Project Page
Code&Model

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



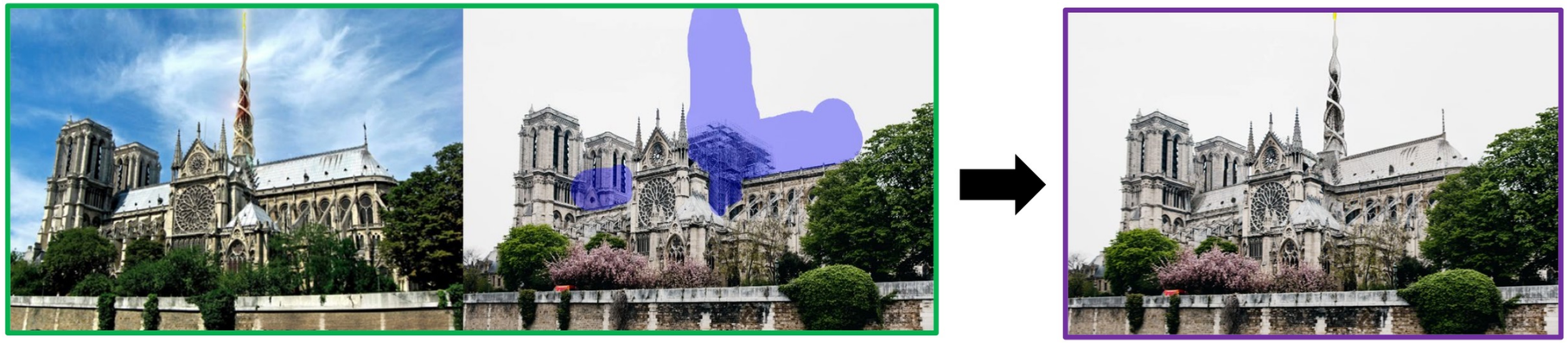


Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, Yanwei Fu

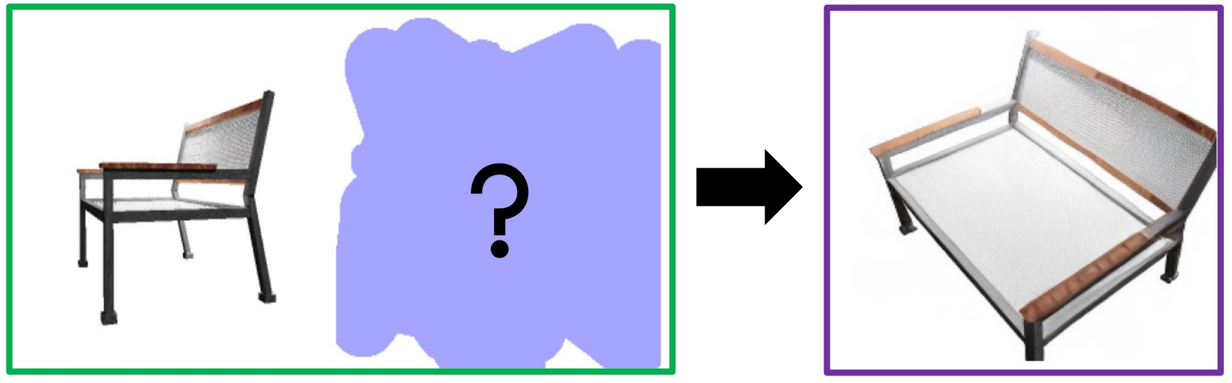
{cjcao20,yncai20,qldong18,yikaiwang19,yanweifu}@fudan.edu.cn

Our Focus: Reference-image(s) Based Generation

- Local Reference: Reference-based Inpainting



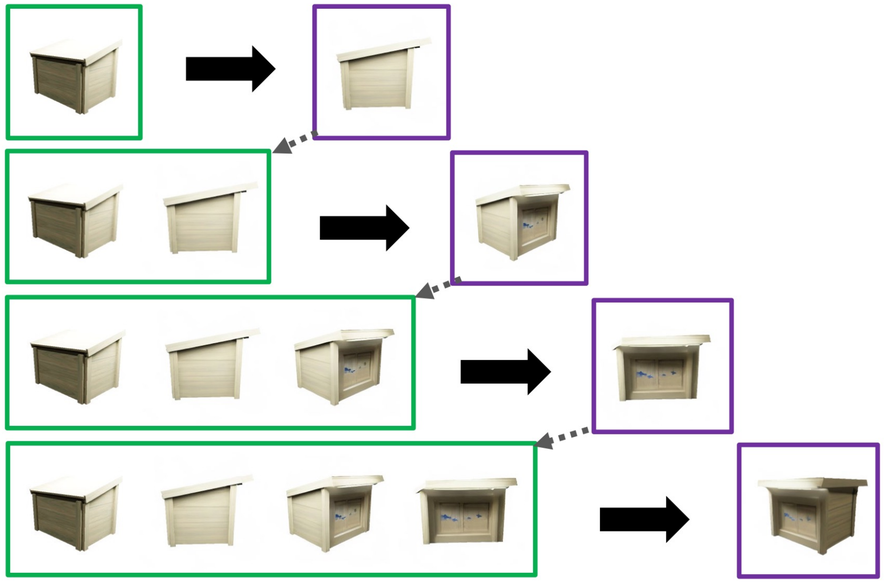
- Global Reference: Novel-View Synthesis



Our Focus: Reference-image(s) Based Generation

- Multi-References

- Multi-Targets



How to Inject Reference Image?

Stable Diffusion
Generation → Inpainting

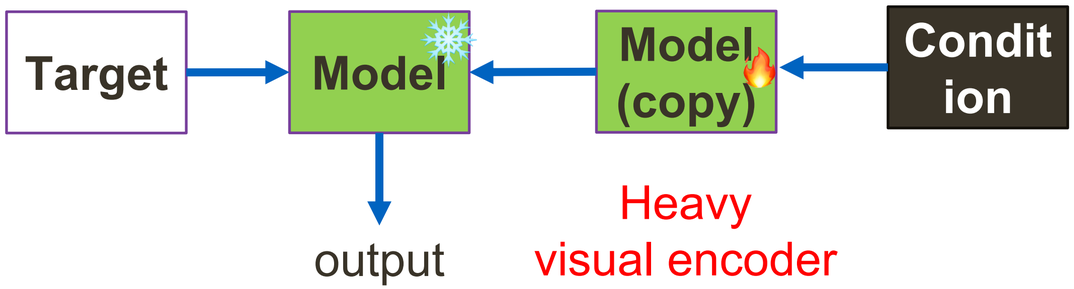
- ◆ Concatenate on the channel dimension
- ● Heavy Training Cost

Inputs of
Generation → Inpainting

Control Net

- ◆ Concatenate on the (internal) output space
- ● Heavy visual encoder.

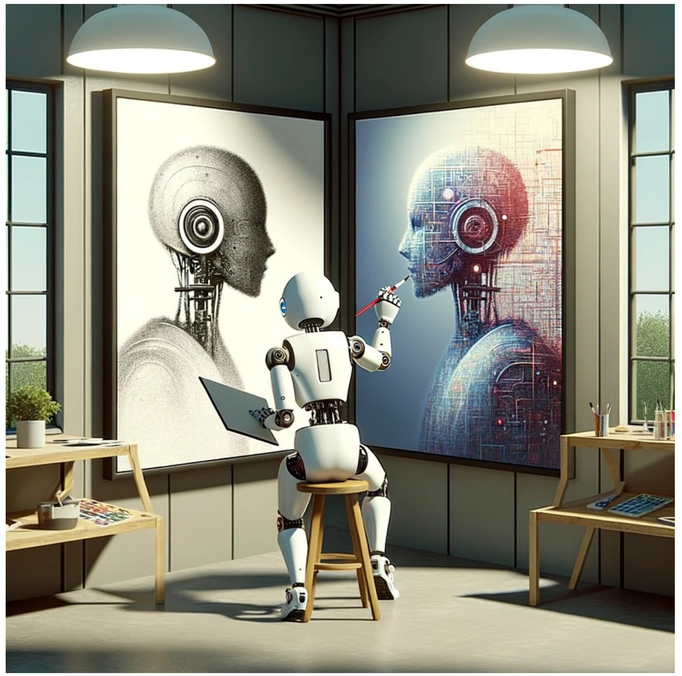
ControlNet



Unified Filling Right Canvas based on Left Reference

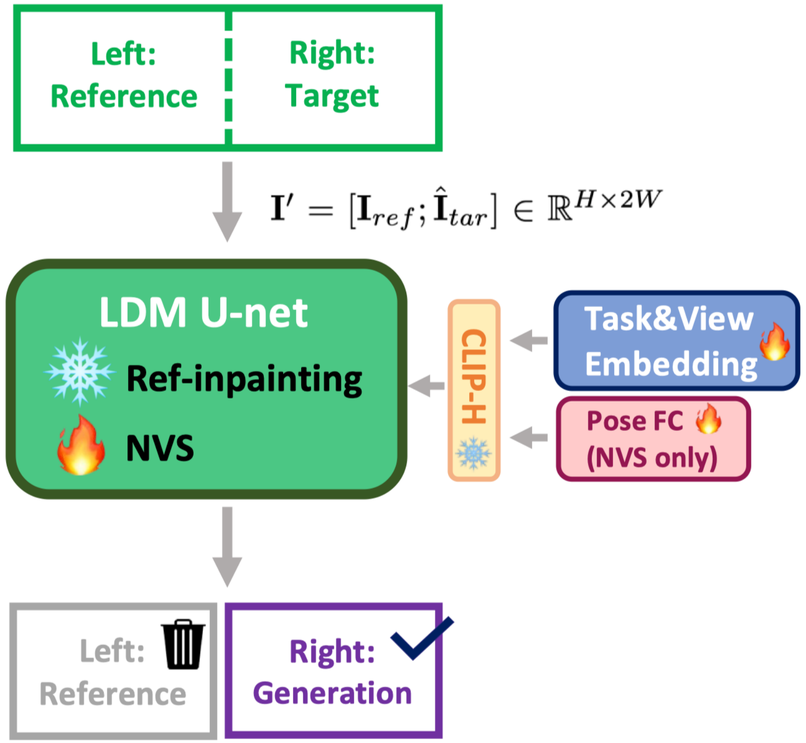
- LeftRefill:
  - Concatenate on the spatial channel, **contextual inpainting**
- Highlights:
  - Lightweight & Small training cost
  - Reuse off-the-shelf Text-to-Image models

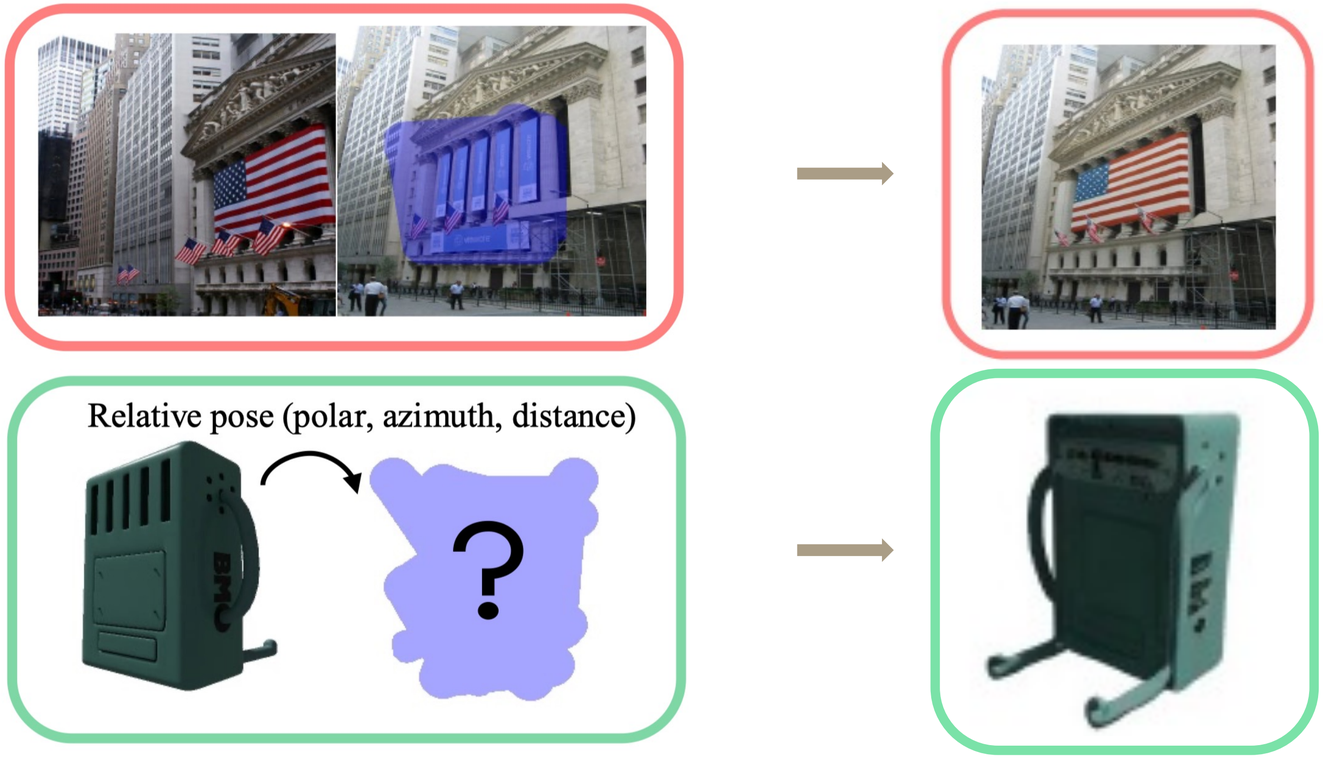
Use the self-attention layers to guide the generation!



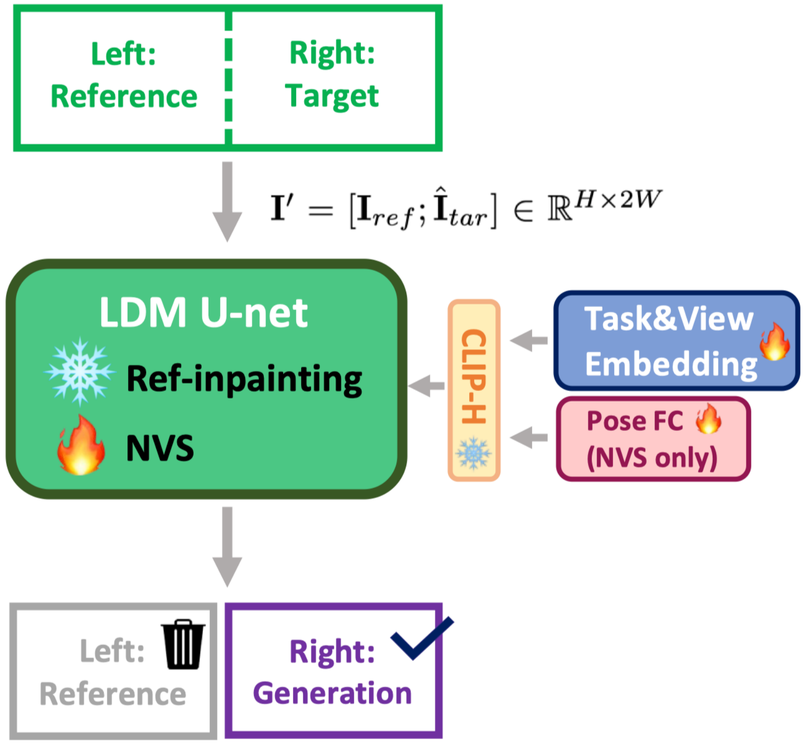
LeftRefill paints right side of canvas referring to left part as a human painter (image generated by DALLE3)

Framework


Left:
Reference
Right:
Target
$\mathbf{I}' = [\mathbf{I}_{ref}; \hat{\mathbf{I}}_{tar}] \in \mathbb{R}^{H \times 2W}$
LDM U-net
Ref-inpainting
NVS
CLIP-H
Task&View
Embedding
Pose FC
(NVS only)
Left:
Reference
Right:
Generation


Relative pose (polar, azimuth, distance)
?

Adapting Text-to-Image Diffusion Models

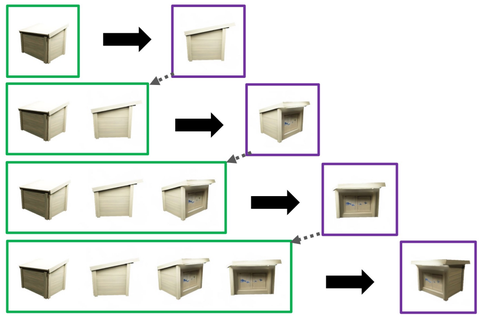


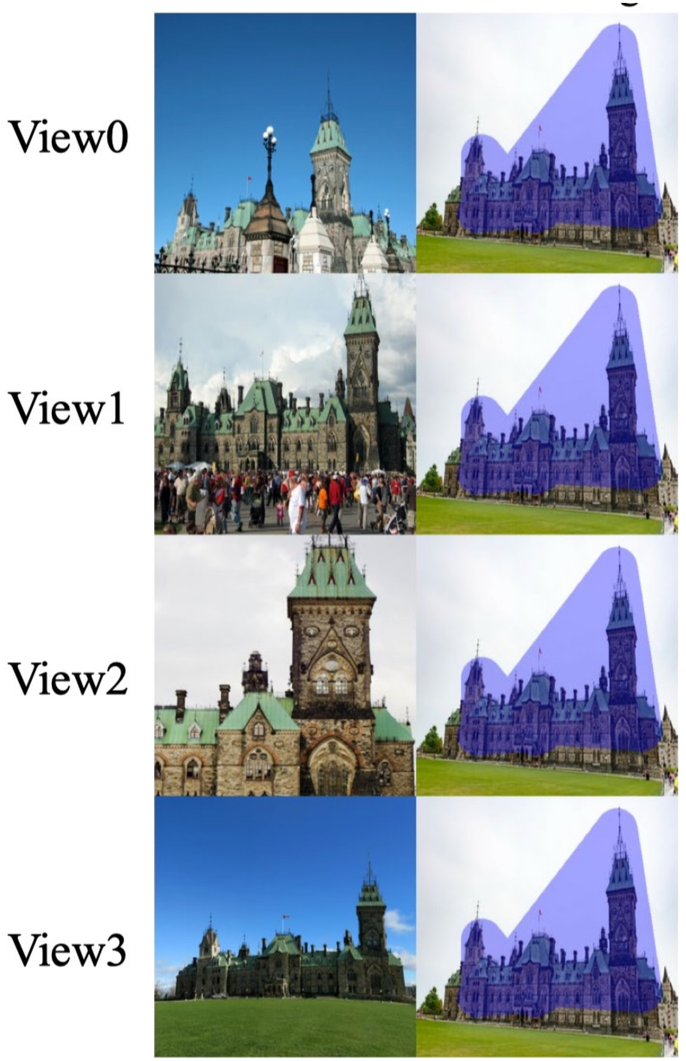
- Input:
  - Stitched reference and target images;
- Condition:
  - Use learnable prompts to simulate text prompts;
    - Task Embedding
    - View Embedding
    - Pose Embedding (NVS only)
- Output:
  - Discard the reference image.

Extension to Multi-View: Data Formulation

Multi-References

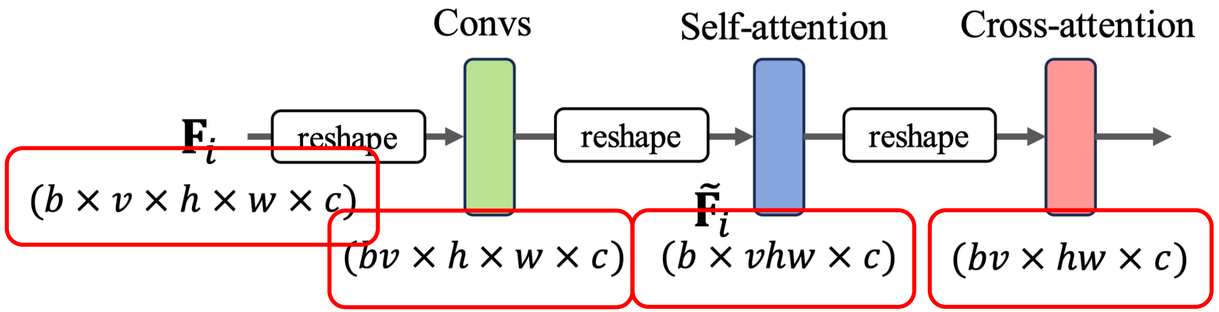
Multi-Targets

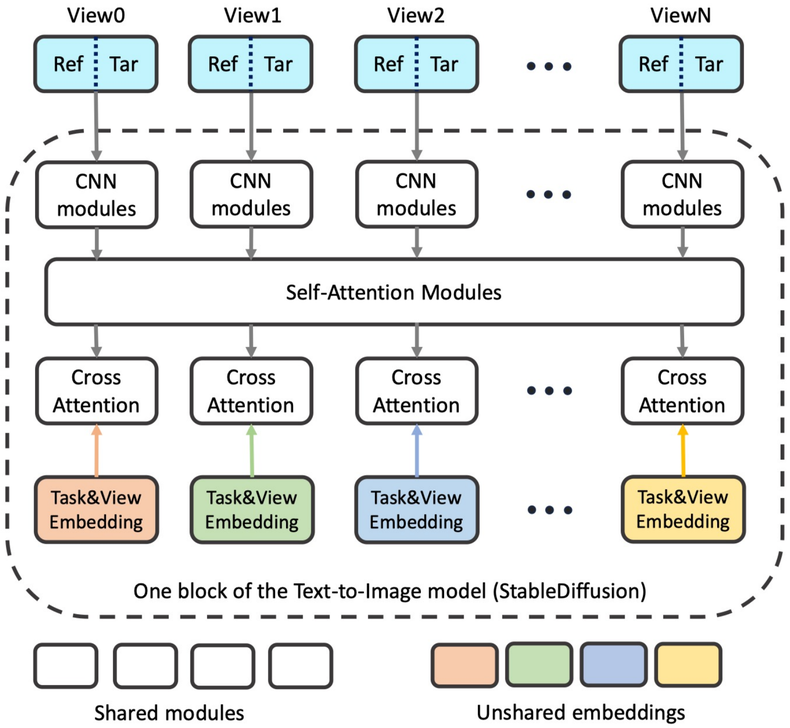


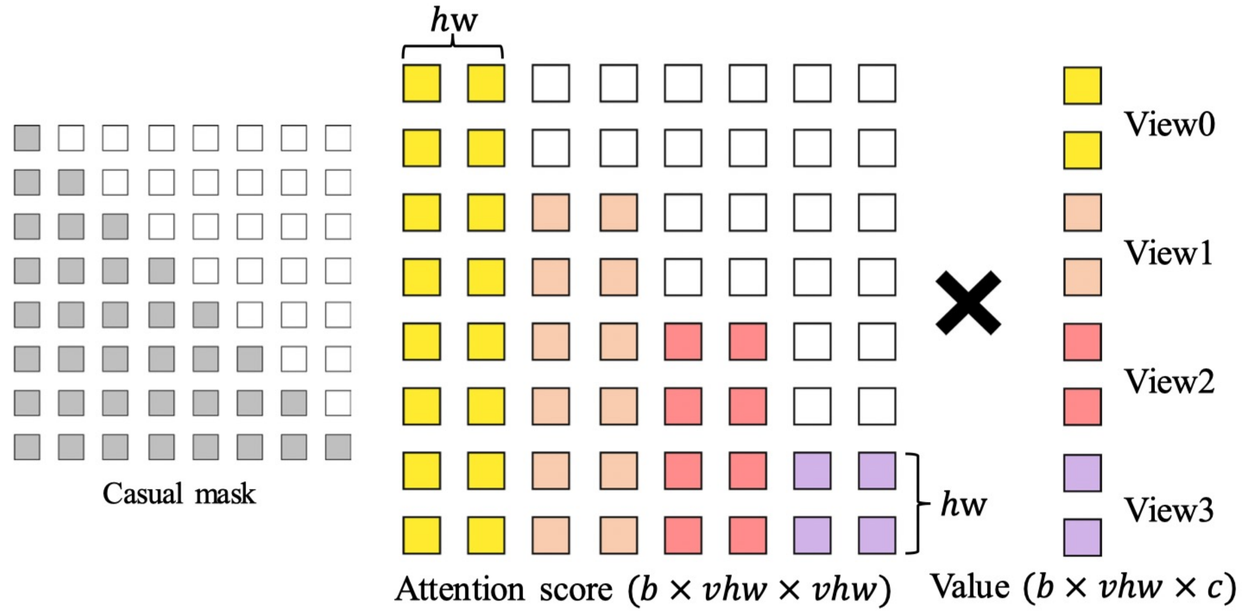


(a) Ref-inpainting (v-to-1)

Extension to Multi-View: Model Formulation

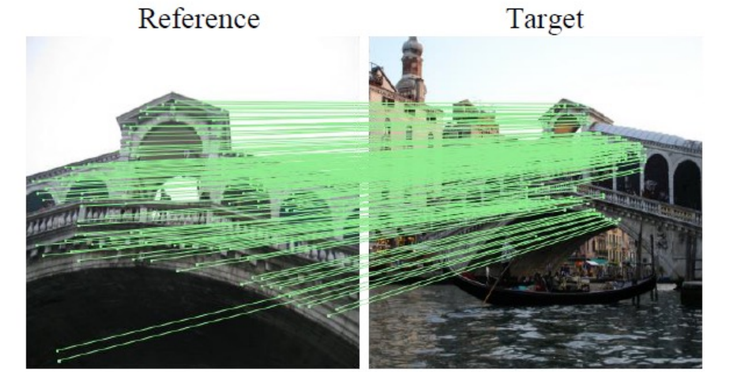




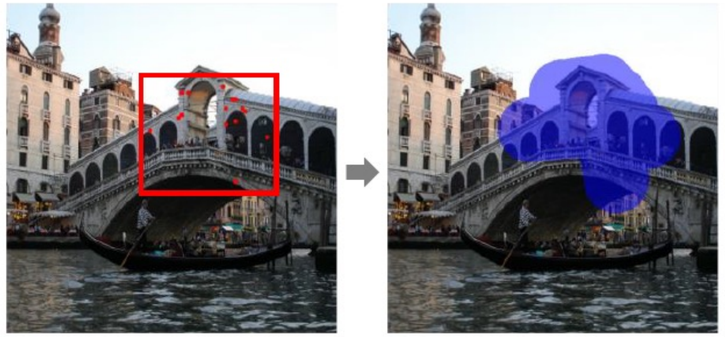


Block causal mask for autoregressive training

Generating Training Masks

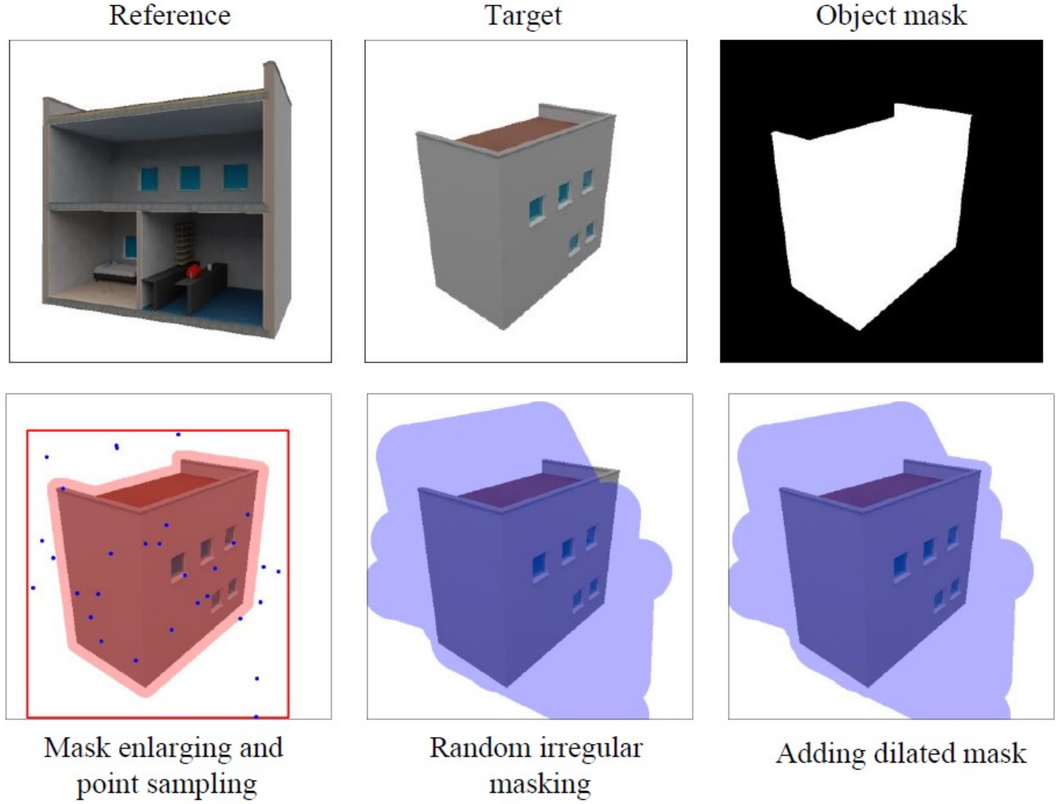


Matching and confidence filtering



Cropping and points sampling Randomly-Painted masking

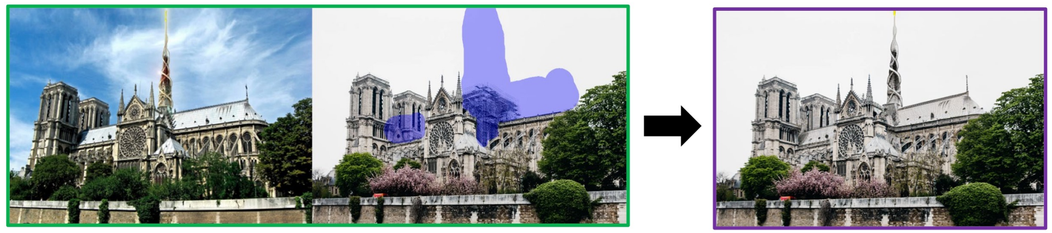
(a) Matching-based Masking



Mask enlarging and point sampling Random irregular masking Adding dilated mask

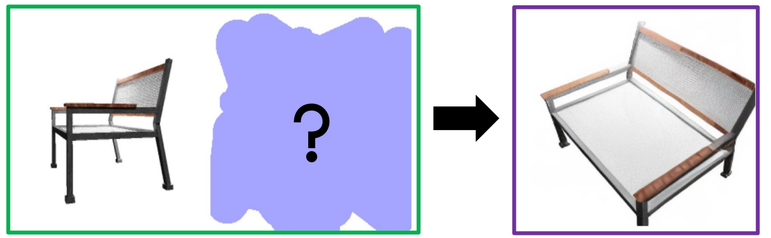
(b) Masking for Objaverse images

Training Setup



Reference-Based Inpainting

- Training Dataset:
  - MegaDepth: 512x512, 80k images (820k pairs);
- Trainable Parameters:
  - Task Embedding: 45 tokens
  - View Embedding: 5 tokens
- Training Configuration:
  - 1-view: 6k steps of batch size 16;
  - 4-view: 16k steps of batch size 64.



Novel-View Synthesis

- Training Dataset:
  - Objaverse: 256x256, 800k scenes;
- Trainable Parameters:
  - Task Embedding: 45 tokens
  - View Embedding: 5 tokens
  - Pose Embedding: 1 token (generated by the 4-channel relative pose input to 2-layer FC)
  - Fine-Tune U-Net
- Training Configuration:
  - 1-view: 80k steps of batch size 48;
  - 4-view: 110k steps of batch size 512.

Ref-Inpainting Results

(a) Reference (b) Masked target (c) SD (d) Control+Match (e) Perceiver (f) Paint-by-Example (g) TransFill (h) **LeftRefill**

Qualitative and multi-view results

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

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	20.125	0.1085	+0.05M
LeftRefill	20.926	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

Novel-View Synthesis Results

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

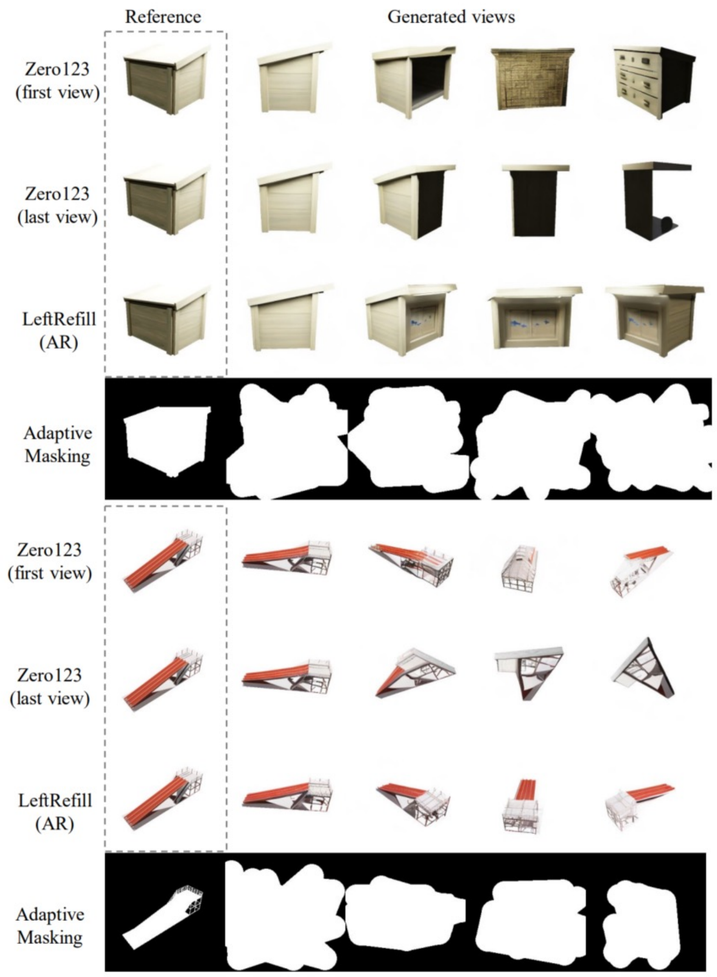
Table 2. Results of 1-view NVS conditioned on different numbers of reference views on Objaverse [10]. 'w.o. AM' indicates NVS results without Adaptive Masking (AM).

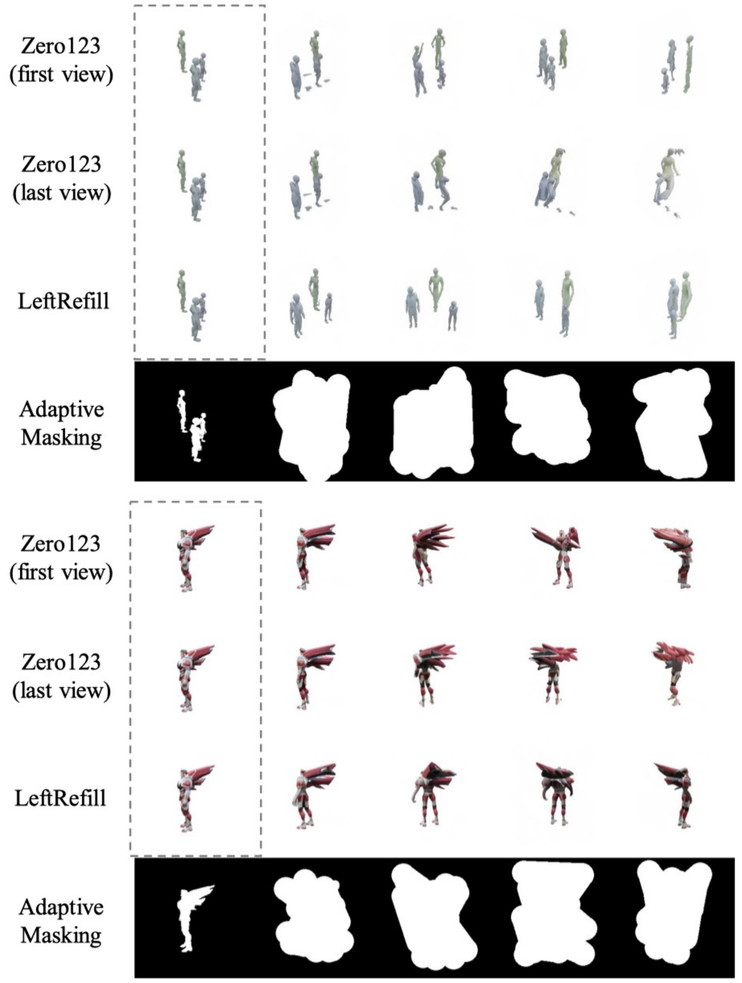
Methods	Ref-View	PSNR↑	SSIM↑	LPIPS↓	CLIP↑
Zero123 [30]	1	19.402	0.858	0.1309	0.7816
LeftRefill (LoRA)	1	19.514	0.869	0.1534	0.7589
LeftRefill (w.o. AM)	1	21.675	0.887	0.1089	0.7959
LeftRefill	1	**21.404**	**0.882**	**0.1151**	**0.7972**
LeftRefill	2	22.935	0.895	0.0871	0.8280
LeftRefill	3	24.107	0.908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

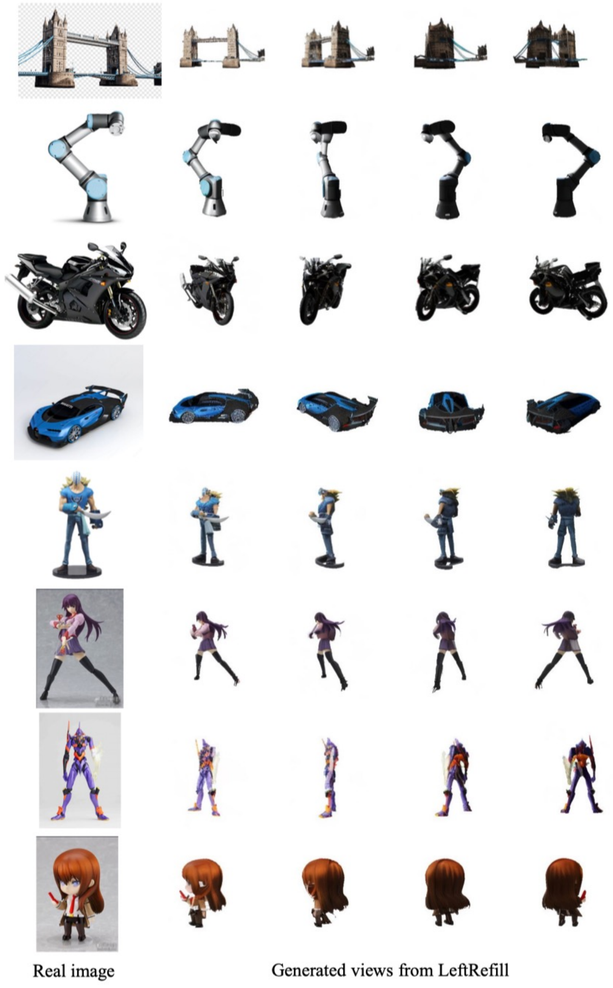
Table 3. Results of 4-view NVS generations based on 1 reference view on Objaverse [10]. P-CLIP means pairwise CLIP score showing consistency of generated views. The reference (Ref) can be categorized into the first ground-truth view and the last generated view, while we also provide the AR results of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

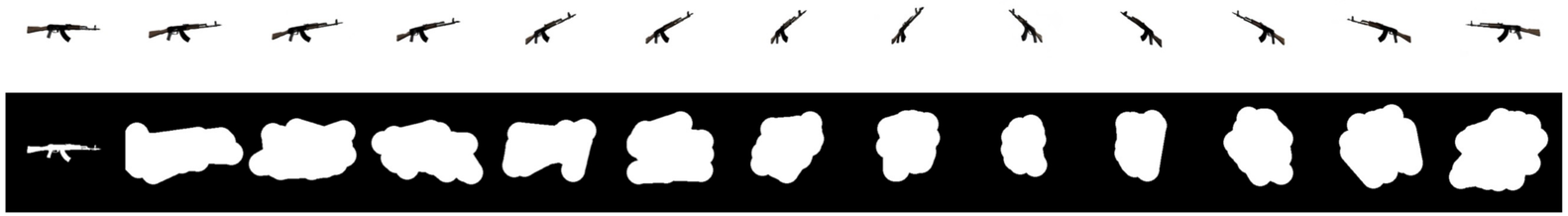
Methods	Ref	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	P-CLIP↑
Zero123	First	19.265	0.855	0.1366	0.7723	0.7756
Zero123	Last	14.621	0.767	0.2569	0.6921	0.7667
LeftRefill	First	**21.573**	**0.883**	**0.1143**	**0.7964**	0.7709
LeftRefill	AR	21.271	0.882	0.1195	0.7882	**0.7958**

Multi-View NVS

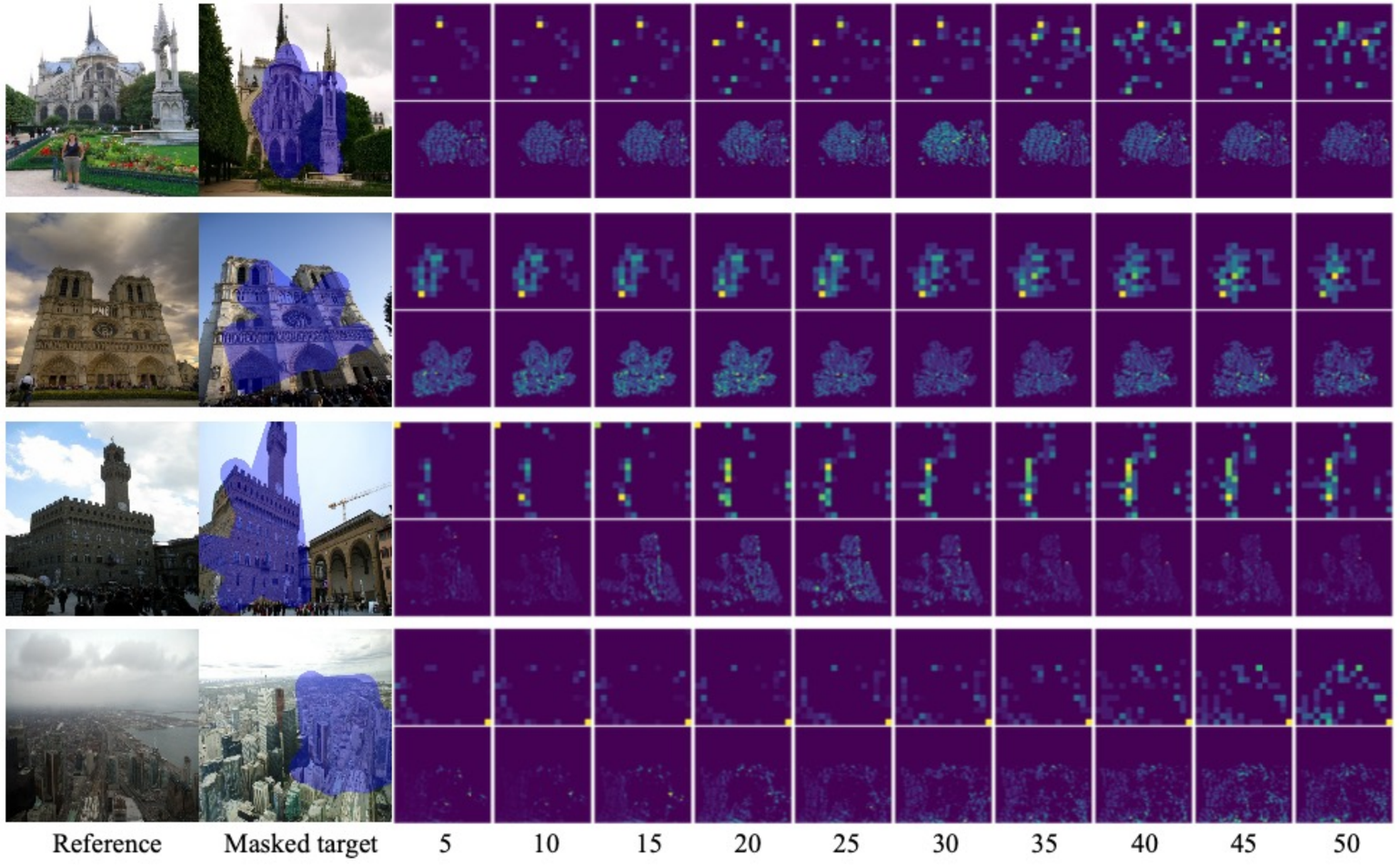








Self-Attention Analysis



Visualization of attention scores in LeftRefill across different DDIM steps.

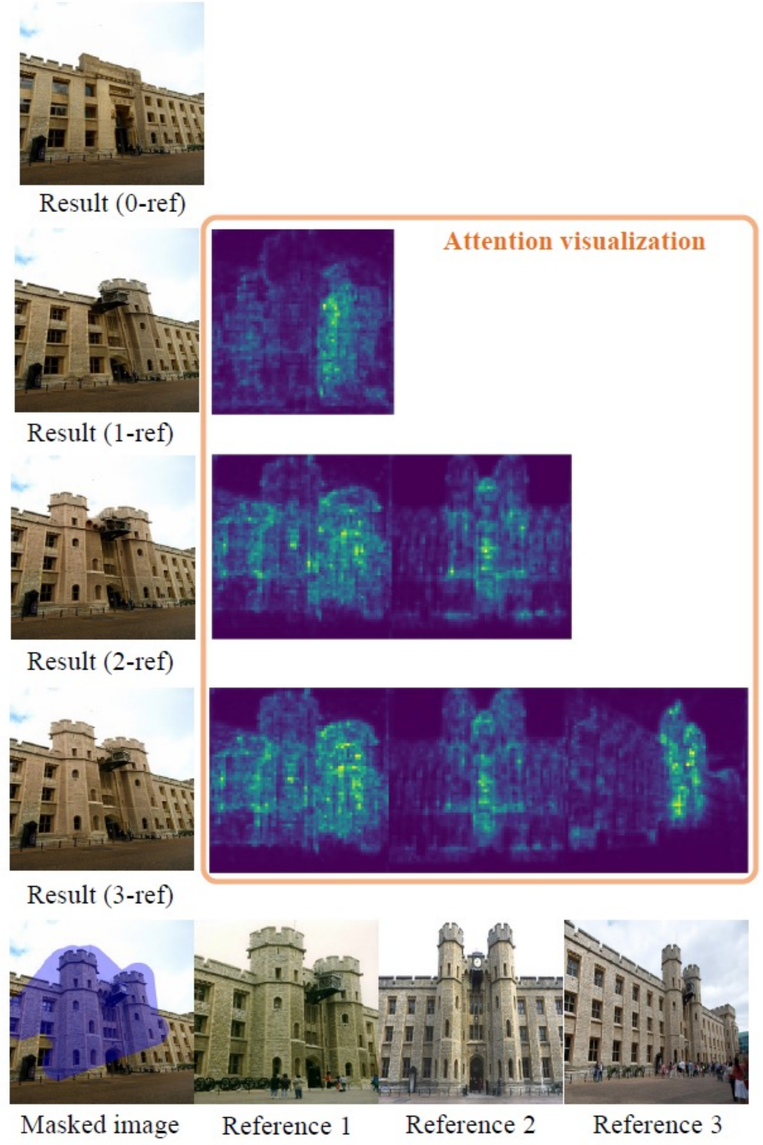


Figure 9. Attention visualization (Algorithm 2) with increased reference views.

Influence of CFG (geometry and texture trade-off)

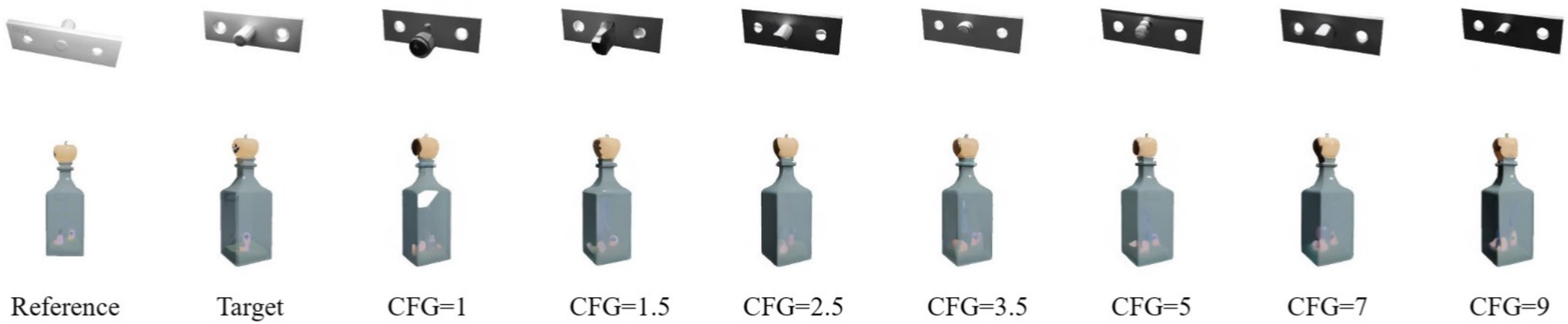


Figure 14. NVS on Objaverse [9] with different CFG weights.

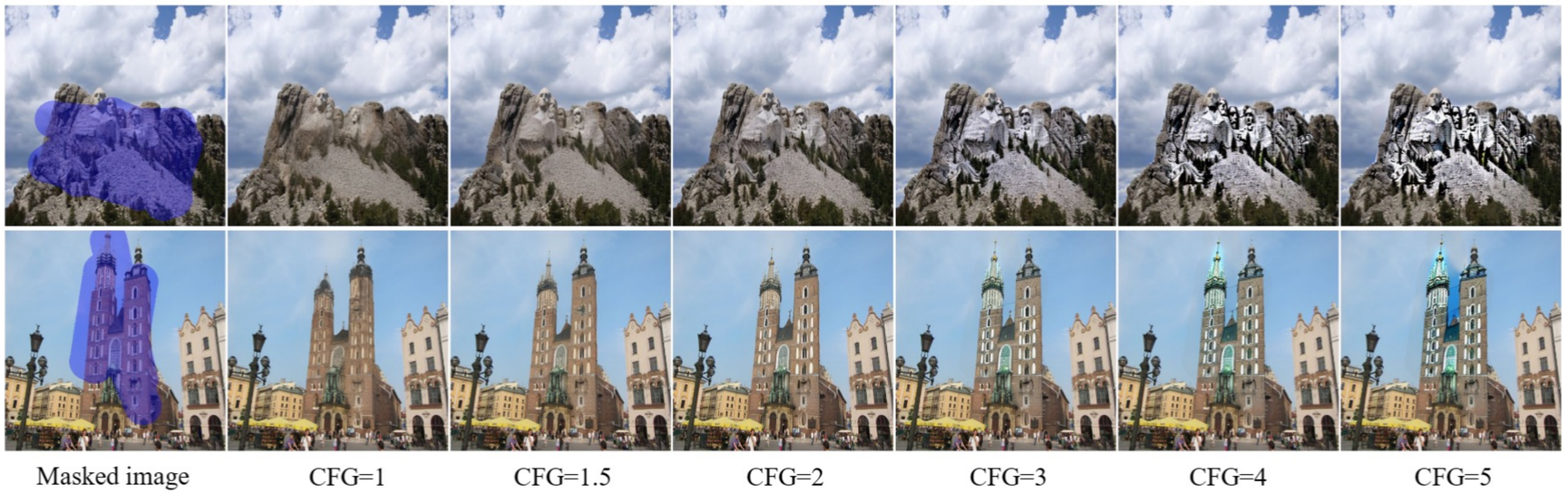


Figure 15. Ref-inpainting results with different CFG weights, which trade-off between structural and textural recoveries.

Summary

- Lightweight and generalized task formulation based on off-the-shelf T2I models;
- Use specific (task, view, pose) prompt tuning to precisely control the generation process;
- LeftRefill could be easily extended to tackle multi-view generation tasks.

Project Page
Code&Model

Thanks!

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





Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, Yanwei Fu

{cjcao20,yncai20,qldong18,yikaiwang19,yanweifu}@fudan.edu.cn