Spin: Diffusion-based Semantic Image Painting Through Independent Information Injection

Dantong Wu[1], Zhiqiang Chen[2], Tianjiao Du[1], Peipei Ran[3], Mengchao Bai[3], Kai Zhang[1]*

[1]Shenzhen International Graduate School, Tsinghua University, China
[2]Institute of Automation, Chinese Academy of Science, China
[3]Media Technology Lab, Huawei, China
wdt22@mails.tsinghua.edu.cn, zhangkai@sz.tsinghua.edu.cn

Abstract

Diffusion models have been utilized as powerful tools for various image editing tasks, including semantic image painting(SIP), which aims to generate content within masked regions conditioned on a reference image or text. SIP, especially those using images as condition, often suffers from three issues: semantic inconsistency, unnatural transitions and style inconsistency, which significantly hinder its practical application. To address these challenges, we propose a novel Semantic image Painting framework with INdependent INformation INjection(Spin). Specifically, we compute a saliency map to segregate the reference image into salient and non-salient components. We then filter out the non-salient information of it during the semantic embedding extraction phrase, and precisely inject the semantic embedding into the masked region instead of the whole image during the semantic generation phrase. Furthermore, we impose an additional style guidance to promote style consistency between background and foreground. Experimental results demonstrate that Spin achieve superior semantic similarity and image coherence across various styles, including realistic, pencil drawings, cartoon, and oil painting. Additionally, Spin offers diversity and editability, and can be integrated into other models that meet our prerequisites.

1 Introduction

In recent years, generative models have rapidly advanced, making editing specific areas of real images more convenient, such as using image inpainting techniques to easily eliminate unwanted objects from designated areas. However, it is a very common scenario where people want to generate specific content in those areas, such as in photo editing and secondary creation of artworks, which is an equally important but understudied aspect. Semantic Image Painting (SIP) aims to fill regions defined by masks in an image according to semantic conditions specified by text or images, ensuring that the background outside the mask remains unchanged, the content within the mask aligns with the semantic conditions, and the image appears visually coherent and complete.

Before the emergence of diffusion models, the methods work best on text-conditioned SIP often leverage the powerful generative power of GAN(Goodfellow et al. 2020). But they can only generate images from a specific domain and can't achieve good preservation of the background. Image-conditioned SIP is divided into several subtasks (Niu et al. 2021), and traditional methods usually only focus on one task while ignoring other aspects. Also, it is difficult to obtain more diverse results while ensuring semantics. With the advancement of diffusion models, notable progress has been made in SIP. Current diffusion-based techniques excel in background preservation, can generate more diverse results, and are not domain-specific.

However, there are still some unsolved issues, due to the simultaneous influence of the semantic condition and the background image on the generated region, which will cause

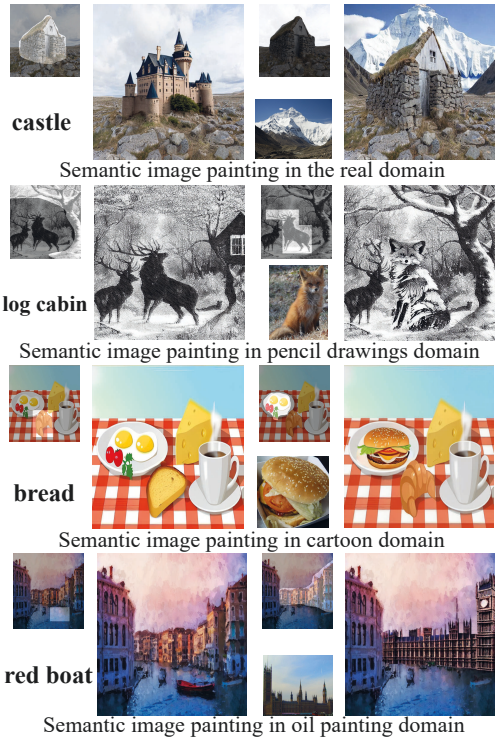

Figure 1: semantic image painting aims to inject semantic condition to the mask region. Our method achieves remarkable results in various domains.

*Corresponding Author.

unnatural results: 1) There is often an insufficient consistency between the main generated content and our desired semantic conditions. 2) The transition between content inside and outside the mask is unnatural, particularly at the boundaries. 3) When the background image is not a image of the real domain, the style of the generated region is not consistent with the background. And above issues are more obvious for image-conditioned SIP. Addressing these issues effectively could significantly enhance the results of SIP and greatly improve its practical application.

The causes of issue 1 and 2 share some similarities, which can be attributed to two main aspects: i) When the semantic condition is specified by a reference image, it includes not only the salient information we aim to preserve but also undesired content, such as background details. This unwanted content may interfere with the generation of the target content, leading to issue 1. Additionally, if undesired elements like background information are generated within the mask, they contribute to issue 2; ii) the cross-attention mechanism of the diffusion model tends to inject the semantic condition into the whole image, so the content generated within the mask reflects only a partial alignment with the semantic condition, contributing to both issues 1 and 2. And issue 3 arises due to insufficient incorporation of the background image information into the mask, which prevents its stylistic elements from being effectively integrated into the foreground. So to address or alleviate the existing issues in SIP, one possible approach is to decouple the semantic and non-semantic parts of the reference image, filtering out non-semantic information while enhancing the injection of semantic details. And the enhancement means that semantic conditions should be guided to be fully injected within the mask. Furthermore, during semantic injection, it is beneficial to constrain the generated content to maintain the similar style as the background image.

In this paper, we propose a framework called Spin, based on text-to-image diffusion models. To accommodate both text and image as semantic condition, we utilize textual inversion technique(Gal et al. 2022) to learn a corresponding text embedding when the semantic condition is specified by an image. To focus only on semantically significant portions, we introduce a saliency focus loss on top of existing textual inversion methods.This ensures that the learned text embedding captures prominent information from the reference image while diminishing attention to non-essential content, such as background information. To inject semantic information into the mask, we design *Attention-based Region Guidance*(ARG). Specifically, we modify the input of the original cross-attention module by using the results of self-attention on the content within the mask as query vectors, which are then cross-attended with text embeddings. This ensures that semantic information is directed into the mask while preserving the original self-attention output, allowing background information to also be injected into the mask. To incorporate the background style into the mask, we use the style of the background image as the target style. We guide the adjustment of the predicted noise by measuring the difference between the style estimated at each step in the pixel space during the diffusion process and the target style. With the above design, we obtain satisfactory results on the SIP task (see Fig.1). Additionally, all our designs are plug-and-play components that can be transferred to other image generation models that meet the criteria, meaning that our performance will improve with the advancement of personalization task for T2I models and the improvement of foundational text-to-image models.

Overall, our key contributions are as follows:

- When the semantic condition is an image, we design a saliency focus loss $\mathcal{L}_{\mathrm{s}}$ to enhance personalization methods, which can learn the semantically significant parts of the reference image while weakening focus about the rest.
- We design the *Attention-based Region Guidance* to make semantic condition injection independent. Combining it with $\mathcal{L}_{\mathrm{s}}$, we achieve both semantic consistency and a harmonious transition.
- We design the style guidance to let background information inject independently, which improves the consistency in style between the masked and unmasked regions.
- Experimental results show that we achieve higher semantic similarity and image harmony than existing methods in four domains. Also, our approach is diverse, editable, model-independent so can be used in other frameworks.

2 Related Work

Personalization for T2I models. To leverage the powerful ability of text-to-image models to generate images similar to the reference image, TI(Gal et al. 2022) learned a text embedding that captures features of the reference image. However, it can't efficiently learn satisfactory embeddings, especially when the distribution of the input image and the training images are significantly different. DreamArtist(Dong, Wei, and Lin 2022) and VCT(Cheng et al. 2023) extended TI to multi-tokens TI, and proposed the latent loss and pixel loss respectively. NeTI(Alaluf et al. 2023) and P+(Voynov et al. 2023) improved the text-conditioning latent space, while DreamBooth(Ruiz et al. 2023) fine-tuned the UNet. They are effective improvements, but all aim to learn the complete information of the reference image while in SIP we only want to learn the salient semantic information of the reference image.

In this paper, we propose a loss that focuses on the salient parts when learning concepts and can be used as reinforcement in all the above methods.

Semantic image painting. Before the advent of diffusion models, significant efforts had already been made to address SIP tasks. BigGAN(Brock, Donahue, and Simonyan 2018) and related works utilized GANs for text-conditioned SIP, while many traditional methods focused on solving specific subtasks of image-conditioned SIP. For instance, Poisson Image Editing(Pérez, Gangnet, and Blake 2023) was one of the outstanding pioneering work focusing on image blending while ignoring other aspects, and DIB(Zhang, Wen, and Shi 2020) has designed the Poisson Gradient Loss based on it. However, they have many limitations and can't achieve diverse and satisfactory results. The fusion of diffusion model and CLIP promotes the progress of text-conditioned SIP, and also enables the image-conditioned SIP to achieve a har-

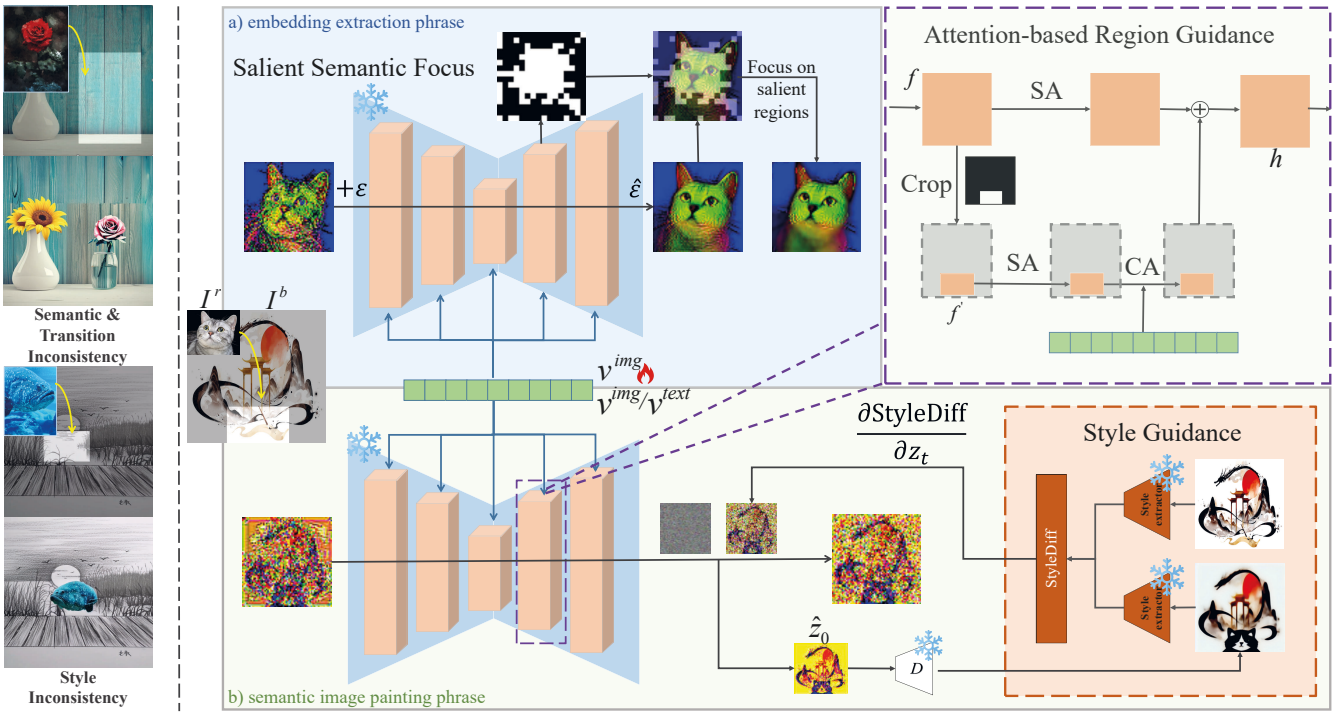

Figure 2: Left: the issues with existing methods. Right: our pipeline. When the semantic condition is an image, we first use our proposed $\mathcal{L}_s$ in the embedding extraction phase to better learn the salienct semantic information of the reference image. Then, in the semantic image painting phase, we use *Attention-based Region Guidance* to guide the learned semantic conditions to be independently injected into the mask, and apply style guidance to update the noise predicted at each step.

monious effect. Blended(Avrahami, Lischinski, and Fried 2022) and BLD(Avrahami, Fried, and Lischinski 2023) introduced gradient guidance to keep the unchanged background outside the mask based on the clip score, allowing the text to affect only the mask region. SmartBrush(Xie et al. 2023), SD-inpaint(Rombach et al. 2022) and PPt (Zhuang et al. 2023) fine-tuned the diffusion model to ensure that the content outside the mask is unchanged and the text only responds inside the mask. They implemented text-conditioned SIP using diffusion models, but can't receive images as input, and thus can't provide more detailed control. Paint(Yang et al. 2023) fine-tuned the original CLIP model and diffusion model, Any(Chen et al. 2023b) represented the target into corresponding features and combined them into the interaction with the background. They performed well in image-conditioned SIP within the real domain, but struggled with tasks in non-real ones. Uni(Yu et al. 2023) could receive both text and images as semantic conditions and had demonstrated good performance in some cross-domain tasks but not all. Also, it often exhibits unnatural transitions at the edges of the mask. TF(Lu, Liu, and Kong 2023) and Magic(Ruiz et al. 2024) achieved impressive cross-domain image synthesis results, but they failed to maintain adequate semantic consistency with reference images. And TF required additional information but failed to produce diverse or editable results.

We propose a method that can receive text or image as semantic conditions and addresses issues of semantic inconsistency, unnatural transitions, and style inconsistency.

3 Method

Given a background image I^b, a mask M, and a reference image I^r or $prompt$, *Spin* generates an image I^g where $(1-M)*I^g \approx (1-M)*I^b$, the content in $M*I^g$ is as close as possible to the semantic information specified in I^r or $prompt$, and the whole image looks harmonious and consistent. In the generation phase, we replace $M*z_t$ with corresponding values obtained by ddim inversion(Song, Meng, and Ermon 2020) of I^b to preserve the background. And other parts of our approach are presented in Fig.2.

3.1 Precise semantic Injection

Semantic Extraction Through Saliency Focus. To obtain salient semantic information from the reference image, we aim for the learned semantic embedding v^{img} to capture the key semantic parts and minimize interference from unwanted elements like background information. To achieve this, we enhance existing personalization methods by introducing a new loss function, referred to as the saliency focus loss $\mathcal{L}_s$:

$$\begin{aligned}\mathcal{L}_s = \|&\hat{z_0} * S + B(\hat{z_0}) * (1-S),\\ &z_0 * S + B(z_0) * (1-S)\|_2\end{aligned} \tag{1}$$

where $\hat{z_0}$ refers to the estimated clean latent given z_t and v^{img}. $B(\hat{z_0})$ includes all operations that weaken $\hat{z_0}$, and in

practice we use Gaussian blurring because it also enhances the salient regions. S is the saliency mask, which marks the salient and desired region in I^r. To compute S, we first extract the self-attention map from a specific layer of the UNet during the learning of v^{img}. Next, we average the self-attention map values across the channel dimension and sum the contributions of each latent pixel to others. The resulting values are then interpolated to match the latent vector's height h and width w, yielding a saliency map denoted as $O^{\text{i,j}}$, where $0 \leq \text{i,j} \leq w, h$. Since the self-attention map captures the relationships between latent pixels, $O^{\text{i,j}}$ reflects the saliency of each latent pixel. So it is subsequently used to define S:

$$S^{\text{i,j}} = \begin{cases} 1 & \text{if } O_{\text{t}}^{\text{i,j}} \geq \gamma \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where γ denotes the threshold, and empirically we set $\gamma = 1$ in this work.

Attention-based Region Guidance. To let the semantic condition respond only within M, we design *Attention-based Region Guidance* to achieve the goal. Technically, we don't change results of the normal self-attention operation. But we make a copy of the feature injected into it, keeping only the values within M to apply self-attention:

$$\begin{cases} f = F(z_t), \\ f' = \{f_{i,j} \mid M(i,j) = 1\} \end{cases} \tag{3}$$

We use $F(z_t)$ to denote all operations performed by the denoising network on the latent vector z_t before applying self-attention. Then the cross-attention operation is performed between it and the text embedding v. The final feature h obtained by the attention operation is:

$$h = SA(f) + CA(SA(f'), v) \tag{4}$$

where SA is self-attention and CA is cross-attention.

Our design is based on the following considerations: during cross-attention stage, semantic information is injected into a set of tokens processed by self-attention. If we only retain the cross-attention values corresponding to the latent vectors within M and discard the values outside M, only the semantic information within M is preserved, which is incomplete. In contrast, our approach ensures the complete injection of semantic information into M. Additionally, we retain the original self-attention output to preserve background details, ensuring they also influence the generation within M.

3.2 Style Guidance

During the denoising process, the background style information is already progressively injected into M through each iteration of the diffusion model. However, in SIP task, to preserve the background, we replace latent vectors outside M. Concatenating features directly at the noise level may produce outputs that deviate from the expected data distribution, potentially causing artifacts or inconsistencies(Avrahami, Lischinski, and Fried 2022), especially when I^b comes from a non-real domain. Thus, the model's inherent denoising capability is insufficient to inject background information into the mask.

We explicitly leverage the background style information to guide the generation process toward the correct manifold. To achieve this, we use the style difference between the currently predicted latent vector and the background vector as guidance. Since no model exists for computing style loss directly on latent vectors, we decode the latent vectors into pixel space to utilize existing models in that domain. Consequently, the style loss can be based on any style loss defined in pixel space, and in our experiment, we use the simplest one:

$$StyleLoss(D(\hat{z}_0), I^b) = \sum_{l=1}^{L} \|G_l(D(\hat{z}_0)) - G_l(I^b))\|^2 \tag{5}$$

where $D(\cdot)$ denotes the VAE decoder, and $G_l(\cdot)$ represents the l-th layer of style extractor. Finally, we can adjust the predicted noise by style guidance:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t) = \epsilon_\theta(\mathbf{z}_t) + s\frac{\partial StyleLoss(D(\hat{z}_0), I^b)}{\partial z_t} \tag{6}$$

$\epsilon_\theta(\mathbf{z}_t)$ represents the noise predicted by the denoising network at time t, while $\tilde{\epsilon}_\theta(\mathbf{z}_t)$ represents the final predicted noise at time t. And s is the style control coefficient that governs the influence of style guidance, which we typically set to 0.1 in our experiments. The repaint technique(Lugmayr et al. 2022) is also commonly used, although not very efficient. The inefficiency lies in adding random noise without guidance, whereas our style guidance offers clear direction. And we can retain its concept of employing guidance techniques multiple times to enhance our effect.

4 Experiments

4.1 Experimental settings

Dataset. We mainly use the dataset proposed by TF(Lu, Liu, and Kong 2023), which contains 332 sets of data from four domains: real domain, pencil drawings domain, cartoon domain and oil painting domain. In the quantitative experiments, we construct the test-set by all data from real domain where metrics are more effective. In the qualitative experiments, we enrich it by collecting images from the Internet since it lacks non-real domains images.

Baselines. We compare with the excellent open source methods mentioned in related work. In personalization of the T2I model, the improvement we propose is based on multi-tokens TI(Gal et al. 2022), so we choose basic TI, multi-tokens TI, DreamArtist and VCT for comparison. The text-conditioned SIP methods we choose for comparison include Blended, BLD, SD-inpaint, PPt and Uni. For image-conditioned SIP, we compare our method with DIB, Any, Paint, Uni and TF.

Metric. For personalizing the T2I model, we want to preserve the information of salient parts of the reference image, that is, the semantic information. So the similarity between the generated image and the reference one is important. We use CLIP$_I$(Ramesh et al. 2022) at different steps as

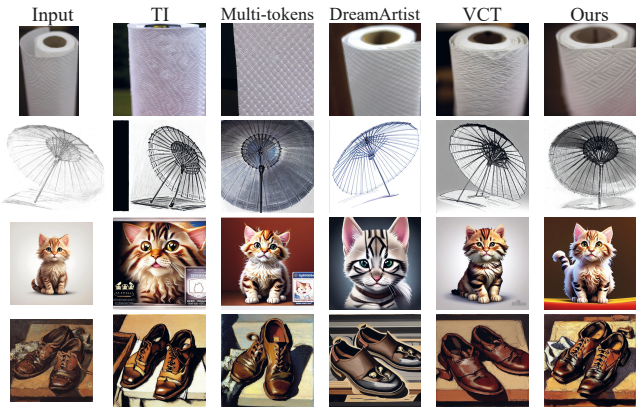

Figure 3: Comparison with existing methods for personalization of the T2I model.

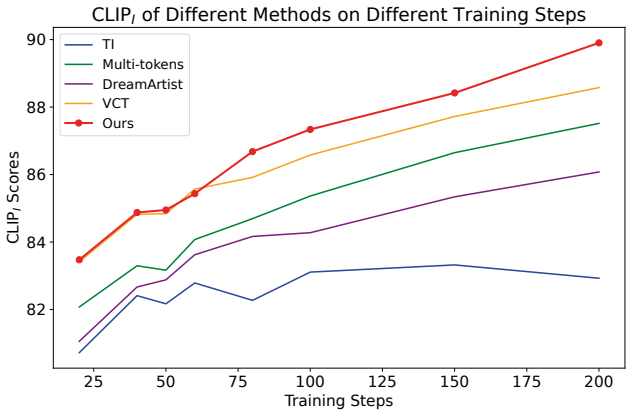

Figure 4: Quantitative analysis of feature concept learning tasks.

the evaluation metric. For SIP task, we use LPIPS$_{BG}$(Zhang et al. 2018) to reflect the perceptual similarity between the masked-out pixels of I^g and I^b. CLIP$_{I/T}$(Ramesh et al. 2022) is used to to reflect the semantic similarity between the masked-in pixels of I^g and I^r or prompt. NIMA(Talebi and Milanfar 2018) and User Score(Croitoru et al. 2023) are used to measure the harmony of the generated images, which reflects the transitions naturalness and the style consistency. The details of User Score and its preferred rate version are in the appendix.

Implementation details. Our experiments are conducted on a single V100 GPU. Most of the baselines are based on version 1.4 of Stable Diffusion(Rombach et al. 2022), so we use the same model for fairness. We adopt DDIM(Song, Meng, and Ermon 2020) with 50 steps for sampling and set the CFG scale(Ho and Salimans 2022) to 7.5. For each reference image, we perform 200 iterations using the AdamW(Kingma and Ba 2014) optimizer with a learning rate of 0.0005.

4.2 Personalization of the T2I model

Qualitative comparisons. We select data from four domains, and generate embeddings with each method using the same number of iterations. We let TI(Gal et al. 2022) optimize 6 tokens to represent the target concept, while other methods optimize all. For Dreamartist(Dong, Wei, and Lin 2022) and VCT(Cheng et al. 2023), we add their proposed pixel loss and latent loss to the original loss respectively. The results are shown in Fig. 3.

The effect of multi-tokens textual inversion is generally better than basic textual inversion, but still not similar enough to the reference image. The introduction of pixel loss is easy to learn sharper image information and distort the generated images, while our proposed $\mathcal{L}_s$ and latent loss can better learn the image information. And we don't overfit the contour information while capture more detailed semantic information.

Quantitative comparisons. The quantitative experimental results are shown in Fig. 4. Considering the training time should not be too long, we train each method for 200 steps. This allows all methods except adding pixel loss to be finished within five minutes. By the 20th iteration, all methods begin to show noticeable effects, with CLIP$_I$ steadily increasing as training progresses. Under the same training steps, the method with multi-tokens has higher scores than the pure TI. The training time is longer while CLIP$_I$ is lower after adding pixel loss, likely due to the overly strict pixel-level constraints. In comparison, our approach and VCT show better learning results. Within the first 60 iterations, our method matches the CLIPI of VCT and eventually surpasses it. By the 150th iteration, it already achieves the best performance of the existing methods at 200th iteration. $\mathcal{L}_s$ strikes a balance between effectiveness and time efficiency.

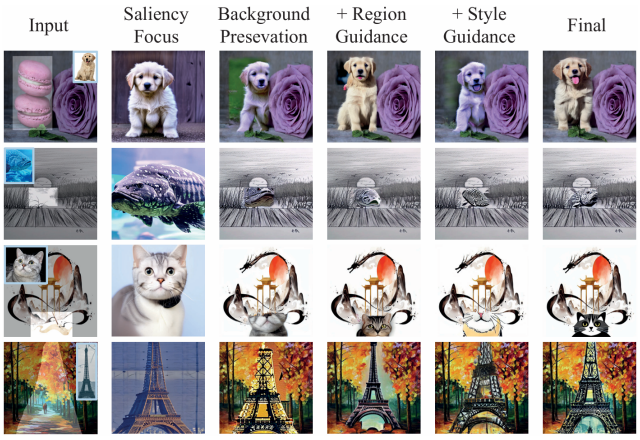

Figure 5: Ablation experiments.

4.3 Semantic image painting

Qualitative comparisons. The results for text-conditioned SIP are illustrated in Fig.6. Our method is the only one that generates harmonious, high-quality and high consistency with text across a variety of cross-domain data. Blended, BLD and Uni encounter issues with unnatural transitions, while SD-inpaint and PPt struggle with generating content that aligns with the semantic conditions. And none of the them achieve a style consistent with the background, which

Figure 6: Comparison of text-conditioned SIP methods

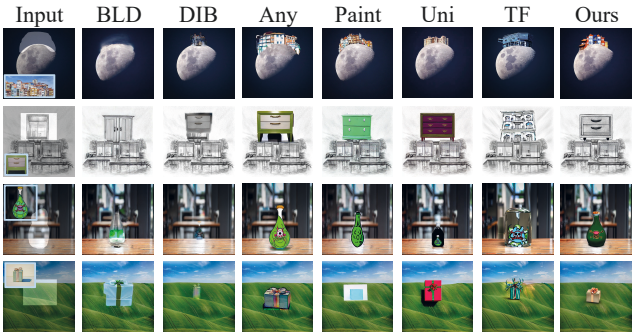

Figure 7: Comparison of image-conditioned SIP methods.

is particularly evident in the second row of Fig.6. Additionally, the image generated by Blended inside the mask contains significant noise and requires long inference times for good results. SD-inpaint struggles to respond to some of the texts, likely due to its inpainting-focused training.

For the image-conditioned SIP, as demonstrated in Fig.7, good generation results are obtained by us even when the background or reference images were from non-real domains. In contrast, Blended is difficult to generate the details clearly because it relies on text descriptions of the reference image, leading to information loss. In the results of DIB, the background color tends to bleed into the foreground, especially noticeable in cross-domain scenarios. This may be an inherent effect of methods based on the Poisson equation. Paint, Uni and TF have difficulty fully preserving the reference image's semantic information. Paint and Uni exhibit unnatural transitions, while Any, Paint, and Uni fail to incorporate the background's style information into the mask. Additionally, TF obtains a harmonious effect by changing the content inside and outside the mask simultaneously, so the preservation of the background cannot be guaranteed. And DIB, Any, and TF require a segmentation mask of the reference image, which adds extra complexity and can be challenging to achieve, especially for images with unclear edges like oil paintings.

Quantitative comparisons. For text-conditioned semantic image painting task, we conduct experiments using the same methods as for qualitative comparisons. The results are shown in Tab.1. Due to the diversity of the diffusion model, we generate 32 images for each image in Blended, and 4 images for BLD, SD-inpaint, PPt, Uni and our work. The one with highest CLIP_T selected as the result, and the obtained scores are shown in Table 1. We achieve the best results in LPIPS_{BG}, CLIP_T and NIMA. The results demonstrate the remarkable ability of our results in terms of background preservation, text coherence and harmonization.

Method	$\text{LPIPS}_{BG}\downarrow$	$\text{CLIP}_T\uparrow$	NIAM $\uparrow$
Blended	0.12	28.87	4.62
BLD	0.19	26.88	4.31
SD-inpaint	**0.06**	27.40	4.72
PPt	**0.06**	26.46	4.82
Uni	0.08	28.66	4.88
Ours	**0.06**	**29.13**	**5.03**

Table 1: Quantitative results for text-conditioned SIP.

Method	$\text{LPIPS}_{BG}\downarrow$	$\text{CLIP}_I\uparrow$	$\text{CLIP}_T\uparrow$	$S^{User}\uparrow$
BLD	0.11	73.25	25.19	5.77
DIB	0.11	77.57	26.84	6.42
Paint	0.13	80.26	25.92	6.22
Uni	0.09	80.97	27.54	6.92
TF	0.10	82.86	28.11	7.18
Ours	**0.06**	**83.70**	**28.80**	**7.27**
cp*	0.01	87.52	29.39	/

Table 2: Quantitative results for image-conditioned SIP. *: copy and paste result.

For image-conditioned SIP, we evaluate the same methods as in the qualitative experiments except for Any. For our work and other work that produces multiple results, we choose the one with highest CLIP_I as the result and report the mean ± std deviation in the appendix. We paste the result of the reference image with the segmentation mask to the relatively appropriate region in M as the upper bound of the metrics. The obtained scores are shown in Table 2. We achieved the best results in the LPIPS_{BG}, CLIP_T, CLIP_I and S^{User}. We achieve sufficient retention of background information, accurate reflection of semantic conditions, and high quality of image generation.

Config	$\text{LPIPS}_{BG}\downarrow$	$\text{CLIP}_I\uparrow$	$\text{CLIP}_T\uparrow$	NIMA $\uparrow$
Saliency Focus	0.63	73.80	26.25	4.87
+Background	0.08	77.14	27.66	4.79
+ARG	**0.06**	82.66	28.02	5.03
+Style Guidance	0.08	78.25	27.69	4.88
Final	**0.06**	**83.70**	**28.80**	**5.07**

Table 3: Ablation study: quantitative comparison of various variants of our framework.

Ablation study. Finally, each sub-part of the method is ablated, as shown in Fig.5 and Tab.3. We use the results ob-

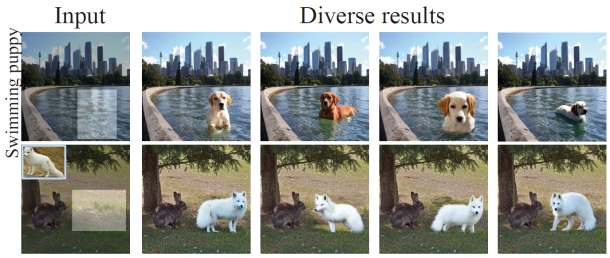

Figure 8: Qualitative results about diversity of Spin.

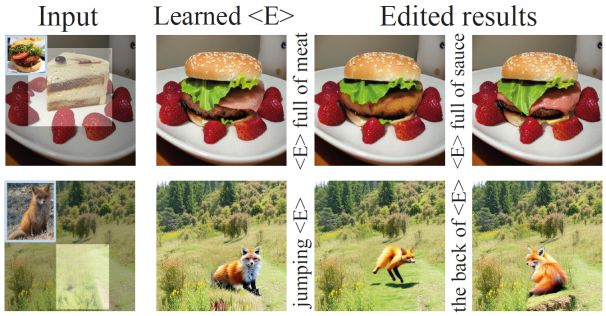

Figure 9: Qualitative results about editability of Spin.

tained by denoising from the initial vector z_T^* and v as our baseline. Since we get v using our proposed $\mathcal{L}_s$, it reflects the salient semantic information of the reference image, so we denote the results as Saliency Focus. When the background is preserved, only partial semantic information is injected into the mask. By adding *Attention-based Region Guidance*, the full semantic information can be injected into M, resulting in a higher CLIP$_I$. Furthermore, incorporating style guidance enhances the harmony between the object in the mask and the background, leading to a higher NIMA score. When both techniques are applied simultaneously, we can obtain satisfactory results.

4.4 Diversity, editability and generality

Since Spin is based on diffusion model and projects the reference image into text embedding space, it can generate diverse and editable results. We compare the diversity and editability of the existing methods for SIP in Tab.4, and the qualitative results are shown in Fig.8 and Fig.9 respectively.

Abilities	BLD	Paint	DIB	TF	Any	PPt	Uni	Ours
Diversity	✔	✔	✗	✗	✔	✔	✔	✔
Editability	✗	✗	✗	✗	✗	✗	✗	✔

Table 4: Diversity and Editability of existing methods

Since our design doesn't depend on a specific model, it can be applied to any framework that satisfies our setting. In order to prove that $\mathcal{L}_s$ is also effective for other methods about personalization of T2I models, we choose the open source work NeTI(Alaluf et al. 2023) to show(Fig.10). While NeTI can already learn the general information of the reference image very well, $\mathcal{L}_s$ improves its ability to capture semantic details in each domain. The ARG and style guidance can be transferred to other generative models that alternately use self-attention and cross-attention block to inject semantic information into the background independently. We select PixArt(Chen et al. 2023a), a model that meets these requirements, to demonstrate the generality(Fig.11).

Figure 10: Qualitative results about generality of $\mathcal{L}_s$.

Figure 11: Qualitative results about generality of Spin.

5 Conclusion

This paper proposes a novel framework, Spin, for semantic image painting. Spin filters out non-semantic information like background during the extraction of semantic embedding from the reference image and precisely injects this information into the masked region during semantic painting. This enables the generated image to be semantically complete while harmonizing with the background. Additionally, by imposing a style consistency constraint, the generated part further aligns with the style of the background image. Experiments demonstrate that our method can generate style-consistent images that harmonize with the background even when the reference image and the target image have different styles, such as realistic, pencil drawing, cartoon, and oil painting. Spin generates diverse and editable results, and is model-independent thus can be applied to other compatible frameworks.

Acknowledgments

This work is supported by the Project from Shenzhen Science and Technology Innovation Commission (KJZD20230923114810022).

References

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

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

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

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

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. 2023a. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

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

Cheng, B.; Liu, Z.; Peng, Y.; and Lin, Y. 2023. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22736–22746.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, Z.; Wei, P.; and Lin, L. 2022. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. 5-An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

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

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Niu, L.; Cong, W.; Liu, L.; Hong, Y.; Zhang, B.; Liang, J.; and Zhang, L. 2021. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*.

Pérez, P.; Gangnet, M.; and Blake, A. 2023. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 577–582.

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

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

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Ruiz, N.; Li, Y.; Wadhwa, N.; Pritch, Y.; Rubinstein, M.; Jacobs, D. E.; and Fruchter, S. 2024. Magic Insert: Style-Aware Drag-and-Drop. *arXiv preprint arXiv:2407.02489*.

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

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

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

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

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

Zhang, L.; Wen, T.; and Shi, J. 2020. Deep image blending. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 231–240.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2023. A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting. *arXiv preprint arXiv:2312.03594*.