Re-Imagining Photoshop - The AI Editor of 2030

Mid Evaluation Report

Team 76

Inter IIT Tech Meet 14.0

1 Approach Outline

1.1 Problem Understanding

The problem statement asks us to imagine how creative tools - especially Photoshop will evolve by 2030 in the world where mobile devices and AI-assisted workflows dominate.Current editing tools are powerful but still heavily dependent on manual operations , complex interfaces , and high computational resources.In contrast , the brief envisions a future where creators interact with images more naturally and effortlessly , using simple prompts , fluid gestures and most minimal hardware. The challenge is to indentify gap's in today's creative ecosystem and propose how AI can fill these gaps making editing faster , more intuitive, and more context aware. We are expected to deliver two workflows that demonstrate this shift : features that are not just "automated version of existing tools", but genuinely rethink how editing should feel when powered by intelligent models.These workflows must be grounded in real user pain points, supported by a clear market rationale , and implemented using open-source AI models capable of region selection,generation, and inpainting.Overall the problem asks us to blend user research,design thinking, and cutting edge AI to build a prototype that reflects the creative experience of 2030-lightweight , intelligent and human-centric.

1.2 Solution Overview

Our solution is built around two complementary workflows that together represent the future of AI-assisted, mobile-friendly image editing. Before entering either workflow, the user can begin by uploading an image or generating one using our user-style personalized LoRA, ensuring a highly customized starting point. From there, the system branches into two specialized pipelines designed to support different creative needs.

1.2.1 Workflow 1 : AI-Enhanced Image editing Tools

The first workflow focuses on intuitive, fine-grained image editing using a suite of advanced open-source AI tools. It includes LeDits++ for image-to-image transformation, enabling users to refine or restyle their images with high fidelity. For artistic transformations, we integrate a style-transfer module that automatically selects the most appropriate style LoRA based on the user's prompt and applies it seamlessly.Region-level editing is supported through Segment Anything (SAM), which allows users to isolate any part of the image and then choose to erase it, inpaint new content, or manipulate it using Inpaint4Drag, a state-of-the-art drag-based deformation model.Additionally, the workflow includes Lightning Drag, which enables users to adjust the direction or orientation an object is facing, and Generative Expand, an outpainting tool that extends scenes while preserving visual coherence. Together, these tools form an intelligent, flexible editing environment that reflects the natural, prompt-driven editing experience envisioned for 2030.

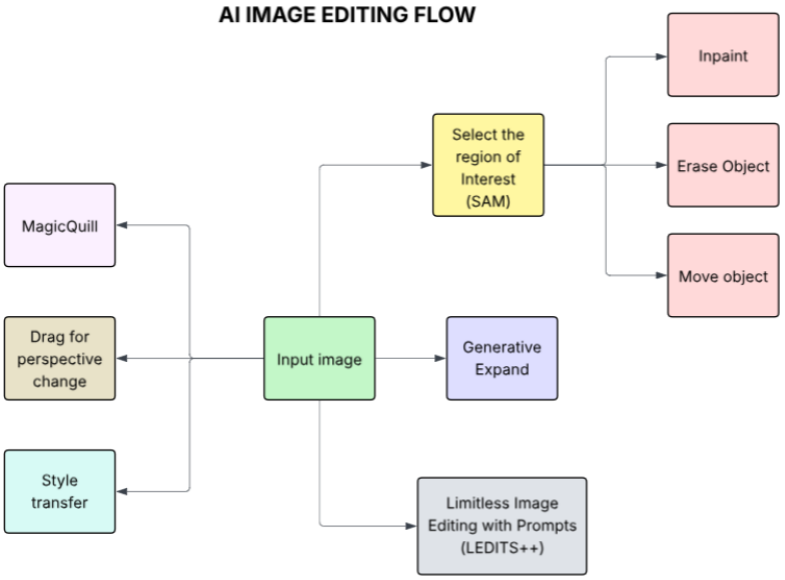


Figure 1: AI Image Editor Workflow. A flowchart illustrating the core capabilities of the editing system, which branches from an input image into specialized modules for region-based editing (using SAM), generative expansion, text-guided editing (LEDITS++), style transfer, and perspective adjustments.

1.2.2 Workflow 2 : Smart Composition and 3D-Aware Object Insertion

The second workflow is designed for high-quality object insertion and blending, enabling users to integrate new elements into a scene with realism and spatial coherence.The process begins with Smart Crop, which prepares and focuses the base image. The user then selects any object image to insert, and the system automatically removes its background, isolating the subject. This extracted object is passed through a 2D-to-3D generation model, which reconstructs a lightweight 3D representation that allows proper orientation, scaling, and positioning relative to the target image. Once the 3D orientation is finalized, the object is composited back into the scene. The blended result is then refined through a relighting model, ensuring that shadows, highlights, and color temperature align with the background. Finally, the combined and harmonized output is delivered, producing an integrated and realistic image with minimal user effort.

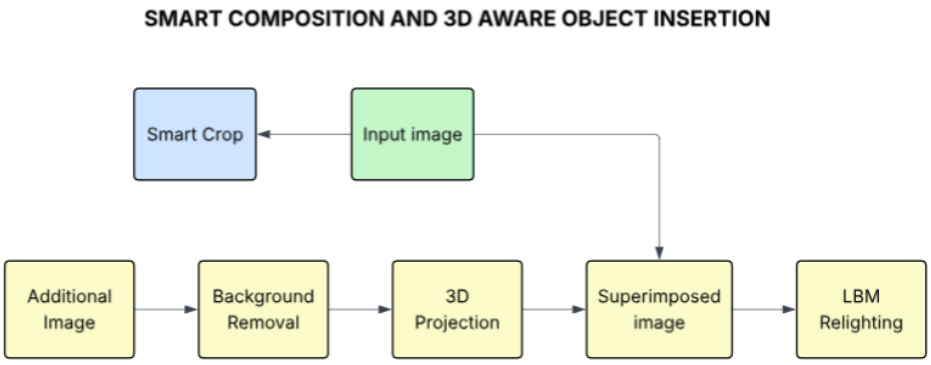


Figure 2: AI Image Editor Workflow. A flowchart illustrating the system's composition and relighting pipeline. It details how an additional image is processed via background removal (BRIA) and 3D projection (SF3D) before being superimposed onto the main input image and enhanced with LBM relighting. The workflow also includes a module for smart cropping.

1.3 Overview of Overall Inference

This section describes the architecture used to deliver a fast, memory-efficient, and scalable image-editing experience. The system combines dynamic model loading, resolution-aware processing, optimized Stable Diffusion backbones, and reusable vision-language components to enable high-quality results with minimal latency.

1.3.1 Dynamic Model Loading and Execution

To reduce GPU memory usage and ensure consistent performance, the system loads each model only when the user selects a specific tool. When a tool is activated, a `POST /load_model` request initializes the model on the GPU. After the user provides an image or prompt, inference is executed using `POST /run`. When the user leaves the tool or returns to the main menu, a `POST /unload` call removes the model from GPU memory. This approach offers three main advantages:

- **Lower memory footprint:** Only the active model occupies GPU resources.
- **Lower latency:** Avoids constantly reloading large models.
- **Higher scalability:** Multiple tools can coexist without exceeding GPU limits.

This load-on-demand strategy is a key enabler for a responsive editing experience, even on limited hardware.

1.3.2 High-Resolution Processing with Efficient Upscaling

The system supports user inputs up to 2K resolution. However, performing all operations directly at full size would significantly increase computation time. Instead, the image is first downsampled to 512×512, where the main editing or generation task is performed. After the operation, a diffusion-based upscaler reconstructs the output back to 2K resolution.

This design provides:

- **Fast processing:** Most compute-heavy work happens at lower resolution.
- **High quality:** The upscaler restores detail and sharpness effectively.
- **Product efficiency:** Users receive high-resolution results without long waits.

This balance between speed and quality is a critical aspect of the overall product experience.

1.3.3 Optimized Stable Diffusion 1.5 for Core Tasks

Stable Diffusion 1.5 (SD 1.5) serves as the backbone for a majority of the system's editing tasks. The model has been optimized internally to improve both speed and accuracy. These optimizations include faster sampling techniques, lighter conditioning paths, and more efficient GPU memory usage. Because SD 1.5 is versatile and reliable across many editing

modes, it is reused for over 60% of all operations in the product. This reuse avoids switching between models unnecessarily, reducing load times and ensuring a smooth workflow. From a product perspective, this creates a consistent user experience with predictable visual quality and optimized performance across tools.

1.3.4 Reusable Auxiliary Models: Moondream and CLIP

Beyond the core diffusion model, the system integrates additional lightweight models such as Moondream and CLIP. These models support semantic understanding, content filtering, and other auxiliary tasks that enhance user experience.

They are initialized once per session and reused across different tools, providing:

- **Faster feature extraction and analysis**
- **Efficient content and safety checks**
- **Reduced overhead from repeated model initialization**

This shared-model approach simplifies the system architecture and improves responsiveness across multiple product features.

2 Our Solution

2.1 Text to Image

2.1.1 SANA 1.6B + Nunchaku: Efficient Text-to-Image Generation

Introduction

SANA is a text-to-image diffusion framework engineered for high-resolution generation (up to 4096×4096) with strong text–image alignment. To further enable deployment on consumer-grade GPUs, the **SVDQuant +** Nunchaku system introduces aggressive 4-bit quantization for both weights and activations, paired with a specialized inference engine. Together, this stack delivers fast, resource-efficient and high-fidelity image generation.

Methodology

- **SANA-1.6B Latent Diffusion Transformer:** SANA performs denoising in a $32\times$-compressed latent space using a Linear-DiT architecture with linear attention and Mix-FFN blocks, enabling efficient high-resolution (up to 4K) generation by avoiding quadratic attention costs.
- **Nunchaku 4-bit SVD-Quantization:** Nunchaku stabilizes low-bit diffusion by decomposing each weight matrix into a low-rank SVD component $U_r\Sigma_r V_r^T$ plus a 4-bit quantized residual $Q(W_r)$.

$$W \approx U_r\Sigma_r V_r^T + Q(W_r)$$

  Combined with dynamic activation rescaling and fused low-bit kernels, it reduces memory bandwidth and preserves near full-precision fidelity during 4K inference on consumer GPUs.

Advantages Over Other Solutions

- **4K Generation at Extreme Speed:** SANA-1.6B with linear attention reduces 4K diffusion latency from ~ 469s to ~ 9.6s, enabling high-resolution generation on consumer GPUs.
- **Massive VRAM Savings:** Nunchaku's 4-bit SVDQuant lowers memory usage to just 25–33% of FP16 while preserving over 95% of full-precision visual quality.
- **Optimized Throughput:** Fused low-rank and low-bit kernels provide additional speed-ups, enabling fast, high-fidelity inference without sacrificing detail.

2.1.2 FLUX.1-dev + Nunchaku: Efficient Text-to-Image Diffusion Transformer (CLOUD PIPELINE)

Introduction

FLUX.1-dev is a high-performance diffusion transformer designed for fast, instruction-aligned image generation with strong semantic control. It operates as an efficient DiT-style architecture optimized for general-purpose synthesis, editing, and visual conditioning. When combined with the Nunchaku quantization framework, FLUX.1-dev becomes deployable on limited-memory GPUs while maintaining near-full-precision output quality.

Methodology

- **Efficient Latent Diffusion Backbone:** A VAE compresses images into a compact latent space where the diffusion transformer performs denoising with reduced dimensionality, preserving semantic and structural detail while

lowering compute.

- **Stable 4-Bit SVDQuant Inference:** Nunchaku decomposes each weight into a low-rank SVD branch plus a 4-bit residual, with dynamic activation scaling and fused low-bit kernels enabling stable, low-VRAM inference without FP16 fallback.

Advantages Over Other Solutions

- **Faster High-Resolution Diffusion:** SANA denoises in a $32\times$ compressed latent space, enabling 4K generation that is far faster than pixel-space transformers while preserving global and local detail. It enables practical high-resolution, prompt-aligned generation.
- **Superior Speed–Memory Tradeoff:** Nunchaku's 4-bit SVDQuant cuts VRAM usage to 25–33% of FP16 while retaining over 95% of full-precision quality, outperforming naive 4-bit or 8-bit quantization methods that destabilize diffusion.

2.1.3 Personalized FLUX.1-dev (LoRA Fine-Tuning) (CLOUD PIPELINE)

Introduction

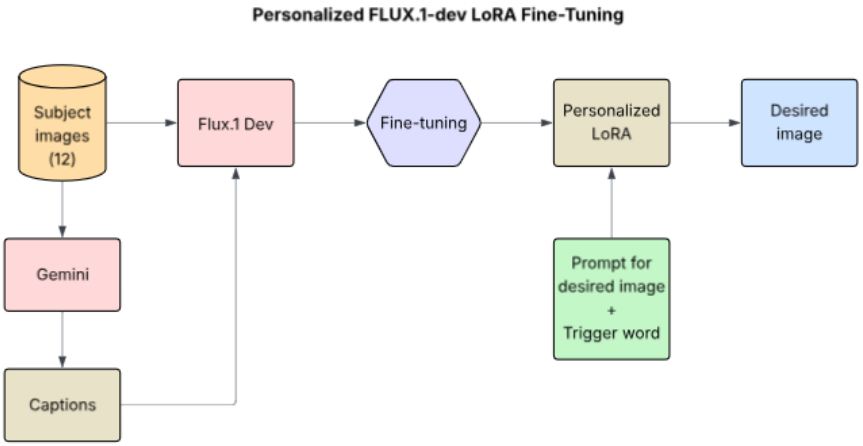
We finetuned FLUX.1-dev, a high-performance diffusion transformer, using the Ostris LoRA training framework to achieve personalized image generation. A small dataset of 12 images of a subject was used to adapt the base model so it can accurately reproduce the identity through natural prompts. The goal is to obtain a lightweight, fast, identity-preserving personalization module.

Methodology

- **Fine-Tuning Strategy:** A LoRA module (rank 16, alpha 16) is applied to FLUX-1-dev's UNet while keeping the text encoder frozen for stability. Training data is loaded from captioned image files and automatically bucketed at 512, 768, and 1024 px. Captions use 5% dropout for better generalization, and VAE latents are cached to disk for faster training.
- **Training Configuration:** The model is trained for 2000 steps, batch size 1, using AdamW-8bit, gradient checkpointing and bf16 precision. A flow-matching noise scheduler matches FLUX's sampling behavior, and EMA (decay 0.99) stabilizes identity consistency throughout training.
- **Sampling & Checkpoints:** Validation samples (1024×1024) are generated every 250 steps using the flow-matching sampler and guidance scale of 4.

Advantages Over Other Solutions

- **Lightweight Personalization Module:** LoRA training keeps the base FLUX.1-dev model untouched while adding only a small rank-16 adapter, enabling fast training and compact checkpoints suitable for real workflows.
- **Identity-Faithful Generation from Minimal Data:** With just 12 reference images, the personalized LoRA reliably reproduces the user's appearance across diverse prompts—no large dataset or additional conditioning models are required.
- **Cost-Efficient Fine-Tuning:** The approach avoids the computational and financial overhead of full-model fine-tuning, reducing GPU memory usage and training time while maintaining high identity fidelity.



2.2 Workflow 1: AI-Enhanced Image Editing Tools

2.2.1 MagicQuill : Intent Aware Stroke-Based Editing

Introduction

Drawing on the brush-based interaction framework introduced in *MagicQuill: An Intelligent Interactive Image Editing System*, this system interprets user intent by transforming add, remove, and color strokes into structured masks and control signals for precise, region-specific editing. These strokes define where new structure should be inserted, where existing content should be removed, and which regions require color adjustments, enabling fine-grained manipulation of both geometry and appearance. Because the system operates directly on the semantics of stroke patterns, it can capture highly intricate editing intentions—**such as producing realistic facial wrinkles from subtle wavy lines drawn on the face**—and translate them into detailed structural modifications. Combined with the original image and the prompt inferred from stroke semantics, the editing module delivers results that faithfully reflect the user's intended changes, supported by **very fast inference** and **high-quality output** enabled by the underlying diffusion-based architecture.

Methodology

- **Stroke-Based Intent Extraction**: The user's brush interactions—including add, remove, and color strokes—are first converted into structured binary masks that encode localized geometric and appearance cues. These masks, together with the original image, are passed to a fine-tuned LLaVA-1.5 model, which is prompted using a "Draw & Guess" formulation to infer the user's underlying editing intention. This enables the system to interpret the semantic meaning of even abstract or partial strokes, producing a concise, context-aware prompt that guides the subsequent editing stage.
- **Intent-Aware Image Editing**: The inferred prompt, original image, and stroke-derived conditions are then fed into a fine-tuned Stable Diffusion 1.5 Editing Processor that follows MagicQuill's dual-branch architecture. An inpainting branch injects content-aware per-pixel features derived from the masked image, while a ControlNet-style control branch incorporates the stroke-generated edge and color conditions to impose precise structural and appearance constraints. By jointly leveraging these two complementary control pathways, the diffusion model synthesizes a final image that aligns closely with the user's intended modifications, achieving high spatial fidelity, fine-grained detail preservation, and rapid inference.

Advantages Over Other Solutions

- **Fine-Grained, Stroke-Level Control**: Unlike instruction-only or mask-only editing systems, the proposed approach enables highly localized manipulation through add, remove, and color strokes. These strokes translate into precise edge and color conditions, allowing users to control geometry, structure, and appearance at a much finer granularity. This enables intricate edits that traditional text-guided or coarse-mask systems cannot capture reliably.
- **Intent-Aware Editing via Multimodal Reasoning**: The integration of a fine-tuned LLaVA-1.5 model with the "Draw & Guess" formulation removes the need for repeated manual prompt engineering. By interpreting the semantics of brush interactions, the system generates contextually accurate intent prompts, resulting in more consistent and meaningful edits. This tight coupling between strokes and intent substantially improves usability compared to instruction-driven methods such as SmartEdit or MGIE.
- **High-Quality, Fast Inference through Dual-Branch Diffusion**: The Editing Processor's dual-branch architecture—combining a content-aware inpainting branch with a ControlNet-style structural conditioning branch—provides strong controllability without sacrificing output fidelity. This design yields edits that adhere closely to user guidance while maintaining high realism.

2.2.2 LEDITS++: High-Speed, High-Fidelity Image Editing on SD 1.5

Introduction

LEDITS++ is a lightweight, text-guided image editing framework designed to transform the capabilities of the original Stable Diffusion 1.5 model. Without requiring any fine-tuning or architectural changes, LEDITS++ enables SD1.5 to achieve edit quality comparable to larger diffusion systems while delivering significantly faster inference. This makes it highly suitable for real-time and on-device editing scenarios.

Methodology

The system is built around three core components. First, a DPM-Solver++ based "perfect inversion" reconstructs the input image with near-zero error in approximately 20 steps. This provides a mathematically accurate mapping from image

space to diffusion noise space, preventing the drift commonly observed in DDIM inversion and dramatically reducing computational overhead. Second, text-driven editing is performed using modified classifier-free guidance, where each editing instruction generates its own independent guidance vector. This allows fine-grained addition or removal of concepts and supports multiple simultaneous edits without cross-interference. Third, LEDITS++ uses implicit semantic masks derived from cross-attention maps and noise-difference maps. These masks automatically localize edits to meaningful regions, ensuring backgrounds, identities, and lighting remain intact.

Advantages over other Solutions

- **Real-Time Performance:** LEDITS++ achieves a 20–25× faster inversion compared to DDIM and eliminates all optimization or fine-tuning steps, reducing latency from minutes to milliseconds. Its mask-based regional editing updates only relevant areas, enabling near–real-time SD1.5 performance even on modest GPUs.
- **High Editing Accuracy:** Zero-drift inversion preserves identity and texture far more reliably than SD1.5 img2img, while semantic masking prevents global distortions. Independent guidance vectors avoid concept entanglement, offering more stable attribute edits than many larger diffusion models.
- **Efficiency Without Scaling:** Through algorithmic improvements rather than model size, LEDITS++ delivers edit quality approaching SDXL and SD 2.x, despite using a much smaller SD1.5 backbone—demonstrating that smarter computation can rival heavy diffusion systems.

2.2.3 Inpaint4Drag: Drag-Based Image Editing using Stable Diffusion 1.5

Introduction

Inpaint4Drag is a drag-based image editing framework that performs structural deformation directly in pixel space rather than in latent space. By separating geometric manipulation from content synthesis, the system warps existing pixels toward user-defined target positions and then inpaints only the newly exposed regions. Treating the editable area as a deformable elastic surface enables high-precision edits, real-time responsiveness, and compatibility with any inpainting model, including SD 1.5.

Methodology

- **Inputs:** A SAM-generated binary mask defines the editable region, and the user provides pairs of handle and target points.
- **Bidirectional Pixel-Space Warping:** The masked region is modeled as an elastic surface and warped using a bidirectional scheme. A forward warp pushes pixels toward target positions, while backward mapping fills any unmapped or stretched areas using valid source pixels. This combination produces a dense, hole-free, and stable geometric deformation that runs in approximately 10 ms, enabling real-time previews.
- **Intelligent Inpainting:** Regions that become exposed after warping are automatically detected and passed along with the warped image and binary mask to an external inpainting model. Because the system uses a standard inpainting interface, it remains fully model-agnostic and works seamlessly with diffusion and non-diffusion architectures.

Advantages

- **Editing Quality and Control:** Inpaint4Drag provides stable, precise point-based manipulation with smooth pose shifts, repositioning, and large structural edits. It achieves lower MD error on DragBench-S/D, competitive or improved LPIPS and SSIM scores, and minimal artifacts even under significant deformations.
- **Lightweight and Fast Performance:** The system remains mobile-friendly and plug-and-play with SD1.5, while offering real-time responsiveness: warping preview at 0.01 s, mask refinement at 0.02 s, and full editing with inpainting at 0.3 s for 512×512 resolution.

2.2.4 LightningDrag: Lightning-Fast Point-Based Image Manipulation

Introduction

LightningDrag is a rapid drag-based image editing approach that achieves high-quality, pixel-precise manipulation. Unlike traditional approaches that rely on time-consuming latent optimization or gradient-based guidance, LightningDrag redefines drag-based editing as a conditional generation task. Users simply select handle points, target points and define editable areas through masks. By training on large-scale video frames containing natural motion transformations, LightningDrag delivers superior accuracy and consistency while generalizing to complex deformations not seen during training.

Methodology

LightningDrag consists of three main technical components:

- **Conditional Generation Backbone:** LightningDrag uses an SD 1.5 Inpainting U-Net, a reference-only appearance encoder, and a 12-layer point-embedding network to maintain identity and enforce precise handle–target correspondence.
- **Video-Based Motion Learning:** Trained on 220k video samples with pose, scale, and translation changes, using optical-flow-based handle selection, CoTracker2 target tracking, and motion masks to learn realistic deformation behavior.
- **Stable Test-Time Refinements:** A noise-prior initialization and point-following CFG with inverse-square decay improve point accuracy and prevent oversaturation during denoising.

Advantages Over Other Solutions

- **Speed & Efficiency**: Compared to methods like DragDiffusion (55 seconds) or DragGAN (1-2 minutes for inversion alone), LightningDrag completes edits in 3 seconds with standard sampling, making it the only drag-based method practical for real-time mobile workflows.
- **Strong Generalization**: Unlike zero-shot methods lacking explicit supervision, LightningDrag learns from large-scale video motion patterns, enabling it to handle diverse scenarios including pose changes, object scaling, translation, and remarkably, local deformations not present in training data.
- **Accuracy & Consistency**: LightningDrag runs at constant speed regardless of dragging distance. On DragBench, LightningDrag achieves the lowest Mean Distance (18.62 vs. 28.5+ for competitors), indicating most effective dragging capability, while maintaining high Image Fidelity (0.885-0.890).

2.2.5 SAM ViT-H (Segment Anything Model): Universal Image Segmentation

Introduction

SAM is a segmentation system that produces high-quality object masks from minimal input cues such as points, boxes, or rough masks. Trained on the massive SA-1B dataset containing over 1 billion masks across 11 million images, SAM demonstrates remarkable zero-shot generalization to new image distributions and tasks.

Methodology

LEDITS++ integrates SAM through three lightweight components: a ViT-B image encoder that processes the image once to produce high-dimensional embeddings while maintaining an optimal 375 MB size for fast, consumer-grade inference; a prompt encoder that represents user-specified positive and negative point prompts using learned and positional embeddings; and a mask decoder that predicts precise segmentation masks by fusing image embeddings with prompt features via cross-attention.

Advantages Over Other Solutions

- **Strong Zero-Shot Precision:** Achieves 77–81% mIoU with accurate, prompt-driven masks that refine easily using simple point or box inputs.
- **High-Quality, Efficient Inference:** ViT-H (2.5B params) delivers state-of-the-art segmentation quality while remaining responsive for real-time editing workflows.

2.2.6 PowerPaint: Unified Diffusion-Based Region Editing using SD 1.5

Introduction

PowerPaint is a unified, instruction-guided diffusion framework for region-based image editing that supports erase, inpaint and outpaint operations within a single model. Instead of relying on separate networks or pipelines for each task, PowerPaint exposes a consistent interface: an input image, a mask, and a text instruction. The model then interprets the editing intent and synthesizes high-quality, context-aware results. By coupling diffusion-based generation with semantic conditioning, PowerPaint preserves global scene structure, maintains visual coherence, and enables both corrective and creative editing in a single system.

Methodology

- **Erase (Background-Preserving Inpainting)**: The masked region is treated as content to be removed and replaced with a clean background. The model encodes the visible portion of the image into a latent representation and uses local context (geometry, texture, and lighting) to regenerate only the missing background. The diffusion process

is constrained to preserve surrounding structures while avoiding the introduction of new objects. This minimizes artifacts such as cloning, repetition or haloing.

- **Inpaint (Instruction-Guided Content Synthesis)** : PowerPaint fills the masked region with new or reconstructed content according to the text instruction. The diffusion model conditions jointly on the unmasked image region and the instruction embedding.
  Cross-attention layers propagate semantic information across the full spatial grid, aligning the generated content with the existing scene layout and illumination. This allows both restoration-style edits and creative edits while keeping the rest of the image intact.
- **Outpaint (Canvas Expansion and Scene Extension):** PowerPaint first expands the canvas and defines the newly added borders as masked regions. The original image serves as a strong conditioning anchor, while the diffusion process synthesizes plausible continuations outside the initial bounds. The model maintains consistency in perspective, style, and lighting, and can optionally follow textual instructions to bias the extended region toward specific content.. This supports high-resolution scene expansion, reframing and context-aware composition.

Advantages Over Other Solutions

- **Unified and High-Quality Editing:** A single model handles removal, inpainting, and outpainting while achieving competitive LPIPS, SSIM, and PSNR scores for globally consistent, realistic edits.
- **Prompt-Driven, Mask-Aware Control:** Supports flexible text prompts and precise region selection, enabling intuitive and real-time editing workflows.
- **Efficient SD 1.5-Based Deployment:** Built on the Stable Diffusion 1.5 backbone and optimized for the IOPaint backend, offering lower memory usage and faster inference compared to heavier diffusion models.

2.2.7 Dynamic Style Transfer with SD1.5 (Semantic LoRA Selection + Prompt Enhancement using Moondream2)

Introduction

This system performs dynamic style-transfer using Stable Diffusion 1.5 by automatically selecting the most semantically appropriate LoRA based on the user's prompt. For the prototype stage, we employ a curated set of 13 stylistic LoRAs—including oil painting, retro game art, cosmic nebula, Studio Ghibli, pencil sketch, Toy-Story-style 3D cartoon, zombie, and other distinctive aesthetics—which serve as the initial style bank.

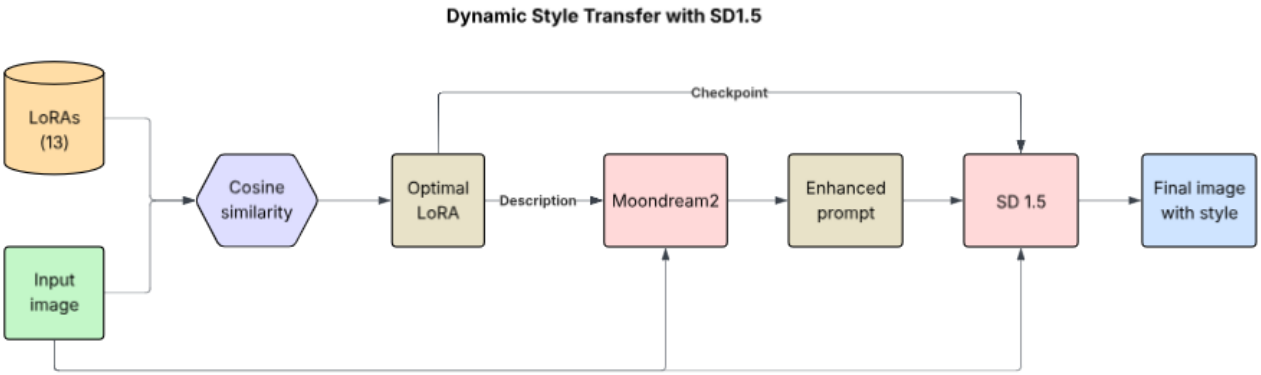
The system analyzes the user's input, identifies the closest stylistic match, and refines the chosen aesthetic using Moondream2 to generate an enhanced LoRA prompt, fully removing the need for manual LoRA selection. Although the current version uses 13 LoRAs for demonstration and evaluation, the architecture is inherently scalable and can be expanded to incorporate dozens or even hundreds of new styles, enabling continuous updates aligned with evolving artistic trends and user preferences.

Methodology

Each LoRA's name, description, and trigger words are embedded using a SentenceTransformer encoder and cached to form a semantic style index. When the user provides a style prompt, it is encoded into the same embedding space, and cosine-similarity search with LoRA descriptions identifies the most relevant one. The selected LoRA is dynamically injected into the Stable Diffusion Img2Img pipeline without reloading the base model, ensuring efficient GPU usage. Input images are preprocessed, prompts are augmented with LoRA-specific metadata and controlled Img2Img generation produces the final stylized output.

Advantages Over Other Solutions

- **Automatic Style Retrieval:** Semantic similarity search selects the most relevant LoRA automatically, removing manual style selection.
- **Fast, Memory-Efficient Switching:** Dynamic LoRA injection enables rapid multi-style generation without reloading the base model, enabling smooth multi-style workflows on consumer GPUs.
- **Improved Stylization Quality:** Moondream2 enhances prompts to better match each style, boosting consistency and reducing prompt-engineering effort.
- **Lightweight, Local Execution:** Built on SD1.5, compact LoRA adapters and Moondream2, the pipeline runs efficiently on consumer GPUs with low VRAM usage.



2.2.8 stabilityai/stable-diffusion-x4-upscaler : Latent Diffusion Model

Introduction

A central limitation of conventional diffusion-based image generators is operating directly in pixel space is computationally expensive, requiring extremely large models and long training cycles. To overcome this, stable-diffusion-x4-upscaler shifts the diffusion process from full-resolution pixel space into a compressed latent space learned by a pretrained autoencoder. This reduces memory and compute requirements while still preserving high-quality visual detail.

Methodology

The latent diffusion pipeline consists of three primary components:

- **Latent VAE Compression:** A variational autoencoder encodes high-resolution images into a compact latent space, retaining semantic and structural information while removing redundancy for efficient denoising.
- **Latent Diffusion via Time-Conditional U-Net:** A U-Net–based denoiser operates directly in latent space, progressively removing noise across timesteps with spatial attention and a hierarchical structure for global–local feature modeling.
- **Cross-Attention Conditioning:** Cross-attention layers inject text, mask, spatial, or style signals into intermediate features, enabling controlled generation. After denoising, the VAE decoder reconstructs the final high-resolution output.

Advantages Over Other Solutions

- **Efficient Latent-Space Generation:** LDMs compress images by 3–12× in spatial resolution, cutting training cost by 2–4× and reducing sampling time by nearly 75% while maintaining strong generative fidelity.
- **High-Quality Outputs:** Despite operating in latent space, the model achieves excellent realism with FID ≈ 4.98 on LSUN Church and FID ≈ 7.23 on CelebA-HQ, matching or outperforming heavier pixel-space diffusion models.
- **Flexible Cross-Attention Conditioning:** Conditioning through text, masks, or semantic cues improves alignment and boosts CLIP-consistency scores by 8–12%, enabling high-resolution, detail-preserving upscaling.

2.2.9 Hybrid CLIP + Moondream2 Defect Analysis System

Introduction

This system performs automated photographic defect analysis by combining CLIP's zero-shot image–text matching with Moondream2's vision-language reasoning. A curated dataset of 101 possible visual defects—including lighting errors, color casts, blur, anatomical distortions, facial imperfections and AI-generation flaws—forms the diagnostic vocabulary. The pipeline identifies the top three most likely defects in an input image using CLIP, then sends them to Moondream2 for verification and detailed explanation.

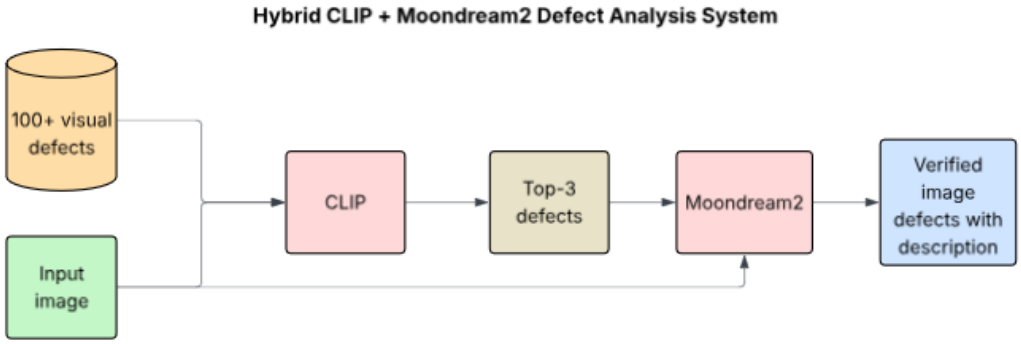
Methodology

- **CLIP-Based Defect Retrieval**: The input image is encoded using CLIP (ViT-B/32), and all 101 defect descriptions are tokenized and embedded. Cosine similarity is computed to rank defect likelihood, and the top-3 highest scoring defects are selected as candidates.
- **Moondream2 Verification & Explanation:** The shortlisted defects are passed, along with the image, to the Moondream2 VLM. The model confirms whether each defect truly appears in the image and provides a natural-language

explanation of what it sees. This hybrid approach ensures both semantic retrieval accuracy (via CLIP) and grounded visual validation (via Moondream2).

Advantages Over Other Solutions

- **High Accuracy with Fewer False Positives:** CLIP retrieves the most likely defects from over 100 categories, while Moondream2 verifies each one, ensuring reliable and context-aware detection.
- **Explainable Defect Analysis:** Moondream2 provides clear, grounded explanations of confirmed defects, producing editor-grade descriptions instead of simple labels.
- **Scalable and Fully Local:** The hybrid pipeline runs efficiently on-device and scales to large defect vocabularies, enabling automated, privacy-preserving quality control for photography and AI-generated images.



2.3 Workflow 2 : Smart Composition and 3D-Aware Object Insertion

2.3.1 A2-RL: Aesthetics Aware Reinforcement Learning for Smart Cropping

Introduction

A2-RL is a weakly supervised image cropping method that formulates cropping as a sequential decision-making process rather than brute-force sliding window search. Unlike traditional methods, A2-RL uses a reinforcement learning agent that intelligently navigates the image space through deliberate actions. Starting from the full image, the agent iteratively adjusts the cropping window's size, position, and aspect ratio until it identifies the most aesthetically pleasing composition.

Methodology

A2-RL models image cropping as a sequential decision-making problem, treating it as a Markov decision process where the agent chooses among 14 actions, including five zoom operations, four translations, four aspect-ratio adjustments, and a termination action. Its aesthetics-aware reward function penalizes aspect ratios outside the 0.5–2.0 range to discourage compositions that violate photographic norms. The model uses a 5-layer CNN (shared with the VFN aesthetic scorer) and an actor–critic architecture with three fully connected layers and an LSTM module for temporal reasoning.

Advantages Over Other Solutions

- **Superior Accuracy:** A2-RL outperforms both sliding-window and learning-based baselines, achieving state-of-the-art IoU scores, even surpassing some fully supervised models that rely on costly bounding-box annotations.
- **Major Speedup Over Classical Methods:** Delivers a 5–40× speed improvement with higher accuracy.

2.3.2 BRIAAI / RMBG-2.0: High-Resolution Background Removal

Introduction

RMBG-2.0 is a state-of-the-art background-removal (foreground segmentation) model developed by BRIA AI, built on top of the BiRefNet architecture for high-resolution dichotomous image segmentation (DIS). BiRefNet combines global semantic understanding with fine-grained detail reconstruction through its bilateral-reference mechanism, enabling precise separation of foreground objects from complex backgrounds.

Methodology

RMBG-2.0 adopts the two-stage BiRefNet architecture, where a transformer-based Global Localization Module first identifies coarse foreground regions using global semantic cues, even in cluttered scenes. This is followed by a Fine Reconstruction Module that refines boundaries through a bilateral-reference mechanism, combining Source Reference and Target Reference signals to recover thin structures, hair strands, soft edges, and intricate textures—achieving far more precise contour reconstruction than conventional segmentation models.

Advantages Over Other Solutions

- **High-Precision Detail Recovery:** Preserves fine structures and handles complex backgrounds with production-grade accuracy.
- **Efficiency:** Maintains clean boundaries on large images while remaining computationally lightweight.

2.3.3 Stable Fast 3D: Single-Image Feed-Forward 3D Reconstruction

Introduction

SF3D reconstructs a complete, textured 3D mesh from a single 2D image. It provides an ultra-fast feed-forward where given one image, it outputs a fully textured mesh—complete with geometry, UV mapping, materials, and normal maps.

Methodology

SF3D combines neural reconstruction with efficient mesh and texture generation:

- **Transformer-Based 3D Prediction** : A transformer network infers 3D structure, material parameters (roughness, metallicity), surface normals, and a latent code for texture generation directly from the input image.
- **Illumination Disentanglement / Delighting :** The method estimates scene lighting baked into the source image and removes it, producing a neutral, albedo-like texture suitable for realistic relighting.
- **Mesh Extraction & UV Mapping**: A differentiable mesh extraction algorithm produces the 3D shape, followed by fast UV unwrapping to generate a proper texture atlas. Textures are then baked onto UV space instead of simple vertex colors, enabling high-fidelity appearance reproduction.
- **Material & Normal Map Generation** : SF3D outputs physically meaningful materials and normal maps, allowing the final mesh to respond correctly to new lighting conditions in downstream renderers.

Advantages Over Other Solutions

- **Real-Time Single-Image Reconstruction:** SF3D produces a fully textured mesh in $\sim$0.5 s—far faster than multi-view or optimization-heavy pipelines that require seconds to minutes.
- **High-Fidelity Geometry and Textures:** The model generates detailed meshes with clean UV atlases, sharp textures, and realistic material properties.
- **Efficient and Production-Ready Output:** UV unwrapping and texture baking are artifact-free and compatible with real-world rendering engines, unlike many neural field–based 3D models that produce noisy, non-rigged, or non-exportable outputs.

2.3.4 LBM Relighting: Single-Step Latent-Space Illumination Transfer

Introduction

Latent Bridge Matching (LBM) provides a fast, one-step solution for image relighting by performing illumination transformation entirely in latent space. Conventional relighting approaches rely on multi-step diffusion or optimization, making them slow and prone to structural drift. LBM reframes relighting as a latent-space transport problem: the model adjusts illumination while preserving geometry, texture, and object boundaries.

Methodology

LBM embeds the input image into a VAE latent space and constructs a stochastic "bridge" between the source illumination and the target lighting condition. Lighting parameters—such as direction, intensity, or environment maps—condition the target latent distribution. A neural denoiser (U-Net) is trained to approximate the drift of this bridge, enabling a direct single-step mapping from the source latent to the target latent without iterative denoising. The decoded output exhibits coherent modifications to shading, highlights, cast shadows, and global illumination while keeping scene structure intact.

Advantages Over Other Solutions

- **Real-Time Relighting:** LBM replaces multi-step diffusion with single-step latent transport, achieving much faster relighting while matching or exceeding the quality of iterative diffusion models.
- **Stable, Realistic Illumination:** It produces smooth shading, consistent shadows, and reliable material appearance across large lighting changes, enabling interactive, low-cost relighting workflows.

2.4 Ethical Considerations

2.4.1 InvisMark : Invisible and Robust Watermarking for AI-generated Image Provenance

Introduction

InvisMark is a watermarking system designed to ensure reliable provenance tracking for AI-generated images. Unlike traditional watermarks that are fragile or detectable, InvisMark embeds an invisible, high-capacity watermark that remains intact even after common image manipulations, addressing growing concerns around deepfakes and content authenticity. It embeds a large 256-bit watermark with no visible artifacts, maintaining extremely high image fidelity even on modern high-resolution AI images.

Methodology

InvisMark embeds a binary watermark using a neural encoder that adds a subtle residual to the image. A paired decoder extracts this watermark. During training, a robustness module applies distortions such as JPEG compression, blur, noise, cropping, and color shifts, encouraging the system to withstand real-world transformations. The loss function balances perceptual quality with extraction accuracy, enabling high-resolution, imperceptible watermarking.

2.4.2 NSFW Content Guardrails using CLIP

Introduction

We designed NSFW Content Guardrails, a fast and adaptive CLIP-based safety screening system that detects a broad spectrum of harmful or sensitive visual content. Unlike traditional moderation pipelines—often limited to sexual-content detection—our approach covers violence, extremism, weapons, drugs, psychological risk, and over 130 nuanced safety categories. By leveraging CLIP's shared image–text embedding space, the system provides far richer coverage with significantly lower computational cost, enabling transparent, customizable, real-time safeguarding for AI-assisted image editing workflows.

Methodology

The system uses CLIP ViT-B/32 to embed images and safety-taxonomy labels into the same semantic space. Incoming images are preprocessed, encoded once and matched against all labels using cosine similarity. The top-3 risk categories are recorded and marked as violations if their similarity exceeds a defined threshold (=0.2), creating a precise and transparent rule-based moderation pipeline. It is 10–50× faster than VLM-based moderation, requires minimal GPU resources, and is easily customizable or expandable.

3 Future Scopes

3.1 Autonomous System

The current system comprises multiple editing modules, each capable of performing a specific image manipulation task. These modules can be integrated into a fully autonomous pipeline through the use of a Large Language Model (LLM)–based orchestrator. Although such agentic systems can be constructed using existing state-of-the-art LLMs, the models that are lightweight enough for deployment on mobile devices still lack the capability to perform complex reasoning processes, such as multi-step decision-making or chain-of-thought planning. As a result, achieving reliable on-device autonomy remains a challenge. Future advancements in compact LLM architectures may enable the development of an end-to-end, on-device agentic editing system.

3.2 Integration of Next-Generation Vision Models

Several upcoming models, such as Chronos-2 and Z-Image, are expected to deliver significant improvements in image generation and editing quality. Once released, these models can serve as stronger backbones for the system, potentially enhancing both performance and efficiency across multiple editing modules. Integrating such advanced architectures may substantially elevate the overall capability and user experience of the application.

3.3 Time-Based Diffusion and Video Editing

Future iterations of the system can integrate temporal diffusion models as an extension of the existing image-editing pipeline. The current modular framework already supports task-specific image processors, and a video module can be added by applying these operations frame-wise while enforcing temporal consistency through a lightweight temporal backbone. This allows the system to reuse existing components—such as segmentation, inpainting, and style transfer—while enabling coherent video editing without major architectural changes.

3.4 Automated Story-to-Manga Generation

A potential future enhancement is the integration of a story-driven manga generation pipeline, where the user provides an initial prompt and an LLM autonomously expands the narrative into scenes while generating corresponding image panels through diffusion models. However, constructing such a system is currently challenging due to limitations in lightweight LLMs, which struggle with long-horizon narrative planning, maintaining character consistency, and coordinating multi-image generation under strict resource constraints on mobile devices. Future advances in compact multimodal models, long-context reasoning, and efficient image-generation backbones may overcome these barriers, making coherent story-to-manga generation viable within the existing product framework.

References

[1] Black Forest Labs. 2024. FLUX.1-dev. Hugging Face. https://huggingface.co/black-forest-labs/FLUX.1-dev

[2] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. 2024. SF3D: Stable Fast 3D Mesh Reconstruction with UV-unwrapping and Illumination Disentanglement. arXiv:2408.00653 [cs.CV] https://arxiv.org/abs/2408.00653

[3] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. 2023. LEDITS++: Limitless Image Editing using Text-to-Image Models. arXiv:2311.16711 [cs.CV] https://arxiv.org/abs/2311.16711

[4] BRIA AI. 2024. BRIA RMBG-2.0: Background Removal Model. Hugging Face. https://huggingface.co/briaai/RMBG-2.0

[5] Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. 2025. LBM: Latent Bridge Matching for Fast Image-to-Image Translation. arXiv:2503.07535 [cs.CV] https://arxiv.org/abs/2503.07535

[6] CivitAI Community. 2024. LoRA Models Database. CivitAI. https://civitai.com

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV] https://arxiv.org/abs/2304.02643

[8] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. 2017. A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping. arXiv:1709.04595 [cs.CV] https://arxiv.org/abs/1709.04595

[9] Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. 2024. SVDQuant: Absorbing Outliers by Low-Rank Components for 4-Bit Diffusion Models. arXiv:2411.05007 [cs.CV] https://arxiv.org/abs/2411.05007

[10] Jingyi Lu and Kai Han. 2025. Inpaint4Drag: Repurposing Inpainting Models for Drag-Based Image Editing via Bidirectional Warping. arXiv:2509.04582 [cs.CV] https://arxiv.org/abs/2509.04582

[11] MIT HAN Lab and Nunchaku Team. 2024. Nunchaku: 4-Bit Diffusion Model Inference. GitHub. https://github.com/nunchaku-tech/nunchaku

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

[13] Sanster. 2024. PowerPaint v2: High-Quality Inpainting and Outpainting. Hugging Face. https://huggingface.co/Sanster/PowerPaint_v2

[14] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent Y. F. Tan, and Jiashi Feng. 2024. LightningDrag: Lightning Fast and Accurate Drag-based Image Editing Emerging from Videos. arXiv:2405.13722 [cs.CV] https://arxiv.org/abs/2405.13722

[15] Stability AI. 2022. Stable Diffusion x4 Upscaler. Hugging Face. https://huggingface.co/stabilityai/stable-diffusion-x4-upscaler

[16] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. 2024. MagicQuill: An Intelligent Interactive Image Editing System. arXiv:2411.09703 [cs.CV] https://arxiv.org/abs/2411.09703

[17] Vikhyat K. 2024. Moondream2: A Tiny Vision Language Model. Hugging Face. https://huggingface.co/vikhyatk/moondream2

[18] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. 2024. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers. arXiv:2410.10629 [cs.CV] https://arxiv.org/abs/2410.10629

[19] Rui Xu, Mengya Hu, Deren Lei, Yaxi Li, David Lowe, Alex Gorevski, Mingyu Wang, Emily Ching, and Alex Deng. 2024. InvisMark: Invisible and Robust Watermarking for AI-generated Image Provenance. arXiv:2411.07795 [cs.CV] https://arxiv.org/abs/2411.07795