

MagicQuill V2: Precise and Interactive Image Editing with Layered Visual Cues

Zichen Liu^{♡,1,2}, Yue Yu^{♡,1,2}, Hao Ouyang², Qiuyu Wang², Shuailei Ma^{3,2}, Ka Leong Cheng², Wen Wang^{4,2}, Qingyan Bai^{1,2}, Yuxuan Zhang⁵, Yanhong Zeng², Yixuan Li^{2,5}, Xing Zhu², Yujun Shen^{†,2}, Qifeng Chen^{†,1}

¹HKUST, ²Ant Group, ³NEU, ⁴ZJU, ⁵CUHK

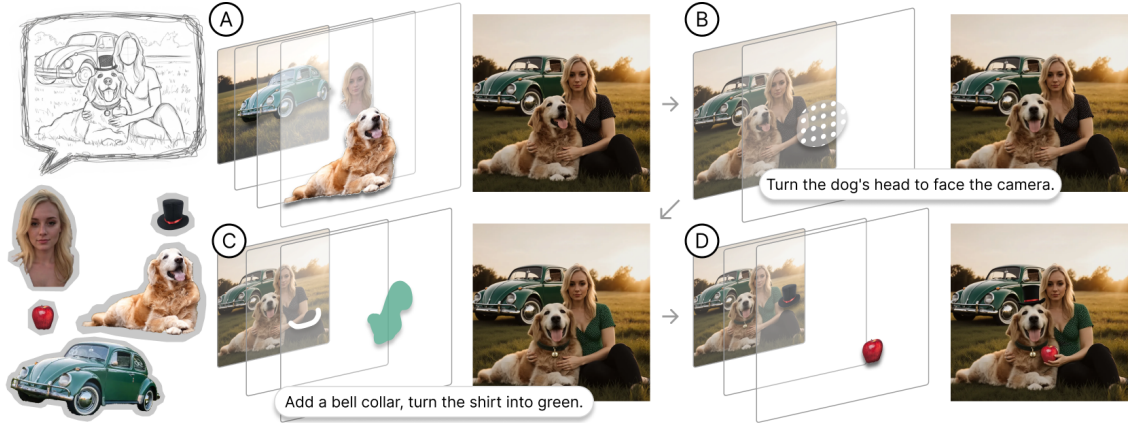


Figure 1. MagicQuill V2 introduces a layered composition framework for precise generative image editing. Users articulate complex intents by stacking independent visual layers in a continuous workflow: (A) composing a base scene with multiple **content layers** (car, lady, dog); (B) restructuring the dog’s pose with a **spatial layer**; (C) modifying visual attributes by sketching a bell collar on the **structural layer** and painting the shirt on the **color layer**; (D) inserting final details like a hat and apple seamlessly. Demo available at [project page](#).

Abstract

We propose *MagicQuill V2*, a novel system that introduces a **layered composition** paradigm to generative image editing, bridging the gap between the semantic power of diffusion models and the granular control of traditional graphics software. While diffusion transformers excel at holistic generation, their use of singular, monolithic prompts fails to disentangle distinct user intentions for content, position, and appearance. To overcome this, our method deconstructs creative intent into a stack of controllable visual cues: a content layer for what to create, a spatial layer for where to place it, a structural layer for how it is shaped, and a color layer for its palette. Our technical contributions include a specialized data generation pipeline for context-aware content integration, a unified control module to process all visual cues, and a fine-tuned spatial branch for precise local editing, including object removal. Extensive experiments validate that this layered approach effectively resolves the user intention gap, granting creators direct, intuitive control over the generative process.

1. Introduction

The advent of diffusion transformers has ushered in a new era of powerful image editing systems, such as the FLUX series [2], Qwen-Image [70], GPT-4o [45], and Nano Banana (Gemini 2.5 Flash Image) [16]. These models exhibit remarkable proficiency in executing complex edits based on natural language or reference images.

However, this reliance on high-level, holistic inputs presents a considerable barrier when users require precise control over the fundamental components of an image. As illustrated in Fig. 1, consider the nuanced task of creating a complex scene: a user starts by composing foundational elements like a vintage car, a person, and a dog (A). Then, the user may want to precisely adjust specific details, such as turning the dog’s head to face the camera (B), modifying attributes like shirt color (C) and inserting fine details like a top hat or an apple (D). Existing conversation-based systems struggle to disentangle these intertwined commands from a single prompt, as they often sacrifice the spatial precision in traditional tools (e.g., Photoshop, GIMP). After all, a user’s creative intent is rarely monolithic; it is a composite of distinct desires concerning *what* to create, *where* to place it, *how* it should be structured, and with *what colors*

[♡]Equal contribution. [†]Corresponding author.

it should be painted.

In this work, we introduce a novel framework built upon the principle of **layered composition**, drawing from the robust and intuitive workflows of professional graphics software. Our approach deconstructs the editing process by mapping each fundamental component of visual creation to its own dedicated, controllable layer. This allows a user to articulate their intent not as an ambiguous command, but as a stack of precise **visual cues**: a **content layer** specified by a foreground reference image to control what appears; a **spatial layer** defined by a mask to control where the editing appears; a **structural layer** guided by an edge map to dictate how it is shaped; and a **color layer** applied with strokes to determine its palette. Within our interactive system, these layers are not static inputs but dynamic elements that can be independently composed, repositioned, and refined. By uniting the generative prowess of diffusion models with the granular control of a layered interface, our approach offers a powerful yet intuitive solution that bridges the gap between semantic convenience and spatial precision.

We build our system upon the strong image editing model FLUX Kontext [2]. To effectively implement our layered design, we introduce several key contributions. First, we propose a specialized data generation pipeline to train the content layer (foreground pieces) for seamless, context-aware content integration, rather than a simple copy-paste operation. Second, we introduce a unified control module designed to process all control cues, including the spatial layer (mask), the structural layer (edge maps), and the color layer (color maps). The control module precisely constrains the geometry and palette accordingly. We additionally fine-tune the spatial layer branch on object removal data, enabling a dedicated and precise removal function. Finally, we design and implement a cohesive, interactive system that enables users to manage and compose these distinct layers intuitively. This system, augmented with tools like SAM [30] for effortless cue creation, realizes a precise, layer-like editing workflow that provides users with fine-grained control and creative freedom.

We demonstrate through extensive experiments that our method (MagicQuill V2) significantly outperforms existing approaches in editing that demands high precision. Our work validates that this layered composition approach effectively bridges the intention gap, granting users direct and powerful control over the generative process.

2. Related Works

2.1. Controllable Image Generation and Editing

While initial text prompts offered creative scope, their inherent ambiguity limited spatial precision. Early pioneering works like ControlNet [76] and T2I-Adapter [43] augmented UNet-based [54, 55] architectures to accept spa-

tial conditions such as Canny edges, depth maps, and human pose, marking a significant step towards granular control. With the architectural shift towards more powerful Diffusion Transformers [46], new control paradigms emerged [3, 4, 28, 42, 64, 65, 79]. OminiControl [64, 65] introduced a parameter-efficient, unified framework by concatenating text, noise, and condition tokens into a single sequence for joint processing. In contrast, EasyControl [79] proposed a modular approach, processing each condition in an isolated branch to enhance plug-and-play compatibility.

Despite their architectural advance, these state-of-the-art methods, and even more recent multi-modal models that accept multiple reference images like Qwen-Image [70], Step1X-Edit [37], GPT-4o [45], Nano-Banana [15], and Seedream [57], share a fundamental limitation. Even when a reference image is provided, its impact on the editing process is governed by the text prompt. This paradigm inherits the natural imprecision of language, as text prompts are inherently inefficient for specifying the exact where, what, and how of a targeted spatial edit. MagicQuill V2 challenges this paradigm by introducing precise visual cues, such as foreground pieces or a sketched edge map, that serve as a direct complement to the textual prompt. These cues provide the specific spatial and appearance information that text cannot, eliminating the inferential gap and creating a direct channel between user intent and generative output.

2.2. Layered Image Synthesis

Layered image synthesis encompasses both layer decomposition [14, 18, 24, 25, 29, 51, 68, 74, 75, 78] (analyzing an image into layers) and layer composition (assembling generative elements), which our work aligns with. This contrasts with the related subtask of image harmonization [6, 10–13, 21, 22, 27, 36, 48, 49, 58, 63, 66, 67, 80], which assumes a pre-existing, static foreground to be adjusted to the scene. In our setting, the “foreground” is not a fixed layer but a generative cue that guides what the model should synthesize, allowing for flexible manipulation and interaction with the context through generative instructions.

Image composition is a broader task that involves inserting objects into a scene while preserving their identity [44]. Representative works include Paint-by-Example [73], TF-ICON [39], AnyDoor [8], MimicBrush [7], Imprint [60], Insert-Anything [59], UniReal [9] and IC-Custom [34]. Although these methods allow users to specify a layout or mask to control where an object should appear, the final appearance is still inferred from a holistic reference image, which constrains fine-grained control. Our approach differs fundamentally: by enabling users to provide pieces of the foreground as direct cues, we offer a more precise channel to guide the editing process.

2.3. Interactive Support for Expressing User Intent

A significant challenge in generative editing is bridging the “intention gap”, the discrepancy between a user’s precise mental image and the ambiguous modalities available to describe it. While early solutions relied on autonomously suggesting or refining prompts via Multimodal Large Language Models (MLLMs) [19, 38], recent Human-AI interaction research has shifted toward more direct, expressive paradigms that allow users to express intent through interactive manipulation rather than just description.

To overcome the linearity of chat-based prompting, some approaches propose tools like Fillable Brushes to reify abstract user intents into direct and interactive graphical “instruments.” [53] Some other approaches, like Sketch-Flex [35], focus on addressing the barrier of artistic expertise, leveraging MLLMs not just for interpretation, but to actively refine novice users’ rough sketches into coherent spatial conditions. Most relevant to our work are approaches that decouple visual attributes for more granular control. FusAIn [47] introduces the concept of “smart pens” that allow designers to explicitly extract and apply specific visual attributes such as object identity, color, or texture from inspiration images.

MagicQuill V2 formalizes these interactive concepts into a layered editing framework for professional precision. Instead of interpreting rough sketches or using abstract instruments, we map fundamental visual cues to explicit layers, allowing users to dynamically compose and refine the entire generative process in real-time.

3. Methodology

Built upon FLUX Kontext [2], our methodology is designed to realize the layered composition framework introduced previously. Our goal is to develop a system that generates a target image x conditioned jointly on a context image y , a natural language instruction c , and a stack of precise layered visual cues L . These cues are mapped directly to the fundamental components of visual creation. Formally, we aim to approximate the conditional distribution:

$$p(x|y, c, \{L_{fg}, L_{control}\}). \quad (1)$$

Here, the conditioning cues are explicitly divided to match our layered framework: The **content layer** (L_{fg}), defining what to create, is specified by one or more foreground pieces F_i . The **control layers** ($L_{control}$) provide explicit control defining where, how, and with what colors to edit. This group consists of **spatial layer** (mask M) for targeted regional editing, **structural layer** (edge map E) for precise geometric guidance, and **color layer** (color map C) for exact color control.

This section details the construction of MagicQuill V2. First, we introduce our data generation pipeline to enable

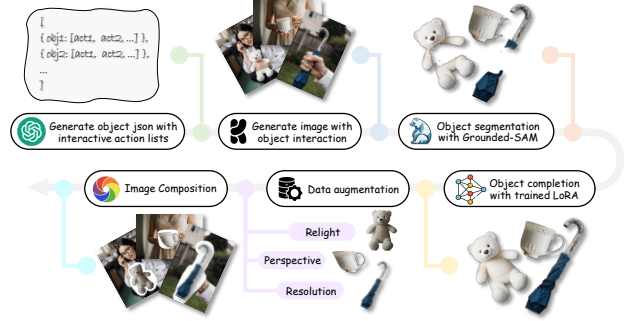


Figure 2. Overview of our Data Construction Pipeline. We start by synthesizing images depicting interactions. From these, we extract foreground objects, restore occluded items using a trained object completion LoRA, apply a series of augmentations (Relight, Perspective, Resolution), and finally composite the augmented object back into the scene to create the training data.

the content layer (Sec. 3.1). Second, we detail our unified control architecture for the control layers (Sec. 3.2). Finally, we present the interactive system that unifies these layers into an intuitive, precise editing tool (Sec. 3.3).

3.1. Enable Content Layer via Foreground Cues

The content layer (L_{fg}), which defines what to create, consists of one or more user-provided foreground pieces (F_i) that function as content anchors and an optional mask to specify editing regions. The core task is to execute the edit by seamlessly integrating the content and generating the surrounding context, rather than a simple copy-paste operation. We enable this capability through a novel data generation pipeline.

Our pipeline begins with the synthesis of a base dataset of 5,000 images depicting diverse object-scene interactions. We employ Qwen3-8B [72] to generate descriptive captions (c) and use Flux.1 Krea [31] to produce high-resolution, photorealistic source images. For each source image, we utilize Grounding SAM [52] to extract the primary object’s mask, yielding our initial set of foreground pieces and their corresponding source images.

A key challenge is that extracted foregrounds are frequently incomplete due to occlusion (e.g., a hand covering part of an apple). Training on such data would lead to simple copy-paste behavior rather than contextual understanding. To mitigate this, we trained a dedicated object completion model, a LoRA [23] fine-tuned on FLUX Kontext [2]. This model was trained on a dataset of 3,000 complete objects by learning to restore them from randomly applied brushstroke masks. By applying our extracted foregrounds, this LoRA produces a complete and whole version of the foreground object from the previously occluded item.

To ensure robustness against the diverse quality of user-provided inputs, we create what we term the augmented foreground object through a series of augmentations. First,

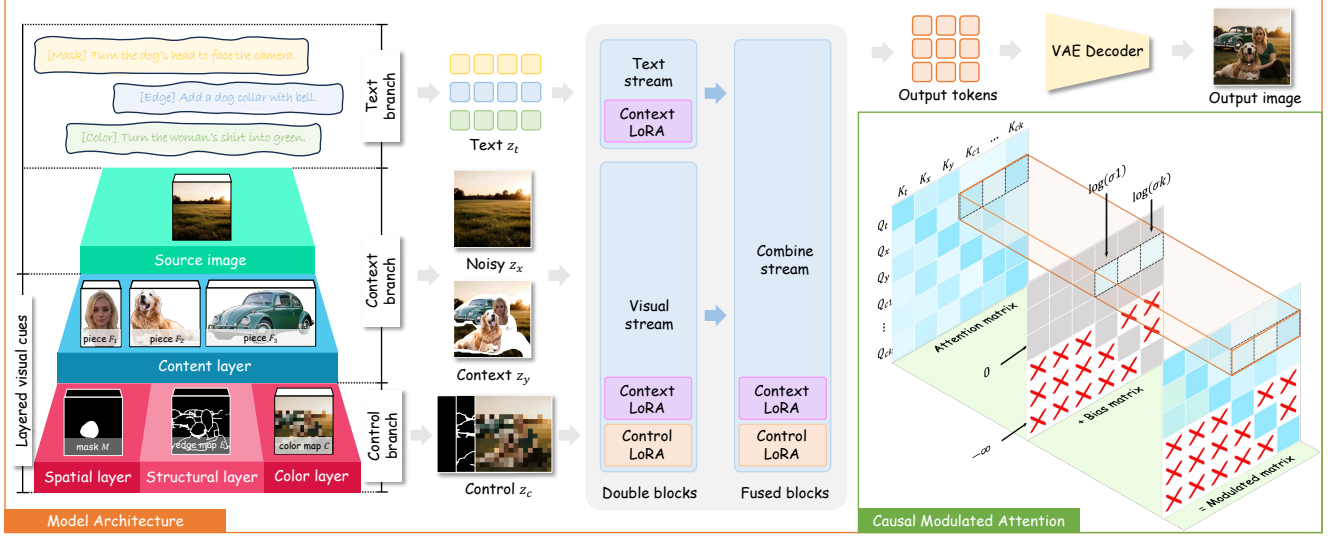


Figure 3. **Overview of MagicQuill V2 Model Architecture.** Our model processes layered visual cues and text instructions through distinct branches. The layered visual cues (left) are divided into a content layer (L_{fg}) and multiple control layers ($L_{control}$: spatial, structural, color), which are encoded into latents (Z_y , Z_c). These are processed alongside text (Z_t) and noisy image (Z_x) latents in a unified control module (middle) adapted with dedicated control modules. A Causal Modulated Attention mechanism applies a bias matrix to the attention logits to precisely manage the influence and isolation of each control cue.

to address photometric inconsistencies, we perform photometric augmentation by applying random lightmaps with ICLight [77], forcing the model to learn lighting harmonization. Second, to simulate inputs of varying detail, resolution augmentation involves randomly downsampling and resizing the object. Third, to account for geometric mismatches, perspective augmentation applies a random perspective transformation, introducing plausible distortions.

In the final stage, we assemble the training triplets. The target (x) is the unmodified source image, and (c) is its descriptive caption. The input image (y) is constructed by first compositing the augmented foreground object back into its original position, and then applying random mask augmentations to the surrounding background region, forcing the model to learn contextual harmonization.

To integrate this new capability, we fine-tune the FLUX Kontext [2] backbone using a Low-Rank Adaptation (LoRA) [23] adapter over the attention layer. The model is trained on the triplets (y , c , x) generated by our pipeline, optimizing the rectified-flow objective shown below:

$$\mathcal{L}_\theta = \mathbb{E}_{t \sim p(t), x, y, c} [\|v_\theta(z_t, t, y, c) - (\epsilon - x)\|_2^2], \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $z_t = (1 - t)x + t\epsilon$.

3.2. Unified Control Module for Control Layers

Control module. To seamlessly integrate the diverse cues comprising $L_{control}$, i.e., spatial layer (mask M), structural layer (edge map E), and color layer (color map C), we introduce a minimal and unified control module.

For computational efficiency, we resize all visual control cues from their original size (H, W) to a fixed low resolution (h, w). To maintain correct spatial alignment, the positional encoding for a resized patch at grid (i, j) is mapped to its original high-resolution coordinates (P_i, P_j) via $P_i = i \cdot \frac{H}{h}$, $P_j = j \cdot \frac{W}{w}$.

Consider the input representations for different branches: text embedding (Z_t), noisy image latent (Z_x), and context image latent (Z_y). In the Multi-Modal Diffusion Transformer (MMDiT) [2, 32], the Query-Key-Value (QKV) transformation for image-related branches is defined as:

$$Q_i, K_i, V_i = W_Q Z_i, W_K Z_i, W_V Z_i, \quad i \in \{x, y\} \quad (3)$$

where W_Q, W_K, W_V are shared projection matrices.

To incorporate the visual cue latents Z_c , we introduce a dedicated conditional branch adapted using Low-Rank Adaptation (LoRA) [23]. The updated QKV features for the Condition Branch are computed by adding the LoRA update directly to the standard projection:

$$Q_c = W_Q Z_c + B_Q A_Q Z_c, \quad (4)$$

$$K_c = W_K Z_c + B_K A_K Z_c, \quad (5)$$

$$V_c = W_V Z_c + B_V A_V Z_c, \quad (6)$$

where A and B are the low-rank decomposition matrices with rank $r \ll d$. The final Query, Key, and Value for Multi-

Modal Attention (MMA) can be expressed as:

$$\mathbf{Q} = [Q_t; Q_x; Q_y; Q_c], \quad (7)$$

$$\mathbf{K} = [K_t; K_x; K_y; K_c], \quad (8)$$

$$\mathbf{V} = [V_t; V_x; V_y; V_c]. \quad (9)$$

To modulate the influence of visual cues on the denoising process, we directly manipulate the attention scores by adding a custom bias matrix, $\mathbf{B}$, to the attention logits:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B} \right) \mathbf{V}. \quad (10)$$

This bias matrix $\mathbf{B}$ is designed to selectively manage information flow. Let $\mathcal{I}_t, \mathcal{I}_x, \mathcal{I}_y, \mathcal{I}_{c_k}$ be the token index sets for the text Z_t , noisy image Z_x , context image Z_y , and the k -th visual cue Z_{c_k} . The bias entry B_{ij} is defined as:

$$B_{ij} = \begin{cases} \log(\sigma_k) & \text{if } i \in \mathcal{I}_x \text{ and } j \in \mathcal{I}_{c_k}, \\ -\infty & \text{if } i \in \mathcal{I}_{c_k}, j \notin \mathcal{I}_{c_k}, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

The $\log(\sigma_k)$ term acts as a user-adjustable guidance scale, where $\sigma_k \geq 0$ is a scalar parameter modulating the attention from the noisy image latents Z_x to the k -th cue Z_{c_k} . When $\sigma_k = 1$, the bias is 0, resulting in the standard, unbiased attention mechanism. As σ_k increases, the positive bias $\log(\sigma_k)$ strengthens the cue’s influence, forcing stricter adherence. Conversely, setting $\sigma_k = 0$ makes the bias $-\infty$, effectively disabling the cue. Simultaneously, the second case in Eq. 11 sets the bias to $-\infty$ for any attention between different control signals to prevent interference.

Training. We train three distinct control branches for the spatial (M), structural (E), and color (C) layers, which are designed to be used independently or cooperatively. For the structural and color layer (E, C), we train the model to generate images from noise conditioned on these cues, rather than on editing pairs. This is achieved by optimizing the rectified-flow objective in Eq. (2), but with $y = \emptyset$ i.e., without the context image y . The model thus learns to approximate $p(x|c)$, where the condition c includes the natural language prompt and the specific control map (E or C). We found that this capability, learned during conditional generation, robustly generalizes to the conditional editing task at inference time, allowing the model to apply structural or color guidance even when a context image y is provided.

In contrast, the spatial layer (mask M) is trained explicitly for local editing, as its purpose is to constrain changes to a specific region. This requires a dataset of (source, target, prompt, mask) tuples, which we generate via self-distillation. First, we use a VLM (Qwen2.5-VL-72B [1]) to generate several plausible local editing prompts for a given source image. Our base FLUX Kontext model then executes

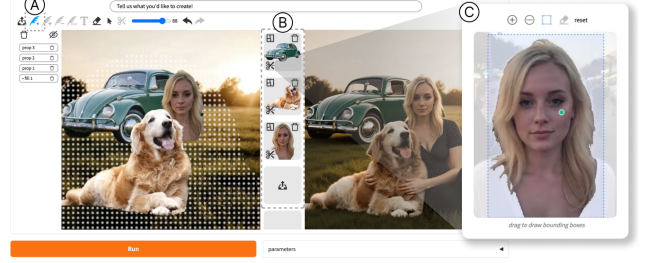


Figure 4. The interactive system interface of MagicQuill V2. (A) The *Fill Brush* in the toolbar allows users to define the spatial layer (mask M) by painting on the canvas. (B) The *Visual Cue Manager* holds content layer cues for drag-and-drop composition. (C) The *Image Segmentation Panel*, triggered from the manager, enables precise cue extraction using SAM-based interactions.

these edits. To derive the mask M , we compute the pixel-wise difference between the source and edited images, apply a threshold, and calculate the convex hull of the changed regions. We then filter this dataset to remove samples with masks that are either too large (global edits) or too small (no significant change). To further enhance the model’s capability for the common task of object removal, we follow Jiang et al. [26] to construct an additional dataset. This involves randomly extracting foreground objects and pasting them back onto arbitrary locations in the same image, strengthening the model’s ability to seamlessly remove content.

3.3. Interactive System Interface

We designed an interactive system, shown in Fig. 4, to unify our layered composition framework. This interface extends the “Idea Collector” from MagicQuill V1 [38]. The main Canvas is supplemented by a *Toolbar* containing a new *Fill Brush* (A). This tool is the primary interaction for defining the spatial layer (mask M), allowing users to simply paint where the edit should occur. The *Toolbar* also retains the similar brush-based tools from V1, enabling users to draw sketches for the structural layer (edge map E) and apply strokes for the color layer (color map C). A *Visual Cue Manager* (B), located right to the canvas, holds all content layer (L_{fg}) visual cues (foreground pieces F_i). Users can drag these cues onto the canvas to define what to generate.

To help users extract these visual cues from source images, each cue in the *Visual Cue Manager* features a “segment” button, which opens an *Image Segmentation Panel* (C). This panel utilizes SAM [30] to precisely segment an object in real-time, using positive/negative dots or a bounding box. The refined cue can then be saved back to the manager. This workflow provides an intuitive method for managing and composing precise visual cues.

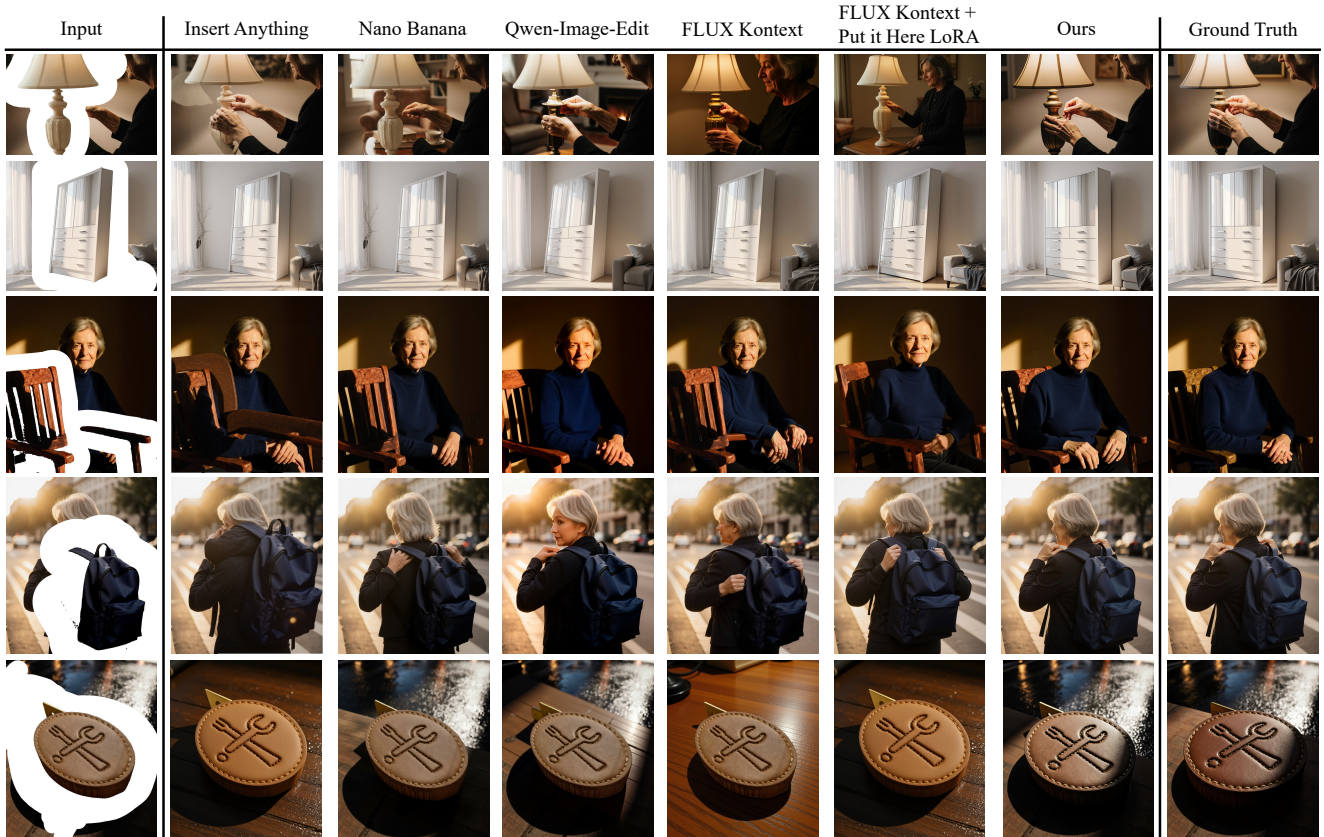


Figure 5. Qualitative comparison of our content layer integration against state-of-the-art methods. Our model (MagicQuill V2) successfully integrates foreground objects by handling complex semantic interactions (rows 3, 4), harmonizing lighting (row 1), and correcting perspective distortions (rows 2, 5). These examples highlight our method’s ability to produce results that are significantly more realistic and faithful than Insert Anything [59], Nano Banana [16], Qwen-Image [1], and the specialized Kontext LoRA “Put it Here” [1].

4. Experiment

We conduct a comprehensive set of experiments to validate the effectiveness of our proposed layered composition framework. Our evaluation is structured to first analyze the fidelity and contextual awareness of the content layer (Sec. 4.1), followed by a detailed assessment of the precision and versatility of the control layers (Sec. 4.2 and Sec. 4.3). Implementation details and more results will be shown in the supplementary materials.

4.1. Analysis of Content Layer

The goal of the content layer (L_{fg}) is to perform context-aware edits around the user-provided foreground cues, seamlessly integrating the foreground pieces into the edited image, rather than a simple “copy-paste” operation. Following the principles of our data construction pipeline detailed in Sec. 3.1, we manually curate a test set of 200 samples. This set is designed to be challenging and is divided into two categories: 100 interaction-based samples and 100 placement-based samples. We evaluate our method against a suite of state-of-the-art models, including InsertAnything

[59], Nano Banana [16], Qwen-Image [70], the base FLUX Kontext model [2], and a popular, community-trained Kontext LoRA “Put it Here” [20] specifically designed for high-fidelity foreground insertion.

Qualitative and quantitative comparison. Fig. 5 provides a comprehensive visual comparison. Our model excels at complex interactions, realistically generating the hand around the backpack (row 4) and the person sitting on the chair (row 3). Furthermore, our relight augmentation leads to superior photometric harmonization, as seen in the lamp and hand’s illumination (row 1) and logo’s correct side-lighting (row 4). The successful geometric correction of the cabinet (row 2) and logo (row 5) validates our use of perspective augmentation.

In contrast, other baselines struggle with these scenarios, altering non-edited regions (row 1) or failing complex interactions (rows 3-4), lighting (row 5), and geometry (row 2). Our method’s consistent success validates that our data pipeline focuses on foreground completion, and augmentations are essential for robust, high-fidelity synthesis. This is confirmed quantitatively in Tab. 1, where our method significantly outperforms all baselines on almost all metrics.

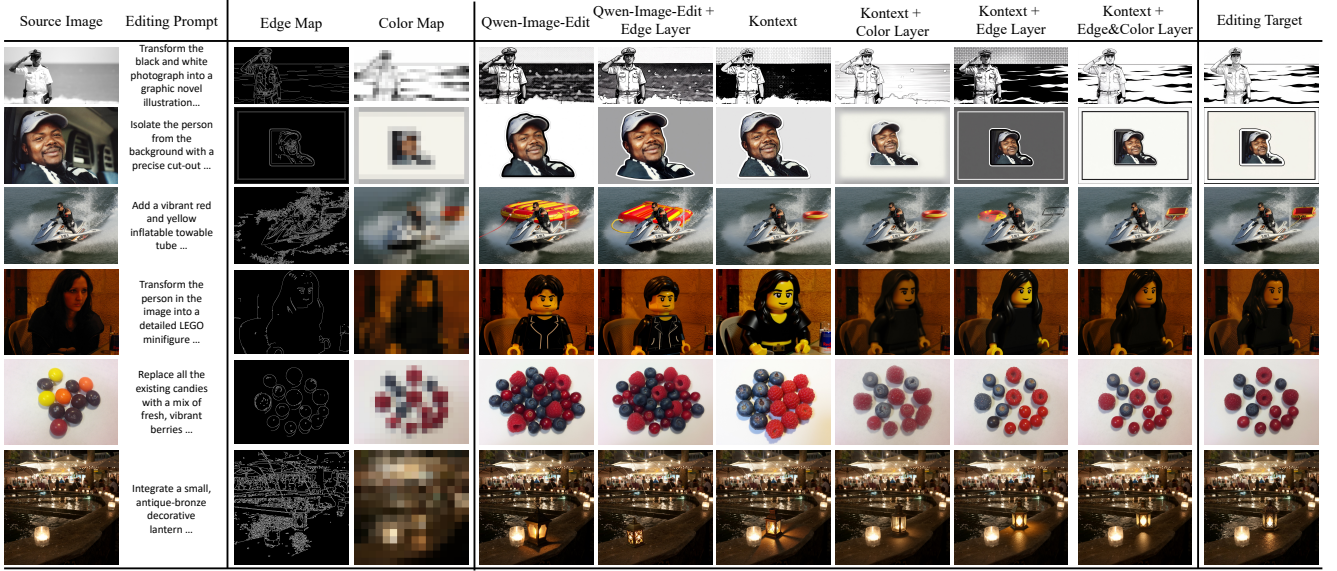


Figure 6. Qualitative comparison for structural (edge) and color layer control. Our full model, by composing both cues, achieves high-precision edits, significantly outperforming baselines.

Table 1. Quantitative comparison for the content layer composition. Our model achieves the best in almost all metrics.

Model	$L_1 \downarrow$	$L_2 \downarrow$	CLIP-I $\uparrow$	DINO $\uparrow$	CLIP-T $\uparrow$	LPIPS $\downarrow$
Insert Anything	0.105	0.039	0.910	0.825	0.327	0.354
Nano Banana	0.105	0.038	0.934	0.891	0.335	0.321
Qwen-Image	0.114	0.042	0.929	0.881	0.334	0.357
FLUX Kontext	0.117	0.045	0.930	0.872	0.337	0.359
Put it Here	0.136	0.054	0.925	0.854	0.335	0.438
Ours	0.061	0.019	0.962	0.930	0.335	0.202

Ablation in Data Construction. We conduct an ablation study to validate the individual components of our data construction pipeline, detailed in Sec. 3.1. We train separate models, each lacking one of the key augmentations, and present the qualitative results in Fig. 7.

Without training on perspective-distorted inputs, the model fails to correct the geometry. As seen, it places the cabinet with its original slant, resulting in a physically implausible scene (row 1). Without relight augmentation, the model defaults to a “pasted-on” look. The subject’s lighting is flat and does not match the bar’s warm, ambient lighting (row 2). When robustness to varied input quality is not learned, the model struggles with low-resolution or low-quality cues, producing a blurry or noisy image. Without our completion LoRA, the model is trained on incomplete foregrounds (e.g., the occluded book). Consequently, it learns a simple “copy-paste” behavior and fails to interact with the foreground.

4.2. Analysis of Structural and Color Layer

To evaluate the structural layer $L_{structural}$ (edge) and the color layer L_{color} (color), we conduct our evaluation on the Pico-Banana-400K benchmark [50], which provides a diverse set of text–image–edit triplets consisting of 35 edit op-

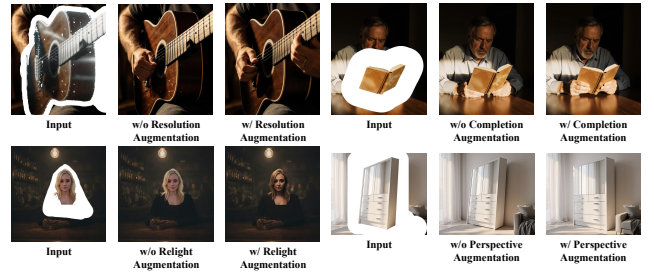


Figure 7. Qualitative ablation of our data construction pipeline (Sec. 3.1). Removing Perspective, Relight, Resolution, or Object Completion augmentations leads to synthesis failures.

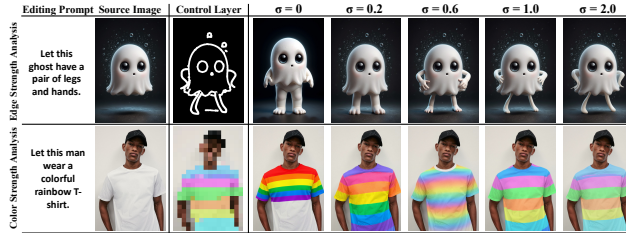
erations across 8 semantic categories. Images from selected benchmark were generated by the state-of-the-art Nano Banana [16] model and verified by Gemini [15] to ensure high fidelity and perfect semantic alignment.

To test MagicQuill V2’s precision and its ability to follow control cues, we sampled 1,000 cases from this benchmark. For each triplet, we simulate a user with a target in mind by extracting the edge map and color map directly from the target image. We then provide these cues to the models. We compare against the latest open-source model Qwen-Image-Edit [70], which includes native edge control module. We also show ablations of our own model with edge and color control layers.

Qualitative and quantitative comparison. The qualitative results in Fig. 6 demonstrate the remarkable precision of our model. Qwen-Image-Edit [70] and FLUX Kontext [2] interpret the instruction to generate plausible edit. However, text prompts alone are often insufficient to capture complex spatial and color intentions. The Edge module of Qwen-Image-Edit attempts to address this, but its control remains ‘soft’ and interpretive, failing to align perfectly.

Table 2. Quantitative comparison for the control layers.

Model	$L_1 \downarrow$	$L_2 \downarrow$	CLIP-I $\uparrow$	DINO $\uparrow$	LPIPS $\downarrow$
Qwen Image	0.132	0.043	0.923	0.871	0.395
Qwen Image (Edge)	0.131	0.042	0.924	0.875	0.387
FLUX Kontext	0.152	0.054	0.908	0.853	0.434
Ours (Edge)	0.107	0.030	0.938	0.909	0.317
Ours (Color)	0.080	0.020	0.943	0.915	0.327
Ours (Edge+Color)	0.080	0.018	0.949	0.930	0.283

Figure 8. Qualitative analysis of the control strength parameter σ . As σ increases, the model’s adherence to the user-provided edge layer (row 1) and color layer (row 2) progressively tightens.

In contrast, our layered approach highlights the complementary value of different control modalities. Our model with the edge layer alone perfects geometry but distorts color, while the color layer alone aligns color but misses structural details. This proves both cues are essential. When composed together, their strengths are combined to perfectly reproduce the user’s intent, consistently matching the target edit. This validates that our layered design provides the crucial bridge to high-fidelity, explicit control. These qualitative findings are directly supported by our quantitative evaluation in Tab. 2, where our combined method performs uniformly better across all metrics.

Analysis of control strength. The control strength parameter σ (Eq. (11)) offers flexible control, which is crucial as user-provided cues, such as hand-drawn sketches, can be rough. Users can tune σ to balance their trust in the cue against the model’s generative priors. We analyze this mechanism in Fig. 8, using manually drawn cues to simulate this scenario. At $\sigma = 0$, the cue is ignored, reverting to the base FLUX Kontext behavior. As σ increases, adherence to the user-drawn edge layer (row 1) and color layer (row 2) progressively tightens. While our default $\sigma = 1.0$ provides a balanced and faithful result, higher values enforce stricter adherence, which can also amplify imperfections in the cue, potentially leading to artifacts.

4.3. Analysis of Spatial Layer

In this section, we evaluate the spatial layer ($L_{spatial}$), which is designed for precise regional edit. We first compare its ability to perform local edits against inpainting-based models. We then evaluate its specialized object removal capability, for which it was explicitly fine-tuned.

Comparison with inpainting models. Our spatial layer is designed to apply edits to masked content, not just regenerate. As shown in Fig. 9, FLUX Kontext fails with global ed-

Table 3. Quantitative comparison for object removal.

Model	$L_1 \downarrow$	$L_2 \downarrow$	LPIPS $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$	FID $\downarrow$
SmartEraser	0.069	0.098	0.196	0.630	21.14	17.03
OmniEraser (Base)	0.058	0.084	0.243	0.660	22.16	19.76
OmniEraser (CN)	0.048	0.084	0.182	0.817	22.96	25.92
Ours	0.042	0.071	0.154	0.840	24.45	16.42

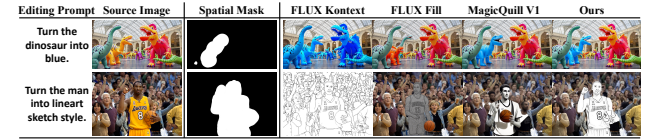


Figure 9. Qualitative comparison for regional editing.

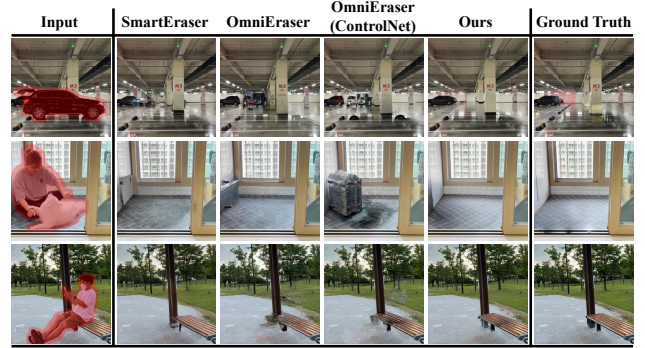


Figure 10. Qualitative comparison for object removal task.

its, while inpainting models like FLUX Fill [32] and MagicQuill V1 [38] ignore the masked region’s content, failing content-aware edits (e.g., Row 2). Our model, however, attends to the content within the mask, successfully applying color adjustment and style transfer while preserving the identity, demonstrating content-aware local editing.

Comparison in object removal. We compare our model with state-of-the-art object removal methods SmartEraser [26] and OmniEraser [69] over 5,000 samples of RORD [56] benchmark. The experimental results are presented in Tab. 3 and Fig. 10. MagicQuill V2 demonstrates superior performance over existing models in erasing target objects.

5. Conclusion

We introduced MagicQuill V2, a novel system that brings the granular control of layered composition to generative image editing. Monolithic text prompts are insufficient for capturing complex user intent. Our method deconstructs this intent into distinct content, spatial, structural, and color layers. We proposed a specialized data pipeline for context-aware content layer composition and a unified control module to precisely manage geometry, color, and regional edits. Extensive experiments validate our approach, demonstrating state-of-the-art performance in content composition, alignment with respect to the control layer, and high-fidelity regional edit. MagicQuill V2 effectively bridges the intention gap, granting users intuitive, powerful, and precise control over the generative process.

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6, 2
- [2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 1, 2, 3, 4, 6, 7
- [3] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18434–18443, 2025. 2
- [4] Ke Cao, Jing Wang, Ao Ma, Jiasong Feng, Zhanjie Zhang, Xuanhua He, Shanyuan Liu, Bo Cheng, Dawei Leng, Yuhui Yin, et al. Relactrl: Relevance-guided efficient control for diffusion transformers. *arXiv preprint arXiv:2502.14377*, 2025. 2
- [5] Caroline Chan, Fredo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. 2022. 2
- [6] Xiuwen Chen, Li Fang, Long Ye, and Qin Zhang. Deep video harmonization by improving spatial-temporal consistency. *Machine Intelligence Research*, 21(1):46–54, 2024. 2
- [7] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. In *NeurIPS*, 2024. 2
- [8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. 2
- [9] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unreal: Universal image generation and editing via learning real-world dynamics. In *CVPR*, 2025. 2
- [10] Zhekai Chen, Wen Wang, Zhen Yang, Zeqing Yuan, Hao Chen, and Chunhua Shen. Freecompose: Generic zero-shot image composition with diffusion prior. *arXiv preprint arXiv:2407.04947*, 2024. 2
- [11] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. Color harmonization. *ACM Trans. Graph.*, 2006.
- [12] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. *arXiv preprint arXiv: 2009.09169*, 2020.
- [13] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 2
- [14] Yusuf Dalva, Yijun Li, Qing Liu, Nanxuan Zhao, Jianming Zhang, Zhe Lin, and Pinar Yanardag. Layerfusion: Harmonized multi-layer text-to-image generation with generative priors. *arXiv preprint arXiv:2412.04460*, 2024. 2
- [15] Google Deepmind. Gemini 2.5 flash image (nano banana). <https://deepmind.google/models/gemini/image/>, 2025. 2, 7
- [16] Google Deepmind. Gemini 2.5 flash image (nano banana). <https://aistudio.google.com/models/gemini-2-5-flash-image>, 2025. 1, 6, 7, 3
- [17] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 2
- [18] Alessandro Fontanella, Petru-Daniel Tudosi, Yongxin Yang, Shifeng Zhang, and Sarah Parisot. Generating compositional scenes via text-to-image rgba instance generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2
- [19] Tsu-Jui Fu, Wenzhe Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 3
- [20] Futurlunatic. Put it here: Kontextv4 lora. <https://civitai.com/models/1808575/put-it-herekontextv4>, 2025. 6, 3
- [21] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14870–14879, 2021. 2
- [22] Guoqing Hao, Satoshi Iizuka, and Kazuhiro Fukui. Image harmonization with attention-based deep feature modulation. In *British Machine Vision Conference*, 2020. 2
- [23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3, 4
- [24] Dingbang Huang, Wenbo Li, Yifei Zhao, Xinyu Pan, Yanhong Zeng, and Bo Dai. Psdiffusion: Harmonized multi-layer image generation via layout and appearance alignment. *CoRR*, abs/2505.11468, 2025. 2
- [25] Runhui Huang, Kaixin Cai, Jianhua Han, Xiaodan Liang, Renjing Pei, Guansong Lu, Songcen Xu, Wei Zhang, and Hang Xu. Layerdiff: Exploring text-guided multi-layered composable image synthesis via layer-collaborative diffusion model. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024. 2
- [26] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24452–24462, 2025. 5, 8, 3
- [27] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. *ICCV*, 2021. 2
- [28] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 2

- [29] Kyoungkook Kang, Gyu Jin Sim, Geonung Kim, Donguk Kim, Seungho Nam, and Sunghyun Cho. Layeringdiff: Layered image synthesis via generation, then disassembly with generative knowledge. *arXiv preprint arXiv:2501.01197*, 2025. 2
- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 5
- [31] Black Forest Labs Krea. Flux krea. <https://github.com/krea-ai/flux-krea>, 2025. 3, 1
- [32] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2023. 4, 8
- [33] Muyang Li*, Yujun Lin*, Zhekai Zhang*, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [34] Yaowei Li, Xiaoyu Li, Zhaoyang Zhang, Yuxuan Bian, Gan Liu, Xinyuan Li, Jiale Xu, Wenbo Hu, Yating Liu, Lingren Li, et al. Ic-custom: Diverse image customization via in-context learning. *arXiv preprint arXiv:2507.01926*, 2025. 2
- [35] Haichuan Lin, Yilin Ye, Jiazhi Xia, and Wei Zeng. Sketchflex: Facilitating spatial-semantic coherence in text-to-image generation with region-based sketches. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2025. Association for Computing Machinery. 3
- [36] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [37] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2
- [38] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13072–13082, 2025. 3, 5, 8, 2
- [39] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2
- [40] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023. 3
- [41] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 3
- [42] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025. 2
- [43] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2
- [44] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021. 2
- [45] OpenAI, :, Aaron Hurst, and Adam Lerer et al. Gpt-4o system card, 2024. 1, 2
- [46] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2
- [47] Xiaohan Peng, Janin Koch, and Wendy E. Mackay. Fusain: Composing generative ai visual prompts using pen-based interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2025. Association for Computing Machinery. 3
- [48] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 1434–1439. IEEE, 2005. 2
- [49] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8808–8816, 2018. 2
- [50] Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing, 2025. 7
- [51] Fabio Quattrini, Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Alfie: Democratising rgba image generation with no \$\$\$s. *arXiv preprint arXiv:2408.14826*. 2
- [52] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3
- [53] Nathalie Riche, Anna Offenwanger, Frederic Gmeiner, David Brown, Hugo Romat, Michel Pahud, Nicolai Marquardt, Kori Inkpen, and Ken Hinckley. Ai-instruments: Embodying prompts as instruments to abstract & reflect graphical interface commands as general-purpose tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2025. Association for Computing Machinery. 3
- [54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 1

- [55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2
- [56] Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Won Jung, and Sung-Jea Ko. Rord: A real-world object removal dataset. In *BMVC*, page 542, 2022. 8
- [57] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 2
- [58] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2
- [59] Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*, 2025. 2, 6, 3
- [60] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058, 2024. 2
- [61] Xavier Soria, Yachuan Li, Mohammad Rouhani, and Angel D. Sappa. Tiny and efficient model for the edge detection generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1364–1373, 2023. 2
- [62] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5117–5127, 2021. 2
- [63] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010. 2
- [64] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 2
- [65] Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. Ominicontrol2: Efficient conditioning for diffusion transformers. *arXiv preprint arXiv:2503.08280*, 2025. 2
- [66] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. *International journal of computer vision*, 103:178–189, 2013. 2
- [67] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 2
- [68] Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. Mulan: A multi layer annotated dataset for controllable text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22413–22422, 2024. 2
- [69] Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, et al. Omnieraser: Remove objects and their effects in images with paired video-frame data. *arXiv preprint arXiv:2501.07397*, 2025. 8
- [70] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 2, 6, 7, 3
- [71] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 2
- [72] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3
- [73] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18381–18391, 2023. 2
- [74] Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. *arXiv preprint arXiv:2411.17864*, 2024. 2
- [75] Jinrui Yang, Qing Liu, Yijun Li, Soo Ye Kim, Daniil Pakhomov, Mengwei Ren, Jianming Zhang, Zhe Lin, Cihang Xie, and Yuyin Zhou. Generative image layer decomposition with visual effects. *arXiv preprint arXiv:2411.17864*, 2024. 2
- [76] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2
- [77] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 4, 1
- [78] Xinyang Zhang, Wentian Zhao, Xin Lu, and Jeff Chien. Text2layer: Layered image generation using latent diffusion model. *arXiv preprint arXiv:2307.09781*, 2023. 2
- [79] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025. 2
- [80] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3943–3951, 2015. 2

MagicQuill V2: Precise and Interactive Image Editing with Layered Visual Cues

Supplementary Material

6. Implementation Details

6.1. Content Layer

Training and Inference Details. We integrate the content layer capability by fine-tuning the FLUX Kontext backbone with a LoRA adapter of rank $r = 32$. The model was trained for 9,000 steps on 8 H20 (140GB) GPUs, using the AdamW optimizer with a constant learning rate of 1×10^{-4} . For inference, the model performs 20 denoising steps, requiring approximately 30 seconds and 30GB of VRAM on a single H20 GPU.

Completion Augmentation. We utilize a dedicated object completion LoRA to restore foreground objects that are occluded in the source images. To train this completion model, we constructed a dataset of 3,000 high-resolution object images with white backgrounds, generated using Flux.1 Krea [31]. We simulate occlusions by applying random white brushstroke masks to these images. As demonstrated in Fig. 11. We train the model by providing pairs of complete, white-background objects (top row) and their counterparts with random brushstroke masks applied (bottom row).



Figure 11. Training data generation for the completion LoRA.

The LoRA is then fine-tuned on FLUX Kontext, with a rank $r = 32$, to reconstruct the original complete object from the masked input, guided by its descriptive caption. The training was conducted on 8 H20 (140GB) GPUs. We used the AdamW optimizer with a learning rate of 1×10^{-4} with 10 epochs of optimization. Qualitative results of this completion model are shown in Fig. 12, demonstrating its ability to restore occluded items extracted from scenes.

Relight Augmentation. We perform photometric augmentation using the SD1.5 [54] version of ICLight [77]. For each foreground object, we first generate a random light map. This map is used as the initial latent in an image-

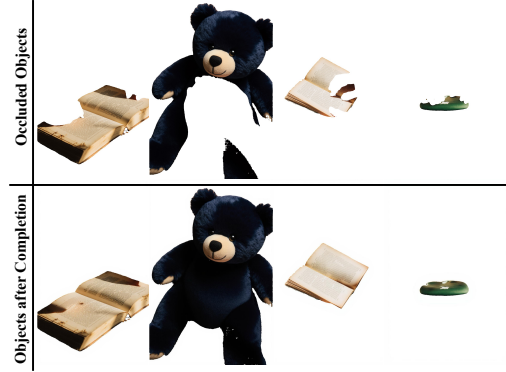


Figure 12. Qualitative results of the completion LoRA.

to-image relighting process. The light map’s color properties are randomly sampled: 50% are grayscale, 30% are low-saturation color, and 20% are high-saturation color, as illustrated in Fig. 13. This forces the model to learn robust lighting harmonization rather than just copy-pasting.

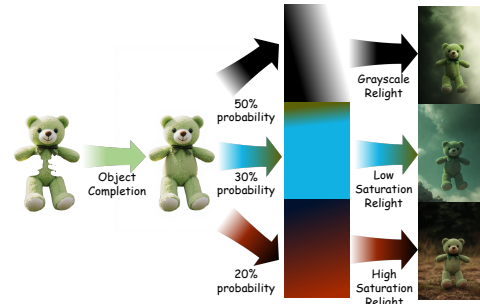


Figure 13. Overview of the relight augmentation pipeline. Completed objects are randomly relit using one of three light map categories to improve photometric robustness.

Perspective Augmentation. To simulate geometric mismatches, we apply a random perspective transformation. Given a foreground object F within its $w \times h$ bounding box, we define its four corner coordinates as the source points $P_{src} = \{(0, 0), (w, 0), (w, h), (0, h)\}$. We sample a maximum perturbation ratio $\rho \sim U(0.1, 0.3)$ and define the maximum displacement values $\Delta x = w \cdot \rho$ and $\Delta y = h \cdot \rho$. We then generate four destination points $P_{dst} = \{(x'_i, y'_i)\}_{i=1}^4$ by adding a random displacement $\delta \sim U([- \Delta x, - \Delta y], [\Delta x, \Delta y])$ to each corresponding source point $p_i \in P_{src}$, clipping the result to stay within the original $w \times h$ boundaries. A 3×3 perspective transformation matrix H is computed by solving the homography that maps P_{src} to P_{dst} . This transformation H is then applied to both the foreground object F and its mask M_F to produce the geometrically augmented F' and M'_F .

Resolution Augmentation. To ensure robustness against varying input qualities, we simulate low-resolution inputs. For each foreground object, we sample a random scaling factor $s \sim U(0.15, 0.9)$. The object is downsampled to s of its original size and then upsampled back to its original resolution using bilinear interpolation.

6.2. Control Layer

Training and Inference Details. We implement the structural, color, and spatial layers using the unified control module, each with a rank of $r = 128$. All visual control cues (e.g., edge maps, color maps) are resized to a fixed 512x512 resolution before being processed. The control modules were trained for about 10,000 steps on 8 H20 (140GB) GPUs, using the AdamW optimizer with a constant learning rate of 1×10^{-4} . For inference, the model performs 20 denoising steps. When a control cue is active, this process requires approximately 45 seconds and 40GB of VRAM on a single H20 GPU. These control modules can be composed and activated simultaneously (e.g., using structural and color cues together for high-fidelity edits) and are also compatible with other LoRAs over FLUX Kontext weights. Following the interactive paradigm of MagicQuill V1 [38], our system automatically selects and provides the appropriate control cues based on the user’s brushstroke type.

Structural Layer. To train the structural layer, we use a large-scale, self-collected dataset, with all images resized to approximately 1000x1000 pixels with detailed captions. For each image, we generate a structural map by randomly selecting one edge and lineart extractor from Canny [17], PidiNet [62], TEED [61], HED [71], and Informative Drawings [5]. As illustrated in Fig. 14, this ensures the model is robust to various styles of structural inputs.

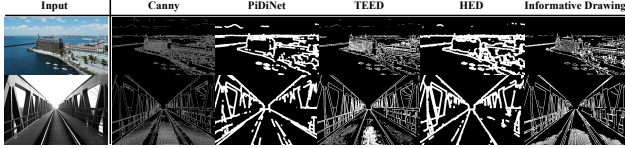


Figure 14. Visual results of different edge/lineart extractors.

The structural layer is trained for conditional generation, optimizing the model to approximate $p(x|c, E)$, where c is the text caption and E is the sampled edge map without the context image. We found that the structure-following capability, learned via conditional generation, robustly generalizes to the conditional editing task at inference time. Following the interactive paradigm of MagicQuill V1 [38], we define binary masks M_{add} and M_{sub} corresponding to the user’s add and subtract brushstrokes, respectively. The subtract brush removes edges from the map, while the add brush introduces new edges. Let E denote the extracted edge map; the final control condition E_{cond} is formally computed as:

$$\begin{aligned} E_{sub} &= E \odot (1 - M_{sub}), \\ E_{cond} &= E_{sub} + M_{add} \odot (1 - E_{sub}). \end{aligned} \quad (12)$$

where $\odot$ denotes the element-wise product. The resulting modified edge map E_{cond} serves as the precise geometric constraint for the generation process.

Color Layer. The training paradigm for the color layer mirrors that of the structural layer. We utilize the same dataset and train the model for conditional generation without the context image context y . During training, the color condition map C is generated by downsampling to 16×16 and resizing back to 512×512 . This removes high-frequency details, forcing the model to learn the mapping from the palette to the final image.

At inference, we adopt the color brushstroke logics from MagicQuill V1 [38]. Each user color stroke is parameterized as a tuple (M_{color}, c, α) , where M_{color} denotes the binary mask of the painted region, c specifies the target RGB color, and $\alpha \in [0, 1]$ (default to 0.4) represents the stroke opacity. Let C denote the initial color map extracted from the source image (or a blank canvas); the updated control condition C_{cond} is computed via alpha blending:

$$C_{cond} = (1 - \alpha \cdot M_{color}) \odot C + \alpha \cdot M_{color} \cdot c, \quad (13)$$

where the specific color c is applied over the region defined by M_{color} with intensity α . This blending mechanism allows users to achieve both subtle color tinting and opaque color replacement.

Spatial Layer. Unlike the structural and color layers, which can be trained on single images via conditional generation, the spatial layer (represented by a binary mask M) requires training on local editing pairs (I_{src}, I_{tgt}, c, M) . The goal is to constrain the model’s generative changes strictly within the user-specified region. To achieve this, we construct a large-scale dataset via a rigorous Self-Distillation Pipeline. We leverage Qwen2.5-VL-72B [1] to propose a set of plausible local editing instructions c for each I_{src} , which are then executed by the base FLUX Kontext [2] model to produce the edited target image I_{tgt} .

To derive robust masks M from these generated pairs, we implement a two-stage difference extraction pipeline. First, a pre-screening step is performed using standard pixel-wise differences. We evaluate the edit magnitude across multiple threshold parameters, filtering out samples where the changing hull area has ratio outside the range $[0.001, 0.75]$ to eliminate insignificant or excessive global edits, as illustrated in Fig. 15. For the retained pairs, the final mask is generated in the CIELAB color space. We compute the Euclidean distance between the aligned Lab vectors to capture perceptual photometric differences, offering superior robustness to compression artifacts compared to standard RGB. To prevent mask fragmentation, we compute the Convex Hull of the detected change regions, ensuring the mask covers the entire semantic object. Finally, this hull is refined using rolling-circle smoothing to simulate the continuous topology of natural human brushstrokes.



Figure 15. Visualization of the data filtering strategy for the Spatial Layer. We calculate the area ratio of the convex hull covering the changed regions between the source and target images. The pipeline automatically rejects failed edits with no significant changes (Row 1) or excessive global changes (Row 2), retaining only high-quality local editing pairs (Rows 3 & 4) for training.

To further augment the model’s capability for object removal, we follow Jiang et al. [26] to create a synthetic dataset. This involves randomly extracting SAM-based foregrounds and pasting them onto arbitrary backgrounds, treating the pre-paste image as the ground truth target.

7. User Study for Content Layer

Standard quantitative metrics often fail to capture perceptual nuances in image composition, such as lighting integration and occlusion handling. To evaluate the human-perceived quality of our method, we conducted a user preference study focusing on the content layer (L_{fg}).

Setup and Methodology. We recruited 30 participants to evaluate 10 editing scenarios sampled from our test set. For each scenario, participants compared results from six methods: Nano Banana [16], Insert Anything [59], FLUX Kontext [2], Kontext + Put It Here LoRA [20], Qwen-Image-Edit [70], and ours. Participants selected the single “best” result based on foreground preservation, visual coherence, and interaction plausibility.

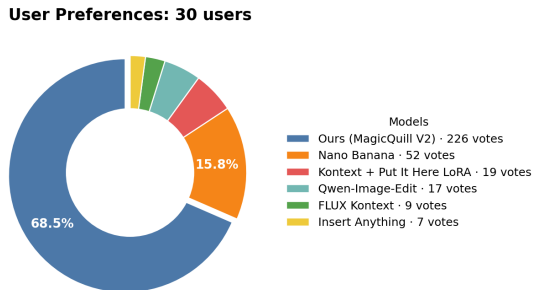


Figure 16. **User Preference Distribution.** MagicQuill V2 (Ours) dominates with **68.5%** of the votes, significantly outperforming the strongest baseline, Nano Banana (15.8%).

Results. As shown in Fig. 16, **MagicQuill V2 achieved a dominant preference rate of 68.5%** (226 votes), significantly outperforming the second-best model, Nano Banana (15.8%, 52 votes). Qualitative feedback indicates that our method consistently delivers results that are physically grounded and harmoniously integrated.

8. Limitation

Despite the significant advancements in precise controllability, MagicQuill V2 has several limitations that present avenues for future research.

Inference Latency. As a diffusion transformer-based model utilizing a heavy backbone (12B) and multiple control adapters, our system prioritizes generation quality over speed. A single edit requires approximately 30-45 seconds on a high-end H20 GPU. This latency creates friction in the “interactive” workflow, as users must wait for feedback. Future work could explore consistency distillation [40, 41] or quantization techniques [33] to achieve near real-time performance for interaction.

Conflict between Modalities. While our layered architecture enables flexible composition, the freedom to stack multiple control layers introduces the potential for contradictory inputs. Conflicts can arise not only between semantic text and visual cues (e.g., a prompt describing a “circle” while the structural layer dictates a “square”) but also between distinct visual layers. For instance, a user might provide a detailed structural layer (edge map) that outlines complex geometry, while simultaneously applying a color layer that contradicts those boundaries. Currently, the model attempts to reconcile these disjointed signals based on learned priors and the per-layer control strength σ . However, when cues are heavily conflicting, this implicit resolution can result in artifacts, where the model fails to satisfy mutually exclusive constraints.

9. Open Source Commitment

To facilitate reproducibility and future research, we are committed to fully open-sourcing our project. We will publicly release the complete codebase, including training scripts, model checkpoints, and the interactive system, including the user interface.