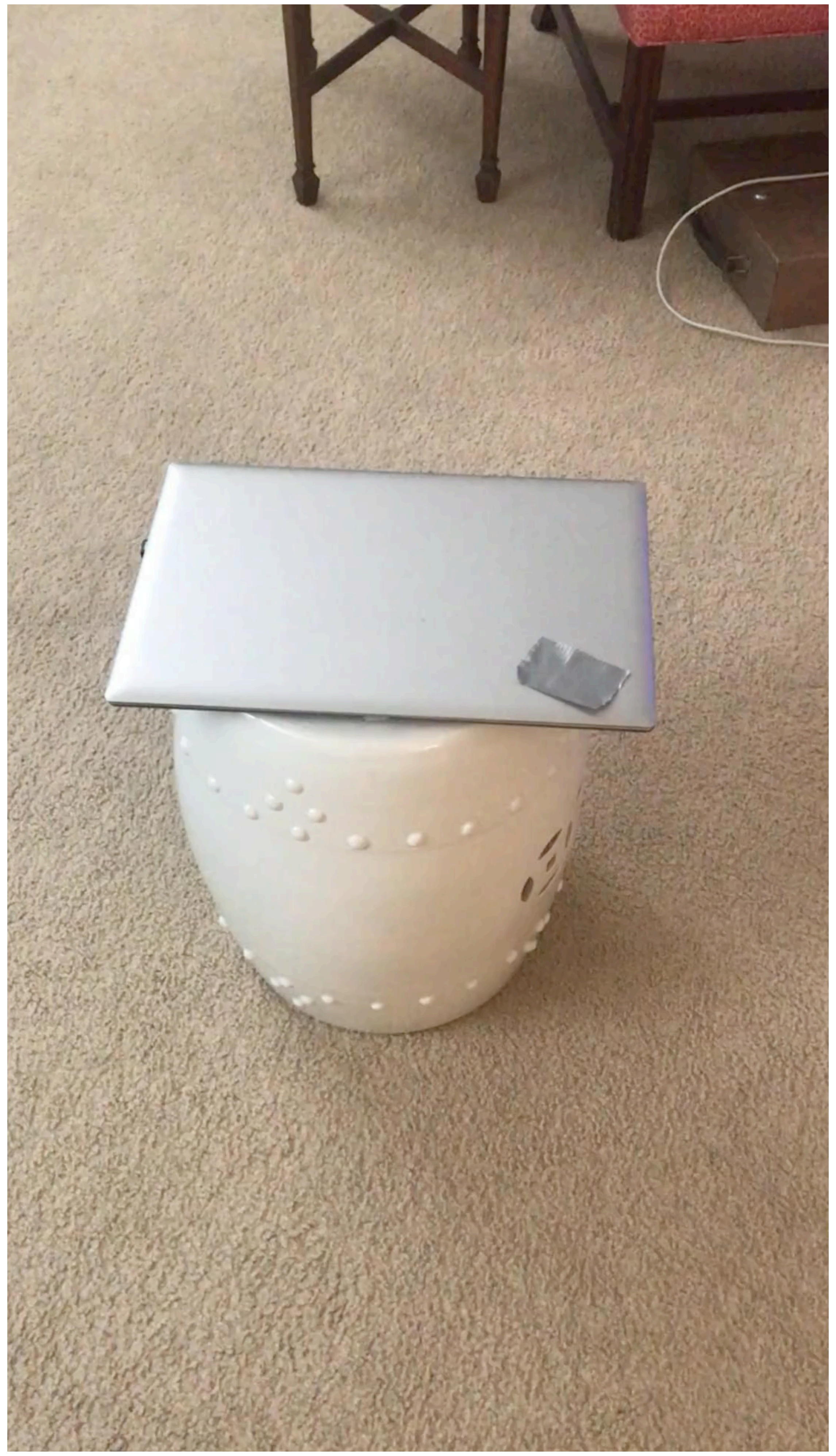
iNVS: Repurposing Diffusion Inpainters for Novel View Synthesis

Yash Kant[1], Aliaksandr Siarohin[2], Michael Vasilkovsky[2], Riza Alp Guler[2], Jian Ren[2], Sergey Tulyakov[2], Igor Gilitschenski[1]

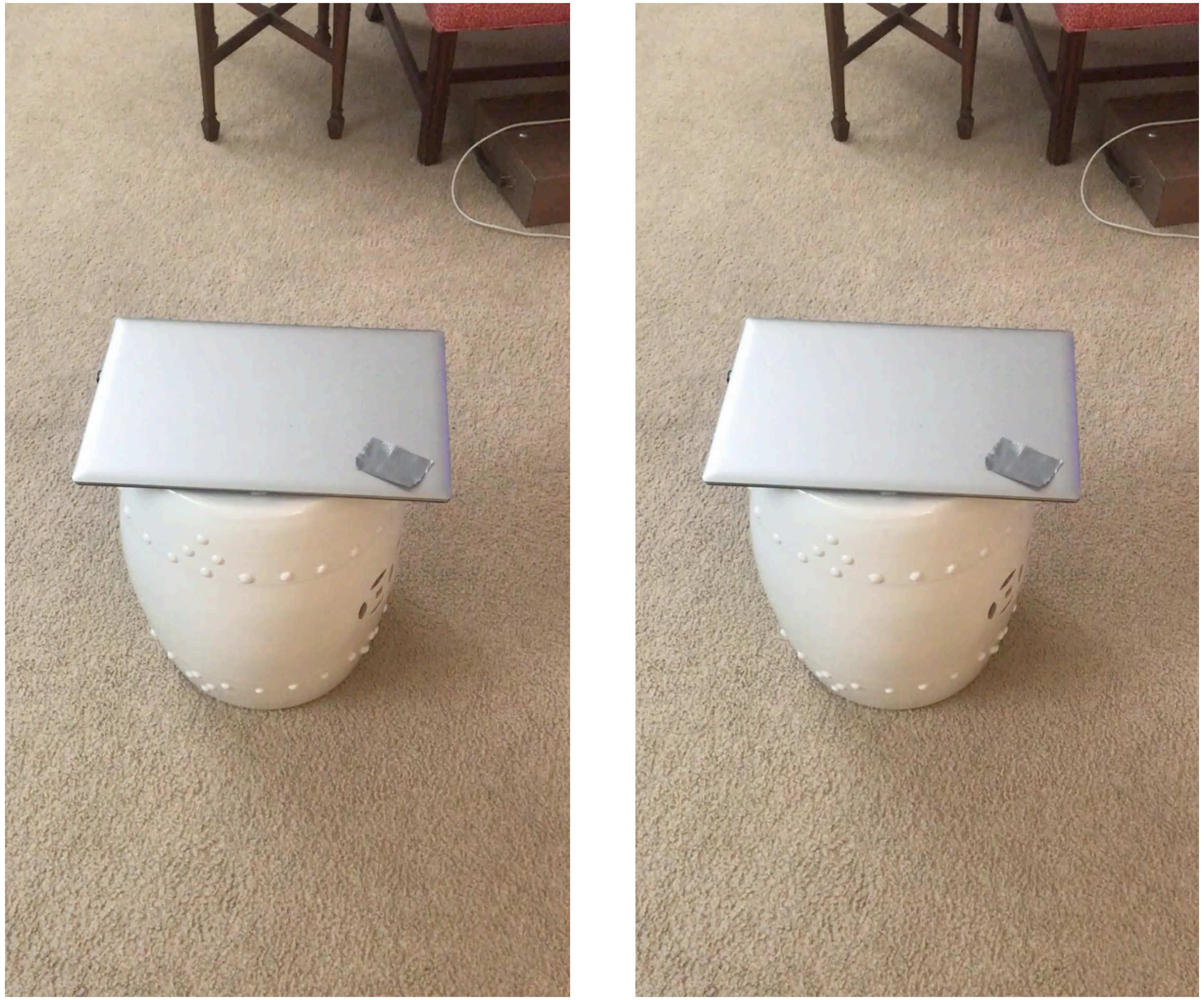
University of Toronto[1]

Snap Research[2]

Task: Given single image of an object, we want to generate it from novel viewpoints.



Given Image



Want Novel Views!

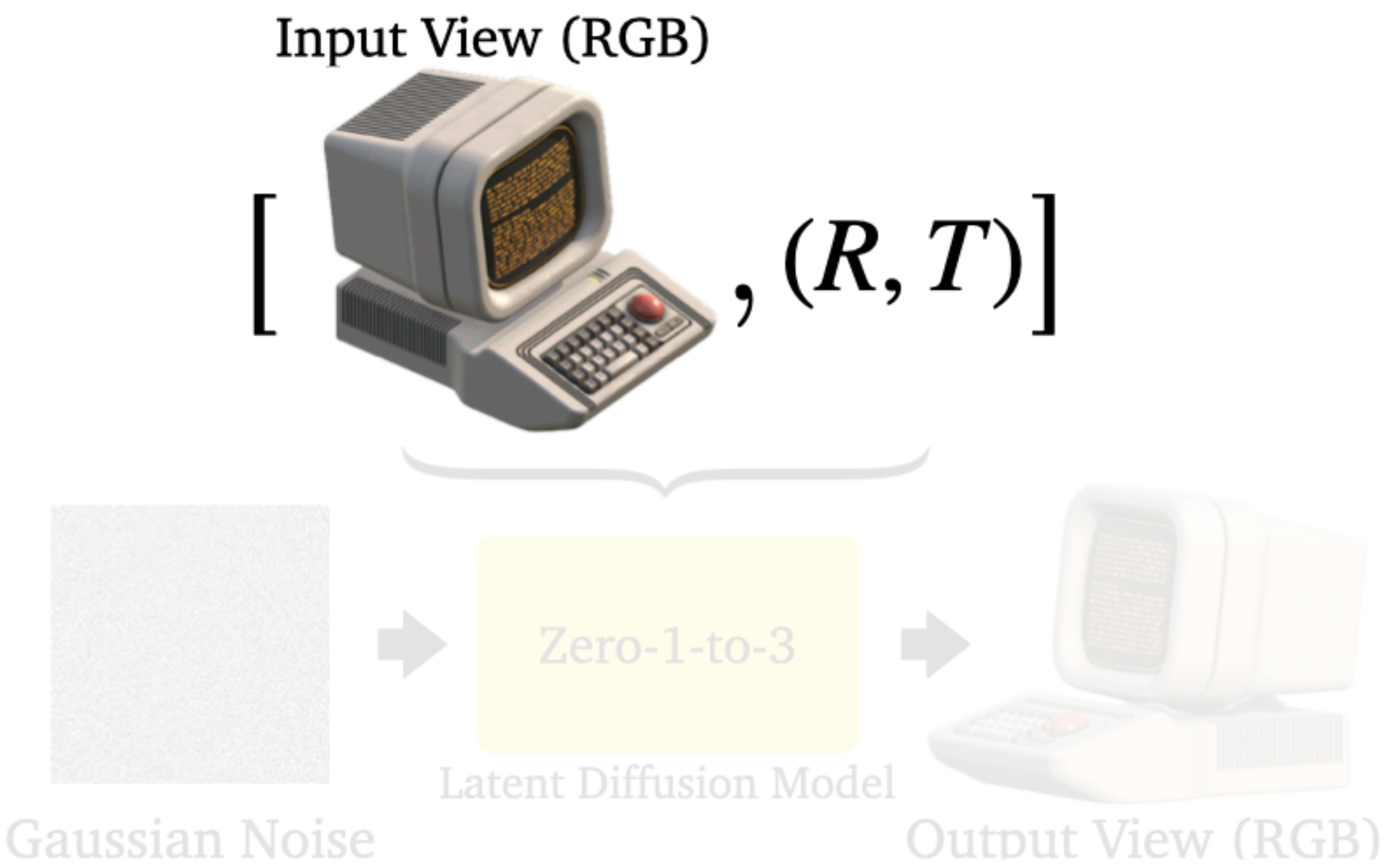
Task: Given single image of an object, we want to generate it from novel viewpoints.

Source View

Target Views

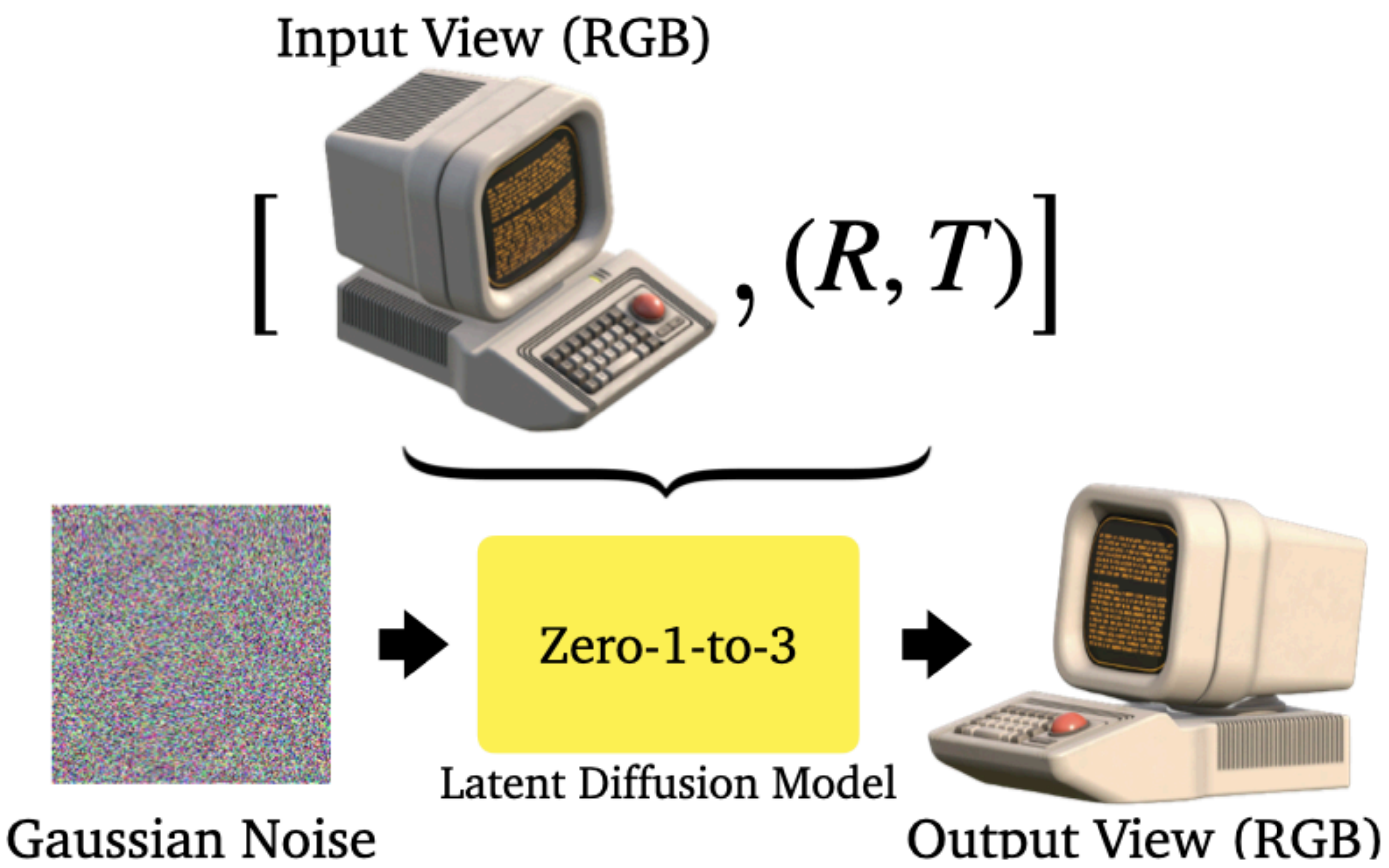
Prior Works: Zero-1-to-3 (ICCV, 2023)

- Encodes the source view using CLIP.
- Relative camera pose is encoded using dense MLP.



Prior Works: Zero-1-to-3 (ICCV, 2023)

- Encodes the source view using CLIP.
- Relative camera pose is encoded using dense MLP.
- Trained a conditional Stable Diffusion to denoise novel views on Objaverse.



Prior Works: Zero-1-to-3 (ICCV, 2023)

- Inefficient reuse of input pixels — sharp details (text and texture) get garbled.

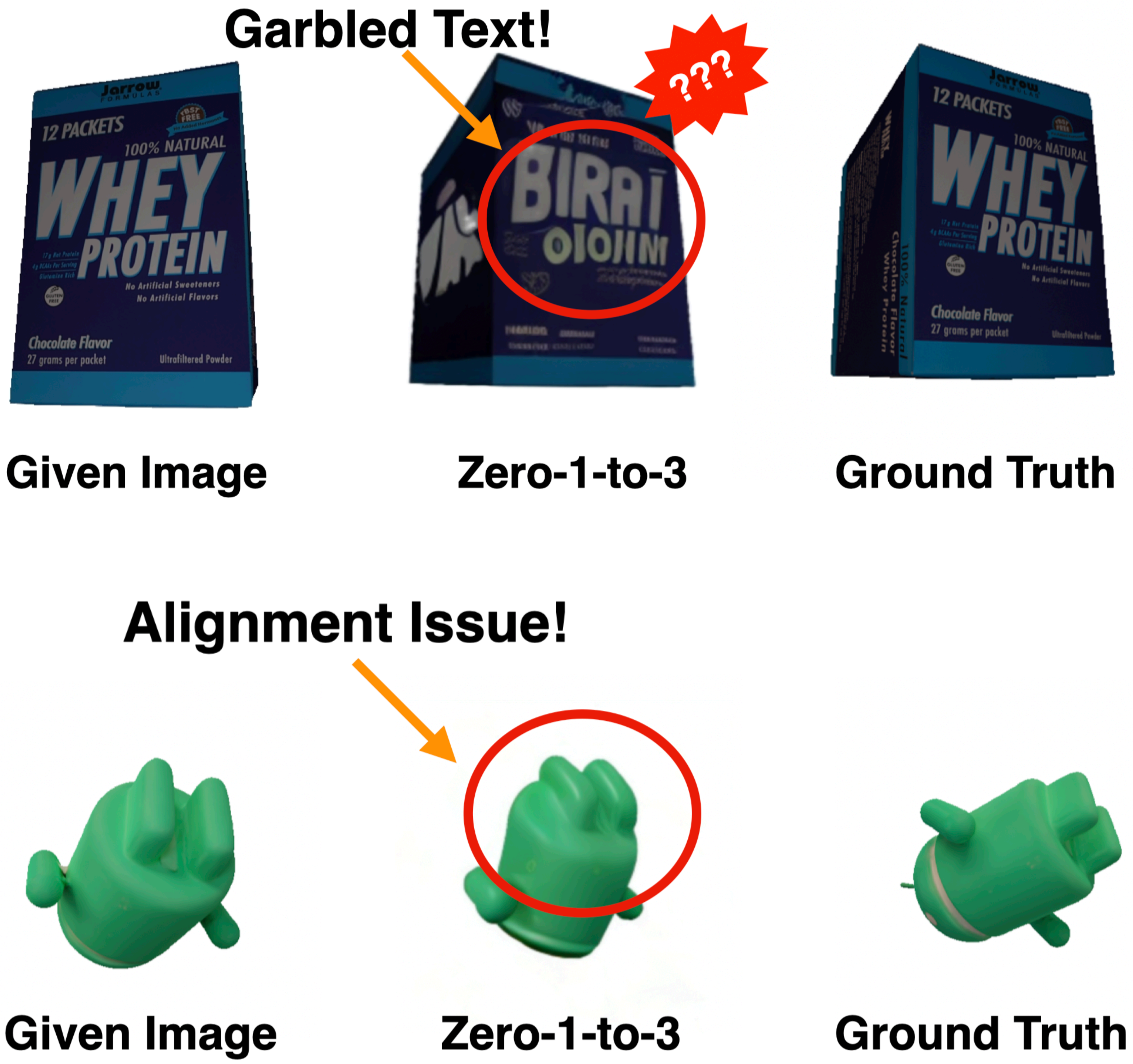
Given Image | Zero-1-to-3 | Ground Truth

Prior Works: Zero-1-to-3 (ICCV, 2023)

- Inefficient reuse of input pixels — sharp details (text and texture) get garbled.
- Camera is encoded in a dense vector — coarse control.

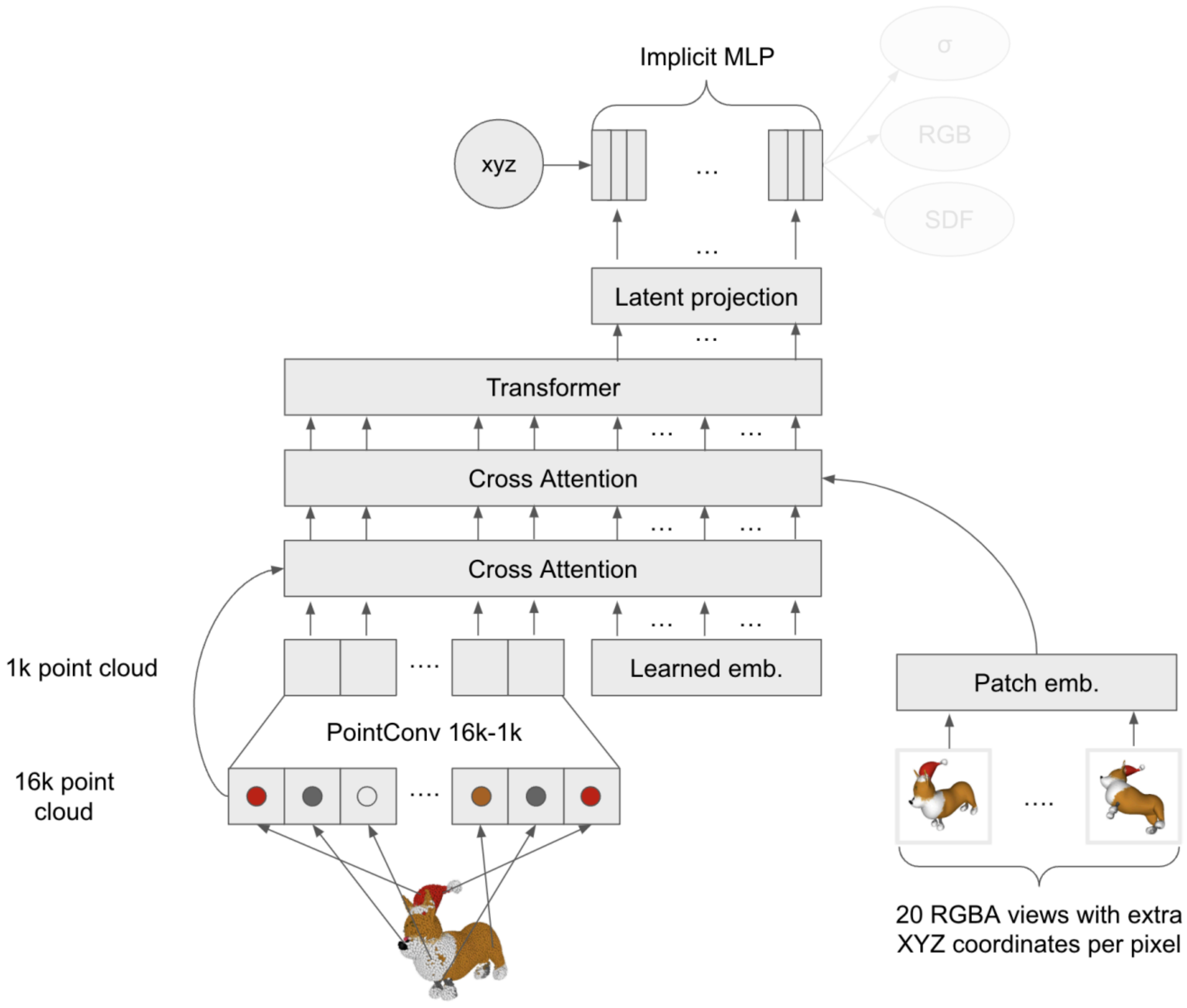
Prior Works: Zero-1-to-3 (ICCV, 2023)

- Inefficient reuse of input pixels — sharp details (text and texture) get garbled.
- Camera is encoded in a dense vector — coarse control.
- 3D consistency is not guaranteed.



Prior Works: Shap-E (OpenAI, 2023)

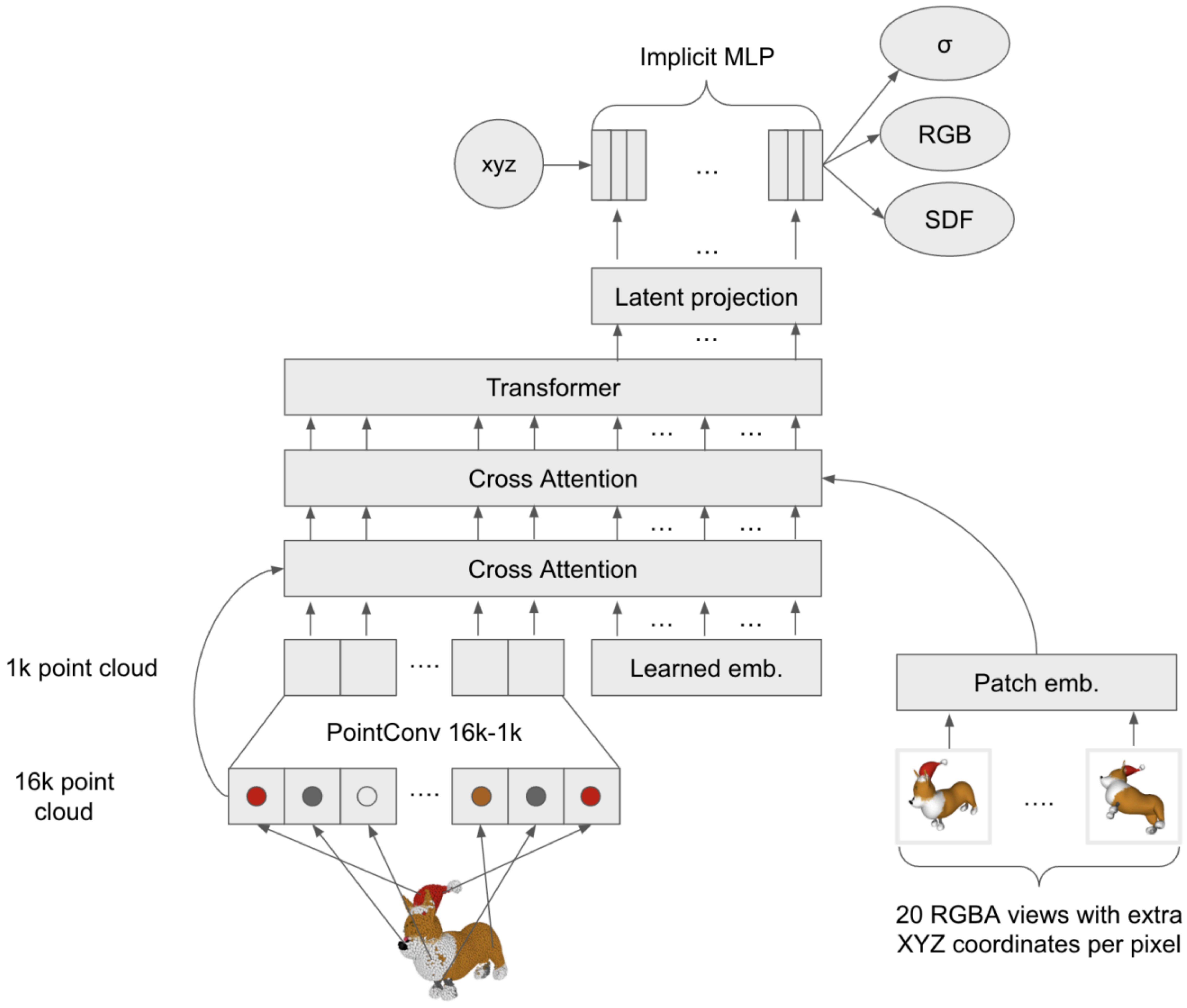
- Hypernetwork-based approach which generates weights of an MLP conditioned on input image.



Prior Works: Shap-E (OpenAI, 2023)

- Hypernetwork-based approach which generates weights of an MLP conditioned on input image.

- This MLP is then queried using points in 3D space to generate occupancy and colour (NeRF).



Prior Works: Shap-E (OpenAI, 2023)

- Instability during training; since the space of possible solutions (weights) is very high dimensional.

Prior Works: Shap-E (OpenAI, 2023)

- Instability during training; since the space of possible solutions (weights) is very high dimensional.

- Inefficient reuse of input pixels — sharp details (text and texture) get garbled.

Prior Works: Shap-E (OpenAI, 2023)

- Instability during training; since the space of possible solutions (weights) is very high dimensional.

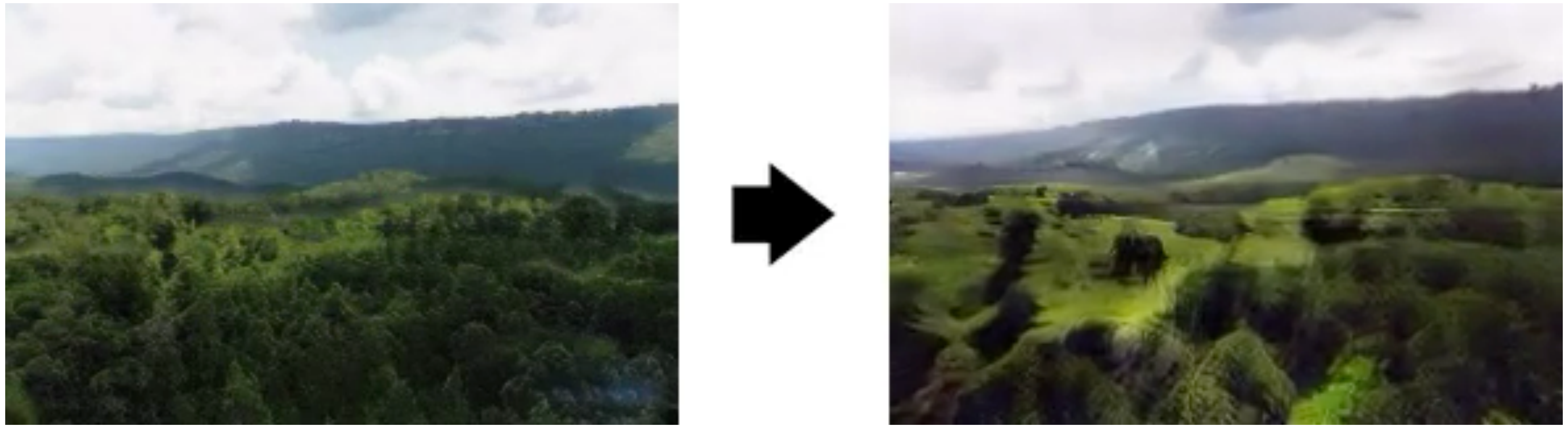
- Inefficient reuse of input pixels — sharp details (text and texture) get garbled.

- Suffers from janus artefacts.

iNVS: Reframing NVS as completion task

Inspired by video generation approaches such as InfiniteNature, we ask —

“can we reframe novel view synthesis as a image completion task rather than generation from scratch?”



InfiniteNature [Andrew Liu, et al. ICCV, 2021]

iNVS: Reframing NVS as completion task

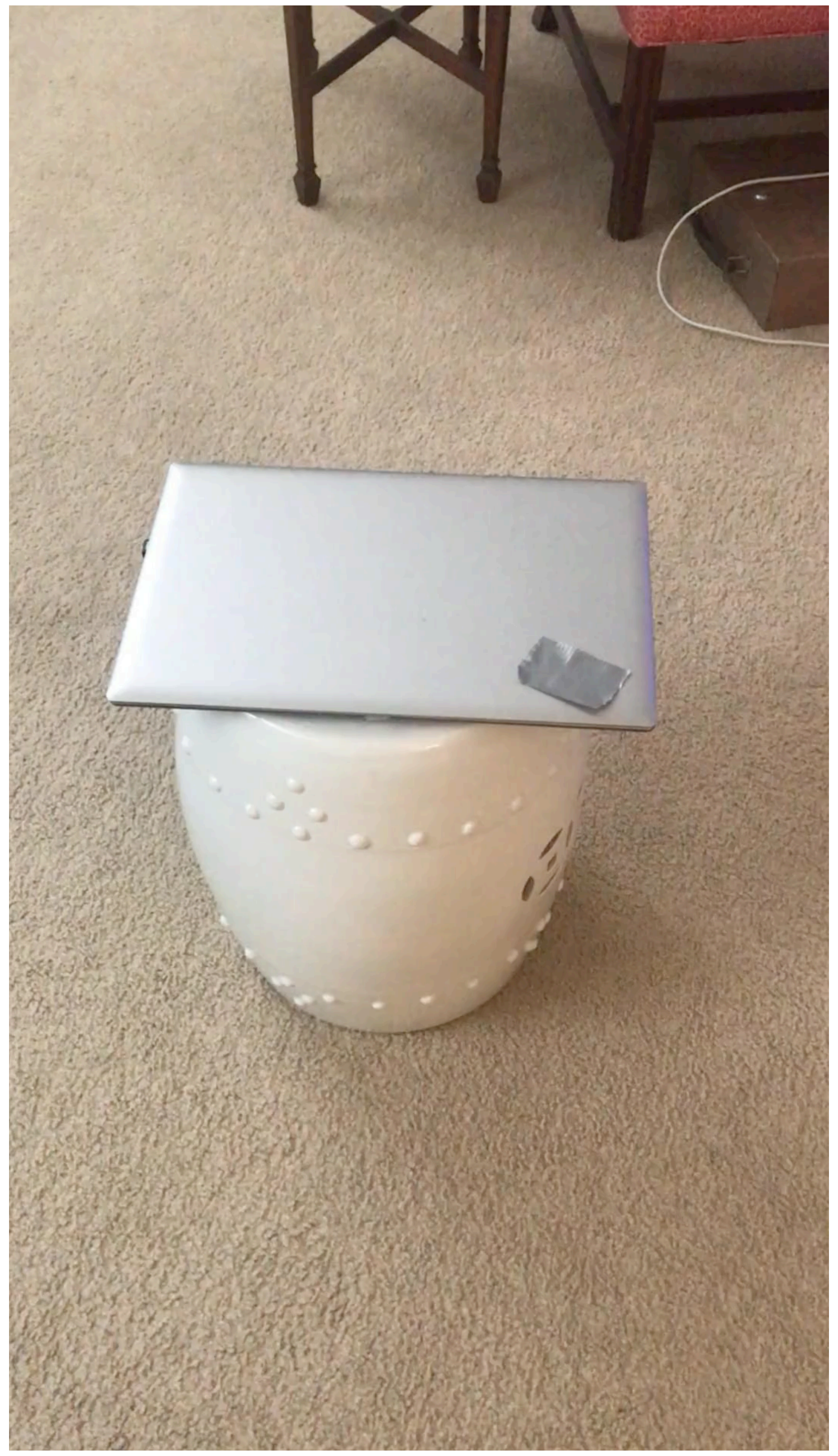
Inspired by video generation approaches such as
Infinite ... ask —

“can we reframe nove ... as a image completion
task rather ... om scratch?”

InfiniteNature [Andrew Liu, et al. ICCV, 2021]

iNVS: Creating Partial Views

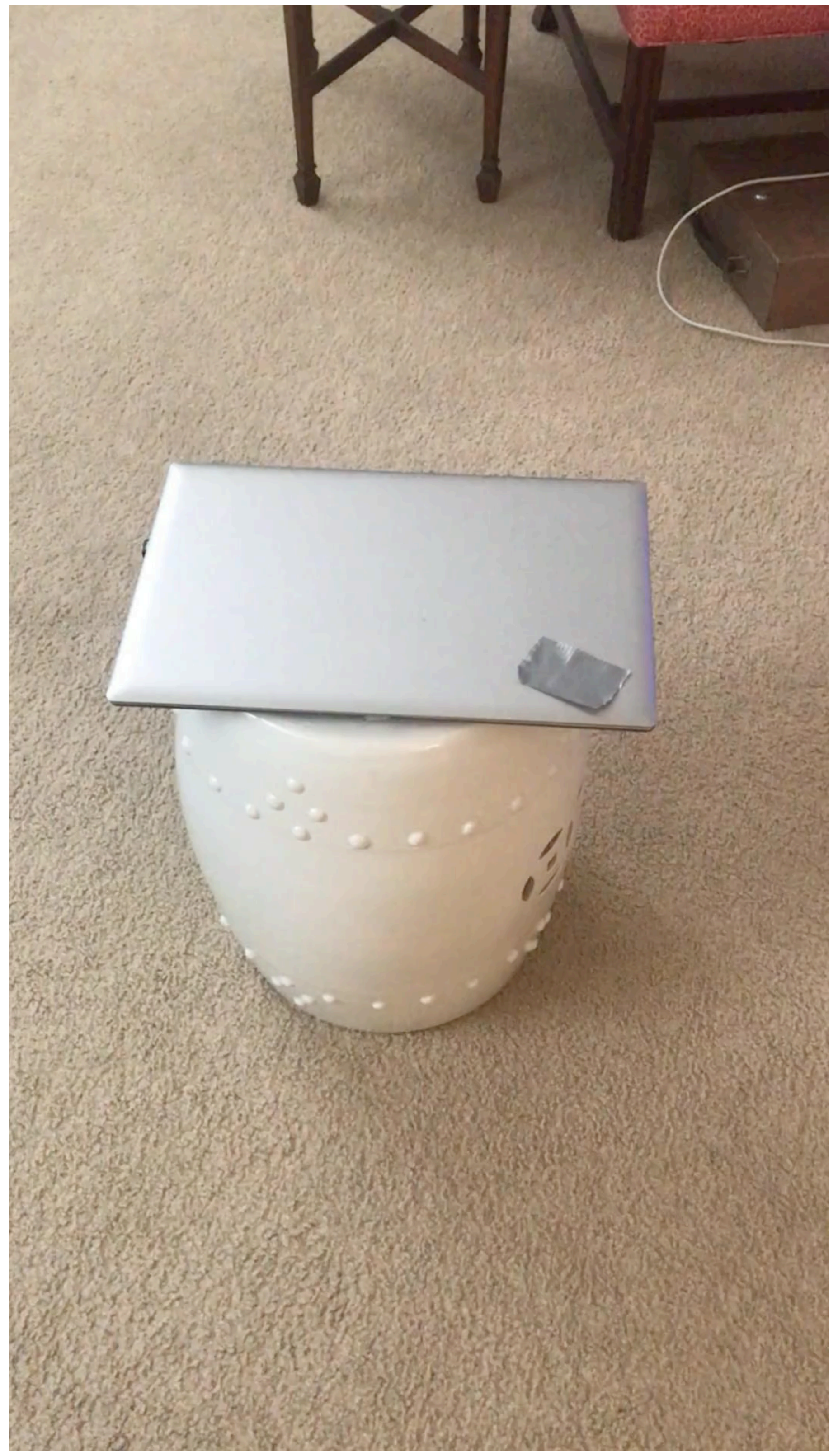
We reuse source image pixels to create a partial image.



Source View

iNVS: Creating Partial Views

We use monocular depth [ZoeDepth] to unproject source pixels in 3D.



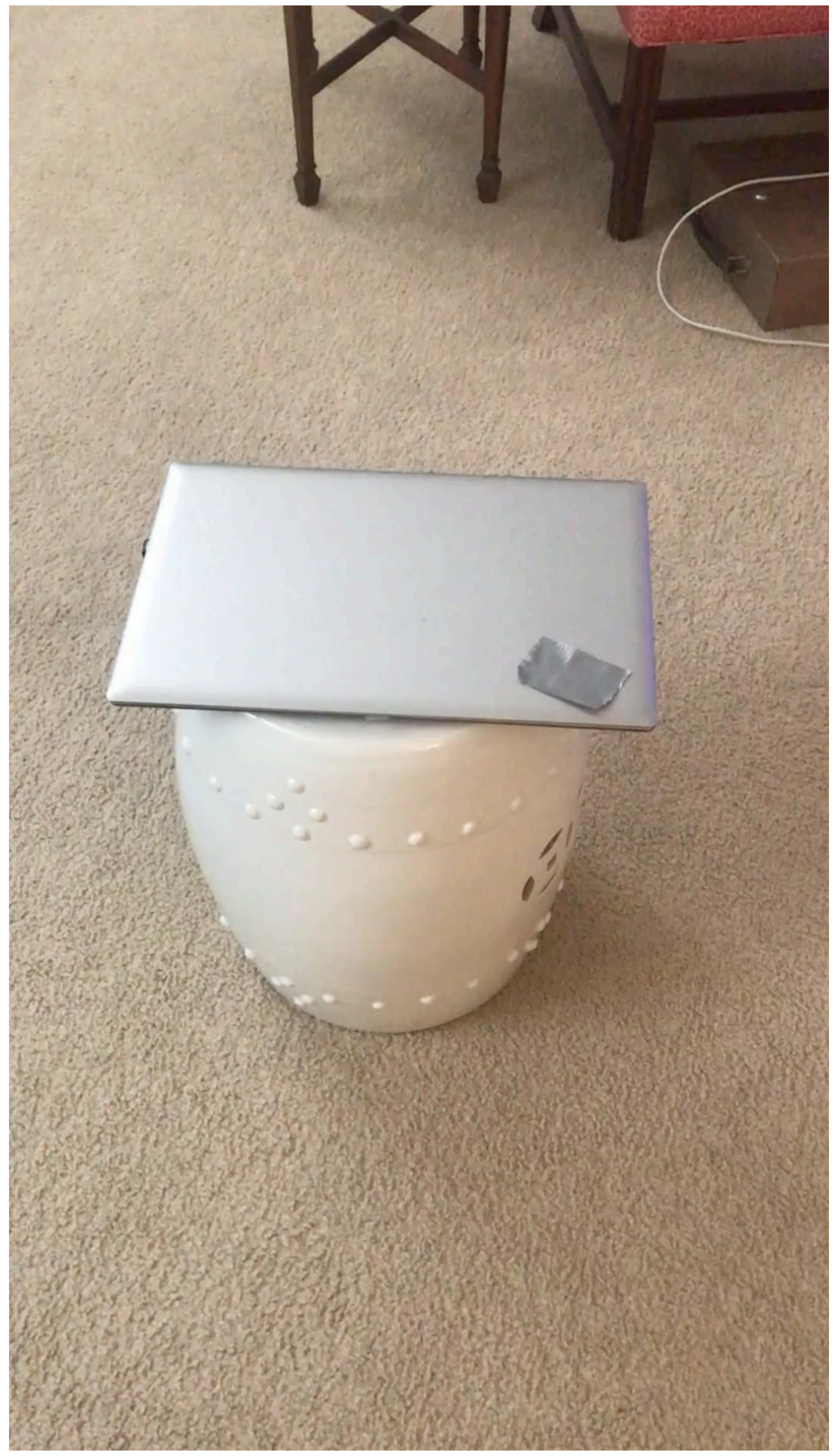
Source View

ZoeDepth

Depth Map

iNVS: Creating Partial Views

We re-project these 3D points back to target view using softmax-splatting and create partial target view.



ZoeDepth

Splatting

Source View **Depth Map** **Partial Target View**

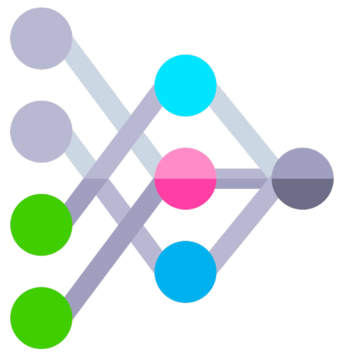
iNVS: Inpainting Partial Views for NVS

iNVS: Inpainting Partial Views for NVS

- Stable Diffusion Inpainter fills in newly discovered regions.

iNVS: Inpainting Partial Views for NVS

- Stable Diffusion Inpainter fills in newly discovered regions.
- We train the Inpainter on Objaverse dataset, to learn 3D completion priors.



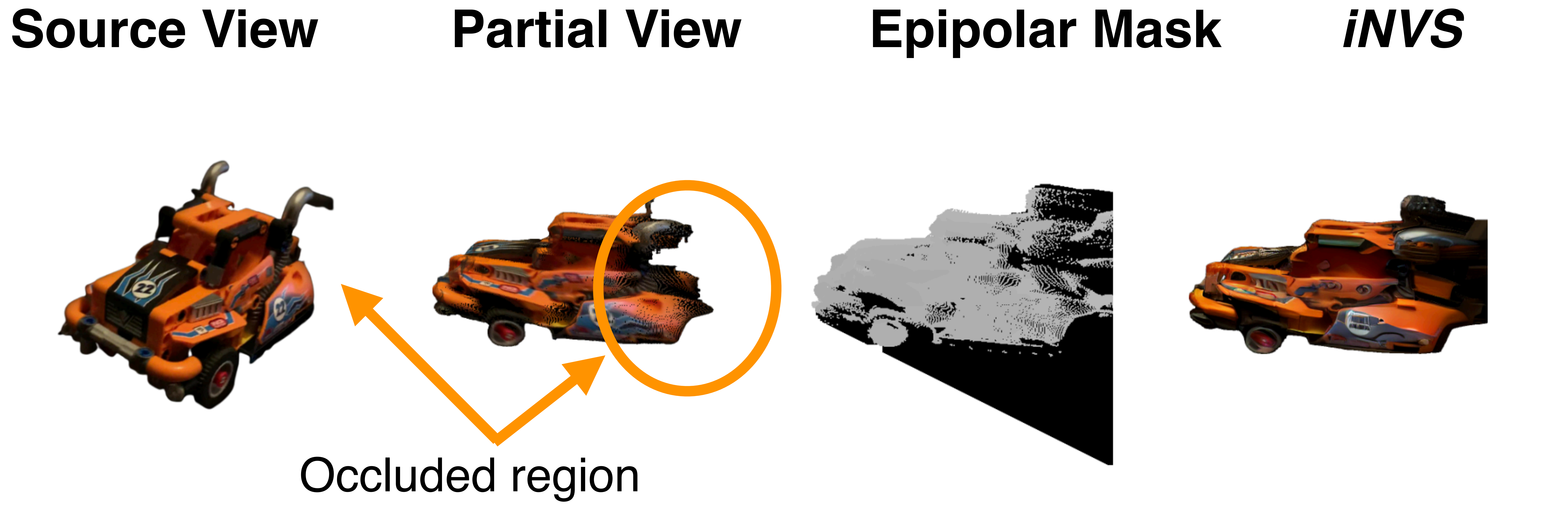
3D-aware Inpainter

Partial Target View

Inpainted View

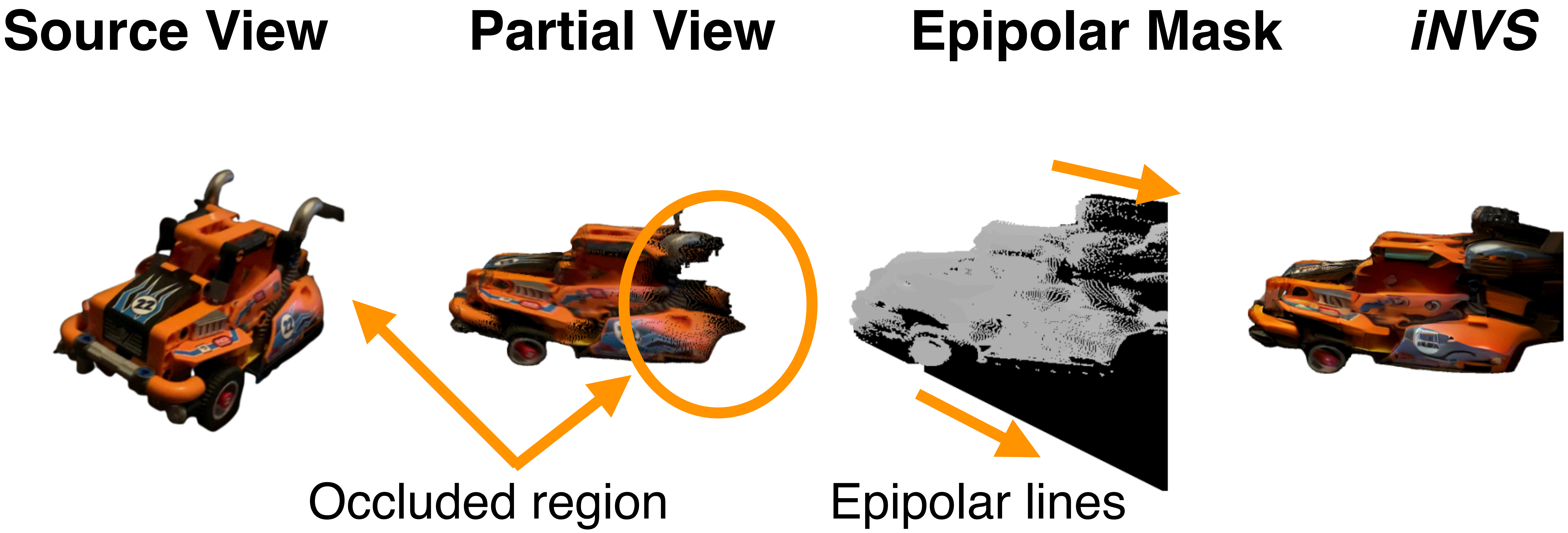
iNVS: Epipolar and Pose-aware Inpainting Mask

We further constrain the inpainting to the areas occluded in source view using epipolar lines.



iNVS: Epipolar and Pose-aware Inpainting Mask

We further constrain the inpainting to the areas occluded in source view using epipolar lines.



iNVS: Epipolar and Pose-aware Inpainting Mask

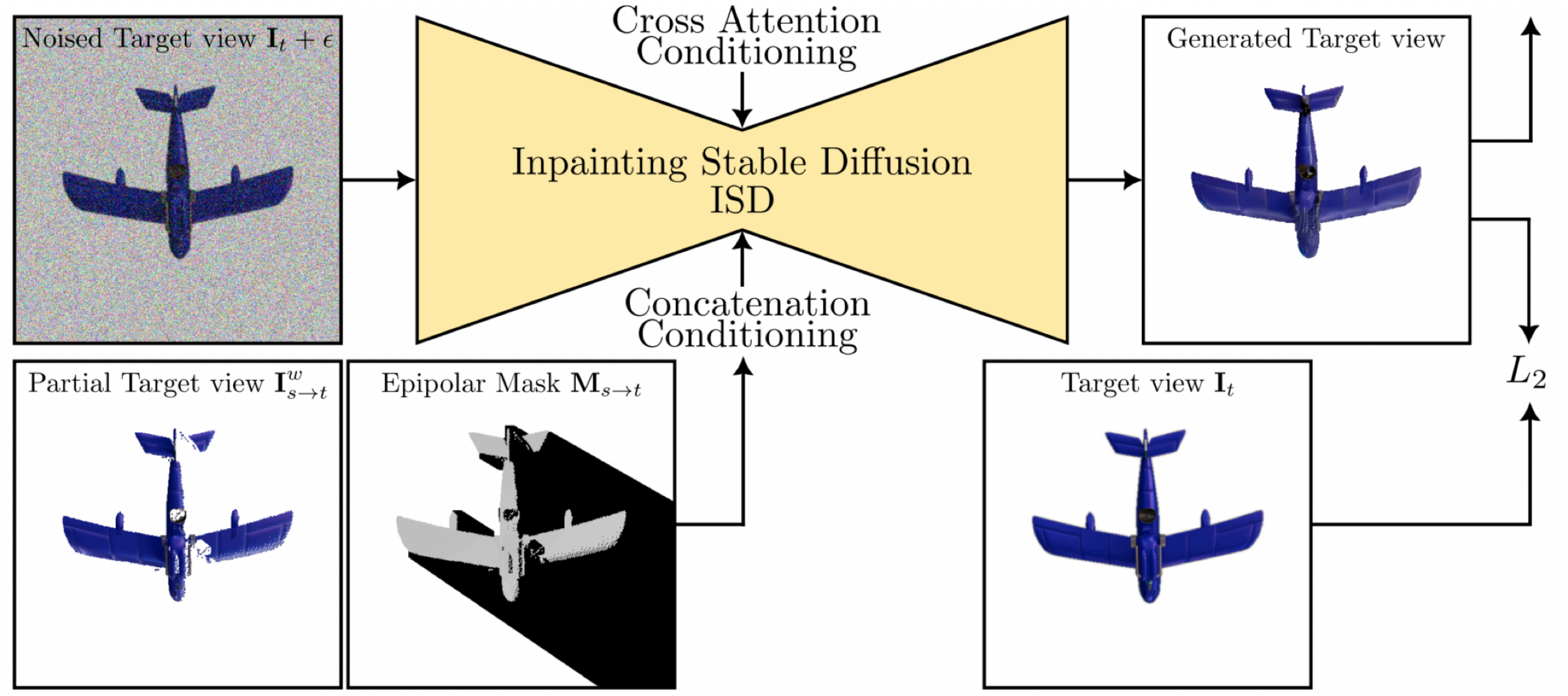
Additionally, we also use a soft-valued (0,1) inpainting mask that conveys the relative angle (0,180) between source and target camera ray.

Source View **Partial View** **Epipolar Mask** ***iNVS***

Soft Inpainting Mask

iNVS: Training Details

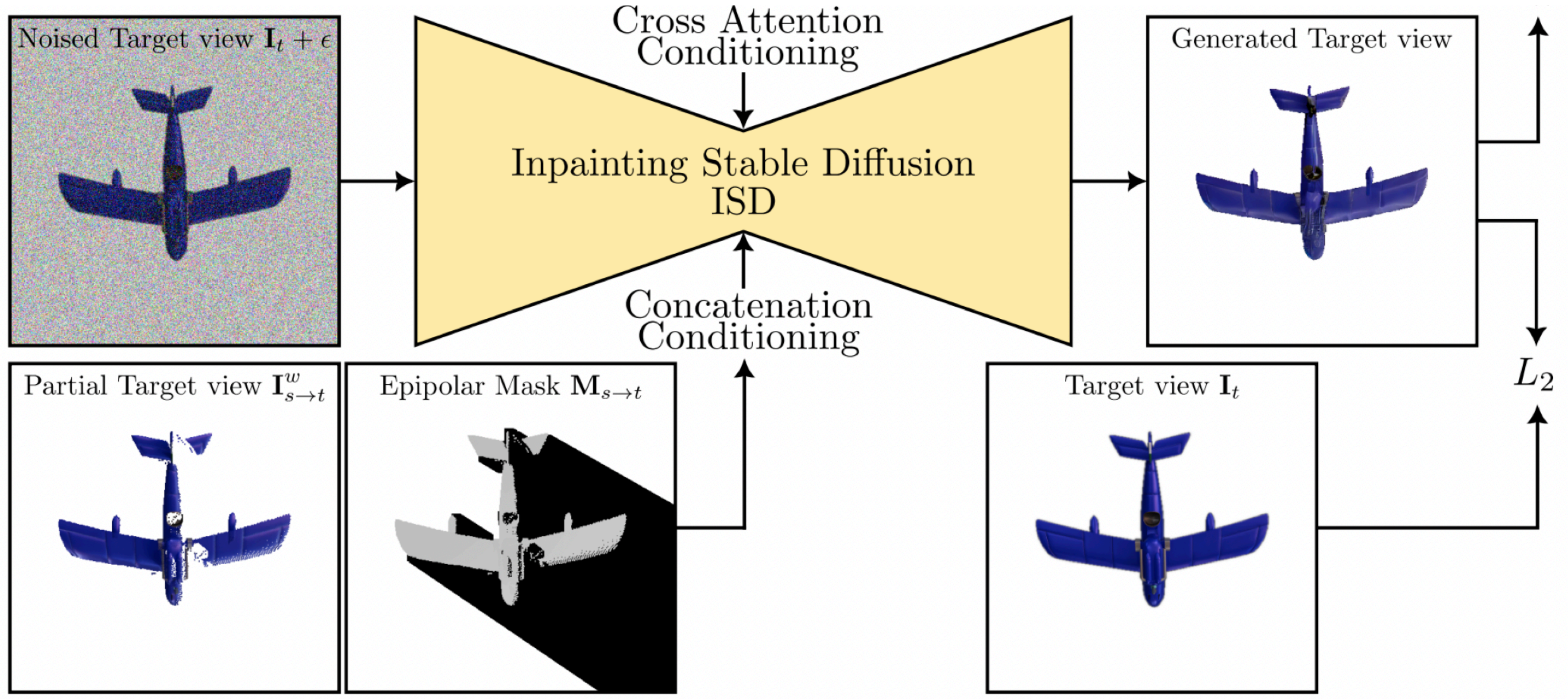
- We fine-tune SD Inpainter network on 96 A100 GPUs for two weeks, on ~ 20M rendered images from Objaverse.



iNVS: Training Details

- We fine-tune SD Inpainter network on 96 A100 GPUs for two weeks, on ~ 20M rendered images from Objaverse.

- Object boundary appears as early as 10% of denoising steps; hence, we sample timesteps with a bias during training.



Early Inference Steps

10% 50%

iNVS: Qualitative Comparison

Given Image

iNVS: Qualitative Comparison

Given Image

Baseline
(Zero-1-to-3)

iNVS: Qualitative Comparison

Given Image

Garbled Text!

Baseline
(Zero-1-to-3)

iNVS: Qualitative Comparison

Garbled Text!

Given Image

Baseline
(Zero-1-to-3)

iNVS
(Ours)

iNVS: Qualitative Comparison

Given Image

Baseline
(Zero-1-to-3)

iNVS
(Ours)

iNVS: Qualitative Comparison

Given Image

iNVS: Qualitative Comparison

Given Image

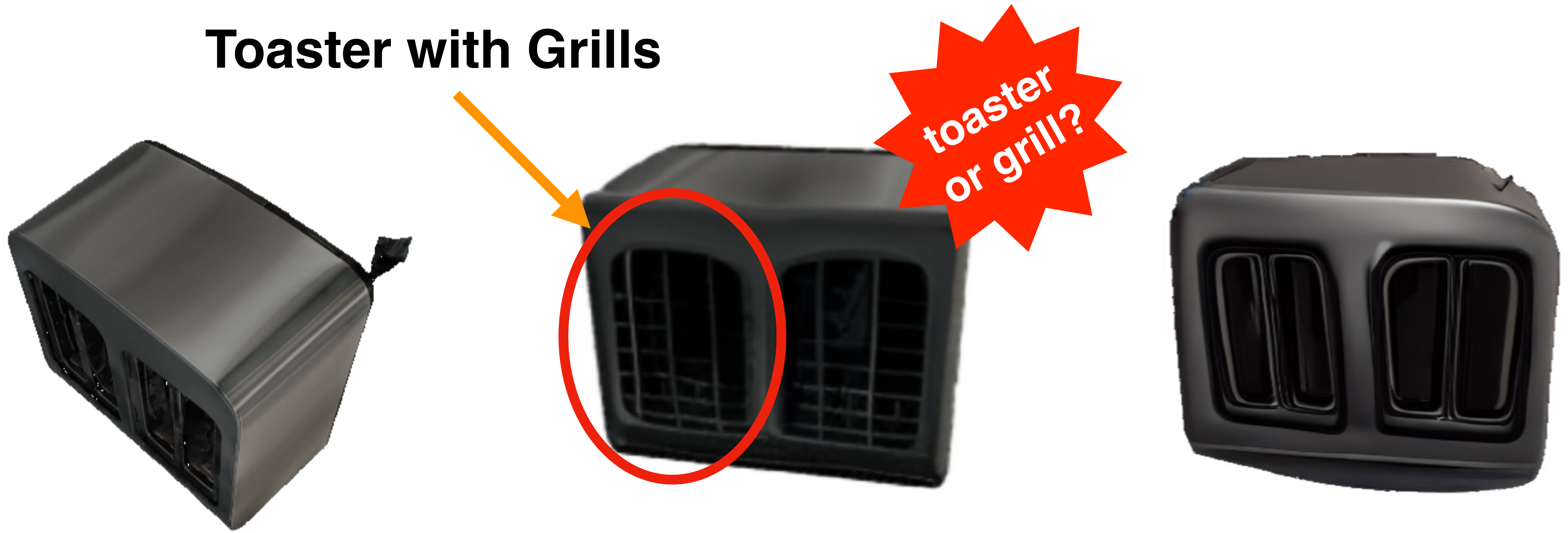
Baseline (Zero-1-to-3)

iNVS: Qualitative Comparison

Given Image

Baseline
(Zero-1-to-3)

iNVS: Qualitative Comparison



Given Image

Baseline (Zero-1-to-3)

iNVS **(Ours)**

iNVS: Qualitative Comparison

Given Image

Baseline (Zero-1-to-3)

***iNVS* (Ours)**

iNVS: Qualitative Comparison

Given Image

Resolution:
256x256

Baseline
(Zero-1-to-3)

Resolution:
512x512

iNVS
(Ours)

iNVS: Quantitative Comparison

iNVS outperforms Zero-1-to-3 on 2/3 metrics on GSO (synthetic)

Method	PSNR ↑	SSIM ↑	LPIPS ↓
iNVS	**18.95**	0.30	**0.24**
Zero-1-to-3	14.74	**0.34**	0.25

Google Scanned Objects

iNVS: Quantitative Comparison

iNVS outperforms Zero-1-to-3 on 2/3 metrics on GSO (synthetic) and CO3D (real-world) datasets.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
iNVS	**18.95**	0.30	**0.24**
Zero-1-to-3	14.74	**0.34**	0.25

Google Scanned Objects

Method	PSNR ↑	SSIM ↑	LPIPS ↓
iNVS	**17.58**	**0.33**	**0.36**
Zero-1-to-3	12.32	**0.33**	0.42

Common Objects in 3D

Failure Mode

Investigating the lower Structural Similarity (SSIM) score, we find some common failure modes.

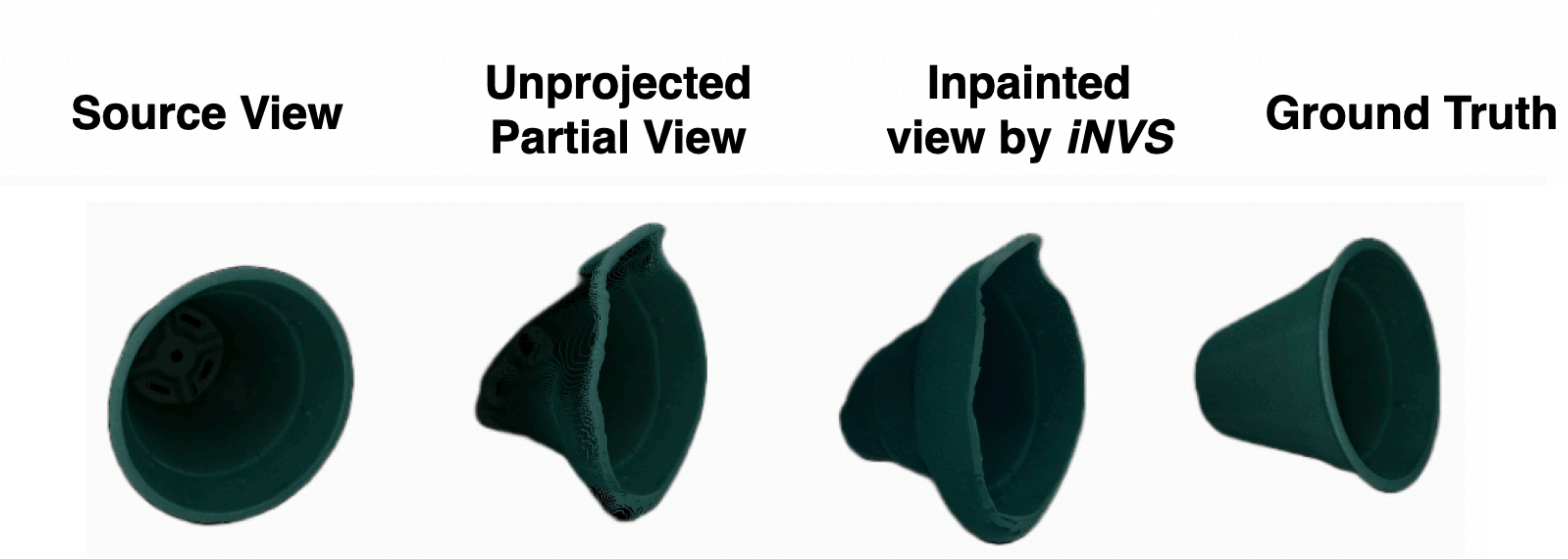
Failure Mode

Investigating the lower Structural Similarity (SSIM) score, we find some common failure modes.

iNVS struggles most when monocular depth estimator generates inaccurate depth.

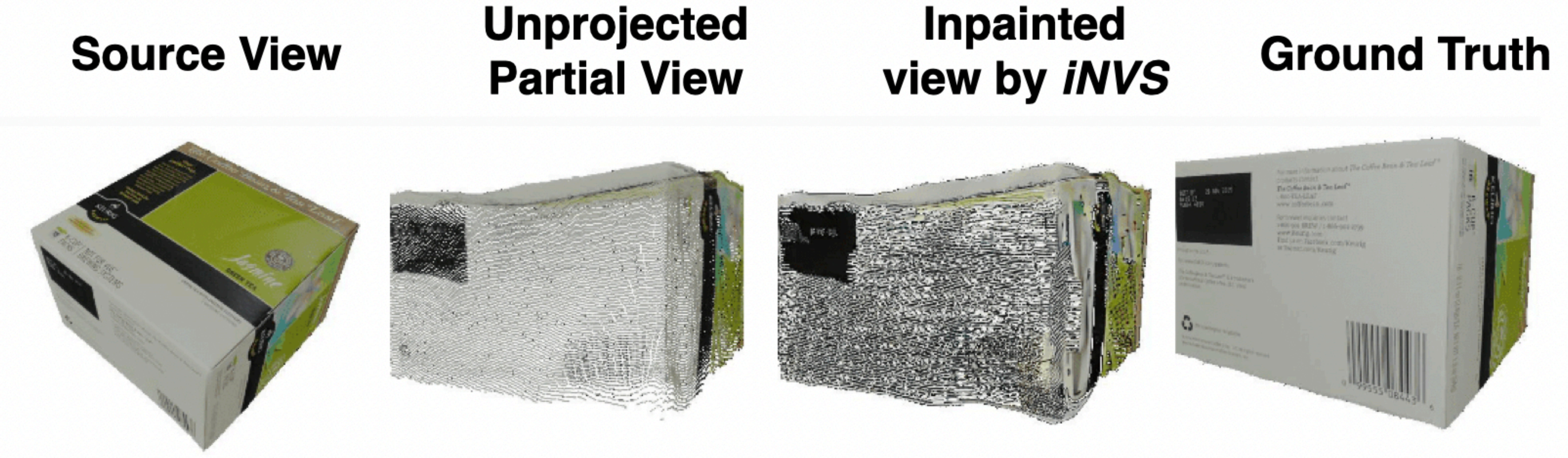
Failure Mode #1: Deformed Partial View

Imprecise depth leads to deformed partial view — difficult to recover from during inpainting.



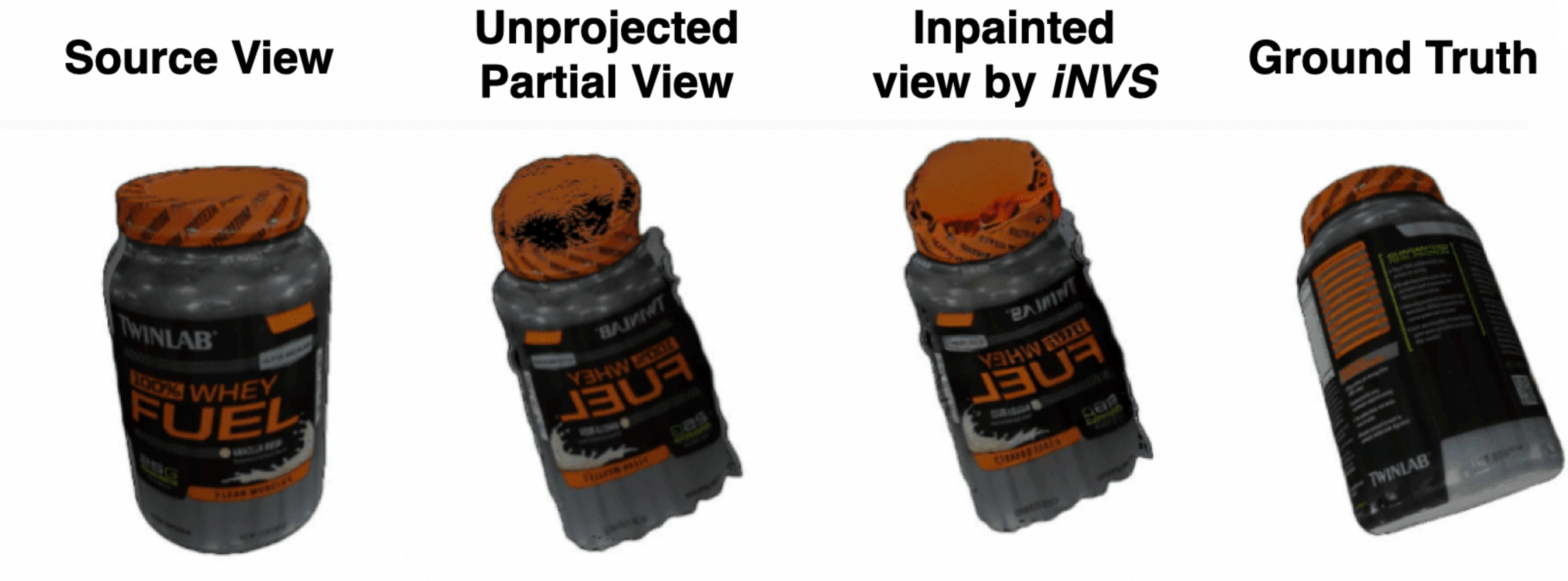
Failure Mode #2: Tiny holes can blend into texture.

Imprecise depth and downsampled inpainting mask occasionally cause reprojection holes to blend in.



Failure Mode #3: Flipped pixels throw-off Inpainter.

Under large-viewpoint changes; we rely on inpainting mask to detect large ray angle changes, but it may fail.



Thanks for listening.

Poster Today @ 6PM!

Webpage: **https://yashkant.github.io/invs**

Yash Kant[1], Aliaksandr Siarohin[2], Michael Vasilkovsky[2], Riza Alp Guler[2], Jian Ren[2], Sergey Tulyakov[2], Igor Gilitschenski[1]

University of Toronto[1]

Snap Research[2]