PIXELS: Progressive Image Xemplar-based Editing with Latent Surgery

Shristi Das Biswas*[1], Matthew Shreve[2], Xuelu Li[2], Prateek Singhal[2], Kaushik Roy[1]

[1]Purdue University
[2]Amazon Fashion
sdasbisw@purdue.edu, {mashreve, xueluli, prtksngh}@amazon.com, kaushik@purdue.edu

Abstract

Recent advancements in language-guided diffusion models for image editing are often bottle-necked by cumbersome prompt engineering to precisely articulate desired changes. An intuitive alternative calls on guidance from in-the-wild image exemplars to help users bring their imagined edits to life. Contemporary exemplar-based editing methods shy away from leveraging the rich latent space learnt by pre-existing large text-to-image (TTI) models and fall back on training with curated objective functions to achieve the task. Though somewhat effective, this demands significant computational resources and lacks compatibility with diverse base models and arbitrary exemplar count. On further investigation, we also find that these techniques restrict user control to only applying uniform global changes over the entire edited region. In this paper, we introduce a novel framework for progressive exemplar-driven editing with off-the-shelf diffusion models, dubbed PIXELS, to enable customization by providing granular control over edits, allowing adjustments at the pixel or region level. Our method operates solely during inference to facilitate imitative editing, enabling users to draw inspiration from a dynamic number of reference images, or multimodal prompts, and progressively incorporate all the desired changes without retraining or fine-tuning existing TTI models. This capability of fine-grained control opens up a range of new possibilities, including selective modification of individual objects and specifying gradual spatial changes. We demonstrate that PIXELS delivers high-quality edits efficiently, leading to a notable improvement in quantitative metrics as well as human evaluation. By making high-quality image editing more accessible, PIXELS has the potential to enable professional-grade edits to a wider audience with the ease of using any open-source image generation model.

Code — https://github.com/amazon-science/PIXELS

Introduction

Language-guided image generation using large text-to-image (TTI) diffusion models trained on web-scale data has made remarkable strides in recent years (Dhariwal and Nichol 2021; Ho, Jain, and Abbeel 2020; Nichol et al. 2021; Ramesh et al. 2022; Ye et al. 2024; Balaji et al. 2022; Shi et al. 2024). As the need for image editing applications for creating novel content became pervasive in the age of social media, pipelines for creative editing have become a popular demand. Powered by these large-scale pre-trained TTI models, editing methods (Chen et al. 2024b; Cao et al. 2023; Mokady et al. 2023) have been proposed that allow manipulating content in images with the guidance of an input text prompt. However, these editing approaches (Brooks, Holynski, and Efros 2023; Hertz et al. 2022) still find it challenging to fit the requirements of complicated practical scenarios. For instance, as shown in Fig. 1, the method needs to modify the porch just enough to circumvent abrupt transitions when editing with the lake-view exemplar by generating intermediate features that realistically connect them, or even in the case of a creative edit imagining a bud seamlessly replace the given dessert. Flexibility to allow any kind of edits is important for real applications like custom scene design, product creation and placement, special effects, etc.

Inpainting approaches (Yu et al. 2023; Xie et al. 2023) take a binary mask and text guidance to regenerate the masked region using TTI models. However, they are limited by the complexity of prompt engineering to accurately reflect user-desired effects, since detailed object appearance features often cannot be feasibly specified by plain text. To this effect, the need for a "universal language" that enables precise and intuitive control has become ubiquitous. To solve these shortcomings, there have been efforts to drive editing through the use of exemplars (Song et al. 2024; Yuan et al. 2023; Chen et al. 2024a,b) which take as input a binary mask to introduce the main subject from an exemplar into the source image. However, such approaches often struggle to deal with local edits and are strictly restricted to using only one exemplar at a time. In addition, they require pairs of object masks from video frames for their training, which are difficult to obtain at scale and restrict their data diversity during training. Other exemplar-based methods (Yang et al. 2023a; Xu et al. 2023) suffer from similar limitations where they train specialized architectures to learn semantic editing of single objects into an image. We believe that instead of encouraging this trend of training custom pipelines on limited datasets, initiatives should be taken towards "reusing" the information learnt by pre-existing image generation models that have already been trained on much larger datasets.

To satisfy these requirements, we propose a method to

*Work conducted at Amazon.

Figure 1: **In-the-wild editing results** produced by our method where users specify to-edit regions in the source image along with exemplars to inspire the edit. Our method changes different regions of the source by different amounts, according to a given non-binary edit map: the darker the region; more flexibility to adapt to the exemplar. This controllability allows us to create gradual spatial changes (e.g., forest-to-beach transition, top left) and transition across an edit realistically.

improve exemplar-based editing to exercise creativity more freely. Our main contributions are: (1) We allow users to define the 'editability' of each pixel in a picture by different strengths efficiently and simultaneously for seamless edits using a dynamic number of in-the-wild reference images. This is mainly achieved by the insight that selectively modifying various regions at different timesteps during a generation model's inference process induces control over their fidelity to the image on a spatial basis. When introducing an exemplar into the source, say for creating a new image that replaces the leaf-strewn ground with a sandy beach (Fig. 1 top-left), we would like to introduce different amounts of changes into the edited regions on the photo, in a controllable manner. This way, source image pixels closer to the edit boundary attain more flexibility to transition to the new exemplar concept harmoniously. This enables finer granularity across an edit, opening ways the user can exert more control over changes than is attainable using previous methods. (2) We show how to extend our algorithm to introduce any number of exemplars into an image, making it a more suitable candidate for general-purpose editing in contrast to the exemplar count restrictions imposed by other models. (3) Our framework only requires adapting the inference process of existing off-the-shelf image generation models (Rombach et al. 2022; Podell et al. 2023) trained on web-scale data, to leverage their rich latent space and pre-learnt knowledge of contextual relationships between objects for realistic edits. This enables a heightened level of refinement to mitigate any artifacts or unnatural transitions for exemplar-driven editing. (4) We do not forsake the original text-guidance capability of these models, leaving it up to the user to exert any form or degree of control. (5) Finally, our approach performs favorably over the prior art for the task, measured by both quantitative metrics and user evaluation.

Related Works

Text-driven image editing: With recent strides in TTI models, language-guided editing has been largely explored. Traditional approaches (Andonian et al. 2021; Gal et al. 2022; Patashnik et al. 2021; Xia et al. 2021) fall back on pretrained-GAN generators (Karras et al. 2020) and text encoders (Radford et al. 2021) to achieve instance-based optimization according to text prompts. However, these approaches are often cost-ineffective due to progressive optimization steps on the image and struggle with complex editing due to the limited modeling capability of GANs. Recently, diffusion models (DMs) have risen as the new state-of-the-art for TTI generation (Song, Meng, and Ermon 2020; Song et al. 2020). A consecutive line of work (Kim and Ye 2021; Ruiz et al. 2023; Kawar et al. 2023) fine-tuned these DMs for each target text. The high computational cost of fine-tuning case-specifically for every image, however, makes them impractical for interactive image editing. Authors in (Avrahami, Fried, and Lischinski 2023) perform a sequence of noise-denoise operations to create local edits given a mask. More recent works like (Hertz et al. 2022; Liu et al. 2023) still suffer from the lack of precise control with text guidance to explicitly convey the user's imagination.

Exemplar-driven image editing: Parallel to the direction of using text guidance to describe desired edits, a more recent line of work alternatively focuses on guidance from visual exemplars. Early works in using images to guide generation (Meng et al. 2021; Seo et al. 2023) focused on modifying the entire image, making them incompatible for edits with local strength control. Methods such as (Yang et al. 2023a) and (Song et al. 2023) extract representations from the reference image using a CLIP encoder to inpaint into a source image while (Ye et al. 2023) trains an image-adapter

Figure 2: **Illustration of transition artifacts in the naive solution**, generating unrealistic edits.

to achieve visual prompt capability for pre-trained text-to-image models. Later works (Xu et al. 2023; Chen et al. 2024a) train specialized architectures with curated objective functions to solve the editing task. While they allow exemplar-guidance, the four major limitations of these methods lie in the fact that firstly, they only allow uniform change across an image, or at most, allow partitioning the picture into an unchanged and a changed region according to a binary map. This is sub-optimal, given that introducing a new concept into the source image with only binary change capability often leads to abrupt transformations between the retained regions and the introduced regions. Secondly, they fail to make use of pre-existing foundational TTI models such as (Rombach et al. 2022; Podell et al. 2023; Razzhigaev et al. 2023) which contain rich latent spaces that can conceptually bridge source and introduced context. Thirdly, these methods often take away the original ability of the model to guide generation using text while consuming large computing resources to specifically train for image editing functionality. Finally, the existing methods are strongly limited by the maximum number of exemplars they have been trained to work with (in most cases, just one). To deal with these shortcomings, we propose a general-purpose image editing approach that leverages off-the-shelf TTI models.

Method

Our goal is to improve exemplar-based image editing along two key dimensions: controllability (down to arbitrary image regions) and quality (adherence and realism). Our proposed method achieves both by enabling users to define the strength of change at each pixel via a non-binary edit map.

Background and Setup. Diffusion models (DMs) are a class of generative models capable of image-to-image translation using a forward process to gradually add Gaussian noise to the image for a fixed Markov chain of t_{ds} steps, and a reverse process to gradually denoise the corrupted image with a learnable model (Ho, Jain, and Abbeel 2020). Denoising strength t_{ds} used controls the timestep at which we start the reverse diffusion process. In this paper, we use their efficient successor, Latent Diffusion Models (Podell et al. 2023) that operate on a lower dimensional representation (latent space), using latent encoders and decoders to translate the input image and output latent across pixel-latent manifolds.

A Naive Solution. A naive solution for this task is to use a binary mask to directly replace the masked pixels in the source image with the exemplar condition. This "surgical latent" is then allowed to naturally denoise through the reverse diffusion process in a TTI model for projecting the latent into the real image distribution. However, applying to test images, we find that the generated edit is far from satisfactory. There exist obvious transition artifacts across the edit boundary, making the result look extremely unnatural, as seen in Fig. 2. We argue that this is because typical image generation models have not been trained with image editing objectives. Hence, with this naive scheme applied during inference, the model fails to adapt the exemplar to the source nor successfully hallucinate how these latents can co-exist realistically when placed next to each other by latent surgery.

Traversing to the Real Image Manifold. Our key hypothesis is as follows: Given a surgical latent $z_{surgical}$, created between the source and exemplar images, we first perturb it with Gaussian noise to induce smoothing out of transition artifacts across the edit. We start by sampling from $z_{surgical}(t_{ds}) \sim N(z_{surgical}, \sigma^2(t_{ds})I)$ (Song, Meng, and Ermon 2020) starting at any denoising strength $t_{ds} \in (0, T)$ (lower the t_{ds}, later its inference begins). This is followed by progressively removing the noise by reverse diffusion. This process allows mapping data from the noised surgical latent distribution to a latent state $z_{surgical}(0)$ (abbreviated as $z(0)$ hereafter) in the manifold of realistic images.

We aim to bound expected squared distance between the initial latent $z_{surgical}(t_{ds})$ and its final state $z(0)$, to estimate the adherence between them. Supposing the latent decoder model is K-Lipschitz, we have that similar latent states lead to similar enough decoded images. Formally, $\|z_{surgical} - z(0)\| \leq K\|x_{surgical} - x(0)\|$, where $x_{surgical}$ is the decoded surgical image with copy-paste artifacts (see edit results in Fig. 2) and $x(0)$ is the generated realistic edit. In this paper, we present an algorithm for progressive editing to enforce constrained traversal between a surgical latent and the generated latent (with desired edit) lying in the real image distribution. This translates to constrained change in the image space, ensuring high adherence between the resulting edit and original images. Following (Yang et al. 2023b), we assume that a finite value B exists s.t $\forall z, B = \sup ||s_\theta(z, \sigma))||^2$ where s_θ is the model learnt during DM training. We present the following lemma:

Lemma 1: For all $p \in (0, 1)$, we have,

$$\mathbb{P}\Big(\mathbb{E}\big[\|z(0) - z_{surgical}(t_{ds})\|^2\big] \leq \sigma^4(t_{ds})B + \sigma^2(t_{ds})\big(k + 2\sqrt{-k\log p} - 2\log p\big)\Big) \geq (1-p) \tag{1}$$

where $z_{surgical}$ has k degrees of freedom. For proof of lemma 1, please refer to the Appendix. It provides an upper bound on the change from the unrealistic surgical latent to the realistic edited image as a function of the denoising strength. We find this to be a two-way street, defining how the expected difference and hence adherence between the surgical latent and the generated latent (measured by their distance in the latent space) can be controlled by t_{ds}, while also allowing intuition on the choice of t_{ds} based on the expected difference between the latents. If the surgical latent is an unrealistic composition, e.g, created between images taken in summer and winter, then even the nearest edited image in the real image manifold could be at a much larger distance, indicating that we must increase the denoising strength to allow farther traversal between the

Figure 3: **Visualization of Algorithm 1 Line 15 over time.** Top: $z_1^t \odot mask_t$, regions copied from a noised version of the input. Bottom: $z_{mix}^{t+1} \odot (1 - mask_t)$, residue regions copied from the U-Net output in previous step. Note how the shifting mask at each timestep controls the inference process - darker the corresponding region in the edit map, the earlier it is copied from the residue. For ease of understanding, images are shown in the pixel space instead of the latent space.

surgical latent and real image distributions. In this case, the distance between the latents increases as denoising strength is increased, encouraging reduced adherence to the exemplar while allowing added flexibility to achieve realism.

Local Strength Control Using Edit Maps. Inspired by (Levin and Fried 2023), our method takes this controlled traversal a step further. Instead of a uniform scalar t_{ds} applied to the entire edited region, our algorithm works with local strength control, defining t_{ds} for every pixel in the edit, to enable greater flexibility and finer-grained control over the output. In the previous example, when progressively editing between summer and winter images such that the scene transitions between seasons across the width of the image, selectively modifying different regions starting at different timesteps during the diffusion inference can control spatial fidelity to the original images. Moving from the summer side to the winter side, pixels are progressively allowed more denoising strength, thus creating a smooth transition across concepts. However, to make our approach compatible with off-the-shelf DMs that only enable a scalar denoising strength setting, we design our method to instead implement this control using a non-binary 'edit map' matrix for specifying the quantity of change at each spatial location.

Algorithm for Progressive Image Editing. Our method takes over the inference algorithm during reverse diffusion steps(Algorithm 1). We start by encoding the source and exemplar images to the latent space (z_1^{init} and z_2^{init} respectively), while the edit map is downsampled to the spatial dimensions of the latent tensors. As examined by (Levin and Fried 2023), we find latent encoders used in (Podell et al. 2023; Razzhigaev et al. 2023) typically encode pixels to the same relative positions. This allows the down-sampled map (μ_d) to align with positions of pixels in the latent representation. The highest edit strength t_{ds}^{max} is set by the maximum number of inference steps. We start with the surgical latent at line 10. At each timestep t in the denoising loop, latents are noised to the current timestep and the regions taken directly from the noisy source decreases, while more of the latent is composed of pixels copied from the denoising output in previous steps. This transition is gradual and controlled by a mask of all points lower than the current threshold determined by normalized timestep count

Algorithm 1: Progressive Image Editing

1: **Input:** x_1 (source image), x_2 (exemplar), μ (edit map), $t_{ds}^{max} = T$ (maximum denoising strength), $p =$"" (prompt)
2: **Output:** $\hat{x}$
3: **procedure** INFERENCE(x_1, x_2, μ, T, p)
4: $z_1^{init} \leftarrow$ ldm_encode(x_1)
5: $z_2^{init} \leftarrow$ ldm_encode(x_2)
6: $\mu_d \leftarrow$ down_sample(μ)
7: $z_1^T \leftarrow$ add_noise(z_1^{init}, T)
8: $z_2^T \leftarrow$ add_noise(z_2^{init}, T)
9: $mask_T \leftarrow \mu_d > 0$
10: $z_{mix}^T \leftarrow z_1^T \odot mask_T + z_2^T \odot (1 - mask_T)$
11: $z_{mix}^T \leftarrow$ denoise(z_{mix}^T, p, T)
12: **for** $t = T - 1$ **to** 0 **do**
13: $z_1^t \leftarrow$ add_noise(z_1^{init}, t)
14: $mask_t \leftarrow \mu_d > (T - t)/T$
15: $z_{mix}^t \leftarrow z_1^t \odot mask_t + z_{mix}^{t+1} \odot (1 - mask_t)$
16: $z_{mix}^t \leftarrow$ denoise(z_{mix}^t, p, t)
17: **end for**
18: $\hat{x} \leftarrow$ ldm_decode(z_{mix}^0)
19: **return** $\hat{x}$
20: **end procedure**

(depicted in Fig. 3). This linearly shifting mask determines when each region begins inference: pixels that start inference earlier (in the darker regions) have lesser "memory" of the source concept and attain more flexibility to adapt to the exemplar concept, while it works the other way around the later you start inference. This gradual exposure is key to how the model can generate intermediate features that naturally connect and transition across the edit seamlessly. Because DMs are not trained on intermediate images with holes, this process mimics its training distribution and gives it advance knowledge of the content in lighter regions. After the loop, the U-net output z_{mix}^0 is decoded to the pixel space as generated edit $\hat{x}$. Our method can be extended to an arbitrary number of exemplars by replacing lines 10 and 15 in Algorithm 1 with nested functions for each new exemplar. See the Appendix for the explicit algorithm.

Spatially Adjusting Noise. A simple baseline involves spatially adjusting the added noise magnitude to generate similar edits to our method. However, we find that multiply-

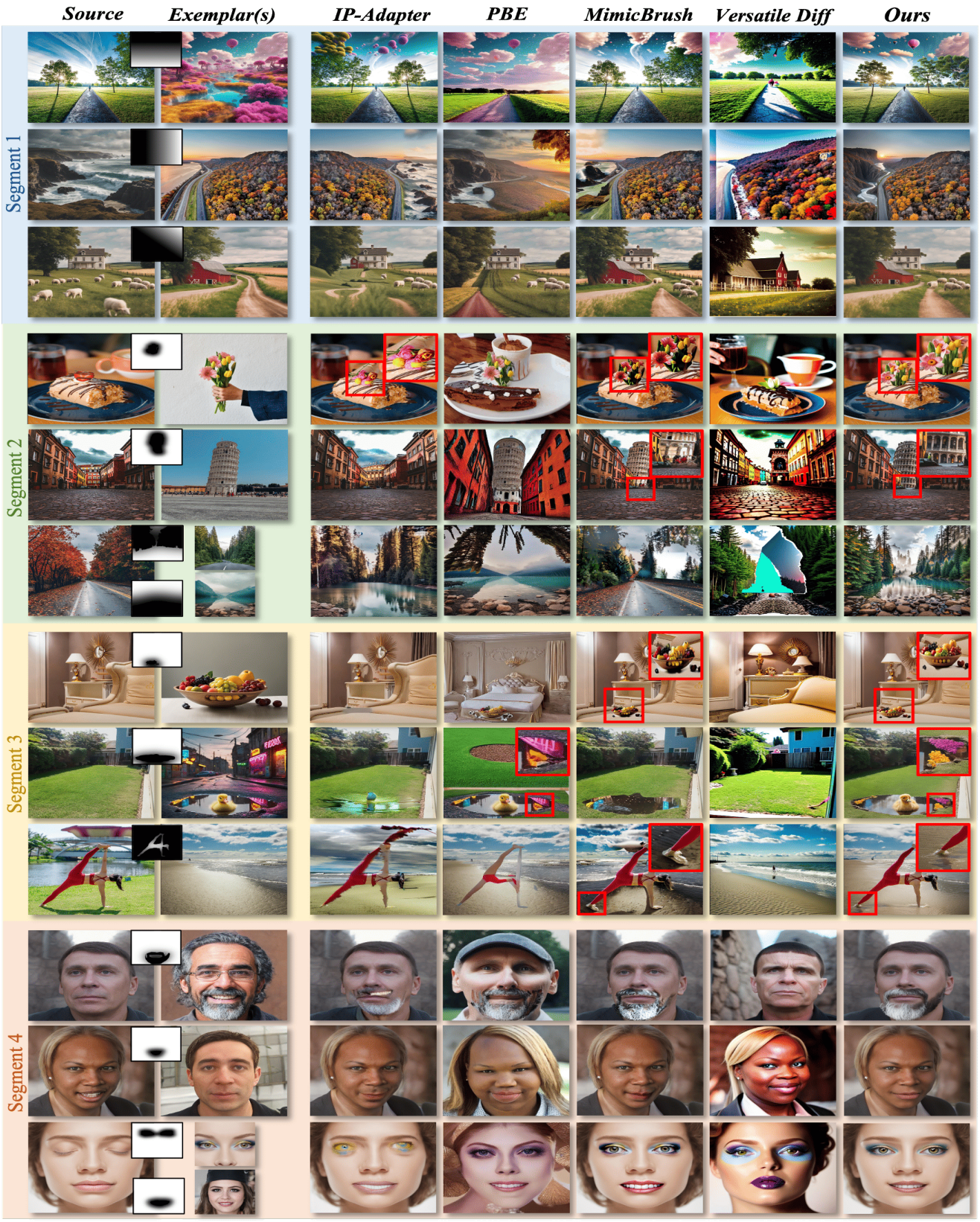

Figure 4: **Qualitative comparisons.** Our method can create realistic edits with high source and exemplar consistency.

(a)

Method	CLIP-I Score (↑)	FID (↓)
Versatile Diffusion (Xu et al. 2023)	68.79	18.174
Paint-By-Example (PBE) (Yang et al. 2023a)	78.67	8.543
IP-Adapter (Ye et al. 2023)	81.32	7.958
MimicBrush (Chen et al. 2024a)	84.95	7.915
PIXELS (ours)	**91.71**	**5.412**

(b)

Method	Adherence	Realism
Versatile Diffusion (Xu et al. 2023)	4.93	4.03
Paint-By-Example (PBE) (Yang et al. 2023a)	3.07	3.10
IP-Adapter (Ye et al. 2023)	3.30	3.50
MimicBrush (Chen et al. 2024a)	2.27	2.97
PIXELS (ours)	**1.40**	**1.40**

Table 1: **(a) Quantitative comparison across methods.** We evaluate the generated image quality using FID score and semantic consistency to the exemplar through CLIP-I score. **(b) User study results.** Average ranking score for semantic adherence and visual realism. 1 is the best, 5 is the worst.

ing the added noise by an edit map tends to cause convergence to random single-color images. This is intuitive since the model has been trained to handle a specified noise distribution, which this scheme disrupts.

Implementation Details. For experiments, we use off-the-shelf SDXL (Podell et al. 2023) for denoising. However, the algorithm can be generalized across any DM (See the Appendix). We do not assume anything about source of the edit maps and find that they can be easily generated by operations like growing and blurring, or simple histogram transformations on binary masks created using tools like Language Segment-Anything, user interaction, automatic depth maps (Miangoleh et al. 2021) etc. Unless stated otherwise, we keep text prompts empty ($p =$"") for all experiments.

Results

We start by comparing to existing works both qualitatively and quantitatively. Next, we demonstrate how our pipeline is compatible with multi-modal prompts for editing.

Evaluation on Exemplar-Driven Editing. We select different categories of edits to evaluate our method. As seen in Fig. 4, our method can deal with edits across different topics and domains. Segment 1 shows examples of scene composition, useful for new image creation where the user wants to perform general edits to compose new scenes. Segment 2 illustrates the application of semantic edits with in-the-wild images. It should be noticed how our method generates seamless results by hallucinating realistic interactions between the edited-concept and the original image. This is made possible by the shifting mask which allows creating river banks and imagining reflection from the trees near the edit boundary while encouraging high source and exemplar fidelity elsewhere. Contemporaries such as PBE (Yang et al. 2023a), IP-Adapter (Ye et al. 2023), and Versatile Diffusion (Xu et al. 2023) do not guarantee fidelity or realism for editing in-the-wild images. MimicBrush (Chen et al. 2024a) generates more realistic results, but we still observe undesired changes to the exemplar and background. Segment 3 illustrates foreground-inpainting. In the third example in this segment, we show a variation to outpaint the background instead of foreground-inpainting and create realistic interactions between the human and the beach, proving our strong generalization ability compared to prior works. In the last segment, we show comparisons in practical applications like face edits, and makeup look curation from inspiration exemplars. For visualizations from contemporaries in more-than-one exemplar setting, we edit iteratively to incorporate each new exemplar. Our method, on the other hand, shows significant superiorities by being able to perform the edit in a single pass, with as little as 0.04% inference memory overhead on the original SDXL. More results in the Appendix.

For fair quantitative comparison, we sample 3000 pairs of random images from Imagenet's validation set (Russakovsky et al. 2015) as inputs to all methods, with a randomly chosen edit map from our database. Since we are the first method to allow non-binary edit maps, we test other methods with the binarized version of the map. To measure the aspects of photo-realism and edit fidelity (to the source and exemplar) independently, we use the FID score (Heusel et al. 2017), which is widely used to evaluate realism in generated images and the CLIP-I score (Radford et al. 2021), measuring similarity between edited region and the exemplar. Table 1a presents quantitative comparison results. Our approach achieves the best performance on both metrics, verifying that it can not only generate high-quality edits but also maintain high fidelity to the original images.

User Study. As quantitative metrics do not fully represent human preferences, we conduct a user study. We let 26 participants assess and rank generated results on 30 groups of samples, with each group containing the source, exemplar(s), edit map(s) and anonymized outputs from all the methods (presented in a random order). Users were asked to rank the outputs from 1 to 5 (1 being best, 5 is worst) on the criterion of adherence and realism. Adherence refers to the ability to preserve the identity of the source and exemplars according to the edit map. Realism considers if the output looks like a natural and seamless edit of good quality. These aspects are evaluated independently. Results are listed in Table 1b; our method earns significantly more preferences.

Bringing Back Text Control. In most existing models for this task, the original text-to-image ability is lost. However, with our proposed algorithm, we can generate edits with multi-modal prompts (exemplars and/or text guidance) since we do not disrupt the original TTI model with training/ fine-tuning. Instead of using empty text prompts like we have been doing, we can use additional language guidance to generate more diverse edits. As seen in Fig. 5, we can edit attributes, change the generation style or do both during denoising by adding simple text descriptions.

Ablation Studies

Analyzing the impact of t_{ds}. We note that for trained diffusion models, there exists an inherent trade-off between adherence and realism metrics with changing denoising strength. Given a source image and increasingly distant exemplars in the latent space (left to right), we aim to create edits that evaluate to at least a minimum realism score (Gu et al. 2020). As visualized in Fig. 6, for the same edit map,

Figure 5: **Results with multi-modal prompts.** Text Prompts: "boat", "watercolor style", "low poly aesthetic, big monster".

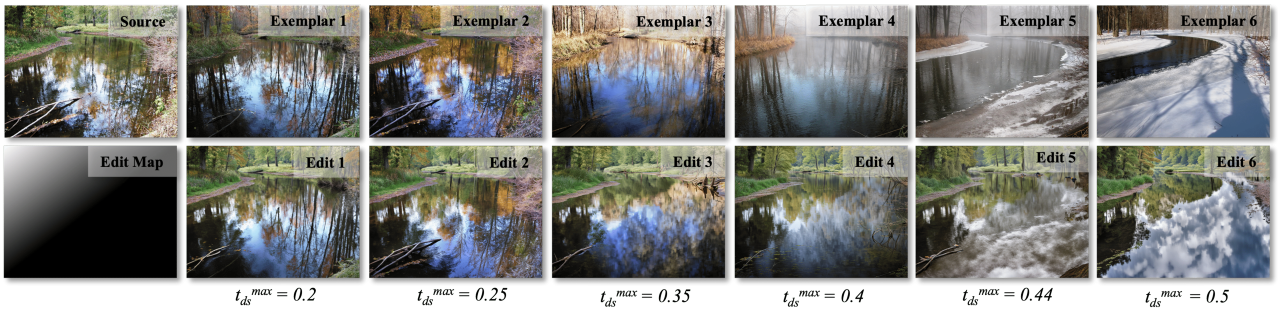

Figure 6: **Trade-off between adherence and realism as source and exemplar images grow further in the latent space (Euclidean distance)**. To reach the minimal realism score, we keep increasing t_{ds}^{max} from left to right as latent distance increases (Lemma 1), causing the edits to show lower fidelity to the exemplar while increasing hallucination to create a realistic scene.

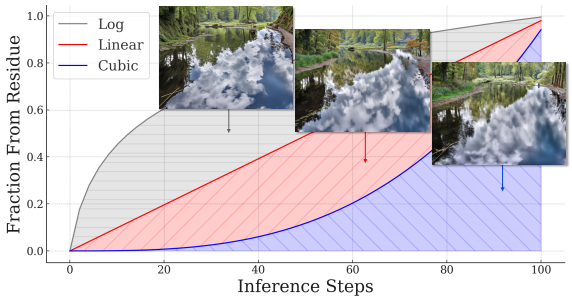

Figure 7: **Thresholding Strategies.** Setting various thresholds allow users to finetune their edits at a given strength.

we increase t_{ds}^{max} to maintain realism in the edit as we go from left to right, allowing the model to exert more hallucination to imagine realistic compositions while becoming less and less adherent to the exemplar. Notice how for exemplar 6, the model leverages the high strength to hallucinate the snowy bank as a cloudy sky for the edit to look natural with high probability (lemma 1), since the surgical latent has an unrealistic composition with winter and summer scenes co-existing in the same image. More details in the Appendix.

Mask Thresholding Strategies. Selecting t_{ds} provides a coarse mechanism for controlling the amount of noise perturbation and hence, the edit strength. We introduce an additional fine adjustment parameter that allows users to precisely modulate the generation results through the choice of an appropriate threshold function, as explored in Fig. 7. At a constant denoising strength, this effectively decides the relative amount of time each noised region corresponding to a brightness level from the edit map spends inside the inference loop (shown as Area Under Curve). Besides the default linear threshold that works well across edits and is hence used for all experiments in this paper, we find that users can switch to Log control to encourage higher fractions of the latent z_{mix}^t to be copied from the U-net residue (hence, more flexibility to hallucinate), or use Cubic control (more Area Above Curve) to rally for copying noisy latent regions from the original image for more steps (hence higher adherence). See the Appendix for more visualizations and details.

Limitations & Potential Impacts

Our method can work with in-the-wild images, allowing for easy photo manipulation and creating content with negative societal impacts. On the other hand, this enables everyday users with little to no artistic expertise to create realistic edits with ease, lowering the barrier to entry for visual content creation. To address this, we will specify permissible uses of the method with appropriate licenses during code release.

Despite the effectiveness of our method, the user creates the edit map for now, using manually drawn masks, depth, or segmentation maps. In the future, tools for automating map generation can help in widespread adoption of our method.

Conclusion

We introduced a novel solution to the exemplar-driven editing task that allows region-wise control over the editing process and demonstrated its superiority over existing baselines. Our method requires no training/fine-tuning, has minimal overhead, can be used with in-the-wild images as well as text prompts, and can work with an arbitrary number of exemplars. This work hopes to bring new inspiration for the community to explore advanced solutions designed by leveraging existing large models for a more sustainable future.

References

Andonian, A.; Osmany, S.; Cui, A.; Park, Y.; Jahanian, A.; Torralba, A.; and Bau, D. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

Avrahami, O.; Fried, O.; and Lischinski, D. 2023. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4): 1–11.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; and Liu, M. 2022. Ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv 2022. *arXiv preprint arXiv:2211.01324*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Chen, X.; Feng, Y.; Chen, M.; Wang, Y.; Zhang, S.; Liu, Y.; Shen, Y.; and Zhao, H. 2024a. Zero-shot Image Editing with Reference Imitation. *arXiv preprint arXiv:2406.07547*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gu, S.; Bao, J.; Chen, D.; and Wen, F. 2020. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 369–385. Springer.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kim, G.; and Ye, J. C. 2021. Diffusionclip: Text-guided image manipulation using diffusion models.

Levin, E.; and Fried, O. 2023. Differential diffusion: Giving each pixel its strength. *arXiv preprint arXiv:2306.00950*.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 289–299.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Miangoleh, S. M. H.; Dille, S.; Mai, L.; Paris, S.; and Aksoy, Y. 2021. Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2085–2094.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Razzhigaev, A.; Shakhmatov, A.; Maltseva, A.; Arkhipkin, V.; Pavlov, I.; Ryabov, I.; Kuts, A.; Panchenko, A.; Kuznetsov, A.; and Dimitrov, D. 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image

diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Seo, J.; Lee, G.; Cho, S.; Lee, J.; and Kim, S. 2023. Midms: Matching interleaved diffusion models for exemplar-based image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2191–2199.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8543–8552.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; and Aliaga, D. 2023. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18310–18319.

Song, Y.; Zhang, Z.; Lin, Z.; Cohen, S.; Price, B.; Zhang, J.; Kim, S. Y.; Zhang, H.; Xiong, W.; and Aliaga, D. 2024. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8048–8058.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2256–2265.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Xu, X.; Wang, Z.; Zhang, G.; Wang, K.; and Shi, H. 2023. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7754–7765.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023a. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yang, Z.; Feng, R.; Zhang, H.; Shen, Y.; Zhu, K.; Huang, L.; Zhang, Y.; Liu, Y.; Zhao, D.; Zhou, J.; et al. 2023b. Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*.

Ye, F.; Liu, G.; Wu, X.; and Wu, L. 2024. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6648–6656.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*.

Yuan, Z.; Cao, M.; Wang, X.; Qi, Z.; Yuan, C.; and Shan, Y. 2023. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*.