Zero-1-to-3: Zero-shot One Image to 3D Object

Ruoshi Liu
Columbia University

Rundi Wu
Columbia University

Basile Van Hoorick
Columbia University

Pavel Tokmakov
Toyota Research Institute

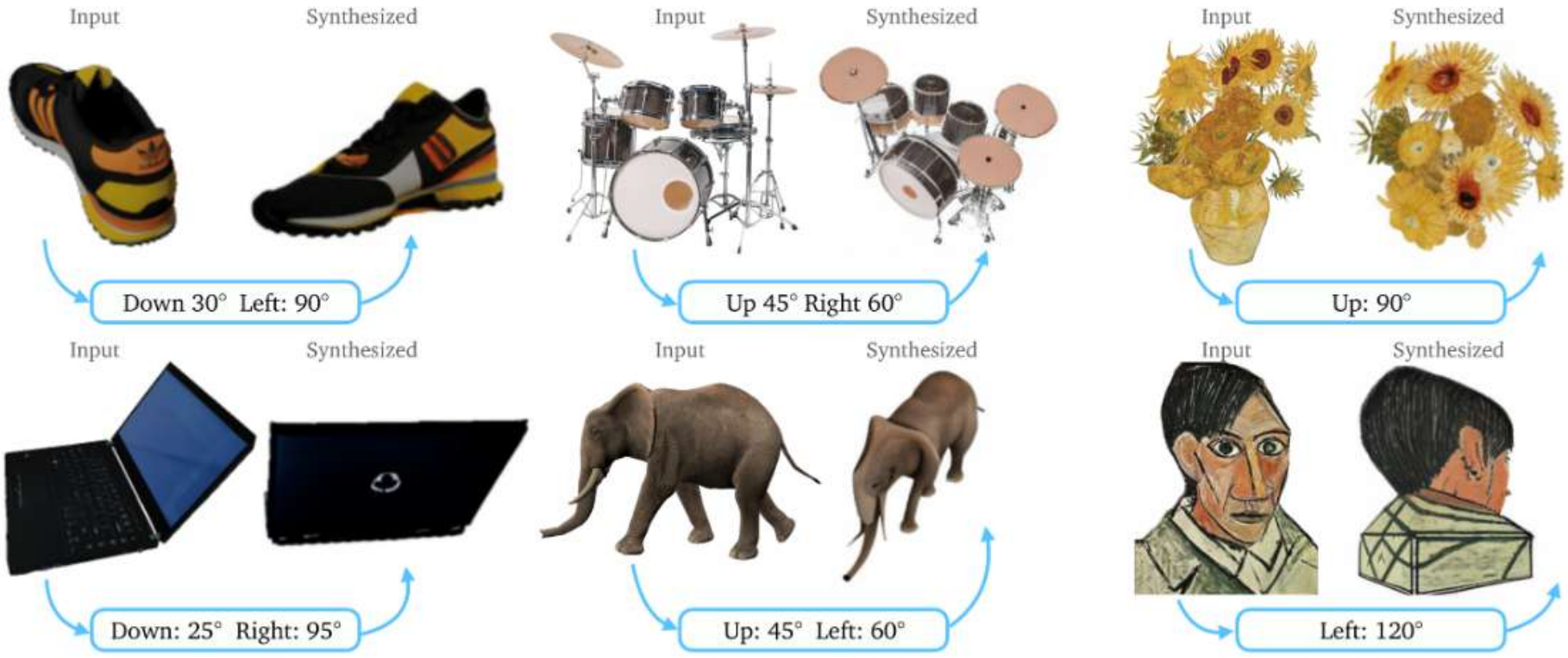
Sergey Zakharov
Toyota Research Institute

Carl Vondrick
Columbia University

TL;DR: We learn to control the camera perspective in large-scale diffusion models, enabling zero-shot novel view synthesis and 3D reconstruction from a single image.

ICCV 2023

Task



$$\hat{x}_{R,T} = f(x, R, T)$$

Challenges

Firstly, although large-scale generative models are trained on a large variety of objects in different viewpoints, the representations do not explicitly encode the correspondences between viewpoints.

Secondly, generative models inherit viewpoint biases reflected on the Internet.

Control Camera Viewpoint

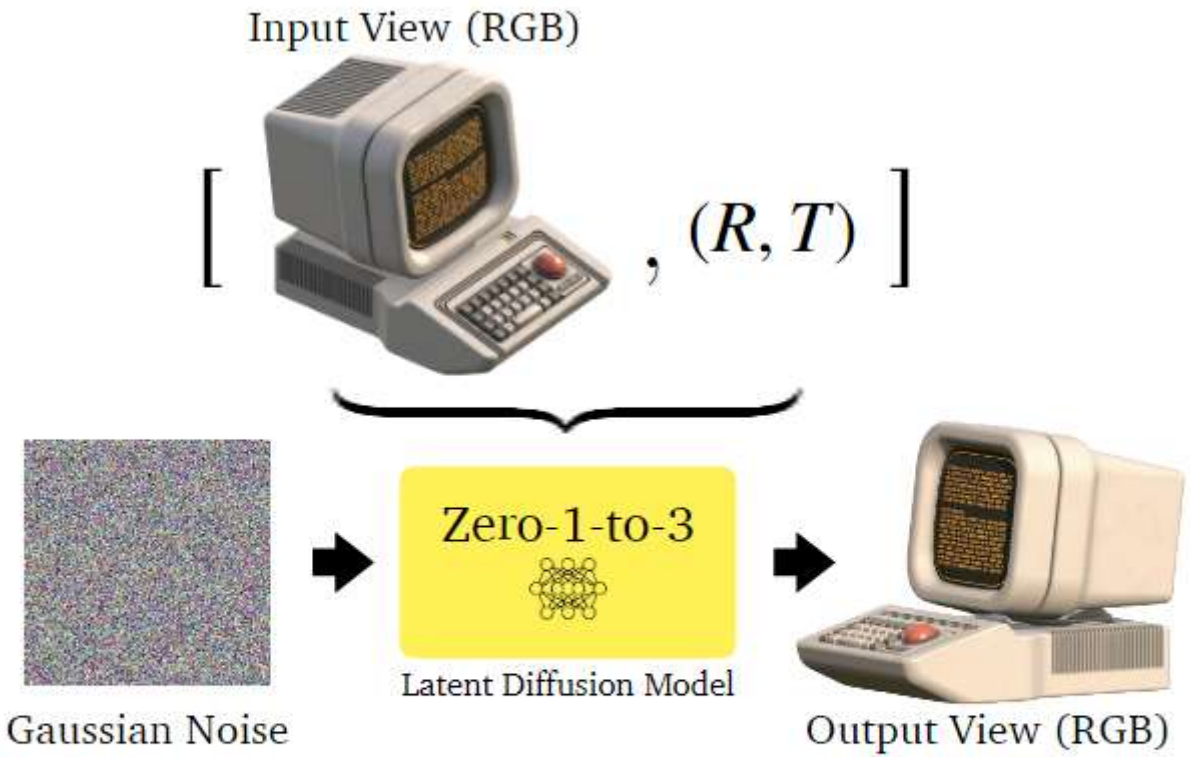


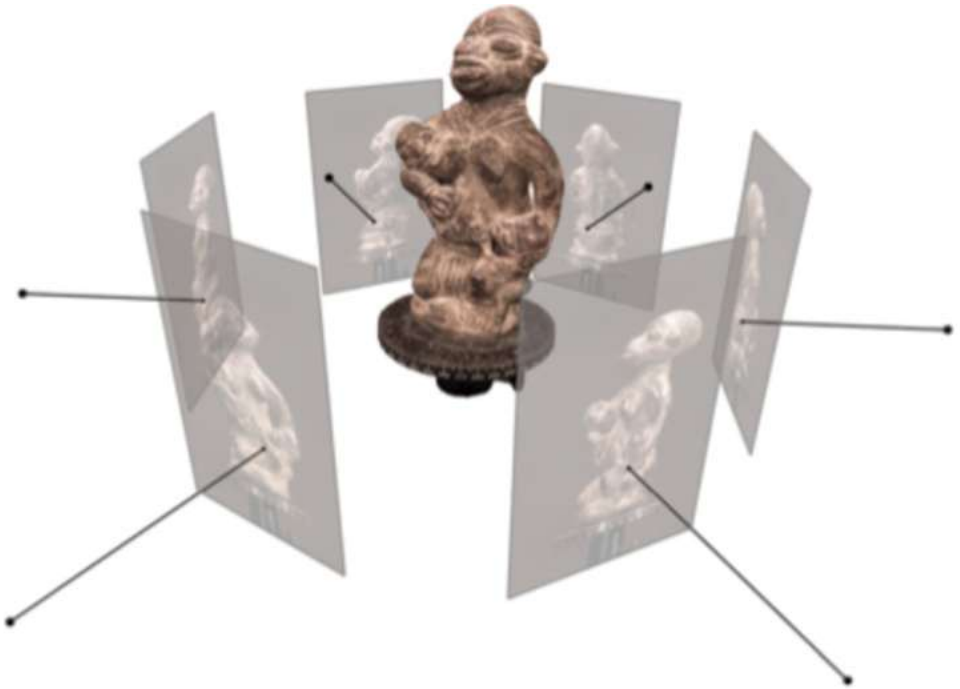
Figure 3: **Zero-1-to-3** is a viewpoint-conditioned image translation model using a conditional latent diffusion architecture. Both the input view and a relative viewpoint transformation are used as conditional information.

$$\hat{x}_{R,T} = f(x, R, T)$$

$$\min_{\theta} \mathbb{E}_{z\sim\mathcal{E}(x),t,\epsilon\sim\mathcal{N}(0,1)}||\epsilon - \epsilon_\theta(z_t, t, c(x, R, T))||_2^2.$$

We concatenate the image CLIP embedding (dimension 768) and the pose vector (dimension 4) and initialize another fully-connected layer ($772 \rightarrow 768$) to ensure compatibility with the diffusion model architecture. The learning rate of this layer is scaled up to be 10× larger than the other layers. The rest of the network architecture is kept the same as the original Stable Diffusion.

Dataset



For each object in the dataset, we randomly sample 12 camera extrinsics matrices M pointing at the center of the object and render 12 views with a raytracing engine. At training time, two views can be sampled for each object to form an image pair $(\mathbf{x}, \mathbf{x}_{R,T})$. The corresponding relative viewpoint transformation (R, T) that defines the mapping between both perspectives can easily be derived from the two extrinsic matrices.

Results

Figure 5: **Novel view synthesis on Google Scanned Objects [10].** The input view shown on the left is used to synthesize two randomly sampled novel views. Corresponding ground truth views are shown on the right. Compared to the baselines, our synthesized novel view contain rich textual and geometric details that are highly consistent with the ground truth, while baseline methods display a significant loss of high-frequency details.

Results

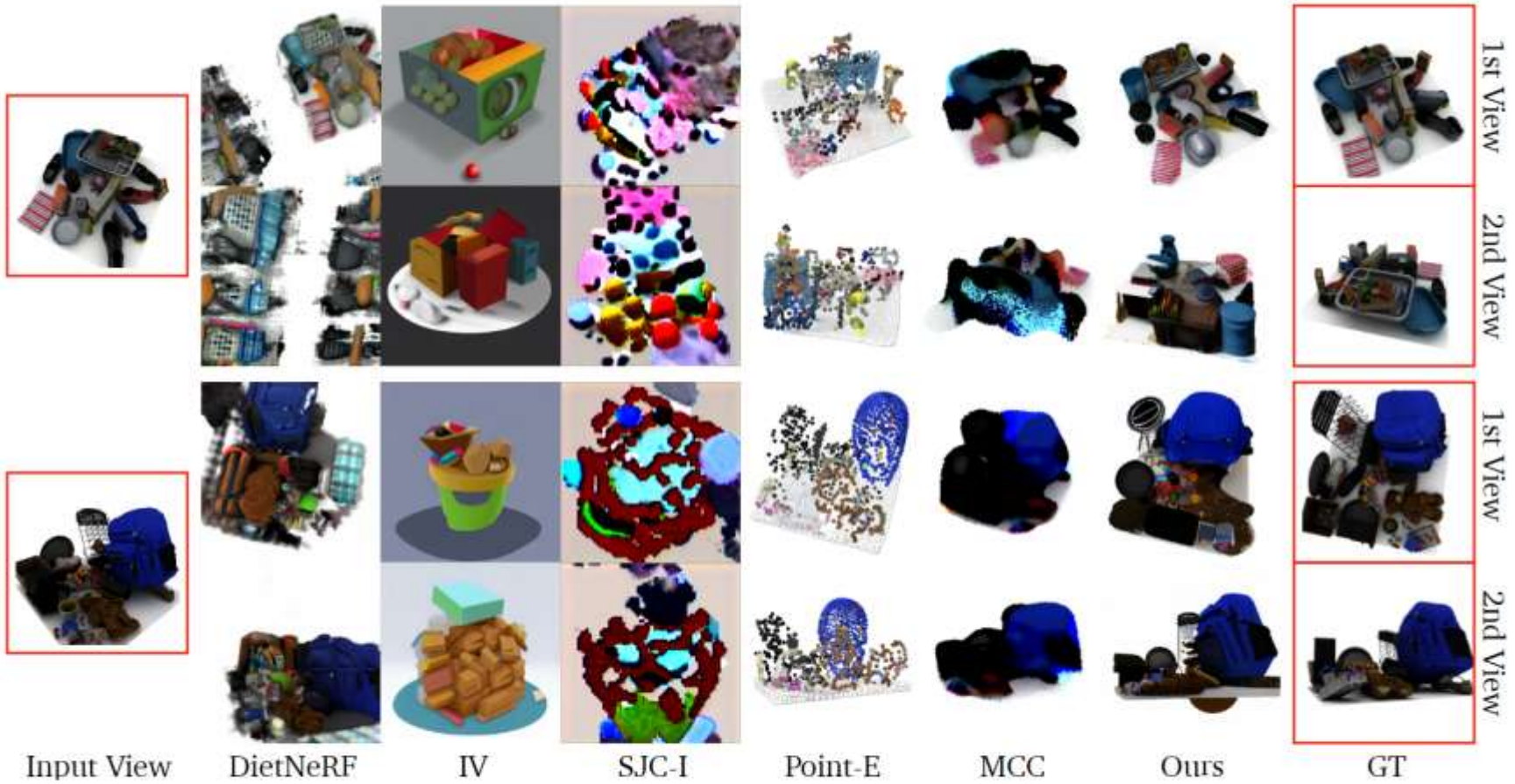


Figure 6: **Novel view synthesis on RTMV [50].** The input view shown on the left is used to synthesize 2 randomly sampled novel views. Corresponding ground truth views are shown on the right. Our synthesized view maintains a high fidelity even under big camera viewpoint changes, while most other methods deteriorate in quality drastically.

Results

Figure 7: **Novel view synthesis on in-the-wild images.** The 1st, 3rd, and 4th rows show results on images taken by an iPhone, and the 2nd row shows results on an image downloaded from the Internet. Our method works are robust to objects with different surface materials and geometry. We randomly sampled 5 different viewpoints and directly showcase the results without cherry-picking. We include more uncurated results in the supplementary materials.

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	Ours
PSNR ↑	<u>8.933</u>	5.914	6.573	**18.378**
SSIM ↑	<u>0.645</u>	0.540	0.552	**0.877**
LPIPS ↓	<u>0.412</u>	0.545	0.484	**0.088**
FID ↓	<u>12.919</u>	22.533	19.783	**0.027**

Table 1: **Results for novel view synthesis on Google Scanned Objects.** All metrics demonstrate that our method is able to outperform the baselines by a significant margin.

	DietNeRF [23]	Image Variation [1]	SJC-I [53]	Ours
PSNR ↑	7.130	6.561	<u>7.953</u>	**10.405**
SSIM ↑	0.406	0.442	<u>0.456</u>	**0.606**
LPIPS ↓	<u>0.507</u>	0.564	0.545	**0.323**
FID ↓	<u>5.143</u>	10.218	10.202	**0.319**

Table 2: **Results for novel view synthesis on RTMV.** Scenes in RTMV are out-of-distribution from Objaverse training data, yet our model still outperforms the baselines by a significant margin.

Improvement: multi-view generation

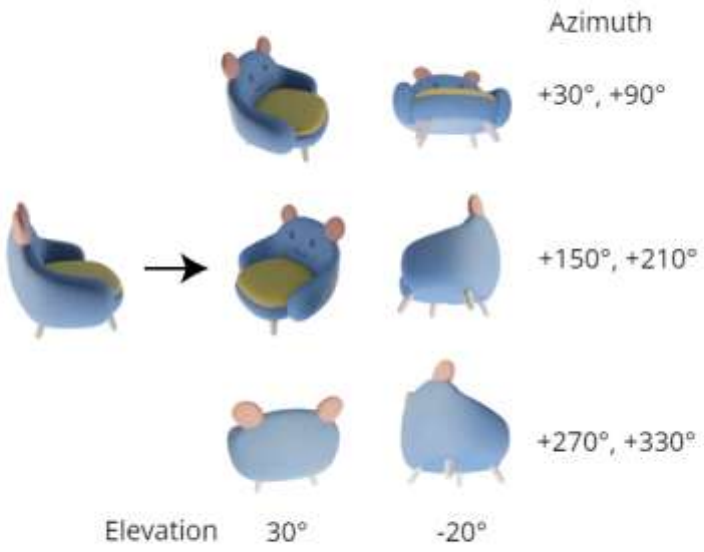


Figure 2. **Layout of Zero123++ prediction target.** We use a fixed set of relative azimuth and absolute elevation angles.

Figure 10: Incorrect elevations lead to distorted reconstruction. Our elevation estimation module can predict an accurate elevation of the input view.

Our reconstruction module requires camera poses for the $4 \times n$ source view images. Note that we adopt Zero123 for image synthesis, which parameterizes cameras in a canonical spherical coordinate frame, (θ, ϕ, r), where θ, ϕ and r represent the elevation, azimuth, and radius. While we can arbitrarily adjust the azimuth angle ϕ and the radius r of all source view images simultaneously, resulting in the rotation and scaling of the reconstructed object accordingly, this parameterization requires knowing the absolute elevation angle θ of one camera to determine the relative poses of all cameras in a standard XYZ frame. More specifically, the relative poses between camera (θ_0, ϕ_0, r_0) and camera $(\theta_0 + \Delta\theta, \phi_0 + \Delta\phi, r_0)$ vary for different θ_0 even when $\Delta\theta$ and $\Delta\phi$ are the same. Because of this, changing the elevation angles of all source images together (*e.g.*, by 30 degrees up or 30 degrees down) will lead to the distortion of the reconstructed shape (see Figure 10 for examples).

Improvement: Local condition

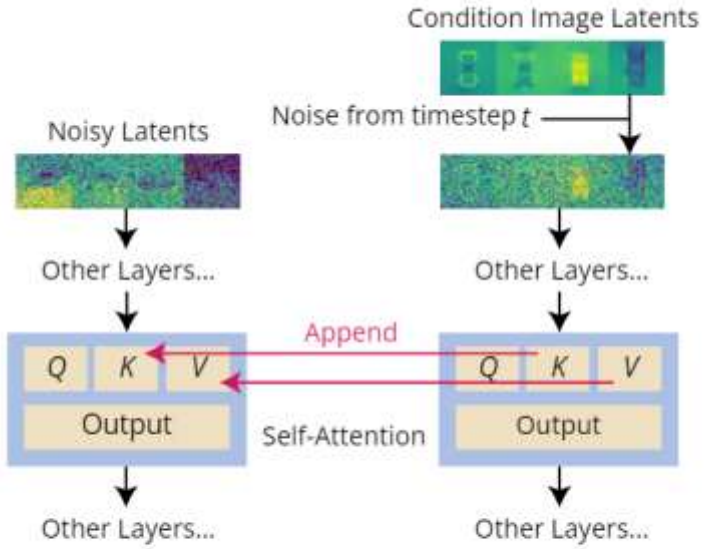


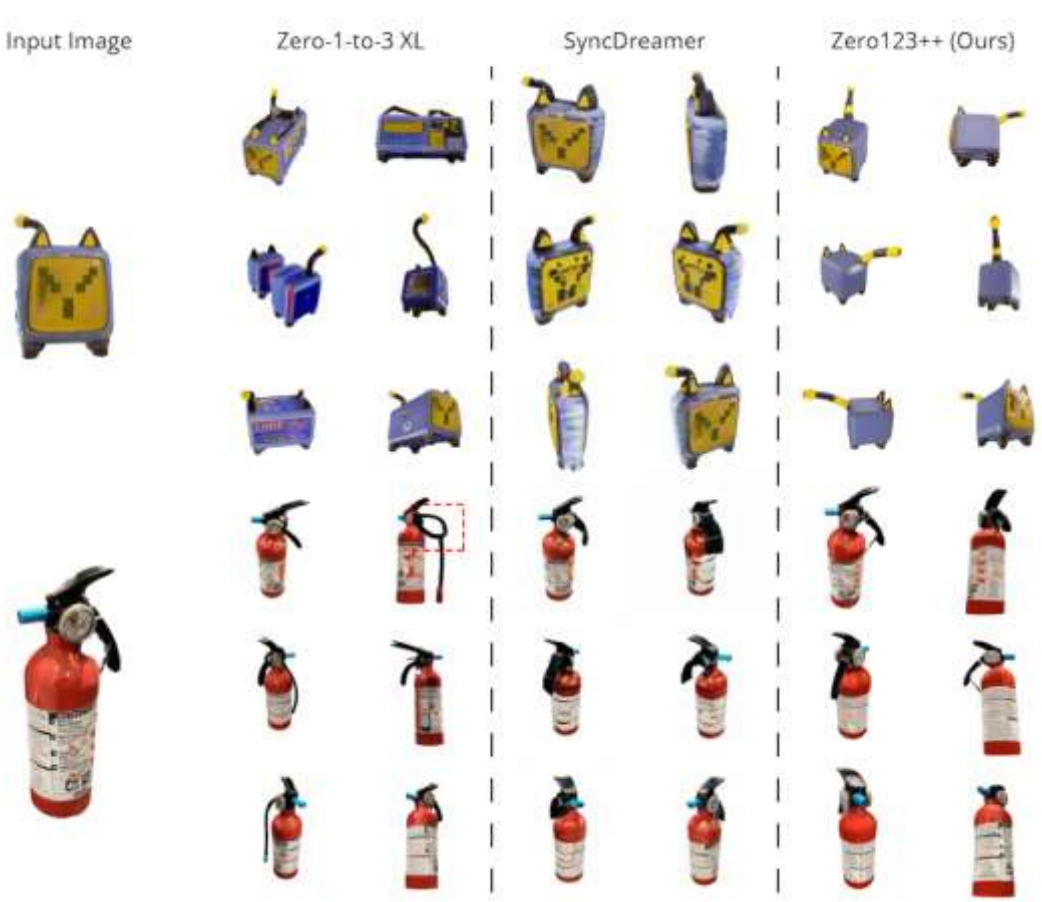
Figure 6. **Reference Attention.** It adds an additional conditioning branch and modifies key (K) and value (V) matrices of the self-attention layers to accept the extra condition image, which can fully reuse Stable Diffusion priors.

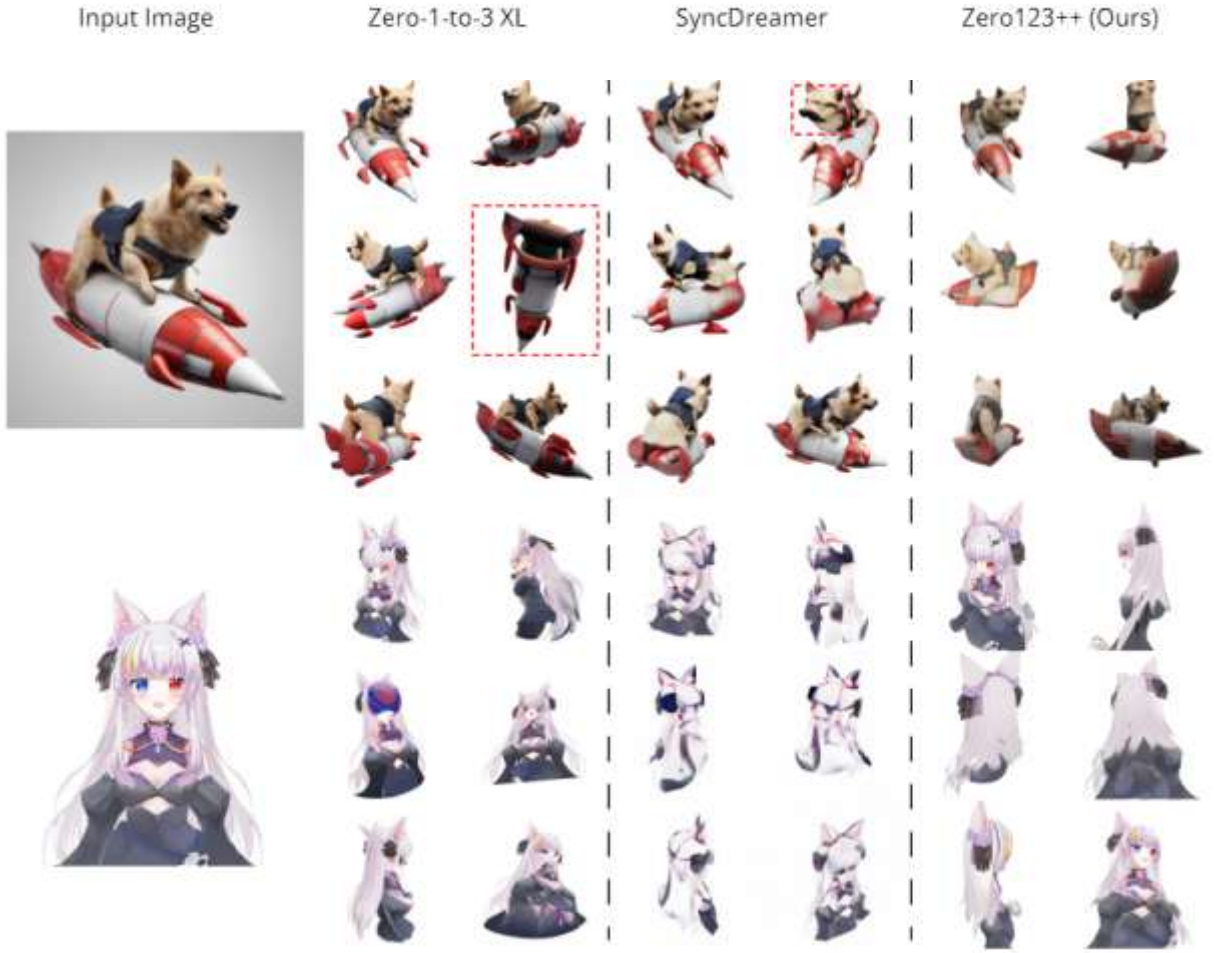
Figure 7. **Comparison on local conditioning.** We train Zero123++ with different levels of scaled reference attention on the ShapeNet Cars dataset. Output coherence with the input image is best on 5x scaled reference attention.

Results

Table 1. Quantitative results of models on our validation split.

Model	LPIPS ↓
Zero-1-to-3	0.210 ± 0.059
Zero-1-to-3 XL	0.188 ± 0.053
Zero123++ (Ours)	**0.177** ± 0.066





Thank you!