Team 2
Final Project Presentation

Sync3D: Single Image Novel View Synthesis via Diffusion Syncing in 3D Space

Asiman Ziyaddinov, Jinhyuk Jang, Prin Phunyaphibarn

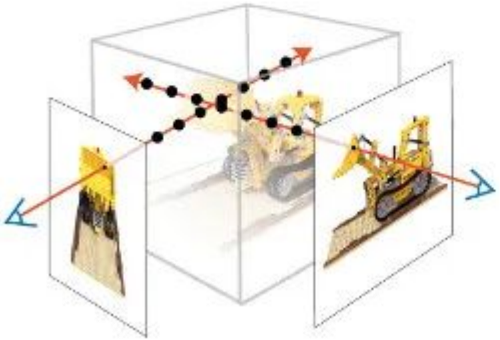
3D Reconstruction (NeRF, 3DGS)

Pros:

+ Simple representation
+ High-Quality Output

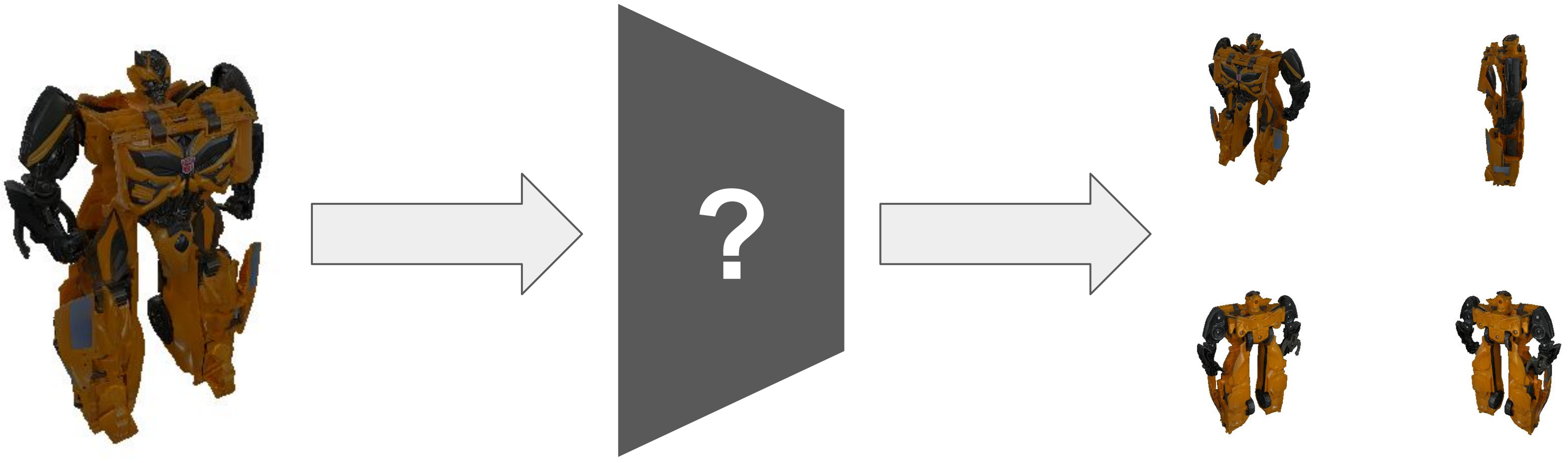
Cons:

- **Dependent on quality of views**
- **Typically requires dense views**



Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.

How can we generate novel views from a single RGB image?



Existing Paradigms in Novel View Synthesis

1. 3D Reconstruction (e.g., NeRF, SinNeRF):

- Encodes scene geometry in a volumetric representation.
- Requires multi-view input or accurate depth maps.

2. Generative Priors (e.g., Zero-1-to-3):

- Learns view synthesis directly from large-scale datasets.
- Outputs are visually compelling but not guaranteed geometrically accurate.

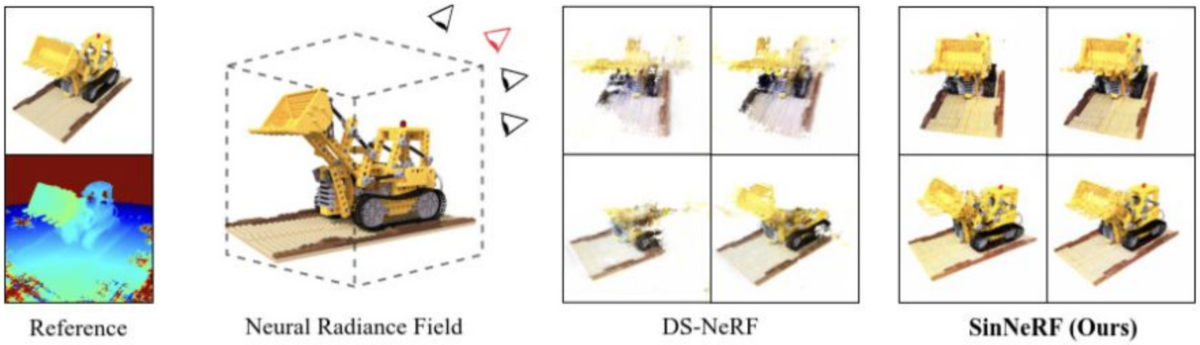
SinNeRF: Training Neural Radiance Fields from a Single Image (ECCV 2022)

Strengths:

+ Enables view-consistent 3D scene reconstruction from single image.

Weaknesses:

- Produces blurry artifacts and broken geometry
- Requires additional cues like accurate depth maps.



TL;DR: Given only a single reference view as input, our novel semi-supervised framework trains a neural radiance field effectively. In contrast, previous method shows inconsistent geometry when synthesizing novel views.

Dejia Xu, et al. 2022, SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image

Zero-1-to-3: Zero-shot One Image to 3D Object (ICCV 2023)

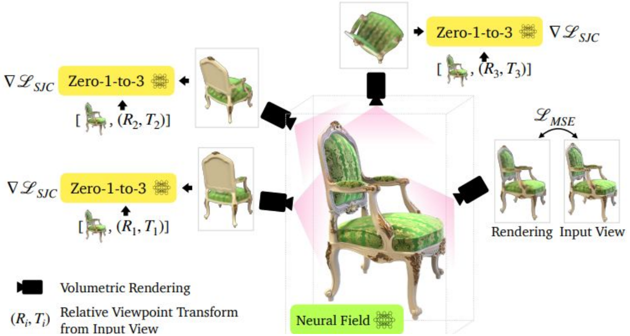
Single-image Novel View Synthesis

Strengths:

+ Data Efficiency
+ Versatile Applications
+ High-Quality Output

Weaknesses:

- Inconsistent Detail
- Dependence on Pre-trained Models



Liu, Ruoshi, et al. "Zero-1-to-3: Zero-shot one image to 3d object." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Single Image Novel View Synthesis:
3D Reconstruction vs. Generative Priors

3D Reconstruction

Pros

(Pre)training-free

multi-view consistent

Cons

Produces blurry artifacts (low quality)

Requires additional information (e.g. depth)

Generative Priors

Pros

High quality

Generalizes to unseen views

Cons

multi-view inconsistent

does not generalize outside training distribution

Combining 3D Representations with Diffusion Priors

Generate novel views using diffusion priors

Enforce multiview consistency by guiding the diffusion process using a unified 3D representation

Recap: Diffusion models progressively denoise an image

Algorithm 1 Diffusion Sampling (DDIM)

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, I)$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad x_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$
4: $\quad x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t)$
5: **end for**
6: **return** $\mathbf{x}_0$

Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models." International Conference on Learning Representations.

Recap: Diffusion models progressively denoise an image

Algorithm 1 Diffusion Sampling (DDIM)

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, I)$

for $t = T, \ldots, 1$ **do**

$\quad x_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$

$\quad x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}x_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(x_t)$

end for

return $\mathbf{x}_0$

How can we guide the diffusion process during the denoising phase?

Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models." International Conference on Learning Representations.

Diffusion Guidance

The Diffusion Process can be guided using the gradient of a loss function

Noisy Images

x_t

"Predicted" Clean Images **(Tweedies)**

$x_{0|t}$

Inject Guidance

Bansal, Arpit, et al. "Universal guidance for diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Diffusion Guidance

1. Compute Tweedies: $x_{0|t} = \frac{1}{\bar{\alpha}_t}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$

2. Update noisy sample using backpropagation:

$$\tilde{x}_t = x_t - \eta \nabla_{x_t} \ell(x_0)$$

3. Denoise the updated sample

Diffusion Guidance

1. Compute Tweedies: $x_{0|t} = \frac{1}{\bar{\alpha}_t}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$

Problem: Gradients with respect to x_t is unstable

2. Update noisy sample using backpropagation:

$$\tilde{x}_t = x_t - \eta \nabla_{\boxed{x_t}} \ell(x_0)$$

3. Denoise the updated sample

Poole, Ben, et al. "DreamFusion: Text-to-3D using 2D Diffusion." The Eleventh International Conference on Learning Representations.

Diffusion Guidance

1. Compute Tweedies: $x_{0|t} = \frac{1}{\bar{\alpha}_t}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$

2. Update noisy sample using backpropagation: **Solution: Take gradients w.r.t $x_{0|t}$**

$$\tilde{x}_t = x_t - \eta \nabla_{\boxed{x_{0|t}}} \ell(x_0)$$

3. Denoise the updated sample

Ye, Haotian, et al. "TFG: Unified Training-Free Guidance for Diffusion Models." The Thirty-eighth Annual Conference on Neural Information Processing Systems.

Diffusion Guidance

Algorithm 2 Diffusion Guidance

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, I)$

for $t = T, \ldots, 1$ **do**

$\quad x_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$

$\quad \tilde{x}_t = x_t - \eta \nabla_{x_{0|t}} \ell(x_0)$

$\quad \tilde{x}_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(\tilde{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\tilde{x}_t)\right)$

$\quad x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\tilde{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\tilde{x}_t)$

end for

return $\mathbf{x}_0$

Diffusion Guidance

Algorithm 2 Diffusion Guidance

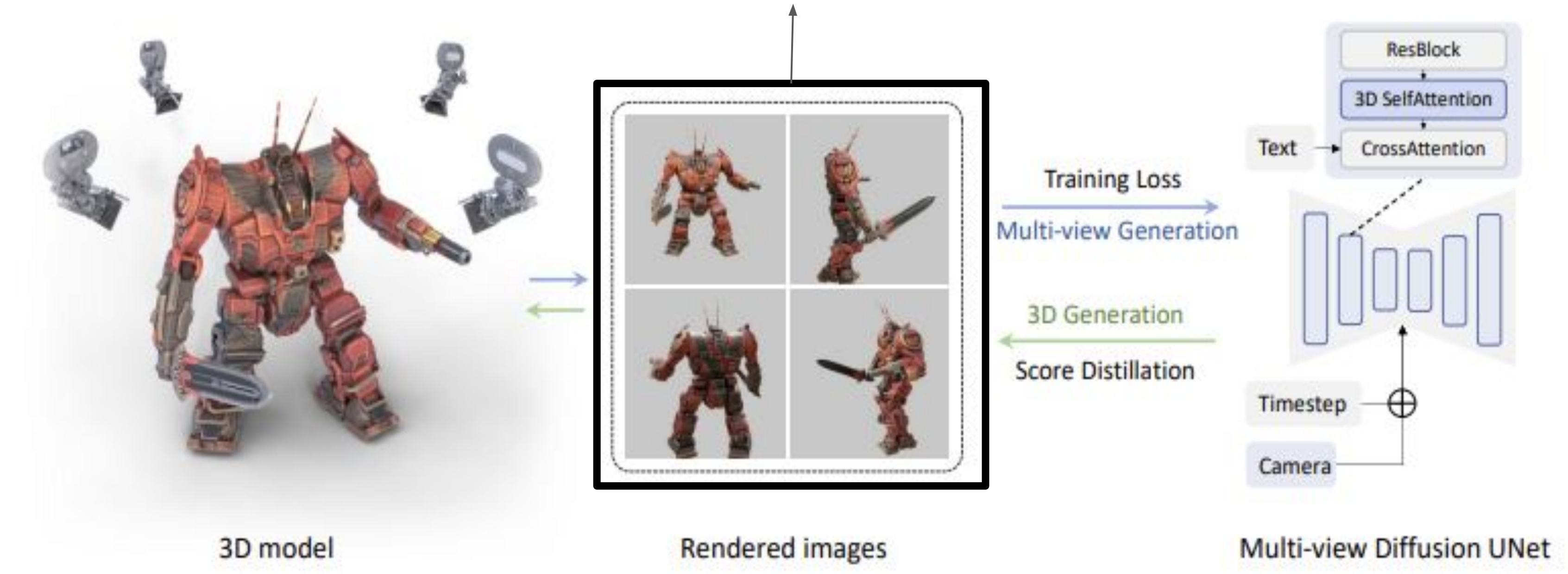
1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, I)$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad x_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t)\right)$
4: $\quad \tilde{x}_t = x_t - \eta \nabla_{x_{0|t}} \ell(x_0)$
5: $\quad \tilde{x}_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(\tilde{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\tilde{x}_t)\right)$
6: $\quad x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\tilde{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\tilde{x}_t)$
7: **end for**
8: **return** $\mathbf{x}_0$

But what loss do we use?

Designing the View-Consistency Loss: Incorporating 3D Priors

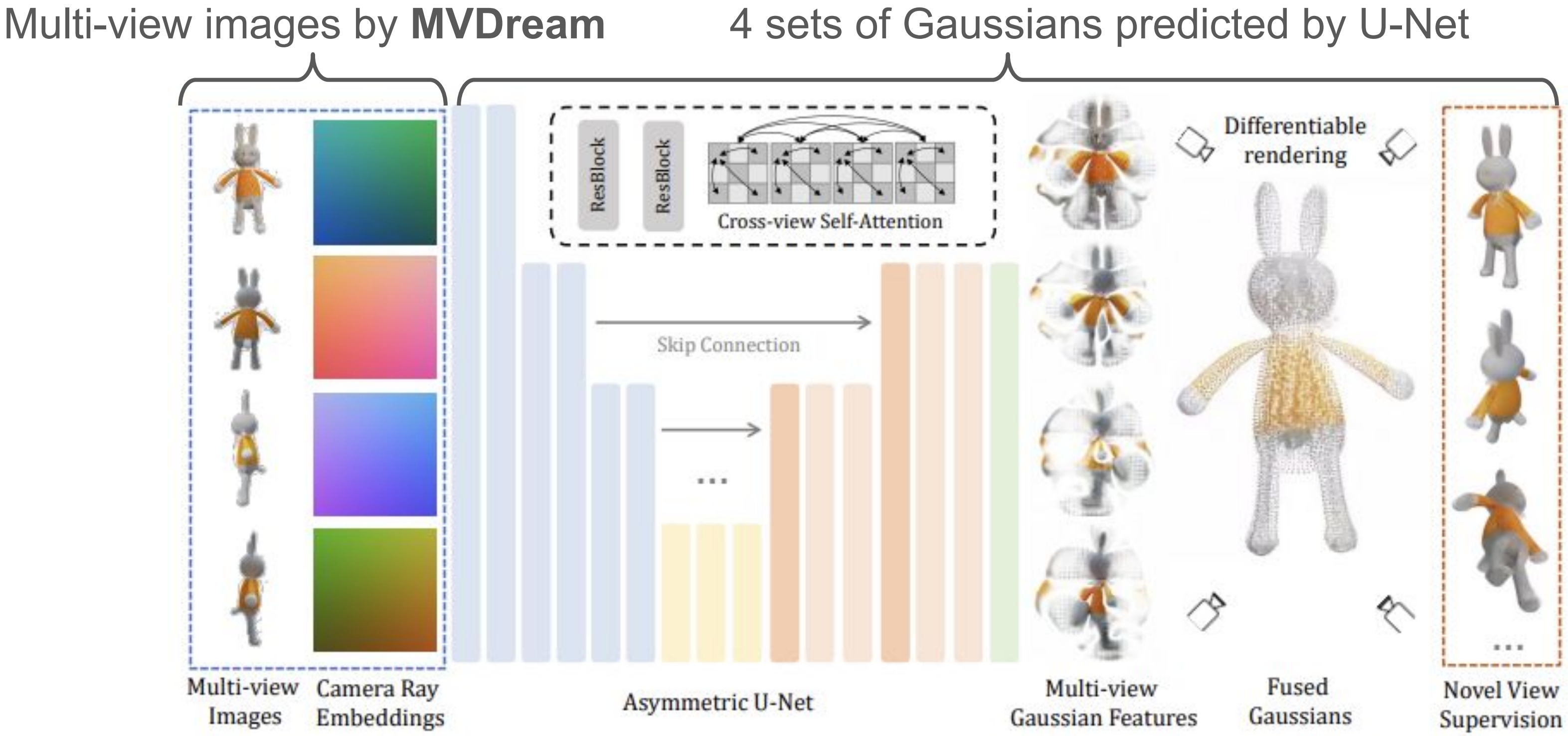
MVDREAM: Multi-view Diffusion for 3D Generation

multi-view images at four orthogonal angles at a fixed elevation



Shi, Yichun, et al. "MVDream: Multi-View Diffusion for 3D Generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024

LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation



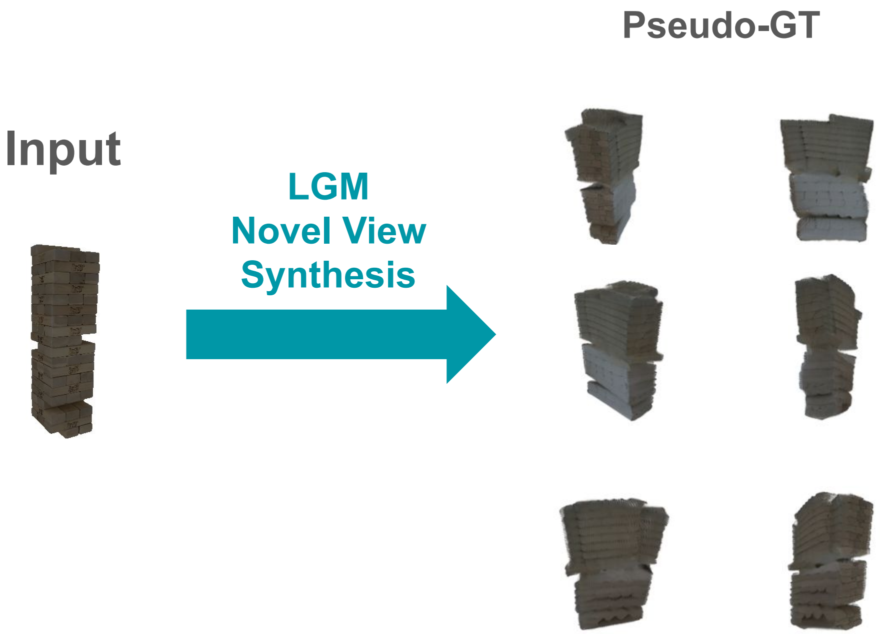
Jiaxiang Tang, et al. "LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation", 2024

LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation

Input

Diffusion Guidance via Pseudo Ground Truth Views

Generate Pseudo Ground Truth Views using LGM



Semantic Guidance

Using MSE captures too many high-level details (LGM produces blurry views)

Use **LPIPS** to capture low-level structure

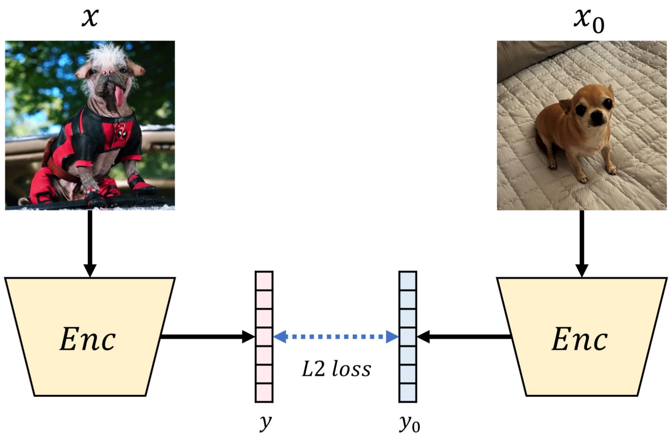


Figure from https://ethanswinery.tistory.com/26

Geometric Guidance

Zero-1-to-3
Predicted
Depth

MSE Loss

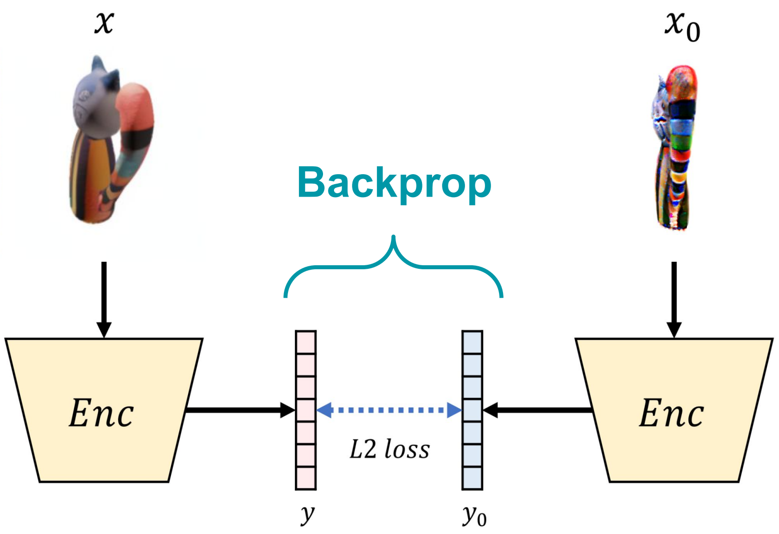
Guidance

Pseudo-GT
Predicted
Depth

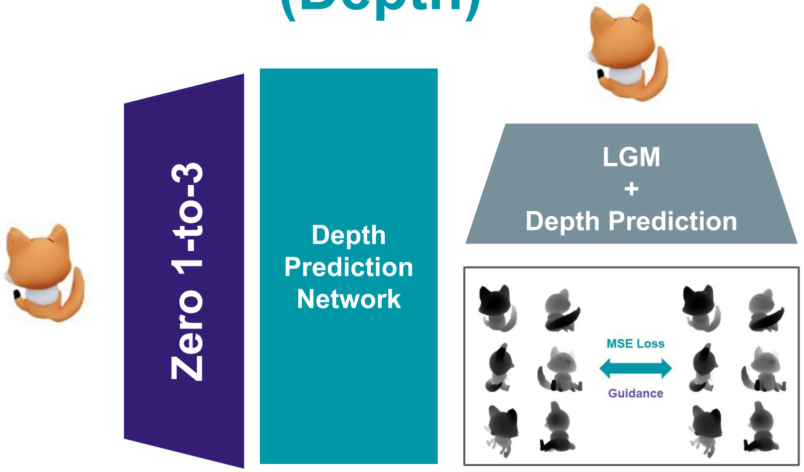
Figure from **Shi, Ruoxi, et al. "Zero123++: a single image to consistent multi-view diffusion base model." arXiv preprint arXiv:2310.15110 (2023).**

Putting it Together

Semantic Guidance (LPIPS)



Geometric Guidance (Depth)



MVDream vs. Our Method

MVDream produces **fixed** views
Our method can generate views from **arbitrary** camera positions/orientation

Experimental Results

Quantitative Results

We evaluate using **Google Scanned Objects** dataset (>1000 scanned objects).

We report the average LPIPS, PSNR, and SSIM of 6 rendered views per object.

Method	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**
Zero-1-to-3	0.211	16.037	<u>0.824</u>
LGM	0.273	14.717	0.819
Ours (w/o UNet Gradients) +LPIPS Guidance	<u>0.199</u>	**16.403**	0.816
Ours (w/o UNet Gradients) +LPIPS Guidance +Depth Guidance	**0.198**	<u>16.397</u>	**0.830**

Downs, Laura, et al. "Google scanned objects: A high-quality dataset of 3d scanned household items." 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022.

Ablation Study: UNet Gradients

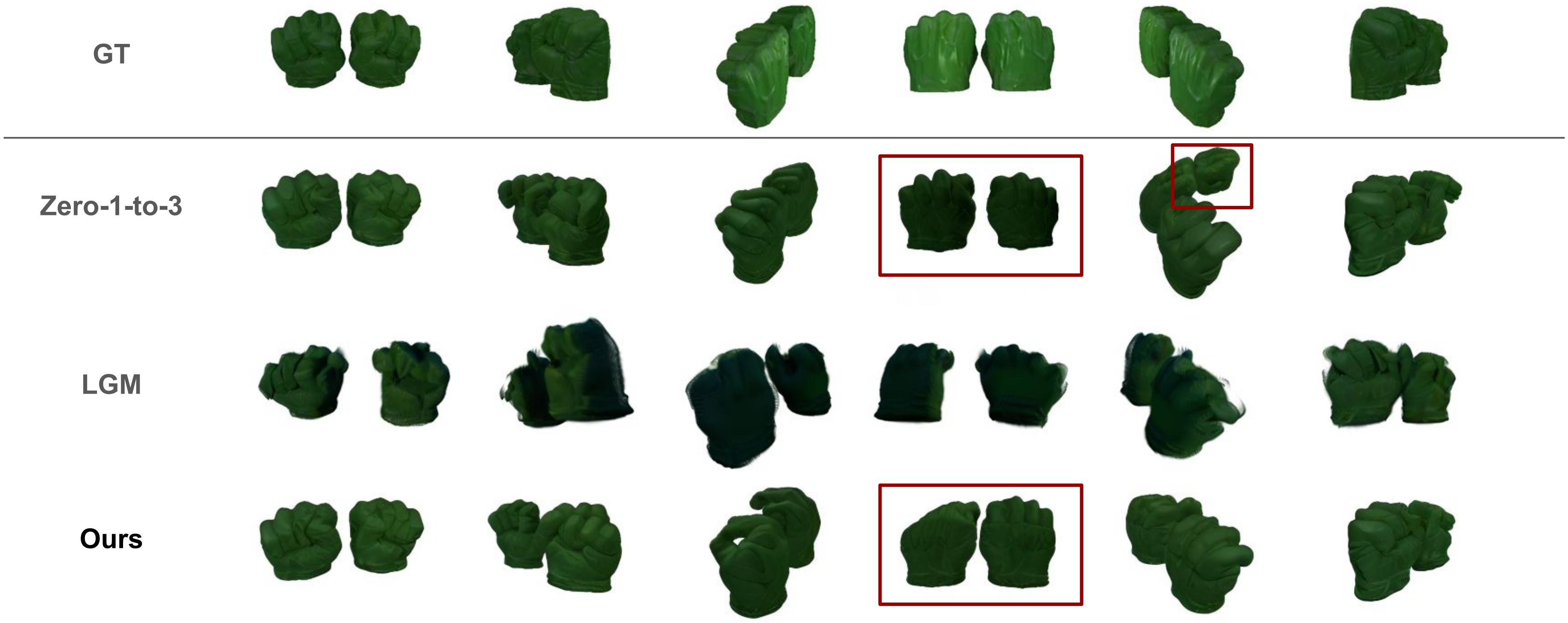
Method	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**
Ours **(w/o UNet Gradients)** +LPIPS Guidance	**0.199**	**16.403**	**0.816**
Ours (w/ UNet Gradients) +LPIPS Guidance	0.202	16.316	0.827
Ours **(w/o UNet Gradients)** +LPIPS Guidance +Depth Guidance	**0.198**	**16.397**	**0.830**
Ours (w/ UNet Gradients) +LPIPS Guidance +Depth Guidance	0.199	16.370	0.829

Ablation Study: LPIPS vs MSE Guidance

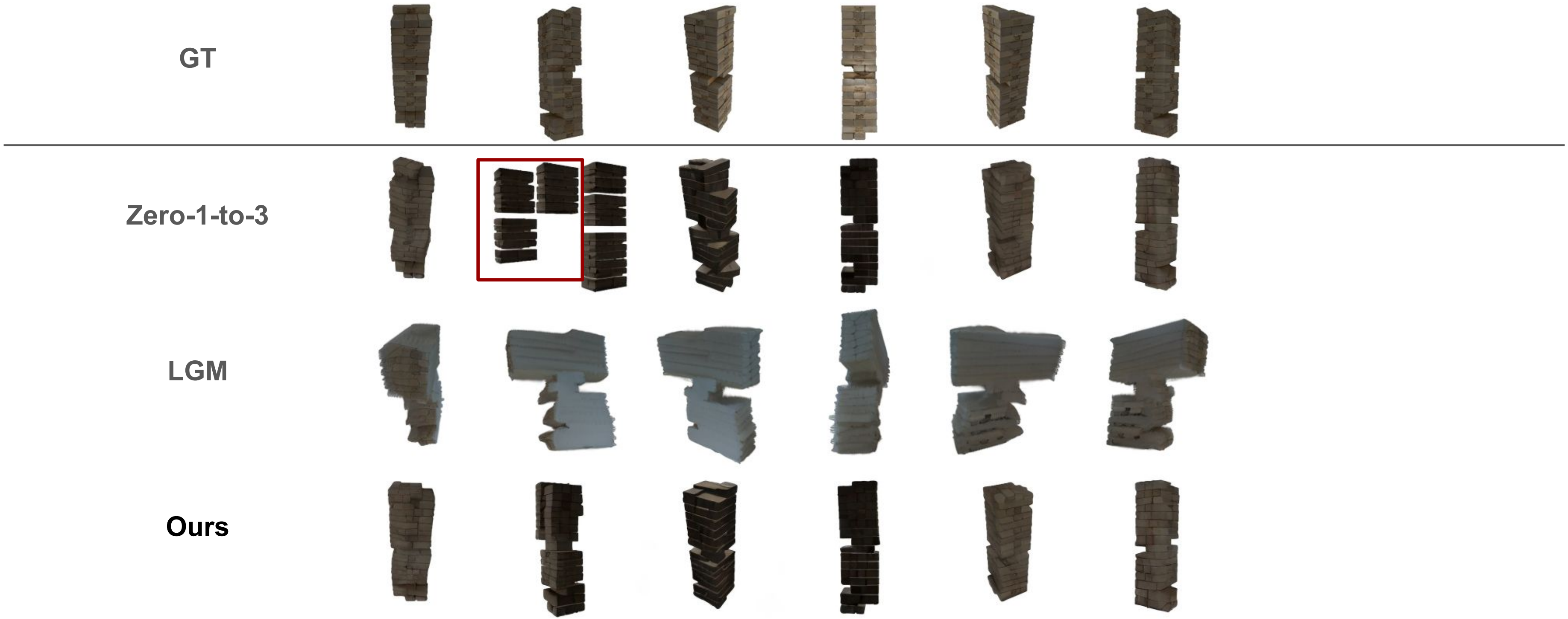
Method	LPIPS ↓	PSNR ↑	SSIM ↑
Ours (w/o UNet Gradients) +LPIPS Guidance	**0.199**	**16.403**	**0.816**
Ours (w/o UNet Gradients) +**MSE** Guidance	0.206	16.225	0.827

Qualitative Results

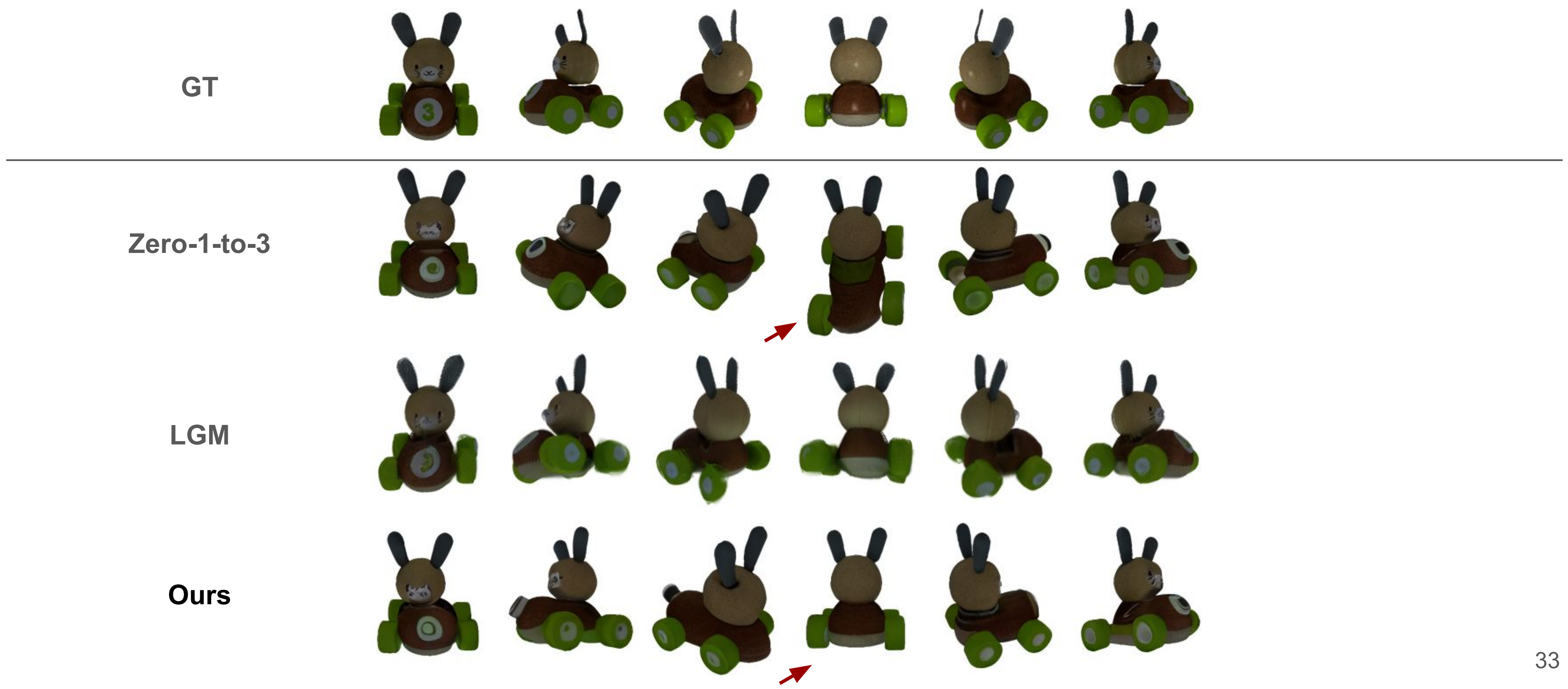
Improves consistency



Reduces hallucination artifacts and improves consistency



Improves view-alignment



Conclusion

We leverage LGM to produce a unified 3D representation which we use to generate pseudo ground truth views to guide the diffusion process via semantic and depth guidance to achieve high-quality multiview-consistent generations.

Limitations

- Diffusion guidance take more time (~1 min. per 6 views)
- More memory intensive–need to load 3 models
- Dependent on quality of LGM and MVDream

Contributions

Prin: Zero-1-to-3 pipeline, LPIPS guidance, and evaluation code

Jinhyuk: Integrate 3D reconstruction (LGM) into the pipeline

Asiman: Depth prediction and depth guidance