NÜWA-LIP: Language-guided Image Inpainting with Defect-free VQGAN

Minheng Ni[1] Xiaoming Li[1] ✉ Wangmeng Zuo[1,2]
[1]Harbin Institute of Technology [2]Peng Cheng Laboratory
mhni@stu.hit.edu.cn csxmli@gmail.com wmzuo@hit.edu.cn

Abstract

Language-guided image inpainting aims to fill the defective regions of an image under the guidance of text while keeping the non-defective regions unchanged. However, directly encoding the defective images is prone to have an adverse effect on the non-defective regions, giving rise to distorted structures on non-defective parts. To better adapt the text guidance to the inpainting task, this paper proposes NÜWA-LIP, which involves defect-free VQGAN (DF-VQGAN) and a multi-perspective sequence-to-sequence module (MP-S2S). To be specific, DF-VQGAN introduces relative estimation to carefully control the receptive spreading, as well as symmetrical connections to protect structure details unchanged. For harmoniously embedding text guidance into the locally defective regions, MP-S2S is employed by aggregating the complementary perspectives from low-level pixels, high-level tokens as well as the text description. Experiments show that our DF-VQGAN effectively aids the inpainting process while avoiding unexpected changes in non-defective regions. Results on three open-domain benchmarks demonstrate the superior performance of our method against state-of-the-arts. Our code, datasets, and model will be made publicly available[1].

1. Introduction

The task of image inpainting, which aims to fill missing pixels in the defective regions with photo-realistic structures, is as ancient as art itself [2]. Despite its practical applications [18, 32, 34], such as image manipulation, image completion, and object removal, the task poses significant challenges, including the effective extraction of valid features from defective input and the generation of semantically consistent results.

With the remarkable success of vision-language learning, language-guided image inpainting has become a promising topic [24, 28, 33], which enables the generation of controllable results with the guidance of text description (see the completed results in Fig. 1). Recently, multimodal pre-training methods based on diffusion and autoregressive models have exhibited impressive capabilities in synthesizing various and photo-realistic images, such as Stable Diffusion [21], Parti [29] and NÜWA [25]. In particular, NÜWA has demonstrated a promising capability for language-guided image generation, suggesting the potential for combining this pre-training schema with VQVAE and Transformer for language-guided image inpainting.

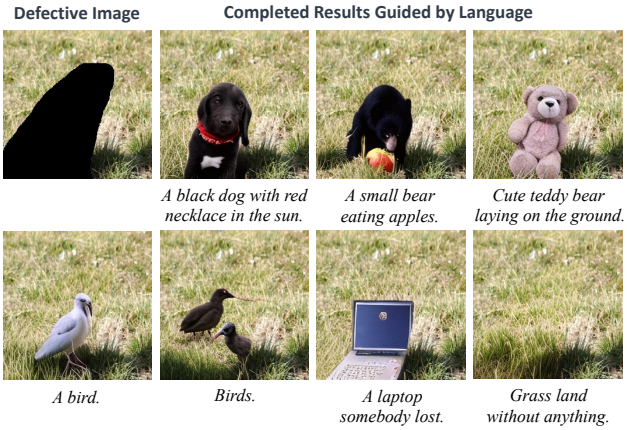

Figure 1. **Language-guided inpainting results via NÜWA-LIP.** Text descriptions provide effective guidance for inpainting the defective image with desired objects. More examples are in the suppl.

It is worth noting that this work focuses on addressing the challenge of processing defective images with some regions filled with zeros (see the first image in Fig. 1). This setting has been widely adopted in previous works [24, 28, 33] and is consistent with real-world corrupted image inpainting. To learn unified representations of the vision and language, it is crucial to ensure that the representation of non-defective regions is accurate and remains unaffected by defective parts. This requirement exacerbates the challenges associated with our method, as it involves the effective extraction of valid features from defective input.

However, existing pre-trained image generative models [21, 25, 29] are usually trained on non-defective images. When used in the image inpainting task, these models encode the whole image and fuse the features from defective regions into the representations of non-defective parts, re-

[1] https://github.com/kodenii/NUWA-LIP

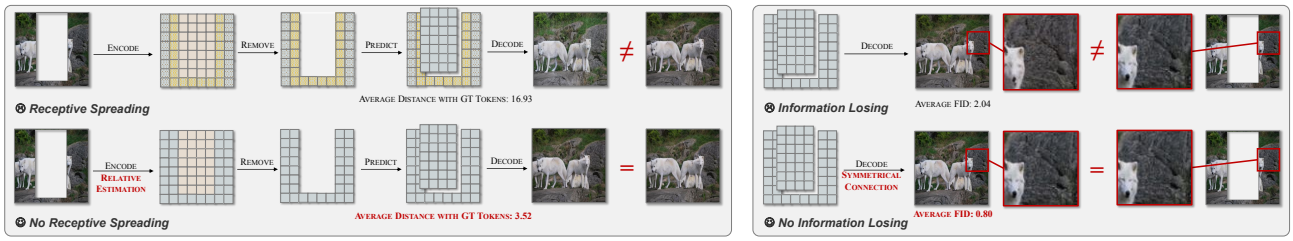

Figure 2. **Illustration of Receptive Spreading and Information Losing.** As for receptive spreading, we can observe an obvious change and a blurry boundary on the reconstructed non-defective region. In terms of information losing, some unexpected changes are also apparent in the output of the non-defective regions.

sulting in what we call *receptive spreading* of the defective region. These types of approaches can severely limit the accuracy of modeling non-defective regions, especially when the defective region is large, and make it difficult to match language descriptions. Furthermore, in the VQVAE-based model, the known non-defective regions are challenging to reconstruct exactly the same as the original image due to the compression of the image into discrete tokens. We refer to this as *information losing* in the non-defective region (see Fig. 2 for an illustration).

To enable effective adaptation of text descriptions to these types of defective images, we propose NÜWA-LIP which leverages a novel defect-free VQGAN (DF-VQGAN) and a multi-perspective sequence-to-sequence module (MP-S2S). In contrast to VQGAN [8], DF-VQGAN incorporates the relative estimation to decouple defective and non-defective regions. This helps to control receptive spreading and obtain accurate visual representations for vision-and-language (VL) learning in MP-S2S. To retain the information of non-defective regions, symmetrical connections replenish the lost information from the features in the encoding procedure. Additionally, MP-S2S further enhances visual information from complementary perspectives, including low-level pixels, high-level tokens, and the text description.

Moreover, we construct three datasets to evaluate the performance of language-guided image inpainting and conduct a comprehensive comparison with existing methods. Experiments demonstrate that NÜWA-LIP outperforms competing methods by a significant margin on all three benchmarks. An ablation study is further conducted to evaluate the effectiveness of each component in our NÜWA-LIP.

The main contributions are summarized as follows:

- To effectively encode the defective input, we propose a DF-VQGAN, which introduces relative estimation to control receptive spreading and symmetrical connections to retain the information of non-defective regions.
- We propose a multi-perspective sequence-to-sequence module for enhancing visual information from complementary perspectives of pixel, token, and text domains.
- We build three open-domain datasets for evaluating language-guided image inpainting. Experiments show that NÜWA-LIP achieves state-of-the-art performance in comparison with the competing methods.

2. Related Work

Vector Quantized Variational AutoEncoder. The Vector Quantized Variational AutoEncoder is a VAE model that compresses the continuous information into discrete latent variables [17]. Several works, such as VQVAE-2 [20], aim to decode the image at a more fine-grained level. To generate images with more vivid structures, VQGAN [8] uses a GAN model [9] to constrain the decoded images indistinguishable from the real ones. However, existing models typically focus on normal images that are not corrupted. In images containing defective regions, especially in inpainting scenarios, these defective regions can affect all discrete latent variables due to receptive spreading. This can result in the color cast or faults in the decoded result. We note that Liu *et al*. [14] proposes partial convolution (PConv) with a modified convolution layer to enhance features from non-defective regions. However, directly applying this design to VQGAN is unsuitable because it cannot effectively encode only the non-defective regions and decode the full image during training according to VAE[2]. Additionally, VQGAN consists of a series of operations, *e.g.*, normalization, which can easily lead to receptive spreading. The loss of information can modify the non-defective region, decreasing the inpainting quality. Therefore, we design a new encoding paradigm specifically for VQGAN on the inpainting task.

Language-guided Image Inpainting. Language-guided image inpainting, as a subfield of text-image synthesis, has attracted tremendous attention. This task aims to fill in the defective regions of an image with text guidance that describes the content of the entire image. Analogous to unconditional image inpainting [24, 28, 33], some works, such as [1, 26, 30], employ generative adversarial networks to handle generic real images. Recently, many works [3, 15, 26] pre-training on large-scale data are presented to improve inpainting ability in the general domain rather than a specific domain. Several methods, such as ImageBART [7] and GLIDE [16], perform this task with a pre-training diffusion model. Following

[2]Further discussion can be found in suppl.

DALL-E [19], other works like NÜWA [25] utilize the autoregressive model. However, existing models can not be directly applied to defective input and suffer from the problems of receptive spreading of defective regions and loss of information in non-defective parts. Moreover, modeling the image from a single perspective condition limits the quality of the inpainting results. Therefore, in this work, we propose the NÜWA-LIP, which utilizes the DF-VQGAN to encode consistent and valid features, and MS-S2S to provide more comprehensive guidance for defective regions.

3. Method

3.1. Problem Formulation

Given an input image $x \in \mathbb{R}^{W\times H\times C}$, a mask matrix $m \in \{0,1\}^{W\times H}$ with value 1 denoting the defective regions, and a piece of natural text t, the task of language-guided image inpainting is to repair the defective regions under the guidance of the text and generate a new image $\hat{y} \in \mathbb{R}^{W\times H\times C}$. In the following, we refer to the input x as the **defective image**, $\hat{y}$ as the **completed result**, and y as the corresponding **ground-truth image**.

In the Bayesian framework, the language-guided image inpainting task can be defined as maximizing the log posterior probability[3], as denoted in Eqn. (1):

$$\log p(y|x,m,t;\omega,\phi,\psi) = \log \frac{p(y|z,x,m;\psi)p(z|x,m,t;\phi)}{p(z|y,x,m;\omega)}, \quad (1)$$

where ω, ϕ, ψ denote model parameters. z represents the latent tokens. By taking the expectation w.r.t a auxiliary density $z \sim q(z|y,x,m;\phi)$ on both sides, the right side of Eqn. (1) can be formulated as:

$$\begin{aligned} &\mathbb{E}_{z\sim q(z|y,x,m;\phi)}[\log p(y|z,x,m;\psi)] - \\ &\mathbb{KL}_{z\sim q(z|y,x,m;\phi)}\left[q(z|y,x,m;\phi)||p(z|x,m,t;\phi)\right] + \\ &\mathbb{KL}_{z\sim q(z|y,x,m;\phi)}\left[q(z|y,x,m;\phi)||p(z|y,x,m;\omega)\right]. \end{aligned} \quad (2)$$

According to VAE, since the third Kullback-Leibler divergence term is always greater than 0, we only need to maximize the first two terms denoted as the Evidence Lower BOund (ELBO). The first expectation term is the reconstruction loss of the completed image. The second term is a Kullback-Leibler divergence loss that ensures the conditional distribution of the latent tokens generated by the VAE encoder should be close to that generated by the auxiliary probability density. The whole framework is shown in Fig. 3. In the following, we will introduce how we model the first term with a defect-free VQGAN (DF-VQGAN) in Sec. 3.2 and the second term with a multi-perspective sequence-to-sequence (MP-S2S) in Sec. 3.3. From the terms of mask m and defective image x in ELBO, we can observe that it is necessary to introduce mask matrix and defective image in the process of Vector Quantized Variational AntoEncoder for image inpainting.

[3]We assume t is conditionally independent of y given z, x, m.

3.2. DF-VQGAN

The modeling of $\mathbb{E}_{z\sim q(z|y,x,m;\phi)}[\log p(y|z,x,m;\psi)]$ in Eqn. (13) can be split into a VAE encoder of $q(z|y,x,m;\phi)$ and a VAE decoder of $p(y|z,x,m;\psi)$. The probabilistic density function $q(z|y,x,m;\phi)$ shows that the probability distribution of z should be conditioned on three variables: ground-truth image y, defective image x, and mask matrix m. In other words, the latent tokens z should not only represent the completed image but also be sensitive to defective images with masked regions. To achieve this, we propose DF-VQGAN, a defect-free VAE model.

Defect-free Encoder. Let $w \times h$ be the length of the quantized latent token sequence and n_z be the dimensionality of each latent token. To obtain an equivalent latent token sequence $z^y \in \mathbb{R}^{w\times h\times n_z}$ of the ground-truth image y, we feed y to encoder E, which consists of several normal operations, such as attention or convolutions:

$$z^y = E(y), \quad (3)$$

The defective regions in x will have adverse effects on the non-defective regions during the encoding phase (see the inconsistent color of the VQGAN column in Fig. 5), making it different from encoding y. Mathematically, if we feed an image x into a convolutional network (taking only one convolution layer as an example), the number of affected latent tokens is equal to the number of 1 in $\mathrm{MaxPool}(m)$, which uses the same stride and kernel size as the convolution. For the networks with the attention or normalization operation, more latent tokens will be affected.

As for the image inpainting task, it is intractable for the original VAE model to avoid the adverse effects from defective regions, because VQGAN cannot encode only the non-defective regions and then decode the full image during the training phase. Besides, VQGAN consists of a series of operations that can result in receptive spreading. Therefore, we propose DF-VQGAN, which splits features of defective regions from non-defective regions and merges them into the features of a full image during training alternately. This can keep the feasibility of training and prevent encoding the non-defective regions with effect from defective regions. By using the mask matrix, we can easily mask the defective parts for attention and convolution operations by padding the defective features and their attention score to zero. However, normalizing the non-defective regions directly would break parallelism. Therefore, we propose a mathematical method to remove the influence from the defective regions. Using the re-estimated mean and variance, the normalization[4] can be formulated as:

$$\mathrm{Norm}^{\mathrm{DF}}(x,m) = \frac{x - \frac{N}{N_m}\mathrm{E}[x]}{\sqrt{\frac{N-1}{N_m-1}\mathrm{Var}[x'] + \epsilon}}, \quad (4)$$

[4]We use Layer Norm as an example to illustrate our method. This equation can be easily extended to other normalizations like Group Norm.

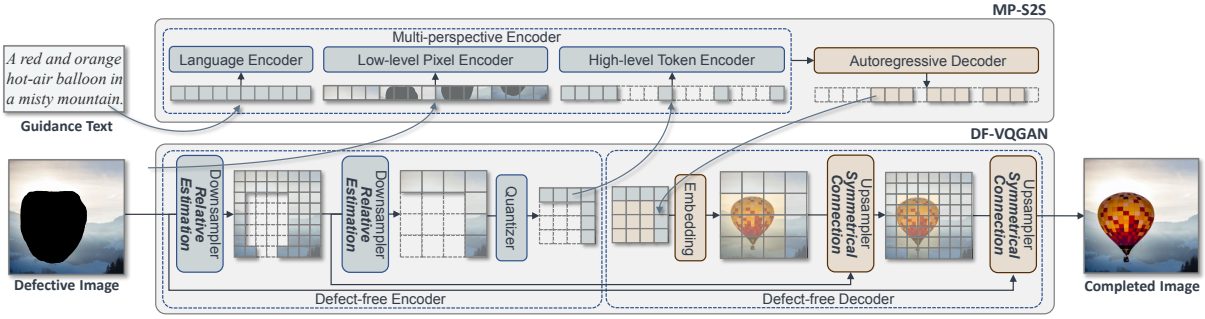

Figure 3. **Overview of NÜWA-LIP during inference.** The encoder of DF-VQGAN generates the tokens from the defective image. Then, tokens located in defective regions will be replaced with those predicted by our MP-S2S. The decoder of DF-VQGAN will generate the completed result based on these tokens.

where N is the number of pixels in x. N_m denotes the number of defective pixels. x' denotes the x with the defective region fulfilled with $\frac{N}{N_m}\mathrm{E}[x]$. $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ represent expectation and variance, respectively.

By introducing mask m, the defect-free convolution can be formalized with:

$$\mathrm{Conv}^{\mathrm{DF}}(x) = W_c^\top (x \odot m) + b, \tag{5}$$

where W_c and b are the shared parameters of the original convolution. With the shared parameter W_a of the original attention, we can mask the attention score of defective regions by:

$$\mathrm{Attn}^{\mathrm{DF}}(x) = \mathrm{Softmax}(x^\top W_a x) \odot m \odot x. \tag{6}$$

We refer to these defect-free operations as ***relative estimation***. By replacing operations in encoder E with ***relative estimation***[5], we get the defect-free encoder E^{DF} and then use it encode the defective image x by:

$$z^x = E^{\mathrm{DF}}(x, m), \tag{7}$$

where $z^x \in \mathbb{R}^{w \times h \times n_z}$ is the latent tokens of x without polluting the non-defective part.

In order to enforce the consistency between z^x to z^y, we merge them into a joint latent tokens sequence $z \in \mathbb{R}^{w \times h \times n_z}$ and optimize them jointly (see illustration in Fig. 4). By applying a $\mathrm{MaxPool}$ operation, the mask matrix m can be resized to m' to match the size of z^x and z^y. Next, we can use $m' \in \mathbb{R}^{w \times h}$ to filter out the non-defective parts of z^x and replace them with the corresponding parts from z^y by:

$$\begin{aligned} m' &= \mathrm{MaxPool}(m), \\ z &= z^y \odot m' + z^x \odot (1 - m'). \end{aligned} \tag{8}$$

Note that the encoding is an iterative process and the input will be gradually transformed into the final latent tokens. During the process, the defective region m also evolves. Thus, we extend Eqns. (3, 7, 8) to the whole iterative process.

[5]For more descriptions of the replacement, see our suppl.

Empirically, encoder E and defect-free encoder E^{DF} can be divided into a series of downsamplers $\{e_i\}_{i=1}^T$ and $\{e_i^{\mathrm{DF}}\}_{i=1}^T$. e_i^{DF} and e share the same parameters, with their difference being whether to use ***relative estimation*** or not.

Now, we are able to mitigate the adverse effect of defective regions, regardless of their shape or scale. Formally, we obtain the merged final latent tokens $z = h_T$ with:

$$\begin{aligned} h_i^y &= e_i(h_{i-1}), \\ h_i^x &= e_i^{\mathrm{DF}}(h_{i-1}, m_{i-1}), \\ m_i &= \mathrm{MaxPool}_i(m_{i-1}), \\ h_i &= h_i^y \odot m_i + h_i^x \odot (1 - m_i), \end{aligned} \tag{9}$$

where the stride and kernel size of $\mathrm{MaxPool}_i$ are the same as those of the convolution in e_i. We set the initial values of the process for training as $m_0 = m$ and $h_0 = y$.

With a learnable codebook B, we can obtain the indexes $\tilde{z} \in \mathbb{R}^{w \times h}$ for matching latent tokens z in B through:

$$\tilde{z}_i = \arg\min_j ||z_i - B_j||^2. \tag{10}$$

By applying Eqn. (10) to z^x or z^y, we can obtain discrete indexes $\tilde{z}^x$ or $\tilde{z}^y$ which will be used for MP-S2S training.

Defect-free Decoder. Following the VQGAN architecture, we first embed the discrete indexes $\tilde{z}$ by looking up the codebook via $z = B[\tilde{z}]$, which are then fed into the decoder G. Different from the encoding process in Eqn. (7), we do not need to perform the defect-free operations, since there are no defective regions in z because this region has been implemented either by encoding from the encoder E in training or by predicting from the MP-S2S in inference.

As shown in Fig. 4, to avoid the information losing of the non-defective region, we propose ***symmetrical connection***. Let $\hat{y} \in \mathbb{R}^{W \times H \times C}$ be the output of the decoder and τ be the mixture coefficient. The hidden state of the non-defective region in the encoder is mixed with the output of the decoder:

$$\begin{aligned} \hat{y}' &= G(z), \\ \hat{y} &= (1 - m) \odot \frac{1}{\tau + 1}(\hat{y}' + \tau x) + m \odot \hat{y}'. \end{aligned} \tag{11}$$

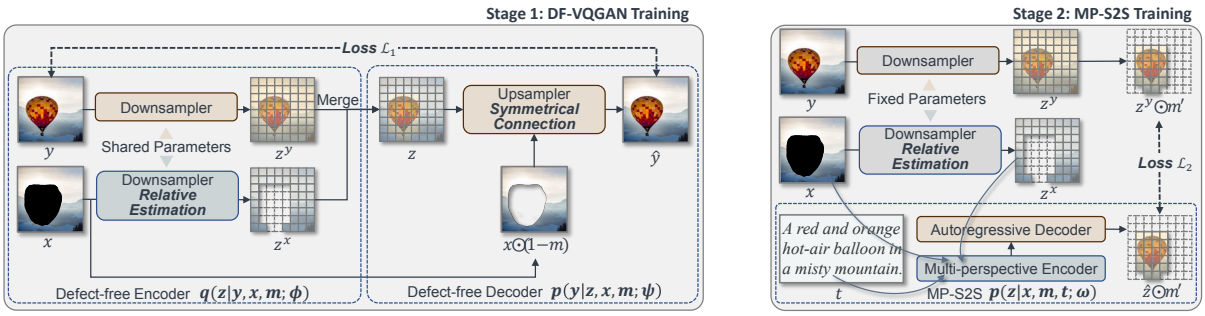

Figure 4. **Overview of NÜWA-LIP during two-stage training.** We adopt a two-stage training process to optimize the whole NÜWA-LIP. Since we only have ground-truth image y in training data, we generate corresponding defective image x by applying a random mask m. For Stage 1, we optimize DF-VQGAN via Eqn. (14) with shared parameters for e and e^{DF} . We use ground-truth image y and defective image x as the inputs of e and e^{DF}, respectively. For Stage 2, we optimize MP-S2S via Eqn. (15) with fixed DF-VQGAN. Here, DF-VQGAN is adapted to generate the training target $z^y \odot m'$ and input latent tokens z^x.

Correspondingly, decoding is a reversed iterative process of encoding. We can divide the decoder G into a series of upsamplers $d_1, d_2, ..., d_T$ that are symmetrical to the downsamplers in E or E^{DF}. Similar to Eqn. (9), we extend Eqn. (11) to the whole iterative process and obtain the final output image $\hat{y} = \hat{h}_0$, which is reconstructed based on the latent tokens z by,

$$\begin{aligned} \hat{h}'_i &= g_{i+1}(\hat{h}_{i+1}) \,, \\ \hat{h}_i &= (1-m_i) \odot \frac{1}{\tau_i + 1}(\hat{h}'_i + \tau_i h_i) + m_i \odot \hat{h}'_i \,, \end{aligned} \tag{12}$$

where we set $\tau_0 = \infty$ to remain non-defective region unchanged and $\tau_i = 1$ for $i \neq 0$ to obtain a smoother mixture. Symmetrically, the initial state of the process is $\hat{h}_T = z$.

Let $\mathrm{sg}[\cdot]$ be the stop gradient function. Following VQGAN, the training objective is:

$$\begin{aligned} \mathcal{L}^V &= ||y - \hat{y}||_2^2 + ||\mathrm{sg}[z] - B[\tilde{z}]||_2^2 + ||\mathrm{sg}[B[\tilde{z}]] - z||_2^2 \,, \\ \mathcal{L}^P &= ||Q(y) - Q(\hat{y})||_2^2 \,, \\ \mathcal{L}^G &= \log D(y) + \log(1 - D(\hat{y})) \,, \end{aligned} \tag{13}$$

where Q and D are CNN-based modules to obtain the conceptual representation of the image and discriminator.

The overall learning objective of DF-VQGAN is:

$$\mathcal{L}_1 = \mathcal{L}^V + \mathcal{L}^P + \mathcal{L}^G \,. \tag{14}$$

3.3. MP-S2S

In Sec. 3.2, we model the probabilistic density function $q(z|y,x,m;\phi)$ using the Kullback-Leibler divergence term $\mathbb{KL}_{z\sim q(z|y,x,m;\phi)}\left[q(z|y,x,m;\phi)||p(z|x,m,t;\phi)\right]$ by a DF-VQGAN encoder, while the key to solving this term relies on the modeling of the density function $p(z|x,m,t;\phi)$. In this section, we propose a multi-perspective sequence-to-sequence module (MP-S2S). It encodes the information from three perspectives, including the input text t, pixel-level defective image x, and its token-level representation z^x, and decodes the latent tokens of the defective regions $\hat{z}$.

Multimodal Encoding. As for guidance text from language modalities, we follow BERT [5] and tokenize the text with BPE and embed them to the representation sequence t, where $t_i \in \mathbb{R}^{n_t}$ denotes the representation of each token and n_t denotes the dimension of the representation. We encode token representation sequence t by $c^{\mathrm{t}} = E^{\mathrm{t}}(t)$.

For the defective image, the model treats it from two perspectives, *i.e.*, high-level token, and low-level pixel representation. For low-level pixel, following ViT [6], we transform the image into a sequence of non-overlapping patches $x^{\mathrm{p}} = (x_1^{\mathrm{p}}, x_2^{\mathrm{p}}, ...)$. Then, we directly encode the image patch sequence to get representation from a low-level pixel perspective by $c^{\mathrm{l}} = E^{\mathrm{l}}(x^{\mathrm{p}})$. For high-level representation, we encode latent tokens z^x by $c^{\mathrm{h}} = E^{\mathrm{h}}(z^x)$. Note that tokens corresponding to defective regions, which can be located by m', will be replaced with a special trainable vector. We use Transformer encoder as the architecture of E^{h}, E^{l} and E^{t}. Thus, the integrated representation at the hidden size dimension can be formulated as $c = [c^{\mathrm{t}}; c^{\mathrm{l}}; c^{\mathrm{h}}]$.

Autoregressive Decoding. The integrated representation c from two modalities can be considered as the condition of the Transformer decoder. The decoder aims to predict the missing latent tokens based on the condition and latent tokens: $P(z_k|z_{<k}, c)$, where z_k is the k-th tokens and $z_{<k}$ is token sequence before the k-th token. Following MASS [23], we only decode the masked tokens which can be located by m' in DF-VQGAN. The training objective of MP-S2S is:

$$\mathcal{L}_2 = -\sum_k \log P(z_k|z_{<k}, c) \,. \tag{15}$$

3.4. Inference Pipeline

As shown in Fig. 3, we use defective image x as the input image of DF-VQGAN. Since the ground-truth image y cannot be accessed during inference, we use the defect-free encoder by setting $m_0 = m$ and $h_0 = x$ to obtain latent tokens z^x of defective image x. Then, we use MP-S2S to predict latent tokens indicated by m' in z^x, based on the guidance of text t, defective image x, and its latent tokens

z^x. Predicted tokens $\hat{z}$ will replace the tokens in z^x by $\hat{z}' = z^x \odot (1 - m') + \hat{z} \odot m'$. The completed result can be reconstructed by the decoder in DF-VQGAN with $\hat{h}_T = \hat{z}'$.

4. Experiments

4.1. Implementation Details

The final pre-trained model has a total of 1.7 billion parameters. We pre-train the whole NÜWA-LIP using 64 A100 GPUs over a period of two weeks. Each image is resized to 256×256. During training, we randomly generate mask metrics m with the mask ratios ranging from 40% to 60%[6]. AdamW optimizer [11] is used with the warm-up ratio of 5% and dropout of 10% for both pre-training and fine-tuning stages. We pre-train DF-VQGAN on ImageNet [4] and MP-S2S on Conceptual Captions [22], respectively. More details can be found in our suppl.

4.2. Experiments Setup

Datasets Description. In order to evaluate the proposed model, we construct three evaluation datasets, namely MaskCOCO, MaskFlickr, and MaskVG. These dataseets are based on MSCOCO [13], Flickr [27] and VG [12], respectively. Unlike domain-specific datasets, these three datasets are open-domain and comprise diverse language descriptions. The mask region of each image is not fixed (see our suppl.).

Evaluation Metric. To assess the quality of the inpainted image using language guidance, we select the FID score [10] as the metric. This score is used to compare the difference in distributions between the generated images and the real-world ones. In addition, we adopt CLIP Score (CS) to measure consistency between vision and language. Furthermore, we utilize PSNR and LPIPS [31] to evaluate the similarity of the pixel and perception domains.

Baselines. We compare NÜWA-LIP with two robust baselines. GLIDE [16] is an effective diffusion-based model for image generation and editing, and we use the public version of GLIDE from the official repository. To be specific, we modify their inference code by applying de-noising and super-resolution processes on regions of the mask instead of rectangle regions. NÜWA [25] is another effective model for vision generation and editing. For this task, we follow the image completion framework of NÜWA and re-implement it to perform this inpainting task. Moreover, we modify NÜWA to NÜWA-P by pasting the non-defective region onto the inpainted result for a comprehensive comparison.

4.3. Overall Results

We first compare NÜWA-LIP with the two baseline models As shown in Tab. 1, NÜWA-LIP achieves the best performance on all datasets. Our proposed method outperforms the highest baseline GLIDE with 1.5 FID on MaskCOCO, 9.4 FID in MaskFlickr, and 0.5 on MaskVG, indicating the effectiveness of NÜWA-LIP in generating photo-realistic and vision-language consistent results. We suggest that the improvement can be ascribed to NÜWA-LIP's ability to maintain the non-defective region unchanged while avoiding inaccurate or incomprehensive encoding. The improvement on fine-tuning NÜWA-LIP on MSCOCO dataset (see NÜWA-LIP (FINETUNE) *v.s.* NÜWA-LIP in Tab. 1), further demonstrates that the domain-specific dataset can benefit NÜWA-LIP like most pre-trained large models. Furthermore, we conduct extra comparisons with classical image inpainting and visual synthesis models in our suppl.

[6]We tried different ratios in training and selected the best, see suppl.

4.4. Effectiveness of DF-VQGAN

To demonstrate the advantages of DF-VQGAN, we train DF-VQGAN and VQGAN using the same training steps and data for two different tasks. We use the official implementation of VQGAN for a fair comparison.

Image Reconstruction. The goal of this task is to transform each complete image into discrete latent tokens and then reconstruct the image based on these tokens. This task can help evaluate the image encoding and decoding performance of VQGAN, which is critical to NÜWA-LIP. As shown in Tab. 2, we can see that DF-VQGAN performs better than VQGAN in the same setting (*i.e.*, resolution and vocab size). This improvement can be ascribed to the fact that the training of DF-VQGAN covers both complete image reconstruction and incomplete image reconstruction (see Fig. 4), which makes the resulting model more robust.

Oracle Inpainting. This task aims to reconstruct the image given the ground-truth discrete tokens of the defective region (encoded from the ground-truth image) and the discrete tokens of the rest region (encoded from the defective image). From Tab. 2, we can see that DF-VQGAN again performs better than VQGAN in the same setting (*i.e.*, resolution and vocab size), which verifies the effectiveness of DF-VQGAN for the image inpainting task. Fig. 5 presents some visualization examples of VQGAN and DF-VQGAN. Compared with our DF-VQGAN, the original VQGAN easily generates structures with color cast or blurry boundaries.

Average Distance. To verify whether DF-VQGAN avoids fusing uncertain information from defective regions, we measure the consistency of z^x and z^y. For each pair containing defective image x and non-defective one y, we obtain their representations z^x and z^y. We only measure the known part of the image, which is indicated by the mask m. Then we calculate the average vector distance between z^x and z^y as a measure of their consistency. We find DF-VQGAN largely improves the consistency of z^x and z^y, providing mathematical evidence for the effectiveness of DF-VQGAN.

Table 1. **Overall results of language-guided image inpainting compared with open-domain pre-trained models.** Human evaluation can be found in suppl., which shows that our NÜWA-LIP again outperforms others on both visual quality and semantic consistency.

Model	MaskCOCO				MaskFlickr				MaskVG			
	FID↓	PSNR↑	LPIPS↓	CS↑	FID↓	PSNR↑	LPIPS↓	CS↑	FID↓	PSNR↑	LPIPS↓	CS↑
GLIDE [16]	13.5	17.25	0.233	28.99	51.9	17.06	0.225	29.16	9.0	18.53	0.210	24.94
NÜWA [25]	21.4	12.91	0.435	28.10	59.5	13.44	0.361	29.29	18.5	14.04	0.362	25.14
NÜWA-P	20.6	14.39	0.323	28.64	54.2	14.89	0.295	29.55	17.7	16.25	0.301	24.78
NÜWA-LIP	**12.0**	**17.35**	**0.233**	**29.39**	**42.5**	**17.47**	**0.220**	**30.76**	**8.5**	**18.70**	**0.209**	**25.05**
NÜWA-LIP (finetune)	**10.5**	17.23	**0.231**	**29.65**	-	-	-	-	-	-	-	-

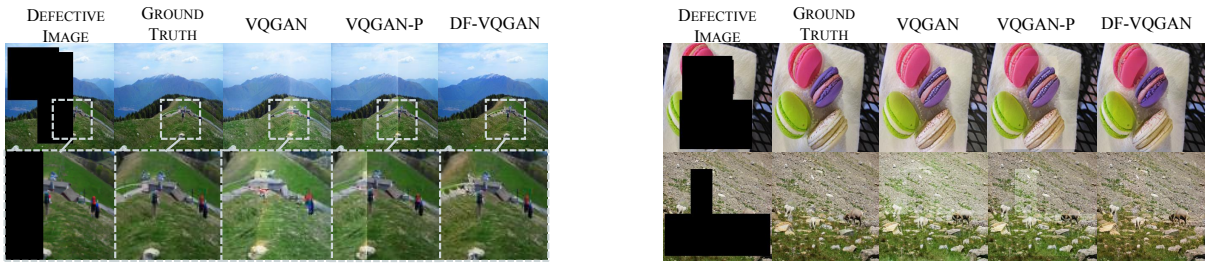

Figure 5. **Illustration on oracle inpainting.** Compared with VQGAN-P, we make a better transition between defective and non-defective regions and keep non-defective regions unchanged.

Table 2. **Comparsion of DF-VQGAN and VQGAN.** Our proposed DF-VQGAN beats the original VQGAN on image reconstruction (Img.Rec), oracle inpainting (Orc.Inp), and average distance (Avg.Dis) under the same settings.

Model	Resolution	Length	Vocab	Img.Rec	Orc.Inp	Avg.Dis
VQGAN	$256^2 \rightarrow 16^2$	256	1024	12.47	16.30	5.75
DF-VQGAN	$256^2 \rightarrow 16^2$	256	1024	**11.16**	**5.56**	**4.54**
VQGAN	$256^2 \rightarrow 16^2$	256	12288	5.48	7.15	9.77
DF-VQGAN	$256^2 \rightarrow 16^2$	256	12288	**5.16**	**2.95**	**4.31**
VQGAN	$256^2 \rightarrow 32^2$	1024	8192	1.47	2.04	16.93
DF-VQGAN	$256^2 \rightarrow 32^2$	1024	8192	**1.38**	**0.80**	**3.52**

In Fig. 5, we compare DF-VQGAN and VQGAN with VQGAN-P, which utilizes the non-defective region from the input image to substitute the reconstruction part of the same region by VQGAN. In comparison to VQGAN-P, our DF-VQGAN has a significantly better transition of the non-defective and inpainted regions, contributing to better boundary consistency. More results can be found in our suppl.

4.5. Ablation Studies

DF-VQGAN. We mainly investigate whether our proposed ***symmetrical connection*** and ***relative estimation*** are beneficial to the inpainting task. To this end, we conduct experiments with two variants of DF-VQGAN/sr and DF-VQGAN/s, which respectively denote the model without both symmetrical connection and relative estimation or only symmetrical connection. We re-train DF-VQGAN under different settings and evaluate oracle inpainting mentioned in Sec. 4.4 on ImageNet. As shown in Tab. 3, we find that ***relative estimation*** reduces the FID by 2.31, while the ***symmetrical connection*** provides a further decrease of 2.49.

Table 3. **Ablation study of DF-VQGAN on ImageNet.** It indicates that both ***relative estimation*** (Rel.Est) and ***symmetrical connection*** (Sym.Con) benefit the performance of DF-VQGAN.

Model	Component		FID↓
	Sym.Con	Rel.Est	
DF-VQGAN/sr	×	×	7.15
DF-VQGAN/s	×	√	5.44
DF-VQGAN	√	√	**2.95**

Table 4. **Ablation study of MP-S2S on MaskCOCO.** Text encoder (Text), the high-level token encoder (High), and the low-level pixel encoder (Low) can well understand the language and defective image from multi-perspectives to improve the performance.

Model	Component			FID↓
	Text	High	Low	
MP-S2S/hl	√	×	×	29.4
MP-S2S/l	√	√	×	27.4
MP-S2S/h	√	×	√	26.8
MP-S2S/t	×	√	√	34.7
MP-S2S	√	√	√	**26.2**
+DF-VQGAN	√	√	√	**11.0**

This indicates that both of these operations are crucial for effectively encoding the defective image.

MP-S2S. To verify whether the multi-perspective benefits language-guided image inpainting, we conduct an ablation study on MP-S2S by removing one or two of its perspectives. MP-S2S/l, MP-S2S/h, and MP-S2S/t denote the MP-S2S module without the low-level pixel encoder, high-level token encoder, and text encoder, respectively. MP-S2S/hl denotes the MP-S2S module with only a text encoder. To exclude the effect of other components, we use the same original

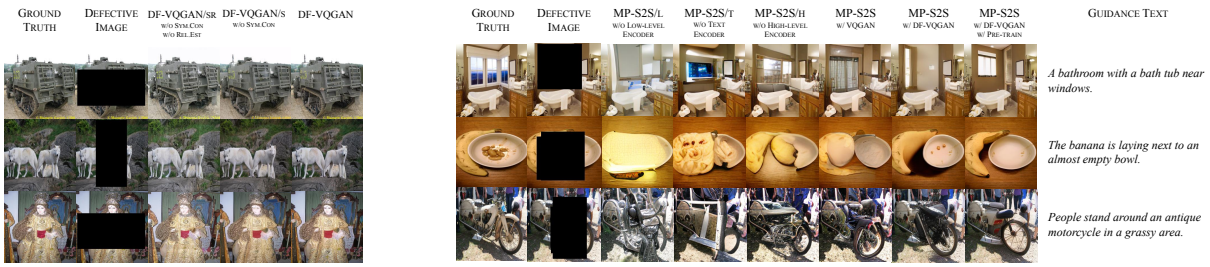

Figure 6. **Qualitative ablations of proposed components for DF-VQGAN and MP-S2S.** We can observe that the model achieves the best performance by combining all proposed components.

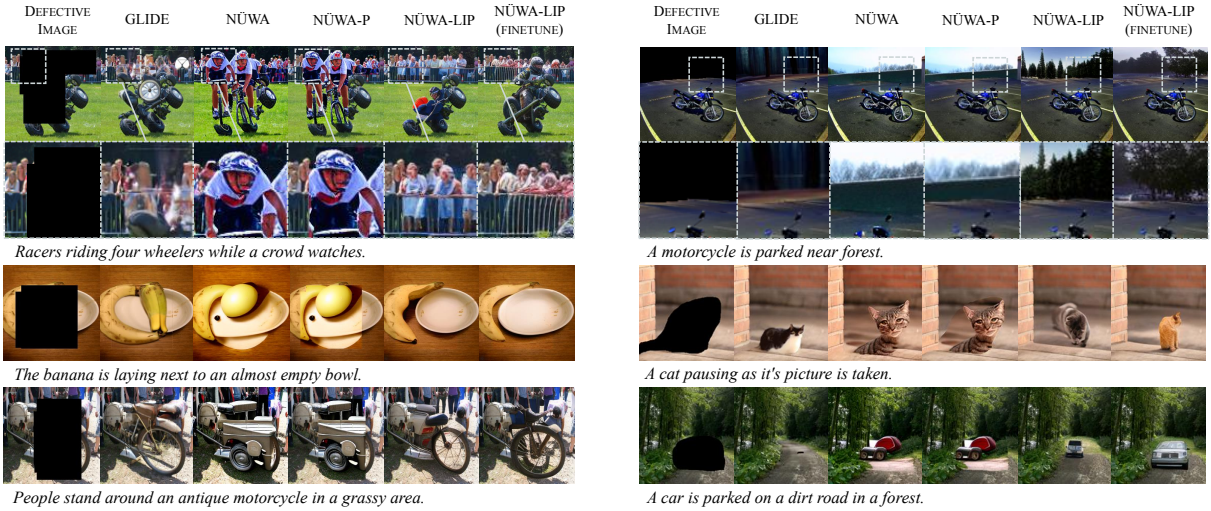

Figure 7. **Language-guided image inpainting results of different models.** NÜWA-LIP achieves the best quality compared with baselines.

VQGAN as a backbone. We re-train the MP-S2S under different settings on MSCOCO and evaluate the performance of language-guided image inpainting results on MaskCOCO. From Tab. 4, we can see that the text perspective provides the highest gain of 8.5, while the high-level tokens and low-level pixels perspective benefit the model with FID gains of 1.2 and 0.6, respectively. When we combine DF-VQGAN with MP-S2S, we obtain a significant improvement of 15.2 on FID, which demonstrates the capability of the whole framework in the language-guided inpainting task. From Fig. 6 we can also find that both our DF-VQGAN and MP-S2S contribute to the best performance of the framework.

4.6. Case Studies

We selected several cases to demonstrate the effectiveness of our proposed method for image inpainting. From Fig. 7, we can observe that 1) compared with NÜWA, NÜWA-LIP accurately preserves the hue of the whole image. 2) Compared with baselines, NÜWA-LIP achieves a better transition between non-defective and completed regions. 3) NÜWA-LIP generates visually and linguistically consistent results with more photo-realistic details. 4) Fine-tuning NÜWA-LIP further improves the quality of the inpainting. We also provide out-of-domain image inpainting results in our suppl.

5. Conclusion

In this paper, we made the first attempt to encode defective images for the language-guided image inpainting. Our NÜWA-LIP consists of a DF-VQGAN that can control receptive spreading and keep information unchanged, and an MP-S2S module that enhances the visual quality from complementary perspectives. With this design, our NÜWA-LIP can effectively adapt the text description into the defective input, making it applicable to real-world corrupted images. Besides, we also constructed three open-domain benchmarks to evaluate the performance of NÜWA-LIP against other competing methods. Experiments show our NÜWA-LIP outperforms these methods by a large margin. This suggests that NÜWA-LIP has great potential to provide users with greater flexibility in image editing and manipulation. However, it is worth noting that fake images can be abused in certain contexts, such as news reporting. As such, we leave it as future work to explore a more trustworthy model.

Acknowledgement. This work was supported by National Key RD Program of China under Grant No. 2021ZD0112100.

References

[1] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[2] 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. 1

[3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[7] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[8] 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. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[14] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 2

[15] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[16] 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, 6, 7

[17] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 2

[18] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021. 1

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 3

[20] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019. 2

[21] 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. 1

[22] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[23] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019. 5

[24] 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. 1, 2

[25] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-

training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021. 1, 3, 6, 7

[26] Xingcai Wu, Yucheng Xie, Jiaqi Zeng, Zhenguo Yang, Yi Yu, Qing Li, and Wenyin Liu. Adversarial learning with mask reconstruction for text-guided image inpainting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3464–3472, 2021. 2

[27] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[28] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 2

[29] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1

[30] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1302–1310, 2020. 2

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[32] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020. 1

[33] 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. 1, 2

[34] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 1