Thinking Slow and Fast: Recent Trends in 3D Generative Models

Varun Jampani
Vice President of Research
Stability AI

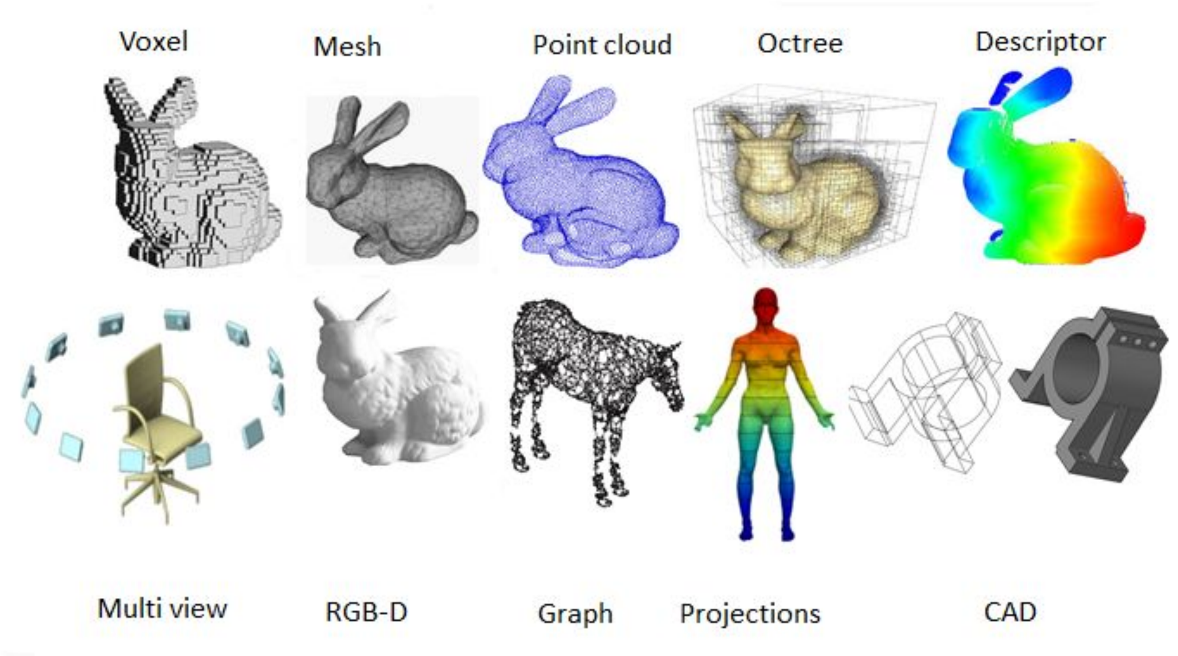
Generative 3D

Text
(or)
Image
(or)
Video

→ **3D Generative Model** →

But, 3D is different from image and language

1. No standard/universal 3D representation



1. Gezawa et al. “A review on deep learning approaches for 3D data representations in retrieval and classifications." *IEEE access* (2020).

But, 3D is different from image and language

1. No standard/universal 3D representation
2. 3D is niche and for power users

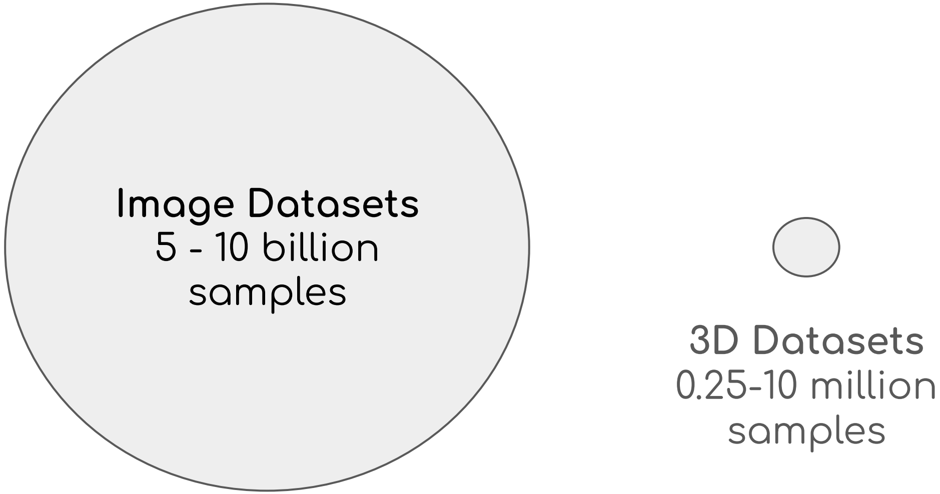
Text

Images or Videos

3D

But, 3D is different from image and language

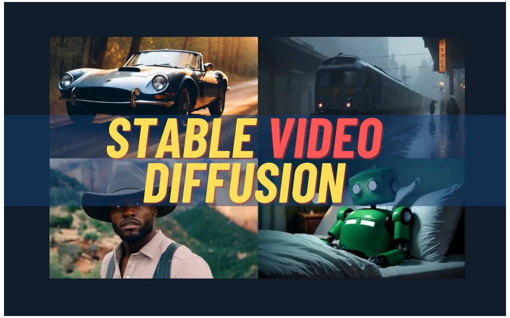
1. No standard/universal 3D representation
2. 3D is niche and for power users
3. **3D data is orders of magnitude smaller**



But, 3D is different from image and language

1. No standard/universal 3D representation
2. 3D is niche and for power users
3. 3D data is orders of magnitude smaller
4. No single/unified generative model exists for different 3D use cases

Image Model



Video Model

?

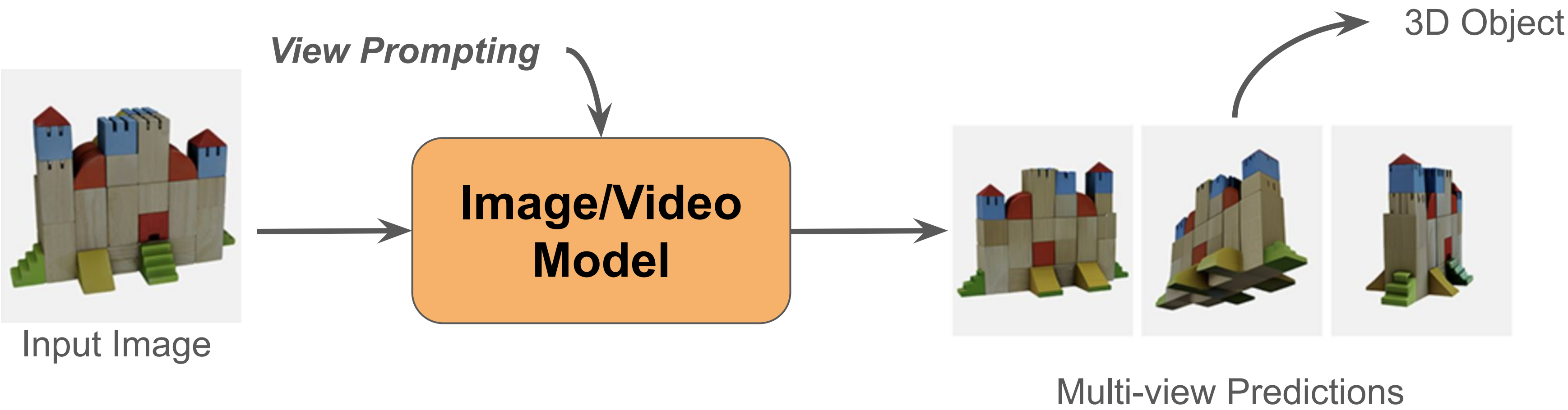
3D Model

Two emerging techniques in 3D Generative Models

1. Multi-view Generation
2. Direct 3D Generation

Two emerging techniques in 3D Generative Models

1. Multi-view Generation

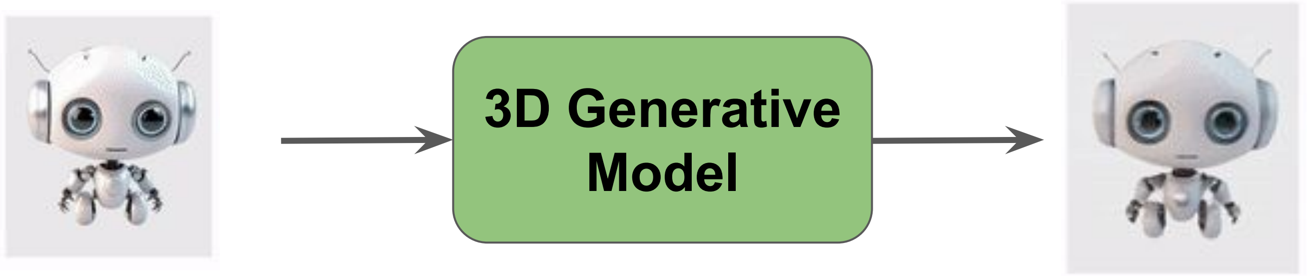


Pros: Leverages image/video models trained on large data and thus have good generalization

Cons: Usually **slow** and also requires further processing to get 3D objects

Two emerging techniques in 3D Generative Models

1. Multi-view Generation
2. Direct 3D Generation



Input Image

Pros: Usually quite **fast** due to direct prediction

Cons: Need good amount of 3D datasets to train and generalize

Multi-View Generation

Multi-view Generation with Image/Video Models

- Score distillation sampling (SDS) Loss
- Reconstruction Losses

1. Text based
2. Camera pose based

View Prompting

3D Object

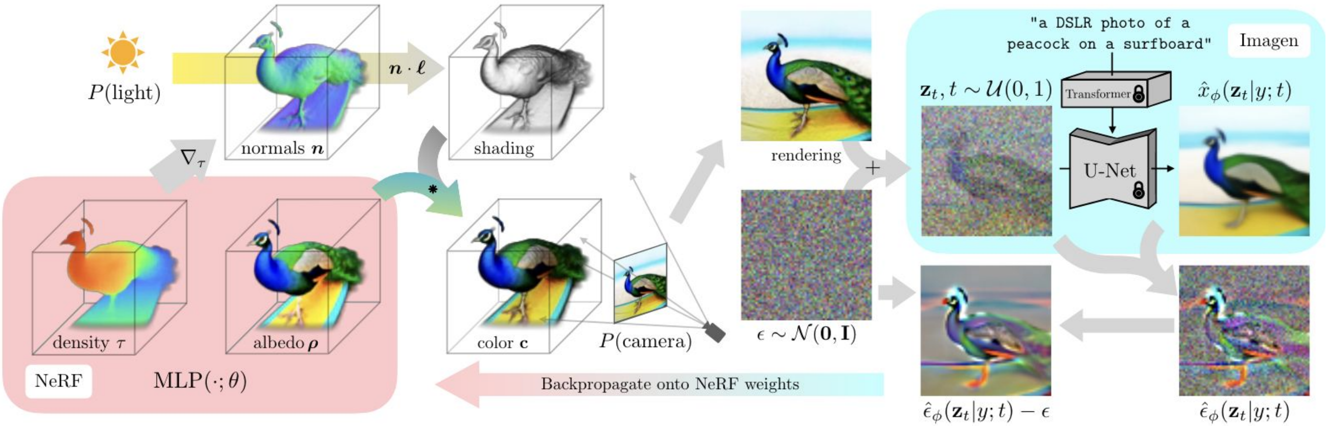
Input Image

Image/Video Model

Multi-view Predictions

SDS Loss with Text-based View Prompting

- DreamFusion [1], Latent-NeRF [2], Magic3D [3]



Score Distillation Sampling (SDS) in DreamFusion [1]

1. Poole et al. DreamFusion: Text-to-3D using 2D diffusion. ICLR 2023
2. Metzer et al. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. CVPR 2023
3. Lin et al. Magic3D: High-Resolution Text-to-3D Content Creation. CVPR 2023

SDS Loss with Text-based View Prompting

- DreamFusion [1], Latent-NeRF [2], Magic3D [3]

Sample results of DreamFusion

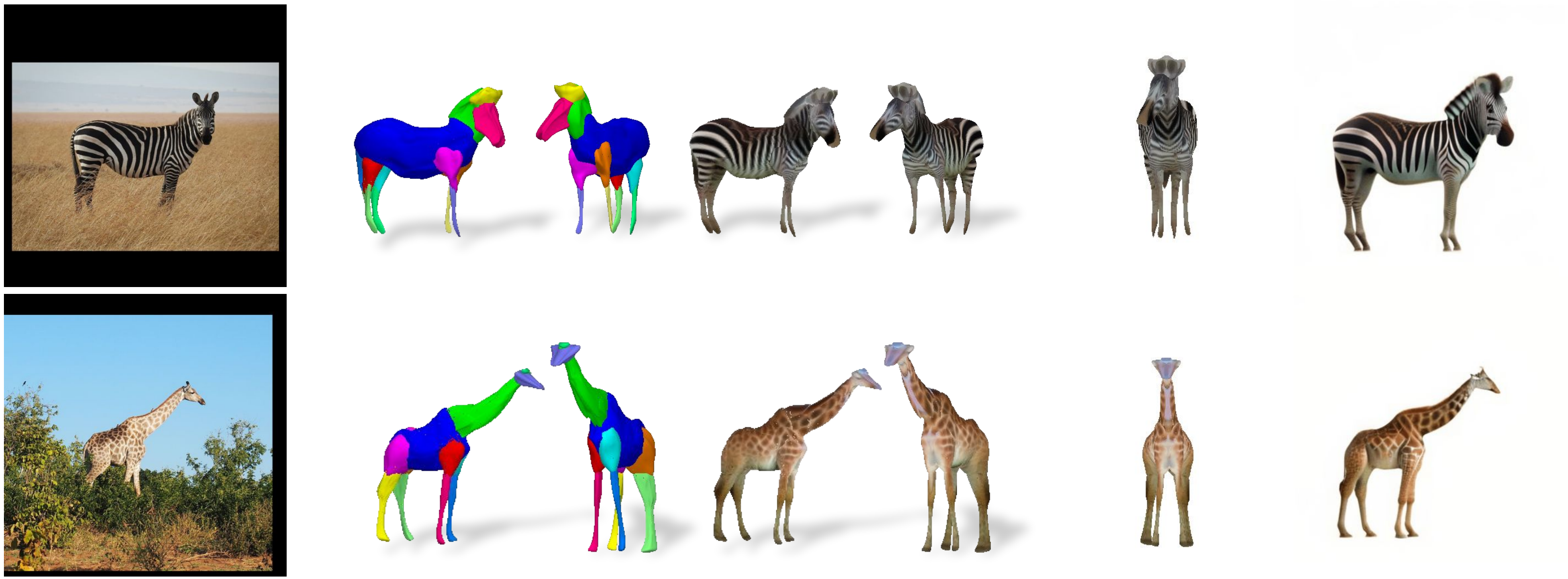
1. Poole et al. DreamFusion: Text-to-3D using 2D diffusion. ICLR 2023
2. Metzer et al. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. CVPR 2023
3. Lin et al. Magic3D: High-Resolution Text-to-3D Content Creation. CVPR 2023

ARTIC3D: Learning Robust Articulated 3D Shapes from Noisy Web Image Collections

Chun-Han Yao, Amit Raj, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein,
Ming-Hsuan Yang, Varun Jampani

NeurIPS'23

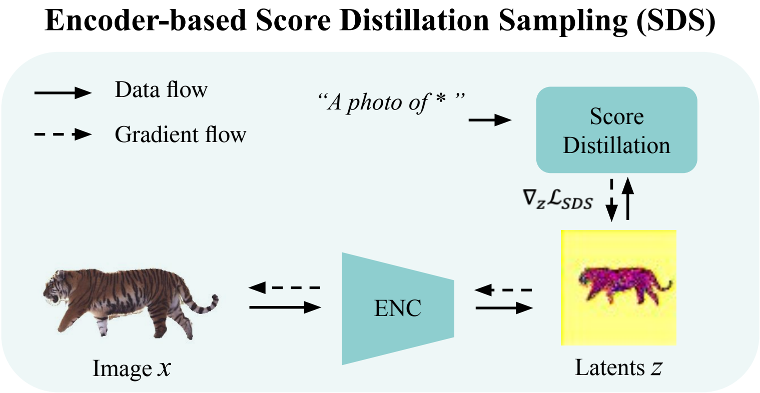
3D Articulated Animals from Noisy Web Images

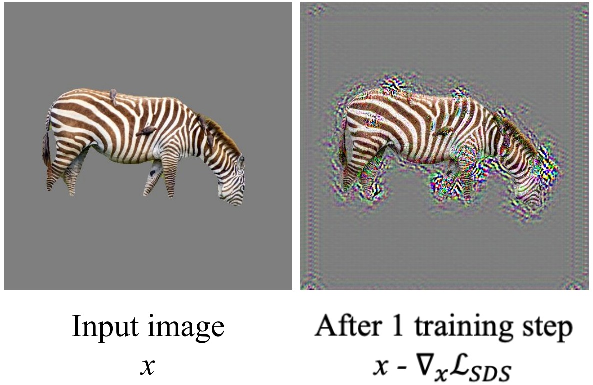


Noisy web images | 3D articulated shapes and texture | Animated | Fine-tuned animation

Issues with standard SDS loss

- Encoder-based SDS [1]: backprop gradients through encoder
    - Noisy gradients
    - High computational costs

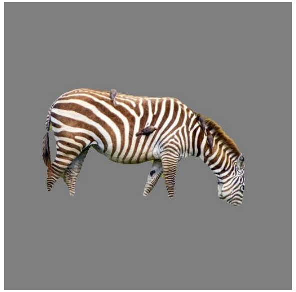




1. Poole et al. "Dreamfusion: Text-to-3d using 2d diffusion." *ICLR* 2023.

DASS: Decoder based Accumulative Score Sampling

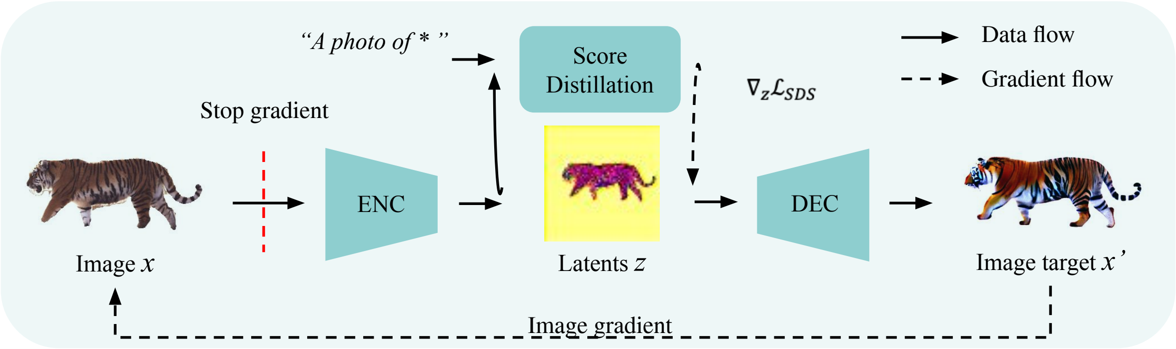
- DASS: decode accumulated latent updates
  - Low memory consumption
  - Clean gradients (stable training)



Input image
x

After 1 training step
$x - \nabla_x \mathcal{L}_{DASS}$

Decoder-based Accumulative Score Sampling (DASS)

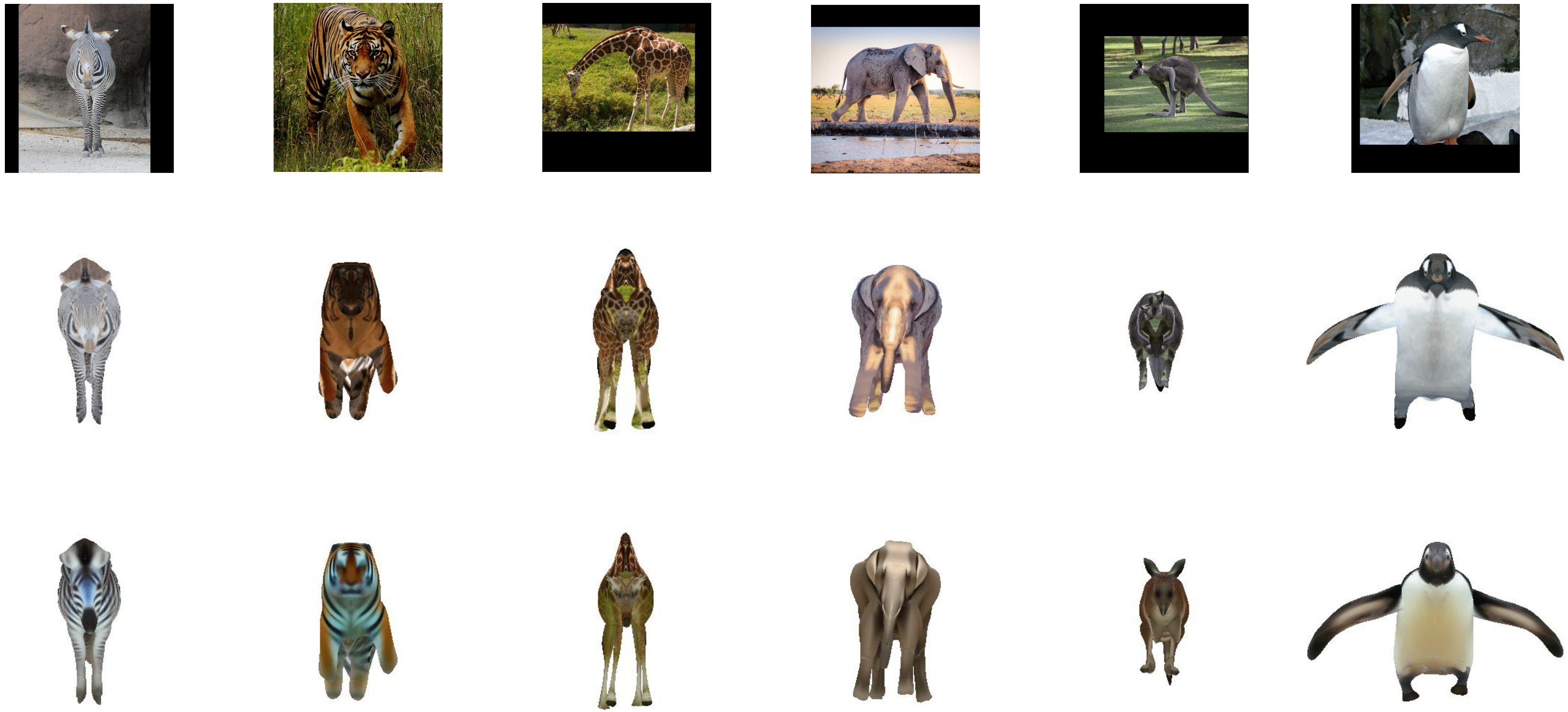


Sample ARTIC3D Results on Web Images

Input images

Hi-LASSIE [1]

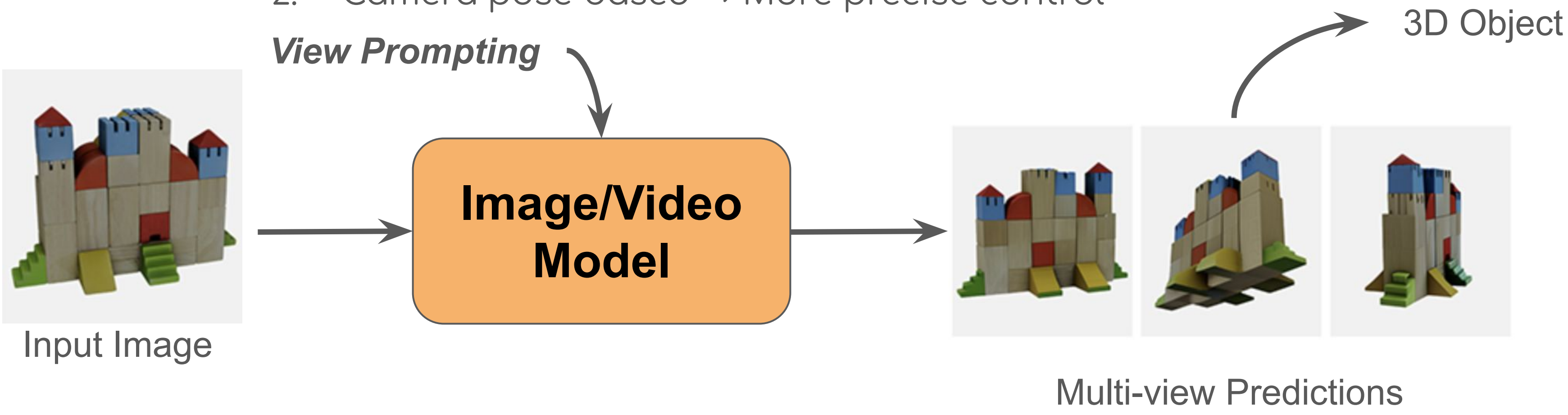
ARTIC3D



1. Yao et al.. "Hi-lassie: High-fidelity articulated shape and skeleton discovery from sparse image ensemble." *CVPR*. 2023.

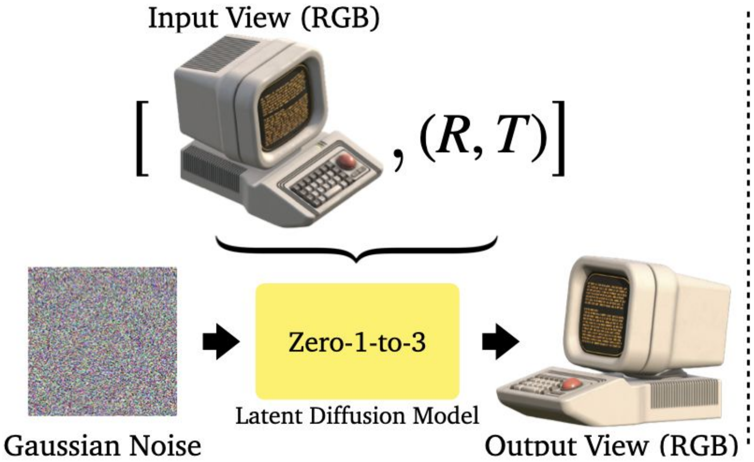
Multi-view Generation with Image/Video Models

1. Text based → Rough Control
2. Camera pose based → More precise control

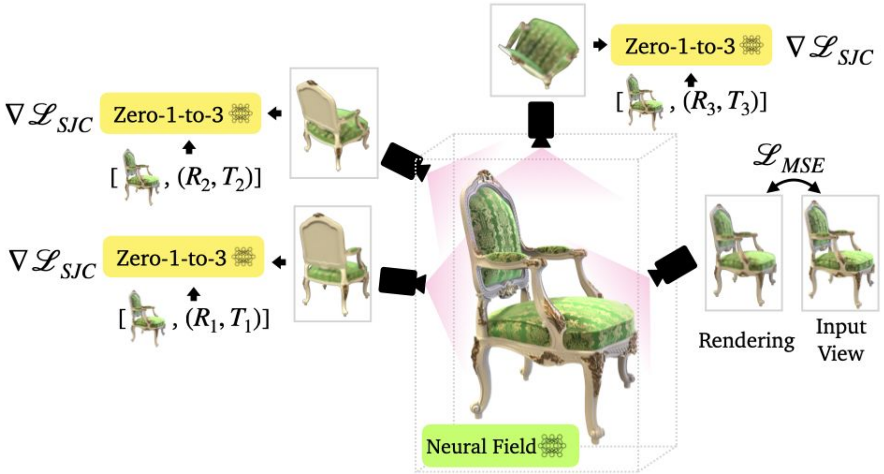


Precise View Control with Camera Conditioning

Zero123 [1], Zero123-XL [2] etc.



Novel View Synthesis



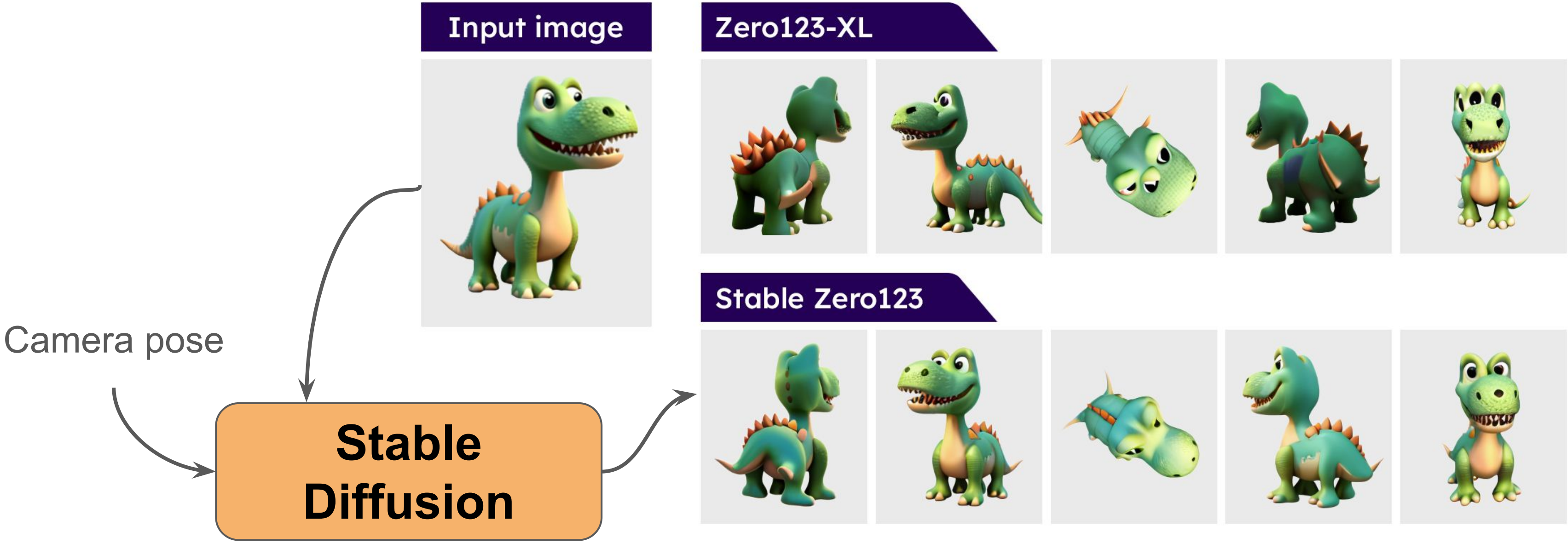
3D Reconstruction

1. Liu et al. Zero-1-to-3: Zero-shot One Image to 3D Object. ICCV 2023
2. Deitke et al. Objaverse-XL: A Universe of 10M+ 3D Objects. 2023

Stable Zero123

Improved training of Zero123

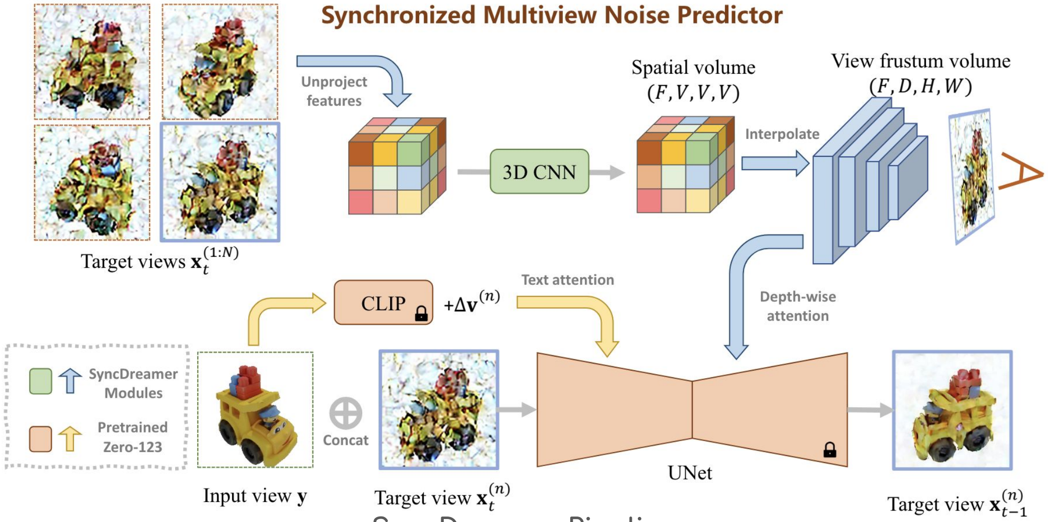
Considerably better than Zero123 and Zero123-XL



Towards Improving Multi-view Consistency

SyncDreamer [1] → Maintain 3D representation during diffusion

MVDream [2] → Always predict views at fixed camera angles

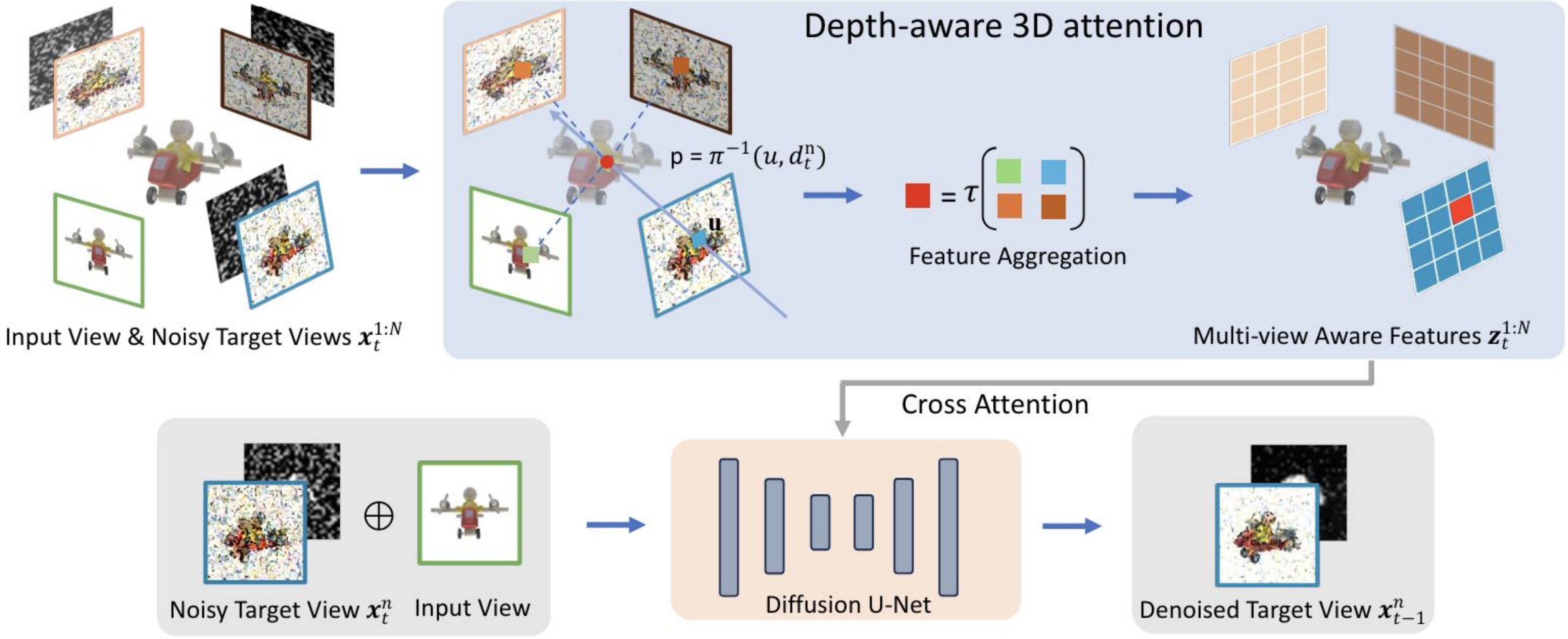


SyncDreamer Pipeline

1. Liu et al. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. ICLR 2024
2. Shi et al. MVDream: Multi-view Diffusion for 3D Generation. 2023

MVD-Fusion: Single-view 3D via Depth-consistent Multi-view Generation

Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, Shubham Tulsiani [CVPR'24]



SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, Varun Jampani

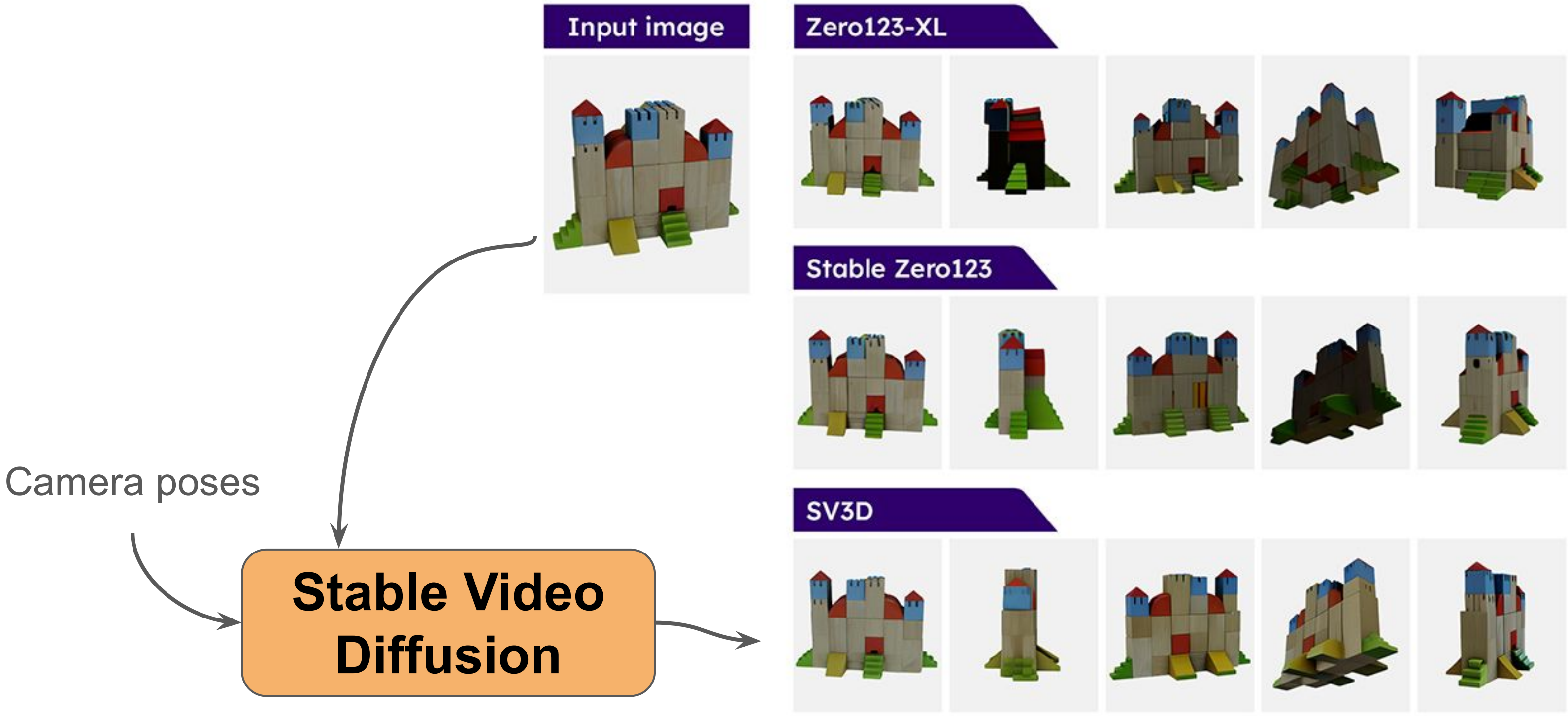
ECCV'24 oral

Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

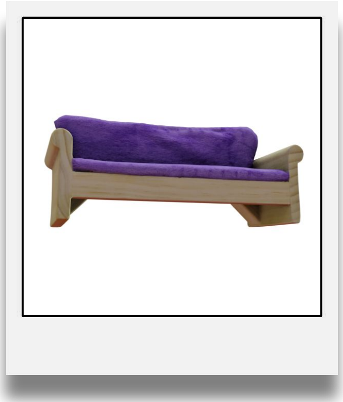
Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, Robin Rombach

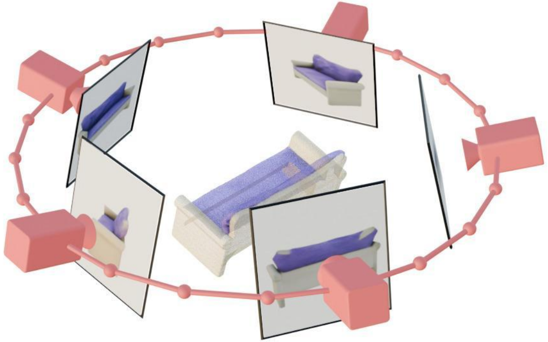
Stable Video 3D (SV3D)

Uses stable video diffusion instead of stable diffusion



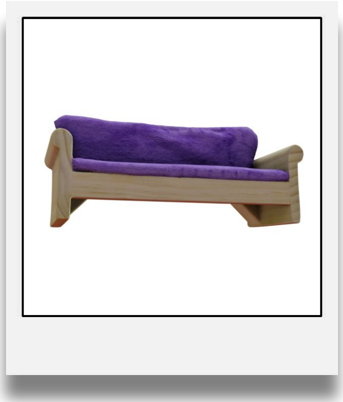
Novel Multi-view Synthesis -- Static Orbits

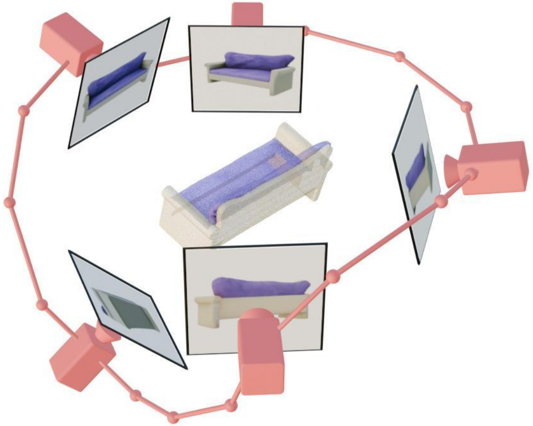




SV3D

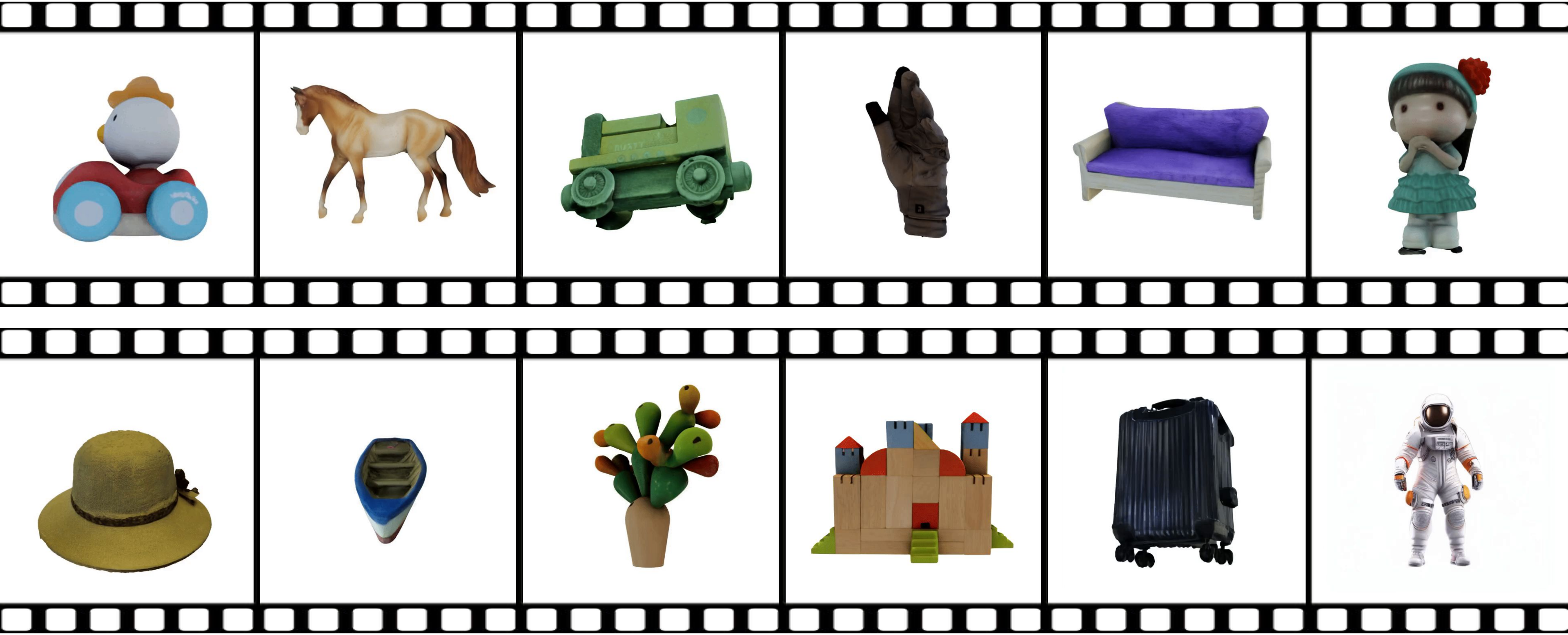
Novel Multi-view Synthesis -- Dynamic Orbits



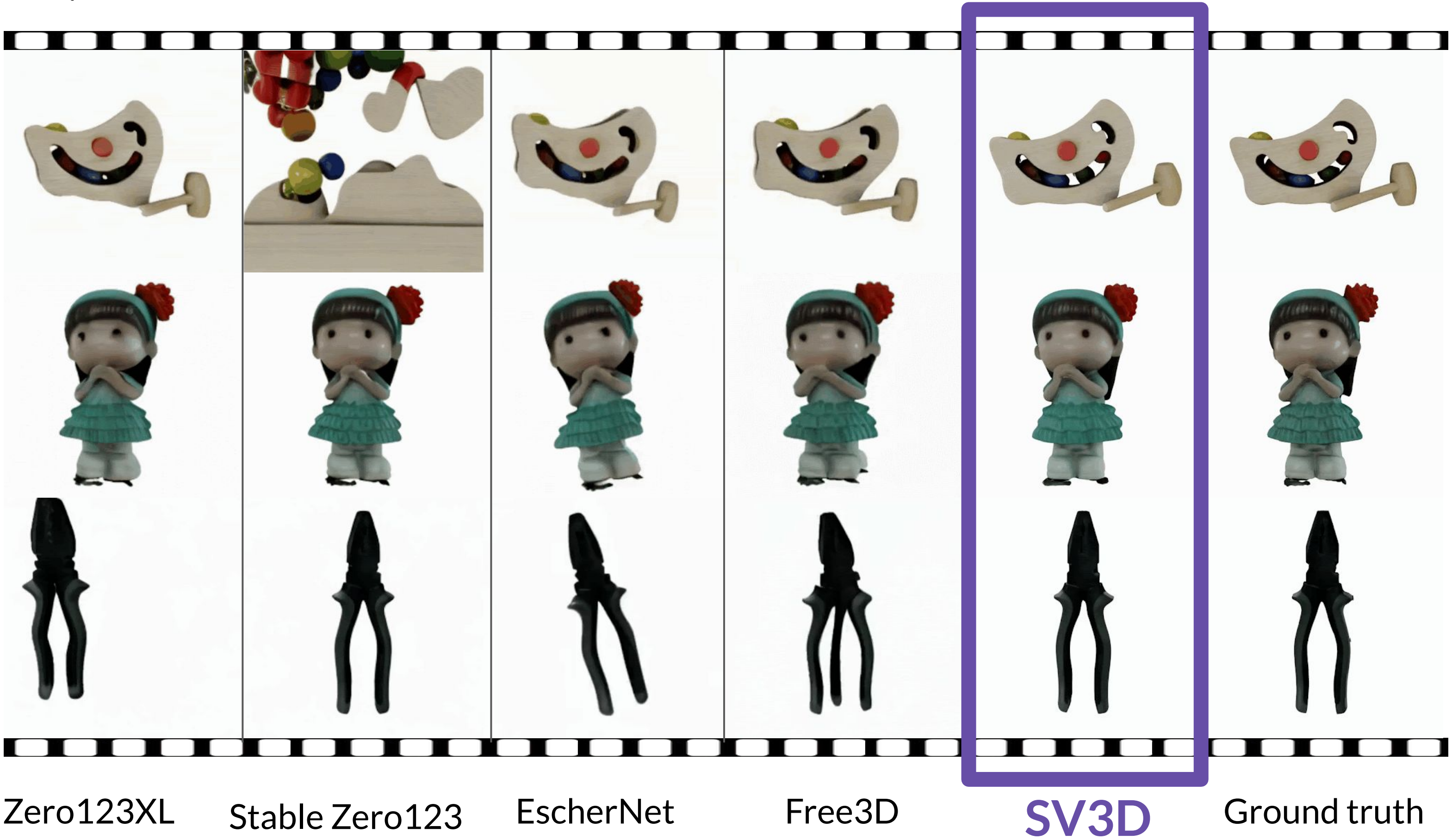


SV3D

Sample Results



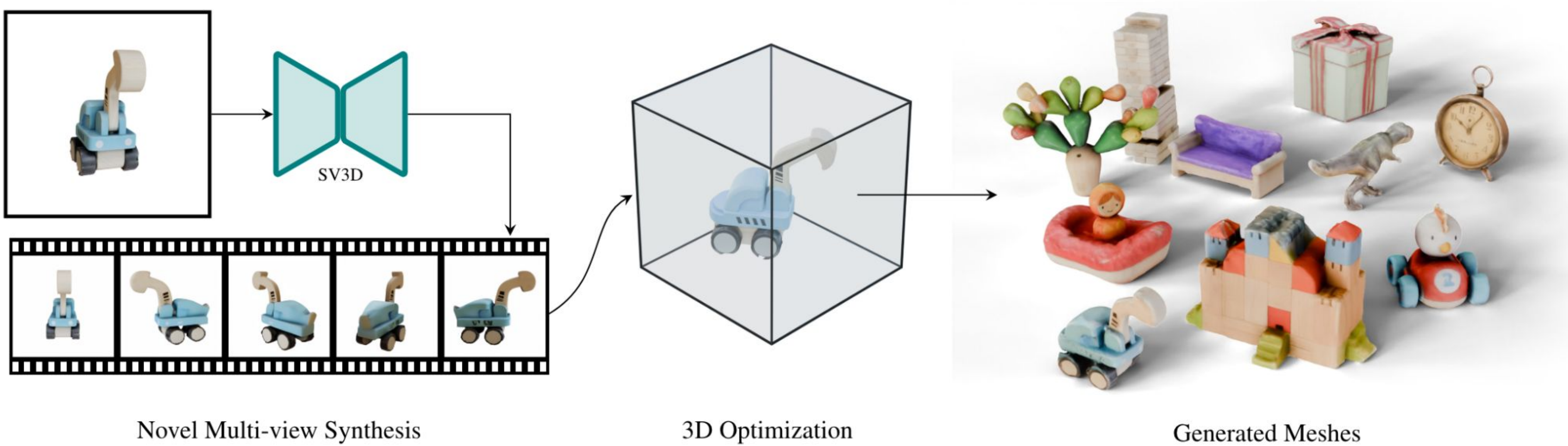
Comparisons



3D Generations using Multi-view Videos

We also propose novel techniques to get 3D objects from generated views

State-of-the-art multi-view and 3D generation results

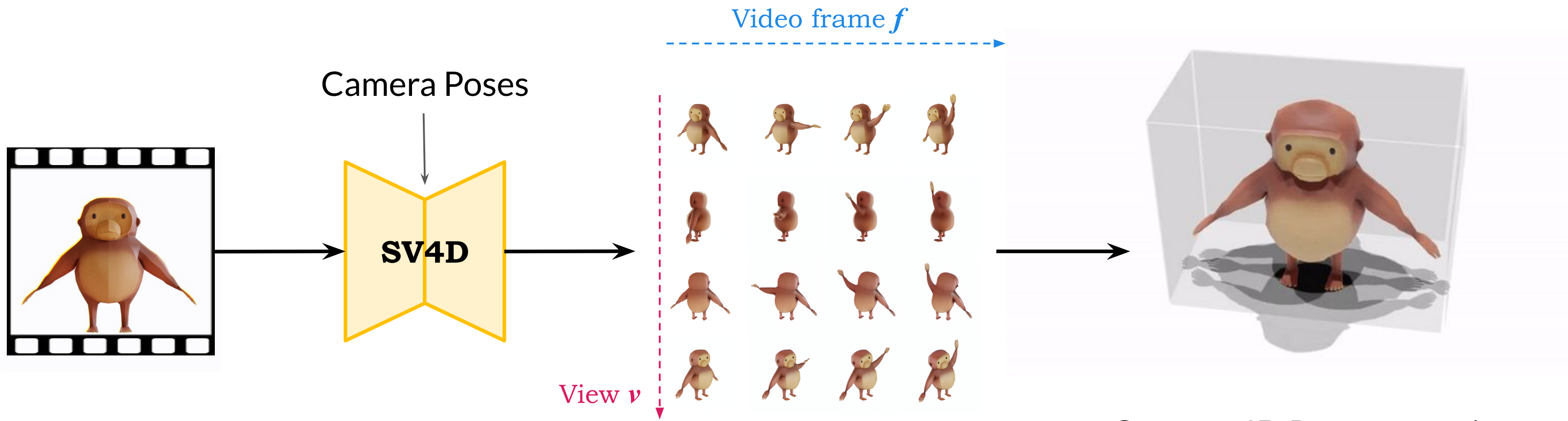


Sample 3D Generations

SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-view Consistency

Yiming Xie*, Chun-han Yao*, Vikram Voleti, Huaizu Jiang^, Varun Jampani^
(*equal contribution, ^equal advising)

SV4D - Novel-view Video Synthesis



Input: Single-view Video

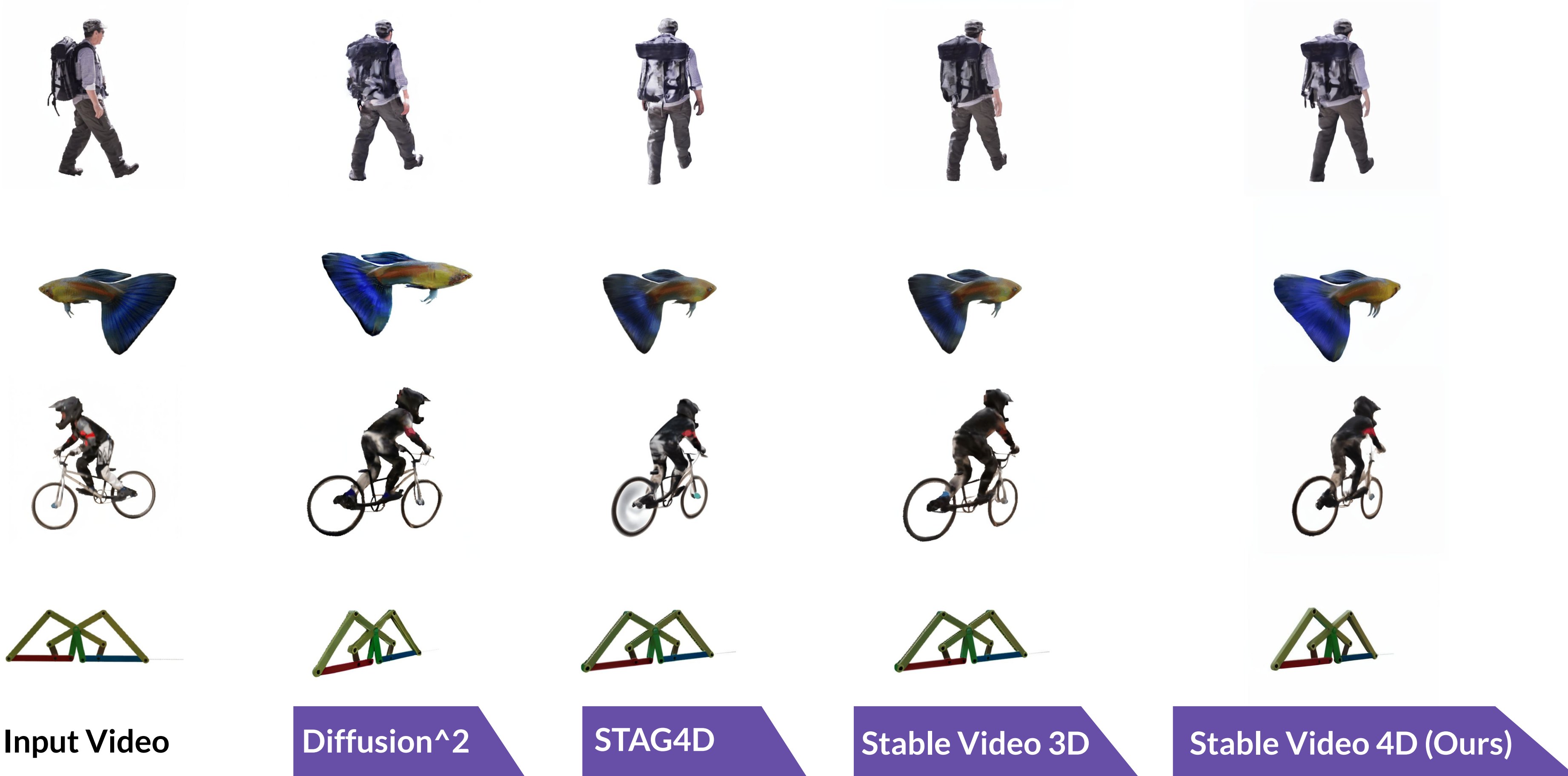
Novel View Videos

Output: 4D Representation
(Dynamic 3D representation, e.g, NeRF, Gaussian Splatting, Mesh...)

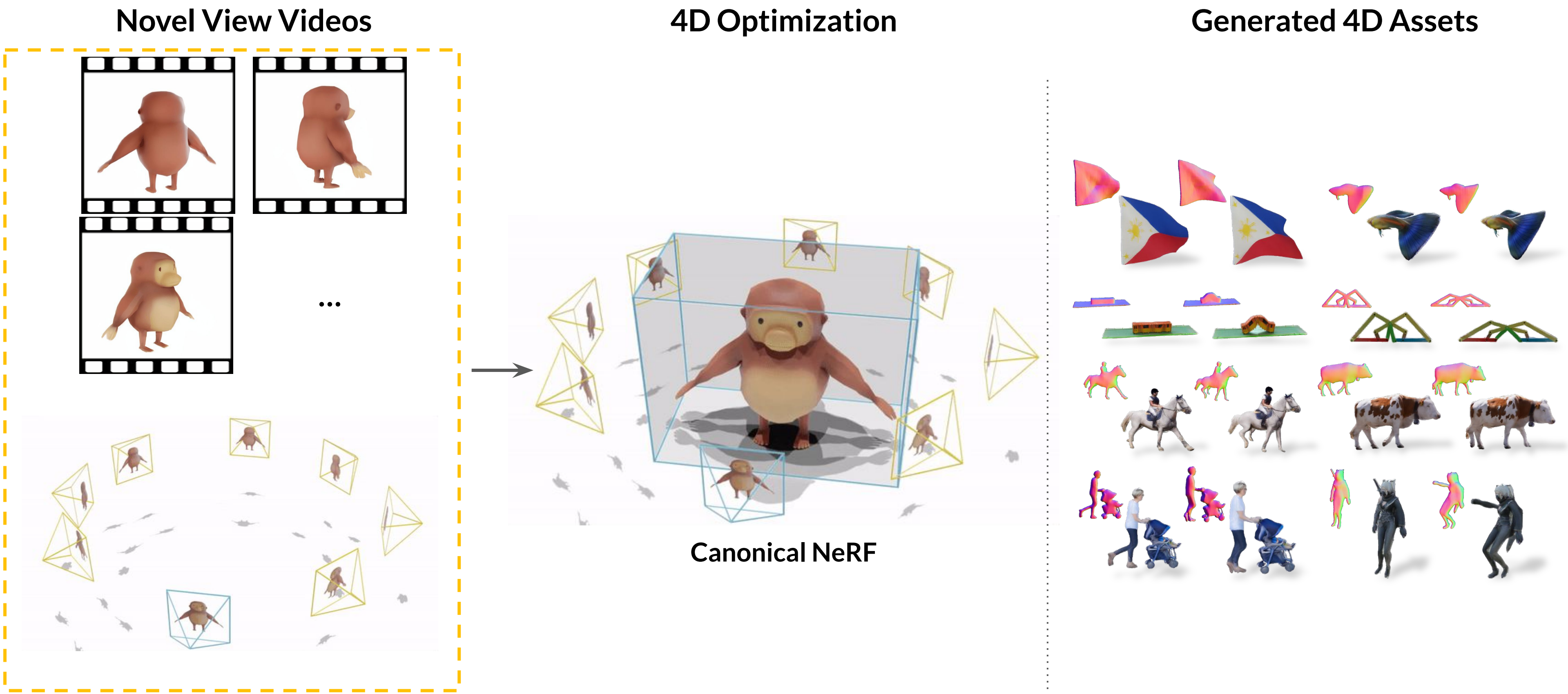
Novel View Video Synthesis

4D Optimization

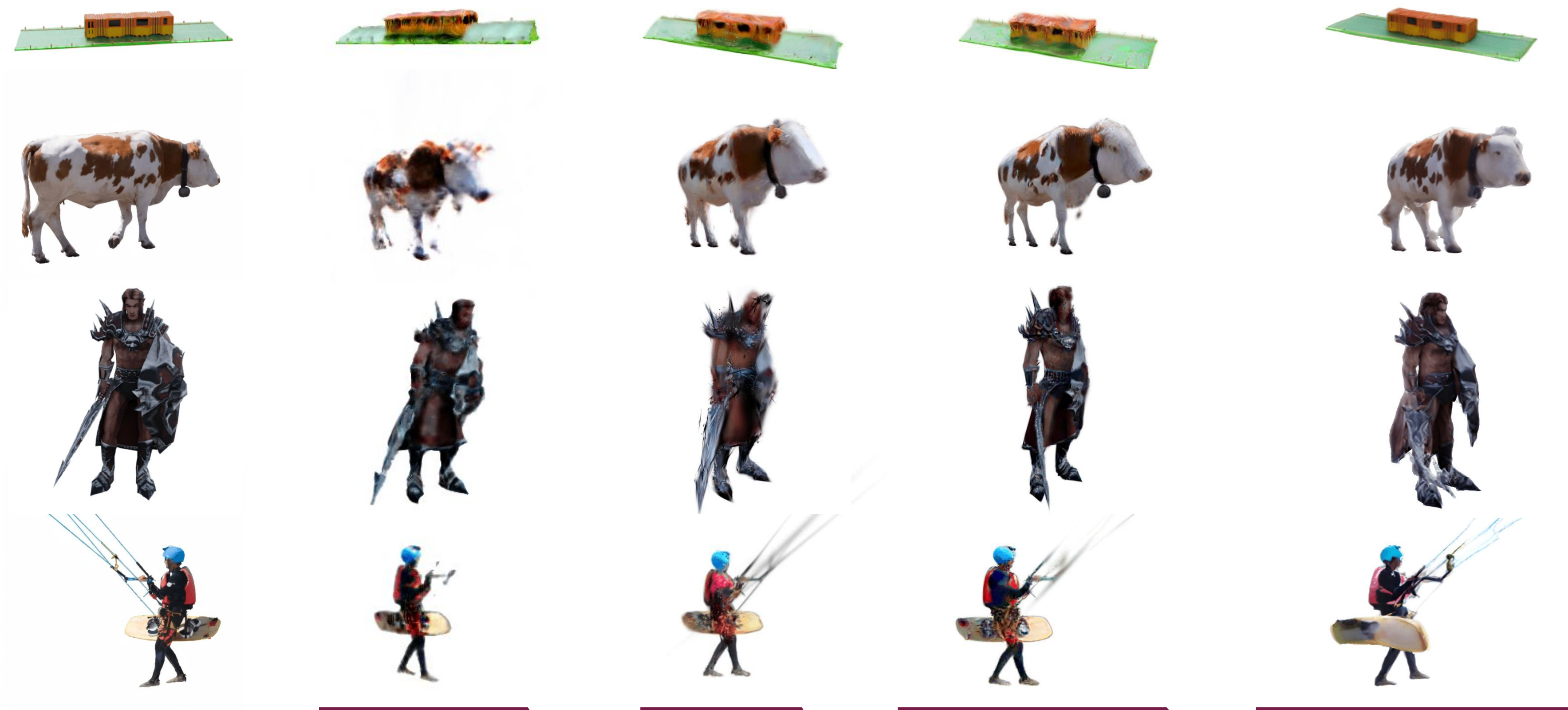
Sample Results - NVS



4D Optimization



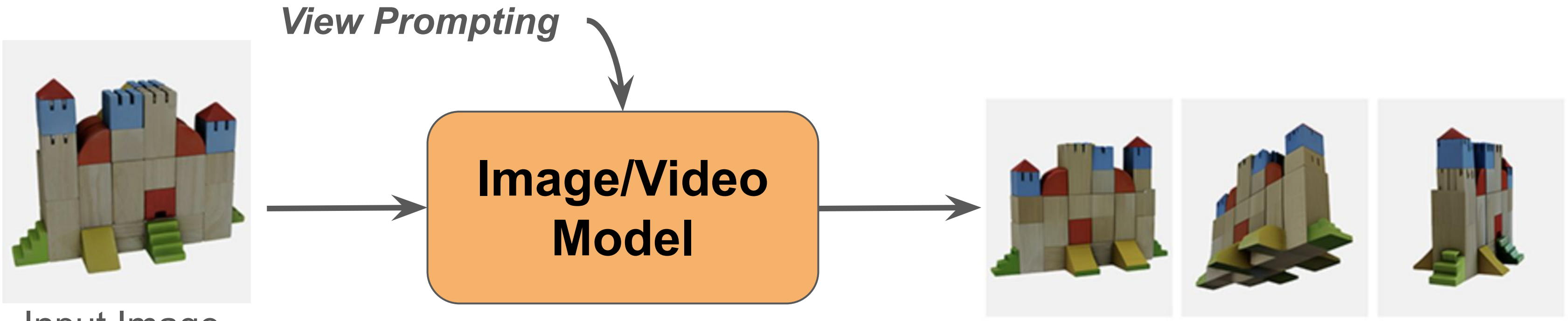
Sample Results - 4D



Input Video | Consistent4D | STAG4D | DreamGaussian4D | Stable Video 4D (Ours)

Multi-view Generation with Image/Video Models

1. Text based (**DreamFusion, ARTIC3D etc.**)
2. Camera pose based (**Stable Zero123, Stable Video 3D, Stable Video 4D etc.**)



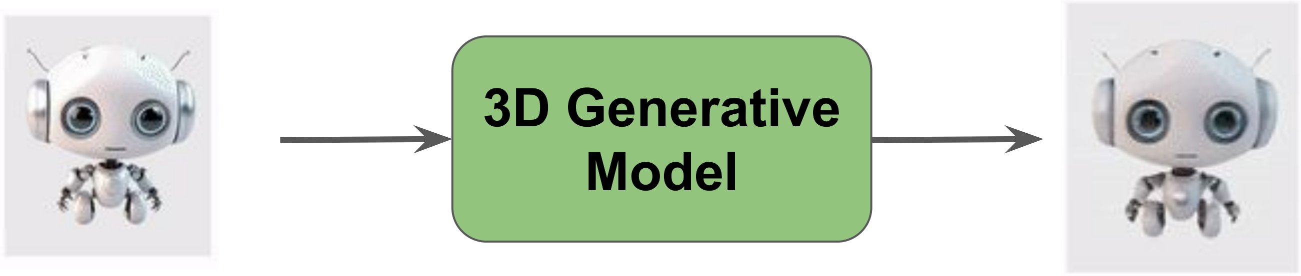
Input Image

Multi-view Predictions

Outlook
- Generalization to scenes, variable number of inputs, unknown cameras etc.
- Making these techniques faster

Direct 3D Generation

Direct 3D Generation

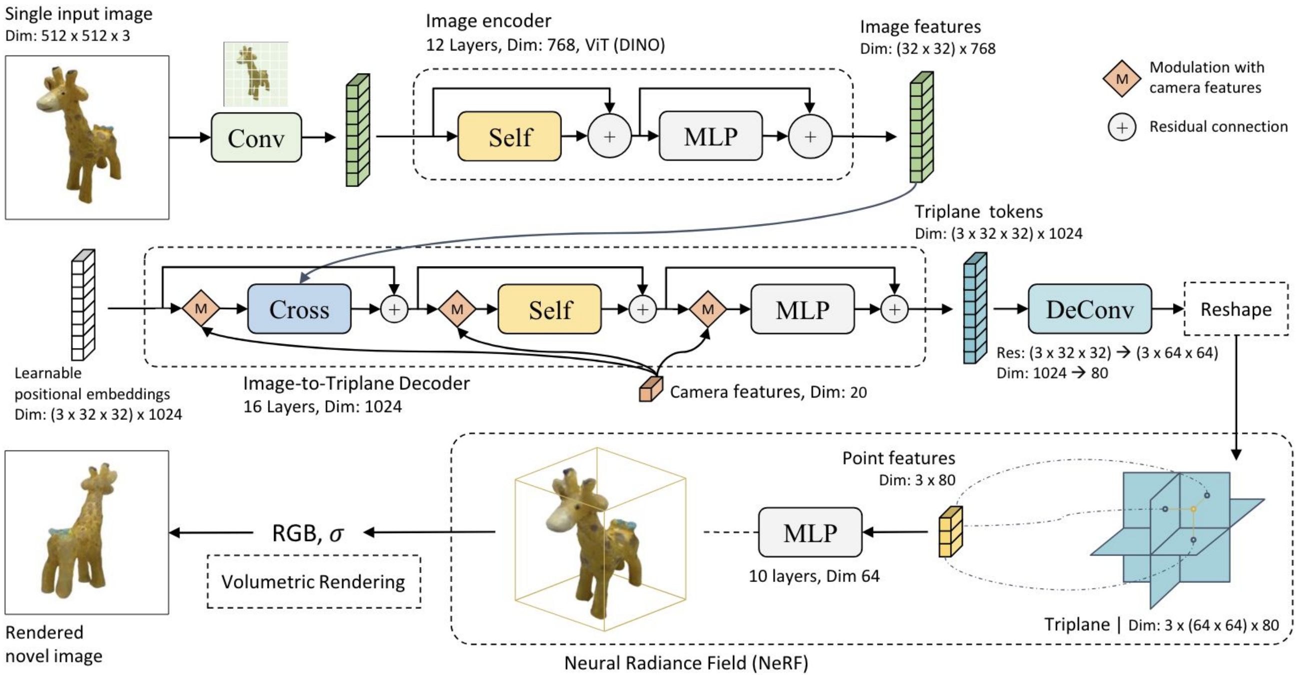


Input Image

Pros: Usually quite fast due to direct prediction

Cons: Need good amount of 3D datasets to train and generalize

LRM: Large Reconstruction Model



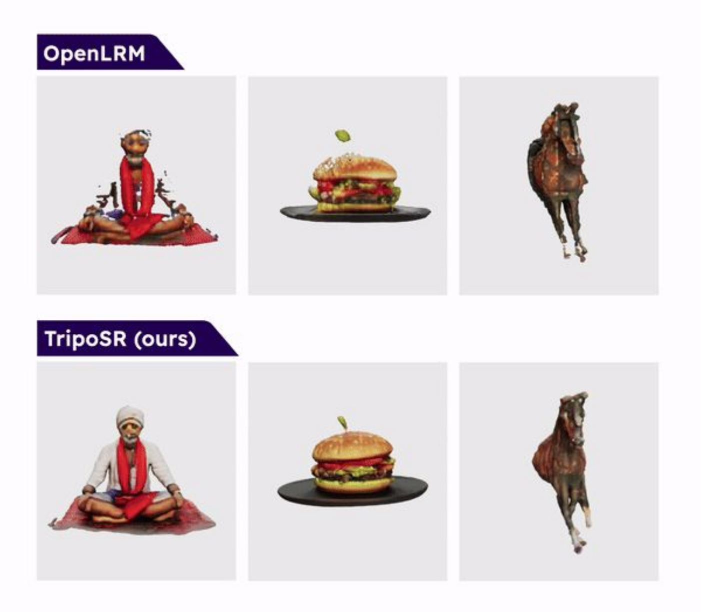
1. Hong et al. LRM: Large Reconstruction Model for Single Image to 3D. ICLR 2024

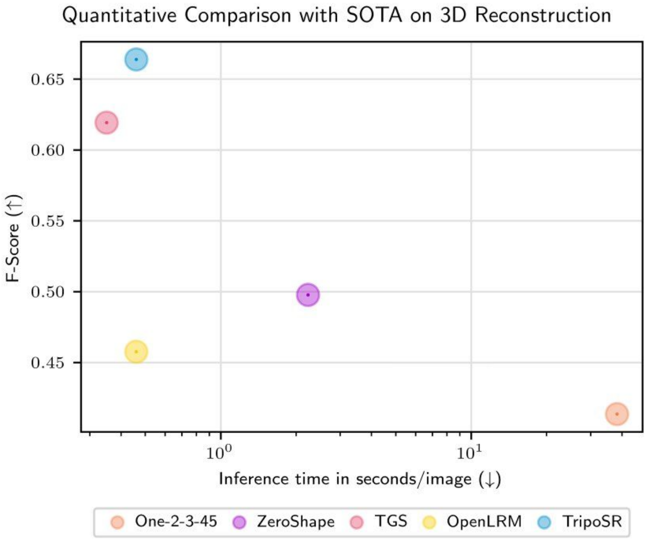
TripoSR: Fast 3D Object Reconstruction from a Single Image

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani*, Yan-Pei Cao*

3D mesh prediction from a single image in <0.5 seconds

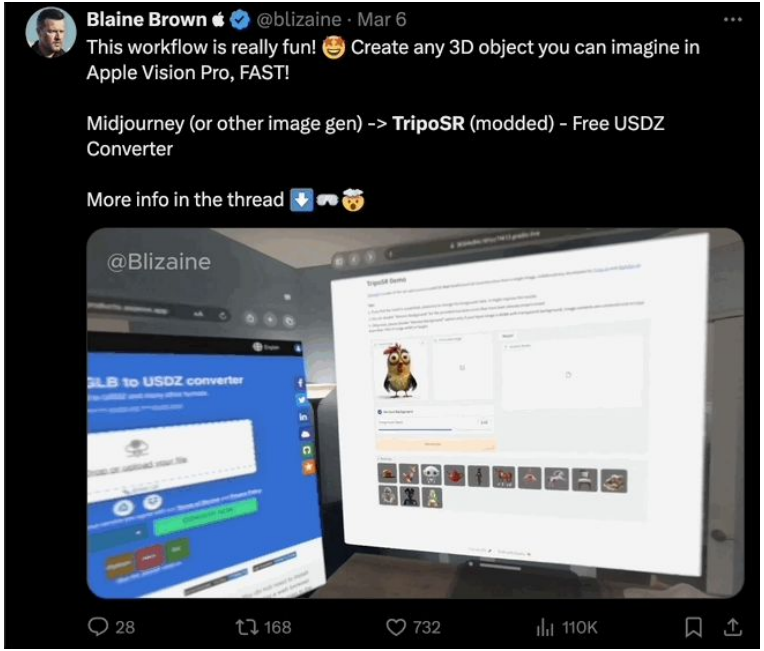
One of the best and fastest 3D generative models among open-source

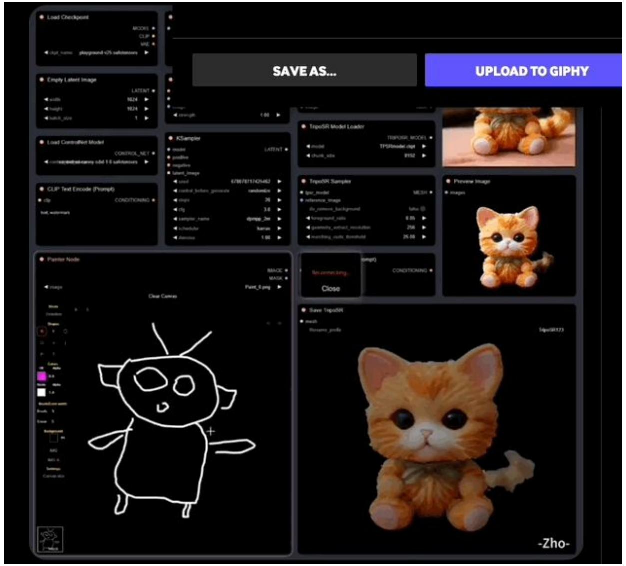




Quick adoption of TripoSR in the community

Several interesting use cases and workflows





SF3D: Stable Fast 3D Mesh Reconstruction with UV-Unwrapping and Illumination Disentanglement

Mark Boss, Zixuan Huang, Aaryaman Vasishta, Varun Jampani

Single Image to Relightable Object

SF3D

Improvements with SF3D

- Illumination disentanglement

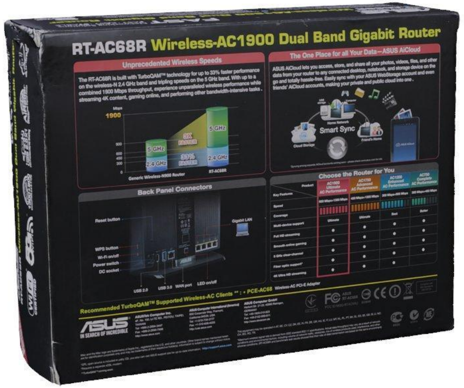
Ground Truth

TripoSR

Ours (SF3D)

Improvements with SF3D

- Illumination disentanglement
- Sharper textures with UV maps (not vertex colors)



Ground Truth

TripoSR

Ours (SF3D)

Improvements with SF3D

- Illumination disentanglement
- Sharper textures with UV maps (not vertex colors)
- Reduce Marching cube artifacts with vertex displacements

Ground Truth

TripoSR

Ours (SF3D)

Improvements with SF3D

- Illumination disentanglement
- Sharper textures with UV maps (not vertex colors)
- Reduce Marching cube artifacts with vertex displacements
- Material properties

Ground Truth

TripoSR

Ground Truth

TripoSR

Improvements with SF3D

- Illumination disentanglement
- Sharper textures with UV maps (not vertex colors)
- Reduce Marching cube artifacts with vertex displacements
- Material properties

Ground Truth

Ours

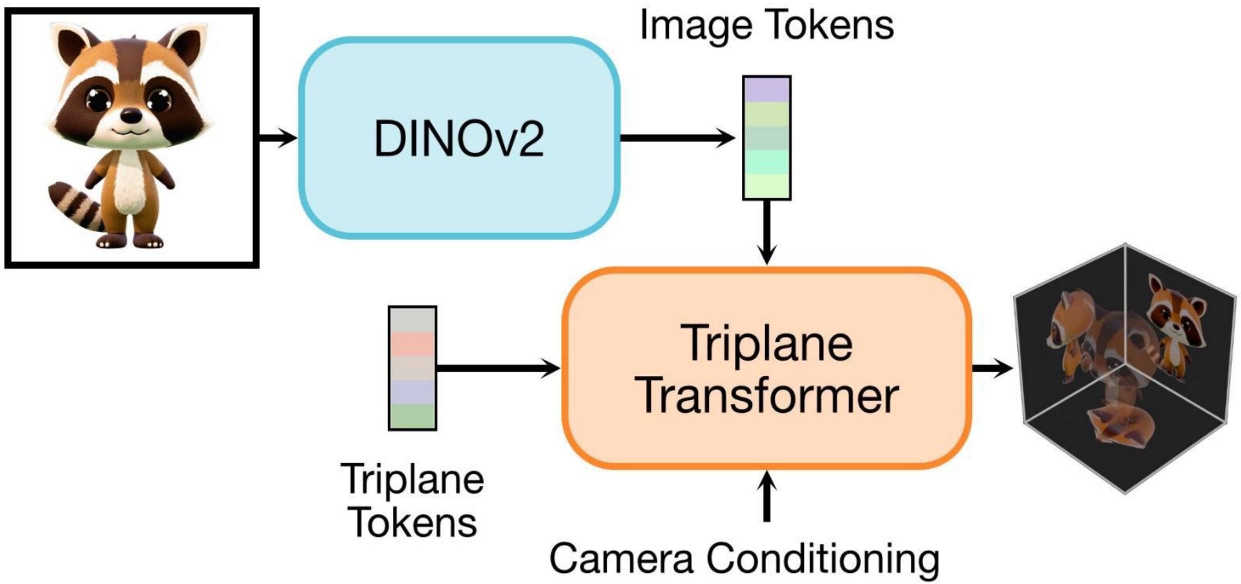
Ground Truth

Ours

SF3D Approach

SF3D Approach

SF3D Approach



Higher resolution triplanes with enhanced transformer

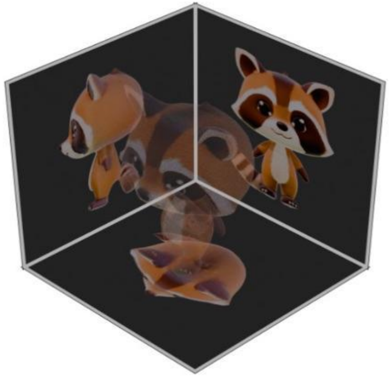
- Previous methods used a low resolution triplane (64 x 64) resulting in grid artifacts and aliasing issues aliasing issues
- We predict high resolution 384 x 384 triplanes with an enhanced transformer

Ground Truth

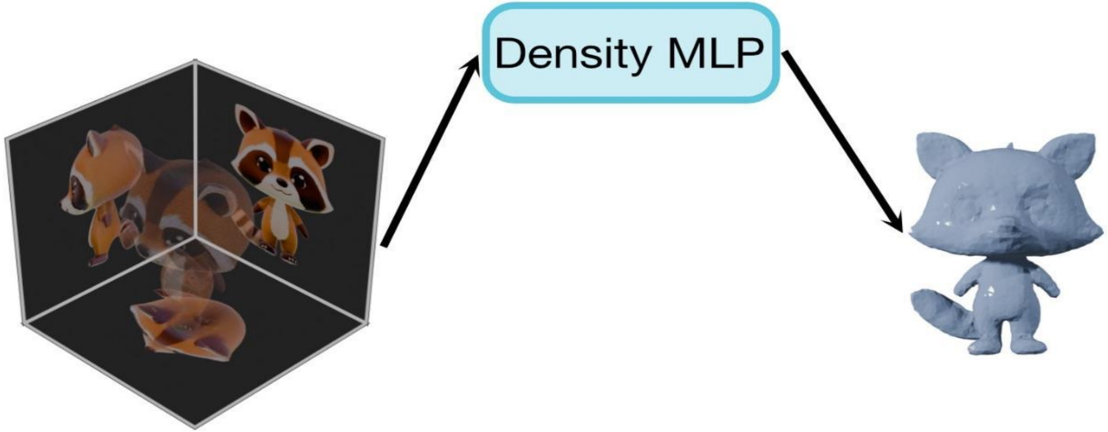
Low Resolution

Ours
(High Resolution)

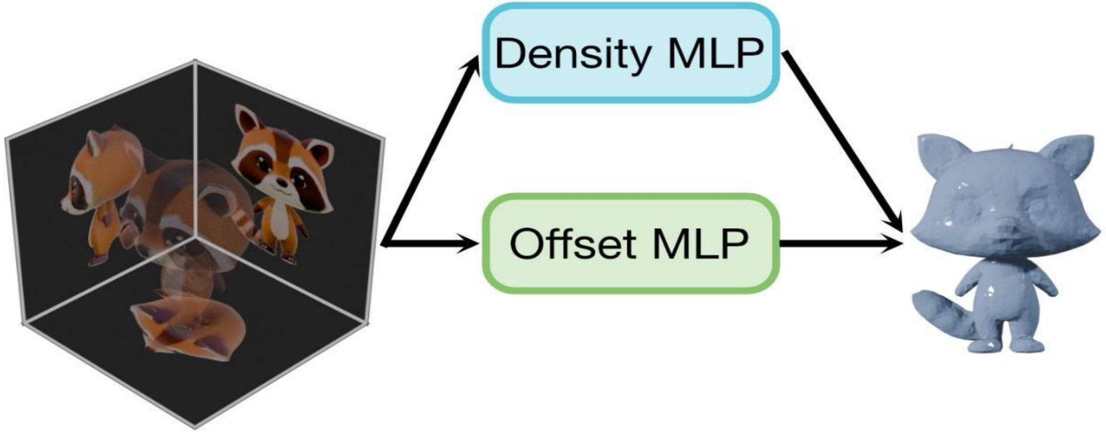
SF3D Approach



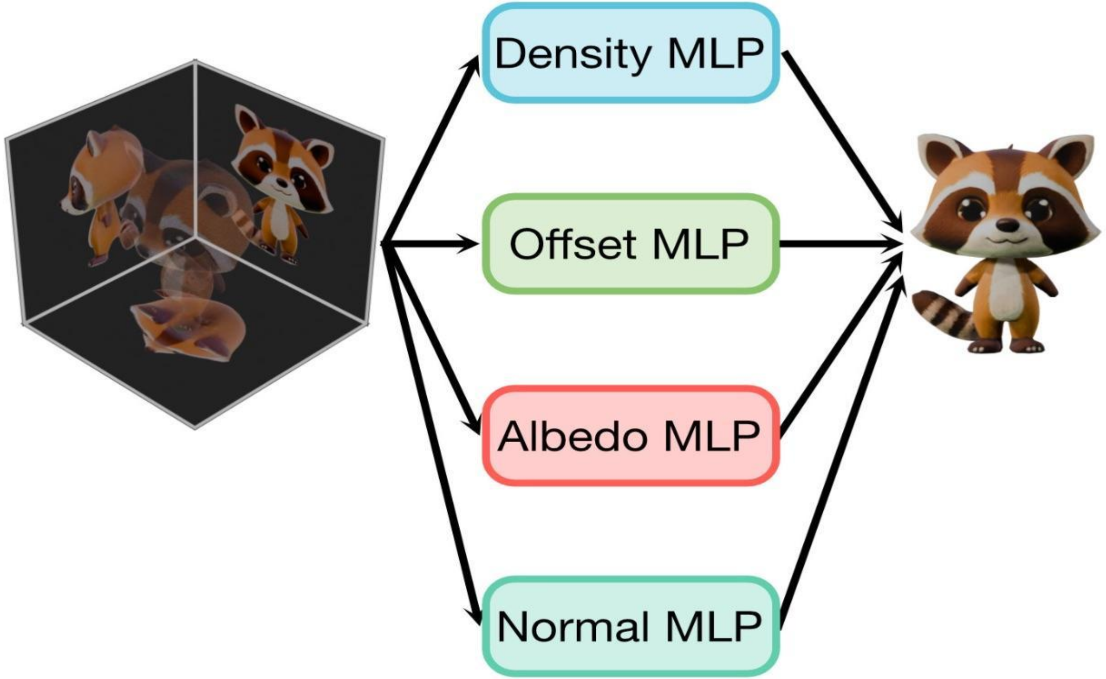
SF3D Approach



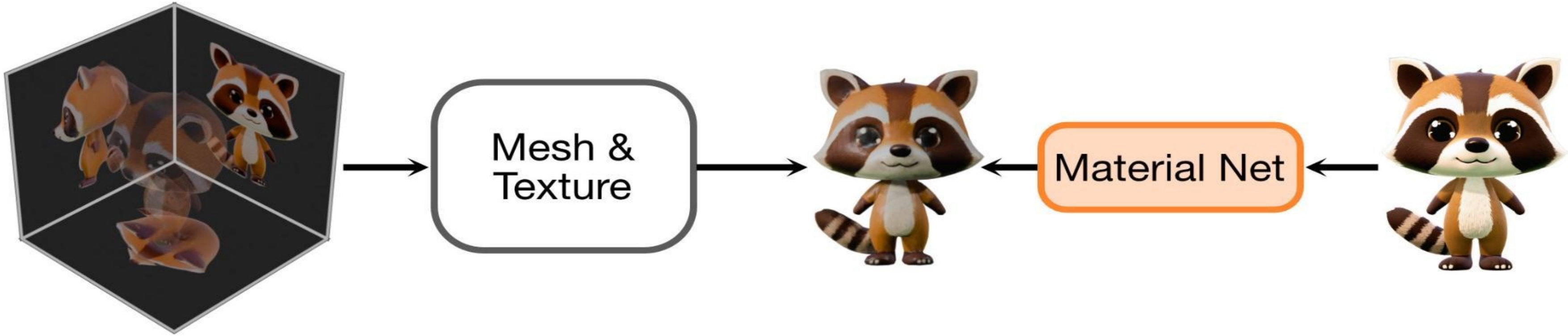
SF3D Approach



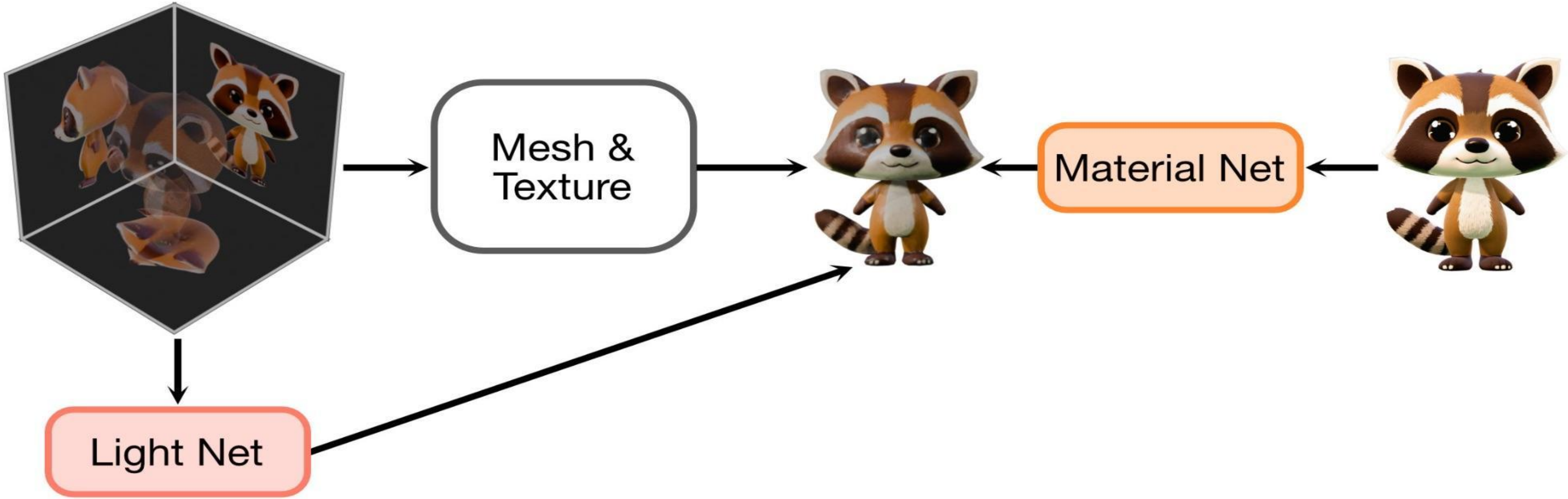
SF3D Approach



SF3D Approach



SF3D Approach



Sample Results

Sample Comparisons

TripoSR

InstantMesh

CRM

Ours (SF3D)

Sample Comparisons

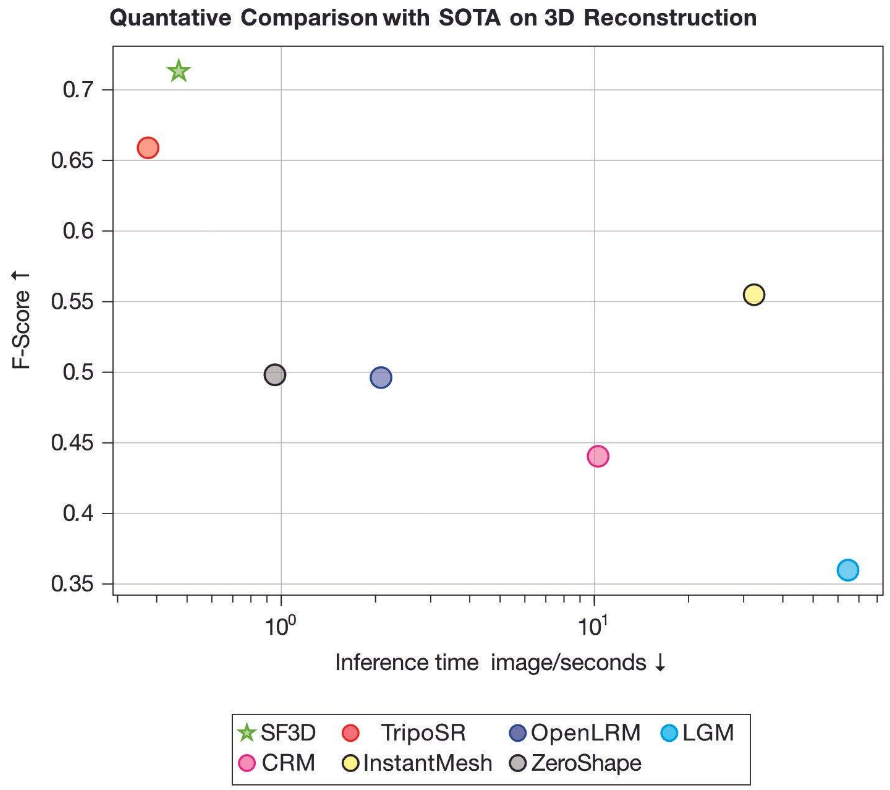
TripoSR

InstantMesh

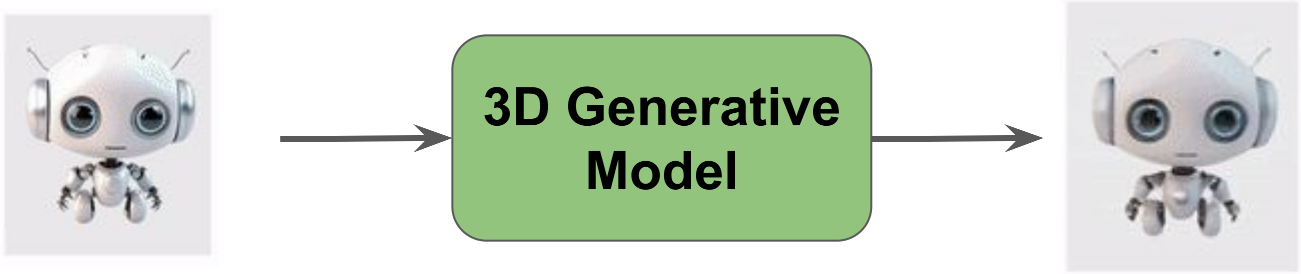
CRM

Ours (SF3D)

Fast (<0.5 seconds) but accurate



Direct 3D Generation - Remarks



Input Image

→ **TripoSR, Stable Fast 3D etc.**

Fast, but requires good amount of 3D datasets

Outlook

- Generalization to scene generation as well as dynamic 3D generation

Concluding Remarks

Two emerging technologies in Generative 3D

- Direct 3D generation → **Fast** but needs lots of 3D data
- Multi-view generation → **Slow** but can generalize well

Outlook

- Combining the strengths of both results in **fast and generalizable** networks
  - Speed of direct prediction approaches
  - Generalization of multi-view generation networks

Thank You

Comments and suggestions are most welcome

varunjampani@gmail.com
varunjampani.github.io