's outputs, incorporating the legs.
   - Ensure consistency across facial and body features while introducing diverse poses like left-facing, frontal, and right-facing
6) **Full-Body Layer:**
   - Use 3/4 shots to create full-body views.
   - Maintain coherence in facial features, body proportions, and appearance. Optionally, add specific backgrounds or environmental details for tailored applications.
7) **Iterative Refinement:**
   - Build each layer upon the last, progressively adding detail and complexity.
   - At each stage, review the outputs to verify alignment with the intended design, refining where necessary to preserve coherence and fidelity.

This iterative, hierarchical process, illustrated in Fig. 3, ensures that the dataset achieves high fidelity and consistency across all levels of detail. By using each layer as input for the next, the methodology allows for seamless refinement, making the resulting dataset highly suitable for LoRA training.

3) Animated Characters: The method for generating datasets of animated characters closely resembles the process for synthetic human figures, following the same iterative and hierarchical structure. Initial AI-generated images of the animated character form the foundation, and each subsequent layer refines and expands the dataset, progressing from close-up facial details to full-body representations. However, a key difference lies in the stylization and exaggerated features inherent to animated characters, which require careful control to preserve their distinct visual traits, such as oversized eyes, exaggerated expressions, or unique proportions.

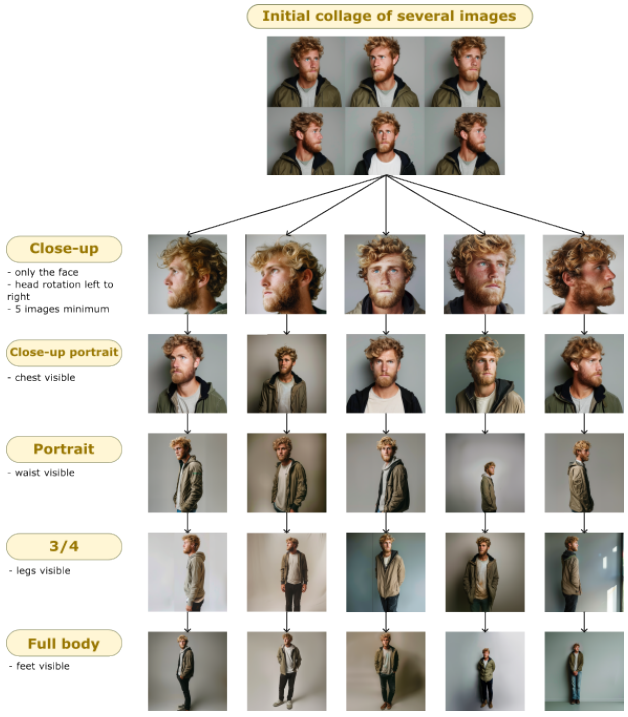

Fig. 3. The hierarchical process of synthetic dataset creation starting with AI-generated images, progressing from close-up facial shots to full-body representations, with each layer refining the previous outputs for consistency and coherence.

A major challenge in this process is ensuring consistency of artistic style across layers, especially when generating new poses or perspectives that deviate significantly from the initial inputs. Furthermore, animated characters often involve non-photorealistic elements (e.g., flat shading, cartoon-like proportions) that are less constrained by real-world physics, requiring AI models to adapt to stylistic nuances rather than replicating realism. Despite these challenges, the iterative refinement process remains effective for maintaining coherence, ensuring the character's unique features and style are preserved across all outputs.

C. LoRA models training

With the structured datasets prepared, the next step involves fine-tuning pre-trained diffusion models using LoRA. The fine-tuning process ensures that the model learns task-specific features, such as consistent traits, while maintaining computational efficiency. For this purpose, different tools and diffusion models are employed based on the target domain, with SimpleTuner serving as a framework for LoRA fine-tuning [15].

Extensive experimentation has shown that the choice of model significantly impacts the quality of generated outputs:

- **SDXL for realistic characters:** SDXL consistently outperforms Flux for generating realistic characters. Its strength lies in capturing intricate details such as facial features, hair textures, and expressive nuances, ensuring

superior results for creating consistent and lifelike characters across diverse poses and lighting conditions.

- **Flux for animated characters:** While Flux may not achieve the same level of realism in facial features as SDXL, it is well-suited for generating anime-style or less texturally detailed characters. Flux's advantage lies in its compositing capabilities, making it easier to generate complete scenes where characters are seamlessly integrated into appropriately styled backgrounds. This makes Flux a strong choice for stylized animations where fine texture details are less critical, but cohesive scene composition is essential.

The choice of configurations for fine-tuning LoRA models is vital in order to achieve best quality:

a) SDXL: The fine-tuning configuration for LoRA models on SDXL is designed to enhance character consistency, detail, and generalization. Training begins with a high-quality base model (we use DreamShaper XL v2.1 Turbo DPM++ SDE [20]). Key parameters include a resolution of 1024×1024, which captures intricate features like facial details and textures, and a learning rate of 1×10^{-6} for both the U-Net and text encoder, ensuring stable convergence. The network dimensions used is 64. The best results are typically achieved between 400 and 600 training epochs.

To prevent overfitting and maintain generalization, a custom regularization dataset is created. A custom GPT-4 prompt [21] is used to generate dense captions from the training dataset. These captions are modular, explicitly separating descriptions of the character's physical appearance from details about the background. Replacing character-specific attributes with generic phrases like "photo of a man" allows for the generation of a diverse regularization dataset. This regularization strategy not only enhances the consistency of character representations but also preserves the model's ability to generate high-quality, detailed backgrounds.

Masking techniques are applied during training to focus the model on the character while preventing it from overfitting to backgrounds. This ensures the final model maintains its ability to generate high-quality backgrounds while improving character-specific features.

b) Flux: Process starts from one of Flux family models (we use Flux.1 Schnell). According to our experiments, training at a resolution of 1024×1024 is effective for preserving details essential for achieving high-quality outputs in animated characters generation tasks. Using a polynomial learning rate schedule with an initial learning rate of 5×10^{-4} and 100 warmup steps provides stable convergence while avoiding early instability. The network dimensions used in the training are set between 16 and 32, providing a balance between model capacity and computational efficiency. Optimal results are typically achieved between 1400 and 1600 training steps.

D. Inpainting

Character images generated with trained LoRA models generally maintain consistency, especially for close-ups and facial details. However, as the framing expands to full-body or 3/4 shots, smaller details, such as facial features, often lose accuracy due to the increased spatial complexity.

To address this, we employ inpainting within an autoreferential feedback loop. Images with low similarity from the training datasets are visually assessed, isolated, and refined using LoRA-based inpainting.

The process begins by masking the face, hair, beard, and ears in the input image. A square region containing the mask is cropped and upscaled to 1024×1024 resolution using the Lanczos algorithm, preserving fine facial details. The mask is also upscaled to match the image dimensions.

The cropped image and mask are converted to latent space and processed using a trained model with custom fine-tuned parameters. The latent output is passed through a second sampler with a lower denoise value (e.g., 0.40) without the mask, improving coherence between the newly generated facial features and the surrounding background.

Finally, the output is composited into the cropped input image using the original mask, resized to its initial dimensions, and integrated back into the input frame. This workflow ensures seamless integration of high-quality inpainted facial details, as shown in Fig.4.

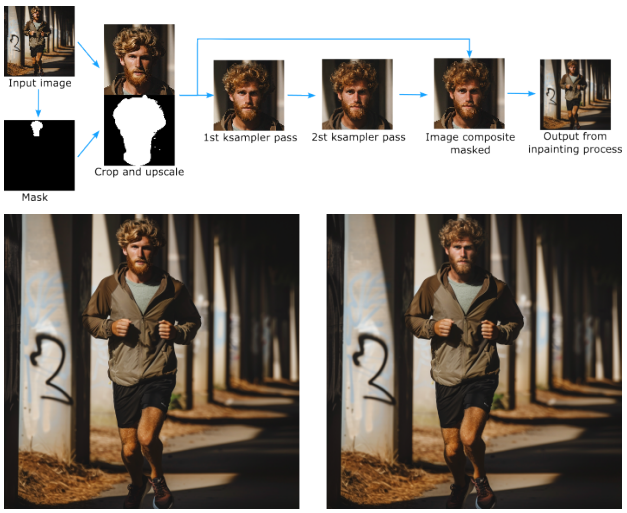

Fig. 4. Inpainting process and example results for 2-step SDXL LoRA. Left: without inpainting. Right: with inpainting

The final dataset incorporates multi-layered generation, supplemented with inpainted faces, to enhance overall consistency. The improved, inpainted results from this process can serve as a refined source for a second LoRA training pass. By using the enhanced dataset for consecutive training, this 2-step approach further improves the model's ability to maintain consistency across different frames, including challenging full-body and three-quarter compositions.

E. Experiments

To validate the proposed framework, we conducted a comprehensive evaluation by following all outlined steps. The process began with generating initial images using MidJourney, which served as the foundation for creating a synthetic dataset

of 135 images through iterative refinement, as detailed in the methodology. Subsequently, two separate LoRA models were trained. The first model was trained on the dataset produced after the first generation step. The second model was trained on an enhanced dataset that underwent an additional feedback loop pass, incorporating inpainting to improve facial details and consistency.

To evaluate the performance of the trained models, we generated four distinct test datasets (with the same starting seed value), each comprising 50 images of full-body and 3/4 body shots. These datasets were created using the DreamShaper XL v2.1 Turbo model. Prompts for generating the images were randomly written by GPT-4 in order to avoid overlap with the training datasets and ensure unbiased evaluation. The datasets were as follows:

1) *1-step LoRA:* Generated using the LoRA trained on 1-step input data without inpainting.
2) *1-step LoRA + inpainting:* Generated using the LoRA trained on 1-step input data with inpainting.
3) *2-step LoRA:* Generated using the LoRA trained on 2-step input data without inpainting.
4) *2-step LoRA + inpainting:* trained on 2-step input data with inpainting.

The consistency of the generated images was assessed by comparing them to three reference portraits from the initial training dataset. Facial similarity was objectively measured by computing the cosine similarity between the facial embeddings of the generated images and the reference portraits from the starting set. For this evaluation, state-of-the-art face recognition models were employed, including Facenet [24], OpenFace [25], DeepID [26], and the DeepFace tool [27], [28]. The improvement between the 1-step LoRA and the 2-step LoRA with inpainting ranges from 0.0996 for Facenet to 0.1868 for OpenFace, demonstrating a significant enhancement. The detailed results are presented in Fig. 5.

Additionally, we include a representative example of image generation for animated characters, derived from a real-world production use case, showcasing results both without and with inpainting. The results demonstrate a substantial improvement in quality metrics when inpainting is utilized, with similarity gains up to 0.1212 for OpenFace. This enhancement can be observed in Fig. 6, which highlights the visual differences between the two approaches. Furthermore, we present Facenet's embedding heatmaps in Fig. 7, offering a deeper analysis of the improvements achieved through inpainting.

The experiments demonstrated that the proposed framework performs effectively, with each consecutive step contributing significant improvements to the results. The integration of inpainting within the feedback loop notably enhanced facial detail and consistency, particularly in full-body and 3/4 body shots, where deviations were most pronounced. By iteratively refining the dataset and leveraging multi-step input data, the final outputs achieved a high level of quality, preserving character identity across diverse poses and expressions. The combination of well-curated datasets, advanced LoRA training methodologies, and systematic post-processing resulted in visually consistent and realistic images, validating the efficacy of the framework for synthetic dataset generation and character preservation tasks.

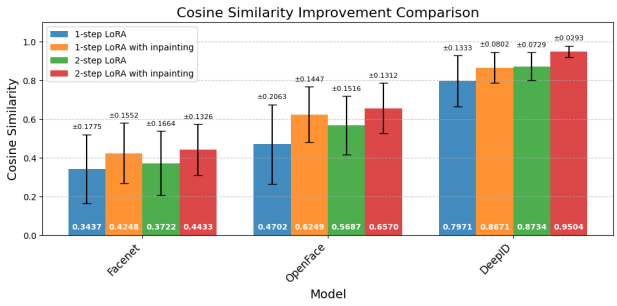

Fig. 5. Mean cosine similarity between generated datasets and the reference images. The plots demonstrate the impact of consecutive steps in the proposed framework, showing consistent improvements in character similarity with the addition of inpainting and multi-step input refinement. Each dataset was evaluated using multiple face recognition models, providing a comprehensive assessment of identity preservation.

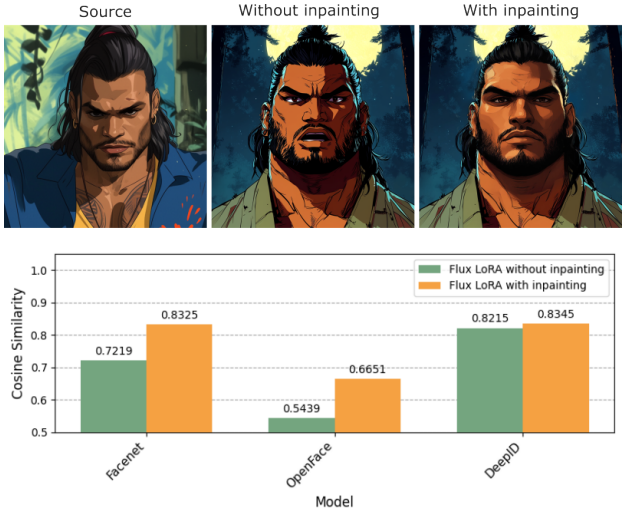

Fig. 6. Cosine similarity between the reference source image and images generated without and with inpainting, demonstrating the impact of inpainting on improving image similarity for animated characters.

IV. Next Steps

While significant progress has been made, further research is needed to overcome existing limitations. Two promising directions leverage advancements in Low-Rank Adaptation (LoRA) and video generation technologies.

The first integrates LoRA with video models, extending its parameter-efficient fine-tuning to both spatial and temporal components. This approach enables consistent character appearances and smooth transitions across sequences, improving temporal coherence efficiently.

The second explores applying multiple LoRAs within a single generation process. By targeting regions like characters, objects, and backgrounds separately, finer control over visual

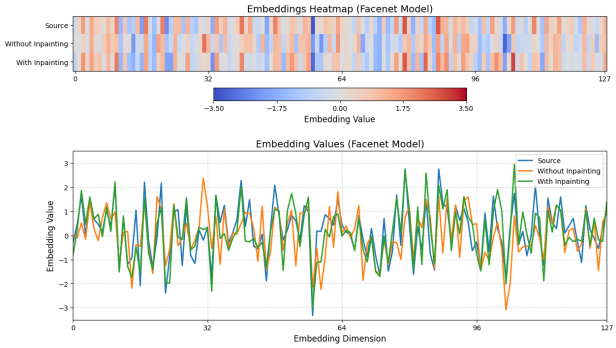

Fig. 7. Detailed analysis of the embeddings of animated characters' faces for the source image and images generated without inpainting and with inpainting, as processed by the Facenet model. The heatmap illustrates the distribution of embedding values across dimensions, while the line plot provides a comparative view of the embeddings. For the Facenet model, the application of inpainting improves the cosine similarity, with an observed increase in the range of 0.05 to 0.20, highlighting its effectiveness in enhancing character representation consistency.

elements can be achieved. Extending this to video generation ensures region-specific adaptations, though challenges like coordination and seamless integration remain.

These directions offer opportunities to enhance consistency, control, and quality in AI-driven video production for filmmaking.

V. Conclusions

This paper addresses one of the most pressing challenges in AI filmmaking: maintaining consistency in character representations across frames and sequences. By leveraging state-of-the-art generative models and Low-Rank Adaptation (LoRA) techniques, we propose a comprehensive framework that spans data preparation, model fine-tuning, and post-processing to achieve high-quality and coherent outputs.

Our work demonstrates that LoRA, when applied effectively, can significantly enhance the consistency and fidelity of AI-generated content. By integrating advanced techniques such as autoreferential feedback loops, we further improved the robustness and precision of generated outputs.

The proposed methodology is validated in real-world applications, particularly in cinematic contexts where character consistency and stylistic coherence are paramount. Our results highlight how structured datasets, iterative refinement, and targeted LoRA fine-tuning can produce outputs that rival traditional methods in quality and coherence.

Acknowledgment

We acknowledge the use of ChatGPT, a tool based on the GPT language model developed by OpenAI, as a tool for improving the language and clarity of this manuscript. ChatGPT was utilized to enhance the writing style, refine sentence structure, and ensure linguistic precision throughout the text.

References

[1] Midjourney, 2022. https://www.midjourney.com
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022 pp. 10674-10685, 2022.
[3] J. Van Sijll, "Cinematic Storytelling: The 100 Most Powerful Film Conventions Every Filmmaker Must Know," Michael Wiese Productions, Studio City, CA, USA, 2005.
[4] L. Braudy and M. Cohen, Eds., "Film Theory and Criticism: Introductory Readings," 7th ed., Oxford University Press, New York, NY, USA, 2009.
[5] B. Block, "The Visual Story: Creating the Visual Structure of Film, TV, and Digital Media," 2nd ed., Focal Press, Burlington, MA, USA, 2007.
[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, 2020.
[7] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna and R. Rombach, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," *arXiv preprint arXiv:2307.01952*, 2023.
[8] W. Peebles, S. Xie, "Scalable Diffusion Models with Transformers," *in Proc. ICCV 2023, pp. 4195-4205*, 2023.
[9] P. Esser et al., "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis," *arXiv preprint arXiv:2403.03206*, 2024.
[10] FLUX https://github.com/black-forest-labs/flux
[11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.
[12] RunwayML, https://runwayml.com
[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.
[14] kohya-ss, "sd-scripts: Training Scripts for Stable Diffusion," GitHub repository, Available: https://github.com/kohya-ss/sd-scripts.
[15] bghira, "SimpleTuner: A General Fine-Tuning Kit Geared Toward Diffusion Models," GitHub repository, Available: https://github.com/bghira/SimpleTuner.
[16] W. Yan, Y. Zhang, P. Abbeel, A. Srinivas, "VideoGPT: Video Generation using VQ-VAE and Transformers," *arXiv preprint arXiv:2104.10157*, 2021.
[17] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, T. Salimans, "Imagen Video: High Definition Video Generation with Diffusion Models," *arXiv preprint arXiv:2210.02303*, 2022.
[18] R. Villegas, M. Babaeizadeh, P. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, D. Erhan, "Phenaki: Variable Length Video Generation From Open Domain Textual Description," *arXiv preprint arXiv:2210.02399*, 2022.
[19] Kling, https://kling.kuaishou.com/en
[20] DreamShaper XL, https://huggingface.co/Lykon/dreamshaper-xl-turbo
[21] ChatGPT, https://chatgpt.com
[22] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models," *arXiv preprint arXiv:2308.06721*, 2023.
[23] Z. Guo, Y. Wu, Z. Chen, L. Chen, P. Zhang, and Q. He, "PuLID: Pure and Lightning ID Customization via Contrastive Alignment," *Advances in Neural Information Processing Systems*, 2024.
[24] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[25] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.
[26] Y. Sun, X. Wang, and X. Tang, "Deep Learning Face Representation from Predicting 10,000 Classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[27] S. I. Serengil and A. Ozpinar, "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules," *Bilisim Teknolojileri Dergisi*, 2024.
[28] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020.