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

This supplemental material mainly contains:

- Discussion with Partial Convolution in Section I
- Details of DF-VQGAN and MP-S2S in Section II
- Details of proposed datasets in Section III
- User Study in Section IV
- More comparisons with other models in Section V
- More comparisons of DF-VQGAN in Section VI
- More comparisons with VQGAN in Section VII
- More inpainting results in Section VIII
- Analyses of failure case in Section IX
- Broader impact and limitations in Section X

I. Difference with Partial Convolution

We note that Liu *et al*. [6] propose a partial convolutional layer (PConv) where the convolution operation is masked and renormalized to be conditioned on only non-defective pixels for image inpainting task. It is defined as:

$$PConv(x) = \begin{cases} \mathbf{W}^T(\mathbf{x} \odot \mathbf{m}) \frac{\text{sum}(1)}{\text{sum}(\mathbf{m})} + b, & \text{if sum } (\mathbf{m}) > 0 \\ 0, & \text{otherwise} \end{cases}, \quad (1)$$

where x is the defective image and m is the mask matrix.

In contrast, we simplify the formulation of our defect-free operation in DF-VQGAN by removing the symmetrical connection, and the defect-free operations on convolution, normalization, and attention layers are defined as:

$$\begin{aligned} \mathrm{Conv}'(y) &= \mathrm{Conv}^{\mathrm{DF}}(x) \odot (1-m) + \mathrm{Conv}(y) \odot m\,, \\ \mathrm{Norm}'(y) &= \mathrm{Norm}^{\mathrm{DF}}(x) \odot (1-m) + \mathrm{Norm}(y) \odot m\,, \\ \mathrm{Attn}'(y) &= \mathrm{Attn}^{\mathrm{DF}}(x) \odot (1-m) + \mathrm{Attn}(y) \odot m\,, \end{aligned} \quad (2)$$

where y represents the ground-truth image and $x = y \odot m$. $\mathrm{Conv}^{\mathrm{DF}}$, $\mathrm{Norm}^{\mathrm{DF}}$, and $\mathrm{Attn}^{\mathrm{DF}}$ are the defect-free operations, which are defined as:

DEFECTIVE IMAGE | GROUND TRUTH | PCONV | DF-VQGAN

Figure A. **Comparison with PConv and DF-VQGAN.** We incorporate the PConv operation into VQGAN and compare the result of oracle inpainting with our DF-VQGAN. We can observe that result of PConv easily generate distorted structures in defective regions.

$$\begin{aligned} \mathrm{Conv}^{\mathrm{DF}}(x) &= W_c^\top(x \odot m) + b, \\ \mathrm{Conv}(y) &= W_c^\top(y) + b, \\ \mathrm{Norm}^{\mathrm{DF}}(x) &= \frac{x - \frac{N}{N_m}\mathrm{E}[x]}{\sqrt{\frac{N-1}{N_m-1}\mathrm{Var}[x'] + \epsilon}}, \\ \mathrm{Norm}(y) &= \frac{y - \mathrm{E}[y]}{\sqrt{\mathrm{Var}[y] + \epsilon}}, \\ \mathrm{Attn}^{\mathrm{DF}}(x) &= \mathrm{Softmax}(x^\top W_a x) \odot m \odot x, \\ \mathrm{Attn}(y) &= \mathrm{Softmax}(y^\top W_a y) \odot y, \end{aligned} \quad (3)$$

where N and N_m are the numbers of all pixels and defective pixels in x, respectively. x' is the revised x with the defective region fulfilling with $\frac{N}{N_m}\mathrm{E}[x]$. $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ denote expectation and variance, respectively. W_c (W_a) is shared parameters between $\mathrm{Conv}^{\mathrm{DF}}$ and Conv ($\mathrm{Attn}^{\mathrm{DF}}$ and Attn).

Comparing Eqn. (1) with Eqns. (2-3), our DF-VQGAN has two differences with PConv: (1) DF-VQGAN is a VAE model, which is trained by reconstructing a full image $\hat{y}$ from a full image y. For adopting VAE in the inpainting task, we need to introduce mask m carefully without destroying the schema of VAE. However, PConv is not a VAE model, and it takes the defective image x as input to predict a inpainted result $\hat{y}$. (2) Instead of performing on the convolution layer only, our DF-VQGAN focuses on three operations, which may easily lead to receptive spreading. This allows our DF-VQGAN effectively learn the valid features from defective input and reconstruct results with high fidelity. To validate the effectiveness of DF-VQGAN, we add the PConv operation to VQGAN and train it with the same setting as DF-VQGAN. We also provide quantitative and qualitative comparison with PConv in Fig. A and Tab. C. We can see that PConv tends to generate modified hue and distorted structures while our DF-VQGAN can generate results with better quality, indicating the effectiveness of our defect-free operation.

II. Details of DF-VQGAN and MP-S2S

DF-VQGAN. It adopts the settings of the vanilla VQ-GAN [4]. The vocab size is set to $8,192$, and the learning rate is 5×10^{-6}. The batch size and the dim of the latent token is set to 200 and 256, respectively. The input resolution is 256×256 and DF-VQGAN encodes the input to 32×32 tokens. We pretrain the DF-VQGAN with ImageNet [3].

MP-S2S. The layer number in encoder $E^{\rm l}$, $E^{\rm h}$ and $E^{\rm t}$ is set to 12, respectively. The layer number in the autoregressive decoder is 24. All Transformers have 20 heads and the hidden size is 1024. We set the learning rate to 5×10^{-4}, and the batch size to 320. The text encoder $E^{\rm t}$ and tokenizer is initialized with the text encoder and tokenizer from pre-trained CLIP [8]. The $E^{\rm l}$ and $E^{\rm h}$ are trained from scratch. We use Conceptual Captions [10] as the pre-training corpus.

III. Details of Proposed Datasets

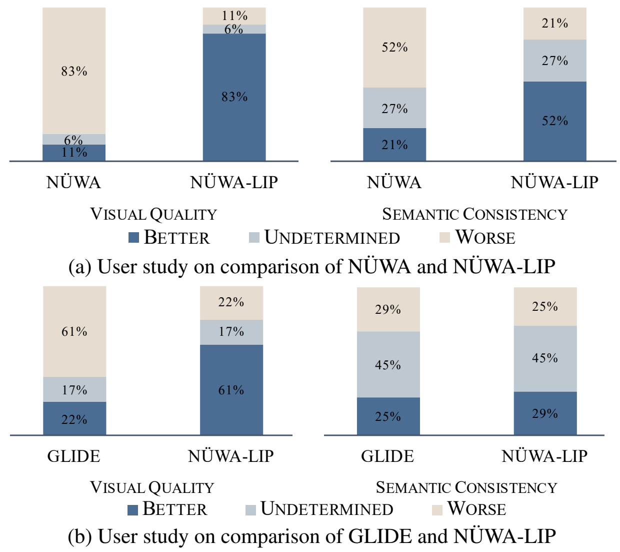
We follow [5] and select the test sets of MSCOCO and Flickr to build our MaskCOCO and MaskFlickr. As for MaskVG, we randomly select $10,000$ samples from the VG dataset. For each image-text pair, the original image and corresponding caption are considered ground-truth images and text descriptions. Each image will be cropped and resized to the resolution of 256×256. The defective image is generated by masking with either one bounding box of the object or a random irregular region. The details of the proposed three datasets are listed in Tab. A. We will release them under a Creative Commons Attribution 4.0 License.

Table A. **Details of the evaluation datasets.**

DATASET	IMAGE-TEXT PAIRS	MASK RATIO
MASKCOCO	5000	31.5%
MASKFLICKR	1000	48.3%
MASKVG	10000	14.6%

IV. User Study

To further evaluate the quality of our NÜWA-LIP and baselines, we conduct a user study from real human perception. We randomly select 500 samples from the MaskCOCO dataset and compare with NÜWA-LIP, NÜWA, and GLIDE



(a) User study on comparison of NÜWA and NÜWA-LIP

(b) User study on comparison of GLIDE and NÜWA-LIP

Figure B. **User study of NÜWA-LIP and baselines.**

on two aspects, *i.e.*, visual quality and semantic consistency. The visual quality focuses on evaluating the structures whether they are photo-realistic or contain distorted details. The semantic consistency assesses whether the inpainted results have semantically consistent content with the language guidance. Volunteers with a computer vision background are required to give a choice about which one has better quality. From Fig. B, we can observe that compared with these competing methods (*i.e.*, NÜWA and GLIDE), our NÜWA-LIP has obvious better performance in both visual quality and semantic consistency, which indicates that our NÜWA-LIP can generate more photo-realistic and consistent results.

V. Comparisons with Other Models

To explore the effectiveness of NÜWA-LIP, we conduct additional experiments with other inpainting models. Specifically, we compare NÜWA-LIP with TDANET [13], a popular non-pre-trained language-guided inpainting model, and MASKGIT [2] and LAMA [11], which are class-conditional and unconditional inpainting models pre-trained on large-scale data, respectively. For MASKGIT, we use CLIP to classify the class of the ground-truth image as the input. As shown in Tab. B, NÜWA-LIP outperforms all these models, showing its effectiveness and the essentials of the language.

Discussion with Stable Diffusion STABLE DIFFUSION is an effective model for visual synthesis tasks. Fig. C shows the difference between STABLE DIFFUSION and most prior image inpainting works [2, 11, 13]. In general image inpainting settings, the input image for the inpainting model is defective or damaged. However, the input image of STABLE DIFFUSION should be normal images without defective

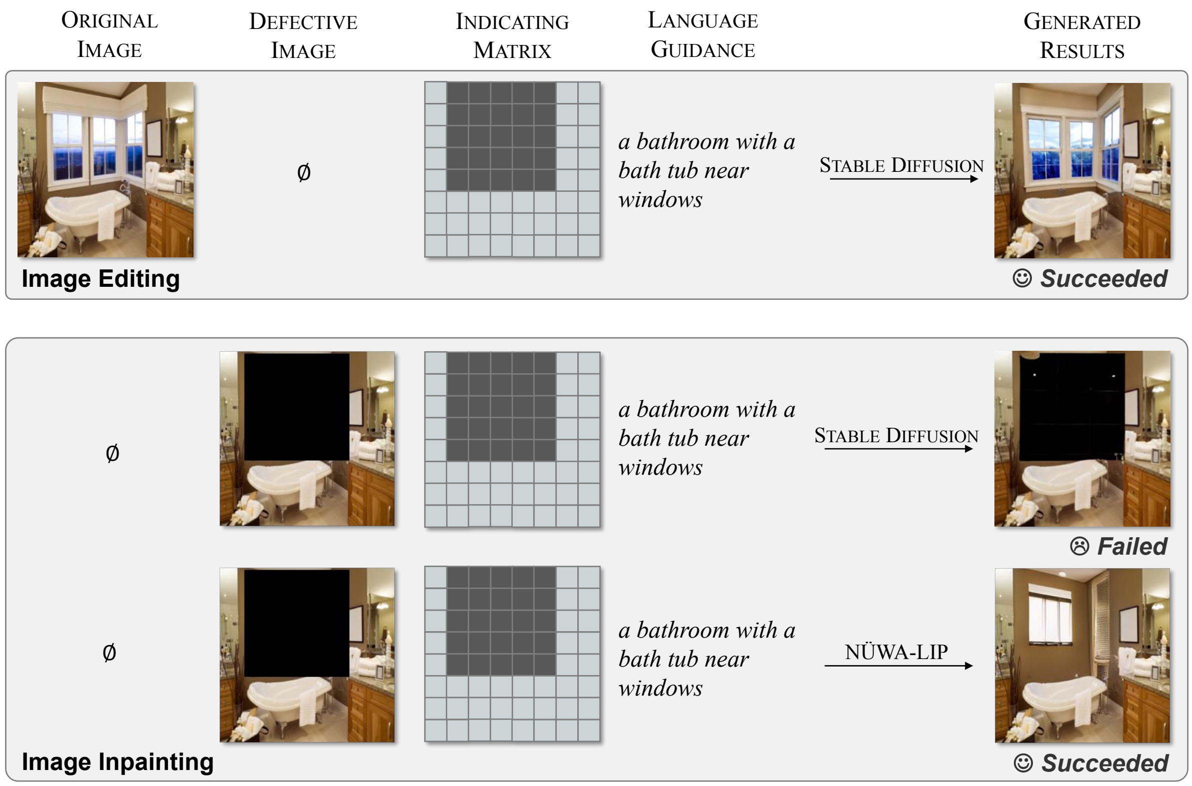


Figure C. **Comparison of STABLE DIFFUSION and NÜWA-LIP pipelines.** Different from most prior works, the input image of STABLE DIFFUSION needs to be well-formed, which is called image editing in most related works.

Table B. **Comparsion with different models on MaskCOCO.** † denotes trained or finetuned on COCO.

MODEL	FID↓	CLIP SCORE↑
STABLE DIFFUSION [9] (IMAGE EDITING)	10.9	30.38
TDANET† [13]	27.2	27.90
LAMA [11]	17.3	24.38
MASKGIT [2]	15.5	27.20
NÜWA-LIP (W/O PRETRAIN)†	**11.0**	**28.74**
NÜWA-LIP	**12.0**	**29.34**
NÜWA-LIP (FINETUNE)†	**10.5**	**29.65**

or damaged regions, which is called image editing in most related works [2, 7, 12]. From Fig. C, we can find that STABLE DIFFUSION is hard to directly handle these types of defective images. For a fair comparison, the input image of STABLE DIFFUSION is set to the ground-truth without corrupted regions, while ours and other baselines take the occluded image as input for the general inpainting task. Here we use their official implementation from DIFFUSERS[1] and checkpoints[2] on this type of defective input. Besides, STABLE DIFFUSION is trained on the LAION-5B dataset, which is 1000× larger than ours. From Tab. B we can observe that our method still obtains comparable performance.

VI. More Comparisons of DF-VQGAN

Table C. **More quantitative comparisons of DF-VQGAN on ImageNet.** DF-VQGAN outperforms VQGAN or VQGAN-P on both image reconstruction (IMG.REC) and oracle inpainting (ORC.INP).

MODEL	RESOLUTION	TOKEN LENGTH	VOCAB SIZE	IMG.REC	ORC.INP
VQGAN	$256^2 \rightarrow 16^2$	256	12288	6.03	7.15
VQGAN-P	$256^2 \rightarrow 16^2$	256	12288	6.03	3.77
PARTIAL CONV.	$256^2 \rightarrow 16^2$	256	12288	6.83	7.14
TS-VQGAN	$256^2 \rightarrow 16^2$	256	12288	5.89	6.47
DF-VQGAN/S	$256^2 \rightarrow 16^2$	256	12288	**5.14**	**5.44**
DF-VQGAN	$256^2 \rightarrow 16^2$	256	12288	**5.16**	**2.95**

To validate whether ***relative estimation*** can avoid receptive spreading of defective regions, we compare VQGAN with DF-VQGAN/S, which is DF-VQGAN without ***symmentrical connection***. In the upper part of Tab. C, we can find that we significantly reduce the FID score from 7.15

[1] https://github.com/huggingface/diffusers

[2] We use the best SD-V1-4 checkpoint.

to 5.44 in the oracle inpainting task. Besides, the gain in image reconstruction can be ascribed to the usage of ***relative estimation*** in improving the robustness of the model.

We further validate whether ***symmentrical connection*** can protect the information of non-defective regions. We compare DF-VQGAN with VQGAN-P, which directly copies and pastes the non-defective region of the image into the generated results. In the bottom part of Tab. C, we can find that we achieve a better FID score (*i.e.*, 2.95 v.s. 3.77) in oracle inpainting task, which indicates that our symmetrical connection can make a significantly better transition between the non-defective region and inpainted part. Meanwhile, we obtain a comparable FID score of 5.16 in the image reconstruction task. The comparison with DF-VQGAN/S shows the benefits of combining ***relative estimation*** and ***symmentrical connection*** in image inpainting task.

Finally, we conduct the comparison with TS-VQGAN [1], which is used in conditional image synthesis to encode an image without defective regions. The goal of TS-VQGAN is to avoid information leaking, which means results are more similar to reference images rather than the target condition. Different from TS-VQGAN, DF-VQGAN works in the image inpainting scenario, in which defective regions and non-defective regions exist at the same time in an image. From Tab. C, we can observe that our approach still outperforms TS-VQGAN with a large margin in oracle inpainting.

VII. More Comparisons with VQGAN

We provide more visual results in Fig. D to compare our DF-VQGAN with vanilla VQGAN and analyze their performance on both defective and non-defective regions. We can find that our DF-VQGAN can well capture the semantic details and generate consistent structures in defective regions. More importantly, our DF-VQGAN can well keep the non-defective content unchanged.

VIII. More Inpainting Results

We provide more inpainting results to show the effectiveness of NÜWA-LIP in Fig F. We can observe that NÜWA-LIP can leverage the text guidance well and generate results with higher fidelity and better consistency.

IX. Failure Case

Although NÜWA-LIP shows effectiveness in most cases, we find that it may fail in some cases like Fig. E, which shows fine-tuning will cause failure in inpainting some rare objects. In most cases, fine-tuning brings impressive improvement in the quality of the inpainted images but may fail in some objects which occurs very little in the fine-tuning dataset. We will balance the distribution of each object and augment these with fewer samples.

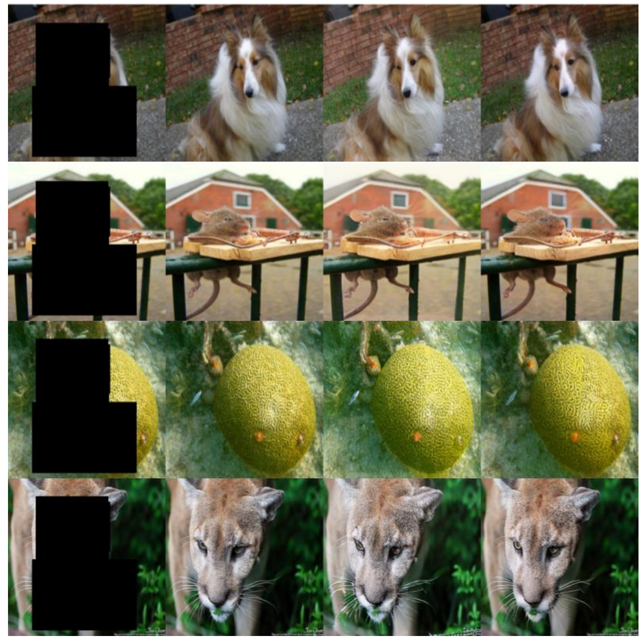


Figure D. **More illustration on oracle inpainting.** DF-VQGAN shows better ability in generating consistent details in defective regions and keeping non-defective regions unchanged.

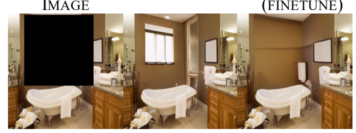


Figure E. **Failure case of NÜWA-LIP.** The failure case may be caused by the rare objects in the fine-tuning dataset.

X. Broader Impact and Limitations

NÜWA-LIP, which is an effective model for language-guided image inpainting, can provide the potential for users to edit and manipulate an image, which may lead to destructive behaviors, *e.g.*, fake images may be abused in some cases like news reporting. We will explore a more trustworthy model to prevent such abuse cases. Besides, as a common issue of autoregressive models, handling an extremely large image would have a much higher computational cost, and may be easy for users to retrain this model.

Figure F. **More inpainting results.** NÜWA-LIP can effectively complete the defective image under the guidance of different texts.

References

[1] Chenjie Cao, Yuxin Hong, Xiang Li, Chengrong Wang, Chengming Xu, Yanwei Fu, and Xiangyang Xue. The image local autoregressive transformer. *Advances in Neural Information Processing Systems*, 34:18433–18445, 2021. 4

[2] 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, 3

[3] 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. 2

[4] 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

[5] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[6] 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. 1

[7] 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. 3

[8] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2

[9] 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. 3

[10] 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. 2

[11] 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. 2, 3

[12] 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. 3

[13] 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, 3