# An Item is Worth a Prompt: Versatile Image Editing with Disentangled Control

**Aosong Feng[1], Weikang Qiu[1], Jinbin Bai[2], Zhen Dong[3], Kaicheng Zhou[3], Xiao Zhang[3*], Rex Ying[1*], Leandros Tassiulas[1*]**

[1] Yale University, New Haven, USA
[2]National University of Singapore, Singapore
[3]Collov Labs
{aosong.feng, weikang.qiu, rex.ying, leandros.tassiulas}@yale.edu, jinbin.bai@u.nus.edu, zhendong@berkeley.edu, {xiao, caseyz}@collov.com

## Abstract

Building on the success of text-to-image diffusion models (DPMs), image editing has emerged as a crucial application for enabling human interaction with AI-generated content. Among various editing techniques, prompt-based editing has garnered significant attention for its capacity to simplify semantic control. However, because diffusion models are typically pretrained on descriptive text captions, directly modifying words in text prompts often results in entirely different generated images, which undermines the objectives of image editing. Conversely, existing editing methods often employ spatial masks to maintain the integrity of unedited regions, but these are frequently disregarded by DPMs, leading to disharmonious editing outcomes. To address these two challenges, we propose a method that disentangles the comprehensive image-prompt interaction into multiple item-prompt interactions, with each item associated with a uniquely learned prompt. The resulting framework, named D-Edit, leverages pretrained diffusion models with disentangled cross-attention layers and employs a two-step optimization process to establish item-prompt associations. This approach allows for versatile image editing by enabling targeted manipulations of specific items through their corresponding prompts. We demonstrate state-of-the-art results in four types of editing operations including image-based, text-based, mask-based editing, and item removal, covering most types of editing applications, all within a single unified framework. Notably, D-Edit is the first framework that can (1) achieve item editing through mask editing and (2) combine image and text-based editing. We demonstrate the quality and versatility of the editing results for a diverse collection of images through both qualitative and quantitative evaluations.

## Introduction

The recent advancements in text-to-image diffusion generative models represent a cutting-edge approach in the field of generative models. By gradually introducing noise into the image, these models facilitate sophisticated image synthesis (Podell et al. 2023; Ruiz et al. 2023; Song, Meng, and Ermon 2020) while preserving semantic alignment with the text prompt. One notable application is image editing, where diffusion models provide unprecedented control over various editing tasks, including inpainting (Nichol et al. 2021; Avrahami, Fried, and Lischinski 2023), text-guided editing (Hertz et al. 2022; Parmar et al. 2023), pixel editing (Mou et al. 2023; Brooks, Holynski, and Efros 2023), etc. Various types of editing can generally be evaluated based on two key criteria: preservation of the original image's information and fidelity or consistency with the target guidance. An effective image editing process should prioritize retaining essential information from the original image while ensuring precise semantic alignment with the intended modifications.

To improve consistency with the target guidance, some work (Yang et al. 2023; Chen et al. 2023; Shen et al. 2023; Xue et al. 2022) encodes reference images by introducing additional trainable encoders to preserve identities of the reference, and adds additional controls to DPMs using methods like ControlNet (Zhang, Rao, and Agrawala 2023). However, such methods cannot incorporate the existing text prompt control flow in DPMs and therefore require large-scale pretraining which is usually costly and domain-specific. To preserve information about the original image and improve harmonization, another line of work fixes diffusion sampling trajectory (by setting random seed or using DDIM) and achieves editing by carefully tuning text prompts (Mokady et al. 2023; Miyake et al. 2023), changing part of the trajectory (Meng et al. 2021), merging trajectories (Lu, Liu, and Kong 2023; Wallace, Gokul, and Naik 2023), or optimizing the latent pixel space (Mou et al. 2023; Shi et al. 2023). This avoids additional pretraining but either relies on careful source prompt design to match the editing region or additional optimization per edit.

In this work, we propose two key techniques aimed at enhancing the aforementioned criteria: (1) **Disentangled Control**: To preserve the original image's information, the editing of a target item should minimally impact surrounding items. The control process from prompt to image should also be disentangled, ensuring that modifications to an item's prompt do not interfere with the control flow of other items. Recognizing that text-to-image interactions occur within the cross-attention layers of attention-based diffusion models, we propose a grouped cross-attention mechanism to disentangle the control flow between prompts and items. (2) **Unique Item Prompt**: To enhance consistency with the guidance (e.g., a reference image), each item should be associated with a unique prompt that directs its generation. These prompts typically involve special tokens or rare words. Pre-

*Corresponding author

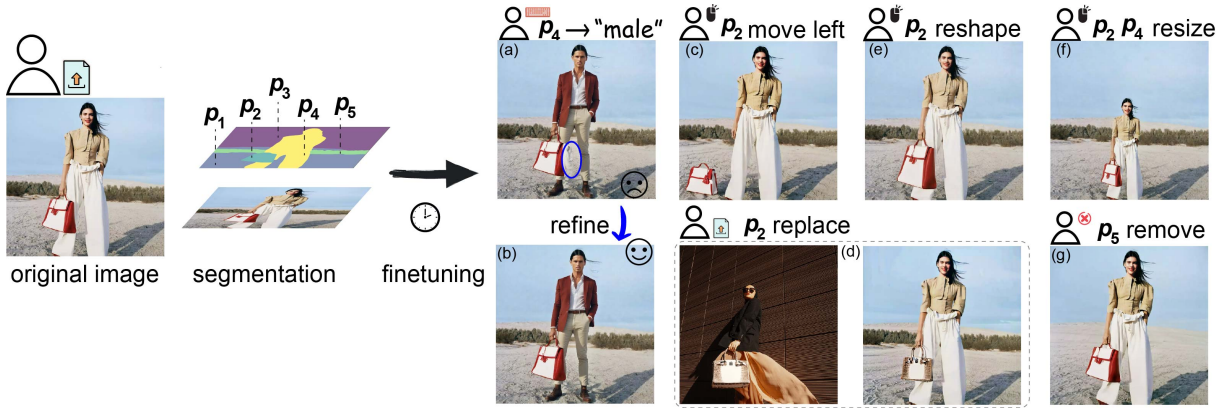


Figure 1: The editing pipeline of using D-Edit. The user first uploads an image which is segmented into several items. After finetuning DPMs, the user can do various types of control, including (a) replacing the model with another using a text prompt; (b) refining imperfect details caused by segmentation; (c) moving bags to the ground; (d) replacing the handbag with another one from a reference image; (e) reshaping handbag; (f) resizing the model and handbag; (g) removing background.

vious works on image personalization, such as Dreambooth (Ruiz et al. 2023) and Textual Inversion (Gal et al. 2022) have explored this concept by representing a new subject with a unique prompt, which is then used for image generation. In contrast, our approach employs independent prompts to define individual items rather than the entire image. Ideally, if each item in the image, with all its details, could be precisely described by a unique English word, users could achieve various editing tasks simply by swapping the current word for the desired one.

By fully harnessing the potential of prompt uniqueness and disentangled control, we introduce a versatile image editing framework for diffusion models called Disentangled-Edit (D-Edit). This unified framework enables a wide range of image editing operations at the item level, including text-based, image-based, mask-based editing, and item removal. As illustrated in Fig. 1, the process begins with segmenting the target image into multiple editable items (in this context, background and unsegmented regions are also referred to as items), each associated with a prompt composed of several new tokens. The associations between prompts and items are established through a two-step fine-tuning process, which optimizes both the text encoder's embedding matrix and the UNet model's weights. To disentangle prompt-to-item interactions, we introduce grouped cross-attention, which isolates attention calculation and value updates. This allows users to achieve various types of image editing by modifying prompts, items, and their associations, as well as by adjusting corresponding masks. This flexibility opens up a wide range of creative possibilities and offers precise control over the editing process. We demonstrate the versatility and performance of our framework across four image editing tasks, utilizing both Stable Diffusion and Stable Diffusion XL. We summarize our contribution as follows:

- We propose to establish item-prompt association to achieve item editing.
- We introduce grouped cross-attention to disentangle the controlling flow in diffusion models.
- We propose D-Edit as a versatile framework to support various image-editing operations at the item level, including text-based, image-based, mask-based, editing, and item removal. D-Edit is the first framework that can do mask-based editing, and perform text and image-based editing at the same time.

## Related Works

**Trajectory-Based Editing**. Because natural language cannot perfectly describe a given image, a single prompt may correspond to multiple sampling trajectories with different random seeds. SDEdit (Meng et al. 2021) achieves editing by sharing the former part of the sampling process to preserve the high-level information like layout and changing the latter part for realistic reconstruction. Diffusion inversion (Mokady et al. 2023; Huberman-Spiegelglas, Kulikov, and Michaeli 2023; Lu, Liu, and Kong 2023) inverts the reverse diffusion process and forces the original and edited trajectory to share the same sampling starting point (picky to the sampling method). Interactions between the two trajectories can then be built by sharing cross/self-attention to preserve the original identity (Tumanyan et al. 2023; Mou et al. 2023). Combined with P2P (Hertz et al. 2022), these methods can achieve fine-grained text-based editing, but it requires an accurate captioning prompt of the original image to reverse the diffusion, and the prompt can only change by a few words. Our methods are agnostic to sampling trajectories, don't require any prior prompt, and support more freedom to change the prompt.

**Image Identity Extraction**. Because image-based editing involves additional modalities for conditioning, the natural thought is to introduce additional encoders to encode the corresponding modalities. Paint-by-example (Yang et al. 2023) trains additional MLP layers following a pretrained CLIP image encoder to encode reference image information. AnyDoor (Chen et al. 2023) design an additional identity extractor to preserve the original item identity. PCDMs (Shen et al. 2023) introduced additional layers to encode the source image and target position. Because of the introduction of the additional trainable module, these models have to train on a

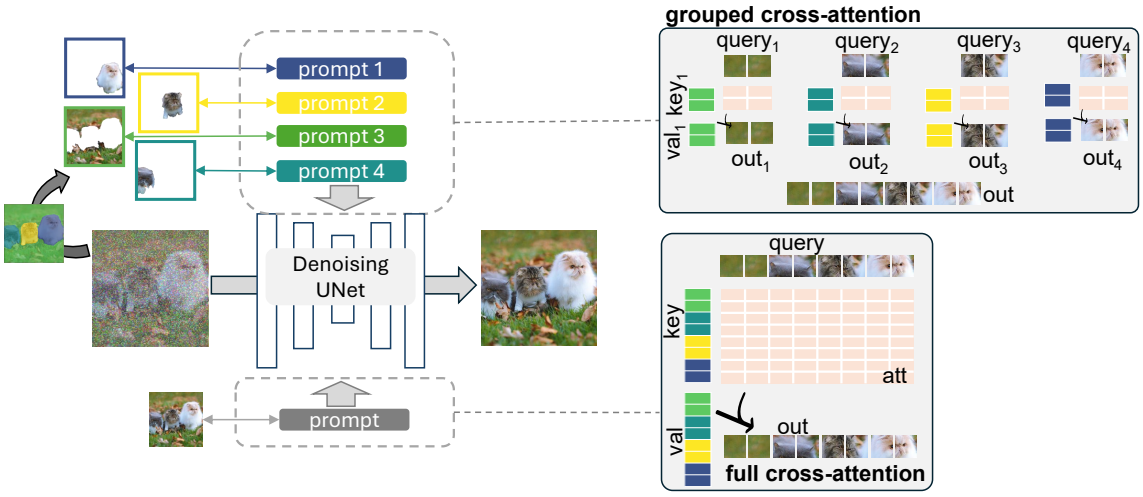


Figure 2: Comparison of conventional full cross-attention and grouped cross-attention. Query, key, and value are shown as one-dimensional vectors. For grouped cross-attention, each item patch only attends to the text prompt assigned to it.

large dataset to perform well on more images. Our method leverages the original pretrained text-encoder and UNet to encode reference images and can therefore process a wider range of images.

**Image as a Word**. Representing images with special tokens has been a popular choice for image personalization. Textual Inversion (Gal et al. 2022) and Dreambooth (Ruiz et al. 2023) represent the original subject with new tokens or rare tokens, the embedding layers or the full model are optimized with a few personalization optimization steps. We follow this line of thought in the image editing context. Instead of learning prompts from images, we learn from items and therefore can be applied when the given image contains multiple items with different subjects. The most similar works to us are SINE (Zhang et al. 2023) and Imagic (Kawar et al. 2023). SINE achieves single-image editing by combining Dreambooth-trained prompt with source prompt using classifier-free guidance. Imagic optimizes the prompt embedding to be aligned with both the input image and the target text and interpolates the learned prompt to achieve editing. Compared to these methods, our framework does not require a captioning prompt in advance and can achieve more types of controls besides text-based editing.

## Method

In this section, we discuss the details of the D-Edit framework. We first review the basics of diffusion models and text-to-image control flow. Next, we show how to establish item-prompt association through the two-step finetuning. Then, we discuss how to utilize the editability of prompts for versatile image editing operations.

### Diffusion Models

Denoising diffusion probabilistic models generate high-quality images by learning to reverse the given forward Markov chain through iterative refinement. During the forward process, it works by gradually adding Gaussian noise to the original data, deriving intermediate latent as

$$z_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon_t \tag{1}$$

with $0 = \alpha_T < \alpha_{T-1} < .. < \alpha_0 = 1$ being the noise schedule, and $\epsilon_t \sim \mathcal{N}(0, \mathbb{I})$. The neural network $f_\theta(z_t, t)$ (like UNet) is introduced to predict the added noise $\epsilon_t$. The predicted noise is then used for sampling by running the reverse process which starts from the pure Gaussian noise $z_T$ and ends at original data $z_0$. Latent Diffusion Model (LDM) is the most widely adopted diffusion model for high-resolution image generation. Given an image $I \in \mathbb{R}^{H\times W\times 3}$, LDM operates in the encoded latent space, $z_0 = E(I)$, and maps the sampled latent representation to the original space using the paired decoder.

**Text-to-Image Control**. A key factor contributing to the success of LDM is its robust ability for text-to-image generation. By introducing the additional condition $y$ as the auxiliary input to $f_\theta(z_t, t, y)$, LDM can generate images according to the designed user prompt. It should be noted that such prompts are usually general textual descriptions, and the final generated image is additionally controlled by the random seed in use for sampling.

Textual prompt controls the image generation through the cross-attention process. Specifically, the given text prompt $P$ containing $W$ words is first encoded by the pretrained text encoder (e.g. CLIP (Radford et al. 2021)) $g_\phi$ into text embedding $c = g_\phi(P) \in \mathbb{R}^{W\times D_c}$ ($W$ is the prompt length and $D_c$ is the embedding dimension). It is then used as input along with the image latent $z_t \in \mathbb{R}^{Z\times D_z}$ (we abuse the notation of model input and layer input) in the UNet cross-attention layer:

$$\begin{aligned} q &= w_q z_t \in \mathbb{R}^{Z\times D} \\ k &= w_k c \in \mathbb{R}^{W\times D} \\ v &= w_v c \in \mathbb{R}^{W\times D} \end{aligned} \qquad \begin{aligned} A &= \text{softmax}(qk^T) \in \mathbb{R}^{Z\times W} \\ O(c, z_t) &= A \cdot v, \end{aligned} \tag{2}$$

where the condition $c$ is encoded into key and value vector while the image input $z_t$ is encoded into query vector.

## Item-Prompt Association

As shown in Eq. 2, the original LDM performs text-image interaction between every token in $c$ and every pixel in $z_t$ through cross-attention matrix $A$. In fact, such token-pixel interactions have been shown disentangled in nature (Tang et al. 2022; Hertz et al. 2022), and the attention matrix $A \in \mathbb{R}^{Z \times W}$ is usually sparse in the sense that each column (token) only attend to several non-zero rows (pixels). For example, during image generation, the word "bear" has higher attention scores with pixels related to the bear region compared to the remaining region.

Inspired by the natural disentanglement, we propose to segment the given image $I$ into $N$ non-overlapped items $\{I_i\}_{i=1}^N$ using segmentation model (same segmentation applied to $z^t$ because of emergent correspondence (Tang et al. 2023)). A set of prompts $\{P_i\}_{i=1}^N$ is adopted to replace the original text prompt $P$. Each prompt $P_i$ is regarded as the textual representation of the corresponding item $z_i^t$ (Details of $P_i$ are discussed in Sec. ). As shown in Fig. 2, we force different items $I_i$ to be controlled by distinct prompt $P_i$ by masking our other items, and therefore any prompt changes in $P_i$ will not influence the remaining item during the cross-attention controlling flow, which is the desired property for image editing. This results in a group of disentangled cross-attentions. For each item-prompt pair ($I_i$, $P_i$), the cross-attention can be written as

$$\begin{aligned} q_i &= w_q z_i^t \in \mathbb{R}^{Z_i \times D} & \text{out}(\{c_i\}, \{z_i^t\}) &= \Sigma_{i=1}^N \text{out}_i(c_i, z_i^t) \\ k_i &= w_k c_i \in \mathbb{R}^{W_i \times D} & A_i &= \text{softmax}(q_i k_i^T) \in \mathbb{R}^{Z_i \times W_i} \\ v_i &= w_v c_i \in \mathbb{R}^{W_i \times D} & \text{out}(c_i, z_i^t) &= A_i \cdot v_i \end{aligned} \tag{3}$$

It should be noted that such disentangled cross-attention cannot be directly used for pretrained LDMs, and therefore further finetuning is necessary to enable the model to comprehend item prompts and grouped cross-attention.

## Linking Prompt to Item

We link prompts to items with two sequential steps. We first introduce the item prompt, consisting of several special tokens with randomly initialized embeddings. Then we finetune the model to build the item-prompt association.

**Prompt Injection**. We propose to represent each item in an image with several new tokens which are inserted into the existing vocabulary of text encoder(s). Specifically, as shown in Fig. 3, we use 2 tokens to represent each item and initialize the newly added embedding entries using Gaussian distribution with mean and standard deviation derived from the existing vocabulary. For comparisons, Dreambooth (Ruiz et al. 2023) represents the image using rare tokens and perfect rare tokens should have no interference with existing vocabulary, which is hard to find. Textual inversion and Imagic insert new tokens into vocabulary where the corresponding embedding is semantically initialized by given word embeddings which describe the image. This adds additional burdens of captioning the original image. We found that it is sufficient to use randomly initialized new tokens as item prompts and such randomly initialized tokens have minimal impact on the existing vocabularies.

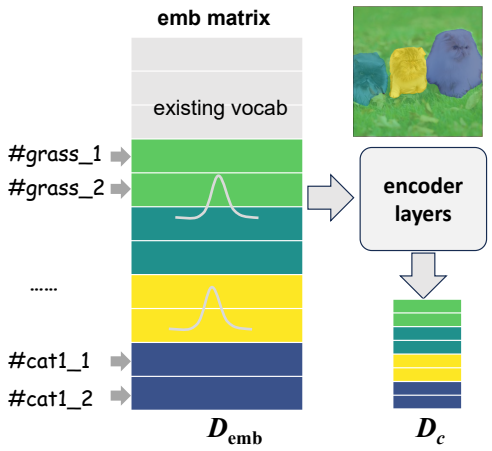


Figure 3: Embedding layer in the text encoder. New tokens are inserted with random initialization.

To associate items with prompts, the inserted embedding entries are then optimized to reconstruct the corresponding image to be edited using

$$\min_e \mathbb{E}_{t,\epsilon} \left[ ||\epsilon - f_\theta(z_t, t, g_\Phi(P))||^2 \right], \tag{4}$$

where $e \in \mathbb{R}^{NM \times D_{\text{emb}}}$ represents the embedding rows corresponding to $N$ items each with $M$ tokens.

**Model Finetuning** . Optimization in the first stage injects the image concept into text-encoder(s), but cannot achieve perfect reconstruction of the original item given the corresponding prompt. Therefore, in the second stage of optimization, we optimize the UNet parameters by running optimization with the same objective function as in Eq. 4. We found that updating parameters solely within cross-attention layers is adequate, as we only disentangle the forward process of these layers rather than the entire model. It should be noted that the optimizations above are running against only one image or two images (target and reference images) if image-based editing is needed.

## Editing with Item-Prompt Freestyle

After the two-step optimization, the model can exactly reconstruct the original image given the set of prompts corresponding to each item, with an appropriate classifier-free guidance scale. We then achieve various disentangled image editing by changing the prompt associated with an item, the mask of an item-prompt pair, and the mapping between items and prompts. We discuss four types of image editing operations that can be achieved by varying item-prompt relationships, summarized in Fig. 4. Details of each operation are discussed in Appendix.

# Experiments

## Experiment Setup

**Training Details**. We implement our D-Edit framework on stable diffusion (SD) v1.5 for 512 $\times$ 512 images and SDXL for 1024 $\times$ 1024 resolution images. Mask2Former (Cheng et al. 2022) is used for segmentation and Grounding DINO

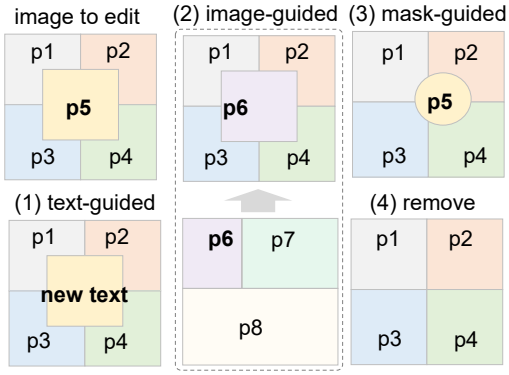


Figure 4: Operations needed for different types of image editing. Each colored item has a unique prompt p.

(Liu et al. 2023) for text-prompted segmentation. The finetuning is performed with Adam optimizer with learning rate $1e$-4 for embedding training, $5e{-}5$ for cross-attention layer training, and $5e{-}5$ for LORA full parameter training. Gradient accumulation is applied to keep the effective batch size to 10 for training robustness. Each image is segmented into 3-8 items by merging excess segments and each item is represented by 1 token for SD and 5 tokens for SDXL. We deploy the default Euler discrete scheduler with sampling step 20 to generate all images during inference. All finetuning and inference are conducted on NVIDIA A6000 GPUs.


Figure 5: Text-guided editing. D-Edit enables selection of any item segmentation and edit using text prompt.

## Text-Guided Editing

Given an input image with appropriate segmentation (no captions needed), we can select any one of the items and replace the learned prompt with the target text prompt. We show such text-guided editing results in Fig. 5. Compared to null-text inversion with Prompt-to-Prompt (P2P), D-Edit can generate more realistic details and have a more natural transition between edited and unedited areas (e.g., the connection between the bike handlebar and the woman's hand) because of disentangled control. The editing of D-Edit is more focused on the target item while the editing of inversion with P2P overflows to other regions (painting in the second example and cheese in the third) using text. Besides, unlike most text-guided methods, D-Edit does not require a caption for the original image which is extremely useful when the scene is hard to describe.

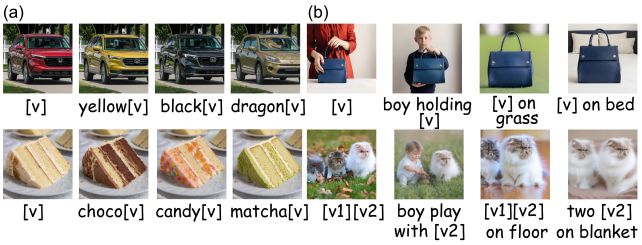


Figure 6: The learned prompt (denoted as [v]) can be combined with words to achieve refinement/editing of the target item. (a) Augment an item prompt with words while keeping other prompts unchanged for editing. (b) Generate the entire image with certain item prompt(s) augmented with text words for personalization.

Then we show the learned item prompts can be combined with normal text words to achieve item refining, besides item replacement. As shown in Fig. 6(a), by combining the learned prompt with adjective words, we can achieve color and texture control of specific items. The preservation of the car and cake shape details after editing indicates the established association between prompts and items through finetuning, while the add-on effect shows the good quality of the learned new prompt in the vocabulary.

In Fig. 6(b), we show the results of item personalization where we generate the image using certain item prompt(s), without using the item-prompt association. This differs from Dreambooth-style personalization in that Dreambooth lacks the capability for item-level customization unless the item is cropped and focused upon, which is usually hard in an image including multiple items. Besides, it requires more images (3-5) with captions for personalization, while our method takes one image without captions. Qualitative results show that learned item(s) can be combined with text to generate personalized images, and the text prompt can be used to personalize the background, number, and position of the item.

For quantitative evaluation, we introduce a new benchmark D-Item(Text) with 100 manually selected multi-item images, where each image is properly segmented into 3-8 items with a segmentation mask. We also include the caption of each image used for baselines, although it is not needed for D-Edit. 2 items of each image are selected and given 5 appropriate target prompts, therefore generating 1,000 item-prompt pair combinations. We adopt CLIP text (CLIP-T) score to represent the semantic alignment of the edited item and target prompt, and LPIPS score to represent consistency with the original images. As shown in Tab. 1, D-Edit outperforms SDEdit (Meng et al. 2021) and P2P with DDIM inversion, especially on LPIPS score which shows improved fidelity to the original images.

## Image-Guided Editing

For image-guided editing, the user can select one item from the reference image and use it to replace one item in the target image. We compare the editing results with baselines including Anydoors (Chen et al. 2023), Paint-by-Example


Figure 7: Qualitative comparison of image-guided editing. D-Edit is compared with Anydoor, Paint-by-Example, and TF-ICON, on item replacement and face swapping.

| | LPIPS↓ | CLIP-T↑ |
|---|---|---|
| P2P | 0.401 | 39.5 |
| SDEdit | 0.432 | 32.1 |
| D-Edit | **0.179** | **42.0** |

Table 1: Text-guided editing: consistency with the original image (LPIPS) and target text prompt (CLIP-T).

| | $\text{LPIPS}_t$↓ | $\text{LPIPS}_r$↓ | CLIP-I↑ |
|---|---|---|---|
| Aydoors | 0.608 | 0.720 | 60.2 |
| PbE | 0.465 | 0.833 | 50.8 |
| D-Edit | **0.340** | **0.701** | **66.4** |

Table 2: Image-guided editing: consistency with the original image ($\text{LPIPS}_t$) and reference image ($\text{LPIPS}_t$ and CLIP-I).

(Yang et al. 2023) and TF-ICON (Lu, Liu, and Kong 2023) when the reference image mainly consists of a single item. As shown in Fig. 7, Paint-by-example can naturally inpaint reference items into the target scene but it falls short in keeping the identity of the reference item (the face and bird example). Anydoors can retain more relevant details from the reference image, yet it may also incorporate undesirable elements in reference, resulting in a less harmonious blend with the target image. For example, the car's original orientation is preserved, causing it to appear out of the parking spot in the target image. Besides, it cannot preserve the face details as in the example. Compared to these methods, D-Edit is capable of seamlessly composing objects into the target while maintaining their identities.

We show more image-based editing results of D-Edit in Fig. 8 and Appendix. D-Edit can work well when the reference image contains multiple items that may be hard to separate (like the bag in hand). Additionally, D-Edit doesn't necessitate the reference item to closely resemble its anticipated appearance in the target image, because blending through the prompt space offers smoother transitions compared to pixel-level manipulation, and the prompt-mask correspondence helps standardize the appearance of the reference item. For example, in the Ultraman example, the reference and target Ultraman can take completely different postures (kneeling v.s. standing).

For quantitative evaluations, based on D-Item(Text) benchmark, we then construct the D-Item(Image) benchmark where each selected item is paired with two reference items from two different reference images, resulting in 400 item-item pairs. Three metrics are considered: $\text{LPIPS}_t$ measures consistency with the original target image; $\text{LPIPS}_r$ and CLIP-Image (CLIP-I) measure alignment with the reference image in low- and high-level feature spaces. As shown in Tab. 2, both D-Edit and Paint-by-Example can achieve high fidelity to the original image, while D-Edit can also preserve the target image better compared with Anydoors.

## Mask-Based Editing and Item Removal

For mask-based editing, we explore four types of operations on the target items, including moving, reshaping, resizing, and refinement. As shown in Fig. 9, D-Edit can edit the shape of the target item by simply editing the corresponding mask. Because of the mask-item-prompt association, the disentangled attention can imagine and fill the new details in the edited regions according to the given item prompt, therefore leading to natural editing results. We also show the post-editing performance in Fig. 10. This can be useful when initial masks from the segmentation model do not cover the whole item, like the missing handle of the handbag and missing straps of the backpack, which will lead to imperfect (image/text/mask-guided) editing results. D-Edit can later fix these mask details and regenerate using the same random seed, and lead to refined results.

D-Edit also enables removing items by deleting the mask-item-prompt pairs. In Fig. 11, by deleting items from the scene image one by one, the resulting blank region will be re-partitioned to nearby masks and join the corresponding item-prompt pair. D-Edit will then use such new associations to imagine the blank regions and therefore lead to reasonable filling results. More visual results can be found in Appendix. To quantitatively assess the item-removed images, we conducted user studies with a group of 15 annotators. They are asked to score 30 pairs of original and item-removed images from 1 to 5 (higher means better), based on quality (how well the region after removal harmonized with the surrounding scene) and fidelity (the reasonableness of the filling content). D-Edit is compared with SDXL inpaint model by inpainting the region where the item is removed with the surrounding item's caption, and results are shown in Tab. 3.

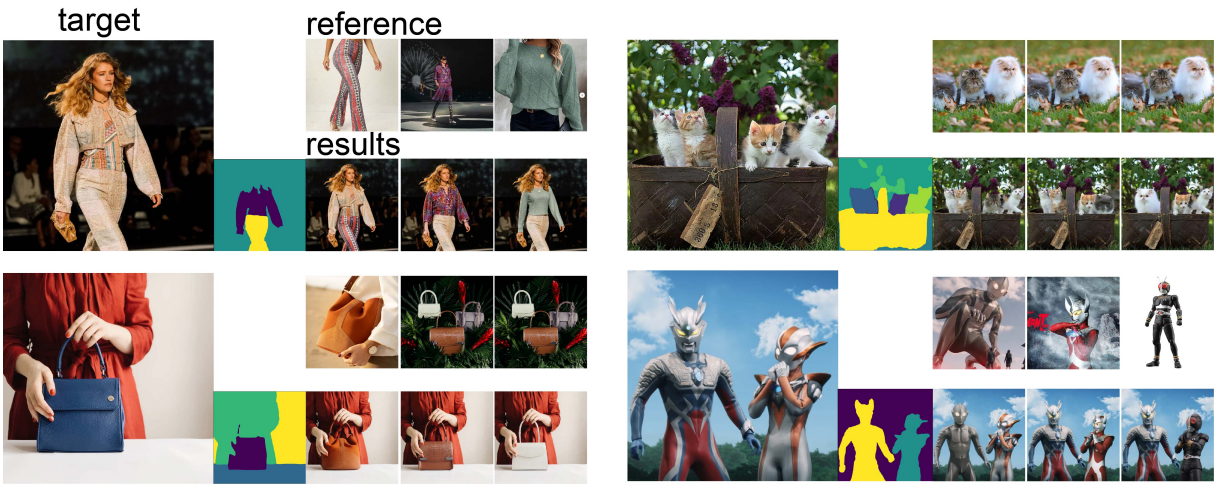


Figure 8: Image-guided editing: Any item in the image can be replaced by another item from the same or different images.

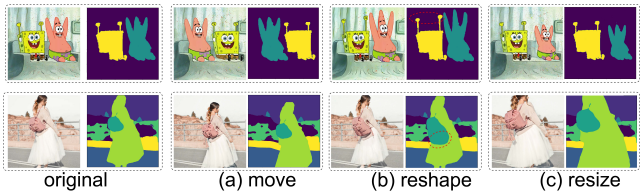


Figure 9: Different types of mask-based editing: (a) Moving/swapping items; (b) reshaping item; (c) Resizing item.


Figure 10: Post-editing refinement can be performed when obtaining imperfect results due to imperfect segmentation.

## Ablation Study

We first study the influence of cross-attention disentanglement. When the cross-attention disentanglement is not used, the learned prompt will affect the entire image and the text-guided editing will be equivalent to the legacy SDXL inpainting. As shown in Fig. 12, when the target item and background are tightly coupled (the hand holding the bag), without disentanglement, the target prompt will take effect based on its own textual semantics and the surrounding item's semantics, therefore leading to poor editing results. This can be avoided by building disentangled item-prompt associations. When the target item can be clearly separated from the background as in the panda example, introducing disentanglement can better preserve the information of the original item, making the editing more controllable. We then study the influence of the number of tokens used to represent each item, and as demonstrated in Tab. 4, 1-5 tokens per item lead to good text-guided editing performance while too many tokens will complicate the embedding training phase and affect the results, and therefore we use 5 tokens in SDXL to generate all results.

| | Quality ↑ | Fidelity ↑ |
|---|---|---|
| SDXL-inpaint | 3.26 | 2.42 |
| D-Edit | **4.01** | **4.44** |

Table 3: Quality and fidelity of editing after removing items.

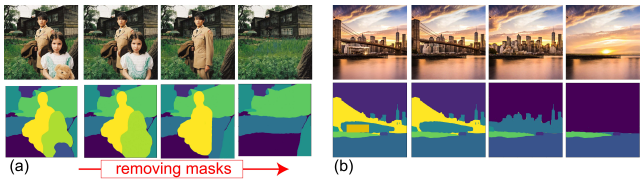


Figure 11: Removing items one by one from the image.

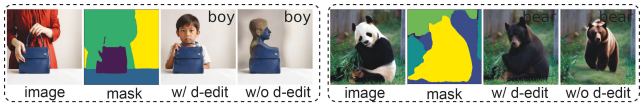


Figure 12: Qualitative comparison of textual-guided editing with and without cross-attention disentanglement

| tokens num. | 1 | 2 | 5 | 10 |
|---|---|---|---|---|
| LPIPS | 0.183 | 0.204 | **0.179** | 0.413 |
| CLIP-T | 38.5 | 38.2 | **42.0** | 30.1 |

Table 4: Text-guided editing with different token numbers.

## Conclusion

In this work, we propose D-Edit as a versatile image editing framework based on diffusion models. D-Edit segments the given image into multiple items, each of which is assigned a prompt to control its representation in the prompt space. The image-prompt cross-attention is disentangled to a group of item-prompt interactions. Item-prompt associations are built up by fintuning the diffusion model which learns to reconstruct the original image using the given set of item prompts.

## Acknowledgments

This work is supported by the U.S. Department of Energy under award DE-FOA-0003264 and the Army Research Office under grant W911NF-23-1-0088.

## References

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

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.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2023. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

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*.

Huberman-Spiegelglas, I.; Kulikov, V.; and Michaeli, T. 2023. An Edit Friendly DDPM Noise Space: Inversion and Manipulations. *arXiv preprint arXiv:2304.06140*.

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.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.

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*.

Miyake, D.; Iohara, A.; Saito, Y.; and Tanaka, T. 2023. Negative-prompt Inversion: Fast Image Inversion for Editing with Text-guided Diffusion Models. *arXiv preprint arXiv:2305.16807*.

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.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2023. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*.

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*.

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

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.

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.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2023. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. *arXiv preprint arXiv:2310.06313*.

Shi, Y.; Xue, C.; Pan, J.; Zhang, W.; Tan, V. Y.; and Bai, S. 2023. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. *arXiv preprint arXiv:2306.14435*.

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

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent Correspondence from Image Diffusion. *arXiv preprint arXiv:2306.03881*.

Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; and Ture, F. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wallace, B.; Gokul, A.; and Naik, N. 2023. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22532–22541.

Xue, B.; Ran, S.; Chen, Q.; Jia, R.; Zhao, B.; and Tang, X. 2022. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *European Conference on Computer Vision*, 300–316. Springer.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. 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.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.