

# SpaRP: Fast 3D Object Reconstruction and Pose Estimation from Sparse Views

Chao Xu  
Sep 29, 2024



# Single view to 3D



Input Image



Stable Fast 3D



Rodin Gen-1 v0.8



Tripo v2.0

# Single view to 3D is **ill-posed**



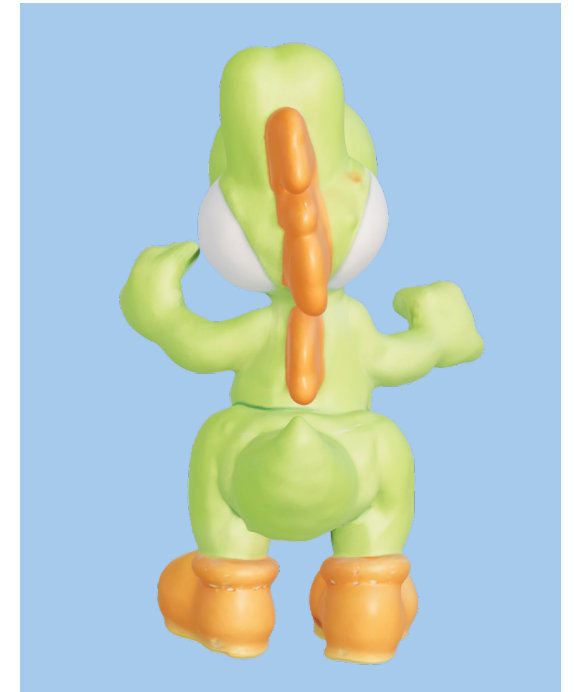
Input Image



Stable Fast 3D



Rodin Gen-1 v0.8



Tripo v2.0

Hallucination is **uncontrollable** and may lead to **undesired** results.

# Sparse views to 3D is **more controllable**



Unposed Input Images

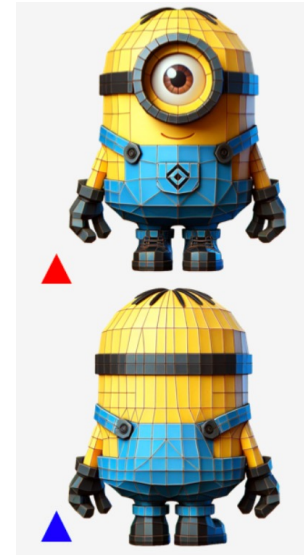
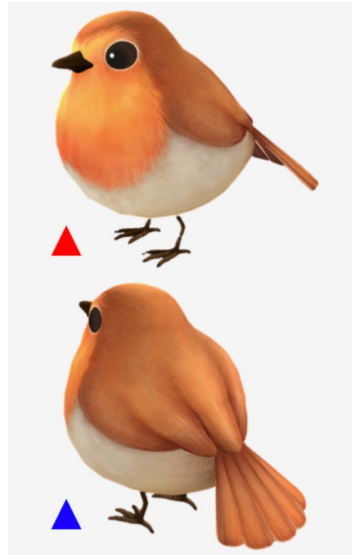


Ours

Our method generates 3D assets closely following the reference unposed images, overcoming the ambiguity inherent in single-view-to-3D.

# More generally ...

Given a few (1~6) unposed images, how to understand their spatial relationship?



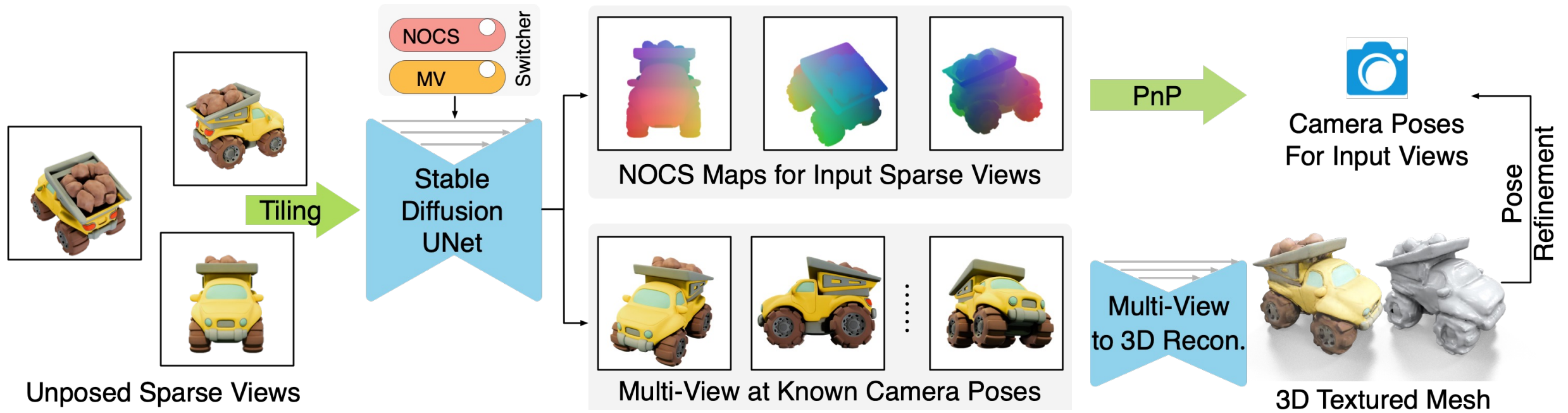




## SpaRP focuses on two tasks:

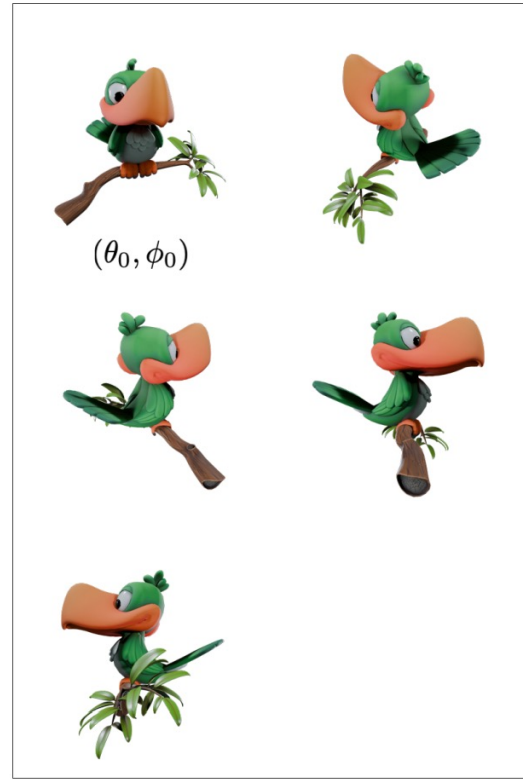
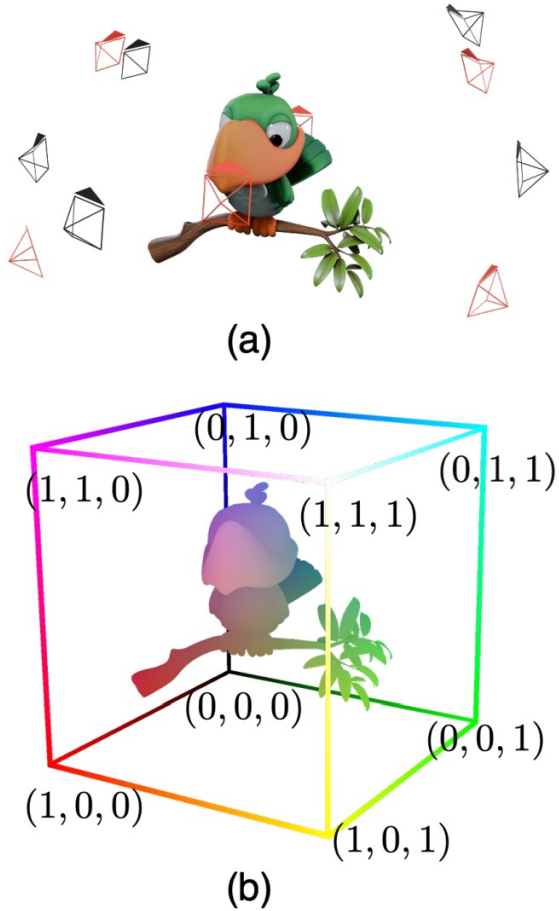
- 3D Reconstruction
- Pose Estimation

# Pipeline

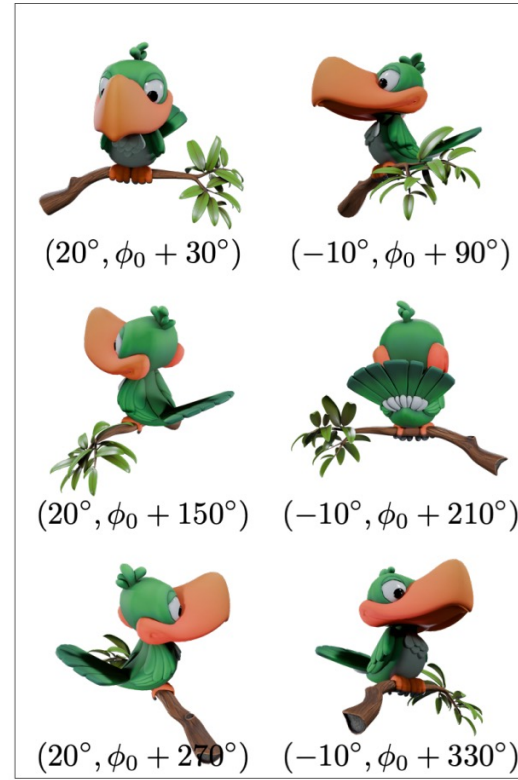


## Pipeline Overview

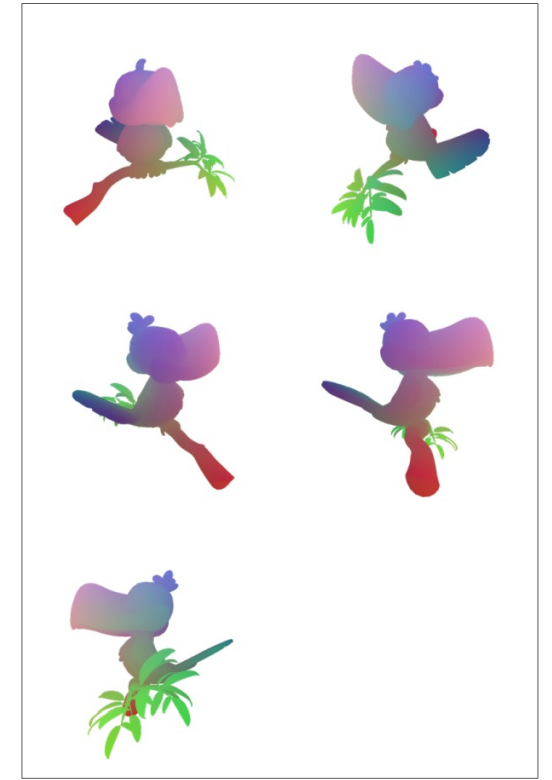
# Pipeline



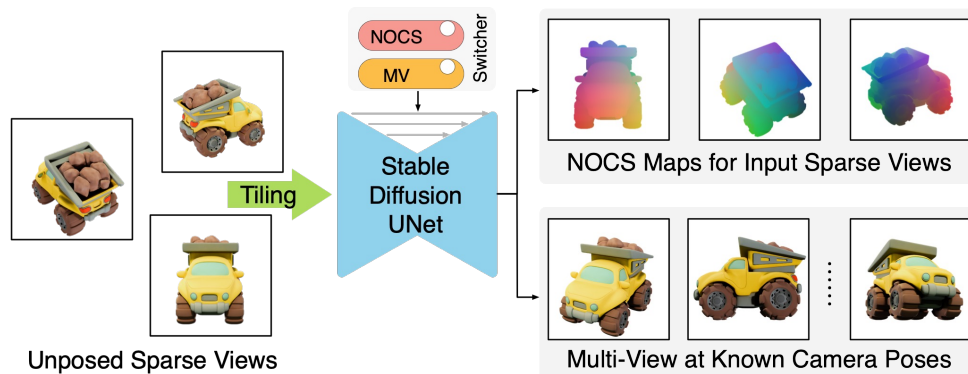
Input (Tiled) Sparse Views



Output (Tiled) Multiviews



Output (Tiled) NOCS Maps



✦ **Tiling** is simple but effective.

✦ **Jointly predict NOCS maps** for pose estimation and **canonical multiviews** for 3D reconstruction.

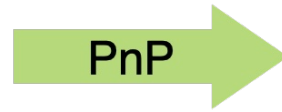
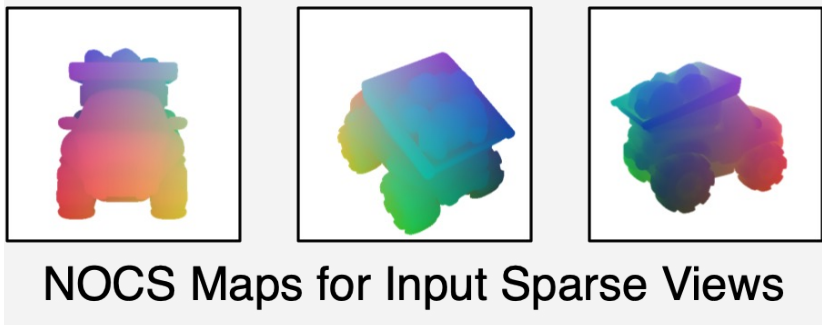


# Pipeline

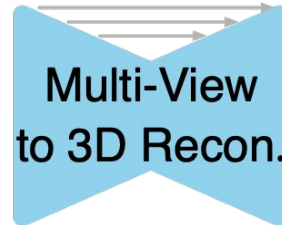
$$\xi_i^{\text{pnp}} = \arg \min_{\xi_i \in \text{SE}(3)} \sum_{j=1}^{m_i} \|\mathbf{p}_{i,j} - \text{proj}(\mathbf{q}_{i,j}, \xi_i)\|_2^2,$$

Camera Pose  
(PnP Ransac)

NOCS pixel location NOCS point coordinate Estimated Pose



Camera Poses  
For Input Views



3D Textured Mesh

Pose  
Refinement

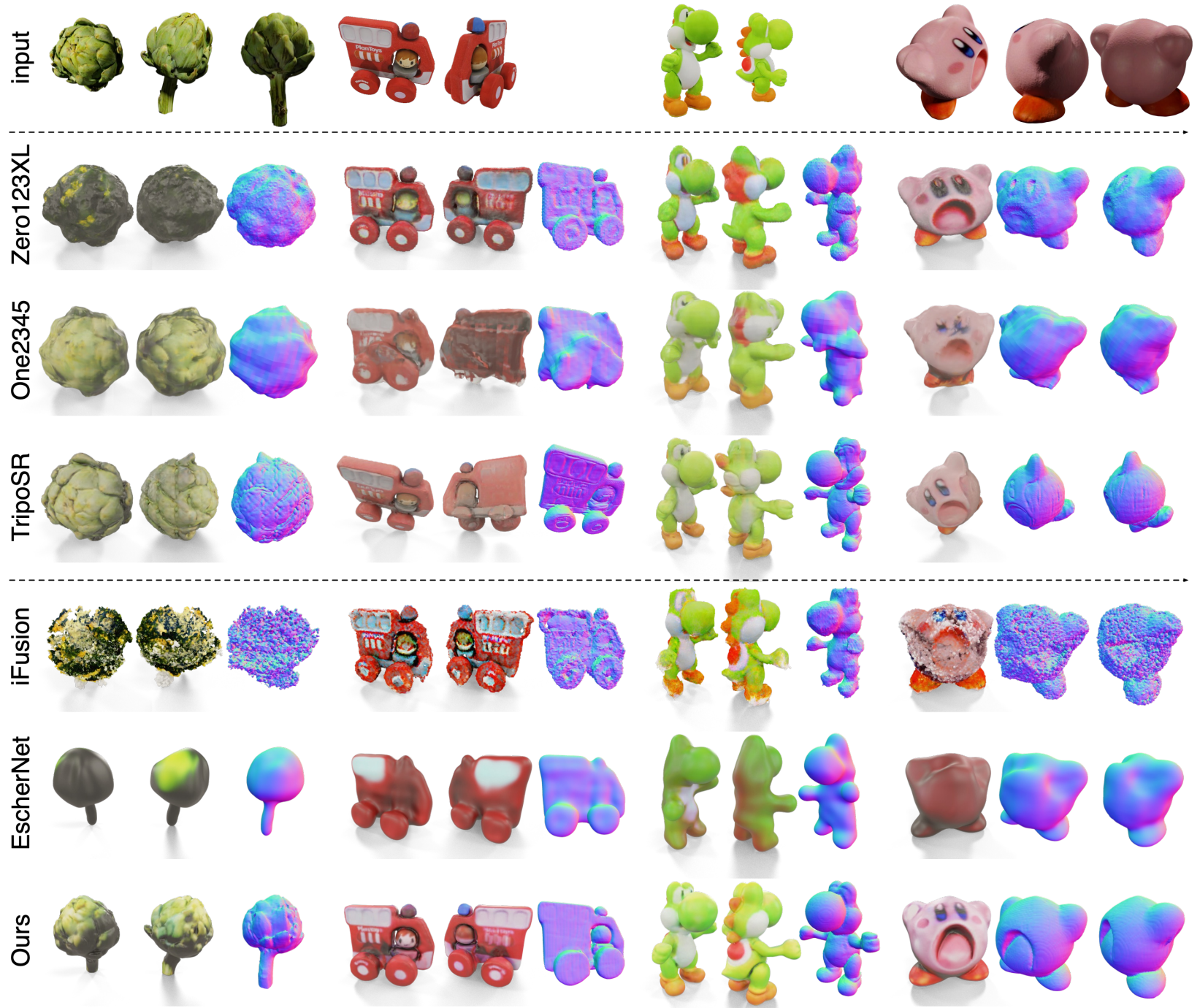
$$\xi_i^* = \arg \min_{\xi_i \in \text{SE}(3)} (\lambda \cdot \mathcal{L}_{\text{mask}}(\mathbf{I}_i^r, \mathbf{I}_i) + \mu \cdot \mathcal{L}_{\text{rgb}}(\mathbf{I}_i^r, \mathbf{I}_i)),$$

**Pose refinement** via differentiable rendering in under a second

Many compatible **multi-view-to-3D** models:  
One-2-3-45++, InstantMesh, GRM, MeshLRM, MeshFormer, etc.

# 3D Recon Comparison

Single Image to 3D  
(only the first image is used)



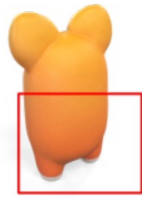
Sparse Images to 3D

## 3D Recon Comparison

single-view  
input



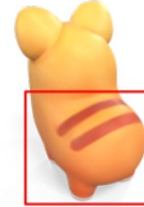
single-view  
output mesh



additional  
input views

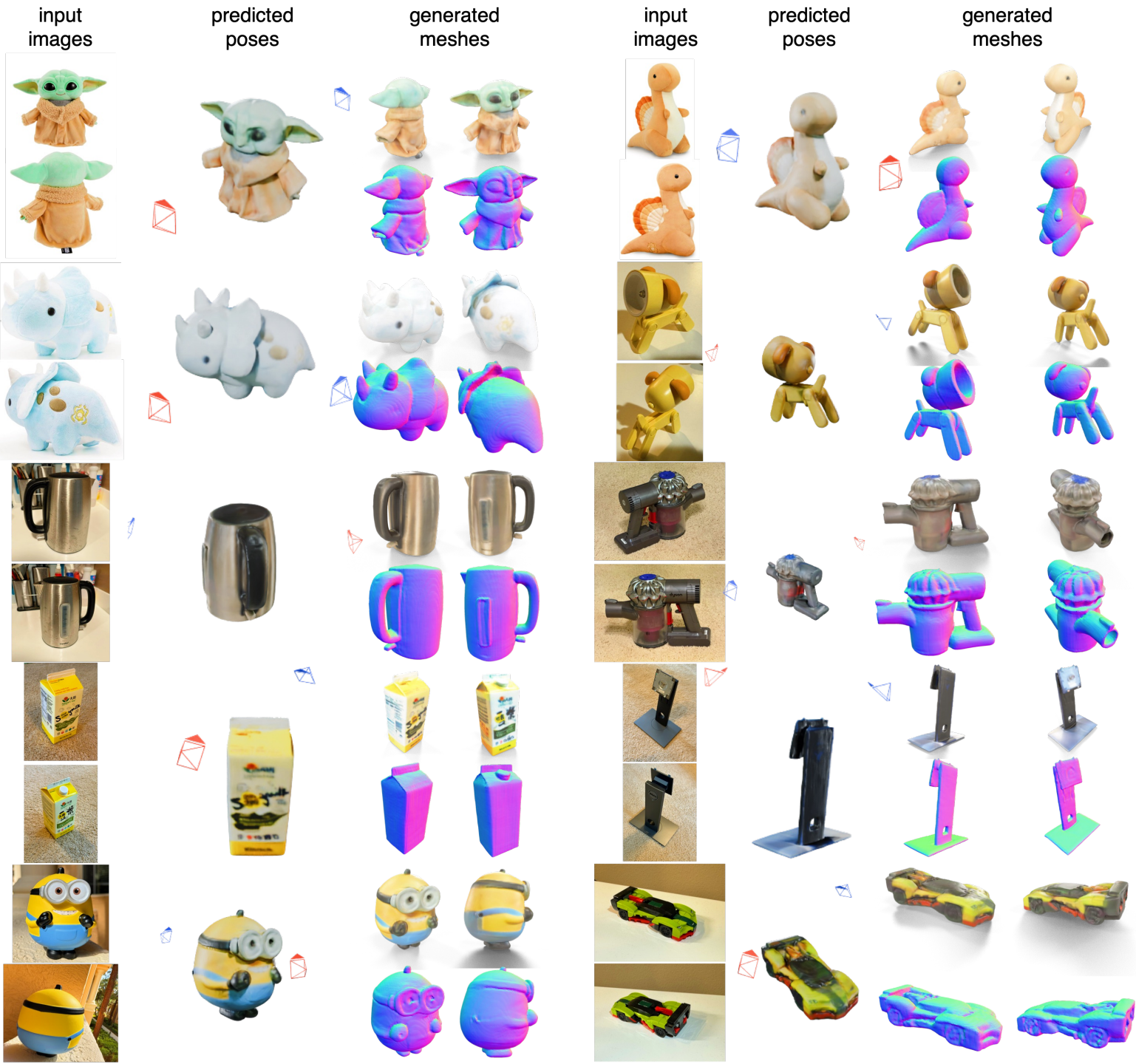


multi-view  
output mesh



More Unposed Views  
Less Unwanted Surprise





# Takeaways:

1. Diffusion models can implicitly learn from unposed sparse views.
2. We can diffuse NOCS, a surrogate representation to predict input poses.
3. Pose prediction and 3D reconstruction can complement each other.

