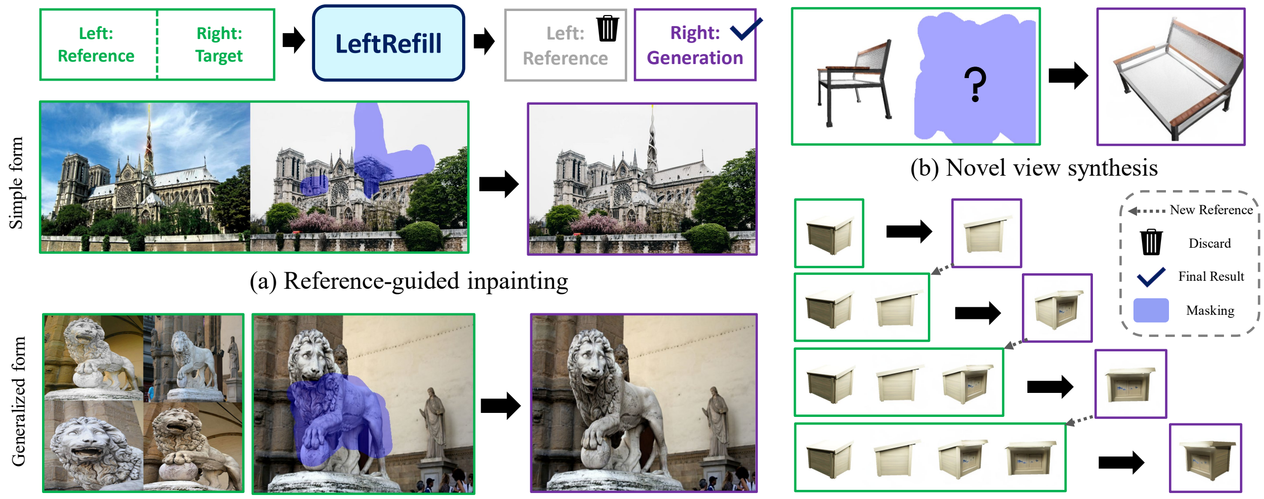
LeftRefill: Filling Right Canvas based on Left Reference through Generalized Text-to-Image Diffusion Model

Chenjie Cao[1,2]*, Yunuo Cai[1], Qiaole Dong[1], Yikai Wang[1], Yanwei Fu[1]†
[1]Fudan University, [2]Alibaba Group



(a) Reference-guided inpainting

(b) Novel view synthesis

(c) Inpainting one target view through multiple references

(d) Generating multiple targets from single view

Figure 1. LetRefill addresses the generation on the right canvas conditioned by left references. We can re-formulate several existing tasks in the LeftRefill manner, including (a) reference-guided inpainting, (b) novel view synthesis. The reference and target can be further extended to multi-view scenes, forming (c) multi-view reference inpainting and (d) multi-view synthesis, respectively. **Green** frames indicate stitched inputs. Reference views are placed on the left side, while masked target views are placed on the right side. **Violet** frames only show enlarged right-side generations produced by LeftRefill. Note that we omit some input details in (c) and (d) for simplicity.

Abstract

This paper introduces LeftRefill, an innovative approach to efficiently harness large Text-to-Image (T2I) diffusion models for reference-guided image synthesis. As the name implies, LeftRefill horizontally stitches reference and target views together as a whole input. The reference image occupies the left side, while the target canvas is positioned on the right. Then, LeftRefill paints the right-side target canvas based on the left-side reference and specific task instructions. Such a task formulation shares some similarities with contextual inpainting, akin to the actions of a human painter. This novel formulation efficiently learns both structural and textured correspondence between reference and target without other image encoders or adapters. We inject task and view information through cross-attention modules in T2I models, and further exhibit multi-view reference ability via the re-arranged self-attention modules. These enable LeftRefill to perform consistent generation as a generalized model without requiring test-time fine-tuning or model modifications. Thus, LeftRefill can be seen as a simple yet unified framework to address reference-guided synthesis. As an exemplar, we leverage LeftRefill to address two different challenges: reference-guided inpainting and novel view synthesis, based on the pre-trained StableDiffusion. Codes&models are released at https://github.com/ewrfcas/LeftRefill.

1. Introduction

Imagine you are a right-handed painter with a task that requires you to draw or modify a target image based on a reference picture. How would you approach it? Intuitively,

[*] Dr. Chenjie Cao is at Alibaba Group. This work was accomplished while Dr. Chenjie Cao was at Fudan University.
[†] Corresponding author: yanweifu@fudan.edu.cn
Other emails: {cjcao20,yncai20,qldong18,yikaiwang19}@fudan.edu.cn

you would likely place the reference image on your left side and paint or modify the target view on the right side conditioned on the left one. If we regard the large Text-to-Image (T2I) models [6, 37, 41, 44, 46, 48] as skillful painters, could they also follow such simple and intuitive task formulation to handle complex reference-guided synthesis tasks? In this paper, we explore this problem with two challenging tasks: 1) Inpainting masked target views conditioned on reference images, *i.e*., reference-guided inpainting (Ref-inpainting) [67, 68] as in Figure 1(a). 2) Generating new views based on known images of specific objects, *i.e*., Novel View Synthesis (NVS) [30] as in Figure 1(b).

It seems straightforward to harness the power of T2I generative models to directly address these image-reference tasks by training additional adapters [21, 36, 64] or replacing textual encoders with visual ones [30, 62] and optimize them for full fine-tuning of the entire T2I model. We should clarify that training these large T2I models with unfamiliar visual encoders is computationally intensive and challenging to converge, particularly when working with limited batch sizes. Additionally, most visual encoders, such as image CLIP [43], tend to emphasize the learning of semantic features rather than the intricately spatial details that are essential for tasks involving Ref-inpainting as verified in our experiments.

To prevent the heavy modification to T2I models mentioned above, we rethink the human painting habit and propose LeftRefill, a unified approach for both Ref-inpainting and NVS. LeftRefill is built upon the inpainting fine-tuned StableDiffusion2.0 (SD) [46][*], which ingeniously reformulates reference-based synthesis as a contextual inpainting process. Specifically, LeftRefill horizontally stitches reference and target views as a whole input. As shown in Figure 1, reference images are positioned on the left side, while masked targets are on the right side. This simple yet effective formulation eliminates the dependency on additional image feature encoders, as both reference and target views have been stitched into the same canvas. As an "experienced painter", LeftRefill is driven by specific instructions, called task and view prompt tuning. These instructions infuse crucial information for specific generative tasks and desired view orders to cross-attention modules, guiding the generation of SD. Note that LeftRefill is a generalized framework, so we can train two LeftRefill models for Ref-inpainting and NVS separately without any test-time fine-tuning.

On the other hand, we emphasize that LeftRefill further enjoys extension into multi-view synthesis scenarios, including image inpainting conditioned on multi-view references in Figure 1(c) and consistent NVS from a single view in Figure 1(d). Formally, we rearrange the tensor shape before the self-attention modules, enabling self-attention to capture cross-view information. Moreover, to tackle the more intricate NVS task, we propose the novel technique of block causal masking, facilitating self-attention-based T2I models in achieving consistent autoregressive (AR) generation. All improvements are integrated into the off-the-shelf SD components, involving no additional image encoder, prompt tuning based on cross-attention, self-attention rearranging and block causal masking. LeftRefill emerges as a unified architecture for addressing reference-guided synthesis, requiring only minor model modifications.

We highlight the key contributions of LeftRefill as follows: 1) *Lightweight and generalized task formulation based on off-the-shelf T2I models:* Benefiting from the novel contextual inpainting formulation and inherent attention mechanisms from SD, LeftRefill provides an efficient solution for reference-guided synthesis without thoroughly re-training T2I models and test-time fine-tuning. 2) *Task and view-specific prompt tuning:* Our work pioneers the use of task and view-specific prompt tuning, allowing for precise control over generative tasks and view orders. 3) *End-to-end Ref-inpainting:* Notably, our LeftRefill addresses the challenging Ref-inpainting end-to-end, without complex 3D geometrical warping and 2D inpainting techniques [65, 67, 68]. 4) *Autoregressive NVS with block causal masking*: For the intricate NVS task, we introduce the novel concept of block causal masking, enabling self-attention-based T2I models to achieve AR generation for geometric consistency.

2. Related Work

Personalization and Controllability of T2I Models. Recent achievements on T2I have produced impressive visual generations [2, 36, 42], which could be further extended into local editing [1, 8, 18]. However, these models could only be controlled by natural languages. As "an image is worth hundreds of words", T2I models based on natural texts fail to produce images with specific textures, locations, identities, and appearances [15]. Textual inversion [15, 35] and fine-tuning techniques [47] are proposed for personalized T2I. Meanwhile, many works pay attention to image-guided generations [26, 34, 58]. ControlNet [64] and T2I-Adapter [36] learn trainable adapters [20] to inject visual clues to pre-trained T2I models without losing generalization and diversity. But these moderate methods only work for simple style transfers. More spatially complex tasks, such as Ref-inpainting, are difficult to handle by ControlNet as verified in Section 4. In contrast, T2I-based exemplar-editing and NVS have to be fine-tuned on large-scale datasets with strong data augmentation [62] and large batch size [30]. Compared with these aforementioned manners, LeftRefill enjoys both spatial modeling capability and computational efficiency.

Prompt Tuning [24, 32, 33] indicates fine-tuning token embeddings for transformers with frozen backbone to preserve the capacity. Prompt tuning is first explored for adaptively learning suitable prompt features for language models rather than manually selecting them for different downstream

[*] https://github.com/Stability-AI/stablediffusion.

tasks [29]. Moreover, prompt tuning has been further investigated in vision-language models [16, 43] and discriminative vision models [23, 28]. Visual prompt tuning in [53] prepends trainable tokens before the visual sequence for transferred generations. Though both LeftRefill and [53] aim to tackle image synthesis, our prompt tuning is used for controlling text encoders rather than visual ones. Thus LeftRefill enjoys more intuitive prompt initialization from task-related textual descriptions.

Reference-guided Image Inpainting. Image inpainting is a long-standing vision task, which aims to fill missing image regions with coherent results. Both traditional methods [3, 9, 17] and learning-based ones [11, 25, 54, 63, 66] achieved great progress in image inpainting. Furthermore, Ref-inpainting requires recovering a target image with one or several reference views from different viewpoints [40], which is useful for repairing old buildings or removing occlusions in popular attractions. But Ref-inpainting usually suffers from a sophisticated pipeline [65, 67, 68], including depth estimation, pose estimation, homography warping, and single-view inpainting. Limited by large holes, the estimated geometric pose is not reliable, largely degrading the pipeline. Thus an end-to-end Ref-inpainting pipeline is meaningful. To the best of our knowledge, we are the first ones to tackle such a difficult reference-guided task with T2I models.

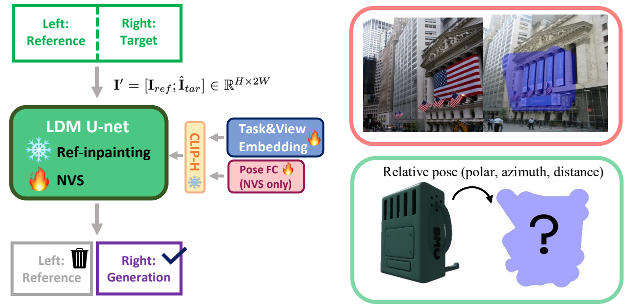
Novel View Synthesis from a Single Image. NVS based on a single image is an intractable ill-posed problem, requiring both sufficient geometry understanding and expressive textural presentation [14]. Many previous works could partially tackle this problem through single view 3D reconstruction [7, 31, 59–61], 2D generative models [39, 45, 52], feature cost volumes [5], and GAN-based methods [4, 38, 51]. However, these manners still suffer from limited generalization or small angle variations. To address this issue, Zero123 [30] uses another image CLIP encoder to inject image features from the reference view to unlock the capacity from 2D diffusion-based T2I models for NVS. But Zero123 requires a large batch size and expensive computational resources to stabilize the training stage with an unknown reference image encoder. Moreover, the image encoder in Zero123 can only tackle one reference image, which fails to generate consistent multi-view images.

3. Method

Overview. LeftRefill is first formulated in Section 3.1. Then we explain using self-attention to capture multi-view correspondence, and AR generation based on block casual masking (Section 3.2). Finally, we discuss the task and view-specific prompt tuning for cross-attention (Section 3.3).

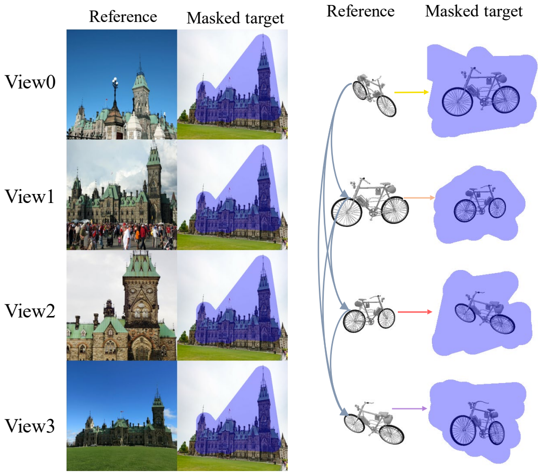
3.1. Framework of LeftRefill

Motivations. As depicted in Figure 2(a), our LeftRefill is built upon the inpainting fine-tuned SD [46]. There are two primary motivations that make us stitch reference and target images together and reformulate both Ref-inpainting and NVS as a contextual image inpainting problem. 1) LeftRefill just considers a single input image, eliminating the requirement of additional image encoders and avoiding major architectural alterations and extensive re-training. 2) Since all T2I models are only pre-trained on single-view images, the left-right stitched input could implicitly reactivate the *essential capacity from large T2I models of modeling the correlation of the single-view image*. Particularly, the left-right stitching input enables the self-attention modules to focus on correct parts from left-side reference at the early stage of diffusion sampling (Figure 10). We also thoroughly evaluate different alterations of reference-guided SD in Section 4.1. LeftRefill substantially outperforms other competitors with high efficiency and negligible trainable weights.



(a) Intuitive overview of LeftRefill (b) Input of Ref-inpainting and NVS

Figure 2. (a) The overview of LeftRefill. Inputs of Ref-inpainting and NVS are shown in (b). Task and view prompt embedding and pose features (optional for NVS) are infused to CLIP-H for cross-attention learning in U-net. For the output of LeftRefill, we discard the left-side reference and take the right-side generation.



(a) Ref-inpainting (v-to-1) (b) NVS (n-to-1, n=1,2,…,v)

Figure 3. Illustration about multi-view training inputs ($v\times H\times 2W$, $v=4$) of LeftRefill, where $v, H, 2W$ indicate the view number, height, and width of stitching images. All views of Ref-inpainting (a) share the same masked target, while the multi-view NVS (b) should be trained with the AR generation.

Single-view Formulation (1-to-1). Thanks to the convo-

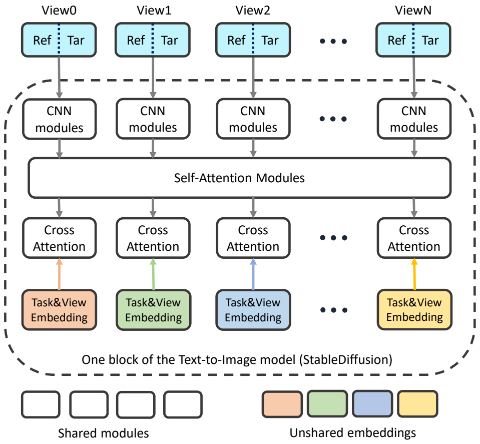
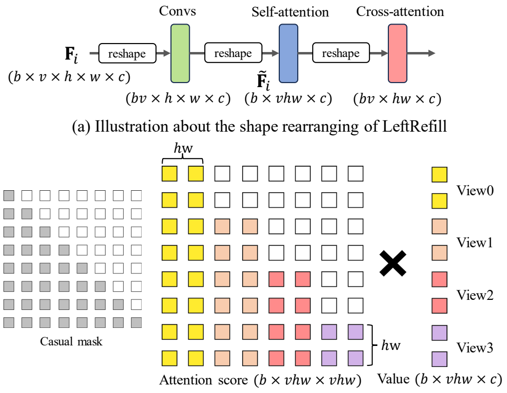


Figure 4. Detailed architecture of LeftRefill for multi-view synthesis. Both CNN and cross-attention modules are encoded separately for each stitched view, while all views share the same self-attention for multi-view correlation learning.


Convs
Self-attention
Cross-attention
$\mathbf{F}_i$
reshape
$(b \times v \times h \times w \times c)$
$(bv \times h \times w \times c)$
$\tilde{\mathbf{F}}_i$
$(b \times vhw \times c)$
$(bv \times hw \times c)$
hw
Casual mask
×
View0
View1
View2
View3
Attention score $(b \times vhw \times vhw)$
Value $(b \times vhw \times c)$

(a) Illustration about the shape rearranging of LeftRefill

(b) Block causal masking used for autoregressive training

Figure 5. (a) Feature rearranging, and (b) block causal masking of LeftRefill, where b, v, h, w, c indicate the batch size, view number, height, width, and channels of features, where w is the width of stitching features (downsampled from $2W$).

lutional U-net architecture in the Latent Diffusion Model (LDM), we can expand the input image in the spatial dimensions without any modification. Let's consider a scenario with a single reference image at first, *i.e.*, 1-to-1. Our input $\mathbf{I}'$ is a stitching image of $\mathbf{I}_{ref}$ and the masked target $\hat{\mathbf{I}}_{tar}$, forming as $\mathbf{I}' = [\mathbf{I}_{ref}; \hat{\mathbf{I}}_{tar}] \in \mathbb{R}^{H \times 2W}$ as shown in Figure 2(a). In practice, we take the reference image on the left side, while the target one is placed on the right side. We just take the right-side output as the final generation while the left-side output is discarded. Note that the diffusion optimization is based on the whole stitched image without any modification. For the masked target $\hat{\mathbf{I}}_{tar}$, we treat Ref-inpainting's input targets as *partially masked* images, while NVS's input targets are *entirely masked* through the objective segmentation and bounding box. More details about the processing of data and masks are discussed in the supplementary.

Multi-view Formulation (V-to-1). For the multi-view references, we stitch each reference with the specific target as shown in Figure 3. All views are learned separately for convolutions and cross-attention, while they share the same self-attention processing as shown in Figure 4 and detailed in Section 3.2. From Figure 3(a), the multi-view Ref-inpainting leverages information from different reference views to repair the same target, *i.e.*, v-to-1, where v means the view number. We simply take the generation of the first view as the final inpainted output in v-to-1. For the multi-view NVS, it could be seen as an AR generation for sequentially consistent view synthesis, as depicted in Figure 3(b), *i.e.*, n-to-1 (n=1,2,...,v). During the inference, we apply the generated targets as new references for the subsequent view synthesis, while the training phase is accomplished parallelly as detailed in the block casual masking of Section 3.2.

Controlling Generation for LeftRefill. Although the self-attention modules in SD have the potential to enable the correlation between left reference and right target, SD is not trained to explicitly activate this capacity. The most intuitive way to guide the SD in capturing the correlation among left-right stitched images is to apply suitable text prompts to drive the diffusion model for the desired generation. However, it is non-trivial to define Ref-inpainting and NVS with natural languages. Furthermore, it is beneficial to have a non-instance-level prompt to generally guide the diffusion model to accomplish specific tasks. To this end, we propose to use prompt tuning to learn task and view-specific prompts as detailed in Section 3.3. Except for the prompt tuning, all weights in LDM are frozen in Ref-inpainting to maintain the proper generalization as shown in Figure 2(a). For NVS, we fine-tune the whole LDM to achieve essentially precise pose control, but LeftRefill enjoys much better convergence compared with other fine-tuning based methods [30].

3.2. Reactivating Self-Attention for Multi-View

As shown in Figure 5(a), given multi-view features $\mathbf{F}_i$ of layer i, all MLP, convolutional, and cross-attention layers encode $\mathbf{F}_i$ separately. We can simply achieve the separate feature encoding by reshaping the view v and batch b dimension together as bv. Before the self-attention, we rearrange the feature shape as $\tilde{\mathbf{F}}_i \in \mathbb{R}^{b \times vhw \times c}$, thus features across different views could be learned together. To further incorporate positional clues for distinguishing different sides of reference and target in NVS, we incrementally add positional encoding P_i to $\tilde{\mathbf{F}}_i$ before each self-attention block as

$$P_i = \gamma_i \cdot \text{cat}([P_v; P_{Fourier}]), \tag{1}$$

where $P_v, P_{Fourier}$ indicate the trainable view embedding and Fourier absolute positional encoding [57] respectively; γ_i is a zero-initialized learnable coefficient for each layer.

For the Ref-inpainting, no masking strategy should be considered in self-attention modules. All reference views share the same target one, thus it is unnecessary for LeftRefill to sequentially repair target views. In contrast, for the multi-view NVS, generating consistent novel views from a single image needs our model to handle the sequential generation with dynamic reference views. For example, the same LeftRefill should accomplish the NVS from one view, two views, and even more. So the AR generation [13, 49, 56] is suitable to formulate this task.

Block Causal Masking. LeftRefill requires a certain fine-tuning for LDM to effectively tackle challenging NVS as shown in Figure 2(a). The intuitive solution is to train an AR-based generative model that can generalize across various view numbers for multi-view synthesis. Converting a pre-trained diffusion model to an AR-based generative model is non-trivial. However, the inpainting formulation of LeftRefill makes this conversion feasible. Specifically, we just need to adjust the masking strategy during the self-attention learning. We propose the block casual masking as shown in Figure 5(b), while the block side-length of each view is hw, matching the size of the stitched reference and target pair. Different from the traditional casual mask which is a lower triangular matrix, the block casual mask enlarges the minimal unit from one token to a $hw \times hw$ block, ensuring reasonable block-wise receptive fields. In practice, all uncolored tokens in the attention score are masked with "$-\inf$" before the softmax operation. The block casual mask can be implemented parallelly and efficiently as in supplementary.

3.3. Task&View Prompt Tuning

The prompt embedding is adopted as the textual branch to CLIP-H [43] of SD, being applied to cross-attention as shown in Figure 2(a). Specifically, we prepare a set of trainable text embeddings for different generative tasks, which are further categorized into task and view prompts. Specifically, task prompt embeddings are shared in the same task, *e.g.*, all views of Ref-inpainting using the same task embeddings. In contrast, different view prompt embeddings are applied to inject various view-order information through cross-attention modules to the specific input view. Though there are only a few trainable parameters (0.05M to 0.065M), we astonishingly find that prompt tuning is sufficient to drive complex generative tasks such as Ref-inpainting, even with a frozen LDM backbone. The trainable task and view prompt embeddings p_t, p_v are initialized as the averaged embedding of the natural task description. The optimization target is:

$$\{p_t, p_v\}_* = \mathop{\arg\min}_{\{p_t,p_v\}} \mathbb{E}\left[\left\|\varepsilon - \varepsilon_\theta\left([z_t;\hat{z}_0;\mathbf{M}], c_\phi(p_t,p_v), t\right)\right\|^2\right], \quad (2)$$

where $\varepsilon_\theta(\cdot)$ is the estimated noise by LDM; $c_\phi(\cdot)$ means the frozen CLIP-H; z_t is a noisy latent feature of input z_0 in step t; $\hat{z}_0 = z_0 \odot (1 - \mathbf{M})$ are masked latent features concatenated to z_t with mask $\mathbf{M}$. Task and view-specific prompt tuning enjoy not only training efficiency but also lightweight saving [24]. For example, we share the same LeftRefill to address Ref-inpainting with different reference views, while only 0.01% additional weights of $\{p_t, p_v\}_*$ are changed for each view condition. In NVS, we further provide relative poses to LeftRefill. Following [30], we calculate the 4-channel relative pose in the polar coordinate for each view, which is encoded by a two-layer FC. Then the pose feature replaces the last padding token in the prompt embeddings before being applied to the CLIP-H.

Table 1. Quantitative results for Ref-inpainting on MegaDepth [27] test set based on matching and manual masks (upper: 1-view; lower: multi-view). 'ExParams': the scale of extra trainable parameters. * means that the uncorrupted ground truth is visible for the matching. 'No stitching': reference and target views are separate without spatial stitching, and only self-attentions are learned across them.

Methods	PSNR↑	SSIM↑	FID↓	LPIPS↓	ExParams
SD (inpainting) [46]	19.841	0.819	30.260	0.1349	+0
ControlNet [64]	19.072	0.744	33.664	0.1816	+364.2M
ControlNet+NewCrossAttn	19.027	0.743	34.170	0.1805	+463.4M
ControlNet+Matching* [55]	20.592	0.763	29.556	0.1565	+364.3M
Perceiver+ImageCLIP [22]	19.338	0.745	32.911	0.1751	+52.01M
Paint-by-Example [62]	18.351	0.797	34.711	0.1604	+865.9M
TransFill [68]	**22.744**	**0.875**	26.291	0.1102	–
LeftRefill (no stitching)	20.489	0.827	$\underline{20.125}$	$\underline{0.1085}$	+0.05M
LeftRefill	$\underline{20.926}$	$\underline{0.836}$	**18.680**	**0.0961**	+0.05M
LeftRefill (2-view)	21.092	0.836	18.389	0.0969	+0.055M
LeftRefill (3-view)	21.356	0.840	16.838	0.0901	+0.06M
LeftRefill (4-view)	**21.779**	**0.847**	**16.632**	**0.0839**	+0.065M

4. Experiments

Datasets. For Ref-inpainting, we use the resized 512×512 image pairs from MegaDepth [27], which includes many multi-view famous scenes collected from the Internet. To trade-off between the image correlation and the inpainting difficulty, we empirically retain image pairs with 40% to 70% co-occurrence with about 80k images and 820k pairs. The validation of Ref-inpainting also includes some manual masks from ETH3D scenes [50] to verify the generalization. For the NVS, we use Objaverse [10] rendered by [30] including 800k various scenes with object masks. We resize all images to 256×256 as [30]. Note that some extreme views with elevation angles less than -10° are filtered due to excessive ambiguity. More details about the masking and datasets are introduced in the supplementary.

Implementation Details. Our LeftRefill is based on the inpainting fine-tuned SD [46] with 0.8 billion parameters. For the task and view prompt tuning, there are 50 trainable prompt tokens at all. We use 90% tokens to represent the task embeddings, while 10% tokens indicate each view respectively. We use the AdamW optimizer with a weight decay of 0.01 to optimize LeftRefill. For the Ref-inpainting, the prompt tuning's learning rate is 3e-5. Moreover, 75% masks are randomly generated, and 25% of them are matching-based masks. For the NVS, LeftRefill could be tested with the adaptive masking strengthened by the foreground segmentation model. Concretely, we first enlarge the reference segmentation mask and generate a coarse target with fewer DDIM steps. We extract the new mask from the coarse tar-

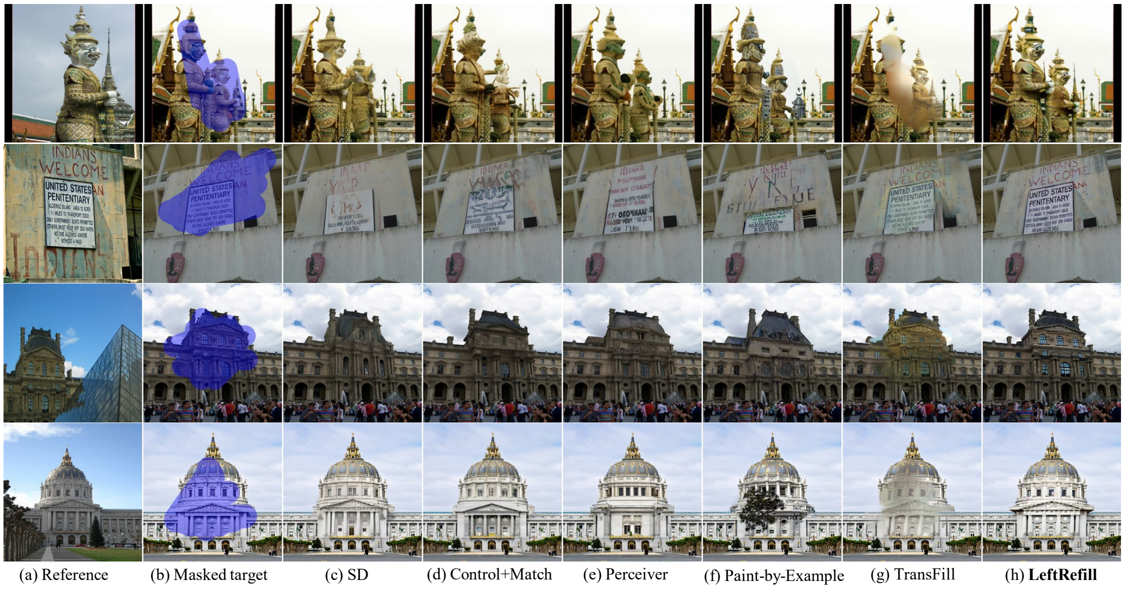


Figure 6. Qualitative Ref-inpainting results on MegaDepth [27]. More results are in the supplementary.

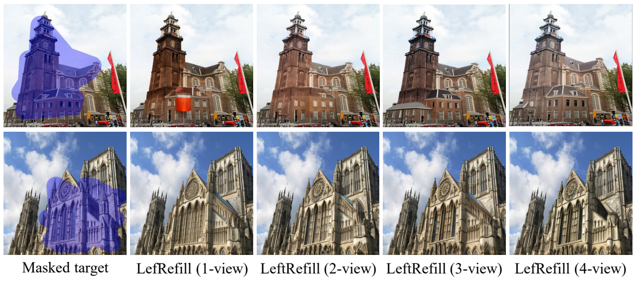


Figure 7. Multi-view Ref-inpainting qualitative results. Increasing the reference view number improves the quality of repaired targets.

get and then further randomly enlarge it as the final target mask. The NVS LeftRefill is trained with 512 batch size and learning rate 3e-5. We show that a simplified LeftRefill with just 48 batch size can also be converged for NVS, which breaks through the limitation of Zero123 [30]. More details about matching-based masks, adaptive masking, and training schedules are discussed in the supplementary.

4.1. Results of Reference-guided Inpainting

Results of One-view Reference. We thoroughly compared the specific Ref-inpainting method [68] and existing image reference-based variants of SD with one-view reference in Table 1 and Figure 6. Note that ControlNet [64] fails to learn the correct spatial correlation between reference images and masked targets, even enhanced with trainable cross-attention learned between reference and target features. Furthermore, we try to warp ground-truth latent features with image matching [55] as the reference guidance for ControlNet, but the improvement is not prominent. Perceiver [22] and Paint-by-Example [62] align and learn image features from Image CLIP. Since image features from CLIP contain high-level semantics, they fail to deal with the fine-grained Ref-inpainting as shown in Figure 6(e)(f). Though TransFill [68] achieves proper results in PSNR and SSIM, it suffers from blur and color difference as in Figure 6(g) with challenging viewpoints. LeftRefill enjoys substantial advantages in both qualitative and quantitative comparisons with negligible trainable weights. Particularly, spatially stitching reference and target views together achieves consistent improvements. Thus it is intuitive and convincing that all U-net modules contribute to improved inpainting results with a whole stitched image, as

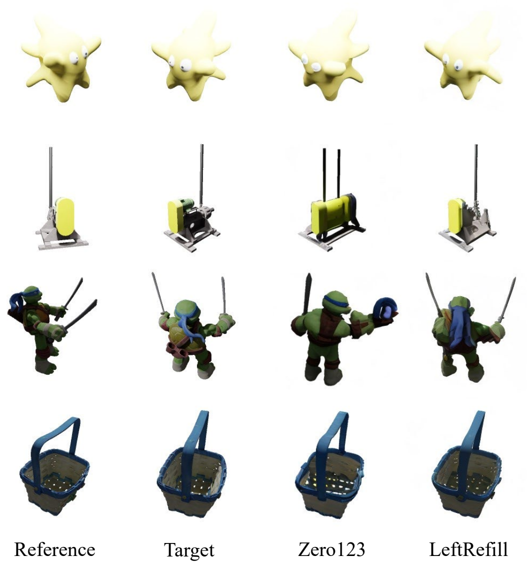


Figure 8. NVS results on Objaverse [10] (row1, 2) and Google Scanned Objects [12] (row3, 4) from a single reference image.

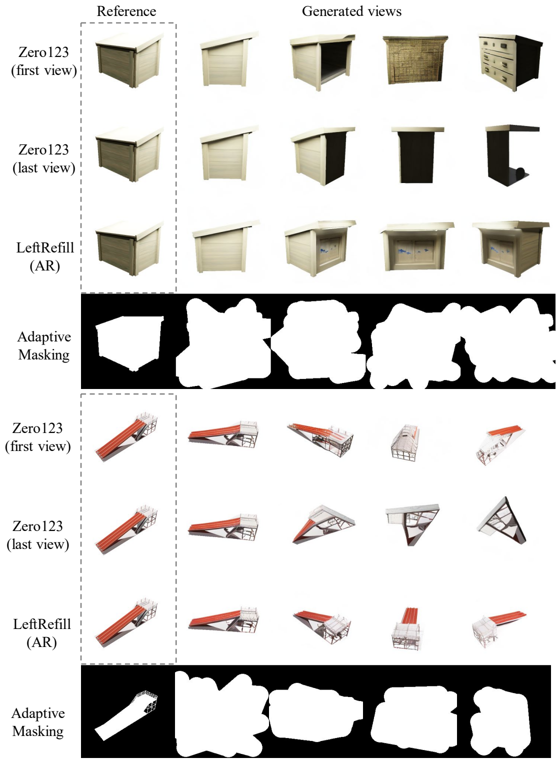


Figure 9. Sequential generative results from a single view. Zero123 [30] are conditioned on real reference (first view) and last generated view (last view), while LeftRefill is based on AR.

opposed to using only attention modules. We further compare LeftRefill with TransFill on their officially provided real-world dataset in the supplementary.

Results of Multi-view Ref-inpainting. We verified models trained with different numbers of reference views in Table 1 (lower). As the number of reference views increases, there is a notable enhancement in inpainting capability. As qualitatively compared in Figure 7, more consistent references lead to robust inpainting results with sensible structures.

Table 2. Results of 1-view NVS conditioned on different numbers of reference views on Objaverse [10]. 'w.o. AM' indicates NVS results without Adaptive Masking (AM).

Methods	Ref-View	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
Zero123 [30]	1	19.402	0.858	0.1309	0.7816
LeftRefill (LoRA)	1	19.514	0.869	0.1534	0.7589
LeftRefill (w.o. AM)	1	21.675	0.887	0.1089	0.7959
LeftRefill	1	**21.404**	**0.882**	**0.1151**	**0.7972**
LeftRefill	2	22.935	0.895	0.0871	0.8280
LeftRefill	3	24.107	0.908	0.0722	0.8432
LeftRefill	4	**24.685**	**0.911**	**0.0634**	**0.8495**

Table 3. Results of 4-view NVS generations based on 1 reference view on Objaverse [10]. P-CLIP means pairwise CLIP score showing consistency of generated views. The reference (Ref) can be categorized into the first ground-truth view and the last generated view, while we also provide the AR results of LeftRefill.

Methods	Ref	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	P-CLIP↑
Zero123	First	19.265	0.855	0.1366	0.7723	0.7756
Zero123	Last	14.621	0.767	0.2569	0.6921	0.7667
LeftRefill	First	**21.573**	**0.883**	**0.1143**	**0.7964**	0.7709
LeftRefill	AR	21.271	0.882	0.1195	0.7882	**0.7958**

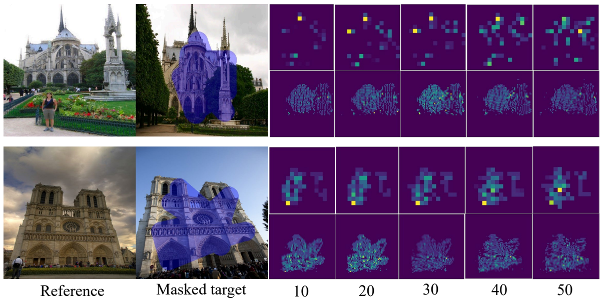


Figure 10. Visualization of attention scores in LeftRefill for Ref-inpainting across different DDIM steps. We show scores from reference views attended by masked regions. The upper row shows attention scores from the 8th self-attention (1/32 scale), while the bottom row shows ones from the 14th self-attention (1/8 scale).

4.2. Results of Novel View Synthesis

Results of Single-view NVS. We compare the quantitative NVS results on Objaverse in Table 2. The qualitative results for both Objaverse and Google Scanned Objects are compared in Figure 8. Without specific annotations, all NVS results are based on adaptive masking. The CLIP score [43] is compared to evaluate the similarity between the generation and the target. Specifically, LeftRefill fine-tuned with the whole LDM enjoys substantial achievements compared to the state-of-the-art competitor Zero123 [30], even though Zero123 might have seen our validation in Objaverse. Moreover, the LoRA-based LeftRefill [21] is still competitive with a very moderate training setting (batch size 48 with 2 A6000 GPUs) in Table 2. We further compare the training log of LeftRefill and Zero123 in the supplementary to investigate the training convergence for NVS. The contextual inpainting-based LeftRefill enjoys a substantially faster convergence and superior image quality. So LeftRefill enjoys a good balance between training efficiency and performance.

Results of Multi-view NVS. Quantitative results of multi-view NVS are shown in the lower of Table 2. Obviously, more reference views lead to better reconstruction quality of LeftRefill. Moreover, additional reference images could substantially alleviate the ambiguity, improving the final results with consistent geometry. Benefited by AR, LeftRefill can be also generalized to synthesize a group of consistent images with different viewpoints from a single view as shown in Figure 9 and Table 3. We introduce the pairwise CLIP score (P-CLIP) to verify the consistency of all generated samples. LeftRefill outperforms Zero123 in most metrics, while AR could prominently improve the consistency with just a little

Table 4. Ablation studies for the setting of prompt tuning in Ref-inpainting. Left: ‘Shallow’ means only prompt tuning to text embedding, while ‘Deep’ indicates tuning additional embedding features to different cross-attention layers (16 layers) in SD. Right: validating the influence of the length of shared (Task) and unshared (View) prompts with 3-view Ref-inpainting.

Prompt Type	Length	PSNR↑	SSIM↑	LPIPS↓	Params
Shallow	25	20.35	0.827	0.104	+0.025M
Shallow	50	**20.49**	0.829	**0.103**	+0.05M
Shallow	75	20.38	**0.830**	0.104	+0.075M
Deep ($\times$16)	25(400)	20.15	0.825	0.106	+0.4M

Task	View	PSNR↑	SSIM↑	LPIPS↓	Params
50	0	21.224	0.838	0.0941	+0.05M
45	5	**21.356**	**0.840**	**0.0901**	+0.06M
25	25	21.127	0.836	0.0950	+0.11M
5	45	20.744	0.832	0.1040	+0.14M
0	50	20.563	0.831	0.1110	+0.15M

Table 5. NVS results of LeftRefill-simple with different reference views with/without incremental Positional Encoding (PE).

Ref-View	PE	PSNR↑	SSIM↑	LPIPS↓
1	×	20.352	0.873	0.132
1	✓	**20.508**	**0.875**	**0.128**
4	×	22.097	0.888	0.099
4	✓	**22.324**	**0.890**	**0.095**

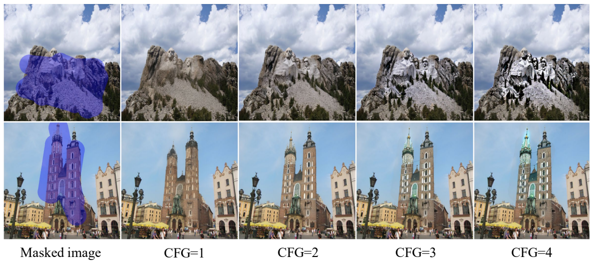


Figure 11. Ref-inpainting results with different CFG weights, which take a trade-off between structural and textural recoveries.

quality degradation. Our method can also be generalized to real-world data as shown in the supplementary.

4.3. Analysis and Ablation Studies

Self-Attention Analysis. We show the visualization of self-attention scores attended by masked regions for Ref-inpainting across different DDIM steps in Figure 10. Self-attention can capture correct feature correlations without any backbone fine-tuning. As diffusion sampling progresses, self-attention modules gradually shift their focus from specific key points to broader related regions, which is convincing and intuitive. Because the key landmarks help to swiftly locate the spatial correlation between the reference and target, while the extended receptive fields further refine the generation for the following sampling steps. More analysis about the attention visualization with increased reference views is shown in the supplementary.

Prompt Settings. The length and depth used in the task and view prompt tuning are explored in Table 4. Different from [23], we find that LeftRefill is relatively robust in the length selection. Thus we select 50 for both Ref-inpainting and NVS. Moreover, the deep prompt with much more trainable prompts for different cross-attention layers does not perform well, which may suffer from a little overfitting. For the multi-view scene, we empirically evaluate the 3-view-based Ref-inpainting performance with various proportions of task&view prompt lengths in the right of Table 4. Increasing the proportion of view tokens initially improves the results, followed by a subsequent decline. We think that a few unshared view tokens contribute valuable view orders, while too many unshared tokens would increase the learning difficulty, leading to an inferior prompt tuning performance.

Incremental Positional Encoding. We incrementally add the concatenation of learnable view embedding and absolute positional encoding to each attention block for NVS (Equation (1)), improving the performance of both single-view and multi-view-based NVS as verified in Table 5.

Classifier-Free Guidance (CFG) [19]. We find that CFG can enhance the performance of Ref-inpainting even without training with prompt dropout as in Figure 11. The adjusting of CFG could be seen as the trade-off between structural and textural recoveries. High CFG scales lead to over-saturated results with superior structure. We empirically set CFG to 2.0 and 2.5 for Ref-inpainting and NVS respectively. More quantitative results and details about the CFG setting of NVS are discussed in the supplementary.

5. Conclusion

In this paper, we propose LeftRefill, formulating reference-based synthesis as inpainting tasks and addressing them end-to-end as a human painter. Benefiting from the prompt tuning and the well-learned attention modules in large T2I models, LeftRefill can address the spatially sophisticated Ref-inpainting and NVS efficiently. Moreover, LeftRefill could be easily extended to tackle multi-view generation tasks. We also propose block casual masking to accomplish NVS with consistent results autoregressively. Comprehensive experiments on Ref-inpainting and NVS show the effectiveness and efficiency of LeftRefill.

Acknowledgements: Yanwei Fu is the corresponding author. Yanwei Fu is with the School of Data Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University, and Fudan ISTBI-ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University. The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2, 2023. 2

[3] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 3

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3

[5] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 3

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[7] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *Advances in neural information processing systems*, 32, 2019. 3

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[9] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–II. IEEE, 2003. 3

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 5, 6, 7

[11] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 3

[12] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 6

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 5

[14] George Fahim, Khalid Amin, and Sameh Zarif. Single-view 3d reconstruction: A survey of deep learning methods. *Computers & Graphics*, 94:164–190, 2021. 3

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[16] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 3

[17] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (ToG)*, 26(3):4–es, 2007. 3

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[19] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 8

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 7

[22] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5, 6

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 3, 8

[24] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2, 5

[25] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 3

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023. 2

[27] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 5, 6

[28] Ning Liao, Bowen Shi, Min Cao, Xiaopeng Zhang, Qi Tian, and Junchi Yan. Rethinking visual prompt learning as masked visual token modeling. *arXiv preprint arXiv:2303.04998*, 2023. 3

[29] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

[30] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2, 3, 4, 5, 6, 7

[31] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 3

[32] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 2

[33] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 2

[34] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 2

[35] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2

[36] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[38] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3

[39] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (ToG)*, 38(6):1–15, 2019. 3

[40] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *proceedings of the IEEE/cvf international conference on computer vision*, pages 4403–4412, 2019. 3

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 7

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[45] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021. 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 5

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[49] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations (ICLR)*, 2017. 5

[50] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 5

[51] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[52] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 3

[53] Kihyuk Sohn, Yuan Hao, José Lezama, Luisa Polania, Huiwen Chang, Han Zhang, Irfan Essa, and Lu Jiang. Visual

prompt tuning for generative transfer learning. *arXiv preprint arXiv:2210.00990*, 2022. 3

[54] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 3

[55] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022. 5, 6

[56] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 5

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[58] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022. 2

[59] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 3

[60] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems*, 2017.

[61] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in neural information processing systems*, 2019. 3

[62] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2, 5, 6

[63] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 1–17. Springer, 2020. 3

[64] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2, 5, 6

[65] Liang Zhao, Xinyuan Zhao, Hailong Ma, Xinyu Zhang, and Long Zeng. 3dfill: Reference-guided image inpainting by self-supervised 3d image alignment. *arXiv preprint arXiv:2211.04831*, 2022. 2, 3

[66] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 3

[67] Yunhan Zhao, Connelly Barnes, Yuqian Zhou, Eli Shechtman, Sohrab Amirghodsi, and Charless Fowlkes. Geofill: Reference-based image inpainting of scenes with complex geometry. *arXiv preprint arXiv:2201.08131*, 2022. 2, 3

[68] Yuqian Zhou, Connelly Barnes, Eli Shechtman, and Sohrab Amirghodsi. Transfill: Reference-guided image inpainting by merging multiple color and spatial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2266–2276, 2021. 2, 3, 5, 6