Digital Mystique_ Advanced Multi-Modal Approach to X-Men Character Transformation Using ComfyUI

Abstract

This research addresses the conflict between identity fidelity and detail generation quality in digital character transformation. Traditional practical makeup and earlier AI techniques struggle to achieve high-fidelity visual changes while preserving core features. This paper proposes a novel hybrid workflow integrating multiple AI models: FLUX.1-Kontext provides structural integrity and identity preservation for the base transformation; precise semantic segmentation masks guide GPT-4o’s image engine to inject high-frequency details (e.g., skin scales); human-assisted composition in Adobe Photoshop synthesizes the outputs; finally, Alibaba’s Wan 2.1 Fun Control enables video synthesis via controllable style transfer. Validated on transforming an actor into X-Men's Mystique using consumer-grade hardware (24GB VRAM), this workflow surpasses single-model approaches in identity consistency, detail richness, and visual coherence. The study offers a practical, efficient solution for film production and a blueprint for AI-assisted complex visual effects engineering.

1. Introduction

1.1. The Challenge of Practical Effects in Character Creation

In cinematic production, the creation of visually striking characters has historically relied upon extensive practical effects. The character of Mystique from the *X-Men* film series serves as a pertinent case study, representing one of the most labour-intensive makeup applications in recent film history. The process required a team of six artists working for up to twelve hours, imposing significant logistical and financial burdens on the production and considerable physical strain on the actor (Failes, 2019). The discomfort experienced by actress Jennifer Lawrence during the eighthour application for *X-Men: First Class* (2011) was so pronounced that it necessitated narrative adjustments in subsequent films to minimise the character's screen time in her transformed state (Failes, 2019). These practical limitations underscore a clear imperative for more efficient, digitally-native solutions.

1.2. Generative AI and the Paradigm Shift in Visual Effects

The rapid maturation of artificial intelligence has catalysed a paradigm shift in digital character creation. The field has progressed from manual, frame-by-frame techniques towards AI-driven automation [user query]. Early AI applications, such as GAN-based texture transfer and DeepFakes, demonstrated potential but often lacked the requisite precision and stability for professional media production (Failes, 2019; Westerlund, 2019; Karras et al., 2020). The recent emergence of powerful language-guided models, including Google's Gemini, OpenAI's GPT-4o, and open-source alternatives such as FLUX.1-Kontext, has introduced new possibilities for high-fidelity, instruction-based image editing [user query]. This technological advancement provides the foundation for this project's central research question: how can these state-of-the-art tools be integrated to address a complex and specific visual effects challenge?

1.3. Contribution: A Synergistic Workflow for Identity-Preserving Transformation

Initial investigations revealed that no single model could adequately address the dual requirements of identity preservation and detail generation. Traditional digital inpainting techniques failed to maintain character likeness, while a comparative analysis of leading generative models highlighted a fundamental trade-off [user query]. Models excelling in structural and identity consistency, such as FLUX.1-Kontext, typically produce textures lacking in fine detail. Conversely, models capable of generating photorealistic detail, such as GPT-4o, often compromise the underlying character identity and composition [user query].

The **principal contribution** of this dissertation is therefore the design and validation of a **synergistic multi-stage workflow**. This engineering-led methodology deconstructs the transformation task into discrete stages, assigning each to the most suitable specialised model. The workflow first establishes a structurally accurate base using FLUX.1-Kontext, then injects high-frequency detail with GPT-4o, and finally integrates these outputs via a manual composition phase. This hybrid approach yields a result superior to that of any individual model, offering a pragmatic and effective methodology for achieving professional-grade visual effects on accessible hardware. This structured pipeline directly addresses prior feedback which noted an "unclear workflow," by defining a coherent and purposeful technical process.[1]

2. Literature Review: Technologies in Programmatic Image and Video Synthesis

This chapter provides a critical review of the technological landscape as of mid-2025, establishing the theoretical and technical rationale for the methodological choices made in this study. This directly responds to feedback requesting more in-depth background research.[1]

2.1. The Evolution from GANs to Language-Guided Transformers

The field of digital image synthesis has evolved rapidly. Generative Adversarial Networks (GANs), while capable of producing high-quality images, were often hampered by training instability (Karras et al., 2020). Diffusion Models subsequently emerged as a more robust alternative. More recently, Transformer-based architectures, particularly those employing novel mechanisms like Flow Matching, have advanced the state-of-the-art, providing the foundation for sophisticated language-guided image editing. This domain, which focuses on manipulating image content via natural language, has been systematically surveyed in recent literature, establishing clear task definitions and evaluation metrics ((https://arxiv.org/abs/2502.10064); Zhang et al., 2025).

2.2. Architectures for High-Fidelity Image Editing

This study's hybrid workflow is predicated on the distinct strengths of two key architectural paradigms.

2.2.1. FLUX.1-Kontext: Contextual Consistency via Flow Matching

Developed by Black Forest Labs, FLUX.1-Kontext is an advanced image editing model built upon a **rectified flow** Transformer architecture ((https://arxiv.org/abs/2506.15742)). It is specifically designed for **"in-context image generation,"** accepting both text and image inputs to guide its output ((https://bfl.ai/announcements/flux-1-kontext)). The model's defining characteristic is its exceptional **character consistency**. Technical documentation confirms its ability to robustly preserve the identity of subjects and objects across multiple, iterative edits with minimal visual drift ((https://arxiv.org/abs/2506.15742);(https://bfl.ai/announcements/flux-1-kontext)). This makes it the ideal instrument for the initial transformation stage, where preserving the actor's likeness is paramount.

2.2.2. GPT-4o: Detail Generation via an Autoregressive Engine

OpenAI's GPT-4o incorporates a native image generation capability that is architecturally distinct from diffusion models. Its official system card describes it as an **autoregressive model** deeply integrated within the omni-modal framework (OpenAI, 2025). This architecture grants it powerful **image-to-image transformation** capabilities and a remarkable proficiency for generating **photorealistic** detail. Crucially, its advanced **instruction-following** ability allows it to interpret nuanced text prompts to generate high-frequency textures, such as skin patterns and specular highlights, making it indispensable for the detail-infusion stage of the workflow (OpenAI, 2025).

2.3. Architectures for Temporally Coherent Video Generation

Extending the static transformation to video requires a model capable of maintaining temporal consistency between frames. The investigation of available models was guided by controllability, performance, and hardware accessibility.

2.3.1. The ControlNet Paradigm

The video synthesis stage is founded on the ControlNet paradigm, which uses auxiliary control signals—such as Canny edges or human pose data—to guide the generation process. This ensures that the structure and motion of the output video remain faithful to the source footage.

2.3.2. Comparative Analysis of Video Generation Models

A review of leading models revealed a significant trade-off between capability and accessibility.

- **VACE (Video All-in-One Creation and Editing):** This framework from Alibaba aims to unify multiple video tasks within a single model (Jiang et al., 2025). While powerful, its recommended hardware requirement of over 32GB of VRAM rendered it unsuitable for this project's consumer-grade setup.[1]
- **Tencent HunyuanCustom:** This multi-modal framework excels at identity preservation but is similarly demanding, with community reports indicating a need for enterprise-level GPUs with up to 80GB of VRAM for effective operation (Hu et al., 2025; [1]).
- **Alibaba PAI Wan 2.1 Fun Control:** In contrast, this model offers a more balanced solution (Alibaba-PAI, 2025). It supports multiple ControlNet inputs, including **style transfer from a reference image**, and its 1.3B parameter version is explicitly designed for consumer GPUs with 12-24GB of VRAM (Alibaba-PAI, 2025).

This analysis demonstrates that while theoretically superior models exist, their hardware prerequisites create a practical accessibility gap. The selection of Wan 2.1 Fun Control was therefore not a compromise, but an optimised decision based on a pragmatic assessment of the project's technical requirements and available resources. This grounding in real-world constraints enhances the practical value and transferability of the proposed methodology.

3. Methodology: A Synergistic Multi-Stage Workflow

To overcome the limitations of individual models and in direct response to feedback requesting greater clarity [1], a structured, multi-stage workflow was designed and implemented. This section details the system architecture and the execution of each stage.

3.1. System Architecture and Data Flow

The workflow architecture is based on a 'divide and conquer' strategy, deconstructing the complex task into four discrete stages implemented within the ComfyUI node-based environment. Figure 1 illustrates the data flow.

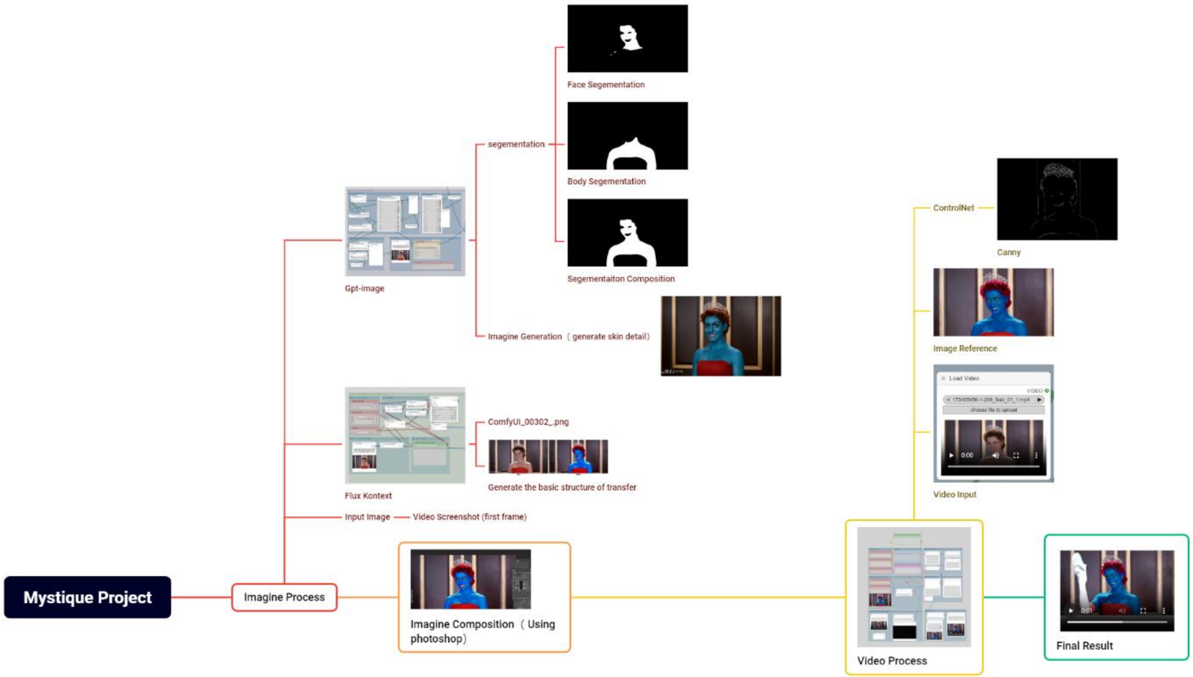
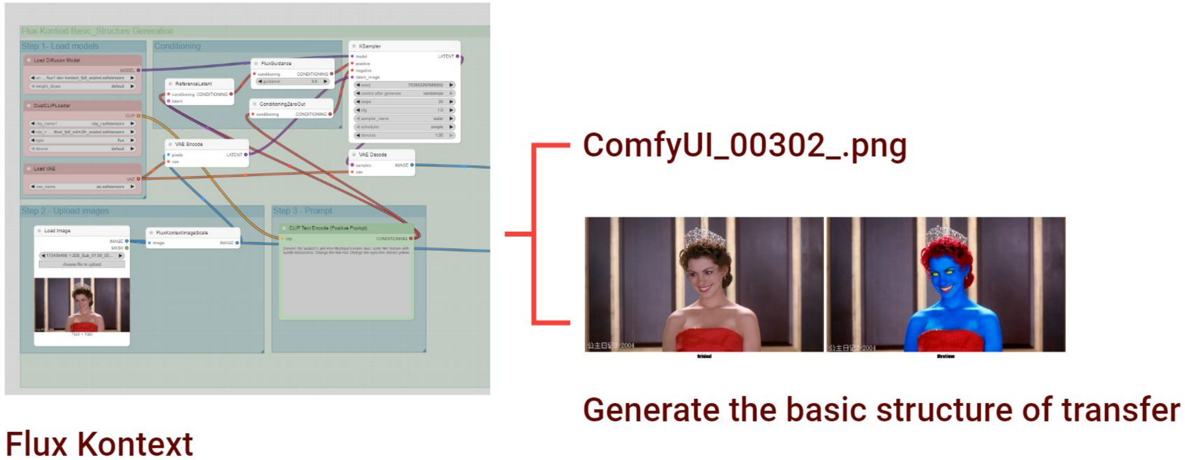


Figure 1: Digital Mystique Hybrid Workflow Architecture 1

The stages are as follows:



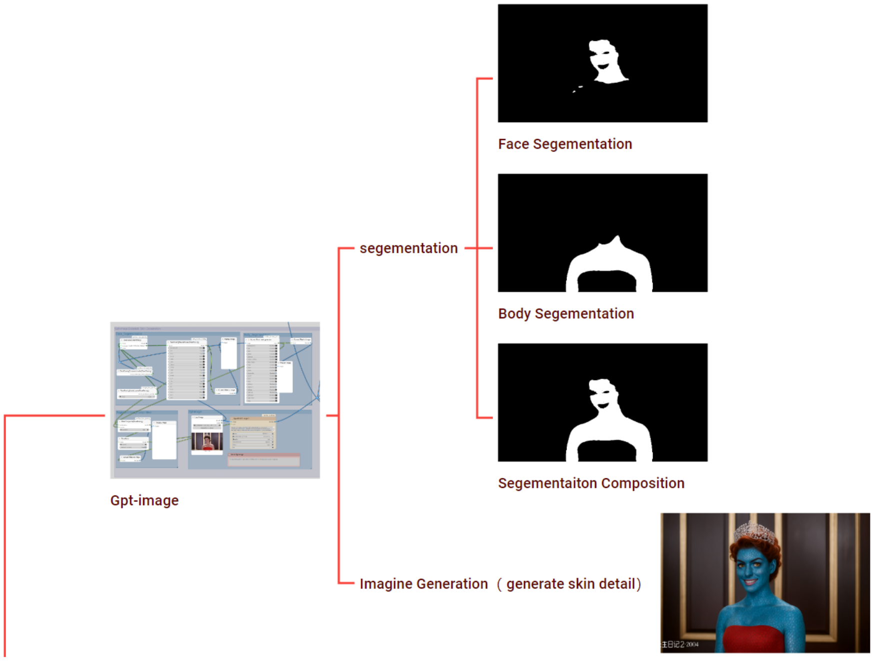
ComfyUI_00302_.png

Generate the basic structure of transfer

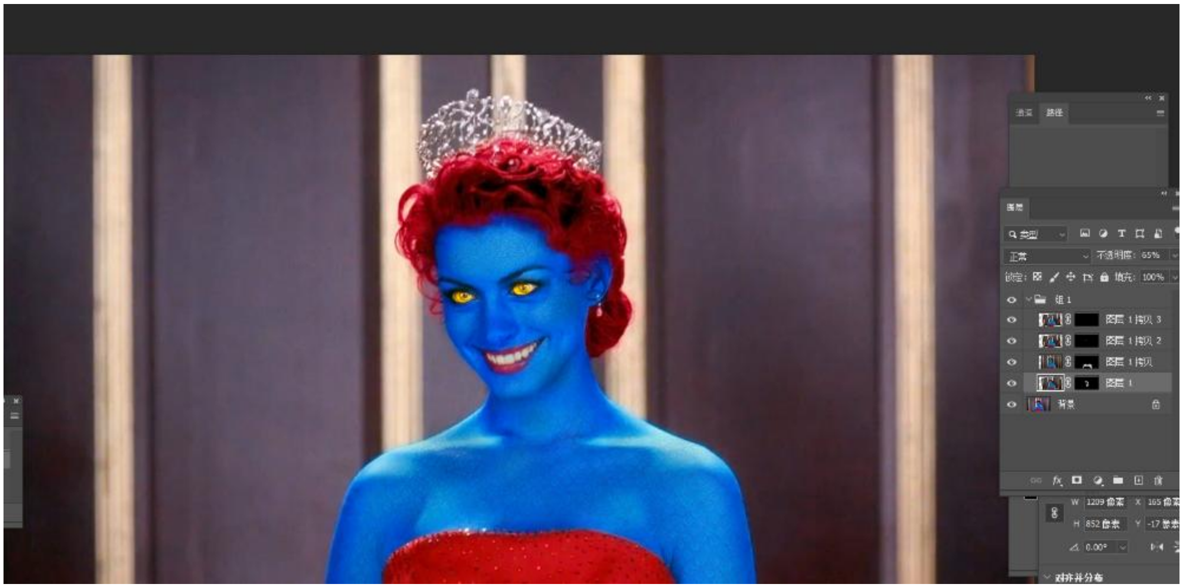
Flux Kontext

Input Image —— Video Screenshot (first frame)

1. **Foundational Transformation:** Utilising FLUX.1-Kontext to perform the primary colour and feature transformation while ensuring complete identity preservation.


Face Segementation
segementation
Body Segementation
Gpt-image
Segementaiton Composition
Imagine Generation （generate skin detail）

2. **High-Frequency Detail Infusion:** Employing GPT-4o, guided by precise segmentation masks, to generate realistic skin textures.



正常

不透明度: 65%

锁定:

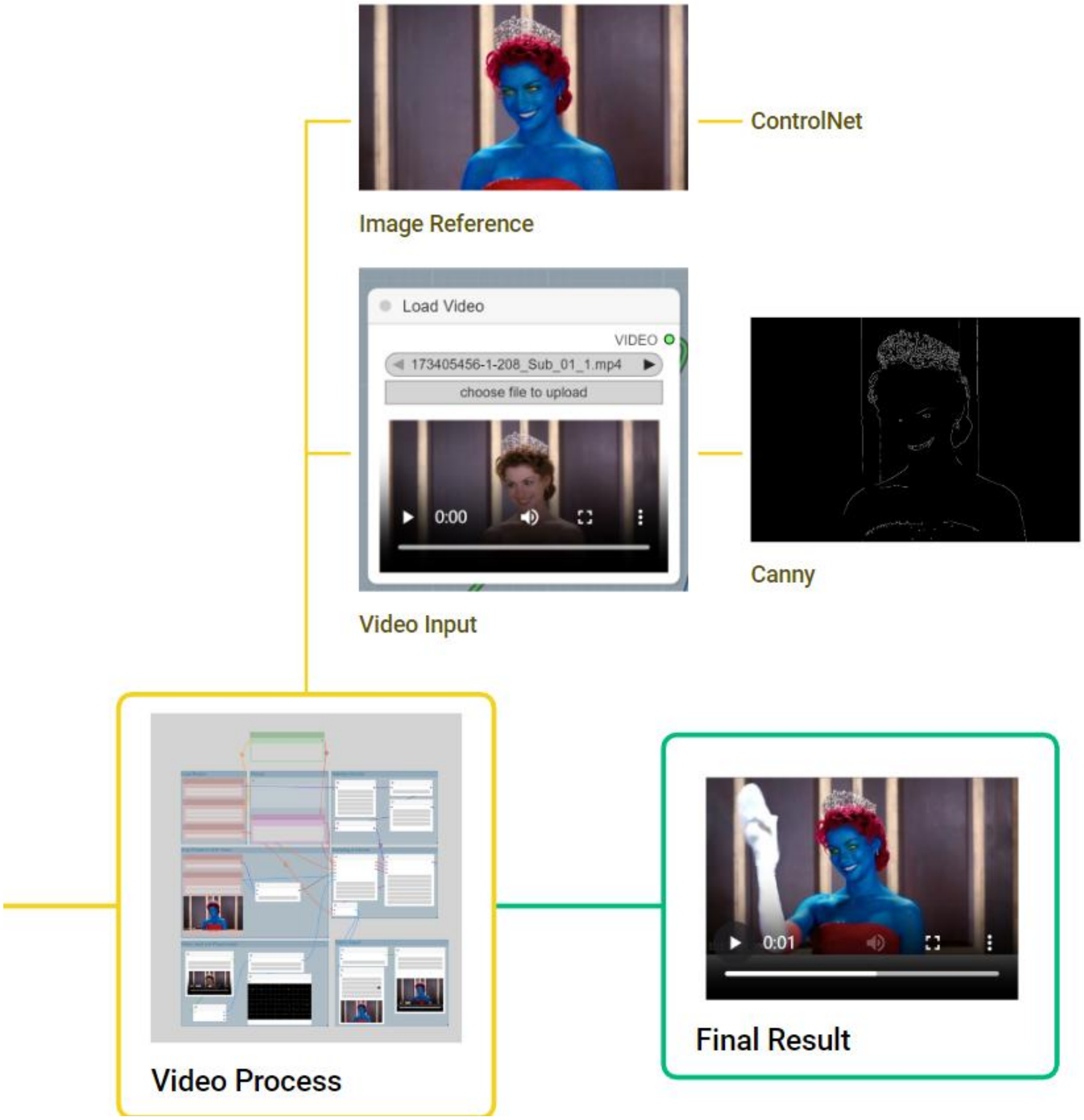
填充: 100%

W 1209 像素 X 165 像素

H 852 像素 Y -17 像素

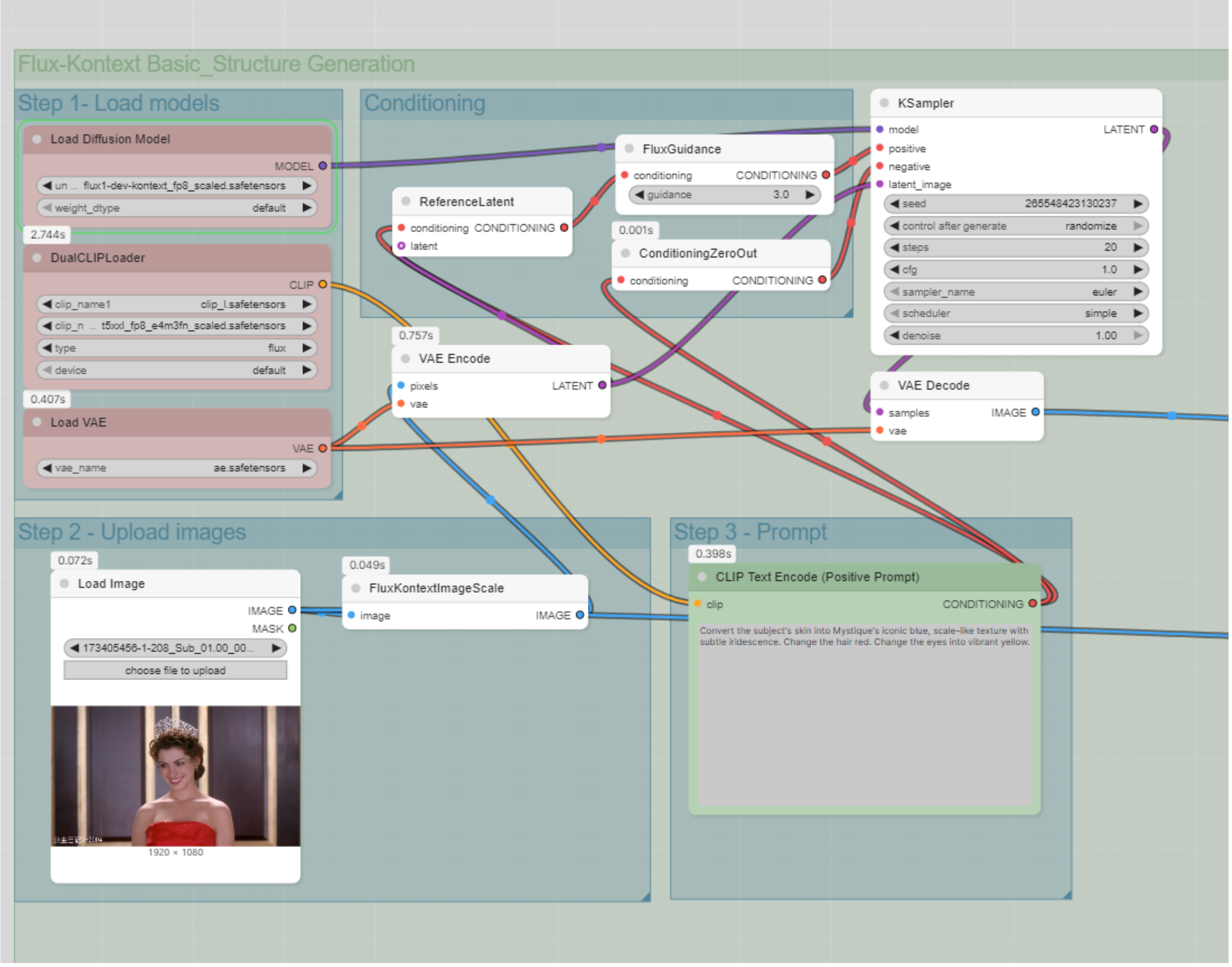
0.00°

3. **Synergistic Composition:** A manual compositing stage in Adobe Photoshop to merge the outputs from the first two stages.



4. **Controlled Video Synthesis:** Using Wan 2.1 Fun Control to animate the final static image by applying it as a style reference to the source video.

3.2. Stage One: Foundational Transformation with FLUX.1-Kontext



Flux-Kontext Basic_Structure Generation

Step 1- Load models

Load Diffusion Model (2.744s)
un ... flux1-dev-kontext_fp8_scaled.safetensors
weight_dtype default

DualCLIPLoader (0.407s)
clip_name1 clip_l.safetensors
clip_n ... t5xxl_fp8_e4m3fn_scaled.safetensors
type flux
device default

Load VAE
vae_name ae.safetensors

Conditioning

ReferenceLatent
FluxGuidance
guidance 3.0
ConditioningZeroOut (0.001s)

VAE Encode (0.757s)

KSampler
seed 265548423130237
control after generate randomize
steps 20
cfg 1.0
sampler_name euler
scheduler simple
denoise 1.00

VAE Decode

Step 2 - Upload images

Load Image (0.072s)
173405456-1-208_Sub_01.00_00...
choose file to upload
1920 × 1080

FluxKontextImageScale (0.049s)

Step 3 - Prompt

CLIP Text Encode (Positive Prompt) (0.398s)
Convert the subject's skin into Mystique's iconic blue, scale-like texture with subtle iridescence. Change the hair red. Change the eyes into vibrant yellow.

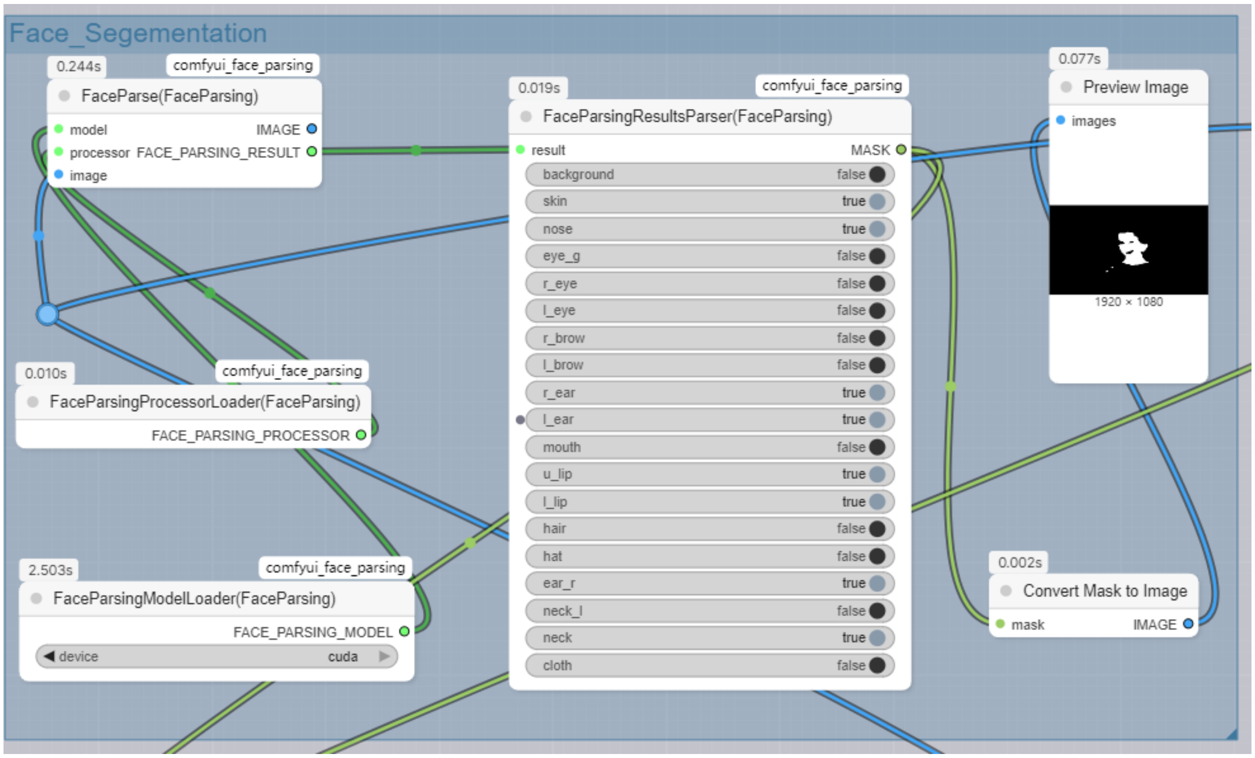
The objective of this initial stage was to produce a structurally correct base image. The primary requirement was to execute the necessary colour changes (blue skin, red hair, yellow eyes) while strictly preserving the actor's identity, expression, and pose from the source image. Within ComfyUI, the FLUX.1-Kontext node was supplied with the source image and a text prompt, such as: " *Convert the subject's skin into Mystique's iconic blue, scale-like texture with subtle iridescence. Change the hair red. Change the eyes into vibrant yellow.*" The output of this stage is an image that is structurally and chromatically correct but lacks fine surface detail, providing a stable foundation for subsequent enhancement [user query].

3.3. Stage Two: High-Frequency Detail Infusion with GPT-4o

This stage was designed to address the textural deficiencies of the Stage One output by leveraging the generative power of GPT-4o.

3.3.1. Precision Masking via Semantic Segmentation

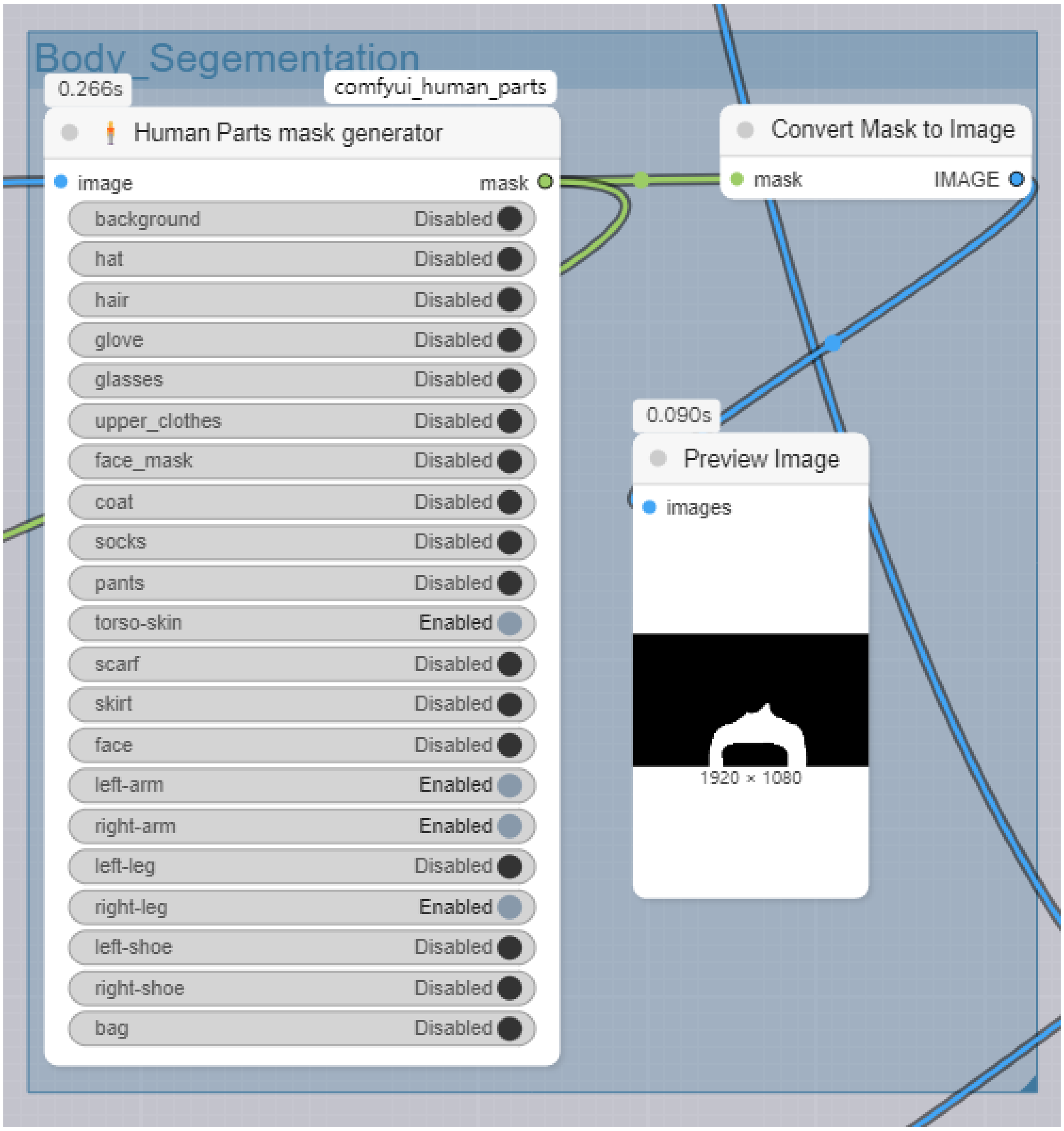
To constrain the powerful but less controllable GPT-4o model, its application was restricted to specific regions defined by high-precision semantic segmentation masks. This prevented the destruction of the character's identity [user query]. A composite mask was generated by combining the outputs of several specialised ComfyUI nodes:



background	false
skin	true
nose	true
eye_g	false
r_eye	false
l_eye	false
r_brow	false
l_brow	false
r_ear	true
l_ear	true
mouth	false
u_lip	true
l_lip	true
hair	false
hat	false
ear_r	true
neck_l	false
neck	true
cloth	false

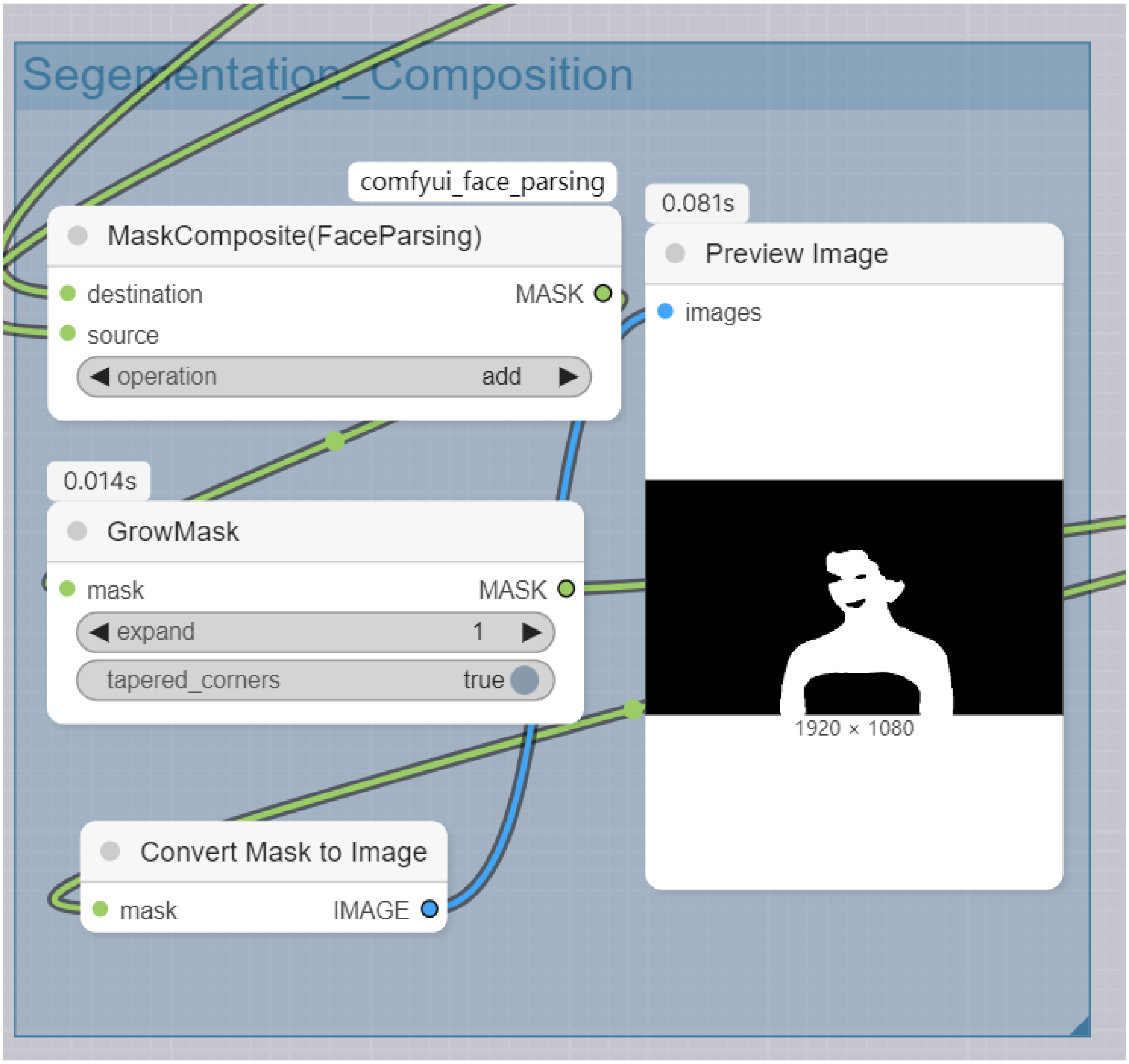
◀ device cuda ▶

- **Face Parsing:** A node utilising the jonathandinu/face-parsing model, based on the Segformer architecture, to accurately segment 19 distinct facial regions, including 'skin', 'hair', and 'eyes' ((https://huggingface.co/jonathandinu/face-parsing)).



background Disabled
hat Disabled
hair Disabled
glove Disabled
glasses Disabled
upper_clothes Disabled
face_mask Disabled
coat Disabled
socks Disabled
pants Disabled
torso-skin Enabled
scarf Disabled
skirt Disabled
face Disabled
left-arm Enabled
right-arm Enabled
left-leg Disabled
right-leg Enabled
left-shoe Disabled
right-shoe Disabled
bag Disabled

- **Body Parsing:** The Human Parts Detector custom node, based on the DeepLabV3+ model, to segment body regions such as 'torso-skin' and arms (CozyMantis, 2024; Li et al., 2020).



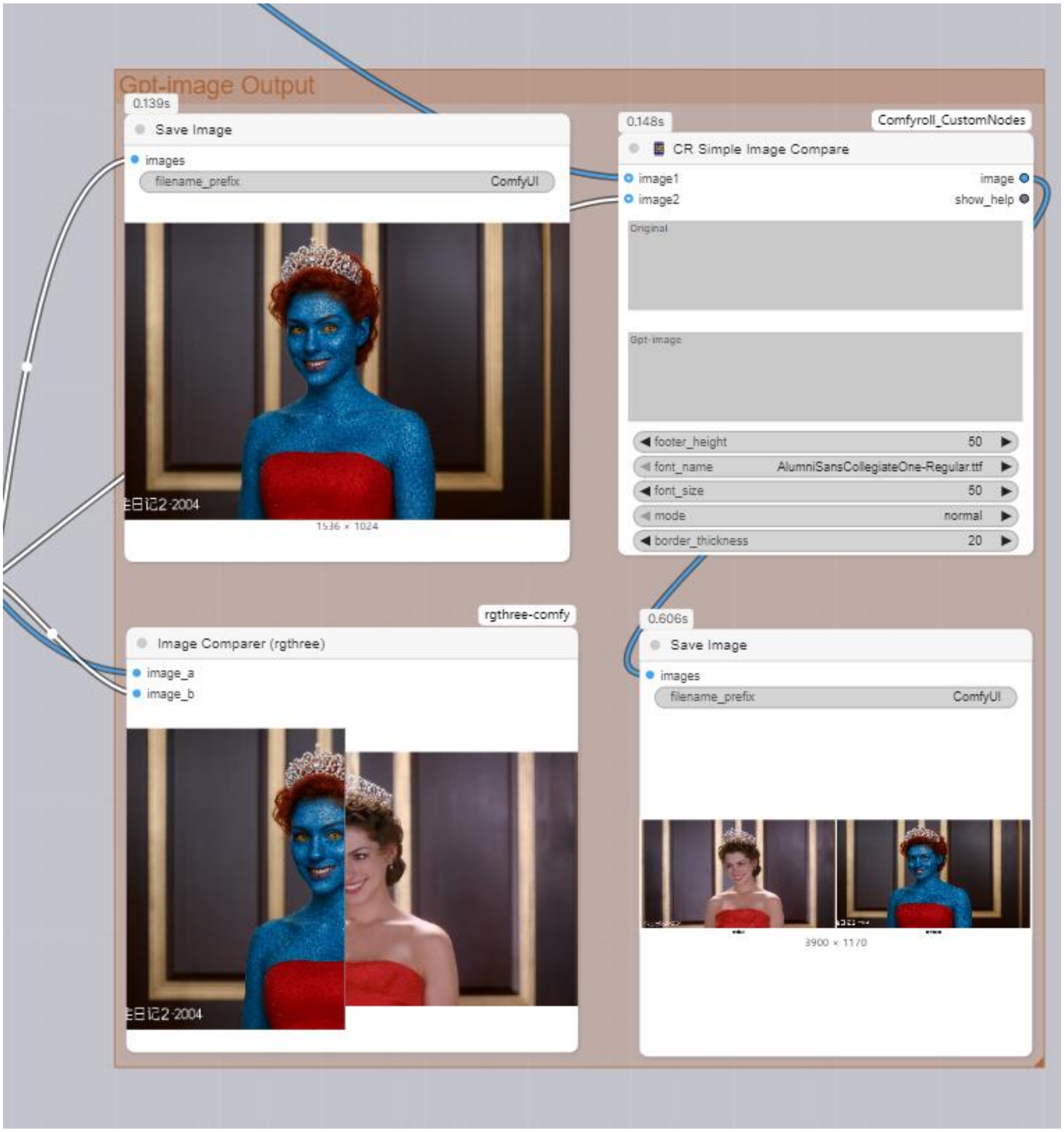
operation add

expand 1

tapered_corners true

- **Mask Composition:** The individual skin masks were logically combined within ComfyUI to create a single, precise mask delineating all target skin areas while excluding non-target regions.

3.3.2. Mask-Guided Detail Generation



Gpt-image Output

0.139s

Save Image

images

filename_prefix ComfyUI

1536 × 1024

0.148s

Comfyroll_CustomNodes

CR Simple Image Compare

image1 image

image2 show_help

Original

Gpt-image

footer_height 50

font_name AlumniSansCollegiateOne-Regular.ttf

font_size 50

mode normal

border_thickness 20

rgthree-comfy

Image Comparer (rgthree)

image_a

image_b

0.606s

Save Image

images

filename_prefix ComfyUI

3900 × 1170

The base image from Stage One, the composite skin mask, and a detail-oriented text prompt were supplied to the GPT-4o image editing process. The prompt focused exclusively on texture, for instance: " *Convert the subject's skin into blue, scale-like texture with subtle iridescence. It is crucial to maintain the original person's distinct facial features, bone structure, and their current expression. The lighting should match the original photo. Change the hair red. Change the eyes into vibrant yellow.*" While this approach yielded high-quality details, minor artefacts and edge mismatches were still observed, confirming the limitations of a fully automated process and necessitating the subsequent manual stage [user query].

3.4. Stage Three: Human-Assisted Synergistic Composition

This manual stage represents a pragmatic engineering decision to achieve a result superior to what is currently possible with fully automated methods. The composition was performed in Adobe Photoshop:

1. The 'structure-correct' image from Stage One was set as the base layer.
2. The 'detail-rich' image from Stage Two was placed as the top layer.
3. The high-precision skin mask was applied as a layer mask to the top layer.

This process effectively transferred the high-fidelity skin texture from the GPT-4o output onto the identity-preserving structure from the FLUX.1-Kontext output, yielding the final, optimised static image.

3.5. Stage Four: Controlled Video Synthesis with Wan 2.1 Fun Control

The final stage animated the static asset. The task was framed as a style transfer problem, using the Wan 2.1 Fun Control model in ComfyUI. The inputs were the original source video and the final static image from Stage Three, which served as the **style reference**. A Canny edge map was extracted from the source video and used as a ControlNet signal to ensure the motion in the output remained consistent with the original performance. Through prompt optimisation, for instance” Mystique wave her hands wearing white long gloves with bright smile “, the model's ability to handle occluding objects (such as gloves) was improved, successfully addressing a specific issue noted in prior feedback.[1] This approach cleverly leveraged the model's strengths, converting a complex character transformation task into a more manageable style transfer operation.

4. Experimental Evaluation and Comparative Analysis

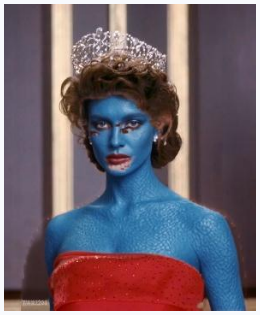
This section presents a detailed qualitative and quantitative comparison of the different methods explored, validating the efficacy of the proposed hybrid workflow and responding to feedback that the previous analysis was unconvincing.[1]

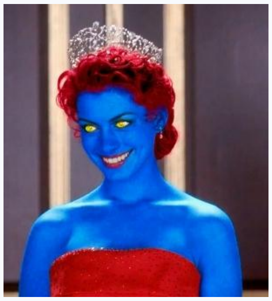
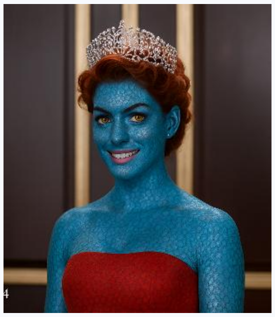
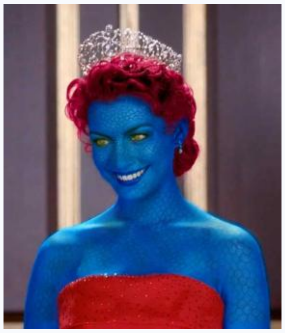
4.1. Image Generation Analysis: Validating the Hybrid Approach

A qualitative comparison of outputs from different methods demonstrates the necessity of the hybrid workflow. Initial attempts using inpainting with Flux Fill resulted in severe facial distortion. Using FLUX.1-Kontext alone achieved perfect identity preservation but lacked realistic skin texture. Conversely, a mask-guided GPT-4o approach produced excellent detail but with noticeable deviation in facial structure. The proposed hybrid workflow successfully synthesised the strengths of the latter two methods, combining the identity fidelity of FLUX.1-Kontext with the rich detail from GPT-4o.

Table 1 provides a systematic summary of these findings.

Table 1: Qualitative Comparative Analysis of Image Generation Methodologies

Method	image	Identity Fidelity	High-Frequency Detail	Structural Consistency	Overall Feasibility
Inpainting (Flux Fill)		Low	Low	Low	Unusable

FLUX.1-Kontext (Standalone)		Very High	Low	Very High	Incomplete
GPT-4o (Mask-Guided)		Medium	Very High	Medium	Incomplete
Proposed Hybrid Workflow		**Very High**	**Very High**	**Very High**	**Optimal**

The analysis clearly indicates that each standalone method possesses critical deficiencies. The proposed workflow systematically overcomes these by integrating the models in a complementary fashion, thereby achieving an optimal result that would be unattainable with any single approach.

4.2. Video Generation Analysis: Temporal Consistency under Hardware Constraints

The selection of a video generation model was primarily constrained by the available hardware. The objective was to maximise temporal consistency on a consumer-grade system. Table 2 compares the leading controllable video generation models, justifying the selection of Wan 2.1 Fun Control.

Table 2: Video Transformation Model Comparison on Consumer Hardware (24GB VRAM)

Video Model	Transformation Quality	Temporal Consistency	Hardware Requirement (VRAM)	Feasibility
Tencent HunyuanCustom	Very High	Very High	> 48GB [1]	Not Feasible
Alibaba VACE	Very High	Very High	> 32GB [1]	Not Feasible
Alibaba Wan 2.1 Fun Control (1.3B)	**High**	**High (moderate motion)**	~12-16GB [3]	**Optimal / Feasible**

The data confirms that while more powerful models exist, their prohibitive hardware requirements render them impractical for this project. The choice of Wan 2.1 Fun Control was therefore the most suitable solution, representing a decision optimised for real-world production constraints. The resulting video demonstrated good transformation quality and temporal coherence for footage containing moderate levels of motion.

5. Discussion, Limitations, and Future Work

5.1. Critical Evaluation of the Final Workflow

This research has successfully demonstrated a workflow that produces high-quality static and dynamic character transformations on consumer-grade hardware. The primary strengths of this approach are the high fidelity of the output, which benefits from the combined advantages of specialised models, and its efficiency. Compared to the extensive time and personnel required for traditional makeup [1], this digital process offers a significant reduction in cost and a substantial improvement in actor comfort. Furthermore, its implementation on a 24GB VRAM GPU confirms the accessibility of this approach for artists and smaller studios without access to enterprise-level hardware.

However, it is important to acknowledge that while digital methods offer superior flexibility, the tactile realism of physical makeup may still be preferable for certain cinematic applications, particularly in close-up shots involving physical interaction with the environment.

5.2. Identified Limitations

Despite its success, the workflow has several limitations which highlight areas for future research:

- **Dependence on Manual Intervention:** The reliance on Adobe Photoshop for the composition stage is the most significant limitation. This manual step, while ensuring quality, prevents full automation and introduces a subjective element. It reflects a current gap in the ability of AI models to seamlessly merge outputs from different generative processes.
- **Video Motion Constraints:** The chosen video model, Wan 2.1 Fun Control, performs well with moderate motion but can produce artefacts during rapid or complex action sequences. This restricts the workflow's applicability in high-intensity scenes.
- **Segmentation Accuracy:** While advanced segmentation models were used, their accuracy can degrade under challenging conditions such as poor lighting or complex object occlusion, sometimes necessitating manual mask refinement.

5.3. Future Work: Towards a Unified, End-to-End Model

The identified limitations suggest clear directions for future research. The most immediate goal is the **elimination of the manual composition stage**. This could be pursued by training a dedicated 'refinement network' capable of learning to intelligently fuse the structural and textural outputs from the first two stages into a coherent whole.

A more ambitious, long-term objective is the development of a **single, unified model** that integrates the contextual awareness of FLUX.1-Kontext with the detail-generation capabilities of GPT-4o. While frameworks like VACE signal a move in this direction (Jiang et al., 2025), they are not yet fully mature and remain computationally prohibitive. The development of an end-to-end model that is both highly capable and accessible on consumer hardware remains the ultimate goal for this area of research.

6. Conclusion

This dissertation has presented a novel, synergistic workflow for digital character transformation, designed to address the critical challenge of maintaining identity fidelity while generating high-quality surface detail. By integrating the complementary strengths of FLUX.1-Kontext, GPT-4o, and Wan 2.1 Fun Control within a structured, multi-stage pipeline, this research has demonstrated that a high-fidelity digital recreation of the 'Mystique' character is achievable on consumer-grade hardware.

The experimental analysis confirms that this hybrid approach yields results superior to those of any single model, offering a pragmatic solution to a persistent problem in generative AI. The contribution of this work is therefore twofold: it provides a practical, efficient, and accessible workflow for a specific visual effects task, and it demonstrates a broader engineering methodology for combining specialised AI tools to overcome their individual limitations. This research serves as a viable blueprint for AI-assisted production and offers valuable insights for the future development of more powerful, unified, and accessible generative models.

References

- Alibaba-PAI (2025) *Wan2.1-Fun-1.3B-Control*. Available at:(https://huggingface.co/alibaba-pai/Wan2.1-Fun-1.3B-Control) (Accessed: 21 May 2025). [1]

- Black Forest Labs (2025a) *FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space*. Available at: https://arxiv.org/abs/2506.15742 (Accessed: 12 July 2025). [4]
- Black Forest Labs (2025b) *Introducing FLUX.1 Kontext and the BFL Playground*. Available at: https://bfl.ai/announcements/flux-1-kontext (Accessed: 12 July 2025). [5]
- Dinu, J. (2024) *face-parsing*. Hugging Face. Available at: https://huggingface.co/jonathandinu/face-parsing (Accessed: 12 July 2025). [6]
- Failes, I. (2019) 'The evolution of digital make-up', *Befores & Afters*, 15 March. Available at: https://beforesandafters.com/2019/03/15/the-evolution-of-digital-make-up/ (Accessed: 12 July 2025). [1]
- Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q. and Lu, Q. (2025) *HunyuanCustom: A Multimodal-Driven Architecture for Customized Video Generation*. Available at: https://arxiv.org/abs/2505.04512 (Accessed: 8 May 2025). [7]
- Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y. and Liu, Y. (2025) *VACE: All-in-One Video Creation and Editing*. Available at: https://arxiv.org/abs/2503.07598 (Accessed: 10 March 2025). [8]
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. and Aila, T. (2020) 'Analyzing and improving the image quality of StyleGAN', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110-8119. [1]
- OpenAI (2025) *Addendum to GPT-4o System Card: 4o image generation*. Available at: https://openai.com/index/gpt-4o-image-generation-system-card-addendum/ (Accessed: 12 July 2025). [9]
- Santos, R., Branco, A., Silva, J. and Rodrigues, J. (2025) *Hands-off Image Editing: Language-guided Editing without any Task-specific Labeling, Masking or even Training*. Available at: https://arxiv.org/abs/2502.10064 (Accessed: 12 July 2025). [10]
- Westerlund, M. (2019) 'The emergence of deepfake technology: A review', *Technology Innovation Management Review*, 9(11), pp. 40-53. [1]
- Zhang, X., et al. (2025) *A Survey on Image Editing with Diffusion Models*. Available at: https://arxiv.org/html/2504.13226 (Accessed: 12 July 2025). [11]