Streamlining Image Editing with Layered Diffusion Brushes

Supplementary Material

6. UI and interaction design

Fig. 8 provides an overview of the user interface. As demonstrated, the UI comprises the following primary sections (each section highlighted with the corresponding number on the image):

1. Model Section
   - This section in the UI enables users to load various model combinations, including pre-trained models, schedulers, and LoRA [25].
2. Generation section
   - This section allows users to either generate a new image using a seed and prompt combination, or upload a real image that will be inverted.
3. Image Canvas
   - This canvas serves as the workspace where users interact with and make edits to images.
4. Editing Section
   - This section provides controls to create and modify different layers to make the desired edits.

We provide the ability to stack and hide/unhide layers, similar to traditional image-editing tools.

When editing a layer, we provide the choice of box mode or brush mode. In box mode, the mask is a square shape controlled by the "brush size" parameter. As the box is dragged around the image, the seed value will automatically increment, providing a continuous stream of new edits. The user may stop dragging when a suitable edit is seen.

In brush mode, the mask is an arbitrary shape that can be added to or subtracted from using a circular brush tool. The size of the brush is controlled by the "brush size" parameter. In this mode, the user can scroll a mouse wheel or use a scrolling gesture to increment or decrement the seed, allowing them to rapidly explore the space of potential edits and return to any edit that appears suitable.

6.1. Hyperparameters

To provide a balance between usability and complexity, we provide control over a number of hyperparameters: number of regeneration steps, "brush strength", brush size and seed number. Each hyperparameter is designed to be largely orthogonal to the other parameters, enabling them to independently affect the appearance of the edit without the need to simultaneously adjust multiple inputs.

- **Number of regeneration steps** (n): An integer value that specifies the number of steps LDB will run to make the edit. Changing n effectively changes the strength of the modification as well as the processing time.
- **Brush Strength** (α): A number that indirectly controls the α value in (Eq. (3)) which controls how strong the initial noise pattern should be. The user-specified alpha, α^*, has a value between 0 and 100, which will be scaled using the following equation:

$$\alpha = \frac{\sqrt{\left|\frac{\alpha^*}{100} \cdot \left(\sigma - 2 \cdot \frac{\mathrm{Cov}(Z_r^{(k)}, Z_0')}{\mathrm{Var}(Z_r^{(k)})}\right)\right|}}{\sqrt{\frac{\sum_{i=1}^{W}\sum_{j=1}^{H}[m_{ij}\neq 0]}{W}}} \tag{3}$$

  where Z_0' and $Z_r^{(k)}$ are the new noise latent and latent for regeneration respectively (as noted in Algorithm 1), σ is the acceptable range for the variance of the $Z_r^{(k)}$ (we used σ=0.25), m is the corresponding mask, and W is the width of Z_{n_k} (W=512).
  This formula is designed to ensure that any fixed value of the user-provided α^* value produces similar effects on the image even as the number of regeneration steps or the brush size/mask size are changed, thus making it more logically independent from the other parameters.
- **Seed Number** (s'): An integer number that will be used for generating the Gaussian noise pattern in the specified region. As with normal image generation, the UI provides buttons to randomize the seed or reuse the previous seed. Moving the box around (in box mode) or using the scroll wheel (in custom mask mode) will adjust the seed automatically.
- **Brush Size** d: An integer value that dictates the radius of the box when utilized in box mode, or the size of the brush in custom mask mode (in pixels).

7. Additional Qualitative Examples

Fig. 9 and Fig. 11 present examples of Type 1 tasks (freeform) and Type 2 tasks (MagicBrush) respectively. All the images were edited by participants during the user study.

8. Ablation Study Details

8.1. Ablation on Mask Strength Control

The magnitude of the edit applied by LDB is jointly governed by the number of edit steps (n) and the mask strength control (α). These parameters control the amount of intermediate noise added to the latent image. Fig. 12 illustrates the effect of varying α. As shown, excessively high α values (right), representing strong edits, prevent the LDM from effectively denoising, leading to artifacts. Conversely, insufficient α results in negligible edits. Furthermore, n and

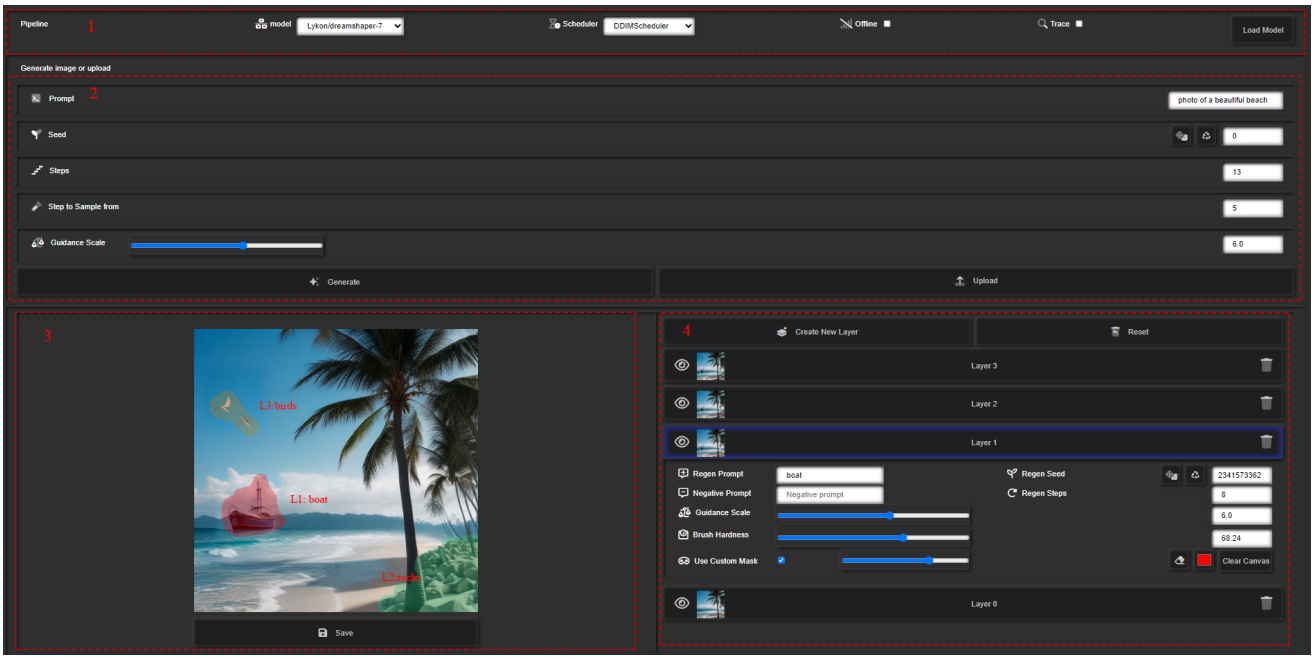


Figure 8. Design of the LDB's UI: The name and functionality of each section are described in the text. In this example, the user has created three layers, visualized on the image canvas, along with the mask and edit prompt. The selected layer in this picture is Layer 1.

α exhibit a coupled relationship. When noise is introduced later in the diffusion process (higher n), the model has less denoising capacity, necessitating a higher α to achieve a noticeable edit. Conversely, with earlier noise injection (lower n), a sufficiently large α is required to prevent the additive noise from being entirely diffused away in the initial denoising steps. Therefore, optimal editing requires careful consideration of both n and α, with α needing adjustment based on the chosen n to balance edit strength and image quality. In our UI, we formulate the translation from α^* to α to decouple these two parameters by factoring in the variance and covariance of the intermediate latent, thus automatically adjusting α when n changes.

8.2. Caching Latents Ablation Metrics

Fig. 13 and Fig. 14 present the graphs for quantitative metrics on the ablation studies as discussed in Sec. 4.3.

9. Video Editing Examples

We integrated LDB with several diffusion image transformers (DiT) and spatio-temporal video generation models. In Fig. 15, we demonstrate examples of video editing by integrating LDB into SVD [7].

10. User Study Details

10.1. Procedure and Task Description

The user cohort comprised four females and three males, with an average age of 30.4 years. Two participants were proficient in image generative models and Stable Diffusion, while the remaining five were graphic design students who used Adobe Photoshop and Illustrator on a daily basis. The study was conducted remotely; participants were provided a link to access the tool.

The study started with a brief introduction to each of the methods. Following this, participants received a short tutorial on how to navigate the user interface (UI). Subsequently, they were provided with a 5-minute window to explore the various options and sections of the tool, becoming familiar with the use of each section.

A dedicated task section was incorporated into the user interface (UI) specifically for the user study. Each type of task comprised three rounds of edits using the three methods: LDB, IP2P, and SDI.

Each user was assigned a unique user ID, and tasks were randomly selected and pre-assigned to users. Throughout the study, users interacted with the task table to load, select, and save each task. An example of the task section is illustrated in Fig. 16.

As mentioned in Section 4.1.1, the user study consisted of two types of tasks: free-form (type 1) and pre-determined (type 2) tasks. For the type 1 tasks, we selected specific

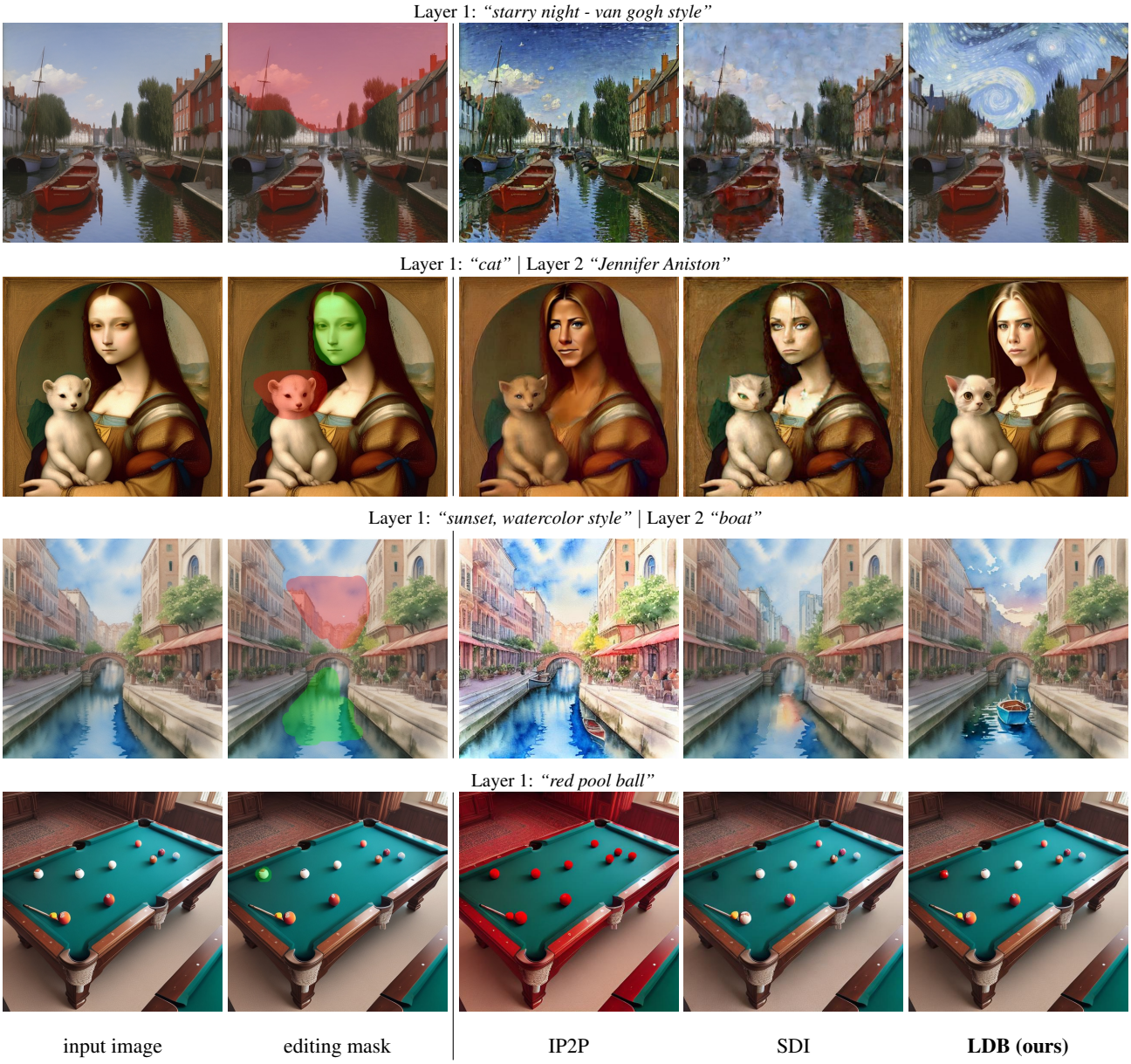


Figure 9. Qualitative results for the freeform part of the user study (Type 1 tasks)

types of edits that showcase various functionalities and capabilities of the system. Here are the description of edit types along with an example used during the user study:

1. Stack layers and create sequential edits (draw with LDB):
   - Input image: photo of a beautiful beach.
   - Layer 1: boat (Introduce a boat in the sea)
   - Layer 2: rocks (Scatter weathered rocks along the shoreline)
   - Layer 3: birds (Populate the sky above the boat with a flock of birds.
2. Modify attributes and features of objects:
   - Input image: portrait of a young man
   - Layer 1: blond (Transform a person's hair color to blond).
   - Layer 2: joker (Perform facial manipulation by swapping one person's face with another's, reshaping identities.)
3. Correct image imperfections and errors:
   - Input image: portrait of a man holding an umbrella
   - Layer 1: remove the rod that is mistakenly placed
   - Layer 2: fix the extra part on the side of the coat
4. Enhance discernibility of similar objects through modification:

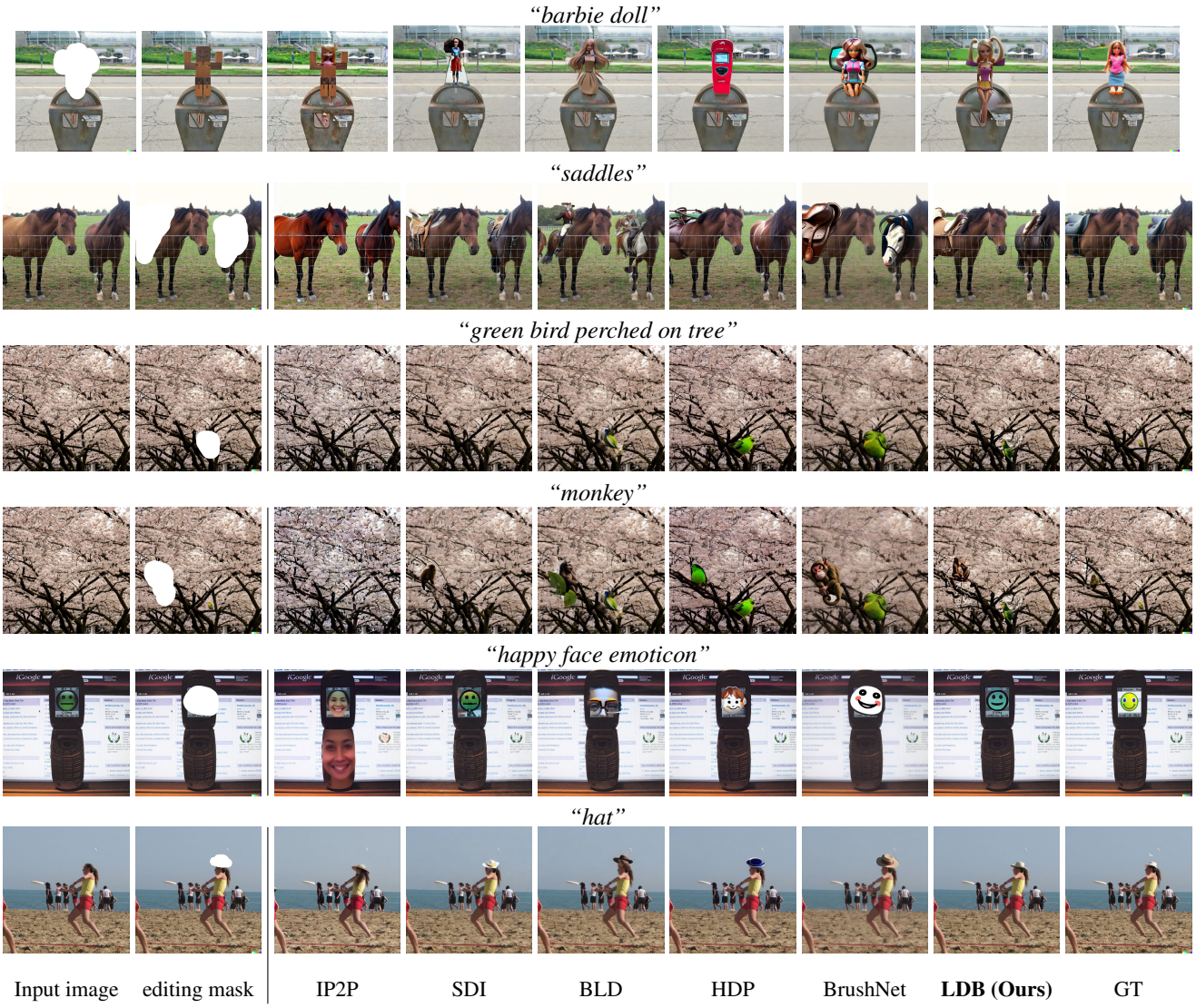


Figure 10. Additional qualitative examples on the MagicBrush dataset. Note that the images are not cherry picked and correspond to the user study (IP2P, SDI, LDB) and quantitative evaluation (BLD, HDP, BrushNet) with default settings.

- Input image: aerial photo of a pool table with balls
- Layer 1: change the colour of a specific ball (third ball from the left) to red

5. Target specific regions for style transfer, refining aesthetics:
   - Input image: Mona Lisa by Leonardo Da Vinci
   - Layer 1: make the left part of the background similar to Van Gogh starry night style.

In our study design, we strategically chose the combination of seeds and prompts to encompass and evaluate these functionalities. Each user was given three seed-prompt items and tasked with creating and editing up to three layers of edits. For the majority of the tasks, N, i.e. the total number of steps for editing was set to $n = 5$. All the images were generated using Dreamshaper-7 [13] and the DDIM scheduler.

For the LDB method, users started by selecting a layer with an existing edit instruction from the task table, then created the corresponding layer in the UI. They had the option of choosing either the box option or the custom mask option. The task was followed by drawing the mask, tweaking the controls or edit prompt if needed, and completing the edit. Once the task was complete, the user saved the edit and moves on to the next task.

Users followed a similar procedure for the IP2P and SDI methods, with the exception of creating layers, as these methods do not incorporate layering capabilities. After completing each layer edit task, users saved the edits, and

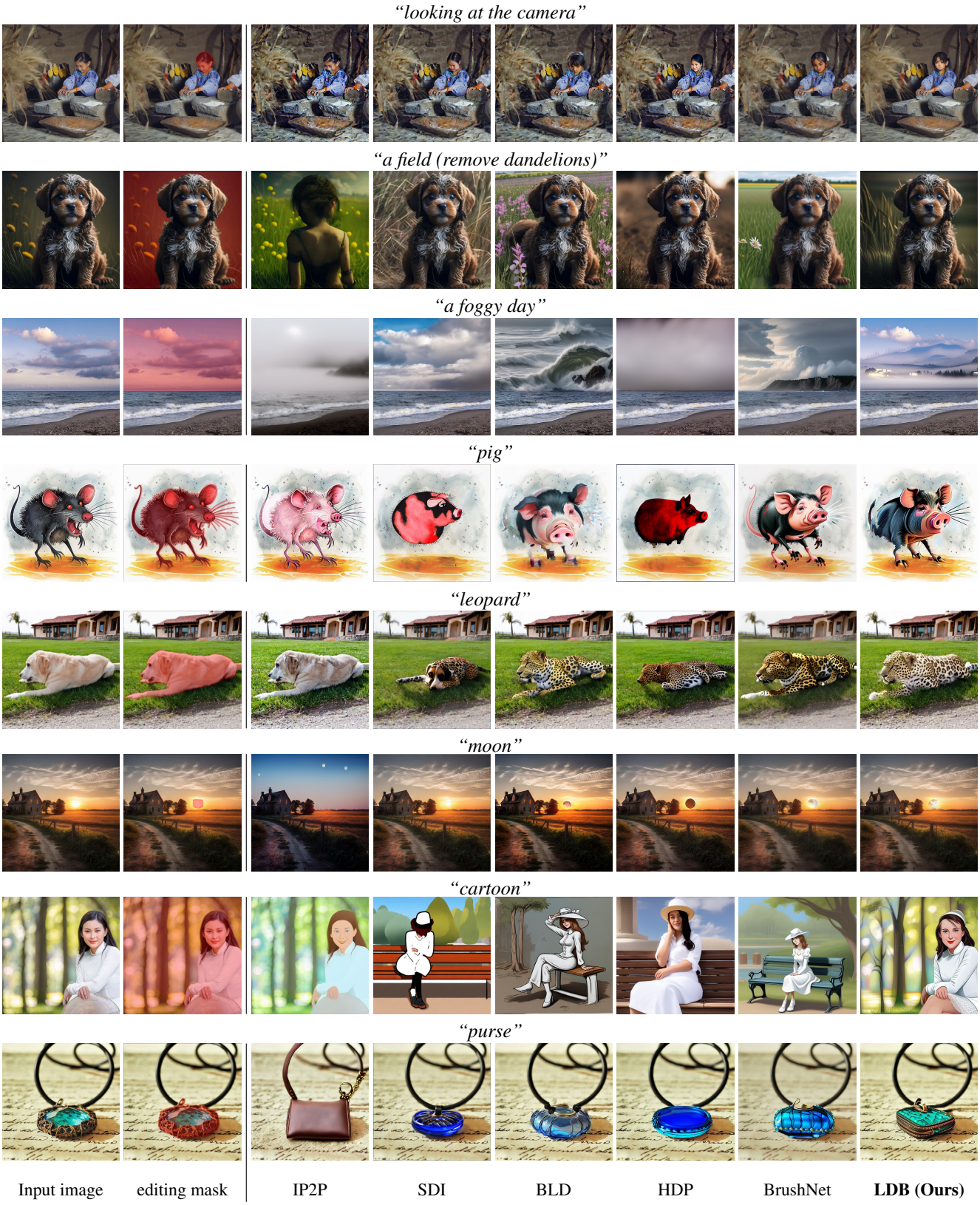


Figure 11. Additional qualitative examples on the PIE-Bench dataset. For all the methods, we used the default settings.

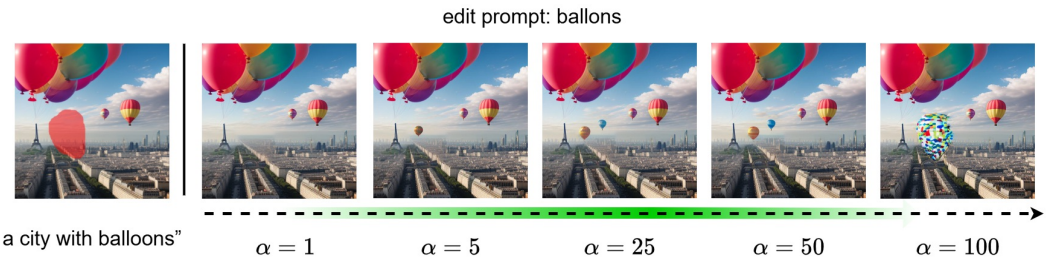


Figure 12. Ablation study on the effect of the strength parameter (α) in LDB. We incrementally increase the mask strength (α) while keeping the mask, seed, and intermediate denoising steps (n) fixed. A value of α that is too large introduces too much noise injection and may cause artifacts, while a value that is too small results in insufficient editing.

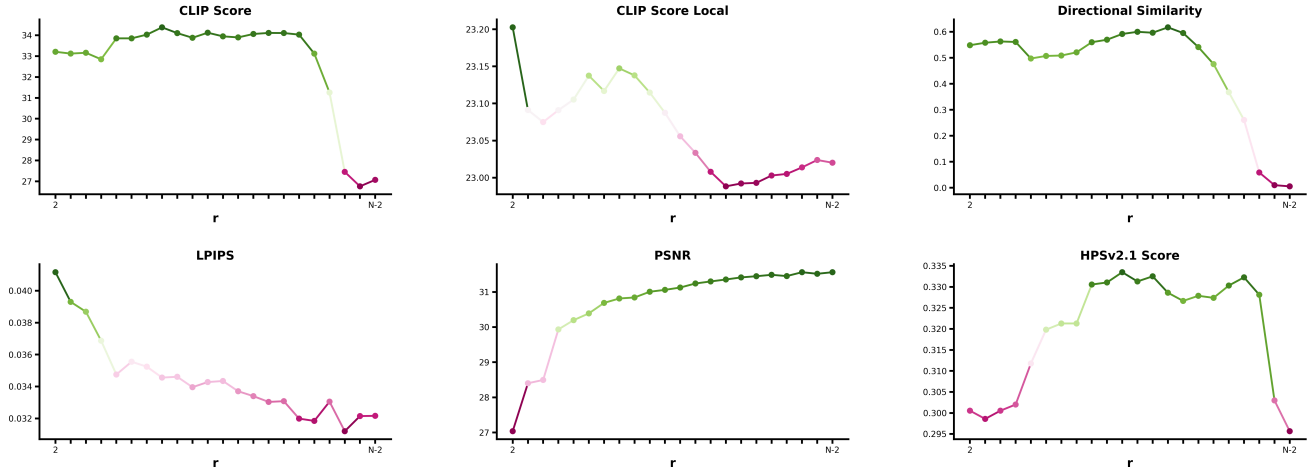


Figure 13. Quantitative evaluation of metrics across different regeneration step values (r). The x-axis represents the regeneration step r, increasing from left to right from 2 to $N - 2$, while the y-axis shows the corresponding score values for each metric.

the user interface (UI) stacked subsequent edits onto the edited image. For IP2P method, users were required to write the instruction prompt and then adjust the image and text guidance scales and regeneration steps to finalize the edit. On the other hand, for the SDI method, users drew a mask and controlled the edit using the strength control. Completion times for each task were recorded for both methods.

Type 2 tasks, corresponding to the MagicBrush dataset [65], were more structured, with the mask, edit prompt, and input images provided by the dataset. MagicBrush utilized crowd workers to collect manual edits using DALL-E 2 [48]. This process involved 5,313 editing sessions and 10,388 editing iterations, resulting in a robust benchmark for instructional image editing. Additionally, the dataset provides manually annotated masks and instructions for each edit and contains up to three layers of edits. Users selected each image, started with the provided mask, could modify the mask if necessary, adjusted the control parameters and prompt, and saved and completed the task for each method.

10.2. Evaluation Survey

After completing the image editing tasks, the participants were asked to complete a three stage evaluation survey. The first part included a System Usability Scale (SUS) form to rate the usability, ease of use, design, and performance of each method. SUS is a standard usability evaluation survey which is widely used in user-experience literature [8]. The participants were presented with 10 questions about each of the methods and were asked to rate each system on a scale of 1 to 5 for each question. A rating of 1 indicated strong disagreement, while a rating of 5 indicated strong agreement. The questions were designed to assess the participants' perceptions of the effectiveness, ease of use, and overall user experience of each tool. Below is the list of the questions:

Q1 I think that I would like to use this tool frequently.
Q2 I found the tool unnecessarily complex.
Q3 I thought the tool was easy to use.
Q4 I think that I would need the support of a technical person to be able to use this tool.
Q5 I found the various functions in this tool were well in-

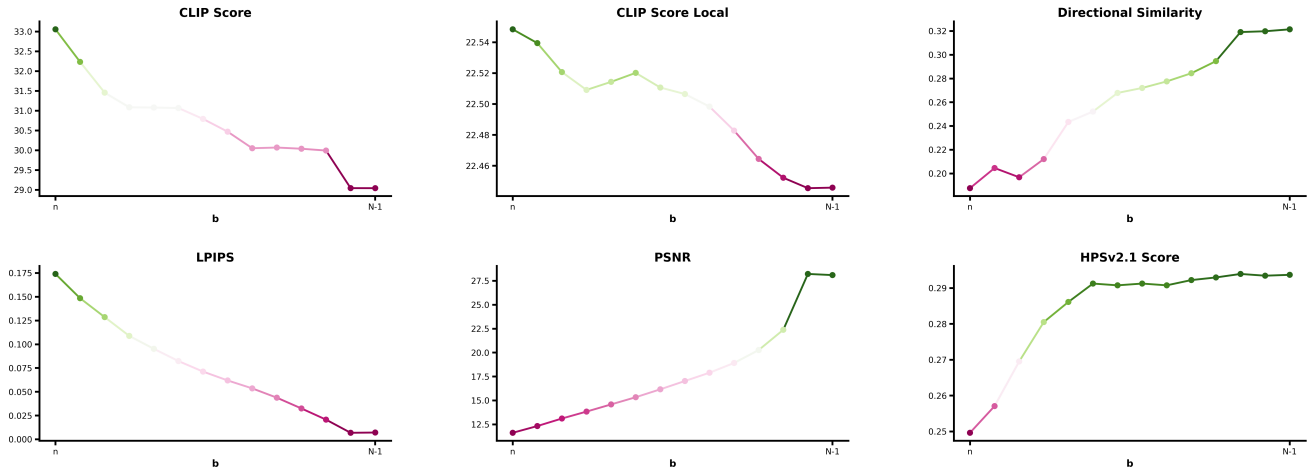


Figure 14. Quantitative metrics for different blending steps (b): The x-axis represents the blending step b, increasing from left to right from $b = n$ to N, while the y-axis shows the corresponding score values for each metric. Smaller b steps lead to poor background protection, while larger b values preserve background integrity and improve edit effectiveness.

tegrated.

Q6 I thought there was too much inconsistency in this tool.

Q7 I would imagine that most people would learn to use this tool very quickly.

Q8 I found the tool very cumbersome to use.

Q9 I felt very confident using the tool.

Q10 I needed to learn a lot of things before I could get going with this tool.

SUS consists of positive and negative phrasing questions. Q2, 4, 6, 8, and 10 are negatively framed, therefore on the chart, red colours means better SUS score and Q1, 3, 5, 7, and 9 are considered positively framed and hence, more green colours demonstrate better score.

The survey was followed by an interview with each participant to gather specific feedback and insights based on their artistic background and experience using the different tools. These processes provided valuable information on the strengths and weaknesses of each tool, as well as how it can be improved to better serve users.

The following multiple-choice questions were also asked for evaluating the performance of each method:

- How much time did it take you to complete the image editing task using the tool you used in this study? [Much less time/About the same/Much more time]
- How did you find each of the tools in terms of effectiveness in achieving the desired edits? [Very effective/Somewhat effective/Neutral/Somewhat ineffective/Very ineffective]
- How does each of the tools you used perform in terms of time to complete the editing task? [Much faster/Somewhat faster/Acceptable/Somewhat slower/Much slower]
- How likely are you to use each of these tools as an AI image editing tool in the future? [Very likely/Somewhat likely/Neutral/Somewhat unlikely/Very unlikely]

The entire study, including filling out the evaluation surveys, took not more than 90 minutes.

Fig. 21 illustrates the outcomes of the post-study CSI survey. Overall, participants expressed positivity towards LDB, indicating that it enhanced their enjoyment, exploration, expressiveness, and immersion, while also deeming the results worth their effort. The CSI score results also show that one participant responded neutrally or negatively to certain aspects, likely due to their being accustomed to the Photoshop tool. Furthermore, there was notable variability in immersion scores, with several participants giving lower ratings. This variability suggests that while some users felt deeply engaged with the tool, others may have encountered challenges or distractions affecting their immersive experience. Analyzing specific factors such as interface design, task complexity, and user preferences could offer insights into enhancing immersion in future iterations of LDB. Despite this variability, the majority of participants found the tool effective and engaging, highlighting its potential usefulness in creative workflows.

One of the most common comments regarding the usability of different methods was that participants found it challenging to find the optimal settings for IP2P and SDI. For example, one user mentioned, “In InstructPix2Pix, increasing the image guidance scale often distorts the edited image too much, and if the text guidance scale is too high, the edited image looks completely different. After many trials and errors, when I find a good combination, the next image behaves differently. Also, SD-inpainting half the times fails to produce a satisfactory result.”

Another user, who is an expert in graphic design, sug-

Figure 15. Video editing examples using LDB and Stable Video Diffusion (SVD). The top row displays frames from an input video generated by SVD. For localized editing, we define a mask on the first frame and apply LDB edits to this initial frame. LDB's caching mechanism is then extended to the temporal dimension within SVD, enabling efficient propagation of edits across subsequent frames. This allows for the creation of multiple editing layers, and even non-sequential fast modifications to different parts of the video by revisiting and adjusting previous layers, while maintaining temporal coherence.

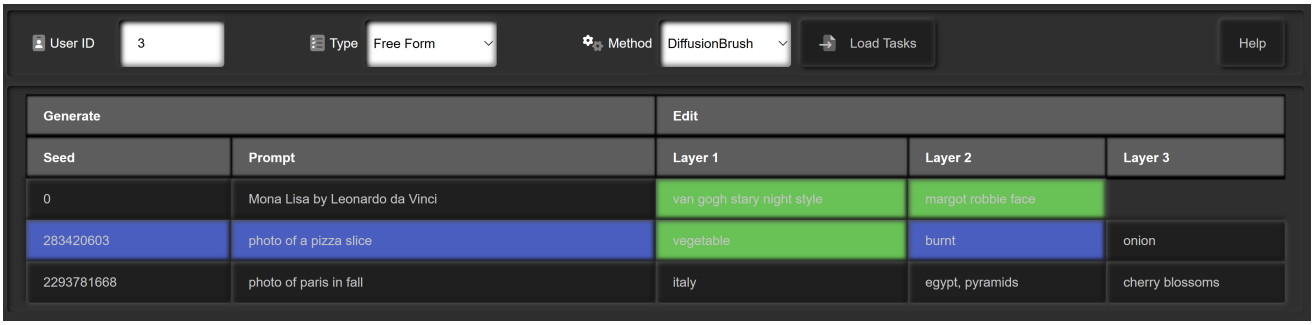


Figure 16. Overview of the tasks section, where users can interact to load, select, and save each task. Tasks that are selected are highlighted in blue, while those completed and saved are highlighted in green.

gested, "Layers are very helpful. I would like to see the control numbers on top of them as I change them, not beside them. Also, having an undo button is crucial and would be very helpful. Additionally, I would suggest adding a blend option to each layer, similar to Photoshop". These suggestions will be taken into consideration for future improvements.

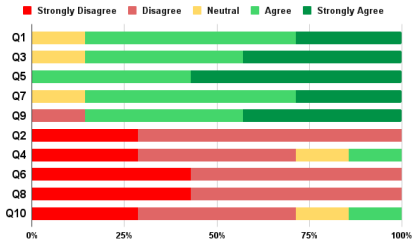


Figure 17. LDB usability

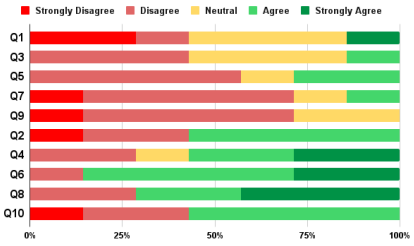


Figure 18. SDI usability

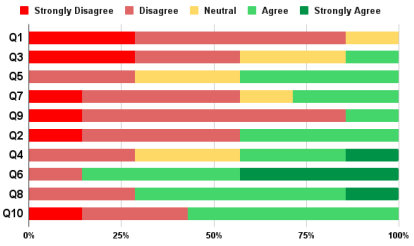
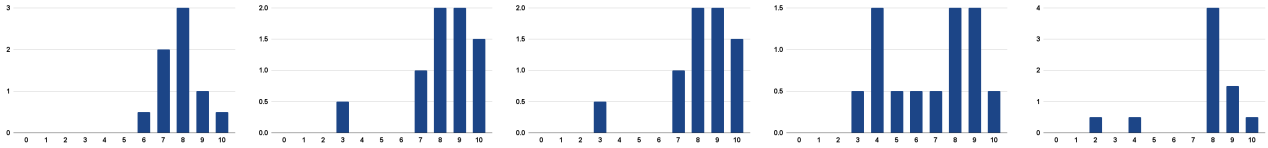


Figure 19. IP2P usability

Figure 20. Results of Q1 - Q10 for the usability of each system among different participants. For odd questions, green colors show more desirable feedback. Even questions are designed with negative wording and more red colors show more favorable feedback.



Enjoyment Expressiveness Exploration Immersion Result Worth Effort

Figure 21. Histogram of the Creativity Support Index from the user study survey.

10.3. System Usability Scale (SUS)

Fig. 20 presents the results of the SUS survey among participants after using LDB, SDI, and IP2P. Based on the bar charts, participants indicated that they are more likely to use LDB compared to IP2P and SD-Inpainting, and that they find it the easiest tool to use. In addition, participants in Q4 expressed that they would not require technical assistance to use the system in the future, indicating its overall good design. These findings were further supported by the interview feedback. For example, when asked about their understanding of the different parameters in the tool, one participant stated: "I believe that I understand the functionality of each parameter. I need to increase the mask strength value if I want to make bigger changes. The tool is quite intuitive and easy to use, and I think I can easily use it without needing any technical support." This feedback highlights that the tool has a user-friendly design and can be easily understood and used by a wide range of users. Based on the survey results, the SUS score for LDB is calculated as 80.35%, while IP2P and SDI achieve a score of 38.21% and 37.5% respectively.

For CSI [11] questionnaire we used all questions, excluding questions about collaboration as it is not relevant for our tool. The CSI measures dimensions of Exploration, Expressiveness, Immersion, Enjoyment, and Results Worth Effort in a tool. CSI helps in understanding how well LDB support creative work overall, as well as pointing out which aspects of creativity support may need attention.

Figure 21 illustrates the outcomes of the post-study CSI survey. Overall, participants expressed positivity towards LDB, indicating that it enhanced their enjoyment, exploration, expressiveness, and immersion, while also deeming the results worth their effort.

11. Initial User Study Insights

We initially developed an earlier version of LDB, called *Diffusion Brush*, with the objective of re-randomizing targeted regions for fine-tuning (*e.g.* fixing small details that were generated incorrectly) and without layering functionalities. Subsequently, we conducted a user study to assess its usability and features and based on the feedback received from this first study, we made significant improvements and revamped the tool. In the first user study, we compared the early version of LDB with SDI and manual editing in Adobe Photoshop [28], involving five expert users.

While the majority of participants acknowledged that Diffusion Brush was faster than manual editing, some participants suggested that even faster editing would be significantly beneficial, aiding in random idea generation for artists. To address this feedback, we incorporated a caching mechanism, as explained in Section 3.1, designed an efficient front-end to communicate with the machine learning backbone, and highly optimized the overall pipeline, achieving as little as 140 ms of inference time for a single edit on a high-end consumer GPU.

Furthermore, a few users struggled with finding the optimal brush strength control, a similar challenge observed in SD-inpainting as well. To address this, we devised a more

generalized approach. While the earlier version of our system also supported multiple masks, these masks were not fully independent, and deleting or hiding them was not possible without performing operations in a specific order. This observation prompted the creation of a more streamlined and flexible mask management system.

Additionally, insights gathered from the first round of interviews indicated the need for further improvement in various aspects of the tool's functionality and user experience. These inputs guided us in refining the tool and enhancing its usability for a wider range of users. Lastly, in the first user study, three participants specifically mentioned this feedback. One participant stated, "I really like the tool as it is right now; it certainly provides value for me in my editing tasks and makes my life easier. But one feature that I would love to see is to be able to tell the system how to make these changes. I still want to use the masking editing, but if I can tell it what to do it would be great.". Based on the findings of the new user study, it is evident that this feature has been well-implemented into the system. All users participating in the current study affirmed the effectiveness of this feature.