Entity-Level Text-Guided Image Manipulation

Yikai Wang*, Jianan Wang*, Guansong Lu, Hang Xu, Zhenguo Li, Wei Zhang, and Yanwei Fu.

Abstract—Existing text-guided image manipulation methods aim to modify the appearance of the image or to edit a few objects in a virtual or simple scenario, which is far from practical applications. In this work, we study a novel task on text-guided image manipulation on the entity level in the real world (eL-TGIM). The task imposes three basic requirements, (1) to edit the entity consistent with the text descriptions, (2) to preserve the entity-irrelevant regions, and (3) to merge the manipulated entity into the image naturally. To this end, we propose an elegant framework, dubbed as *SeMani*, forming the *Se*mantic *Mani*pulation of real-world images that can not only edit the appearance of entities but also generate new entities corresponding to the text guidance. To solve eL-TGIM, SeMani decomposes the task into two phases: the semantic alignment phase and the image manipulation phase. In the semantic alignment phase, SeMani incorporates a semantic alignment module to locate the entity-relevant region to be manipulated. In the image manipulation phase, SeMani adopts a generative model to synthesize new images conditioned on the entity-irrelevant regions and target text descriptions. We discuss and propose two popular generation processes that can be utilized in SeMani, the discrete auto-regressive generation with transformers and the continuous denoising generation with diffusion models, yielding SeMani-Trans and SeMani-Diff, respectively. We conduct extensive experiments on the real datasets CUB, Oxford, and COCO datasets to verify that SeMani can distinguish the entity-relevant and -irrelevant regions and achieve more precise and flexible manipulation in a zero-shot manner compared with baseline methods. Our codes and models will be released at https://github.com/Yikai-Wang/SeMani.

Index Terms—Image Manipulation, Auto-regressive Generation, Diffusion Models, Semantic Alignment.

1 Introduction

There are various active branches of image manipulation, such as style transfer [5], image translation [6], [7], and Text-Guided Image Manipulation (TGIM), by taking advantage of recent deep generative architectures such as GANs [8], VAEs [9], auto-regressive models [10] and diffusion models [11]. Particularly, the previous TGIM methods either operate some objects by text instructions [12]–[14], such as "adding" and "removing" in a simple toy scene, or manipulating the appearance of objects [15] or the style of the image [16], [17]. In this work, we are interested in a novel challenging task of entity-Level Text-Guided Image Manipulation (eL-TGIM), which is to manipulate the entities on a natural image given the text descriptions, as shown in Fig. 3. eL-TGIM takes as inputs the image to be manipulated, a word prompt to locate the interested entity, and a target text description to manipulate the entity. Basically, eL-TGIM imposes three requirements: (1) to edit the entity consistent with the text descriptions, (2) to preserve the entity-irrelevant regions, and (3) to merge the manipulated entity into the image naturally. Critically, our eL-TGIM is much more difficult than the vanilla TGIM task, as it demands manipulation ability at the fine-grained entity level. Thus, it is nontrivial to directly extend previous methods to the eL-TGIM task, as they can not effectively identify and edit the properties of entities.

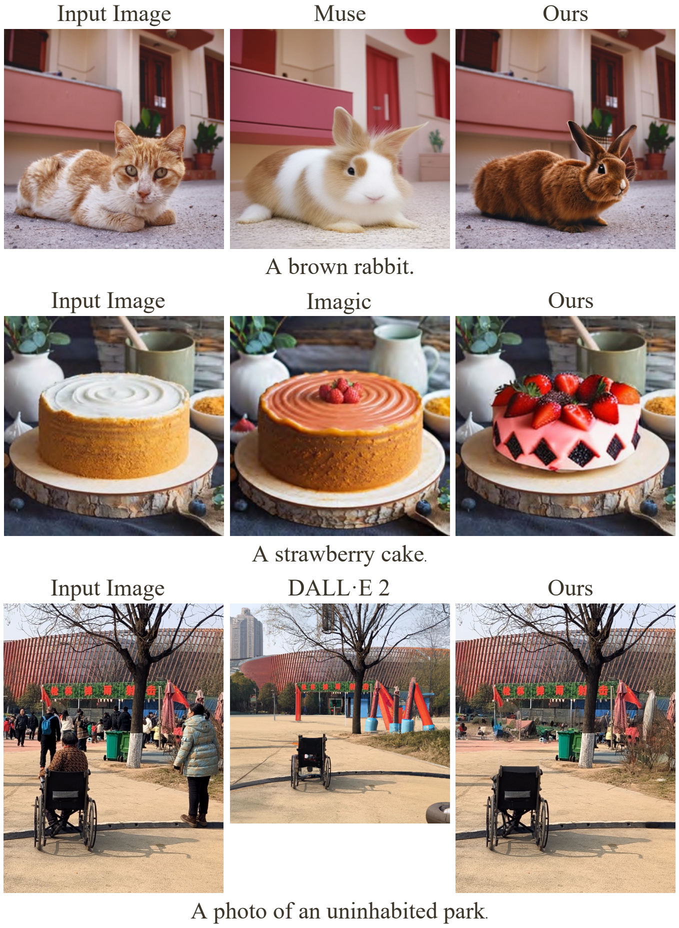


Fig. 1. Comparison between our method and powerful image editors, including transformer-based architectures (Muse [1]), diffusion-based pixel-level model (Imagen [2] based Imagic [3]) and latent-level model (DALLE2 [4]). Analyses are in the second paragraph of Sec. 1.

- *Yikai Wang and Jianan Wang contribute equally.*
- *Yikai Wang, Jianan Wang, and Yanwei Fu are with the School of Data Science, Fudan University. E-mail: {yikaiwang19, jawang19, yanweifu}@fudan.edu.cn*
- *Guansong Lu, Hang Xu, Zhenguo Li, and Wei Zhang are with Huawei Noah's Ark Lab. E-mail: {luguansong, xu.hang, li.zhenguo, wz.zhang}@huawei.com*

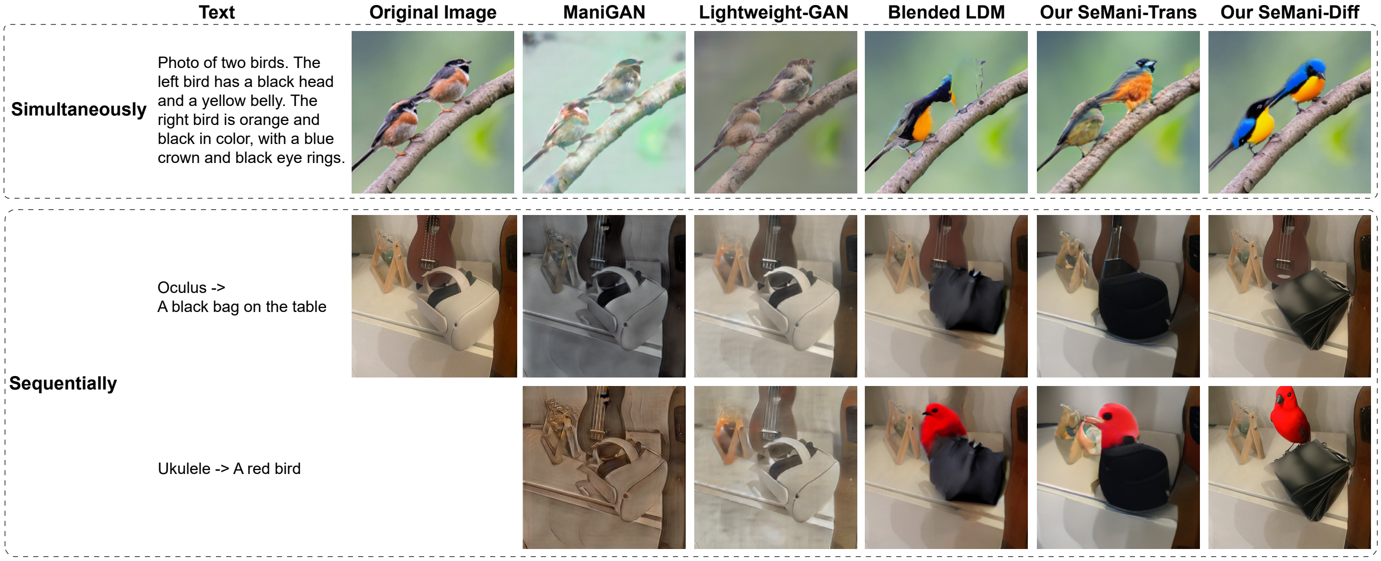

Text
Original Image
ManiGAN
Lightweight-GAN
Blended LDM
Our SeMani-Trans
Our SeMani-Diff
Simultaneously
Photo of two birds. The left bird has a black head and a yellow belly. The right bird is orange and black in color, with a blue crown and black eye rings.
Sequentially
Oculus -> A black bag on the table
Ukulele -> A red bird

Fig. 2. Results of manipulating multiple objects simultaneously or sequentially. Blended LDM uses the masks generated by SeMani as it requires user-provided masks. SeMani can manipulate different objects consistently with different texts, while competitors cannot.

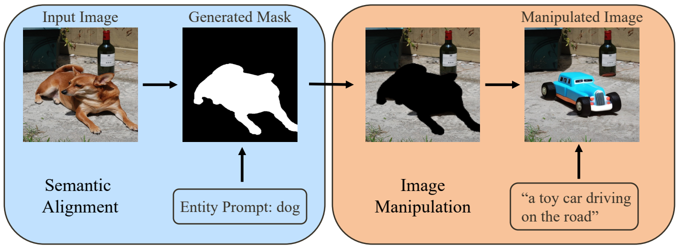


Fig. 3. Pipeline of SeMani for entity-Level Text-Guided Image Manipulation, which is decomposed into the semantic alignment phase and image manipulation phase. In the semantic alignment phase, we focus on the entity-relevant region of the input image given the entity prompt and generate the entity mask to locate the entity in the image. In the image manipulation phase, we generate new images via the target description while preserving the entity-irrelevant regions of the image.

To demonstrate the difficulty of eL-TGIM, we adopt several recent powerful image editors, including transformer-based architectures (Muse [1]), diffusion-based pixel-level model (Imagen [2] based Imagic [3]) and latent-level model (DALLE2 [4]). Results are shown in Fig. 1, where the left column is the input image, the middle column is the manipulation results by other methods, and the right column is our manipulation results. The corresponding text description is at the bottom. Results of Muse and Imagic are reported in their paper, and results of DALLE2 are generated using the official API. Powerful editing algorithms can only satisfy partial requirements of eL-TGIM, which can generate new images that are consistent with text description with the cost of neglecting some essential components for eL-TGIM. For example, Muse and Imagic will change the background (see the door of the first row and the cup of the second row), and DALLE2 requires a user-provided mask to detect the regions to be manipulated and can only generate square images. Our method consistently satisfy the three requirements of eL-TGIM with only text guidance.

Generally, the major obstacle of the TGIM task lies in distinguishing which parts of the image to change or not change. To tackle this problem, existing TGIM methods [18]–[21] propose many different manipulation mechanisms, such as word-level discriminator [19], [21] and text-image affine combination module [20], to differentiate the candidate editing regions from the other image parts. These methods unfortunately are still very limited to be applied to manipulate the entities in nature images. For example, Fig. 2 shows that previous methods can only manipulate the texture/color of an object or require user-provided masks to locate the entity-relevant region, while they fail to perform reasonable entity-level manipulation from text descriptions.

To this end, we propose a novel framework of Semantic Manipulation (SeMani), which decomposes the eL-TGIM task into the semantic alignment phase and image manipulation phase. Imitating the human activities of image editing, we first identify the entity-relevant region corresponding to the entity prompt. The entities should be open vocabulary and not limited to a fixed set of categories. Our target in this phase is to generate the mask of the entity. With this mask, we can focus on the entity and preserve entity-irrelevant regions when manipulating the image. In the second phase, we perform image manipulation with powerful generative models. The target of this phase is to generate new images that are consistent with the target text description and entity-irrelevant regions.

To implement SeMani, we propose two variants that view the image from different perspectives. Specifically, our SeMani-Trans view the image as a discrete token sequence and propose a token-wise semantic alignment module to locate the entity-relevant tokens and perform manipulation on the token sequence in an auto-regressive manner. On the other hand, our SeMani-Diff focuses on the continuous pixel space and directly provides a pixel-level semantic alignment with a denoising generation with diffusion models to perform image manipulation.

SeMani-Trans. Recently, transformer-based models [22]–[24] have been proposed for image synthesis and have shown great expressive power. We thus present a transformer model for generation, by first learning an auto-encoder to down-sample and quantize an image as a sequence of discrete image tokens and then fit the joint distribution of this sequence with a transformer-based auto-regressive model. Furthermore, to successfully identify the entities for editing, our semantic alignment modules include a patch-level segmentation model and a Contrastive Language-Image Pretraining (CLIP [25]) model. The former model helps to locate the entities that existed in the image and the latter model helps to identify the text-relevant image tokens given the textual guidance. Thus our SeMani-Trans generation model can manipulate the image locally and preserve the irrelevant contents to a greater extent, as in Fig. 2. On the other hand, we repurpose the CLIP model as one type of semantic loss to further boost the visual-semantic alignment between the input textual guidance and the manipulated image. Essentially, such semantic loss, proposed in our SeMani-Trans, is complementary to token-wise classification loss, and thus efficiently serves as a pixel-level supervision signal to train our model.

SeMani-Diff provides a continuous alternative to SeMani-Trans and solves several issues that existed in SeMani-Trans.

(1) The auto-regressive generation pipeline limits the available knowledge for the generation model to utilize in the eL-TGIM task. In SeMani-Trans, we encode the image to a token sequence, where the prediction is performed step by step along the sequence. When the entity-relevant region is in the latter part of the sequence, SeMani-Trans could utilize most of the knowledge in the sequence to provide a more precise and consistent manipulation. However, when the entity-relevant region lies in the early part of the sequence, or extremely at the start of the sequence, SeMani-Trans could only utilize very little knowledge to help generation. This inherent drawback of auto-regressive generation limits the generation capacity of SeMani-Trans. To solve this issue, we adopt the recently fast-developed denoising diffusion probabilistic models (DDPMs) [11] to perform generation given the knowledge of the whole image instead of a uni-direction. DDPMs construct the connection between random noises and real images, performing generation via the denoising process. In each step, the knowledge of the whole image can be utilized by DDPMs to perform generation.

(2) The locally calculated similarity is sub-optimal for the semantic alignment module. In SeMani-Trans, the similarity between the visual features of entities and semantics is calculated via an averaged visual token similarity with the semantic token. However, the visual feature of each token considers more local patterns instead of global patterns, and a simple average may not well-extract the relation between tokens of the entity. On the other hand, the global feature for the entire entity is preferable, but the shape of the entity has various types. However, CLIP-like models are trained on images with square shapes. To this end, we strike the balance between local similarity and global similarity and propose to fine-tune the CLIP model with entity image, where entity-irrelevant regions are masked to ensure that only patterns of the entity are extracted. With this modification, the visual features of the entity will include more global and thus semantic information, which benefits the alignment with words.

With these modifications, our SeMani-Diff first adopts the segmentation model to locate several entities in the image, then the entities will be encoded via the fine-tuned CLIP model. The semantic alignment module will select the most-possible entity based on the consistency with the entity prompt in a more global way. Then, SeMani-Diff uses DDPMs to manipulate the entity-relevant region with entity-irrelevant regions and target text descriptions.

We evaluate both SeMani-Trans and SeMani-Diff on multiple datasets including CUB [26], Oxford [27], and COCO [28]. Quantitatively, qualitatively, and user-study comparisons against previous methods demonstrate the superiority of SeMani in all three requirements of eL-TGIM.

Contributions. In summary, our contributions are:

- We introduce a new task, entity-Level Text-Guided Image Manipulation (eL-TGIM) which aims to manipulate entities of the image with only text descriptions.
- To solve eL-TGIM, we propose an elegant SeMani framework, that decomposes the eL-TGIM into the semantic alignment phase and image manipulation phase.
- We propose a transformer-based framework with discrete token-wise semantic alignment and generation, named SeMani-Trans, which can not only manipulate the texture/color of a single object but also manipulate the structure of an object and manipulate multiple objects.
- We further improve the generation process and semantic alignment module by proposing SeMani-Diff, which runs prediction with knowledge of the whole images instead of a uni-direction direction and extracts visual features of entities in a more global way.
- We quantitatively and qualitatively evaluate our method on the CUB, Oxford, and COCO datasets, achieving better results against baseline methods.

Extension. Our conference version of this work was published in [29] as an oral paper. Compared with [29], we have the following extensions.

- We generalize the ideology of [29] and show that the decomposition of semantic alignment and image manipulation is essential for eL-TGIM.
- We analyze several limitations of models in [29], and propose the corresponding improvements for better manipulation capacity.
- Based on the improvements, we propose a new SeMani-Diff framework with a more global semantic alignment module and a better generation pipeline to utilize knowledge of unmasked regions.
- Our proposed SeMani-Diff shows the superior quantitative, qualitative, and human evaluation performance of the eL-TGIM task on CUB, Oxford, and COCO datasets.

2 Related Work

Text-to-image generation focuses on generating images to visualize what texts describe. There are many good GAN-based models [30]–[33]. Li *et al.* [34] further introduce a word-level discriminator network to provide the generator network with fine-grained feedback. Besides GANs, some works also explore applying transformer-based networks for text-to-image generation [35]–[37]. Recently, denoising diffusion probabilistic models (DDPMs) [11] is fast-developed for text-to-image generation tasks. DDPMs bridge the connection between the distribution of real images and the random Gaussian distribution, such that one can start with random Gaussian noise and iteratively denoise it to generate the image. Many variants of DDPMs, including GLIDE [38], Cascaded Diffusion [39], Imagen [40], DALLE2 [4],LDM [41], to name a few, have achieved state-of-the-art performances on text-to-image generation task. In contrast, rather than generating images according to texts, we focus on entity-level manipulating images given texts.

Text-guided image manipulation has attracted extensive attention as it enables the users to flexibly edit an image with natural language [12]–[21], [42], [43] . Particularly, Li *et al.* [20] introduces a multi-stage network with a novel text-image combination module to generate high-quality images. Li *et al.* [21] propose a new word-level discriminator along with explicit word-level supervisory labels to provide the generator with detailed training feedback related to each word, achieving a lightweight and efficient generator network. Recently, due to the good synthesizing capability of StyleGAN, researchers devote to image manipulation by pre-trained StyleGAN models [42], [44]. Patashnik *et al.* [44] adopt the CLIP model for semantic alignment between text and image, and propose mapping the text prompts to input-agnostic directions in StyleGAN's style space, achieving interactive text-driven image manipulation. Text2LIVE [45] introduces an edit layer to composite the generation results with the image to preserve the information. The edit layer is directly predicted by a U-Net model.

Diffusion models also show promising performance in text-guided image manipulation. Blended Diffusion [46] adopts the user-provided mask and target text description for manipulation. The target text is utilized via the gradient of CLIP loss to the diffusion outputs to guide the generation of diffusion models. Some works rely on the textual inversion technique [47] to represent the image via the text embedding and then achieve manipulation by an interpolation between target descriptions and inverse text embedding [3]. Others [48] rely on the inversion of the image that learns a noise that can be transformed into the input image via diffusion models, and then achieve manipulation with this noise. Prompt-to-Prompt [49] controls the cross-attention layer of diffusion models and injects it with weights that correspond to target text to achieve manipulation. Prompt-to-Prompt relies on the inversion of an image or improved textual inversion [50] to implement manipulation on real images. On the contrary, our SeMani can directly manipulate real images.

Semantic image synthesis aims to generate a photo-realistic image from a semantic label. Isola *et al.* [6] propose a unified framework based on conditional GANs [51] for various image-to-image translation tasks, including $Semantic\ labels \leftrightarrow photo$, $Edges \rightarrow Photo$, $Day \rightarrow Night$, and so on. Chen and Koltun [52] adopt a modified perceptual loss to synthesize high-resolution images to tackle the instability of adversarial training. Wang *et al.* [53] propose a novel adversarial loss and a new multi-scale generator and discriminator architectures for generating high-resolution images with fine details and realistic textures. Park *et al.* [54] propose a spatially-adaptive normalization layer to modulate the activation using input semantic layouts and effectively propagate the semantic information throughout the network. Such works enable users to synthesize images with a finite number of semantic concepts associated with the semantic labels, while our method focuses on manipulating the input images according to the input texts, which is more flexible and with an unlimited number of semantic concepts.

Vision and language representation learning models [25], [55]–[62] learn cross-modal representations for various down-stream tasks, including image-text retrieval, image captioning, visual grounding, and so on. They adopt the network architecture of ResNets [63] and/or Transformers [10], [64], and mainly use two kinds of learning tasks for pre-training: cross-modal contrastive learning and masked language modeling. Specifically, the recent CLIP [25] model is trained on a large-scale dataset and shows superior performance on zero-shot tasks. We repurpose the CLIP model to help train our framework for eL-TGIM.

3 Methodology

3.1 Overview

Problem formulation. Entity-Level Text-Guided Image Manipulation (eL-TGIM) aims to perform image editing in the real world, with a focus on the entity level manipulation guided by text. eL-TGIM imposes three basic requirements:

1) To edit the entity consistent with the text descriptions;
2) To preserve the entity-irrelevant regions;
3) To merge the manipulated entity into the image naturally.

Formally, eL-TGIM takes as input the entity prompt word $\boldsymbol{e}$, the target text description $\boldsymbol{T}$, and the original image $\boldsymbol{X}$ from the real world, i.e., $\boldsymbol{X} \sim p(\boldsymbol{X})$ where $p(\boldsymbol{X})$ indicates the distribution of real-world images, The target is to generate a new image $\tilde{\boldsymbol{X}}$ that follows the above three requirements.

SeMani. In this paper, we propose SeMani, forming the Semantic Manipulation for eL-TGIM. Inspired by the human activities for image editing, we decompose the eL-TGIM into the semantic alignment and image manipulation phases.

In the semantic alignment phase, we aim to generate a mask $\boldsymbol{M} \in \{0,1\}^{\boldsymbol{X}}$ for the original image $\boldsymbol{X}$, such that the resulting $\boldsymbol{M} \odot \boldsymbol{X}$ only contains the interested entity and all other regions are masked, where $\odot$ indicates element-wise multiplication. To achieve this, SeMani first adopts a segmentation model to generate a series of masks $\{\boldsymbol{M}_i\}$ such that each mask indicates an entity existed in the image

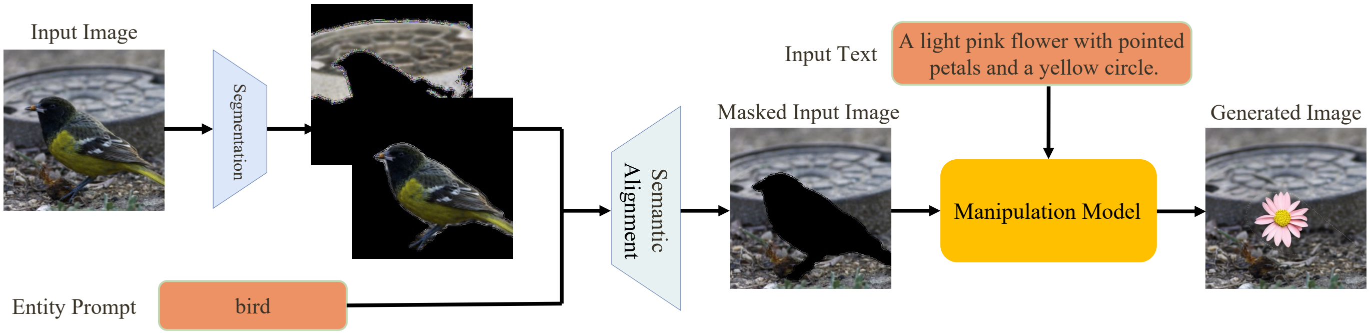


Fig. 4. Framework of our SeMani. We achieve eL-TGIM by first using a segmentation model to distinguish entities that existed in the image, and then adopting a semantic alignment model to identify the prompt-relevant entity. Then the masked image is provided to condition the manipulation model with target text descriptions to perform semantic manipulation.

$\boldsymbol{X}$. Then, given the entity prompt $\boldsymbol{e}$, we propose a semantic alignment module to identify the most possible entity via

$$\boldsymbol{M}_e := \mathrm{argmax}_{\boldsymbol{M}_i} \ \mathrm{Sim}(\boldsymbol{M}_i \odot \boldsymbol{X}, \boldsymbol{e}), \tag{1}$$

where the similarity function $\mathrm{Sim}(\cdot)$ is defined with a modified CLIP model specifically for architectures we used, which will be introduced later. Note that the entity prompt here should be *open vocabulary* and not limited to a fixed set of categories for better practical utilization.

In the image manipulation phase, SeMani takes as input the masked image $\boldsymbol{M}_e \odot \boldsymbol{X}$ and the target text description $\boldsymbol{T}$ to generate new image $\tilde{\boldsymbol{X}}$ with a deep manipulation model.

With this framework, as illustrated in Fig. 4, SeMani achieves eL-TGIM for input images with text guidance. In SeMani, the three requirements of eL-TGIM become:

$$\begin{aligned} & \max \mathrm{Sim}(\tilde{\boldsymbol{X}}, \boldsymbol{T}) \\ & s.t.\ \boldsymbol{M}_{\boldsymbol{e}} \odot \boldsymbol{X} \approx \boldsymbol{e}, \\ & \quad (1 - \boldsymbol{M}_{\boldsymbol{e}}) \odot \tilde{\boldsymbol{X}} = (1 - \boldsymbol{M}_{\boldsymbol{e}}) \odot \boldsymbol{X}, \\ & \quad \tilde{\boldsymbol{X}} \sim p(\boldsymbol{X}), \end{aligned} \tag{2}$$

where $\approx$ indicates the consistency between entity prompt and visual object, $=$ indicates pixel-level equivalent, and $\tilde{\boldsymbol{X}} \sim p(\boldsymbol{X})$ indicates $\tilde{\boldsymbol{X}}$ follows the distribution of the real image, thus satisfying the third requirement of eL-TGIM.

Implementing SeMani. To implement SeMani, we resort to the two popular perspectives of viewing images, the discrete and the continuous perspectives.

The discrete perspective of viewing images is inspired by the recent development of natural language processing, mostly BERT [65] and GPT [66]. Researchers want to introduce the success of NLP models to the vision field by transforming the image into a visual token sequence. Then one can treat the image as a visual "sentence" and perform NLP techniques on the visual sentence to do vision tasks. In this paper, we follow this direction and propose a discrete variant of SeMani, SeMani-Trans, to achieve eL-TGIM discretely. Specifically, we train an auto-encoder to learn a discrete codebook such that any image can be transferred into a discrete token sequence with this codebook vocabulary. For the semantic alignment module, a token-level CLIP model is required to calculate token similarity. For image manipulation, an auto-regressive prediction for generation is implemented to achieve generation on the token space.

The continuous perspective of viewing images is the most straightforward way to model images. Models that follow this perspective directly run on the pixel-level of images and can perform more precise control for semantic alignment and image manipulation. For the semantic alignment, as the comparison is between masked image $\boldsymbol{M}_i \odot \boldsymbol{X}$ and entity prompt $\boldsymbol{e}$, a fine-tuning of CLIP model is needed to better extract the visual features of images with masks. For image manipulation, we resort to the recently fast-developed denoising diffusion probabilistic models [11] (DDPMs) to perform generation by denoising from the pixel-level random noise iteratively. These modules form our continuous variant of SeMani, SeMani-Diff.

In the following, we introduce the architectures and training details of SeMani-Trans and SeMani-Diff, respectively.

3.2 SeMani-Trans

3.2.1 Architectures

The autoencoder consists of three components, a convolutional encoder E, a convolutional decoder G and a codebook $\boldsymbol{Z} \in \mathbb{R}^{K \times n_z}$, containing K n_z-dimensional latent variables. All of them are learnable. Given an image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times 3}$, E encodes the image into a two-dimensional latent feature map $\boldsymbol{Q} \in \mathbb{R}^{h \times w \times n_z}$. The codebook is utilized to quantize the latent feature map by replacing each pixel embedding with its closest latent variables within the codebook element-wisely as follows:

$$\hat{\boldsymbol{Q}}_{ij} = \arg\min_{\boldsymbol{z}_k} \parallel \boldsymbol{Q}_{ij} - \boldsymbol{z}_k \parallel^2 . \tag{3}$$

For reconstruction, the decoder G takes the quantized latent feature map $\hat{\boldsymbol{Q}}$ as input and returns an generated image $\hat{\boldsymbol{X}}$ close to the original image, i.e., $\hat{\boldsymbol{X}} \approx \boldsymbol{X}$.

Auto-regressive generation. For image generation, the quantized feature map $\hat{\boldsymbol{Q}}$ can be modeled as a sequence of discrete tokens, denoted as a sequence of discrete token indices $\boldsymbol{I} \in \{0, \ldots, K-1\}^{h \times w}$. Each token roughly corresponds to an image patch of the size $\frac{H}{h} \times \frac{W}{w}$. Thus, the prediction of a token sequence is equivalent to synthesizing an image. In practice, we refer to uni-directional Transformer [10] to predict the image sequence autoregressively as follows:

$$P(\boldsymbol{I}_{\leq i}|\boldsymbol{T}) = \prod_{j}^{i} P(\boldsymbol{I}_j|\boldsymbol{I}_{<j}, \boldsymbol{T}), \tag{4}$$

where $\boldsymbol{T}$ is the text tokens of the caption paired with image $\boldsymbol{X}$.

To introduce positional information of the two modalities in Transformer, we learn two sets of positional embeddings. One is axial positional embeddings [67] for the visual sequence from a spatial grid. The other is sequence embeddings as BERT [65] for text sequence.

3.2.2 Training with Language and Vision Guidance

Main task. One consequent training idea is masked sequence modeling by optimizing the loss for the paired text and image tokens. However, unlike most existing vision-and-language models [68]–[70] taking detected regions as an image sequence, our model accepts patch sequence, which will be an inexact alignment with text. Moreover, fine-grained correspondences of image patches and attribute tokens are difficult to be aligned. For example, aligning "a red crown" and "a red belly" within the detected bird needs to precisely recognize not only the color but also the body parts. To avoid noisy training signals, we do not adopt masked sequence modeling. Instead, our auto-regressive task minimizes cross-entropy losses for the reconstruction of text and image tokens, respectively [35],

$$\mathcal{L}_{txt} = -\mathbb{E}_{\boldsymbol{T}_i} \log P(\boldsymbol{T}_i|\boldsymbol{T}_{<i}), \tag{5}$$

$$\mathcal{L}_{img} = -\mathbb{E}_{\boldsymbol{I}_i} \log P(\boldsymbol{I}_i|\boldsymbol{I}_{<i}, \boldsymbol{T}). \tag{6}$$

Language guidance. The transformer model determines the basic image tokens at the top level, and the autoencoder model holds the convolutional decoder complementing the texture in detail at the bottom level. Training these two models separately implies splitting the generation stream stiffly. To this end, we propose a semantic loss for the token prediction such that the model not only considers the downstream decoding but also improves the ability to capture the relation between text and image.

The CLIP [44] is a vision-and-language representation learning model, trained with 400 million image-text pairs, and has shown excellent visual-semantic alignment capability by achieving superb performance on the task of zero-shot image classification. It is optimized by a symmetric cross-entropy loss over the cosine similarities of a batch of image and text embeddings. We leverage CLIP to guide our token prediction, through

$$\mathcal{L}_{CLIP} = 1 - D(G(\hat{\boldsymbol{I}}), \boldsymbol{T}), \tag{7}$$

where D is the cosine similarity between the CLIP embeddings of its two arguments. We adopt the straight-through estimator [71] for the gradient back-propagation.

Vision guidance. With the text descriptions, our model can replace an entity with other specific entities. For only editing the appearance of an entity, we need to provide the model with the prior information on the original entity's shape. Specifically, we convert the image to grayscale and append the quantized grayscale image tokens $\boldsymbol{V} \in \{0, \ldots, K-1\}^{h\times w}$ to the text sequence as another condition for the tokens to be manipulated. The grayscale image token sequence $\boldsymbol{V}$ shares the positional embeddings with the image token sequence $\boldsymbol{I}$, for the same modality and spatial positions. For the identities of vision guidance and input text token sequence, we append two special separation tokens [BOV] and [BOT] to the beginning of them respectively. We apply the cross entropy loss on the vision guidance tokens as well,

$$\mathcal{L}_{gray} = -\mathbb{E}_{\boldsymbol{V}_i} \log P(\boldsymbol{V}_i|\boldsymbol{V}_{<i}). \tag{8}$$

We randomly select 50% samples to train with vision guidance. The total loss to train the transformer is a combination of the four losses, which can be divided into two parts, including auto-regressive and semantic losses as

$$\mathcal{L}_{ar} = \lambda_1\mathcal{L}_{img} + \lambda_2\mathcal{L}_{gray} + \lambda_3\mathcal{L}_{txt}, \tag{9}$$

$$\mathcal{L}_{total} = \mathcal{L}_{ar} + \lambda_4\mathcal{L}_{CLIP}, \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are the balancing coefficients.

3.2.3 Inference with Entity Guidance

We design a semantic alignment module to locate the image patches to be manipulated by input text automatically in the inference phase. The semantic alignment module is a two-step module, (1) to find the tokens of every entity and (2) to select the text-relevant entities to be manipulated, where each step is based on a strong existing model.

In the first step, we refer to entity segmentation [72] to recognize each entity on the original image $\boldsymbol{X}$, as Fig. 4 shows. The segmentation is implemented on the original image size, and we use the bilinear interpolation to resize the binary mask map of each entity to the same size of latent feature map $\boldsymbol{Q}$. The pixels whose values are larger than 0 represent that the tokens at the same position belong to the entity. In our preliminary experiments, we compare the bilinear interpolation with max-pooling for finding the entity tokens. The max-pooling dilates the tokens for the bilinear interpolation, however, due to the stack of convolutions in the first stage, the receptive field of the tokens by max-pooling is beyond the entity area and overlaps with other entities. Thus, we use bilinear interpolation to map the segment mask and token mask for a more precise alignment.

In the second step, we set a text prompt word to select the relevant entities. We leverage the FILIP [73], a CLIP-style model optimized by token level similarity, to calculate the similarities between image token and text token. For example, as Fig. 4 shows, we set "bird" as a prompt word to search the bird entities in the image, and then we average the similarities between tokens of each entity and the prompt word "bird". The entities whose similarities are higher than θ are text-related entities.

3.3 SeMani-Diff

In this section, we analyze several limitations of SeMani-Trans and propose new models to solve these issues.

From uni-direction to multi-directions. In SeMani-Trans, we train the transformer model to autoregressively generate image tokens in the uni-direction. However, this is sub-optimal in image manipulation tasks. Specifically, when the entity-relevant region is in the later part of the sequence, the model will well-generate the tokens with the help of an informative sub-sequence. But when the entity-relevant region is in the former part of the sequence, the generation model cannot utilize the information of the later unmasked tokens. In this scenario, the manipulation capacity is limited.

To this end, we propose to use the multi-direction generation process to fully utilize the information of unmasked regions. Specifically, we adopt the recently fast-developed denoising diffusion probabilistic models (DDPMs). Unlikely autoregressive models, DDPMs directly encode the information of the entire image.

From local semantic similarity to global. In SeMani-Trans, we adopt FILIP to calculate the similarity of each visual token to the prompt word and then average them as the similarity between the entity and the prompt word. However, the token-level feature extracts more local information, and a simple average may not well-extract the global semantics of the entity. The original CLIP model is trained with global semantics, but it cannot well-extract the feature of an entity. This is due to that the shape of the entity has various types, but the CLIP is trained on the image of a square shape. If we directly use CLIP to extract features of entity images with masks, the resulting embedding is sub-optimal. Thus, a fine-tuning of CLIP on the masked image is required for better similarity calculation.

With the above two improvements, we now can design a more powerful framework for the entity-level text-guided image manipulation task, dubbed as SeMani-Diff. SeMani-Diff takes as inputs an original image, a prompt of the entity, and a target text description. We first segment the image into several entities and then use the fine-tuned CLIP model to calculate the similarity between the entities and the prompt. The most similar entity will be identified as the target entity to manipulate and forming the masked input image as the visual condition for the generation model. Then, conditioned on the target text description, we adopt the diffusion models to manipulate the image.

Compared with SeMani-Trans, SeMani-Diff enjoys better semantic alignment thanks to the globally extracted visual features. Further, the DDPMs enjoy superior generation capacity compared with the autoregressive prediction pipeline due to the better utilization of unmasked regions of the image. Note that the overall ideology is shared between SeMani-Trans and SeMani-Diff. We argue that the pipeline of first locating the entity-relevant region via the semantic alignment module and then implementing manipulation via local editing is crucial for eL-TGIM. The techniques we use for each module are of course not limited to specific architectures. In the following, we introduce the architectures and training of SeMani-Diff in detail.

3.3.1 Architectures

Formally, DDPMs construct two random processes to connect the distribution of real image $\boldsymbol{X}_0 \sim p(\boldsymbol{X}_0)$ with diagonal Gaussian $\boldsymbol{X}_T \sim \mathcal{N}(0, \sigma\mathcal{I})$. In the forward/diffusion process from $\boldsymbol{X}_0$ to $\boldsymbol{X}_T$, DDPMs gradually add random Gaussian noises, forming a Markov chain as

$$q\left(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}\right) \coloneqq \mathcal{N}\left(\boldsymbol{X}_t; \sqrt{\alpha_t}\boldsymbol{X}_{t-1}, (1-\alpha_t)\mathcal{I}\right). \tag{11}$$

When the noise added in each step is small enough, and the process runs a long time enough (with a large T). The final state $\boldsymbol{X}_T$ can be well-approximated by $\mathcal{N}(0, \sigma\mathcal{I})$ and the posterior distribution $p(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t)$ can also be approximated by a Gaussian distribution. DDPMs adopt the neural network to learn the posterior via

$$p_\theta\left(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t\right) \coloneqq \mathcal{N}\left(\mu_\theta\left(\boldsymbol{X}_t\right), \Sigma_\theta\left(\boldsymbol{X}_t\right)\right). \tag{12}$$

With this denoising step, we can generate a real image from the random Gaussian noise iteratively. In each step, unlike autoregressive predictions, DDPMs directly estimate the less-noised image $\boldsymbol{X}_{t-1}$ simultaneously. In this manner, it can well utilize the information of the overall image.

We leverage the latent diffusion models [41] (LDM) that balance the high-resolution image generation and computation cost by introducing DDPMs in the latent space. Similar to SeMani-Trans, LDM also adopts an auto-encoder with a different target of compressing the image instead of getting discrete token indices. Thus the dimension of the resulting latent variables is 3 instead of 1024 in transformers. Then a time-conditional UNet [74] is performed on the latent variables to learn denoising.

To incorporate the masked image and the target text description as the conditions for generation, we concatenate the masked image with the noise as the input to the diffusion model and adopt cross-attention layer [10] to inject text description. Then the posterior now becomes:

$$\begin{aligned} &p_\theta\left(\boldsymbol{X}_{t-1} \mid \boldsymbol{X}_t, \boldsymbol{T}, \boldsymbol{M}_{\boldsymbol{e}}\right) \\ \coloneqq &\mathcal{N}\left(\mu_\theta\left(\boldsymbol{X}_t, \boldsymbol{T}, \boldsymbol{M}_{\boldsymbol{e}}\right), \Sigma_\theta\left(\boldsymbol{X}_t, \boldsymbol{T}, \boldsymbol{M}_{\boldsymbol{e}}\right)\right). \end{aligned} \tag{13}$$

3.3.2 Training with Masked Image and Language Guidance

The training of DDPMs can be formulated as first randomly sampling an image from the training set and randomly generating a Gaussian noise by randomly selecting a noise level t. As we can derive $\mu_\theta, \Sigma_\theta$ from the noise ε [75], we can directly train the network to learn the error estimation task. Specifically, conditioned on the masked image and text description, the network can be trained to estimate the noise added to the image via the L2 loss function as

$$\mathcal{L} \coloneqq \mathbb{E}_{t\sim[1,T], \boldsymbol{X}_0 \sim q(\boldsymbol{X}_0), \varepsilon \sim \mathcal{N}(0,\mathcal{I})}\left[\left\|\varepsilon - \varepsilon_\theta\left(\boldsymbol{X}_t, \boldsymbol{T}, \boldsymbol{M}_{\boldsymbol{e}}, t\right)\right\|^2\right]. \tag{14}$$

Classifier-free guidance. Ho and Salimans [76] introduce classifier-free guidance for better generation of diffusion models. This technique is also utilized in our DDPMs. Specifically, when training with Eq. (14), we randomly replace the text guidance and mask guidance with empty guidance ϕ, performing the unconditional generation. When inference, the final estimation of ε is

$$\hat{\varepsilon} = \varepsilon_\theta(\boldsymbol{X}_t \mid \phi) + s \cdot (\varepsilon_\theta(\boldsymbol{X}_t \mid \boldsymbol{T}, \boldsymbol{M}_{\boldsymbol{e}}) - \varepsilon_\theta(\boldsymbol{X}_t \mid \phi)), \quad (15)$$

where s controls the scale for conditioning.

3.3.3 Inference with Global Entity Guidance

We strike the balance between local and global semantic similarities via OVSeg [77]. Specifically, we first utilize a segmentation model to distinguish entities in the image. Then for each entity, we pad the other regions with 0 values, forming the same shape of the squared image as the input image. Then this entity image with masks is input to the CLIP model to extract more global visual features. However, the original CLIP model is trained on natural images without masks. Thus a fine-tuning stage is needed for CLIP to adapt to images with masks.

To fine-tune the CLIP model, OVSeg collects masked image and entity name pairs from the existing image-caption dataset. Then the pre-trained CLIP model is fine-tuned on this paired dataset to adapt to the masked images. Specifically, instead of fine-tuning the CLIP model, OVSeg introduces a mask prompt $\boldsymbol{P}$ such that the input of the CLIP model becomes $\boldsymbol{X} \odot \boldsymbol{M} + \boldsymbol{P} \odot (1 - \boldsymbol{M})$ instead of $\boldsymbol{X} \odot \boldsymbol{M}$, then $\boldsymbol{P}$ is learned to better fit the CLIP model and the parameters of CLIP are frozen. This mask prompt tuning [78] technique is beneficial as the training of mask prompt is easier and preferable when we only have a small training set, compared with fine-tuning the CLIP model.

With this newly fine-tuned CLIP model, the visual feature of the entity can be better extracted, and thus the semantic similarity is more global and semantic than the original local averaged similarity. The other parts of the semantic alignment module are the same as in SeMani-Trans.

4 EXPERIMENTS

4.1 Experimental Setup

Competitors. As there are only a few methods that can be directly adapted to the task of eL-TGIM, we compare the most related methods that can be used for eL-TGIM without significant modifications. Specifically, we compare SeMani-Trans and SeMani-Diff against ManiGAN [20], Lightweight-GAN [21], and Blended LDM [79]. Note that Blended LDM works for the user-provided mask. In our experiments, we use the mask generated by our semantic alignment model as the input of Blended LDM. Thus the comparison of Blended LDM and our methods is mostly on the image manipulation phase. Results of competitors are reproduced using the code/model released by the authors.

Datasets. Following common practice, we conduct experiments on three public datasets, including CUB [26], Oxford [27], and the more complicated COCO [28] datasets. The CUB and Oxford are two datasets about birds and flowers respectively. CUB contains 8855 training images and 2933 testing images while Oxford has 7034 training images and 1155 testing images, in which each image has 10 captions. There are at least 80 categories of objects with different shape structures and appearances on COCO images, forming the 80k training images and 40k testing images. Thus, COCO is a more complicated dataset than CUB and Oxford, not only in the understanding of the correspondence between the image and text but also in image manipulation on the entity level. We preprocess these datasets as in [31], [32].

Quantitative metrics. To evaluate the quality of manipulated images, we use the Inception Score (IS) [80] as the quantitative evaluation metric. To evaluate the visual-semantic alignment between the text descriptions and manipulated images, we calculate the cosine similarity between their embeddings extracted with CLIP text/image encoders, called CLIP-score. Besides, we conduct an image-to-text retrieval experiment and report Recall@N for quantitative comparison. In the image-to-text retrieval, for each manipulated image, the text candidates consist of the input text, which serves as the positive sample, and 99 randomly sampled descriptions as negative samples. Such 100 text candidates are sorted in descending order according to their cosine similarity with the manipulated image. Recall@N calculates the percentage of images, whose positive sample occurs within the top-N candidates. As we use the ViT-B/32 CLIP model during training SeMani-Trans, for a fair comparison, we refer to the ResNet50 CLIP model to compute the CLIP-score. Additionally, following [19], to compare the quality of the content preservation, we compute the L2 reconstruction error by forwarding images with positive texts.

The higher the IS, the higher quality of the manipulated images. Higher CLIP-score and R@N indicate better visual-semantic alignment between the input texts and the manipulated images. The lower the L2 error, the higher content preservation quality.

Hyper-parameters of SeMani-Trans. The model at the first stage inherits from the VQGAN [24] pre-trained on ImageNet, where the codebook size is 1024, the image size is 256×256, and the latent feature map size is 16×16. In the second stage, our transformer has 24 layers, 8 heads with 64 dimensionalities for each head. We replace the traditional Feed-Forward Network (FFN) with a GEGLU [81] variant, which adds a Gated Linear Units (GLU) [82] with GELU [83] activation to the first hidden layer of FFN. We use Byte-Pair Encoding [84] to tokenize the text, with vocabulary size 49408. We limit the text length to 128 and learn a padding token for each position as DALL·E. Our transformer has 152M parameters, a little larger than BERT-Base 110M. The hyper-parameters of autoregressive loss $\lambda_1, \lambda_2, \lambda_3$ are set to $7/9, 1/9, 1/9$ and λ_4 of language guidance loss is 5 for all the datasets. The CLIP model for the semantic loss is ViT-B/32. For the semantic alignment module, we use the entity segmentation model based on Swin-L-W7 and the FILIP-large [73] model for similarity computation. The similarity threshold θ is 0.163. For a good initialization of the transformer, we pre-train our transformer on CC12M [85] without language and vision guidance. We use AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.96$ to train 12 epochs with batch size 112. The learning rate linearly ramps up to $6 \times e^{-4}$ for the first 5k iterations and is halved whenever training loss does not decrease for 50000 iterations. With the same

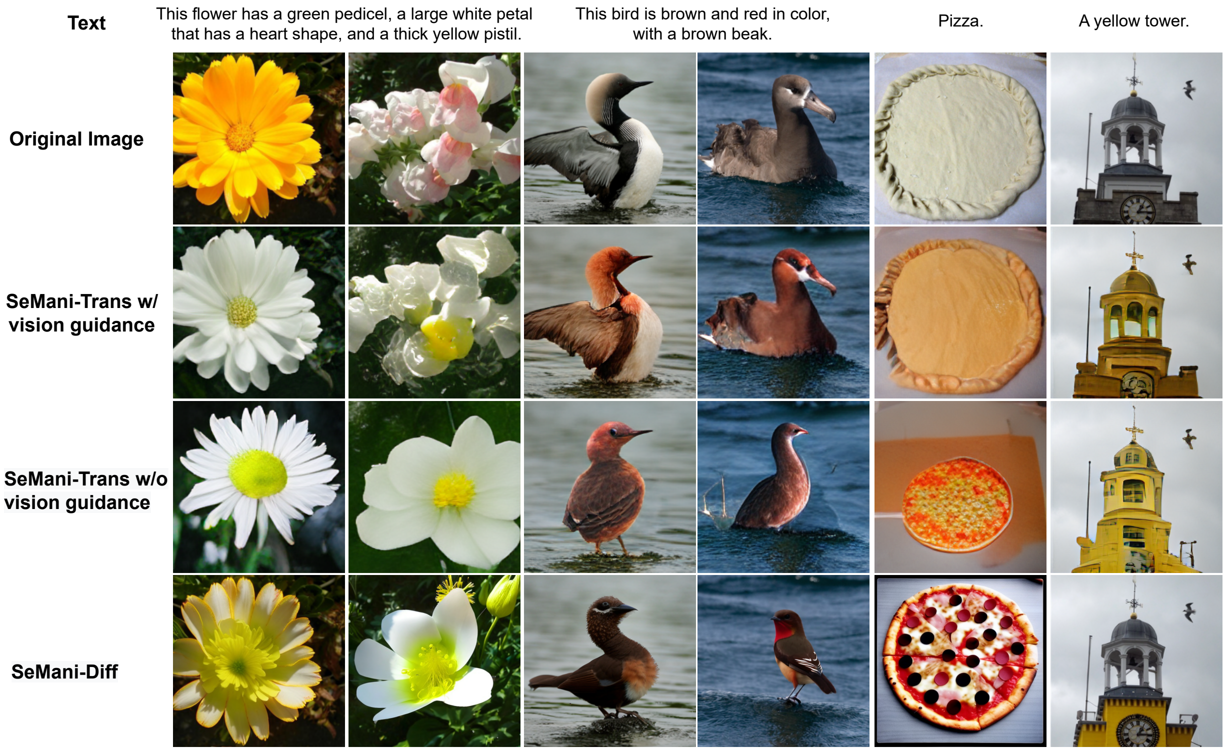

Text
This flower has a green pedicel, a large white petal that has a heart shape, and a thick yellow pistil.
This bird is brown and red in color, with a brown beak.
Pizza.
A yellow tower.
Original Image
SeMani-Trans w/ vision guidance
SeMani-Trans w/o vision guidance
SeMani-Diff

Fig. 5. Manipulation results with (w/) and without (w/o) vision guidance. SeMani-Trans preserves the structure of the entity when w/ vision guidance. SeMani-Trans and SeMani-Diff flexibly perform manipulation according to the text w/o vision guidance. The prompt word for the "Pizza." is "dough".

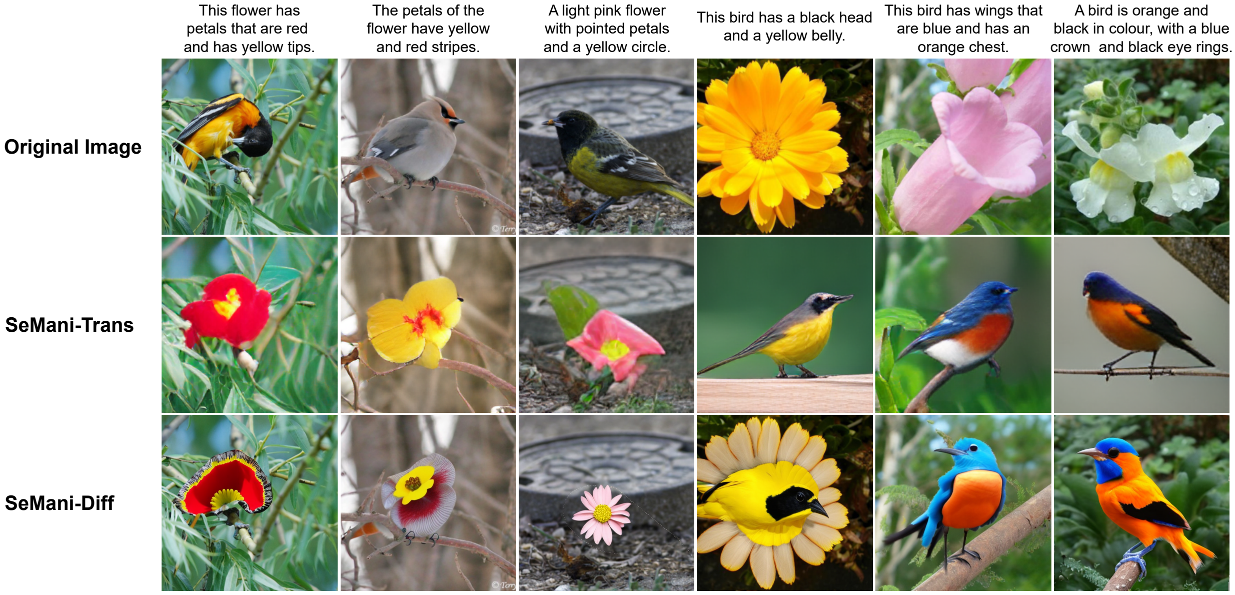

This flower has petals that are red and has yellow tips.
The petals of the flower have yellow and red stripes.
A light pink flower with pointed petals and a yellow circle.
This bird has a black head and a yellow belly.
This bird has wings that are blue and has an orange chest.
A bird is orange and black in colour, with a blue crown and black eye rings.
Original Image
SeMani-Trans
SeMani-Diff

Fig. 6. Manipulation results from bird to flower and flower to bird with our proposed SeMani.

TABLE 1
Quantitative comparison between ManiGAN [20], Lightweight-GAN [21], Blended LDM [79], and our SeMani-Trans and SeMani-Diff. IS: Inception Score. CLIP-score: averaged cosine similarity with CLIP embeddings. R@10: recall within the top 10 candidates. L2-error: L2 reconstruction error. Higher IS, CLIP-score, and R@10 indicate better performance, while lower L2-error is better. The best results are in bold.

Model	CUB				Oxford				COCO			
	IS	CLIP-score	R@10	L2-error	IS	CLIP-score	R@10	L2-error	IS	CLIP-score	R@10	L2-error
ManiGAN	4.19 ± 0.04	21.30	10.49	0.05	4.37 ± 0.11	21.59	14.21	0.02	22.65 ± 0.40	11.91	14.50	0.03
Lightweight-GAN	4.66 ± 0.06	18.88	10.00	0.13	4.35 ± 0.09	17.55	11.58	0.12	24.80 ± 0.94	**13.65**	14.49	0.03
Blended LDM	**5.95 ± 0.08**	23.34	38.17	0.02	4.21 ± 0.09	22.11	21.19	0.01	27.84 ± 0.63	13.02	**28.64**	0.01
SeMani-Trans	5.02 ± 0.11	23.56	34.82	**0.01**	**4.50 ± 0.06**	**23.34**	**36.49**	0.03	21.45 ± 0.41	13.10	21.32	0.02
SeMani-Diff	5.13 ± 0.07	**24.03**	**45.51**	0.02	4.30 ± 0.08	22.05	17.37	**0.01**	**32.98 ± 0.61**	12.28	24.73	**0.01**

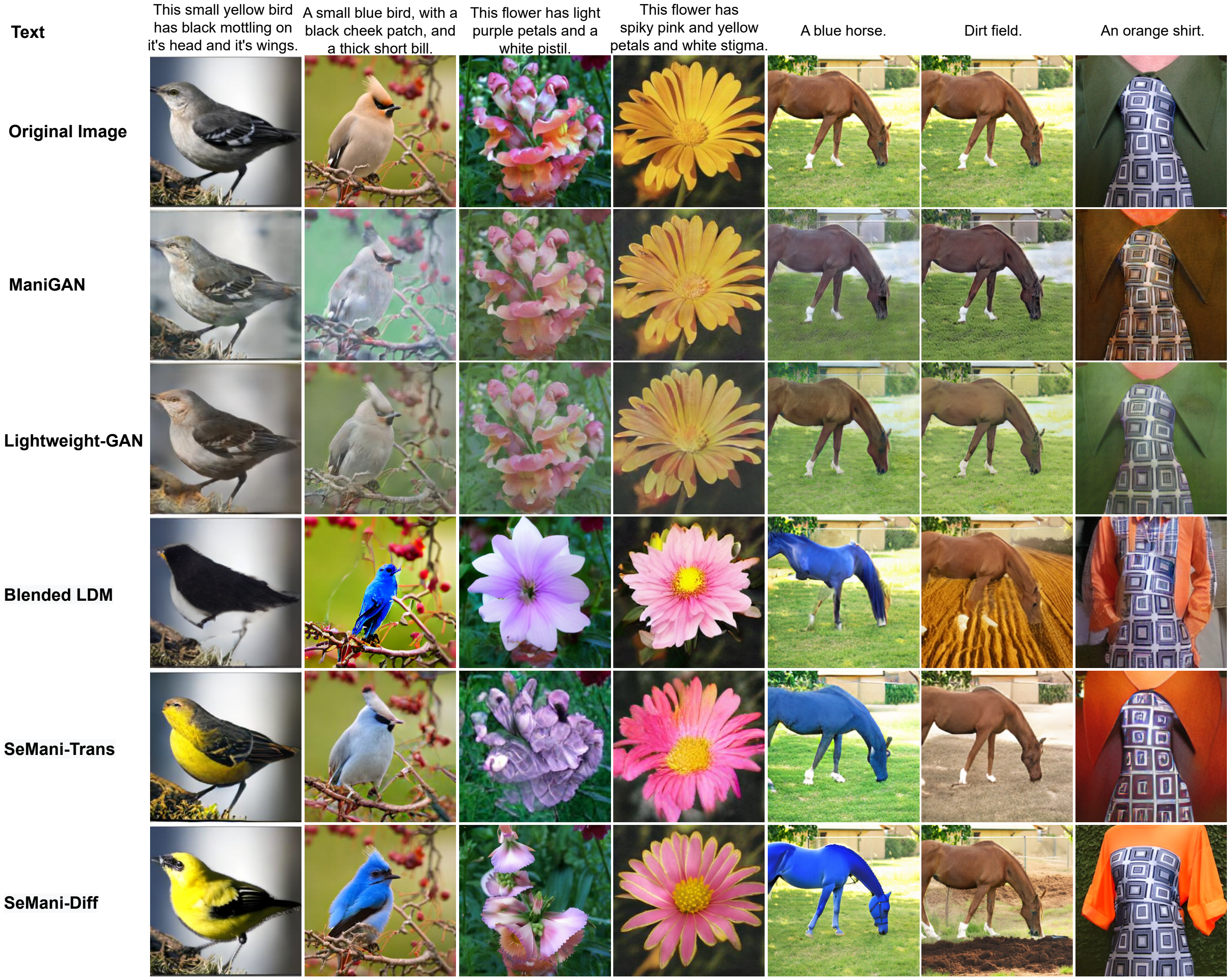

Text
This small yellow bird has black mottling on it's head and it's wings.
A small blue bird, with a black cheek patch, and a thick short bill.
This flower has light purple petals and a white pistil.
This flower has spiky pink and yellow petals and white stigma.
A blue horse.
Dirt field.
An orange shirt.
Original Image
ManiGAN
Lightweight-GAN
Blended LDM
SeMani-Trans
SeMani-Diff

Fig. 7. Qualitative comparison of different methods on the CUB, Oxford, and COCO datasets. SeMani-Trans uses vision guidance to manipulate the images. Note that Blended LDM uses the entity mask provided by SeMani-Diff as it requires user-provided masks.

optimizer, we fine-tune our model on the three datasets with the same two steps. The first step fine-tunes the model without vision guidance. The second step adds the vision guidance into training with 50% samples. Each step lasts 500 epochs with batch size 96 and the learning rate linearly ramps up to $5 \times e^{-4}$ for the first 1k iterations and is halved when training loss does not improve for 10 epochs.

Hyper-parameters of SeMani-Diff. In SeMani-Diff, we adopt the text-guided inpainting model of latent diffusion models [41] trained on LAION-5B [86] dataset as our generative model, as it is more suitable for entity manipulation task. The diffusion model has 866M parameters. We use the segmentation model and fine-tuned CLIP model in OVSeg [77]. In the image manipulation phase, we adopt the DDIM [87] generation process with 50 steps with a classifier-free guidance scale $s = 9$.

As discussed in Section 3.2.3, we set a word prompt for the entity to be manipulated in the inference phase. Particularly, CUB and Oxford have specific category images, where we set “bird” and “flower” as the prompt word respectively. COCO contains various category entities, and we randomly set a prompt word based on the original caption of each image to ensure the entity exists in the image and randomly select a caption of other images as the target text description. Almost all the prompt words of COCO are the nouns of their text in the following experiments and we will clarify the prompt words for special examples.

4.2 Main Results

In this section, we first qualitatively verify the manipulation ability of our model to edit or change the entity on the CUB, Oxford, and COCO datasets. As Fig. 5 shows, SeMani-Trans can manipulate the images with the same object structure providing the vision guidance, i.e. the grayscale image, as prior shape information. Without the constraint of the vision guidance, our model generates a different entity corresponding to the text description in place of the original entity. SeMani-Trans merges the generation ability to the manipulation without any user manual mask but only the guidance of input text, where most existing models fail. SeMani-Diff enjoys better background preservation compared with SeMani-Trans (for example the first column). SeMani-Diff also enjoys high-fidelity manipulation consistent with the provided text description. As SeMani-Diff does not have a vision guidance module, it can not preserve the shape of the entity.

We also conduct an experiment of SeMani-Trans that trained on the mixture of datasets CUB and Oxford to verify a wider manipulation than on the same category. As shown in Fig. 6, SeMani-Trans generates reasonably manipulated entities which are corresponding to the text and fit the background, in both bird-to-flower and flower-to-bird settings. For example, in the third column from the left, SeMani-Trans not only generates a flower consistent with the description but also complements the upper left corner of the manipulated flower with a leaf, which shows that SeMani-Trans also learns a combination of the object information and the background. As SeMani-Diff is trained on the large-scale dataset, it also has the capacity of generating a new category.

4.3 Comparison with the State of the Art

Quantitative results. Table 1 shows the quantitative comparison of our method against previous methods, including ManiGAN [20], Lightweight-GAN [21], and Blended LDM [79]. Note that Blended LDM uses the entity mask provided by SeMani-Diff as it requires user-provided masks. **(1)** Compare SeMani-Trans with GAN competitors: On CUB and Oxford datasets, our SeMani-Trans achieves better results than other models on almost all metrics, except for the L2-error on the Oxford dataset, where SeMani-Trans is competitive with ManiGAN. It demonstrates that our method can generate high-quality manipulated images (IS), which are consistent with the text descriptions (CLIP-score and R@10) , and preserve the content of original images (L2-error). For the more complicated dataset, COCO, SeMani-Trans outperforms the ManiGAN and Lightweight-GAN on the R@10 and L2-error and achieves a competitive CLIP-score. The IS of our method is competitive with ManiGAN and Lightweight-GAN. However, as Fig. 7 shows, within many text-guided manipulation cases, ManiGAN and Lightweight-GAN both change the images slightly, more like applying a filter, while SeMani-Trans conducts manipulation according to the text. Typically, the former one is easier to generate high-quality images than the latter and this is why their IS are a bit higher than our method. **(2)** On CUB, SeMani-Diff enjoys a much better CLIP-score and R@10 compared with Blended LDM, while on Oxford the performance diverges. In COCO, SeMani-Diff has a much better IS score than Blended LDM, which can be confirmed in Fig. 7 that Blended LDM introduces unrealistic edges around the manipulated entity (for example the first and second columns) while SeMani-Diff won't. **(3)** In summary, SeMani-Trans and SeMani-Diff show superior or comparable eL-TGIM capacity compared with other methods.

TABLE 2
The average rank of user study between ManiGAN [20], Lightweight-GAN [21], Blended LDM [79], and our SeMani-Trans and SeMani-Diff. Cons., Pre., and Fid. are the abbreviations for consistency, preservation, and fidelity, respectively. Note that Blended LDM uses the entity mask provided by SeMani-Diff as it requires user-provided masks. The lower rank indicates better performance.

Model	Con. (↓)	Pre. (↓)	Fid. (↓)
ManiGAN	3.27	3.76	2.60
Lightweight-GAN	4.49	4.43	4.28
Blended LDM	2.79	2.40	3.16
SeMani-Trans	2.68	2.93	3.03
SeMani-Diff	**1.78**	**1.48**	**1.93**

Qualitative results. As Fig. 7 shows, compared with the original images, ManiGAN directs the images toward the semantics of text closer than Lightweight-GAN but changes the background style further from the original as well. Lightweight-GAN preserves the irrelevant contents better than ManiGAN while failing in transforming the text-relevant regions according to the descriptions. Blended LDM can generate entities that are consistent with text description, but will introduce unrealistic edges around the entities. SeMani-Trans outperforms them both on background preservation and foreground manipulation. As the second and third columns from the right of the Fig. 7 show, our SeMani-Trans can manipulate the horse and the field respectively on one image, while the baseline methods only change the whole image style with the text. SeMani-Diff enjoys a higher quality of manipulated results.

User study. We conduct a user study experiment to obtain the subjective evaluation of humans. Specifically, we randomly select 20 testing cases and perform eL-TGIM using our algorithms and competitors. Then for each case, we design three different perspectives of quality assessment: (1) *Consistency* which reflects the first requirement of eL-TGIM, to check whether the manipulated image is consistent with the text description; (2) *Preservation* which reflects the second requirement of eL-TGIM, to ensure that the manipulated image will preserve the text-irrelevant regions; (3) *Fidelity* which generalizes the third requirement of eL-TGIM, to judge whether the manipulated image is indistinguishable from real pictures. Then the 20 groups are divided into 5 questionnaires, each of which contains $4 \times 3 = 12$ ranking questions to rank the manipulated images (the lower the better) according to the above three perspectives. Participants are invited to answer one or several questionnaires, and 45 valid questionnaires were recovered. The screenshot of the questionnaires can be founded in the appendix, and one of the questionnaires can be found via https://www.wjx.cn/vm/P2IJ5RF.aspx for convenience.

Results of averaged ranking results are shown in Table 2. While SeMani-Trans show comparable results compared with diffusion models, our SeMani-Diff achieves significantly superior performance compared with other methods. This further validates the effectiveness of SeMani and its compatibility with diffusion models.

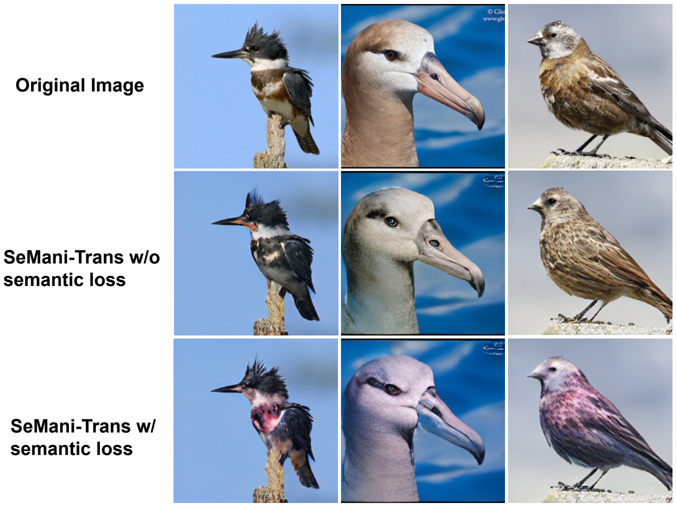


Fig. 8. SeMani-Trans w/ and w/o semantic loss on CUB dataset. The text is "This particular bird has a belly that is purple and gray.".

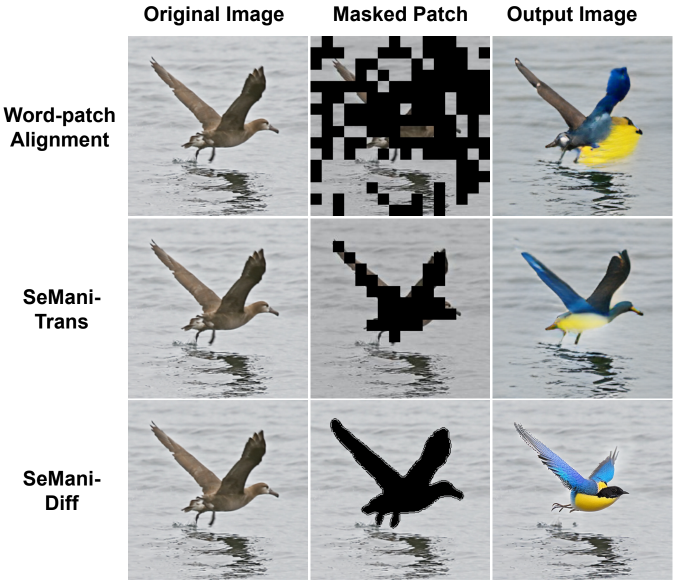
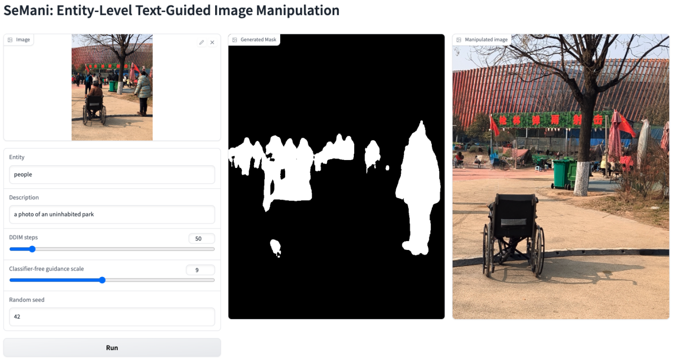


Fig. 9. Qualitative comparison of our methods with our semantic alignment mechanism and word-to-patch alignment on the CUB dataset. The text is "This bird has wings that are blue and has a yellow belly.".

4.4 Ablation of SeMani

Semantic loss in SeMani-Trans. A comparison of SeMani-Trans trained with and without the semantic loss is shown in Fig. 8. The model trained with the semantic loss manipulates the bird as gray and purple, while the model trained without the semantic loss neglects the purple. It implies that semantic loss helps the model capture the relation between image and text.

Semantic alignment. Word-patch alignment is a technique to align a pair of text and image tokens used in many multi-modality transformer methods [25], [73]. The word-patch alignment begins from the word tokens to sort the patch tokens, which takes the image patches separately and neglects the information of the entity tokens as a whole during the alignment. Thus the selected image tokens may well scatter within or around an entity area. Manipulating these scattered tokens gets a messy image, where the foreground stays while the background changes. A comparison between word-patch alignment and our semantic alignment method is shown in Fig. 9. The two methods share the same similarity threshold θ for sorting the image tokens. As Fig. 9 shows, our semantic alignment selects the image patches corresponding to the bird precisely, while the word-patch alignment misses some patches corresponding to the bird and selects a few patches which belong to the background. With the inaccurate patches selected by word-patch alignment, only the right-wing turn to blue, and the yellow leaks out. Although the color of the two manipulated images both match the description, the qualitative result by the semantic alignment is better, resulting from more precise edited locations. Besides, the SeMani-Diff enjoys a more precise alignment in the pixel space, indicating the superiority of the fine-grained control of manipulating region.


SeMani: Entity-Level Text-Guided Image Manipulation

Entity: people

Description: a photo of an uninhabited park

DDIM steps: 50

Classifier-free guidance scale: 9

Random seed: 42

Fig. 10. Illustration of our interface for eL-TGIM. The left part is the input area, where users can input an *image*, an *entity* to manipulate, a text *description* to guide the manipulation, and several hyper-parameters to modify. Then our model will output the generated mask of the entity in the middle part, and the manipulated image in the right part.

4.5 Interface of SeMani

We design an interface of SeMani for users to perform eL-TGIM with little effort. Our interface is constructed with Gradio [88]. As shown in Fig. 10, the left part of the interface is the input area, where users can upload an original *image*, an *entity* to manipulate, and a text *description* to guide the manipulation. We also provide several hyper-parameters of the generation model to modify to better match the needs of users. Then our model will output the generated mask of the entity in the middle part, and the manipulated image in the right part. A demo video of using the interface to generate results in Fig. 1 is in the supplementary material.

5 Conclusion

For the first time, this paper studies a new task – entity-level text-guided image manipulation. To tackle this task, we propose a novel framework – SeMani for the semantic manipulation of eL-TGIM. Two variants of SeMani from discrete and continuous viewing of images are proposed, respectively. SeMani-Trans proposes token-wise semantic alignment and manipulation, while SeMani-Diff directly performs semantic alignment and image manipulation at the pixel-level. Experiments on CUB, Oxford, and COCO validate the superiority of SeMani.

References

[1] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein *et al.*, "Muse: Text-to-image generation via masked generative transformers," *arXiv preprint arXiv:2301.00704*, 2023. 1, 1

[2] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, 2022. 1, 1

[3] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," *arXiv preprint arXiv:2210.09276*, 2022. 1, 1, 2

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022. 1, 1, 2

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423. 1

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134. 1, 2

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232. 1

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. 1

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 2014. 1

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. 1, 2, 3.2.1, 3.3.1

[11] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015. 1, 1, 2, 3.1

[12] A. El-Nouby, S. Sharma, H. Schulz, D. Hjelm, L. E. Asri, S. E. Kahou, Y. Bengio, and G. W. Taylor, "Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 304–10 312. 1, 2

[13] T. Zhang, H.-Y. Tseng, L. Jiang, W. Yang, H. Lee, and I. Essa, "Text as neural operator: Image manipulation by text instruction," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1893–1902. 1, 2

[14] T.-J. Fu, X. Wang, S. Grafton, M. Eckstein, and W. Y. Wang, "Iterative language-based image editing via self-supervised counterfactual reasoning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4413–4422. 1, 2

[15] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu, "Language-based image editing with recurrent attentive models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8721–8729. 1, 2

[16] H. Wang, J. D. Williams, and S. Kang, "Learning to globally edit images with textual description," *arXiv preprint arXiv:1810.05786*, 2018. 1, 2

[17] W. Jiang, N. Xu, J. Wang, C. Gao, J. Shi, Z. Lin, and S. Liu, "Language-guided global image editing via cross-modal cyclic mechanism," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2115–2124. 1, 2

[18] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5706–5714. 1, 2

[19] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: manipulating images with natural language," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 42–51. 1, 2, 4.1

[20] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "Manigan: Text-guided image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7880–7889. 1, 2, 4.1, 1, 4.3, 2

[21] B. Li, X. Qi, P. Torr, and T. Lukasiewicz, "Lightweight generative adversarial networks for text-guided image manipulation," *Advances in Neural Information Processing Systems*, vol. 33, 2020. 1, 2, 4.1, 1, 4.3, 2

[22] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017. 1

[23] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in neural information processing systems*, 2019, pp. 14 866–14 876. 1

[24] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 873–12 883. 1, 4.1

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 4.4

[26] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011. 1, 4.1

[27] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729. 1, 4.1

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 1, 4.1

[29] J. Wang, G. Lu, H. Xu, Z. Li, C. Xu, and Y. Fu, "Manitrans: Entity-level text-guided image manipulation via token-wise semantic alignment and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 707–10 717. 1

[30] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1060–1069. 2

[31] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324. 2, 4.1

[32] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915. 2, 4.1

[33] ——, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018. 2

[34] B. Li, X. Qi, T. Lukasiewicz, and P. H. Torr, "Controllable text-to-image generation," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 2065–2075. 2

[35] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021. 2, 3.2.2

[36] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *arXiv preprint arXiv:2105.13290*, 2021. 2

[37] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis," *arXiv preprint arXiv:2108.08827*, 2021. 2

[38] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*, 2022. 2

[39] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation." *J. Mach. Learn. Res.*, vol. 23, pp. 47–1, 2022. 2

[40] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, 2022. 2

[41] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695. 2, 3.3.1, 4.1

[42] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2256–2265. 2

[43] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410. 2

[44] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094. 2, 3.2.2

[45] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in *European Conference on Computer Vision*. Springer, 2022, pp. 707–723. 2

[46] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 208–18 218. 2

[47] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022. 2

[48] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *International Conference on Learning Representations*, 2022. 2

[49] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022. 2

[50] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," *arXiv preprint arXiv:2211.09794*, 2022. 2

[51] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014. 2

[52] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1511–1520. 2

[53] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807. 2

[54] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. 2

[55] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," *arXiv preprint arXiv:2102.05918*, 2021. 2

[56] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588. 2

[57] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 647–12 657. 2

[58] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning," *arXiv preprint arXiv:2012.15409*, 2020. 2

[59] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120. 2

[60] J. Lin, R. Men, A. Yang, C. Zhou, M. Ding, Y. Zhang, P. Wang, A. Wang, L. Jiang, X. Jia *et al.*, "M6: A chinese multimodal pretrainer," *arXiv preprint arXiv:2103.00823*, 2021. 2

[61] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137. 2

[62] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *International Conference on Learning Representations*, 2019. 2

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 2

[64] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020. 2

[65] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 3.1, 3.2.1

[66] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018. 3.1

[67] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019. 3.2.1

[68] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019. 3.2.2

[69] M. Ni, H. Huang, L. Su, E. Cui, T. Bharti, L. Wang, D. Zhang, and N. Duan, "M3p: Learning universal representations via multitask multilingual multimodal pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3977–3986. 3.2.2

[70] M. Zhou, L. Zhou, S. Wang, Y. Cheng, L. Li, Z. Yu, and J. Liu, "Uc2: Universal cross-lingual cross-modal vision-and-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4155–4165. 3.2.2

[71] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013. 3.2.2

[72] L. Qi, J. Kuen, Y. Wang, J. Gu, H. Zhao, Z. Lin, P. Torr, and J. Jia, "Open-world entity segmentation," *arXiv preprint arXiv:2107.14228*, 2021. 3.2.3

[73] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," 2021. 3.2.3, 4.1, 4.4

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241. 3.3.1

[75] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. 3.3.2

[76] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3.3.2

[77] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," *arXiv preprint arXiv:2210.04150*, 2022. 3.3.3, 4.1

[78] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 2022, pp. 709–727. 3.3.3

[79] O. Avrahami, O. Fried, and D. Lischinski, "Blended latent diffusion," *arXiv preprint arXiv:2206.02779*, 2022. 4.1, 1, 4.3, 2

[80] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016. 4.1

[81] N. Shazeer, "Glu variants improve transformer," *arXiv preprint arXiv:2002.05202*, 2020. 4.1

[82] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941. 4.1

[83] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016. 4.1

[84] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015. 4.1

[85] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568. 4.1

[86] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4.1

[87] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020. 4.1

[88] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ml models in the wild," *arXiv preprint arXiv:1906.02569*, 2019. 4.5