Tuning-Free Inversion-Enhanced Control for Consistent Image Editing

Xiaoyue Duan[1,2*†], Shuhao Cui[2*], Guoliang Kang[1,4], Baochang Zhang[1,3,4,5], Zhengcong Fei[2], Mingyuan Fan[2], Junshi Huang[2‡]

[1]School of Automation Science and Electrical Engineering, Beihang University, China
[2]Meituan
[3]Hangzhou Research Institute, Beihang University, China
[4]Zhongguancun Laboratory, Beijing, China
[5]Nanchang Institute of Technology, Nanchang, China
{LittleMoon, bczhang}@buaa.edu.cn, {hassassin1621, kgl.prml, feizhengcong, fanmingyuan01, junshi.huang}@gmail.com

Abstract

Consistent editing of real images is a challenging task, as it requires performing non-rigid edits (*e.g.*, changing postures) to the main objects in the input image without changing their identity or attributes. To guarantee consistent attributes, some existing methods fine-tune the entire model or the textual embedding for structural consistency, but they are time-consuming and fail to perform non-rigid edits. Other works are tuning-free, but their performances are weakened by the quality of Denoising Diffusion Implicit Model (DDIM) reconstruction, which often fails in real-world scenarios. In this paper, we present a novel approach called Tuning-free Inversion-enhanced Control (TIC), which directly correlates features from the inversion process with those from the sampling process to mitigate the inconsistency in DDIM reconstruction. Specifically, our method effectively obtains inversion features from the key and value features in the self-attention layers, and enhances the sampling process by these inversion features, thus achieving accurate reconstruction and content-consistent editing. To extend the applicability of our method to general editing scenarios, we also propose a mask-guided attention concatenation strategy that combines contents from both the inversion and the naive DDIM editing processes. Experiments show that the proposed method outperforms previous works in reconstruction and consistent editing, and produces impressive results in various settings.

Introduction

In recent years, remarkable progress in text-to-image (T2I) generation has been witnessed, as evidenced by the impressive results by powerful models (Ramesh et al. 2021; Ding et al. 2022; Ramesh et al. 2022). These large-scale T2I models (*e.g.*, Stable Diffusion (Rombach et al. 2022)) are capable of generating diverse and high-quality images that match the given text descriptions. Moreover, by leveraging T2I models, we can also perform text-guided image editing, as demonstrated by recent works (Nichol et al. 2021; Parmar et al. 2023). These works involve various forms of text guidance and task settings, which highlight the need to address the challenges and opportunities in field.

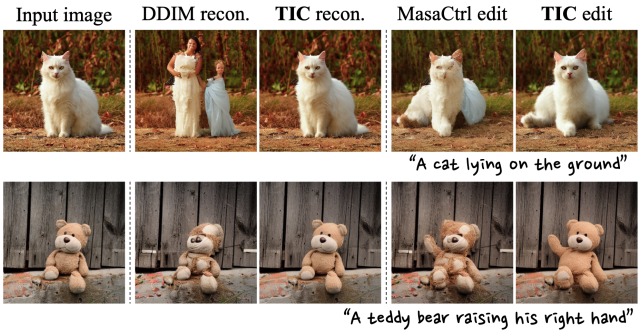

Figure 1: Naive DDIM reconstruction may result in bad cases (col. 2). Therefore, querying the contents from the naive DDIM reconstruction process during editing (*i.e.*, MasaCtrl) lead to unreasonable editing results (col. 4). By enhancing the self-attention layer with contents from inversion, our method accurately reconstruct the input image (col. 3), and thus achieve content-consistent editing (col. 5).

In the realm of text-guided image editing, a crucial and practical task is to achieve consistent image editing. Consistent image editing, as demonstrated in recent works such as (Meng et al. 2021; Hertz et al. 2022), involves preserving the identity of objects and background details in the input image while modifying only certain non-rigid attributes of the objects (*e.g.*, changing posture). Nevertheless, this requirement remains challenging, as existing text-guided editing methods can hardly solve the problems of consistent editing effectively. For example, some methods (Brooks, Holynski, and Efros 2022; Tumanyan et al. 2022) are capable of preserving the structure or layout of the input image when performing style transfer or object replacement, but they can barely preserve the content or identity of the object, resulting in inconsistent editing. Imagic (Kawar et al. 2022) can preserve the original attributes of the object, which allows for non-rigid editing. Yet, it requires fine-tuning the entire T2I model and optimizing the textual embedding for each input image, which is not acceptable for real-world applications in terms of efficiency.

To achieve efficient consistent image editing, subsequent

*These authors contributed equally.
†Work done during an internship in Meituan.
‡Corresponding author.

works (Mokady et al. 2022; Dong et al. 2023; Han et al. 2023) avoid fine-tuning the entire model, and learn the structural information of the input image by tuning either the unconditional or the conditional embedding of the classifier-free guidance based on a pivotal latent trajectory obtained through DDIM inversion (Ho, Jain, and Abbeel 2020). However, tuning the embedding for each edit is still time-consuming, and cannot achieve complex non-rigid editing. Recent works, *e.g.*, MasaCtrl (Cao et al. 2023), achieve consistent editing without fine-tuning any part of the model, but it may introduce artifacts in real-image editing scenarios, and its performance is largely constrained by DDIM reconstruction quality, as shown in Fig. 1.

In this paper, we present a novel approach called Tuning-free Inversion-enhanced Control (TIC) for consistent image editing. Our approach is based on the theoretical analysis of the reconstruction error between the DDIM inversion and sampling processes. We find that the reconstruction error is mainly caused by the bias in the predicted noises resulted from an inaccurate timestep approximation assumption. To address this issue, we propose to enhance the DDIM sampling process by incorporating features from the DDIM inversion process. Specifically, we enhance the self-attention layers by replacing the key and value features in the sampling process with the corresponding features in the inversion process, thus allowing the model to focus on important pixels in the image features. Our method is tuning-free, and outperforms existing editing methods in both image reconstruction and consistent editing of real images, as demonstrated in Fig. 1.

To extend our method to more general and diverse editing scenarios, we further propose a mask-guided attention concatenation strategy, which achieves a good balance between fidelity and editability by querying contents from both the inversion process and the naive DDIM editing process in the mask-guided editing areas. We also demonstrate the effectiveness of our method by integrating it into controllable diffusion models to further enhance the structural layout of the input image. Experiments show the effectiveness and applicability of our method in various settings.

Overall, our contributions can be summarized as: **1)** We conduct theoretical analysis on the reconstruction error of DDIM. Based on the analysis, we propose Tuning-free Inversion-enhanced Control (TIC) to achieve accurate reconstruction and consistent editing of real images. **2)** We extend TIC to more general editing settings by proposing a mask-guided attention concatenation strategy, which achieves a good balance between fidelity and editability. We also demonstrate its effectiveness when integrated with controllable diffusion models. **3)** We demonstrate the versatility and applicability of the proposed method by conducting experiments under both qualitative and quantitative settings.

Related Work

Text-to-image generation. Generating images by text descriptions are mainly based on architectures of Generative Adversarial Networks (GANs) (Reed et al. 2016; Zhang et al. 2018; Brock, Donahue, and Simonyan 2018; Tao et al. 2022), auto-regressive generation (Ding et al. 2021) and diffusion models (Song and Ermon 2019; Nichol and Dhariwal 2021; Ramesh et al. 2022). Early methods based on GANs (Xu et al. 2018; Zhang et al. 2021; Zhou et al. 2022) align the text descriptions and image contents through multi-modal vision-language learning, but can only achieve impressive results on specific domains. By adopting large-scale models and datasets, auto-regressive generation (Ramesh et al. 2021; Ding et al. 2022; Yu et al. 2022) obtain powerful results for open-domain text descriptions. More recently, diffusion models (Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021; Gu et al. 2022; Ho and Salimans 2022) achieve state-of-the-art synthesis results in terms of image quality and diversity. Conditioned on the text prompt, various text-to-image diffusion models (Nichol et al. 2021; Ramesh et al. 2022; Zeng et al. 2023a,b) can synthesize image contents highly consistent with the textual description.

Text-guided image editing. Text-guided image editing is a challenging task that involves manipulating images based on natural language descriptions. Previous methods based on GANs (Nam, Kim, and Kim 2018; Li et al. 2020; Xia et al. 2021; Patashnik et al. 2021; Pan et al. 2023), or auto-regressive models (Crowson et al. 2022) have achieved some success on domain-specific datasets, but their applicability are limited. Recently, the development of diffusion models provide a more flexible space and a more efficient way for editing, while following a simpler setup (Meng et al. 2021). Some works (Nichol et al. 2021; Avrahami, Fried, and Lischinski 2022) leverage extra masks to edit specific regions of the image, while others (Kim, Kwon, and Ye 2022; Brooks, Holynski, and Efros 2022) can edit global aspects of the image by directly modifying the text prompt.

Consistent image editing. Consistent image editing refers to the process of editing images without altering their main components. Imagic (Kawar et al. 2022) allows for various non-rigid editing by directly modifying the prompts. Hertz et al. (2022) utilizes cross-attention or spatial features to edit both global and local aspects of the image by modifying the text prompt. Later, Mokady et al. (2022); Dong et al. (2023) use an initial DDIM inversion as an anchor for optimization, which only tunes the prompt embeddings used in classifier-free guidance. Recently, MasaCtrl (Cao et al. 2023) combines the contents from the source image and the layout synthesized from text prompt to synthesize or edit the desired image. However, these methods either require fine-tuning, which is time-consuming and fail to perform non-rigid edits, or are highly constrained by the DDIM reconstruction quality. For fast and effective editing, a tuning-free method with high reconstruction quality is in need.

Methodology

Given a real image $\mathcal{I}$ and a target prompt $\mathcal{P}^*$, consistent image editing aims to perform non-rigid edits to $\mathcal{I}$ (*e.g.*, changing postures) to make its visual content comply with the textual content in $\mathcal{P}^*$, while preserving the original texture and identity of the main components in $\mathcal{I}$. To preserve the texture and identity, we believe that an accurate reconstruction of the input is a basic guarantee for consistent image editing.

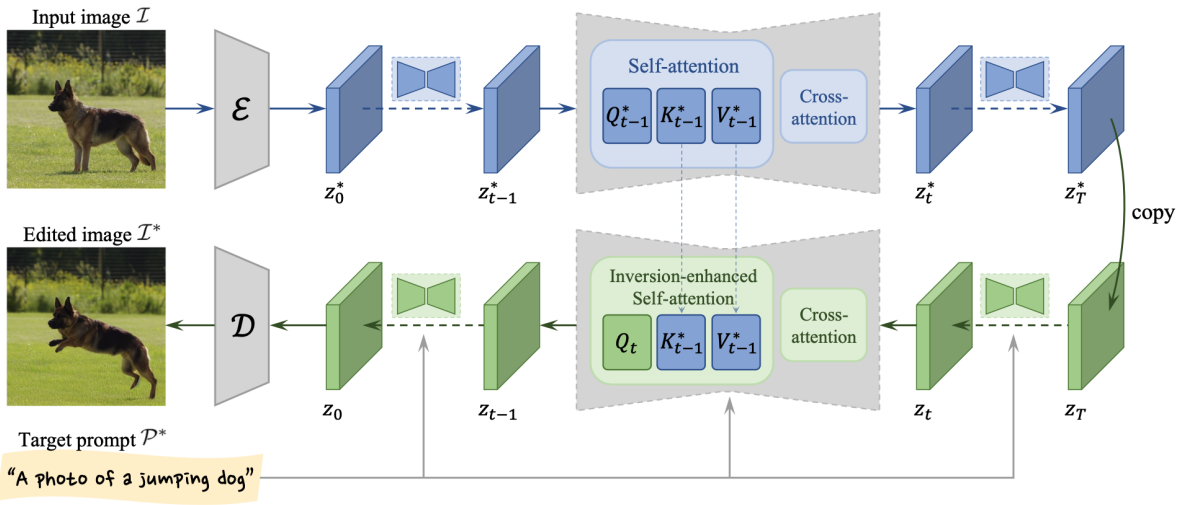

Figure 2: Overview of Tuning-free Inversion-enhanced Control (TIC). Our method first performs DDIM inversion on a given real image to obtain a series of features in the self-attention layers (1st row). These features from inversion contain valuable texture information of the input, which we adopt to enhance the self-attention layers to perfectly reconstruct the input, and achieve non-rigid and content-consistent image editing with the guidance of the target prompt (2nd row).

In this section, we first analyze the limitations of DDIM reconstruction in Latent Diffusion Models (LDMs), which may result in failed reconstruction of the input. Based on the analysis, we then propose Tuning-free Inversion-enhanced Control (TIC) (see Fig. 2 for an overview), which adopts features from inversion to enhance the sampling process. With TIC, we can perfectly reconstruct the input image, and thus perform non-rigid consistent image editing. Finally, we introduce the extensions of TIC for more typical editing, including a mask-aware attention concatenation strategy, and an integration to controllable diffusion models.

Limitations of DDIM Reconstruction

Following the framework of Latent Diffusion Models (Rombach et al. 2022), a pre-trained encoder $\mathcal{E}$ maps the input image $\mathcal{I}$ to a latent representation z_0, and the input image can be then reconstructed using a decoder $\mathcal{D}$, *i.e.*, $\mathcal{I} \approx \mathcal{D}(\mathcal{E}(\mathcal{I}))$, which can be regarded as the upper bound for image reconstruction.

The deterministic DDIM sampling (Song, Meng, and Ermon 2020) adopts the following denoising process in LDMs:

$$\frac{z_{t-1}}{\sqrt{\alpha_{t-1}}} = \frac{z_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \epsilon_t, \quad (1)$$

where ϵ_t denotes the noise prediction process at timestep t. A random Gaussian noise z_T is gradually removed to generate an image latent z_0 by applying Eqn. 1 for T steps. However, as the Gaussian noise is randomly sampled, the generated image by DDIM sampling can be far different from the input one. For better image reconstruction, DDIM inversion is suggested to reverse DDIM sampling, based on the assumption that the ordinary differential equation (ODE) process can be reversed within the limit of small steps:

$$\frac{z_t^*}{\sqrt{\alpha_t}} = \frac{z_{t-1}^*}{\sqrt{\alpha_{t-1}}} + \left(\sqrt{\frac{1}{\alpha_t} - 1} - \sqrt{\frac{1}{\alpha_{t-1}} - 1}\right) \cdot \epsilon_{t-1}^*, \quad (2)$$

where the superscript * denotes features in the inversion process. However, in practice, a slight error is incorporated at each timestep t.

Naive DDIM reconstruction first performs DDIM inversion in Eqn. 2 for T steps to obtain a pivotal trajectory of latents $\{z_t^*\}_{t=0}^T$. Then, starting from $z_T = z_T^*$, DDIM sampling is performed based on Eqn. 1 to obtain the reconstructed trajectory $\{z_t\}_{t=T}^0$. This naive DDIM inversion and reconstruction process is known to provide a rough approximation of the original image (Song, Meng, and Ermon 2020) according to the assumption of the ODE process mentioned above. However, with a larger value of T (*e.g.*, $T = 50$), the errors may be accumulated, leading to failed reconstructions.

To analyze the error in each reconstruction step (*i.e.*, from t to $t-1$), we assume the starting latent is the same, *i.e.*, $z_t = z_t^*$. Then, based on Eqn. 1 and 2, the reconstruction error between z_{t-1} and z_{t-1}^* can be calculated as:

$$\frac{z_{t-1} - z_{t-1}^*}{\sqrt{\alpha_{t-1}}} = \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \left(\epsilon_t - \epsilon_{t-1}^*\right), \quad (3)$$

where α_t is a fixed value with certain t. Then the reconstruction error at step t can be formulated as:

$$z_{t-1} - z_{t-1}^* = \mathcal{C}_t \cdot \|\epsilon_t - \epsilon_{t-1}^*\|, \quad (4)$$

where $\mathcal{C}_t$ is a constant at step t. This indicates that the main source of reconstruction error comes from the difference between the predicted noise in the inversion process ϵ_{t-1}^* and that in the sampling process ϵ_t at each step. This difference is resulted from an inaccurate timestep approximation assumption from t to $t-1$. Therefore, reducing the error between the predicted noises at each step is the key to accurate reconstruction.

Tuning-Free Inversion-Enhanced Control

Since the noise prediction is conducted by the U-Net ϵ_θ, during the inversion process, ϵ_{t-1}^* can be calculated as $\epsilon_{t-1}^* =$

$\epsilon_\theta(z^*_{t-1}, t-1, \phi)$, where z^*_{t-1}, $t-1$ and ϕ denote the latent input, the timestep and a null-text input, respectively. To reduce the error between ϵ_t and ϵ^*_{t-1} at each step, we aim to obtain a sampling trajectory $\{z_t\}^0_{t=T}$ close to the pivotal trajectory $\{z^*_t\}^T_{t=0}$ for accurate reconstruction. To this end, we propose Tuning-free Inversion-enhanced Control (TIC), which utilizes the features from the inversion process to enhance the sampling process and achieve a precise reconstruction of the original image as:

$$\epsilon_t = \epsilon_\theta(z_t, t, \mathcal{C}; z^*_{t-1}), \tag{5}$$

where we introduce an extra feature from the inversion process z^*_{t-1} into the sampling process to reduce the difference between ϵ_t and ϵ^*_{t-1}, as similar values of ϵ_t and ϵ^*_{t-1} ensure a more accurate reconstruction of the input.

To determine which specific feature to adopt for TIC, we analyze the internal structure of the U-Net ϵ_θ. U-Net is composed of stacked convolutional residual blocks (He et al. 2016) and transformer blocks (Vaswani et al. 2017). More specifically, each transformer block mainly consists of a self-attention layer and a cross-attention layer. The attention mechanism in U-Net is known to play a significant role in capturing long-term dependencies and contextual information in the input sequence. The self-attention layers, particularly, allow the model to focus on important pixels of the image features, and thus can provide indications on how the noise should be *added* (in the inversion process) or *removed* (in the sampling process). Therefore, to introduce an extra input feature z^*_{t-1} into ϵ_θ, we propose to adopt features from the self-attention layers.

In self-attention layers, the query, key and value features (denoted as Q, K and V, respectively) are projected from the spatial features. For convenience, we denote the corresponding self-attention features at timestep t as (Q^*_t, K^*_t, V^*_t) for the inversion process, and (Q_t, K_t, V_t) for the sampling process. Inspired by the *cross-attention layers*, where the key and value features receive extra information from the text inputs to achieve feature extraction and fusion, we believe that the key and value features in the *self-attention layers* can also receive additional texture information from the input feature in the inversion process, *i.e.*, z^*_{t-1}. Since (K^*_{t-1}, V^*_{t-1}) are self-attention features obtained with z^*_{t-1} as the input, they can be directly adopted to enhance the self-attention layers during sampling. Thus, we directly replace the key and value features (K_t, V_t) in the sampling process with the corresponding features (K^*_{t-1}, V^*_{t-1}) in the inversion process, resulting in inversion-enhanced self-attention layers. Accordingly, the noise can be then calculated as:

$$\epsilon_t = \epsilon_\theta(z_t, t, \mathcal{C}; \{Q_t, K^*_{t-1}, V^*_{t-1}\}), \tag{6}$$

where $\mathcal{C}$ is the null-text input ϕ for reconstruction, and the target prompt $\mathcal{P}^*$ for consistent editing. With this feature replacement, the whole framework of TIC is implemented as shown in Fig. 2.

We summarize our algorithm for consistent editing in Alg. 1. We first perform DDIM inversion based on Eqn. 2 for T steps to obtain the pivotal trajectory $\{z^*_t\}^T_{t=0}$. During the process, we save the key and value features (K^*_t, V^*_t) in the self-attention layers at each step. Then, starting with $z_T = z^*_T$, the sampling process is performed with the guidance of $\mathcal{P}^*$ based on Eqn. 1 to generate the edited image. During the sampling process, the key and value features (K_t, V_t) in the self-attention layers are replaced with the corresponding features (K^*_{t-1}, V^*_{t-1}) from the inversion process for each step. Note that we do NOT perform inversion-enhanced self-attention for all layers or all denoising steps, since such operation in the early steps or the shallow layers of the U-Net disrupts the layout formation of the target image. Following (Cao et al. 2023), the proposed inversion-enhanced self-attention is only performed when $l > l_0$ and $t > t_0$, where l_0 and t_0 are the start timestep index and layer index for TIC, respectively.

The proposed TIC can achieve accurate reconstruction of the input image without the need for any fine-tuning, effectively solving the problems of naive DDIM reconstruction. Based on this, TIC successfully performs non-rigid edits with the guidance of the given texts, while maintaining high content consistency with the input.

Algorithm 1: Tuning-free Inversion-enhanced Control (TIC) for Consistent Image Editing

Input: A real image $\mathcal{I}$, a target prompt $\mathcal{P}^*$, and the start timestep index t_0 and layer index l_0 of TIC
Output: The edited image $\mathcal{I}^*$

$KV_{\text{list}} = [], z^*_0 = \mathcal{E}(\mathcal{I})$
for $t = 0, 1, ..., T-1$ **do**:
  $\epsilon^*, \{Q^*_t, K^*_t, V^*_t\} \leftarrow \epsilon_\theta(z^*_t, t, \phi)$
  $KV_{\text{list}}[t] \leftarrow \{K^*_t, V^*_t\}$
  $z^*_{t+1} \leftarrow \text{Inverse}(z^*_t, \epsilon^*)$
end for
$z_T = z^*_T$
for $t = T, T-1, ..., 1$ **do**:
  $\{K^*_{t-1}, V^*_{t-1}\} \leftarrow KV_{\text{list}}[t-1]$
  $\{Q_t, K_t, V_t\} \leftarrow \epsilon_\theta(z_t, t, \mathcal{P}^*)$
  if $t > t_0$ **and** $l > l_0$:
    $\epsilon_t = \epsilon_\theta(z_t, t, \mathcal{P}^*; \{Q_t, K^*_{t-1}, V^*_{t-1}\})$
  else:
    $\epsilon_t = \epsilon_\theta(z_t, t, \mathcal{P}^*; \{Q_t, K_t, V_t\})$
  $z_{t-1} \leftarrow \text{Sample}(z_t, \epsilon_t)$
end for
return $\mathcal{I}^* = \mathcal{D}(z_0)$

Extensions of TIC

Mask-guided TIC with attention concatenation. The proposed TIC ensures fidelity to the input, but as it only queries contents from the inversion of the input, it can hardly generate new contents that do not exist in $\mathcal{I}$ (*e.g.*, turn the dog in $\mathcal{I}$ into a cat). This is not friendly for general editing, which often requires new contents (*e.g.*, object replacement). On the contrary, naive DDIM queries almost no content from inversion, and it can generate new contents beyond the original image with the guidance of the prompt. Therefore, we

	VAE recon. (upper bound)	DDIM recon.	NTI recon. (iter=250)	NTI recon. (iter=500)	PTI recon. (iter=250)	PTI recon. (iter=500)	**TIC recon. (ours)**
PSNR (↑)	**27.17**	22.95	25.70	26.69	26.71	26.92	**27.11**
SSIM (↑)	**0.7886**	0.6840	0.7631	0.7797	0.7810	0.7844	**0.7864**
Time (↓)	-	**5.56**	97.86	149.34	89.89	134.53	**5.56**

Table 1: Reconstruction quality (measured by PSNR and SSIM) and time (s per image) of different methods. The proposed TIC outperforms all other methods in reconstruction quality, and is much more efficient than the tuning-based methods (*i.e.*, Null-Text Inversion (NTI) and Prompt-Tuning Inversion (PTI)) under different number of iterations.

propose to concatenate the key and value features in DDIM sampling process with the corresponding features in the inversion process, *i.e.*, $[K_t; K^*_{t-1}]$ and $[V_t; V^*_{t-1}]$, with $[\cdot;\cdot]$ as the concatenation operation. Then the self-attention is formulated as $(Q_t, [K_t; K^*_{t-1}], [V_t; V^*_{t-1}])$. In this way, the model queries both the content faithful to the target prompt, and the content reconstructing the details of the input.

In practice, we aim to edit the desired parts of the image, while preserving the details of other parts. Inspired by previous works (Hertz et al. 2022), we adopt the cross-attention maps to create a binary mask that distinguishes the parts to be edited from the parts to be preserved. Specifically, at each step t, we average the cross-attention maps with the spatial resolution of 16×16 across all heads and layers, resulting in a map $A_t \in \mathbb{R}^{16\times16\times N}$, where N is the number of textual tokens of $\mathcal{P}^*$. We then obtain an averaged cross-attention map for the tokens related to the objects or parts that we want to edit, which is then binarized to obtain the mask. For the editing parts, we adopt the attention concatenation strategy proposed above to query contents from both inversion and naive DDIM editing, balancing fidelity and editability. For other parts that do not require editing, we only query contents from inversion to preserve original details of the input.

Integration to controllable diffusion models. Our method can also be easily integrated with existing controllable image synthesis methods (*e.g.*, ControlNet (Zhang and Agrawala 2023) and T2I-Adapter (Mou et al. 2023)) to better preserve the layout and structure of the input for general editing. Specifically, we first obtain the controllable image map (*e.g.*, depth map) of the input. Then, following the same pipeline as ControlNet, we integrate the controllable image features into the editing process of the mask-guided TIC mentioned above. Through this approach, we achieve more general editing while almost completely preserving the layout and shape of the input. We demonstrate the effectiveness of this combination in the following experiment part.

Experiments

Experiment Setup

Implementation details. We adopt the text-conditional Latent Diffusion Model (Rombach et al. 2022) (also known as Stable Diffusion) with the publicly available checkpoint v1.4. For the DDIM schedule, we perform both inversion and sampling for 50 steps, and retain the original hyperparameter choices of Stable Diffusion. The classifier-free guidance (CFG) scale is set to 7.5 for editing. The step and layer to start TIC is set to $t_0 = 4$ and $l_0 = 10$, respectively.

Baselines and dataset. For reconstruction and editing, we compare TIC with the following baselines: **1)** VAE (Rombach et al. 2022), which directly decodes the latent of the input image without DDIM inversion or sampling, and is commonly considered as the upper bound of reconstruction for LDMs. **2)** DDIM (Song, Meng, and Ermon 2020). **3)** SDEdit (Meng et al. 2021), which is an image-to-image method with the strength value set to 0.5. **4)** Prompt-to-Prompt (P2P) (Hertz et al. 2022), for which we adopt the attention refinement controller to perform non-rigid editing of real images. **5)** Null-Text Inversion (NTI) (Mokady et al. 2022), which learns the structural information of the input image into the unconditional embedding of CFG to maintain layout consistency. The total number of fine-tuning iterations for the unconditional embedding is set to 250 or 500 (*i.e.*, 5 or 10 for each sampling step). **6)** Prompt-Tuning Inversion (PTI) (Dong et al. 2023), which is similar to NTI, but fine-tunes the conditional embedding of CFG instead of the unconditional one. The total number of tuning iterations is also set to 250 or 500 (*i.e.*, 5 or 10 for each sampling step). 7) MasaCtrl (Cao et al. 2023), which is also tuning-free and can perform non-rigid editing as our method. Unlike our method, it queries contents from the naive DDIM reconstruction process of the source image. Other hyperparameters of these methods are set to their default values.

For the dataset, we evaluate the reconstruction quality of VAE, DDIM, NTI, PTI and our method on 200 randomly selected images from the MS-COCO 2017 validation set (Lin et al. 2014). As both NTI and PTI require fine-tuning of the text embedding, we randomly choose one out of five captions for each image from the MS-COCO dataset as the input text prompt. For editing, we perform consistent editing of real images obtained online.

Comparisons on Content-Consistent Image Editing

Reconstruction quality. We first quantitatively evaluate the reconstruction quality of different inversion-based methods on 200 randomly selected images from the MS-COCO validation set. We measure the reconstruction quality by Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM), and efficiency by reconstruction time (Time). As provided in Table 1, the reconstruction quality of our method is significantly superior to DDIM reconstruction, attaining a level of reconstruction that is comparable to VAE, which serves as an upper bound for reconstruction. In addition, compared to the tuning-based methods (*i.e.*, NTI and PTI), our method is tuning-free and much more superior, in terms of both the image reconstruction quality and time.

Consistent image editing. In Fig. 3, we compare the proposed TIC with the baselines on consistent editing of real

Figure 3: Consistent image editing results of different editing methods on real images. Compared to other methods, our method can perform non-rigid edits without introducing artifacts into the image, while maintaining content consistency.

images. Our method can perform non-rigid edits to the postures (the 1st and 2nd rows), facial expressions (the 3rd row) or certain attributes (the 4th row), while completely preserving the foreground and background contents of the original image, while other methods cannot achieve both. For example, in the 1st row, DDIM can generate results that comply with the target prompt $\mathcal{P}^*$ (*i.e.*, make the person raise his right hand), but the background details are significantly changed. SDEdit and P2P neither make the person raise his hand, nor generate contents that are consistent with the input. This is because they try to keep the original layout or the object shape unchanged by leveraging the layout information encoded in the cross-attention maps (P2P) or the input image (SDEdit). Compared to SDEdit and P2P, NTI and PTI can preserve more content and identity of the original image, as they learn the structural information of the input by fine-tuning the unconditional (NTI) or conditional (PTI) embedding of the classifier-free guidance. However, they still introduce artifacts into the background, and fail to perform non-rigid edits. Finally, MasaCtrl can make the person raise his right hand, but it introduces artifacts in the background (see the areas in the red circles), which disrupts consistency. Since DDIM does not always guarantee perfect reconstruction, querying the contents from the reconstruction process of the source image may lead to unsatisfactory results.

Our method can also perform consistent editing when there are multiple objects in the input image. For example, for the results in the 4th row of Fig. 3, our method can turn half a glass of milk into a full glass, while keeping the biscuit and the desk surface unchanged. DDIM, P2P and MasaCtrl can also generate a full glass of milk, but they either significantly change the original content, or introduce a large amount of artifacts. SDEdit, NTI and PTI, on the other hand, completely fail to turn the half glass of milk into a full one. More results are included in the supplementary material.

Extensions for General Text-Guided Editing

Results of mask-guided TIC with attention concatenation. We evaluate the performance of the mask-guided TIC with attention concatenation to demonstrate its effectiveness in more general editing settings. We provide editing results in Fig. 4 of our method and naive DDIM editing. Under each row, we provide the editing prompt and the cross-attention map for generating the guiding mask, which distinguishes the parts to be edited from the parts to be preserved.

As analyzed before, by only querying contents from inversion of the input, TIC can hardly generate new contents that do not exist in the input (see the 3rd column of Fig. 4). With the proposed mask-guided attention concatenation strategy, TIC successfully generates new textual contents in $\mathcal{P}^*$ while better preserving the original contents in terms of parts that do not require editing (the 4th column). For example, for images in the 1st and 2nd rows of Fig. 4, the mask-guided TIC successfully edits the target objects (*i.e.*, turning a branch into a flower for the 1st row, and a white horse into a black one for the 2nd row) while preserv-

Figure 4: Text-guided image editing results of the proposed mask-guided TIC with attention concatenation, which better preserves the contents in areas that do not require editing.

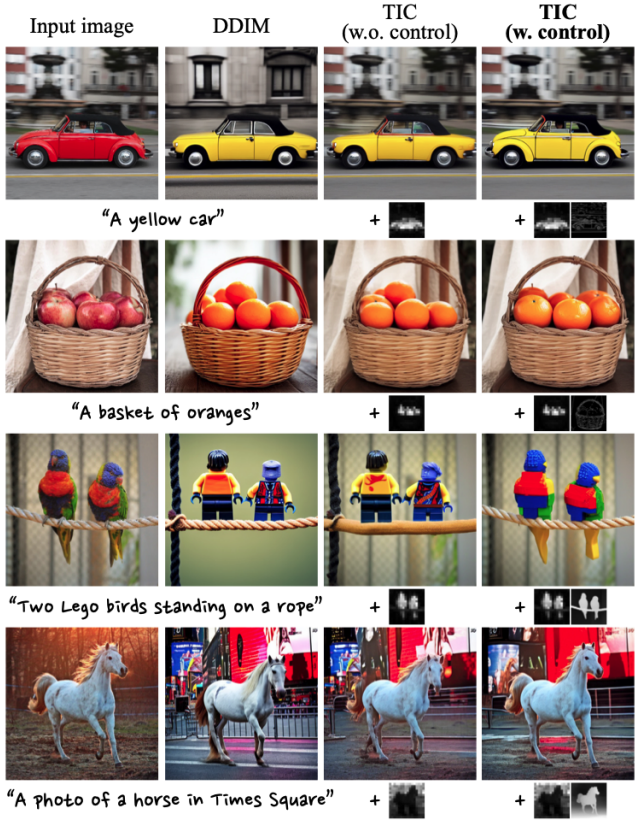

Figure 5: Text-guided image editing results of DDIM and the proposed mask-guided TIC (w. or w.o. additional control). By integrating the mask-guided TIC with canny or depth control, our method further enhances fidelity of the editing to both the contents and the layout of the input image.

ing the details in other parts. DDIM, on the other hand, can barely retain the details of the input for the unedited parts. Besides, when editing the background of the images in the 3rd and 4th rows, DDIM changes the attributes of the foreground drastically, while our method preserves the appearance of the foreground. From the results, TIC is capable of editing both the foreground and the background thanks to the mask for guidance. The results demonstrate that querying the contents from inversion plays an important role in reconstructing the contents in areas that do not need editing.

Results with ControlNet. By integrating the mask-guided TIC with attention concatenation proposed above into controllable diffusion models (*e.g.*, ControlNet (Zhang and Agrawala 2023)), our method can further enhance fidelity by preserving the layout and shape of the original image. In our experiment, we adopt the canny or depth map to extract layout information from the input. In Fig. 5, we show qualitative results with ControlNet. Under each row, we also provide the target prompt, the cross-attention map for generating the guiding mask, and the canny map or depth map for ControlNet (only for the results in the 4th column).

From the results, we observe that DDIM significantly altered the details of the original image. For example, for the results in the 1st row, the background of the synthesized image by DDIM (the 2nd column) is largely different from that in the input. With the mask-aware attention concatenation strategy, TIC (the 3rd column) almost perfectly preserves the details of the background, but the shape of the generated yellow car in the foreground is slightly different from that in the input (*i.e.*, the body of the generated car is a little longer). By introducing canny control, TIC (the 4th column) further ensures the consistency of the car's shape and other attributes, only changing its color. The editing results in other rows show similar trends. The results demonstrate that with the combination of this mask-guided attention concatenation strategy and additional control maps, our method effectively combines the layout synthesized by ControlNet with the target prompt and the content in the input image, further enhancing fidelity without losing editability.

Conclusion

In this paper, we propose Tuning-free Inversion-enhanced Control to perform consistent editing of real images. By introducing features from the DDIM inversion process into the sampling process, our method outperforms previous ones in both accurate reconstruction and content-consistent editing of real images. We demonstrate the versatility and applicability of TIC in various experimental settings. We believe that our approach will be a valuable tool for more applications in both image and video generation, which we point to as future work.

Acknowledgments

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016 and Grant 92370114. The work was also supported by the National Key Research and Development Program of China (Grant No. 2023YFC3300029) and "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268 and ATR key laboratory grant 220402.

References

Avrahami, O.; Fried, O.; and Lischinski, D. 2022. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2022. InstructPix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint arXiv:2211.09800*.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. arXiv:2304.08465.

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

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; et al. 2021. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 34: 19822–19835.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. *arXiv preprint arXiv:2204.14217*.

Dong, W.; Xue, S.; Duan, X.; and Han, S. 2023. Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models. *arXiv preprint arXiv:2305.04441*.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 10696–10706.

Han, L.; Wen, S.; Chen, Q.; Zhang, Z.; Song, K.; Ren, M.; Gao, R.; Chen, Y.; 0003, D. L.; Zhangli, Q.; et al. 2023. Improving Tuning-Free Real Image Editing with Proximal Guidance. *CoRR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

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

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

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

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2426–2435.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. 2020. Manigan: Text-guided image manipulation. In *CVPR*, 7880–7889.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

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

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. *arXiv preprint arXiv:2211.09794*.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*.

Nam, S.; Kim, Y.; and Kim, S. J. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. *NeurIPS*, 31.

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

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *ICML*, 8162–8171. PMLR.

Pan, X.; Tewari, A.; Leimkuhler, T.; Liu, L.; Meka, A.; and Theobalt, C. 2023. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Parmar, G.; Singh, K. K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot Image-to-Image Translation. *arXiv preprint arXiv:2302.03027*.

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

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

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

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *ICML*, 1060–1069. PMLR.

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

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

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*, 16515–16525.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, 2256–2265.

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

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. *arXiv preprint arXiv:2206.10789*.

Zeng, B.; Li, S.; Feng, Y.; Li, H.; Gao, S.; Liu, J.; Li, H.; Tang, X.; Liu, J.; and Zhang, B. 2023a. IPDreamer: Appearance-Controllable 3D Object Generation with Image Prompts. *arXiv preprint arXiv:2310.05375*.

Zeng, B.; Li, S.; Liu, X.; Gao, S.; Jiang, X.; Tang, X.; Hu, Y.; Liu, J.; and Zhang, B. 2023b. Controllable Mind Visual Diffusion Model. *arXiv preprint arXiv:2305.10135*.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, 833–842.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI*, 41(8): 1947–1962.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*.

Zhou, Y.; Zhang, R.; Gu, J.; Tensmeyer, C.; Yu, T.; Chen, C.; Xu, J.; and Sun, T. 2022. Tigan: Text-based interactive image generation and manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3580–3588.