International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02334-x

UniCanvas: Affordance-Aware Unified Real Image Editing via Customized Text-to-Image Generation

Jian Jin[1] · Yang Shen[1] · Xinyang Zhao[1] · Zhenyong Fu[1] · Jian Yang[1]

Received: 28 April 2024 / Accepted: 17 December 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

The demand for assorted conditional edits on a single real image is becoming increasingly prevalent. We focus on two dominant editing tasks that respectively condition on image and text input, namely subject-driven editing and semantic editing. Previous studies typically tackle these two editing tasks separately, thereby demanding multiple editing processes to achieve versatile edits on a single image. However, fragmented and sequential editing processes not only require more user effort but also further degrade the editing quality. In this paper, we propose UniCanvas, an affordance-aware unified framework that can achieve high-quality parallel subject-driven and semantic editing on a single real image within one inference process. UniCanvas innovatively unifies the multimodal inputs of the editing task into the textual condition space using tailored customization strategies. Building upon the unified representations, we propose a novel inference pipeline that performs parallel editing by selectively blending and manipulating two collaborative text-to-image generative branches. Customization enables the editing process to harness the strong visual understanding and reasoning capability of pre-trained generative models for affordance perception, and a unified inference space further facilitates more effective affordance interaction and alignment for compelling editing. Extensive experiments on diverse real images demonstrate that UniCanvas exhibits powerful scene affordance perception in unified image editing, achieving seamless subject-driven editing and precise semantic editing for various target subjects and query prompts (https://jinjianrick.github.io/unicanvas/).

Keywords Real image editing · Pre-trained model customization · Text-to-image generation · Diffusion model · Affordance perception

1 Introduction

Do you envision featuring your cherished plushie in the landscapes you have explored and documented, as if it were there with you? Would you like your recently adopted pet dog to join in the previously captured family photos? Moreover, how about transforming these processed photos into artistic oil paintings and adorning your pet dog in stylish attire?

Performing multiple creative edits on a single real image is increasingly in demand. Our work is dedicated to two primary editing tasks, namely subject-driven editing Yang et al. (2023); Song et al. (2023); Chen et al. (2024b); Lu et al. (2023a) and semantic editing Couairon et al. (2023); Mokady et al. (2023); Kawar et al. (2023); Zhang et al. (2023b), which perform image manipulations conditioned on image and text guidance. Subject-driven editing aims to generate a specific subject in a specified region of an input image, while semantic editing is tasked with modifying the image according to textual descriptions. Previous studies typically focus on a single

Communicated by Ran He.

✉ Zhenyong Fu
z.fu@njust.edu.cn

✉ Jian Yang
csjyang@njust.edu.cn

Jian Jin
jinj@njust.edu.cn

Yang Shen
shenyang_98@njust.edu.cn

Xinyang Zhao
zhaoxy@njust.edu.cn

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

Published online: 14 January 2025

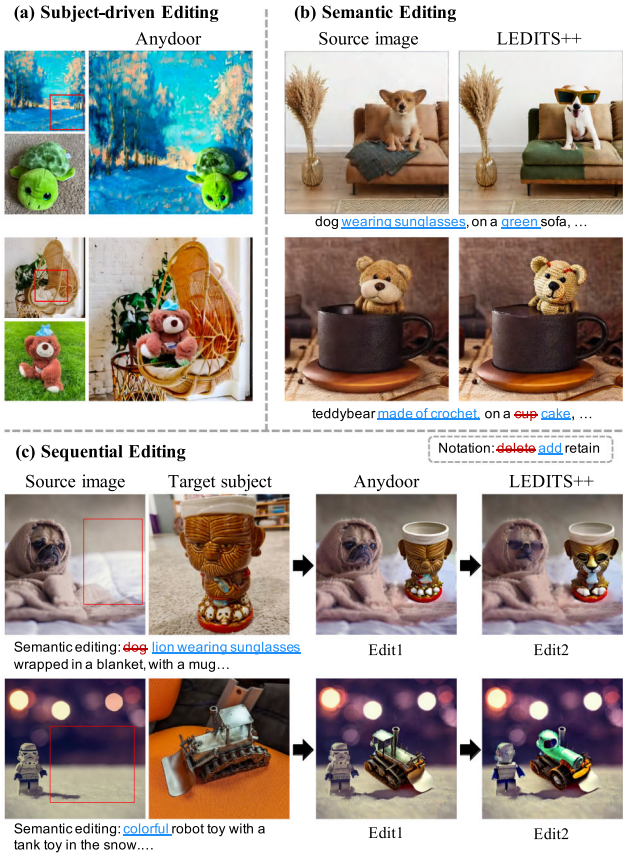

Fig. 1 Limitations of existing methods. We provide visual examples of the latest subject-driven editing method, Anydoor Chen et al. (2024b), and the semantic editing method, LEDITS++ Brack et al. (2024). Existing methods exhibit poor affordance perception and deliver unsatisfactory results. For subject-driven editing in **a**, the blended subject appears incongruous with the background, as if it were directly pasted onto the image. The semantic editing in **b** struggles to locate the editing regions and perform the correct edits. Existing methods must perform sequential edits to achieve unified editing. As shown in **c**, sequential editing causes critical issues, such as error accumulation and deterioration of prior results due to later edits (i.e., 'mug' in the 1^{st} row and 'tank toy' in the 2^{nd} row). Please refer to Figs. 9 and 10 for more examples of the limitations of existing methods

type of editing task. Therefore, to achieve subject-driven and semantic editing on a single image, we need to perform a sequential editing process using multiple editing methods. However, a fragmented editing process involving various methods requires more user effort in preparation (e.g., environment configuration, pre-trained model preparation, etc.) and execution. Worse yet, sequential editing can lead to the accumulation of reconstruction and editing errors. Besides, the later editing process can further deteriorate the previous results, further degrading the editing quality, as shown in Fig. 1c.

Additionally, even for single-type image editing, existing methods deliver unsatisfactory results. To achieve compelling subject-driven editing and semantic editing, in addition to maintaining fidelity of the source image and the given subject, several key issues need to be addressed. Firstly, the edited content should exhibit appropriate visual properties (e.g., geometric attributes, visual domain, lighting conditions, and visual depth), thereby achieving semantic and geometric compatibility with the background scene. Besides, the model should synthesize natural edited-unedited interactions that comply with physical laws. For instance, reflection takes place for reflective surfaces, and shadow parameters should be consistent with background objects. Additionally, the model should perceive the semantic context in the source image to precisely locate the editing region in semantic editing. All of these issues entail the model possessing effective affordance perception Kulal et al. (2023); Gupta et al. (2011); Wang et al. (2017), which enables a semantic understanding of the image and subject to achieve reasonable and realistic editing. Previous methods Cong e al. (2020); Xue et al. (2022); Azadi et al. (2020); Lin et al. (2018); Liu et al. (2020); Yang et al. (2023); Kulal et al. (2023); Lu et al. (2023a); Chen et al. (2024b); Song et al. (2023) generally exhibit unsatisfactory affordance perception on subject-driven editing and semantic editing, thereby hindering high-quality editing. For instance, as shown in Fig. 1, the blended object appears incongruous with the background context in subject-driven editing, while semantic editing methods struggle to accurately locate multiple editing regions.

Our work strives to perform parallel subject-driven editing and semantic editing on a single real image with effective affordance perception (Fig. 2). Formally, given an arbitrary real image, our aim is to seamlessly render a specific subject into a designated region of the image, while also enabling precise and effortless semantic edits on both the source image and the blended subject.

Recently, significant advancements have been achieved in the field of customized text-to-image generation Ruiz et al. (2023); Gal et al. (2023a); Kumari et al. (2023). Given a few images of a custom concept as reference, model customization fine-tunes large-scale pre-trained models to implant the visual concept, binding the visual concept to specific textual prompt. We seek to achieve affordance-aware unified editing based on text-to-image model customization, inspired by its two key merits. First, a unified image editing task takes two modalities as input, *i.e.,* textual prompts and images. Text-to-image model customization can convert images into textual conditions, unifying two types of input to the same modality, thereby paving the way for unified editing. Furthermore, large-scale pre-trained models exhibit powerful generative priors for visual understanding and reasoning. Model customization implants the visual concept into pre-trained models, providing a potential solution to unlock the models' capacity for scene affordance perception in image editing.

Fig. 2 Given a source real image and a target subject specified by several reference images, UNICANVAS can seamlessly render the target subject into a designated region of the source image, while simultaneously being able to perform semantic edits on the resultant image in a precise and effortless manner

In this paper, we propose UNICANVAS, a compact framework built upon customized text-to-image generation for unified image editing. UNICANVAS consists of two editing-specific customization strategies and a novel inference pipeline. Concretely, UNICANVAS firstly fine-tunes the pre-trained text-to-image model to bind the target subject and source image with special textual prompt. The vanilla fine-tuning strategy is proposed for text-to-image generation and is unsuitable for image editing. Therefore, we tailor distinct fine-tuning strategies for these two components, endowing them with ideal inference-time properties. Building upon the unified representations, the proposed inference pipeline achieves parallel editing with two collaborative text-to-image generative branches, namely the subject branch and the image branch. Conditioned on the bound textual prompts, the subject branch is tasked with generating the target subject in the specified region with high visual fidelity, while the image branch is responsible for faithfully reconstructing the source image. To achieve subject-driven editing, these two generative branches are concurrently forwarded within the diffusion model and integrated during the latent denoising steps using a Selective Blending Module (SBM). SBM employs cross-attention maps to dynamically determine aggregation weights of two branches, facilitating coherent and seamless image blending. Meanwhile, by simply modifying the conditional prompts, we can perform semantic editing on both the blended subject and the source image.

UNICANVAS innovatively unifies the multimodal inputs of the editing task into the textual condition space using tailored customization strategies. Customization enable the inference process to harness the strong visual understanding and reasoning capability of large-scale generative models for affordance perception. The subject and source image are generated by dual collaborative branches that interact mutually, facilitating more effective affordance interaction and alignment between the two components for seamless blending. Leveraging the powerful generative priors of pre-trained large-scale models, UNICANVAS substantially enhances rationality and compatibility between the blended subject and source image in terms of geometry and semantics. Furthermore, the text-to-image generation paradigm of UNICANVAS provides a unified interface for subject-driven editing and semantic editing, enabling more precise and convenient image manipulation simply through appropriate textual guidance.

We conduct extensive experiments on various target subject and real image pairs with different target regions. Experimental results demonstrate that UNICANVAS exhibits strong capability in scene affordance perception, enabling it to generate realistic and reasonable subject and interactions in compatibility with the background scene. Therefore, UNICANVAS achieves seamless subject-driven editing and precise semantic editing, even in challenging scenes such as cross-domain blending. More applications like subject replacement and spatial controllable customized text-to-image generation can also be achieved under the framework of UNICANVAS. Overall, the contributions of this work are summarized as follows:

1. We introduce UniCanvas, a unified framework capable of performing parallel subject-driven and semantic editing on a real image in a single inference process. To the best of our knowledge, this is the first work to achieve parallel image editing conditioned on both image and text guidance.
2. UniCanvas innovatively unifies the multimodal inputs of the editing task into the textual condition space using tailored customization strategies. Building upon the unified representations, we propose a novel inference pipeline that performs parallel editing by selectively blending and manipulating two collaborative text-to-image generative branches.
3. Extensive experiments demonstrate that UniCanvas achieves high-quality unified editing for real images. Notably, UniCanvas excels in scene affordance perception, a capability in which previous methods fall short but is critical for reasonable and coherent editing.

2 Related Works

2.1 Text-to-Image Generation

Text-to-image (T2I) generation aims to generate visually convincing images based on textual descriptions. Following the seminal work Mansimov et al. (2015), numerous text-to-image methods have been subsequently proposed. Early approaches Reed et al. (2016); Zhang et al. (2017); Xu et al. (2018); Li et al. (2019) utilized generative adversarial networks (GANs)Goodfellow et al. (2014) to convert natural language into images, primarily focusing on small-scale input scenarios. Autoregressive models such as DALL-E Ramesh et al. (2021), Cogview Ding et al. (2021), NUWA Wu et al. (2022), and Parti Yu et al. (2022), reframe text-to-image generation as a sequence-to-sequence problem. These methods exploit auto-regressive transformers as generators and output sequences of image tokens. While autoregressive methods bolster performance with large-scale textual inputs, they are confronted with challenges such as computational overhead and the accumulation of sequential errors Zhang et al. (2023). Recent advancements in text-to-image generation utilize diffusion models (DMs) Sohl-Dickstein et al. (2015) as the generative backbone. These models generate images through a denoising task while incorporating text conditions during the denoising process. Models like GLIDE Nichol et al. (2022) and Imagen Saharia e al. (2022) generate images at a high-dimensional pixel level. Another line of research, including Stable Diffusion Rombach et al. (2022) and DALL-E 2 Ramesh et al. (2022), trains the diffusion model within a low-dimensional latent space.

2.2 Subject-Driven Image Editing

There are several research lines of subject-driven or reference-based image editing, including image composition and inpainting-based methods, which aim to generate a specific subject in a specified region of the source image. The image composition methods Cong e al. (2020); Xue et al. (2022); Azadi et al. (2020); Lin et al. (2018); Liu et al. (2020); Lu et al. (2023b) cut the foreground from one reference image and paste it on the background image to produce a composite image Niu et al. (2021). These methods often focus on a specific aspect of the composition problem, such as image matting Xu et al. (2017), image harmonization Cong e al. (2020); Xue et al. (2022), geometric correction Azadi et al. (2020); Lin et al. (2018), and shadow generation Liu et al. (2020), to make the composite image more realistic. However, these methods yield composite images that lack affordance perception and subject diversity, resulting in unsatisfactory performance in terms of geometric and semantic compatibility. There have been recent works Yang et al. (2023); Kulal et al. (2023); Lu et al. (2023a); Chen et al. (2024b); Song et al. (2023) that generate a specific subject in the target region of the background using image inpainting. However, inpainting-based methods denoise the target region and generate objects according to the given text prompt, which discard the important structural and semantic information Hertz et al. (2023); Couairon et al. (2023), thereby hindering the affordance perception. This further results in strong artifacts, such as generating partial subjects or inconsistent and distorted content within the target region Xie et al. (2023).

2.3 Semantic Image Editing

Semantic image editing aims to modify an image based on instructions given in natural language Couairon et al. (2023). Some GAN-based methods optimize either the image or its latent representation based on a high-level multimodal objective to edit images Crowson et al. (2022); Couairon et al. (2022); Patashnik et al. (2021), while others discover latent space directions in a pre-trained GAN for semantic edits Härkönen et al. (2020); Collins et al. (2020); Shen et al. (2020). Diffusion model has recently demonstrated powerful capabilities in semantic image editing. Diffusion-based semantic editing methods can be primarily classified into three categories: training-free methods Choi et al. (2021); Meng et al. (2021); Cao et al. (2023); Couairon et al. (2023); Mokady et al. (2023); Tumanyan et al. (2023), training Brooks et al. (2023); Zhang et al. (2023a), and fine-tuning Kawar et al. (2023); Zhang et al. (2023b). Prox-edit Han et al. (2024) proposes proximal guidance and incorporates it to negative-prompt inversion with cross-attention control. Prompt-to-Prompt Hertz et al. (2023)

finds that cross-attention layers are crucial for linking the image layout to prompt words, proposing to control the edited image through attention map injection. Imagic Kawar et al. (2023) finds a prompt embedding that aligns with the input image and performs editing by interpolating between the image embedding and the target prompt embedding. Nguyen Nguyen et al. (2024) takes image pairs as visual prompting, which is inverted into editing instructions to perform the same edit on new images. These methods typically address semantic editing of a single main object but struggle to perform complex editing involving multiple inconspicuous objects.

2.4 Customized Text-to-Image Generation

Customized generation aims to incorporate a novel concept, as described by a few user-provided examples, into pre-trained text-to-image models. This enables the adapted models to generate high-quality and diverse images of the new concept, guided by textual prompts. Pioneering works Gal et al. (2023a); Ruiz et al. (2023) incorporate the novel concept by fine-tuning the pre-trained model. Text Inversion Gal et al. (2023a) inverts the new concept into the embedding of a special prompt token for customization, and it only optimizes the token embedding during fine-tuning. DreamBooth Ruiz et al. (2023) binds the new concept with a rare-token identifier by fine-tuning the entire diffusion model. It also regularizes the adapting process with a class-specific prior preservation loss to prevent over-fitting and language-drift. These seminal studies have garnered widespread attention of customized text-to-image generation. Recent efforts have focused on improving customization quality Alaluf et al. (2023); Tewel et al. (2023) and developing more efficient methods Kumari et al. (2023); Gal et al. (2023b); Han et al. (2023) for customized generation. Another line of research Wei et al. (2023); Xiao et al. (2023); Chen et al. (2024a) utilizes a learning-based paradigm for customized generation, which reduces inference-time costs but sacrifices generation quality. More recently, multi-concept customized generation Kumari et al. (2023); Liu et al. (2023); Han et al. (2023); Gu et al. (2024) has been introduced, aiming to integrate multiple customized concepts into a single output image. In this paper, we achieve unified real image editing based on text-to-image model customization.

3 Preliminaries and Task Descriptions

3.1 Preliminaries

3.1.1 Customized Text-to-Image Generation

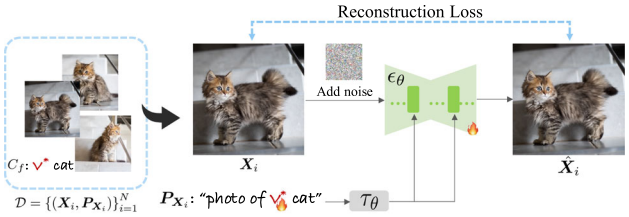

Fig. 3 The conventional text-to-image model customization (fine-tuning) process

A custom concept C_f is incorporated into a pre-trained text-to-image model by fine-tuning the model using N image-prompt pairs $\mathcal{D} = \left\{(\boldsymbol{X}_i, \boldsymbol{P}_{\boldsymbol{X}_i})\right\}_{i=1}^{N}$, where $\boldsymbol{X}_i$ and $\boldsymbol{P}_{\boldsymbol{X}_i}$ are reference images and the corresponding textual prompts of C_f. The fine-tuning process is illustrated in Fig. 3a. We utilize Latent Diffusion Models (LDMs) Rombach et al. (2022) as the generative backbone. Prompt $\boldsymbol{P}_{\boldsymbol{X}_i}$ is firstly projected to an intermediate representation $\mathbf{c} = \tau_\theta(\boldsymbol{P}_{\boldsymbol{X}_i}) \in \mathbb{R}^{M \times d_c}$ by a text encoder τ_θ. $\mathbf{c}$ is then injected into the LDMs via the cross-attention mechanism Vaswani et al. (2017), serving as the condition for the reconstruction of $\boldsymbol{X}_i$. The reconstruction process is regularized by the following squared error loss:

$$\mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),\mathbf{c},t}\left[w_t\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2\right], \tag{1}$$

where $\mathbf{z}_t := \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}$ is the noised latent code at timestep t, $\mathbf{z}$ is the clean latent code of the training data, and w_t, α_t, σ_t are terms that determine the loss weight and noise schedule; ϵ_θ is a denoising autoencoder implemented using a conditional U-Net Ronneberger et al. (2015).

At inference, given the query text $\boldsymbol{P}_{C_f}$ containing the learned concept C_f, the customized LDMs can generate the corresponding images conditioned on $\mathbf{c} = \tau_\theta(\boldsymbol{P}_{C_f})$. Concretely, an initial noise map $\mathbf{z}_T \sim \mathcal{N}(0, 1)$ is iteratively denoised from timestep $t = T$ to $t = 1$ in the latent space with each individual step corresponding to:

$$\mathbf{z}_{t-1} = \mathbf{z}_t - \gamma\epsilon_\theta\left(\mathbf{z}_t, \tau_\theta(\boldsymbol{P}_{C_f})), t\right), \; t = T, \ldots, 1, \tag{2}$$

where γ is the step size. Then $\mathbf{z}_0$ is decoded to image space using an decoder f_D to generate the target image $\hat{X} = f_D(\mathbf{z}_0)$.

3.1.2 Cross Attention Layers in LDMs

The cross-attention layer utilizes latent image features $\mathcal{I}$ and text embeddings $\mathcal{T}$ to compute queries $Q = \ell_Q(\mathcal{I})$, keys $K = \ell_K(\mathcal{T})$, and values $V = \ell_V(\mathcal{T})$, using three projection layers: ℓ_Q, ℓ_K and ℓ_V. Then the attention maps $\mathcal{M} \in \mathbb{R}^{C \times H \times W}$ are calculated as:

$$\mathcal{M} = \textbf{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \in \mathbb{R}^{C\times H\times W}, \tag{3}$$

where d is the dimension of the queries and keys, and C, H, W are the channel dimension, height, and weight of $\mathcal{M}$, respectively. The final output of the cross-attention layer is computed as $\mathcal{I} = \mathcal{M}V$. Each token in the conditional prompt is associated with an attention map $\mathcal{M}$, which determines the spatial layout and geometry of the corresponding textual semantics in the generated image Hertz et al. (2023).

3.2 Task Descriptions and Notations

Given an arbitrary real image $\boldsymbol{X}_s$, we endeavor to perform subject-driven editing and semantic editing of $\boldsymbol{X}_s$ within a unified framework. The subject C_f for subject-driven editing is specified by several reference images $\{(\boldsymbol{X}_i, \boldsymbol{P}_{\boldsymbol{X}_i})\}_{i=1}^{N}$ (N is typically $3 \sim 5$). At inference, given a user mask M_U specifying the target region $\mathcal{R}_f$, subject-driven editing aims to seamlessly render the subject C_f into $\mathcal{R}_f$ of $\boldsymbol{X}_s$, generating a composite image $\boldsymbol{X}_c$. In $\boldsymbol{X}_c$, we necessitate faithful reconstruction for $\boldsymbol{X}_s$ while preserving key identifying features of C_f. Besides, it is crucial to ensure geometric and semantic compatibility between the generated subject and the background context. Additionally, we strive to further perform semantic editing on $\boldsymbol{X}_c$ in a precise and effortless manner, preserving the maximal amount of details from the original image after convincing editing. For clarity, we provide a concise overview of the notations used in this work in Table 1.

4 Method

We propose UniCanvas, which simultaneously achieves subject-driven editing and semantic editing on real images through text-to-image model customization.

UniCanvas unifies the modeling processes of the source image and target subject within the framework of model customization. These two components are implanting into a pre-trained text-to-image model using corresponding image-prompt pairs, binding them with unique textual descriptions. As stated in Sect. 3.2, the ideal inference-time properties differ for these two components. The target subject C_f is expected to exhibit the capability to be rendered in the specified region with high concept fidelity, while the source image needs to be faithfully reconstructed. Therefore, we separately tailor the fine-tuning strategies for source image (Sect. 4.1) and target subject (Sect. 4.2) to achieve their desired properties. Notably, these two fine-tuning schemes are mutually compatible to implant two components into a pre-trained model simultaneously.

At inference, leveraging the bound textual descriptions and target region as conditions, UniCanvas performs the subject-driven editing (Sect. 4.3) and semantic editing (Sect. 4.4) through customized text-to-image generation.

4.1 Source Image Customization

We fine-tune the pre-trained text-to-image model with the single image-prompt pair to implant the source image $\boldsymbol{X}_s$ Zhang et al. (2023b); Han et al. (2023). For a source image $\boldsymbol{X}_s$ containing K primary visual concepts, the prompt for $\boldsymbol{X}_s$ is constructed as follows:

$$\boldsymbol{P}_{\boldsymbol{X}_s} = [\mathbf{T}_1][\mathbf{E}_1][\mathbf{T}_2][\mathbf{E}_2]\cdots[\mathbf{T}_K][\mathbf{E}_K], \tag{4}$$

where $[\mathbf{T}_n] \in \mathbb{R}^d$ ($n = 1, \ldots, K$ represents the indices of the visual concepts in $\boldsymbol{X}_s$.) are learnable concept-specific context, $[\mathbf{E}_n] \in \mathbb{R}^{t\times d}$ ($n = 1, \ldots, K$, t is the token length of $[\mathbf{E}_n]$) are class descriptors of the K primary objects in $\boldsymbol{X}_s$, and d is dimension of token embedding. For instance, the prompt of the source image illustrated in Fig. 4b is designed as "$V^*_{s_1}$ `bed` $V^*_{s_2}$ `table` $V^*_{s_3}$ `vase`", where $V^*_{s_i}$ ($i = 1, 2, 3$) represent concept-specific context. To ensure faithful reconstruction during inference, no extra data augmentation or regularization is employed in fine-tuning. During inference, we use the entire prompt $\boldsymbol{P}_{\boldsymbol{X}_s}$ as the textual condition for the reconstruction of $\boldsymbol{X}_s$.

4.2 Target Subject Customization

To enable the generation of a customized subject in a specified region $\mathcal{R}_f$ with high fidelity, we propose Region-Aware Customization (RAC) strategy for subject fine-tuning. RAC modifies the original text-to-image reconstruction task by incorporating generation region $\mathcal{R}_f$ of subject as an additional condition, alongside Region Variability Augmentation (RVA) for data construction and a dedicated novel prompt scheme.

4.2.1 Region Variability Augmentation

RAC receives the subject generation region as an additional condition, thus we need to provide the generation region of each reference image for fine-tuning. The original generation region is delineated by the bounding box of the subject in the images. However, due to the limited quantity of reference images, the implanted subject tends to favor restricted sizes of target region at inference, posing problems such as significant subject distortion or subject omission for small target regions. Therefore, we propose a data augmentation strategy called Region Variability Augmentation (RVA) to enhance the generalization and robustness of the implanted subject

Table 1 A concise overview of the notations

Notations	Descriptions
X_s	Source image for editing
C_f	Target subject for subject-driven editing
$\mathcal{R}_f$	Target region for subject-driven editing
$\boldsymbol{X}_c$	The edited image
$\mathcal{M}$	Attention map in cross-attention layers
T	The number of denoising steps
Fine-tuning stage	
$\{(\boldsymbol{X}_i, \boldsymbol{P}_{\boldsymbol{X}_i})\}_{i=1}^{N}$	Reference image set of C_f, where $\boldsymbol{X}_i$ is the i-th reference image and $\boldsymbol{P}_{\boldsymbol{X}_i}$ is the fine-tuning prompt of $\boldsymbol{X}_i$
$\boldsymbol{P}_{\boldsymbol{X}_s}$	Fine-tuning prompt for $\boldsymbol{X}_s$
$\mathbf{T}$	Concept-specific context in the prompt
$\mathbf{E}$	Class descriptors in the prompt
$\mathbf{I}$	Image-specific context in the prompt
$\boldsymbol{X}_i^a$	Augmented image of $\boldsymbol{X}_i$
$\boldsymbol{L}_{\boldsymbol{X}_i^a}$	Layout of image $\boldsymbol{X}_i^a$
Inference stage	
$\boldsymbol{P}_{C_f}$	Query prompt for inference
M_U	Binary user mask indicating $\mathcal{R}_f$
$\mathcal{B}$	Inference-time generative branch
$\mathcal{T}$	Prompt token set
$\mathrm{M}_{\mathcal{B}}$	Dynamic aggregation mask of $\mathcal{B}$
$\mathbf{S}$	Softmax operation
$\hat{\mathcal{F}}$	Output features of cross-attention layers

in different region conditions. Specifically, we crop the subject from the original reference image $\boldsymbol{X}_i$ along its bounding box. The cropped subject is then randomly scaled down to $0.4 - 1.0\times$ and padded with zero pixels to restore it to the original size, resulting in an augmented image $\boldsymbol{X}_i^a$. $\boldsymbol{X}_i^a$ consists of the foreground region $\mathcal{R}_f$ and the background region $\mathcal{R}_b$ (i.e., the augmented region), as illustrated in Fig. 4. The position of $\mathcal{R}_f$ within the augmented image is also randomized. We define the layout of $\boldsymbol{X}_i^a$ as:

$$\boldsymbol{L}_{\boldsymbol{X}_i^a}^{p} = \begin{cases} 1, & p \in \mathcal{R}_f \\ 0, & p \in \mathcal{R}_b \end{cases}, \tag{5}$$

where $\boldsymbol{L}_{\boldsymbol{X}_i^a}^{p}$ represents the value of layout $\boldsymbol{L}_{\boldsymbol{X}_i^a}$ at pixel p.

4.2.2 Prompt Scheme Design

Previous works Ruiz et al. (2023); Kumari et al. (2023); Gal et al. (2023a) construct fine-tuning prompts from manually crafted prompt templates (e.g., "`A photo of {}`") for reference images. The manually designed part in prompts is frozen during fine-tuning, which may provide imprecise guidance for model adaptation, especially with region variability augmentation. Therefore, we replace the manually crafted prompt templates with learnable, image-specific contexts for each reference image. Concretely, the conditional prompts forwarded to the text transformer are structured as follows:

$$\boldsymbol{P}_{\boldsymbol{X}_i} = [\mathbf{I}]_1^i \cdots [\mathbf{I}]_M^i\, [\mathbf{T}]_1 \cdots [\mathbf{T}]_S\, [\mathbf{E}]\, [\mathbf{I}]_{M+1}^i \cdots [\mathbf{I}]_{M+L}^i, \tag{6}$$

where the prompt $\boldsymbol{P}_{\boldsymbol{X}_i}$ for the image $\boldsymbol{X}_i$ is composed of three components: image-specific context vectors $[\mathbf{I}]_m^i$ $(m = 1, \ldots, M+L) \in \mathbb{R}^d$ with context length of $M+L$, concept-specific context vectors $[\mathbf{T}]_s$ $(s = 1, \ldots, S) \in \mathbb{R}^d$ with context length of S, and the concept descriptor $[\mathbf{E}] \in \mathbb{R}^{t \times d}$. *Image-specific Context* $[\mathbf{I}]$ are continuous context vectors, which are specific to each image and independent across images. These vectors can be optimized end-to-end, adaptively modeling the concept-irrelevant content in each reference image for more precise guidance of model fine-tuning. *Concept-specific Context* $[\mathbf{T}]$ models the concept-relevant content, which is learnable and shared across all reference images. We denote this as "V_f^*". *Concept Descriptor* $[\mathbf{E}]$ is the word embedding of a coarse class noun (denoted as "`<noun>`") of the concept Ruiz et al. (2023), e.g., embedding of noun "`teddybear`". We further define the indicator vector:

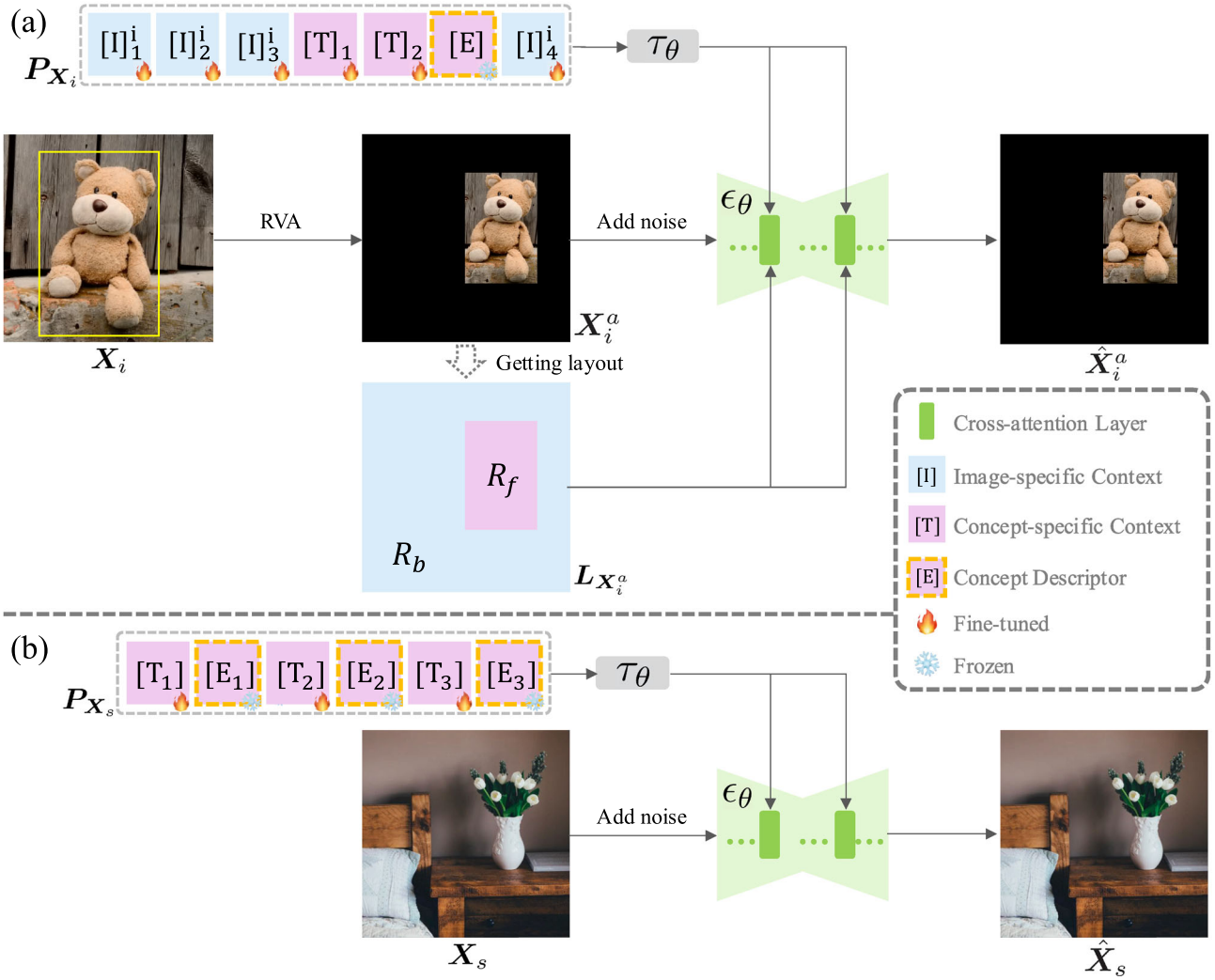

Fig. 4 Overall pipeline of the fine-tuning process. We illustrate this process with a simple example. (**a**) Target subject customization. We introduce Region-Aware Customization (RAC) strategy to target subject fine-tuning. RAC incorporates the generation region $\mathcal{R}_f$ of the subject as an additional condition, alongside Region Variability Augmentation (RVA) for data construction and a dedicated novel prompt scheme. (**b**) Source image customization. The model is fine-tuned with a single image-prompt pair

$$\delta^k_{\boldsymbol{P}_{\boldsymbol{X}_i}} = \begin{cases} 1, \ k \in [\mathbf{T}], [\mathbf{E}] \\ 0, \ k \in [\mathbf{I}] \end{cases}, \tag{7}$$

where $\delta^k_{\boldsymbol{P}_{\boldsymbol{X}_i}}$ indicates that token k corresponds to foreground (i.e., 1) or background (i.e., 0) of the reference image. The learned concept-specific context, in combination with the concept descriptor, is utilized to generate the subject at inference.

4.2.3 Region-Aware Customization

Conditioning on prompt $\boldsymbol{P}_{\boldsymbol{X}_i}$ and layout $\boldsymbol{L}_{\boldsymbol{X}^a_i}$, we aim to reconstruct $\boldsymbol{X}^a_i$. As detailed in Sect. 3.1.2, the attention maps determine the spatial layout of images, allowing for modification to control the shape and location of the generated objects Hertz et al. (2023); Xue et al. (2023); Kim et al. (2023). Therefore, to constrain the spatial distribution of the reconstructed subject, we rectify the corresponding attention maps as follows:

$$\begin{cases} \mathcal{M}_{k,p} += \alpha(\max(\mathcal{M}_{k,p}) - \mathcal{M}_{k,p}), \ \delta^k_{\boldsymbol{P}_{\boldsymbol{X}_i}} = \boldsymbol{L}^p_{\boldsymbol{X}^a_i} \\ \mathcal{M}_{k,p} -= \alpha(\mathcal{M}_{k,p} - \min(\mathcal{M}_{k,p})), \ \delta^k_{\boldsymbol{P}_{\boldsymbol{X}_i}} \neq \boldsymbol{L}^p_{\boldsymbol{X}^a_i} \end{cases}, \tag{8}$$

where $\mathcal{M}_{k,p}$ is attention score of token k at pixel p, max and min operations return the maximum and minimum values for each query, and α is a parameter related to region area. Therefore, for attention maps corresponding to concept-specific tokens and class nouns, we increase the attention score in

region $\mathcal{R}_f$ while decreasing it in region $\mathcal{R}_b$. Conversely, for attention maps corresponding to image-specific tokens, the attention score in region $\mathcal{R}_b$ is increased, and in region $\mathcal{R}_f$ it is decreased. Therefore, the objective for subject customization is modified from Eq. (1) as:

$$\mathbb{E}_{\mathbf{z},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),\mathbf{c},t}\left[w_t\|\epsilon-\epsilon_\theta(\mathbf{z}_t,\boldsymbol{P}_{\boldsymbol{X}_i},\boldsymbol{L}_{\boldsymbol{X}_i^a},t)\|_2^2\right]. \tag{9}$$

This fine-tuning strategy enables the model to be better jointly controlled by textual prompts along with explicit positional conditions, achieving the generation of subjects in specified regions with high subject fidelity.

4.3 Subject-Driven Editing

The overview of the subject-driven editing process is illustrated in Fig. 5. UniCanvas achieves subject-driven editing with two generative branches: the subject branch $\mathcal{B}_f$ and the image branch $\mathcal{B}_s$. These two branches share same model parameters and collaborate to generate target images at inference. Concretely, the subject branch is conditioned on the target region $\mathcal{R}_f$ specified by the user mask M_U and a textual prompt $\boldsymbol{P}_{C_f}$, tasked with rendering the target subject C_f described by $\boldsymbol{P}_{C_f}$ in the region $\mathcal{R}_f$. On the other hand, the image branch, conditioned on the prompt $\boldsymbol{P}_{\boldsymbol{X}_s}$, is responsible for faithfully reconstructing the source image $\boldsymbol{X}_s$. These two generative branches are simultaneously forwarded within the diffusion model and selectively integrated during the latent denoising steps using a Selective Blending Module (SBM) to produce the final output.

4.3.1 Selective Blending Module

To achieve coherent and seamless blending, we introduce a *Selective Blending Module* (SBM) to each cross-attention layer within the U-Net, as depicted in Fig. 5.

The text-to-image generation paradigm enables the designation of specific semantic content by utilizing corresponding textual tokens. Hence, we first construct the token sets for the two branches, each containing the tokens of content that the respective branch is responsible for generating. The token set for $\mathcal{B}_f$ is $\mathcal{T}(\boldsymbol{P}_{C_f})=\{$"$V_f^*$","<noun>"$\}$, and the token set for $\mathcal{B}_s$ consists of all tokens in $\boldsymbol{P}_{\boldsymbol{X}_s}$, denoted as $\mathcal{T}(\boldsymbol{P}_{\boldsymbol{X}_s})$.

Before the dual-branch blending, in consistent with the fine-tuning stage, we need to constrain the spatial distribution of the target subject to the target region by rectifying the attention maps using Eq. (8). The indicator vector and layout are obtained according to $\mathcal{T}(\boldsymbol{P}_{C_f})$ and M_U. We only perform this rectification in the early stage of denoising ($0.3\times T$).

Since the attention maps $\mathcal{M}$ determines the spatial layout of the corresponding textual semantics in the generated image, we take the attention maps corresponding to the tokens in the token sets $\mathcal{T}(\boldsymbol{P}_{C_f})$ and $\mathcal{T}(\boldsymbol{P}_{\boldsymbol{X}_s})$ to get the dynamic blending mask $\mathrm{M}_{\mathcal{B}_f}^{t,i}$ and $\mathrm{M}_{\mathcal{B}_s}^{t,i}$ for branches $\mathcal{B}_f$ and $\mathcal{B}_s$:

$$\mathrm{M}_{\mathcal{B}_f}^{t,i}=\sum_{k\in\mathcal{T}(\boldsymbol{P}_{C_f})}\mathcal{M}_k^{t,i},\mathrm{M}_{\mathcal{B}_s}^{t,i}=\sum_{k\in\mathcal{T}(\boldsymbol{P}_{\boldsymbol{X}_s})}\mathcal{M}_k^{t,i}, \tag{10}$$

where $\mathcal{M}_k^{t,i}$ is the attention map of token k.

Then, we apply the spatial softmax $\mathbf{S}_s$ to enhance the sharpness of each mask's spatial distribution, followed by a pixel-wise softmax $\mathbf{S}_c$ across two branches to control the overall blending intensity. The resulting blending mask $\hat{\mathrm{M}}^{t,i}$ is as follows:

$$\hat{\mathrm{M}}^{t,i}=\left(\hat{\mathrm{M}}_{\mathcal{B}_f}^{t,i},\hat{\mathrm{M}}_{\mathcal{B}_s}^{t,i}\right)=\mathbf{S}_c\left(\mathbf{S}_s\left(\mathrm{M}_{\mathcal{B}_f}^{t,i}\right),\lambda\mathbf{S}_s\left(\mathrm{M}_{\mathcal{B}_s}^{t,i}\right)\right), \tag{11}$$

where λ controls the relative blending strength of the two branches.

$\hat{\mathrm{M}}^{t,i}$ determines the spatially varying impacts of each branch dynamically. Therefore, the selective blending of the two branches is performed with weighting by $\hat{\mathrm{M}}^{t,i}$:

$$\hat{\mathcal{F}}_{\mathcal{I}}^{t,i}=\hat{\mathcal{F}}_{\mathcal{B}_f}^{t,i}\odot\hat{\mathrm{M}}_{\mathcal{B}_f}^{t,i}+\hat{\mathcal{F}}_{\mathcal{B}_s}^{t,i}\odot\hat{\mathrm{M}}_{\mathcal{B}_s}^{t,i}, \tag{12}$$

where $\hat{\mathcal{F}}_{\mathcal{B}_f}^{t,i}$ and $\hat{\mathcal{F}}_{\mathcal{B}_s}^{t,i}$ are the unprocessed outputs from the subject branch and the image branch at the i-th cross-attention layer during denoising step t, and $\odot$ represents pixel-wise multiplication,. This selective integration facilitates coherent and seamless blending, as well as ensuring the preservation of the background within the target region.

To further preserve the source image content outside of $\mathcal{R}_f$, we calculate the output of SBM as follow:

$$\hat{\mathcal{F}}^{t,i}=\hat{\mathcal{F}}_{\mathcal{I}}^{t,i}\odot\mathrm{M}_U+\hat{\mathcal{F}}_{\mathcal{B}_s}^{t,i}\odot(1-\mathrm{M}_U), \tag{13}$$

where M_U is the user mask specifying the target region $\mathcal{R}_f$. We use $\hat{\mathcal{F}}^{t,i}$ to update the subject branch output $\hat{\mathcal{F}}_{\mathcal{B}_f}^{t,i}$.

4.4 Semantic Editing

UniCanvas generates both the target subject and source image conditioning on textual prompts, which are respectively generated by two branches and then adaptively integrated to obtain the final synthesis results. This framework not only achieves superior image blending but also allows for flexible semantic editing of the content in both branches simply by modifying the corresponding textual prompt.

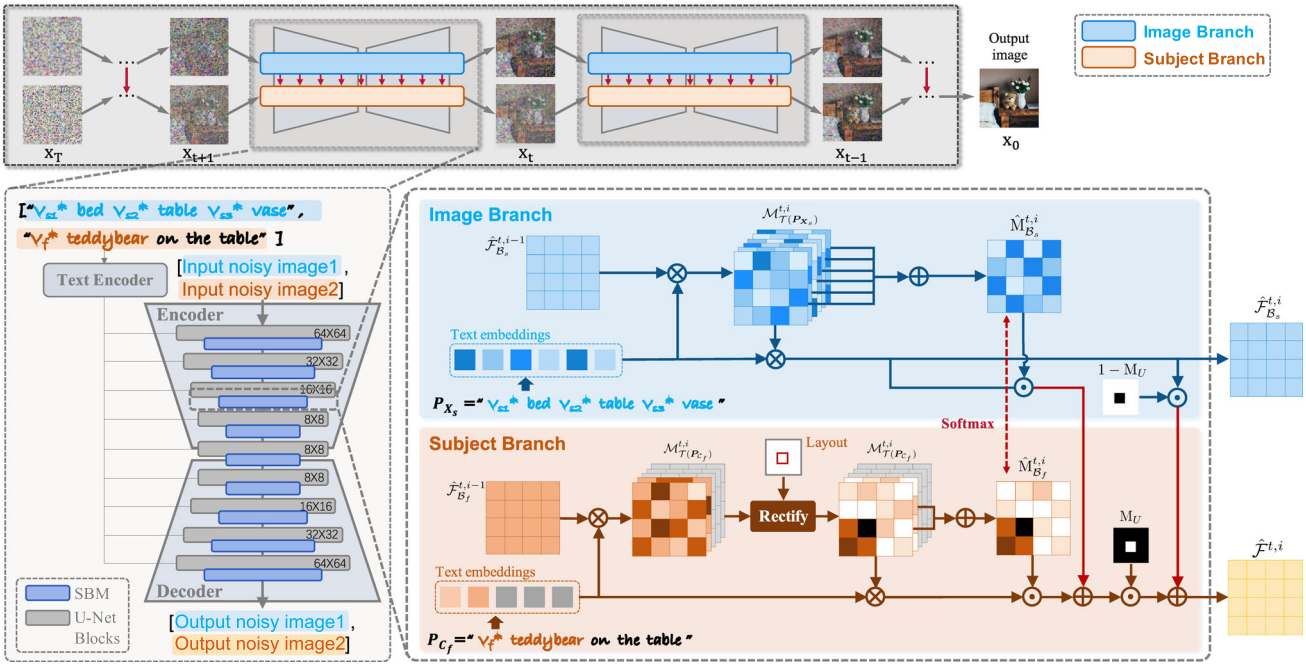

Fig. 5 Overall framework of the inference process. We illustrate this process with an example. UNICANVAS achieves unified editing with two collaborative text-to-image generative branches, namely the subject branch and the image branch. The subject branch is conditioned on $\boldsymbol{P}_{C_f}$ and is tasked with generating the target subject in the specified region, while the image branch is conditioned on $\boldsymbol{P}_{X_s}$ and is responsible for faithfully reconstructing the source image. These two generative branches are integrated using a Selective Blending Module (SBM) at each cross-attention layer to achieve subject-driven editing. SBM employs cross-attention maps to dynamically determine aggregation weights of two branches. Semantic editing can be performed on both the blended subject and the source image by making corresponding textual modifications to $\boldsymbol{P}_{C_f}$ and $\boldsymbol{P}_{X_s}$

4.4.1 Semantic Editing of Blended Subject

To perform edits such as property modifications or accessorization on the blended subject, we simply need to make the corresponding adjustments to $\boldsymbol{P}_{C_f}$. For instance, let us consider the example depicted in Fig. 5. If we want to decorate the teddybear with hat, we need to alter $\boldsymbol{P}_{C_f}$ = "V_f^* `teddybear on the table`" to $\boldsymbol{P}^{'}_{C_f}$ = "V_f^* `teddybear wearing hat on the table`". Then, the token set of the subject branch needs to be updated from $\mathcal{T}(\boldsymbol{P}_{C_f})$ = { V_f^*, "`teddybear`" } to $\mathcal{T}(\boldsymbol{P}^{'}_{C_f})$ = {V_f^*, "`teddybear`", "`wearing`", "`hat`"}, thereby altering the dynamic aggregation mask of the subject branch to encompass the edited content in the generated image.

4.4.2 Semantic Editing of Background Scene

Regarding the background scene, semantic edits such as artistic style variations or object replacement can be achieved by simply making corresponding modifications to $\boldsymbol{P}_{X_s}$. For instance, to alter the color of the vase in the background scene to red, we modify the textual condition from $\boldsymbol{P}_{X_s}$ = "V_{s1}^* `bed` V_{s2}^* `table` V_{s3}^* `vase`" to $\boldsymbol{P}^{'}_{X_s}$ = "V_{s1}^* `bed` V_{s2}^* `table red vase`". Accordingly, we alter the token set of the image branch from $\mathcal{T}(\boldsymbol{P}_{X_s})$ to $\mathcal{T}(\boldsymbol{P}^{'}_{X_s})$, which contains all tokens in $\boldsymbol{P}^{'}_{X_s}$. We denote the intersection of $\mathcal{T}(\boldsymbol{P}_{X_s})$ and $\mathcal{T}(\boldsymbol{P}^{'}_{X_s})$ as $\mathcal{T}_I$ (i.e., which contains the unchanged tokens {"V_{s1}^*", "`bed`", "V_{s2}^*", "`table`", "`vase`" }). The reconstructed background image conditioned on the modified textual prompt $\boldsymbol{P}^{'}_{X_s}$ may suffer from significant undesired structural and content alterations. Since UNICANVAS reconstructs the background scene through a text-to-image generation process, textual editing methods such as Prompt-to-Prompt Hertz et al. (2023) can be utilized to facilitate the preservation of unedited content by injecting the attention maps of the unedited prompt $\boldsymbol{P}_{X_s}$. Specifically, before editing, we perform the denoising diffusion implicit model (DDIM) Song et al. (2021) sampling conditioned on $\boldsymbol{P}_{X_s}$ to reconstruct the source image. DDIM is one of the most widely used diffusion frameworks. The DDIM sampling process, which generates images from initial noise, becomes deterministic by setting the noise variance to 0. This ensures that the same output is produced when provided with the same initial noise. We store the cross-attention maps $\mathcal{M}_{ori}$ of the tokens in $\mathcal{T}_I$ at every DDIM sampling step. At inference, we conduct DDIM sampling conditioned on the edited text prompt $\boldsymbol{P}^{'}_{X_s}$. For the unchanged tokens

p (i.e., tokens in $\mathcal{T}_I$), we replace their associated attention maps $\mathcal{M}^{p,t}$ with the corresponding attention maps in $\mathcal{M}^{p,t}_{ori}$ at timestep t:

$$\hat{\mathcal{M}}^{p,t} = \begin{cases} \mathcal{M}^{p,t}_{ori}, \text{ if } t > t_\tau \text{ and } p \in \mathcal{T}_I \\ \mathcal{M}^{p,t}, \text{ otherwise} \end{cases}, \tag{14}$$

where $\hat{\mathcal{M}}^{p,t}$ is the modified attention maps, and the modification is applied only before timestep t_τ. The original attention maps contain spatial layout and geometry information of the source image, which facilitates the preservation of the structural details.

5 Experiments

5.1 Experiment Setup

5.1.1 Dataset

We compile a new dataset comprising a total of 104 samples. Each sample in the dataset consists of a real image, a target subject, a user mask, along with query prompts. The real images are collected from websites that allow redistribution, encompassing a wide variety of categories such as indoor scenes and natural landscapes. The target subjects are derived from existing customized generation works Gal et al. (2023a); Ruiz et al. (2023); Kumari et al. (2023). These subjects cover a diverse array of categories, spanning pets, toys, plushies, etc., each of which is specified by several reference images.

5.1.2 Evaluation Metrics

Subject-driven editing aims to render a given subject in a specified region of the source image with high subject fidelity. Besides, the generated subject should be reasonable and compatible with the background scene. Therefore, we use the following four metrics to evaluate the quality of subject-driven editing across three aspects. 1) LPIPS Zhang et al. (2018), which measures the LPIPS perceptual distance between the input image and the generated image. 2) Quality Score(QS) Gu et al. (2020), which evaluates the authenticity and quality of the generated image. 3) Local CLIP Radford et al. (2021) and DINO Caron et al. (2021) score. We crop the images along the target region and calculate the local CLIP and DINO similarity between the cropped image and reference images, denoted as $S^{\mathrm{I}}_{\mathrm{CLIP}}$ and S_{DINO}, respectively. These two metrics both gauge visual alignment, and higher values indicate that the blended subject is more similar to reference images. Compared with CLIP, DINO can better capture the unique features of each subject, thereby better reflecting fine subject similarity rather than coarse class similarity. Semantic editing aims to precisely modify an image based on a query prompt while preserving the maximum amount of details from the original image We evaluate the overall quality of semantic editing using the following two metrics: 1) LPIPS, the LPIPS distance between the source image and the edited image. 2) $S^{\mathrm{T}}_{\mathrm{CLIP}}$, which evaluates the prompt fidelity by measuring the average cosine similarity between the CLIP embeddings of the query prompt and the edited image.

5.1.3 Implementation Details

We implement our method using PyTorch and conduct all experiments on NVIDIA RTX 4090 GPUs with 24GB of memory. We employ the released Stable Diffusion V1.4 Stable diffusion (2022) as initialization to provide a robust image prior. Stable Diffusion is a powerful text-to-image LDM that is pre-trained on 512×512 images from the LAION dataset Schuhmann et al. (2021), and its latent code has a spatial size of 64×64.

We fine-tune the pre-trained model with a batch size of 6. During fine-tuning, we jointly optimize the learnable token embedding along with a subset of parameters in cross-attention layers Kumari et al. (2023). The base learning rate of model parameters and concept embeddings is 10^{-5}, which is scaled by the batch size to yield a learning rate of 6×10^{-5}. For target subjects, all reference images are first resized to 512×512. We set the length of concept-specific tokens to 3. We set the length of the prefix prompt as $M = 1$ and the length of the suffix prompt as $L = 1$, initializing them with "a" and ".", respectively. The learning rate of the image-specific contexts is set to 6×10^{-4}, which is ten times the base learning rate. As for source image, we set the learning rate of the image-specific contexts to 6×10^{-4}. The source image and the reference images of the target subject are utilized to jointly fine-tune the pre-trained model. At inference, images are generated using 50 DDIM sampling steps with a classifier-free guidance scale of 6 for all compared methods. The generated results have an image resolution of 512×512 with a latent dimension of 64×64.

5.2 Qualitative Evaluation

To demonstrate the effectiveness of UniCanvas for unified image editing, we present sample generations covering a variety of real images.

5.2.1 Subject-Driven Editing

In Fig. 6, we pair different source images and target subjects for subject-driven editing, with the target region indicated by the red bounding box. As observed, UniCanvas precisely renders the target subject to the specified target region while consistently preserving its key identifying visual features.

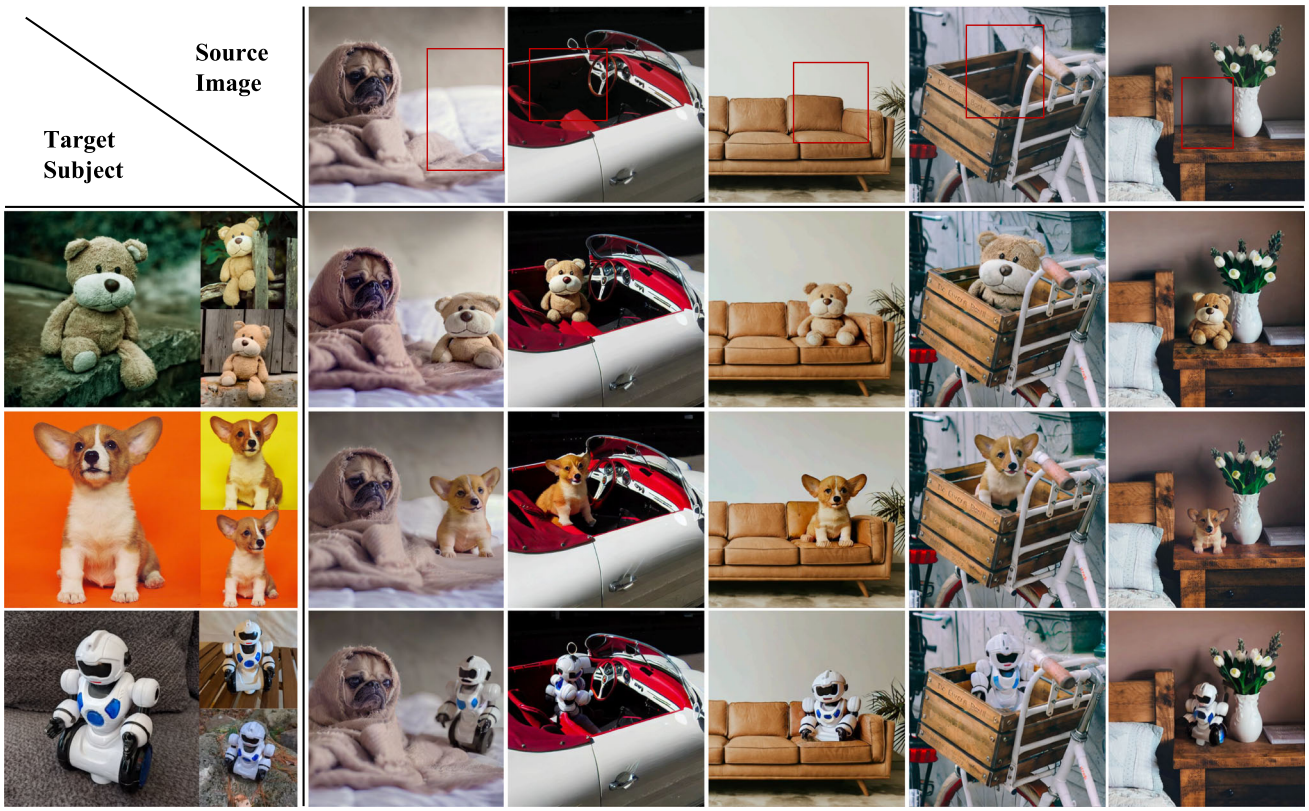

Fig. 6 Subject-driven Editing. We pair different source images and target subjects for subject-driven editing. Reference images of target subjects are shown on the left, and source images are shown on the top. The target region is indicated by the red bounding box. UNICANVAS can seamlessly render target subjects into the target region with high subject fidelity

Most importantly, our method exhibits powerful capabilities in affordance perception, achieving coherent and seamless visual composition. Perceiving varying background contexts, including factors such as viewpoints and illumination conditions, the generated target object can adaptively adjust its pose and lighting features to ensure geometric and semantic harmony and compatibility with the background scene. Furthermore, our method demonstrates strong generalization and robustness to various region conditions, maintaining high visual fidelity of the target subject even when dealing with challenging editing involving small target regions.

UNICANVAS can also achieve cross-domain subject-driven editing, seamlessly blending target subjects into specific contexts across diverse domains. We show sample generation in Fig. 7. The reference images of the target subjects are from the photorealism domain, while the source images are from domains of oil painting, watercolor painting, and an unknown artistic style. As observed, UNICANVAS adaptively adjusts and switches domains of the target subjects to align with background domain while preserving their identities. Additionally, UNICANVAS embellishes blended objects with supplementary background-related elements (e.g., the water splashes at the intersection of dog's paws and the stream's surface.), generating more harmonious images.

5.2.2 Semantic Editing

By modifying the conditioning prompt of the subject branch, we can achieve the corresponding edits on the blended subject. We show examples in Fig. 8. Edits such as property modification and accessorization are applied to the blended subjects. As we can see, UNICANVAS demonstrates editing capabilities on various target subjects guided by different textual prompts. The edited subjects faithfully adhere to the editing instructions while effectively preserving their key identifying features.

Meanwhile, UNICANVAS enables flexible editing of the background scene by modifying the conditioning prompt $\boldsymbol{P}_{X_s}$. Figure 8 also presents sample generations of background editing. We conduct various edits such as object modification and artistic style variations on the background scene. As observed, UNICANVAS showcases the convincing ability to edit diverse background scenes with corresponding textual guidance. Besides, the resultant images maintain high authenticity and structural integrity, preserving the original

Fig. 7 Cross-domain Composition. Each row displays an example, with the target regions indicated by red bounding boxes. The reference images of the target subjects are all from the photorealism domain. The source images are from domains of oil painting, watercolor painting, and an unknown artistic style, respectively. UniCanvas adaptively adjusts and switches domains of the target subjects, achieving seamless and harmonious blending (Color figure online)

geometric and semantic details of the source image in the unedited portion. In $\boldsymbol{P}_{X_s}$, the subjects in the source image are represented as "V_s^* <noun>". Modifying "<noun>" to "<new-noun>" while preserving "V_s^*" can blend properties of "<noun>" into the generated "<new-noun>", as illustrated in the second and third rows of "Edit I" in Fig. 8.

Semantic edits do not introduce any undesired artifacts in subject-driven editing. The edited images still maintain both semantic and geometric compatibility between blended subjects and background scenes.

5.3 Comparisons

5.3.1 Baselines

For subject-driven editing, we select 6 related approaches as baselines for comparison: 1) Custom Diffusion Kumari et al. (2023), a multi-concept customized text-to-image generation method. 2) Paint-By-Example Yang et al. (2023). 3) ObjectStitc Song et al. (2023). 4) SycoNet Niu et al. (2023), the state-of-the-art image harmonization method. 5) Anydoor Chen et al. (2024b), the state-of-the-art reference-based editing method, generates the given subject in the specified region of the source image using inpainting. 6) DreamCom Lu et al. (2023a), a simple baseline that fine-tunes the pretrained text-guided inpainting model for subject-driven editing. As for semantic editing, UniCanvas aims to perform intuitive semantic manipulation using only text, so we compare our results to text-only semantic editing methods: 1) DiffEdit Couairon et al. (2023). 2) Null text inversion Mokady et al. (2023). 3) Imagic Kawar et al. (2023). 4) SINE Zhang et al. (2023b). 5) LEDITS++ Brack et al. (2024). 6) InfEdit Xu et al. (2024).

For methods that only receive one image as reference, we traverse all available references, sequentially utilizing each image as conditions for generation. Subsequently, we select the best result from these generated outputs for comparison.

5.3.2 Quantitative Comparisons

For each competing method, we randomly generate 6 results for each editing case and report the average value for comparison.

Table 2 presents the quantitative comparison results of subject-driven editing. The performance of Custom Diffusion on region-based metrics, including local $S^{\mathrm{I}}_{\mathrm{CLIP}}$ and local S_{DINO}, is consistently low because it lacks the capability to generate target objects in designated regions. Besides, Custom Diffusion exhibits a low LPIPS score as it is unable to faithfully reconstruct the background scene when generating multiple concepts together. Paint-by-Example and ObjectStitc show low S_{DINO}, indicating that they struggle to capture fine-grained identifying features of the given subject. The outputs of SycoNet obtain high $S^{\mathrm{I}}_{\mathrm{CLIP}}$ and S_{DINO} since it directly segments and copies the target subject from the reference image before harmonization. The images generated by ObjectStitc and DreamCom appear implausible according to the QS. Our approach achieves high-quality subject-driven editing, generating the given subject in the specified region with high visual fidelity.

We provide the quantitative results of semantic editing in Table 3. While LEDITS++ achieves better similarity to the source image, it exhibits a low CLIP score. This suggests that LEDITS++ performs less meaningful edits, aligning poorly with the query prompt while leaving the source image largely unchanged. In contrast, InfEdit achieves a high CLIP score, but its high LPIPS score indicates significant and undesired alterations to the source image. Our method demonstrates significantly higher textual CLIP alignment, indicating more meaningful editing that aligns better with the editing prompt.

5.3.3 Qualitative Comparisons

We present qualitative comparisons to further visually demonstrate the effectiveness of UniCanvas over the baselines in subject-driven editing and semantic editing. For each example, we randomly generate 6 results using each com-

Fig. 8 Unified editing. Each row presents an example, with the target regions indicated by red bounding boxes. $\boldsymbol{P}_{\boldsymbol{X}_s}$ and $\boldsymbol{P}_{\boldsymbol{C}_f}$ are the textual guidance of image branch and subject branch, respectively. UNICANVAS provides a unified image manipulation interface for subject-driven editing and semantic editing. In addition to rendering target subjects into the target region with high fidelity, UNICANVAS can perform semantic editing on both the target subject and background scene simply by making corresponding modifications to $\boldsymbol{P}_{\boldsymbol{X}_f}$ and $\boldsymbol{P}_{\boldsymbol{C}_s}$ (Color figure online)

peting method and select the best one from them for visual comparison.

We show qualitative comparisons of subject-driven editing in Fig. 9. For images synthesized by Custom Diffusion, the background scenes exhibit dramatic variations in the overall structure and layout. The faithful reconstruction of the background image requires overfitting, which leads to the omission of the target subject. Additionally, Custom Diffusion is unable to specify the spatial distribution of the blended object. The results generated by SycoNet show inconsistency between the composited subject and the background image. The underlying cause is that SycoNet directly copies the foreground subject from the reference image, which lacks variability in poses and articulations. However, the appearance of the subject in the reference image usually mismatches with various background contexts, resulting in implausible images. Paint-by-example, Anydoor, and DreamCom all use inpainting to generate the subject in a specified region, where Paint-by-example and Anydoor is training-free and DreamCom is fine-tuned on the given target subject. Paint-by-example falls short in preserving the key distinguishing features of the target subject and sometimes generates unnatural objects. These inpainting-based methods discard the content information within the target region, leading to an alteration of the background in some cases. Furthermore, the generated object often appears incongruous with the background context in terms of semantic and geometric consistency, which shows subpar global and local affordance perception. Our method demonstrates powerful affordance perception, achieving reasonable and realistic editing in various scenes. For instance, the domain perception in row 1, the geometric perception in row 2, and the reflection perception in row 5, which are unattainable by other methods.

We present visual comparison results of UNICANVAS with semantic editing baselines in Fig. 10. The source image for semantic editing is generated by UNICANVAS. All of these

Table 2 Quantitative comparisons

Algorithm	LPIPS (↓)	QS (↑)	$S^{\mathrm{I}}_{\mathrm{CLIP}}$ (↑)	S_{DINO} (↑)
Custom Diffusion Kumari et al. (2023)	0.758	71.11	0.6815	0.3264
Paint-by-Example Yang et al. (2023)	0.2466	72.30	0.7662	0.5260
ObjectStitc Song et al. (2023)	0.2212	62.64	0.7649	0.5981
SycoNet Niu et al. (2023)	0.2448	78.81	0.8067	0.6337
DreamCom Lu et al. (2023a)	**0.2028**	71.05	0.7226	0.6124
Anydoor Chen et al. (2024b)	0.2734	74.38	0.7728	0.6062
UniCanvas (ours)	0.2492	**82.47**	**0.8145**	**0.6538**

Bold value represents the best results
The quantitative comparison results of our method and subject-driven editing baselines

Table 3 The quantitative comparison results of our method and semantic editing baselines

Algorithm	LPIPS (↓)	$S^{\mathrm{T}}_{\mathrm{CLIP}}$ (↑)
DiffEdit Couairon et al. (2023)	0.274	0.3201
Null text inversion Mokady et al. (2023)	0.2874	0.3218
Imagic Kawar et al. (2023)	0.2497	0.3123
SINE Zhang et al. (2023b)	0.3067	0.3082
LEDITS++ Brack et al. (2024)	**0.1543**	0.3095
InfEdit Xu et al. (2024)	0.471	0.3306
UniCanvas (ours)	0.2616	**0.3324**

Bold value represents the best results

editing tasks encompass semantic editing for both the target subject and the background scene. Methods like SINE are limited to edit one object within a single editing procedure, thereby requiring iterative operations encompassing multiple editing procedures to attain the target editing. As observed, the baselines commonly exhibit inadequate affordance perception and context understanding, thereby failing to disentangle distinct semantics in the source image and align them with the textual editing instructions. Therefore, these methods struggle to precisely locate the multiple editing regions and accurately perform the corresponding edits, while also inducing undesired structural and texture changes to the unedited regions. Besides, the blended subject suffers from undesired alterations during semantic editing, leading to a quality degradation of preceding subject-driven editing. Contrarily, UniCanvas precisely identifies multiple semantic regions and matches them with corresponding editing instructions, achieving a wide range of realistic and meaningful editing. Moreover, UniCanvas accomplishes semantic editing alongside subject-driven editing in a compact pipeline rather than a two-stage editing process, preventing fidelity degradation of the blended subject.

5.3.4 User Study

To further evaluate the proposed method from the perspective of human perception and preferences, we conduct a user study to gather subjective assessments of editing quality from users. We perform paired tests comparing the proposed UniCanvas with subject-driven editing baselines and semantic editing baselines, involving a total of 52 participants without relevant backgrounds. In each comparison, users are presented with the source image $\boldsymbol{X}_s$ with annotations of the target region, reference images of the target subject C_f, the textual prompt for semantic editing, and two corresponding generations from the two compared methods (ours and the baseline) in a random order. In comparing the quality of subject-driven editing, users are asked to select the better image by answering the question: "Which image achieves a more realistic and natural integration between the blended subject and background scene, with the subject within in the target region closely resembling the provided reference images? When comparing semantic editing quality, users are asked to select the better image with the question: "Which image better achieves the requested edit while preserving most of the original details?" Comparison results are collected for the evaluation, and the aggregated result is shown in Table 4. As observed, our method exhibits a dominant user preference over other methods in terms of both the quality of subject-driven editing and semantic editing. This further validates the effectiveness of the proposed UniCanvas in the unified image editing task.

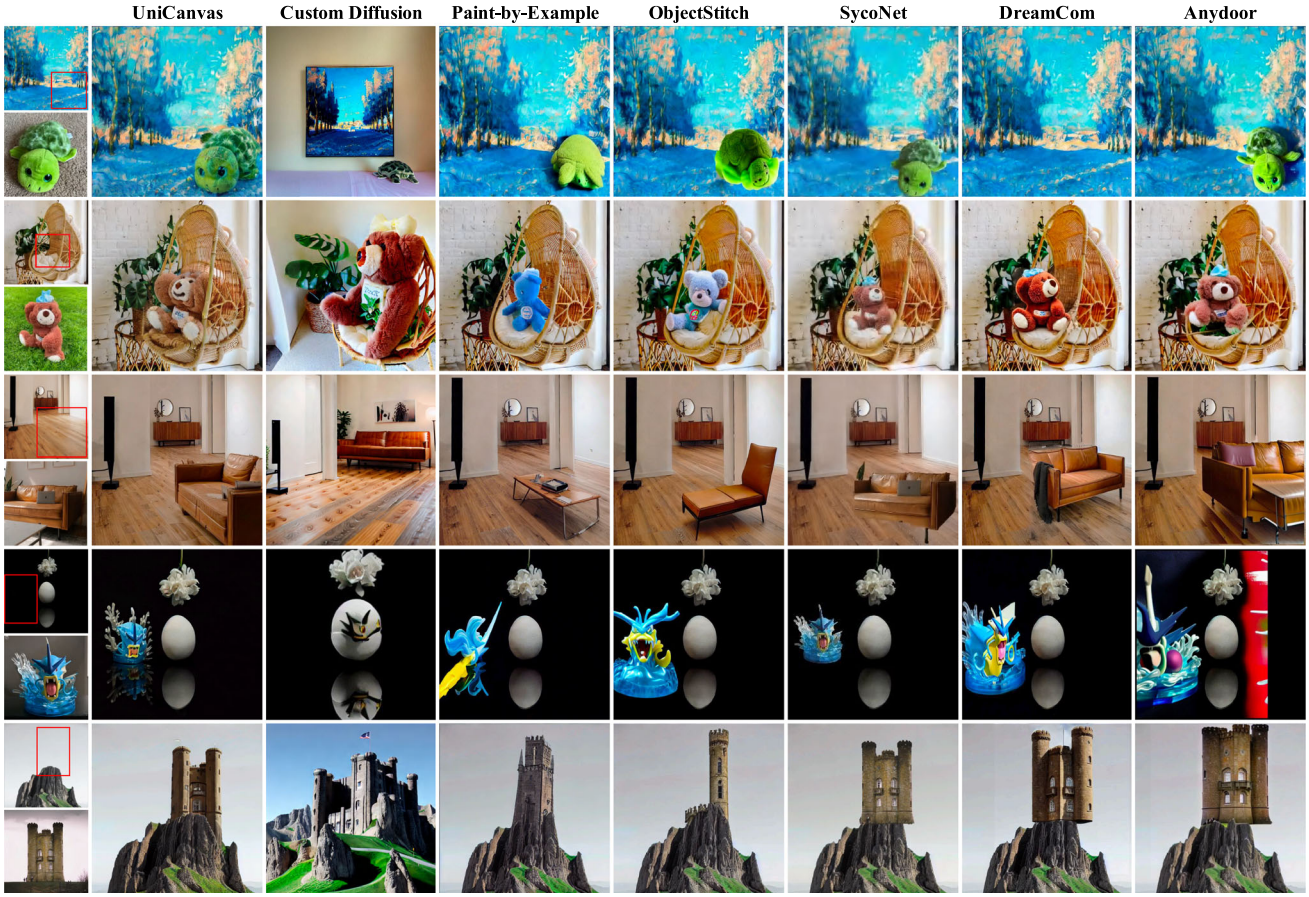

Fig. 9 Visual comparisons of UNICANVAS with subject-driven editing baselines. The first column presents the source image (top) and the target subject (bottom), with the target regions indicated by red bounding boxes. Our method demonstrates powerful affordance perception. For instance, the domain perception in row 1, the geometric perception in row 2, and the reflection perception in row 4, which are unattainable by other methods (Color figure online)

5.4 More Applications

5.4.1 Subject-Driven Editing with Extended Settings

Multiple Subject-driven Editing All examples of subject-driven editing in the experiments above involve a single subject with one target region. UNICANVAS can further achieve multiple subject-driven edits by simultaneously generating multiple subjects in various designated regions within a single inference process, which is challenging for existing subject-driven methods. We present sample generations in Fig. 11. As shown, our approach can simultaneously render multiple target subjects in their respective regions with high subject fidelity, while maintaining geometric and semantic harmony with the background scene. However, we observe that the reconstruction quality of the source image slightly degrades as the number of subjects increases.

Subject-driven Editing with Irregular Target Regions In the examples above, the target regions for subject-driven editing are defined by bounding boxes. In this section, we conduct additional experiments on subject-driven editing using irregular target regions derived from hand-drawn sketches. Sample generations are present in Fig. 12. As demonstrated, UNICANVAS also support irregular, hand-drawn masks for subject-driven editing, delivering compelling results. This provides users with greater flexibility and choice in practical applications.

5.4.2 Subject Replacement

Subject replacement aims to replace a specified subject C_r in a source image $\boldsymbol{X}_s$ with a target subject C_t. UNICANVAS can accomplish subject replacement without requiring a precise segmentation mask of the replaced subject C_r. Specifically, UNICANVAS firstly implants C_t and $\boldsymbol{X}_s$ into the model as the foreground subject and source image. To achieve subject

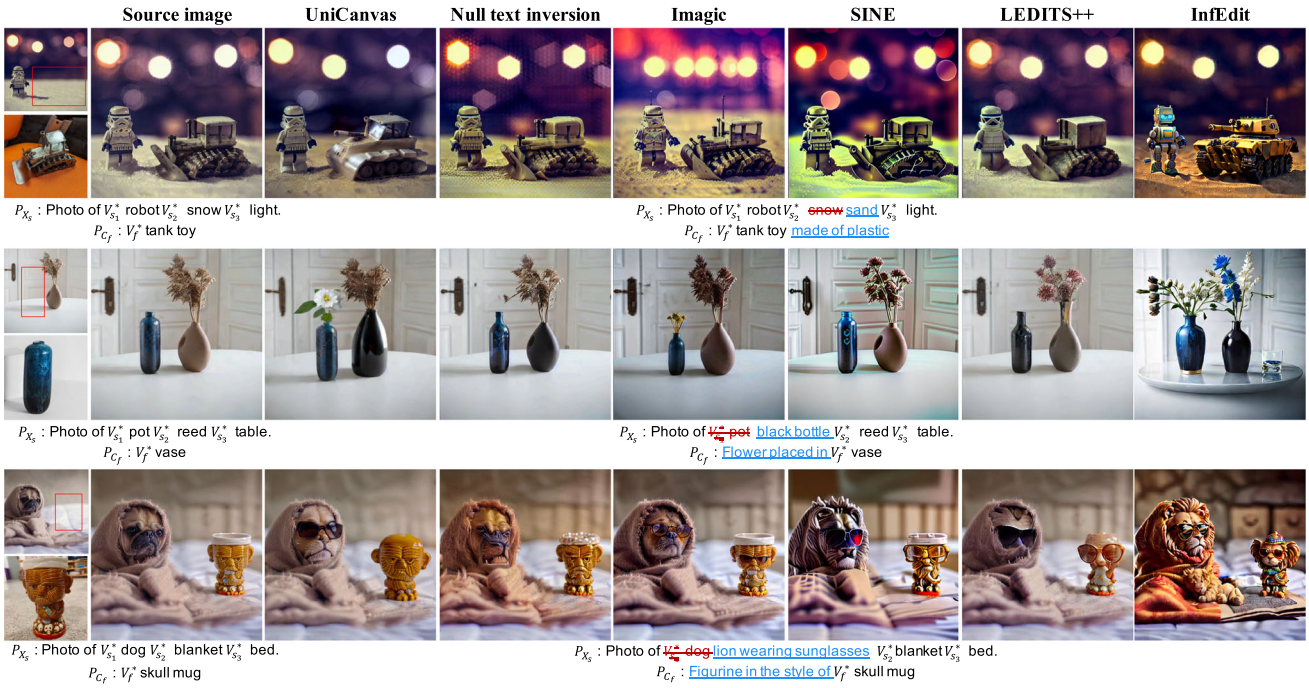

Fig. 10 Visual comparisons of UniCanvas with mask-free semantic editing baselines. The first column presents the source images (above) with the target region indicated, and the target subjects (below). The source images (second column) for semantic editing are obtained using UniCanvas. UniCanvas precisely identifies multiple semantic regions and achieves a wide range of realistic and meaningful editing, surpassing the capabilities of other methods

Fig. 11 Multiple Subject-driven Editing. Each row displays an example, with the target regions indicated by red bounding boxes. UniCanvas can simultaneously generating multiple subjects in various designated regions with high subject fidelity (Color figure online)

Table 4 User study

Algorithm	UniCanvas	
	Subject-driven Editing	Semantic Editing
Custom Diffusion Kumari et al. (2023)	**100** %	-
Paint-by-Example Yang et al. (2023)	**86.54** %	-
ObjectStitc Song et al. (2023)	**83.65** %	-
SycoNet Niu et al. (2023)	**79.81** %	-
DreamCom Lu et al. (2023a)	**72.23** %	-
Anydoor Chen et al. (2024b)	**75.96** %	-
DiffEdit Couairon et al. (2023)	-	**77.88** %
Null text inversion Mokady et al. (2023)	-	**68.27** %
Imagic Kawar et al. (2023)	-	**74.04** %
SINE Zhang et al. (2023b)	-	**84.62** %
LEDITS++ Brack et al. (2024)	-	**75.96** %
InfEdit Xu et al. (2024)	-	**80.77** %

Bold value represents the best results
In each paired comparison, our method is preferred ($\geq$ 50%) over the baseline methods in terms of either subject-driven editing quality or semantic editing quality. The proposed UniCanvas demonstrates an overwhelming user preference over other methods

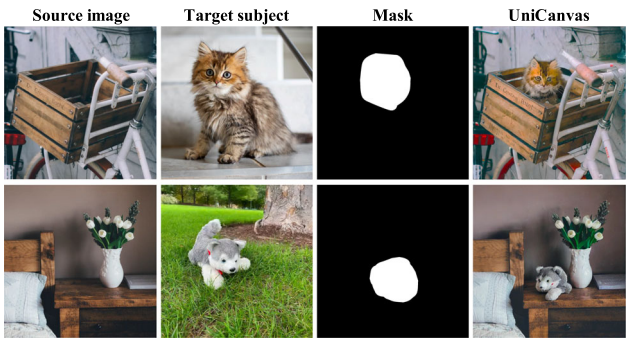

Fig. 12 Subject-driven Editing with Irregular Target Regions. The irregular target regions for editing are derived from hand-drawn sketches, and UniCanvas delivers compelling results

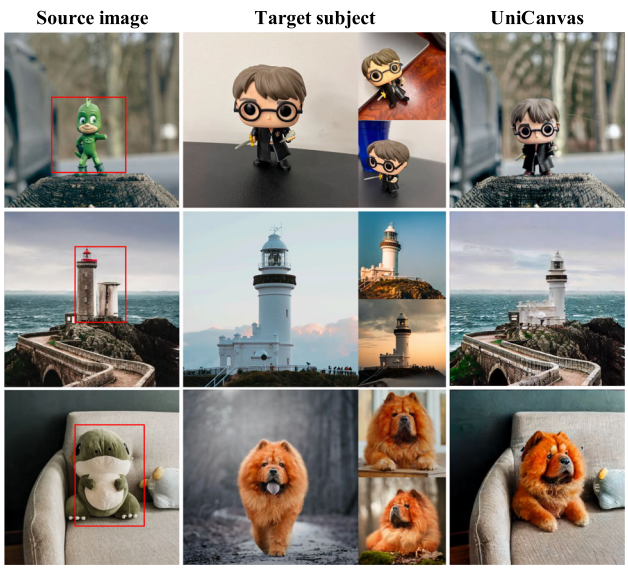

Fig. 13 Subject replacement. Each row displays an example, with the target regions indicated by red bounding boxes. UniCanvas can seamlessly replace any specified subject in the source image with a customized target subject. The key identifying features of the target subject are well preserved in the resulting images (Color figure online)

replacement, we simply need to set the target region of subject C_t to approximately cover the region of C_r and perform the regular UniCanvas inference process. As shown in Fig. 13, the target subject C_t seamlessly replaces the specified subject C_r in the source image $\boldsymbol{X}_s$ and maintains high visual fidelity, generating synthesis images that are free from any undesired feature residue or fusion artifacts. Besides, there are no strict constraints on the categorical or shape relationship between the object to be replaced and the target subject. For instance, in the last row of Fig. 13, the replaced subject and the replacing subject are respectively "`dinosaur plushie`" and "`dog`", which exhibit significant differences in category and appearance.

5.4.3 Spatial Controllable Customized Text-to-image Generation

Existing customized generation methods face challenges in controlling the spatial distribution of the target subject in the generated images. UniCanvas can mitigate this issue. By replacing the customized prompt $\boldsymbol{P}_{\boldsymbol{X}_s}$ with a regular textual prompt as the background condition, UniCanvas can achieve spatially controllable customized text-to-image gen-

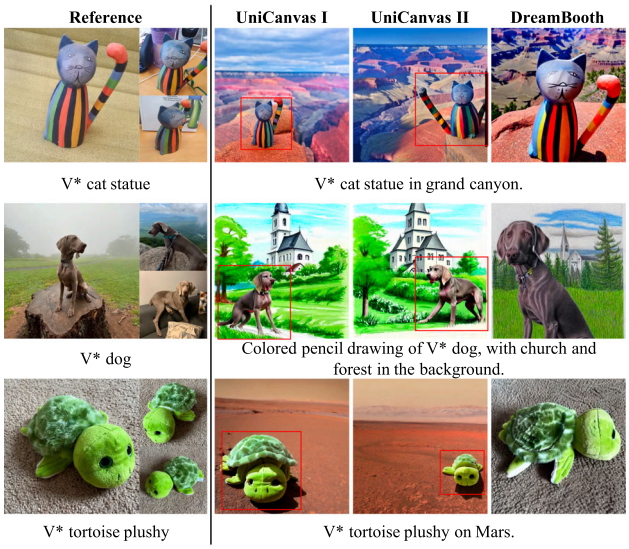

Fig. 14 Spatial Controllable Customized Generation. Each row presents an example, with the reference images shown in the first column and the text prompts indicated beneath the images. For each example, we display two sample generations with different target regions, with the target regions indicated by red bounding boxes. UNICANVAS enhances the controllability of previous customized text-to-image generation methods by generating the target concept in a specified region with high concept and prompt fidelity. Furthermore, UNICANVAS can render the target concept into contexts where previous customized generation methods struggle (Color figure online)

eration, enabling control over the size and location of the generated subject. We present sample generations in Fig. 14, with the red bounding box indicating the target region of the customized subject. As we can see, the target subject can be seamlessly rendered into specified region with various visual contexts, significantly enhancing the controllability over generated images. Furthermore, previous customized generation methods struggle to render the target subject in contexts with low co-occurrence probability Ruiz et al. (2023), where the pre-trained model may not acquire sufficient knowledge during the pre-training process. As exemplified in the last row of Fig. 14, DreamBooth fails to render the customized concept "`V* tortoise plushy`" into the context "`on Mars`". This is likely due to the pre-trained model having a low generative prior in the context "`tortoise plushy on Mars`". Under the dual-branch framework of UNICANVAS, the target subjects and the specified contexts, which have low co-occurrence probability, are individually generated and then integrated during the denoising process to produce the final output. Therefore, UNICANVAS can effectively address this inherent limitation of previous customized generation methods.

5.5 Ablation Studies

We conduct ablation studies to further evaluate the impact of distinct components of UNICANVAS, which consists of a fine-tuning strategy and an inference-time editing process.

5.5.1 Fine-Tuning Strategy

To empower the customized model with the ability to render the target subject in the specified region with high visual fidelity, we introduce a Region-Aware Customization (RAC) strategy, which comprises components including Region Variability Augmentation (RVA) and Image-Specific Prompt (ISP). We construct three variants to gradually ablate and validate them. 1) we eliminate the entire region-aware customization strategy and customize the target subject using the original text-to-image reconstruction fine-tuning, which we donate as *"w/o RAC"*. 2) In the *"w/o RVA"* scenario, we discard the region variability augmentation and leave the reference images unaltered. The region inside the bounding box of the subject is treated as the foreground region, while the outside region is regarded as the background region. 3) In the *"w/o ISP"* scenario, we replace the image-specific prompt with manually crafted templates (e.g., "`A photo of {}`"), which corresponds to the background region in Eq. (8).

The ablation results are presented in Fig. 15. In the absence of RAC, we observed the disappearance of target objects in the majority of examples, where the target subjects fail to blend into source images. In a few examples without RAC, the target subjects can be generated but exhibit a significant deterioration in subject fidelity. Further ablation of RVA and ISP revealed the causes of these observed phenomena. The absence of RVA results in the generation of target subjects with low fidelity, indicating that RVA can enhance the generation generalization of the implanted subject to local regions, thereby improving subject fidelity. Moreover, we observe the absence of target subjects in the majority of the *"w/o ISP"* scenario, demonstrating that ISP can greatly improve the co-occurrence probability of the implanted subject with the source image.

5.5.2 Inference-Time Editing

During inference, UNICANVAS generates the target subject and the source image separately using two branches, which are adaptively integrated by weighting with dynamic aggregation masks in each cross-attention layer. 1) In the *"w/o dual branches"* scenario, we generate images using a single branch with the text prompt $\boldsymbol{P}_u =$ "$\boldsymbol{P}_{\boldsymbol{X}_s}, \boldsymbol{P}_{C_f}$". In Eq. (8), $\boldsymbol{P}_{\boldsymbol{X}_s}$ is linked to the background region, whereas $\boldsymbol{P}_{C_f}$ is associated with the foreground region. 2) In the *"w/o dynamic mask"* scenario, we verify the necessity of dynamic aggregation masks by ablating it and directly adding two branches.

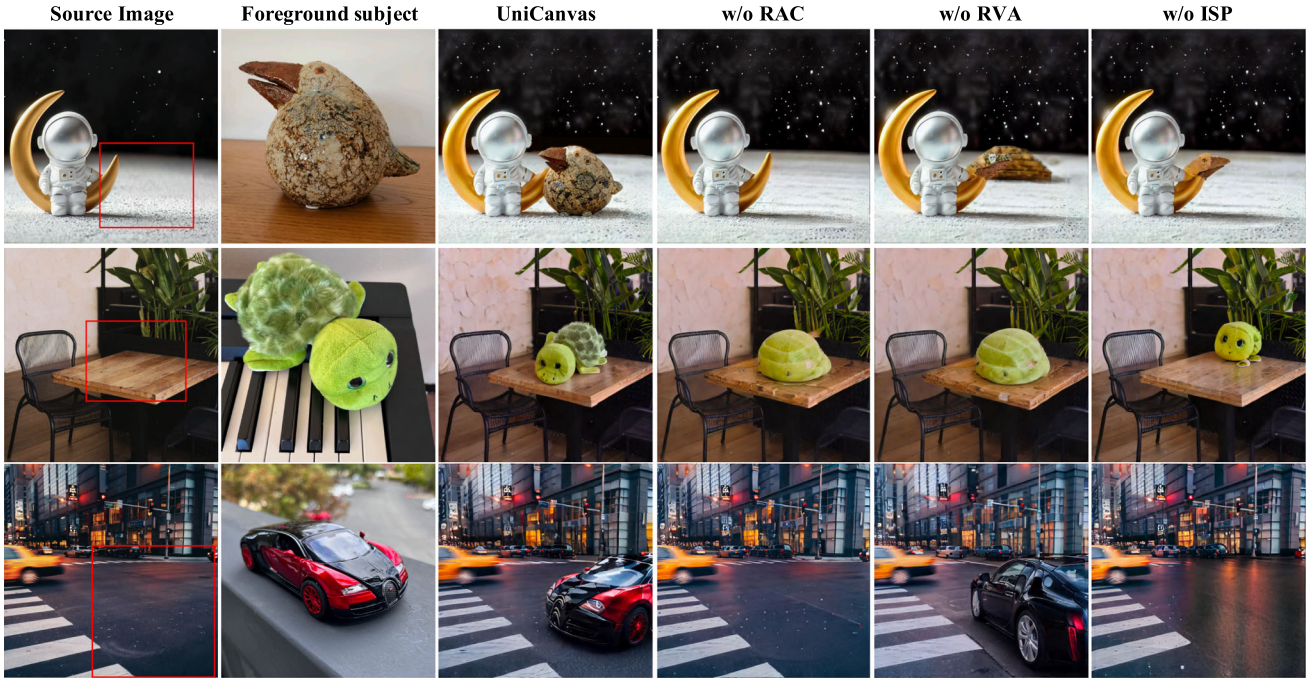

Fig. 15 Ablation study on the fine-tuning strategy of UniCanvas. Each row displays an example, with the target regions indicated by red bounding boxes. We demonstrate the necessity of the Region-Aware Customization (RAC) strategy, Region Variability Augmentation (RVA) and Image-Specific Prompt (ISP) in the *"w/o RAC"*, *"w/o RVA"*, and *"w/o ISP"* scenarios, respectively (Color figure online)

Fig. 16 Ablation study on the inference-time editing of UniCanvas. Each row showcases an example, and the target regions are marked by red bounding boxes. We evaluate the impact of the dual branch framework, dynamic aggregation mask, and deep integration in the *"w/o dual branches"*, *"w/o dynamic mask"*, and *"w/o deep integration"* scenarios, respectively (Color figure online)

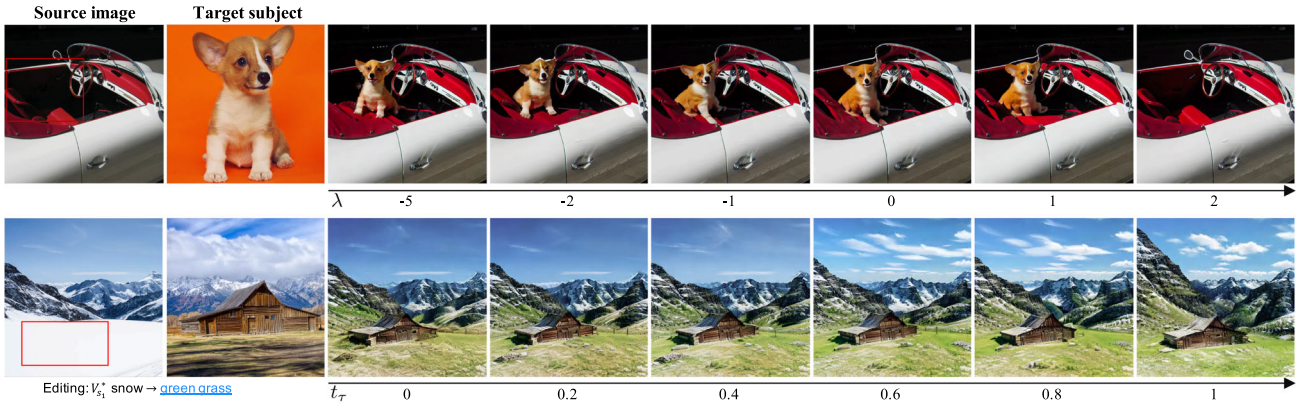

Fig. 17 Sensitivity of λ and t_τ. The model's performance is robust with respect to both λ and t_τ. UNICANVAS can function well across a large parameter interval

3) In the *"w/o deep integration"* scenario, we integrate two branches by only adding the final predicted noise of the U-Net.

The results are presented in Fig. 16. In the *"w/o dual branches"* scenario, the target subject can be successfully generated in the specified region, while the source image fails to be faithfully reconstructed. The variation of the condition prompt (i.e., from the bound prompt $\boldsymbol{P}_{\boldsymbol{X}_s}$ to the prompt $\boldsymbol{P}_u$) severely hinders the reconstruction of background image. The dynamic aggregation mask influences the subject-driven editing quality. For complex backgrounds like the first example in Fig. 16, the absence of aggregation masks results in the target subject suffering from severe fidelity distortion and generating undesired fusion artifacts. Regarding simple backgrounds like the second example in Fig. 16, the subject exhibits residual background from the subject branch. These results demonstrate that dynamic aggregation masks play a key role in maintaining subject fidelity and background preservation within the target region. Without employing deep integration, the source image is unable to be precisely reconstructed, indicating that successful injection of the source image requires multi-scale integration.

5.5.3 Parameter Sensitivity

We introduce additional hyperparameters, λ and t_τ in Eq. 11 and Eq. 14, respectively. In this section, we perform sensitivity experiments to evaluate their effect on model performance. Specifically, we vary λ over the values $-5, -2, -1, 0, 1, 2$ and t_τ over 0, 0.2, 0.4, 0.6, 0.8, 1.0. We present visual examples in Fig. 17. The parameter λ controls the injection strength of the target subject. Results indicate that the model is robust across a wide range of λ, blending the target subject into the target region with high fidelity while preserving the background content. However, the model exhibits performance degradation when λ reaches extreme values. The background content in the target region is altered if λ is too small, while a large λ leads to the unsuccessful blending of the target subject. No notable variation is observed over a broad parameter range for small t_τ. When t_τ approaches 1, the model suffers from undesired structural and content alterations. Overall, the model's performance is not sensitive to either λ or t_τ, and it can function well across a large parameter interval.

5.6 More Comparisons and Discussions

In Sect. 5.3, we focus our comparisons on the most recent semantic editing methods that are closely related to our work. There are several other distinct lines of research in semantic editing. A recently emerging type of editing technique is instruction-based method Brooks et al. (2023); Fu et al. (2024); Li et al. (2024), which edits images based on instructional prompts. These methods use direct instructions (e.g., "remove the sunglasses") to guide the editing process, rather than describing the desired outcome (e.g., "a cat in a garden"). Additionally, a classic research line in semantic editing is GAN-based methods Karras et al. (2019); Wang et al. (2022); Bobkov et al. (2024) that utilize Generative Adversarial Networks (GANs) Goodfellow et al. (2014) as their generative backbone. However, the editing direction in GAN-based methods is typically uncontrollable, requiring additional guidance, such as CLIP Radford et al. (2021), to facilitate specific semantic edits based on textual prompts Crowson et al. (2022). We present visual comparisons between UNICANVAS and the above methods in Fig. 18. As observed, VQGAN+CLIP demonstrates inferior performance compared to diffusion-based models in semantic editing, particularly in editing complex scenes and exhibiting low control precision. ZONE may collapse in some editing tasks that drastically alter the source image and produce significant artifacts. Additionally, it also struggles to accurately

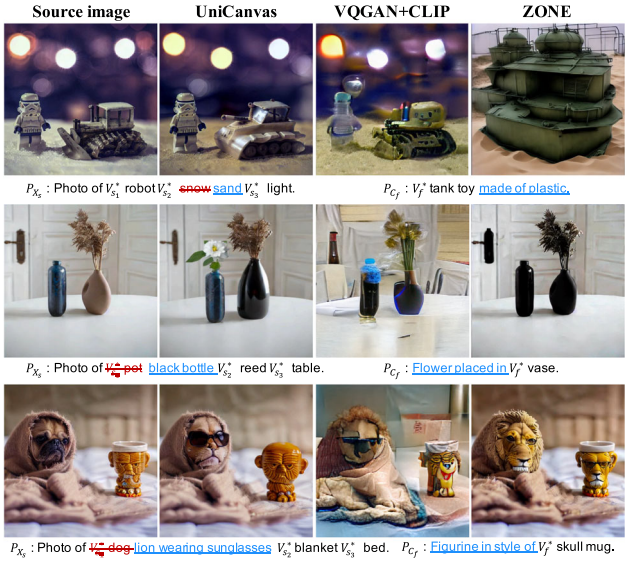

Fig. 18 More Comparison. We compare UNICANVAS with several recent GAN-based methods and instruction-based editing methods. These methods struggle to accurately locate and match multiple editing regions with the given instructions

locate and match multiple editing regions with the given instructions. These results further demonstrate the advantages of the proposed method.

5.7 Limitations and Failure Cases

The experimental results on extensive and diverse editing examples strongly validate the effectiveness of our method across various image editing tasks. While our method performs well in most cases, it has limitations in some challenging editing scenarios. We present several failure cases in Fig. 19. Our method falls short in editing tasks when the target region is extremely small, a limitation that is also observed with other baselines. Empirical observations reveal that the concept fidelity of the blended subject degrades when the ratio of the bounding box area to the entire image area is less than 10%. One potential solution is to magnify the local region for editing and then scale it back to its original size post-editing. Additionally, our method may generate artifacts when semantic conflicts arise between the target object and the scene within the target region of the source image, leading to issues such as fidelity degradation and subject omission. Furthermore, UNICANVAS falls short of state-of-the-art performance in source image reconstruction. This could be enhanced by using more powerful generative models such as SDXL Podell et al. (2023).

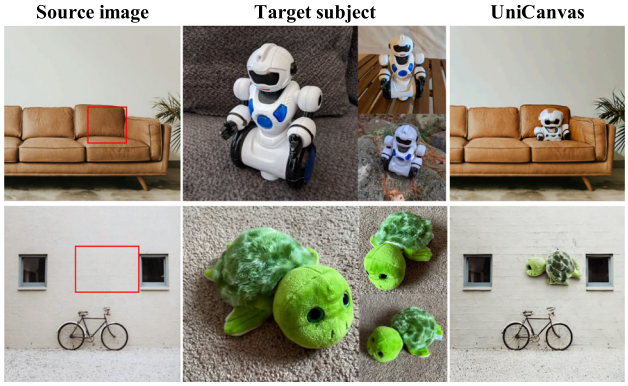

Fig. 19 Limitations and Failure Cases. UNICANVAS faces challenges when 1) the target region is too small, or 2) semantic conflicts exist between the target object and the background scene within the target region, causing issues such as fidelity degradation or subject omission

6 Conclusion

In this paper, we propose UNICANVAS, a compact framework built upon customized text-to-image generation for affordance-aware unified image editing. UNICANVAS provides an integrated interface for subject-driven editing and semantic editing, allowing for multiple high-quality image manipulations on a single real image simply using one inference process. We conduct extensive experiments on various target subject and real image pairs with different target regions. Experimental results demonstrate that UNICANVAS exhibits strong capability in scene affordance perception, enabling it to simultaneously achieve seamless subject-driven editing and precise semantic editing, even in challenging scenes such as cross-domain editing. More applications like subject replacement and spatial controllable customized text-to-image generation can also be achieved under the framework of UNICANVAS.

Potential Societal Impact

The approach presented in this study involves editing and manipulating image content, which could potentially lead to certain societal impacts. One major issue with real image editing and manipulation is the potential misuse to create misinformation and fake content, misleading the public and eroding trust. Additionally, the ability of the proposed method to learn and reproduce visual concepts may raise concerns about copyright infringement. Altering copyrighted images without permission could lead to legal disputes, which may be difficult to detect and pursue. The misuse of this editing method could also give rise to privacy and ethical concerns. On the positive side, high-quality image editing unlocks vast opportunities in the fields of artistic creation and

design, allowing users to explore new forms of visual expression. We hope that the potential negative societal impacts mentioned above can be offset by the new creative possibilities offered by these image editing methods.

Acknowledgements The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China under Grant Nos. U24A20330, 62361166670, 62276132, and 61876085.

Data Availability The source images in Table 2, Table 3, and Table 4 are collected from websites that allow redistribution, and the foregound subjects are derived from existing customized generation works, including Textual Inversion Gal et al. (2023a), DreamBooth Ruiz et al. (2023), and Custom Diffusion Kumari et al. (2023).

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Alaluf, Y., Richardson, E., Metzer, G., & Cohen-Or, D. (2023). A neural space-time representation for text-to-image personalization. *ACM TOG, 42*(6), 1–10.

Azadi, S., Pathak, D., Ebrahimi, S., & Darrell, T. (2020). Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision, 128*, 2570–2585.

Bobkov, D., Titov, V., Alanov, A., & Vetrov, D. (2024). The devil is in the details: Stylefeatureeditor for detail-rich stylegan inversion and high quality image editing. In *CVPR* (pp. 9337–9346) (2024)

Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., & Passos, A. (2024). Ledits++: Limitless image editing using text-to-image models. In *CVPR* (pp. 8861–8870) (2024)

Brooks, T., Holynski, A., & Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *CVPR* (pp. 18392–18402).

Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., & Zheng, Y. (2023). Masactrl: tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV* (pp. 22560–22570) (2023)

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *ICCV* (pp. 9650–9660).

Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M. W., Cohen, W. W. (2024). Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems* **36**.

Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., & Zhao, H. (2024). Anydoor: Zero-shot object-level image customization. In *CVPR* (6593–6602).

Choi, J., Kim, S., Jeong, Y., Gwon, Y., & Yoon, S. (2021) Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938.

Collins, E., Bala, R., Price, B., & Susstrunk, S. (2020). Editing in style: Uncovering the local semantics of gans. In *CVPR* (5771–5780).

Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., & Zhang, L. (2020). Dovenet: Deep image harmonization via domain verification. In *CVPR* (8394–8403).

Couairon, G., Grechka, A., Verbeek, J., Schwenk, H., Cord, M. (2022). Flexit: Towards flexible semantic image translation. In *CVPR* (pp. 18270–18279)

Couairon, G., Verbeek, J., Schwenk, H., & Cord, M. (2023). Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*.

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., & Raff, E. (2022). Vqgan-clip: Open domain image generation and editing with natural language guidance. In *ECCV* (pp. 88–105).

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., Tang, J. (2021). Cogview: Mastering text-to-image generation via transformers. In *NeurIPS* (19822–19835).

Fu, T. J., Hu, W., Du, X., Wang, W. Y., Yang, Y., Gan, Z. (2024). Guiding instruction-based image editing via multimodal large language models. In *ICLR*.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*.

Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., & Cohen-Or, D. (2023). Designing an encoder for fast personalization of text-to-image models. In *Siggraph*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS* (pp. 2672–2680).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014) Generative adversarial nets. In *NeurIPS* (Vol. 27).

Gu, S., Bao, J., Chen, D., & Wen, F. (2020). Giqa: Generated image quality assessment. In *ECCV* (pp. 369–385).

Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al. (2024). Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS* (pp. 15890–15902).

Gupta, A., Satkin, S., Efros, A. A., & Hebert, M. (2011). From 3d scene geometry to human workspace. In *CVPR* (pp. 1961–1968).

Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., & Yang, F. (2023). Svdiff: Compact parameter space for diffusion fine-tuning. In *ICCV* (pp. 7323–7334).

Han, L., Wen, S., Chen, Q., Zhang, Z., Song, K., Ren, M., Gao, R., Stathopoulos, A., He, X., Chen, Y., et al. (2024). Proxedit: Improving tuning-free real image editing with proximal guidance. In *WACV* (pp. 4291–4301).

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). Ganspace: Discovering interpretable gan controls. In *NeurIPS* (pp. 9841–9850).

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross attention control. In *ICLR*.

Karras, T., Laine, S., & Aila, T. (2019) A style-based generator architecture for generative adversarial networks. In *CVPR* (pp. 4401–4410).

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., & Irani, M. (2022). Imagic: Text-based real image editing with diffusion models. In *CVPR* (pp. 6007–6017).

Kim, Y., Lee, J., Kim, J. H., Ha, J. W., Zhu, J. Y. (2023). Dense text-to-image generation with attention modulation. In *ICCV* (pp. 7701–7711).

Kulal, S., Brooks, T., Aiken, A., Wu, J., Yang, J., Lu, J., Efros, A. A., Singh, K. K. (2023). Putting people in their place: Affordance-aware human insertion into scenes. In *CVPR* (pp. 17089–17099).

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., & Zhu, J. Y. (2023). Multi-concept customization of text-to-image diffusion. In *CVPR* (pp. 1931–1941).

Li, B., Qi, X., Lukasiewicz, T., Torr, P. (2019). Controllable text-to-image generation. In *NeurIPS* (pp. 2065–2075).

Li, S., Zeng, B., Feng, Y., Gao, S., Liu, X., Liu, J., Li, L., Tang, X., Hu, Y., Liu, J., et al. (2024). Zone: Zero-shot instruction-guided local editing. In *CVPR* (pp. 6254–6263).

Lin, C.H., Yumer, E., Wang, O., Shechtman, E., Lucey, S. (2018). St-gan: Spatial transformer generative adversarial networks for image compositing. In *CVPR* (pp. 9455–9464).

Liu, D., Long, C., Zhang, H., Yu, H., Dong, X., & Xiao, C. (2020). Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR* (pp. 8139–8148).

Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., & Cao, Y. (2023). Cones: Concept neurons in diffusion models for customized generation. arXiv preprint arXiv:2303.05125.

Lu, L., Zhang, B., Niu, L. (2023). Dreamcom: Finetuning text-guided inpainting model for image composition. arXiv preprint arXiv:2309.15508.

Lu, S., Liu, Y., Kong, A. W. K. (2023). Tf-icon: Diffusion-based training-free cross-domain image composition. In *ICCV* (pp. 2294–2305).

Mansimov, E., Parisotto, E., Ba, J. L., Salakhutdinov, R. (2015). Generating images from captions with attention. In *ICLR*.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J. Y., & Ermon, S. (2021). Sdedit: Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073.

Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *CVPR* (pp. 6038–6047.

Nguyen, T., Li, Y., Ojha, U., & Lee, Y. J. (2024). Visual instruction inversion: Image editing via image prompting. *NeurIPS, 36*, 9598–9613.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M. (2022). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML* (pp. 16784–16804).

Niu, L., Cao, J., Cong, W., & Zhang, L. (2023). Deep image harmonization with learnable augmentation. In *ICCV* (pp. 7482–7491).

Niu, L., Cong, W., Liu, L., Hong, Y., Zhang, B., Liang, J., & Zhang, L. (2021) Making images real again: A comprehensive survey on deep image.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV* (pp. 2085–2094).

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & 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 *ICML* (pp. 8748–8763).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I. (2021). Zero-shot text-to-image generation. In *ICML* (pp. 8821–8831).

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *ICML* (pp. 1060–1069).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *CVPR* (10684–10695).

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI* (pp. 234–241).

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR* (pp. 22500–22510).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS* (36479–36494).

Stable diffusion. https://huggingface.co/CompVis/stable-diffusion-v-1-4-original (2022). Version 1.4

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114.

Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020). Interpreting the latent space of gans for semantic face editing. In *CVPR* (pp. 9243–9252).

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML* (pp. 2256–2265).

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In *ICLR*.

Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., & Aliaga, D. (2023). Objectstitch: Object compositing with diffusion model. In *CVPR* (pp. 18310–18319).

Tewel, Y., Gal, R., Chechik, G., & Atzmon, Y. (2023). Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH* (pp. 1–11).

Tumanyan, N., Geyer, M., Bagon, S., & Dekel, T. (2023). Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR* (pp. 1921–1930).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *NeurIPS* (pp. 5998–6008).

Wang, T., Zhang, Y., Fan, Y., Wang, J., & Chen, Q. (2022). High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11379–11388).

Wang, X., Girdhar, R., & Gupta, A. (2017). Binge watching: Scaling affordance learning from sitcoms. In *CVPR* (2596–2605).

Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., & Zuo, W. (2023). Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV* (pp. 15943–15953)

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., & Duan, N. (2022). Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV* (pp. 720–736).

Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S. (2023). Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431.

Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR* (pp. 22428–22437).

Xu, N., Price, B., Cohen, S., & Huang, T. (2017). Deep image matting. In *CVPR* (pp. 2970–2979).

Xu, S., Huang, Y., Pan, J., Ma, Z., & Chai, J. (2024). Inversion-free image editing with language-guided diffusion models. In *CVPR* (pp. 9452–9461).

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR* (pp. 1316–1324).

Xue, B., Ran, S., Chen, Q., Jia, R., Zhao, B., & Tang, X. (2022). Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV* (pp. 300–316).

Xue, H., Huang, Z., Sun, Q., Song, L., & Zhang, W. (2023). Freestyle layout-to-image synthesis. In *CVPR* (pp. 14256–14266).

Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., & Wen, F. (2023). Paint by example: Exemplar-based image editing with diffusion models. In *CVPR* (pp. 18381–18391).

Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al. (2022). Scaling autoregressive models for content-rich text-to-image generation. *TMLR*.

Zhang, C., Zhang, C., Zhang, M., & Kweon, I.S. (2023). Text-to-image diffusion model in generative AI: A survey. arXiv preprint arXiv:2303.07909.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV* (pp. 5907–5915).

Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS* (pp. 31428–31449).

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (pp. 586–595).

Zhang, Z., Han, L., Ghosh, A., Metaxas, D. N., Ren, J. (2023). Sine: Single image editing with text-to-image diffusion models. In *CVPR* (pp. 6027–6037).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.