Digital Humans – Winter 24/25

Lecture 13_3 – Diffusion Models in 3D Reconstruction

Prof. Dr. Gerard Pons-Moll

University of Tübingen / MPI-Informatics

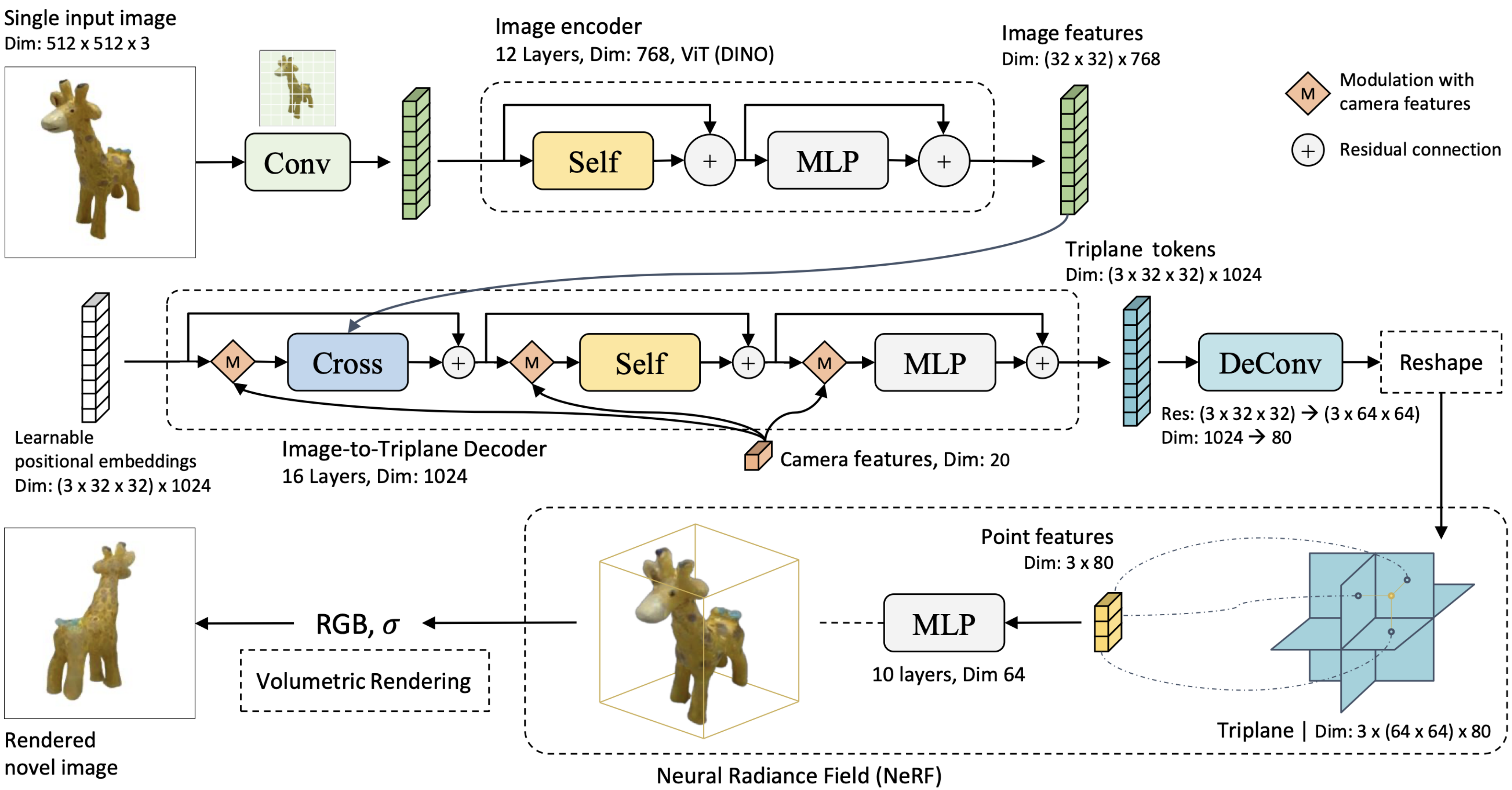
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Main contents

- **3D Reconstruction from single Image**
- 2D Diffusion Model for Novel-view Synthesis
  - Novel-view Diffusion Models
  - Multi-view Image Diffusion Models
- Sync 2D Diffusion & 3D Reconstruction

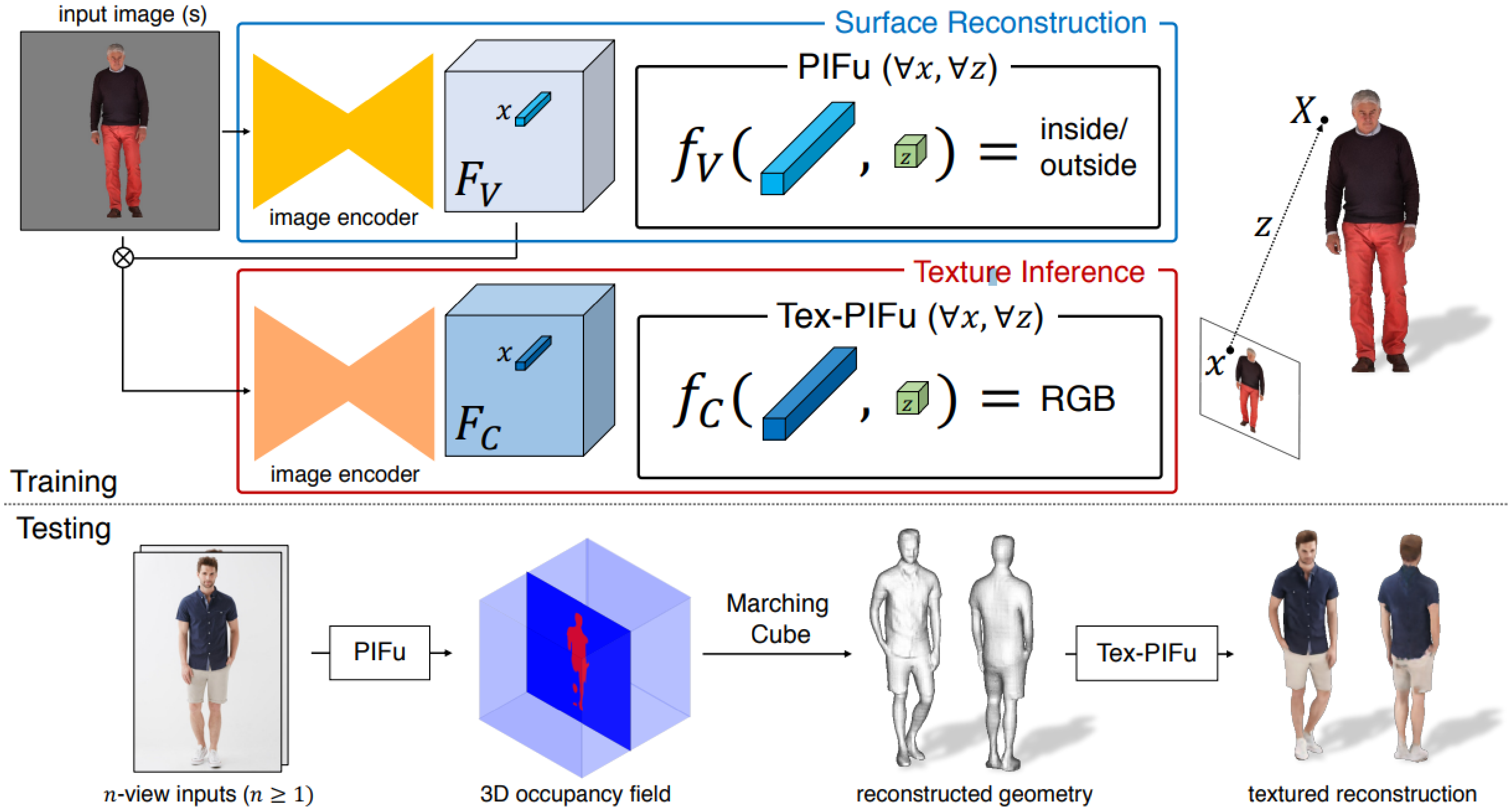
Reconstruct 3D from Single Image

- LRM: Regress NeRF tri-plane features from a RGB image



Reconstruct 3D from Single Image

- PiFU: regress occlusion field from a RGB image.



Reconstruct 3D from Single Image

- Limitation: no generative power, blurry occluded region.

Input Image

OpenLRM, He et al.

Reconstruct 3D from Single Image

- Limitation: no generative power, blurry occluded region.

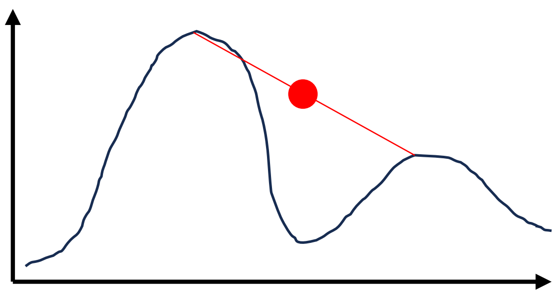
Input Image

SiTH (CVPR 2024)

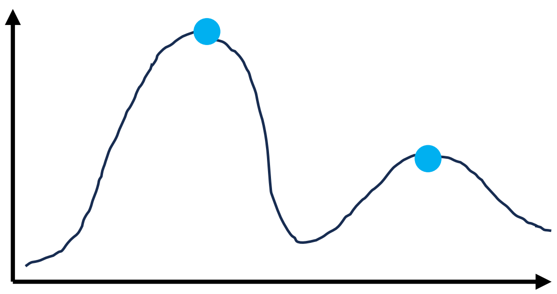
SiFU (CVPR 2024)

Motivation for generative model

- Goal: single view reconstruction is ill-posed.
- Deterministic model might collapse to average value.



Deterministic: learn an average



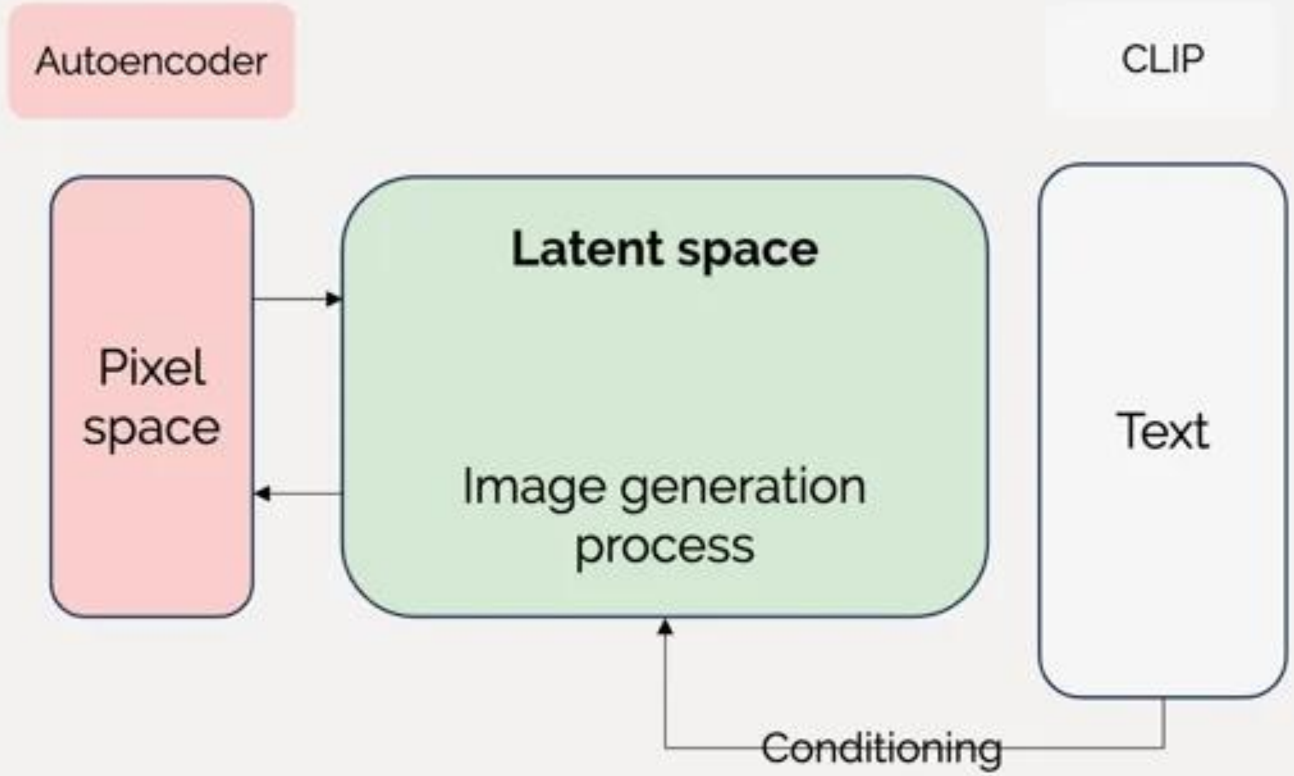
Generative model: learn a distribution

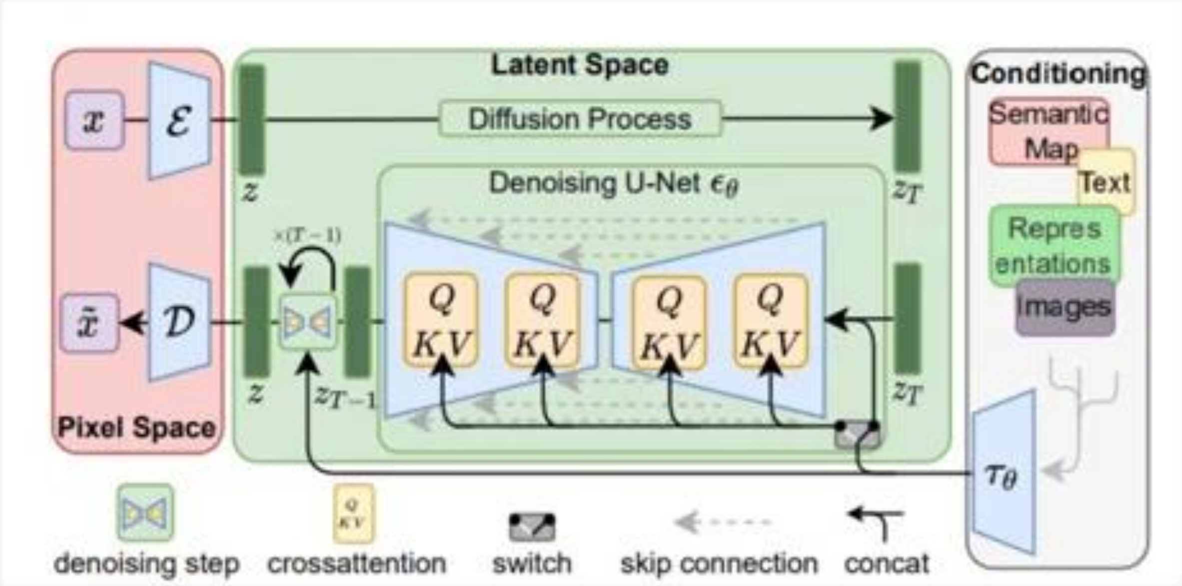
- We should learn a distribution of all possible configurations instead of simply regression.
- Diffusion model for conditional generation!

Main contents

- 3D Reconstruction from single Image
- **2D Diffusion Model for Novel-view Synthesis**
  - Novel-view Diffusion Models
  - Multi-view Image Diffusion Models
- Sync 2D Diffusion & 3D Reconstruction

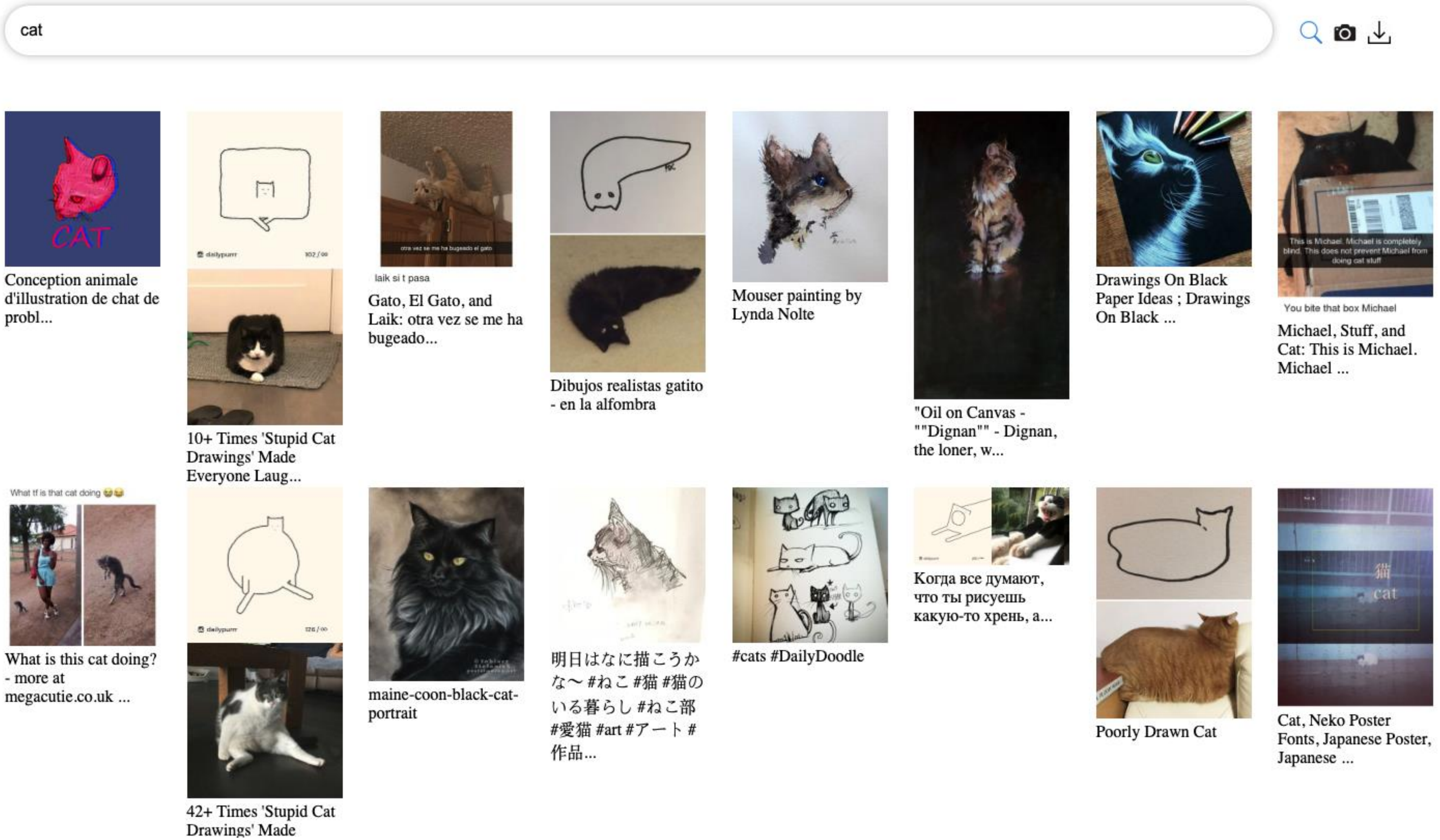
Image Diffusion Model





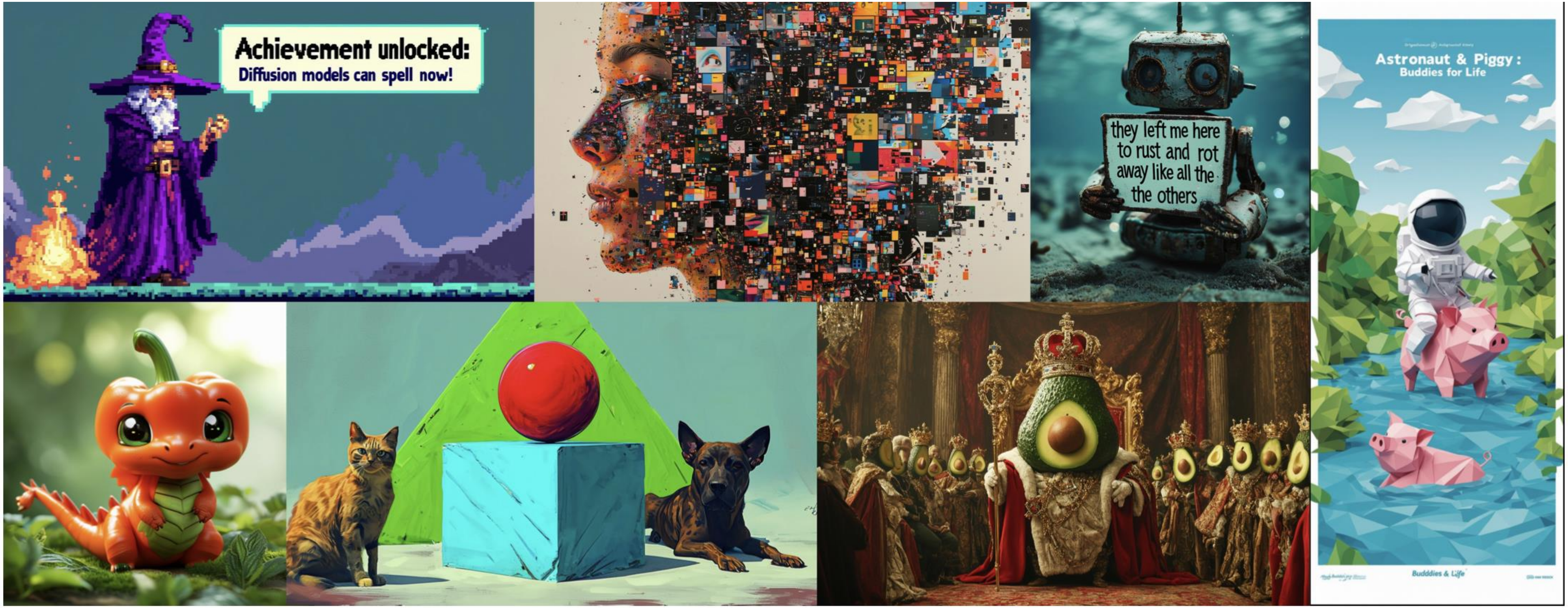
High-Resolution Image Synthesis with Latent Diffusion Models, Rombach et al, 2022

Large-scale Training



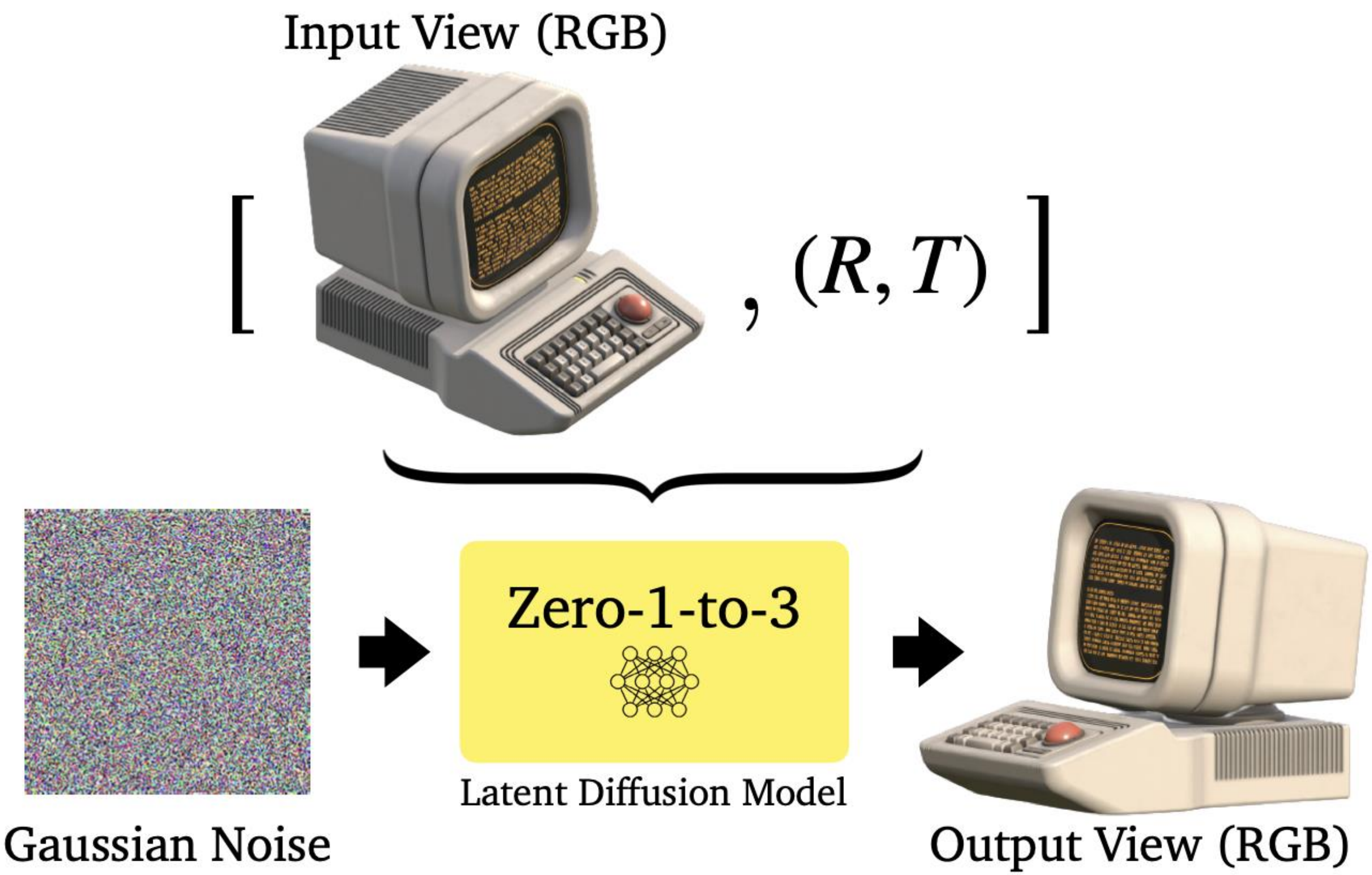
Laion5B:
- **5 Billion images**
- **With Text annotations**

Superior image generation quality



Novel-view Diffusion Model

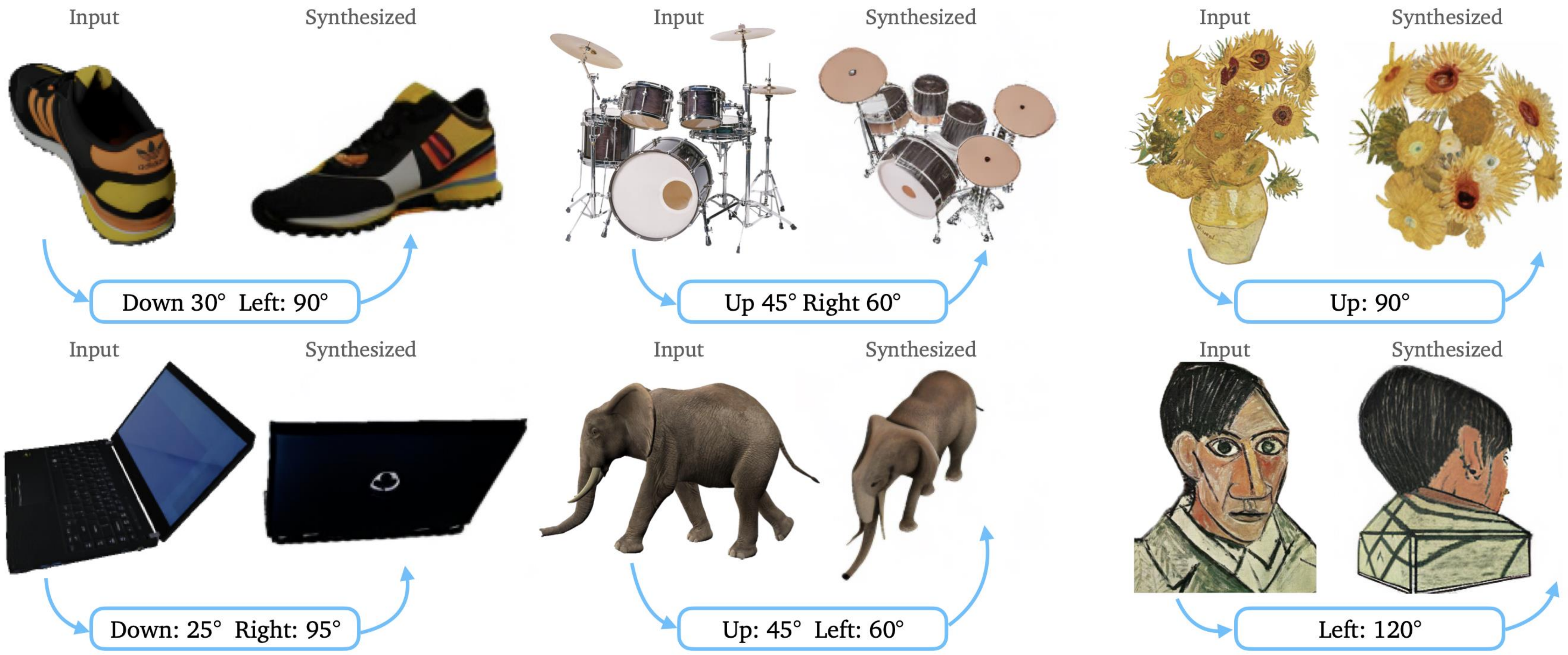
- Leverage Image diffusion prior, generate desired novel view image



Large-scae Training

Objaverse:
- **800K 3D objects**

Novel-view Diffusion Model



Novel-view Diffusion Model

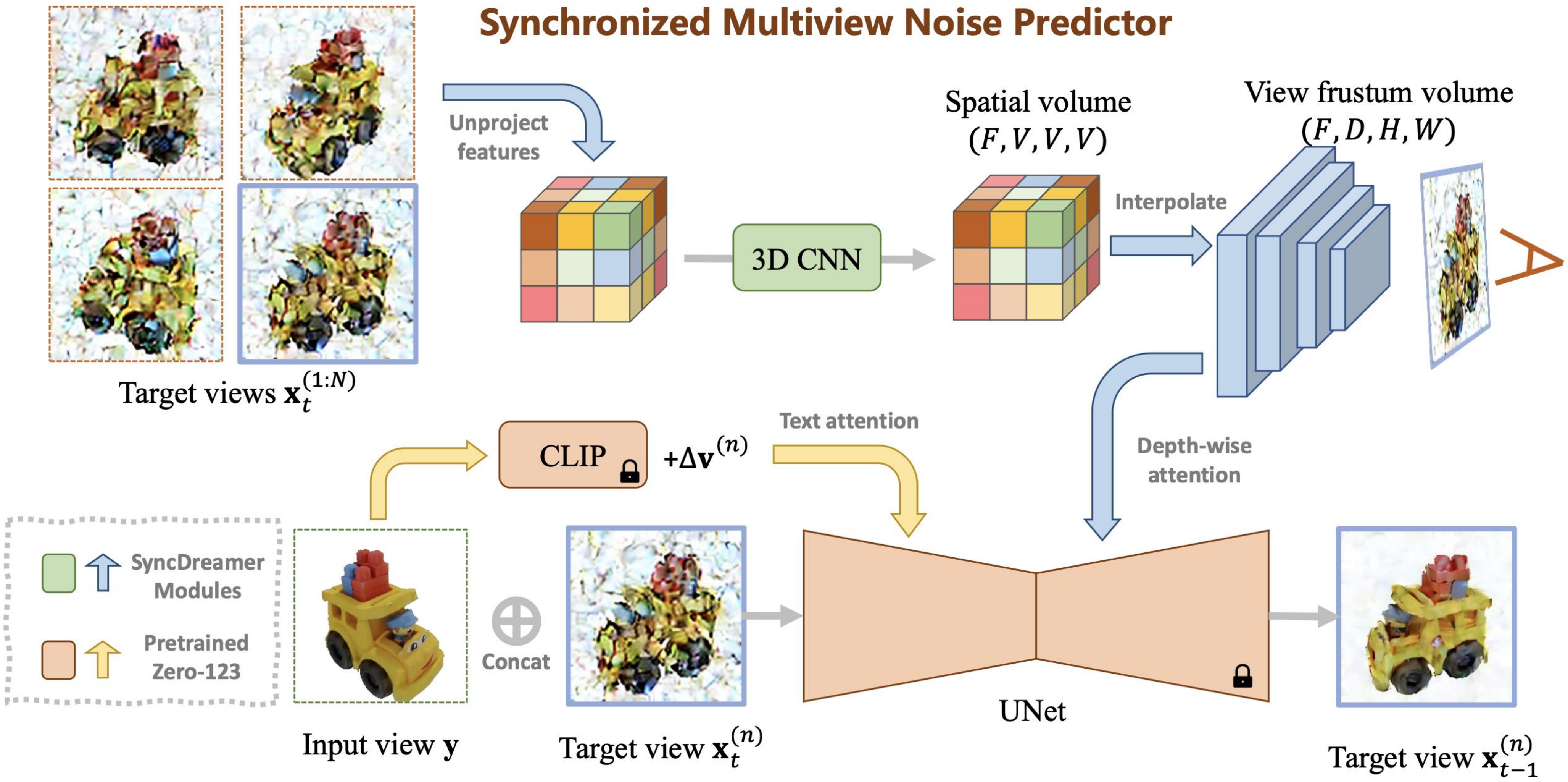
- Limitation: each view generated individually, inconsistent across generation

Input Image

Zero123-XL, Liu et al.

Multi-view Diffusion Model

- Generate Multiple Views simultaneously, have more consistency



Multi-view Diffusion Model

- Compared to Single-view Diffusion, the multi-view diffusion models are more consistent across generated novel views

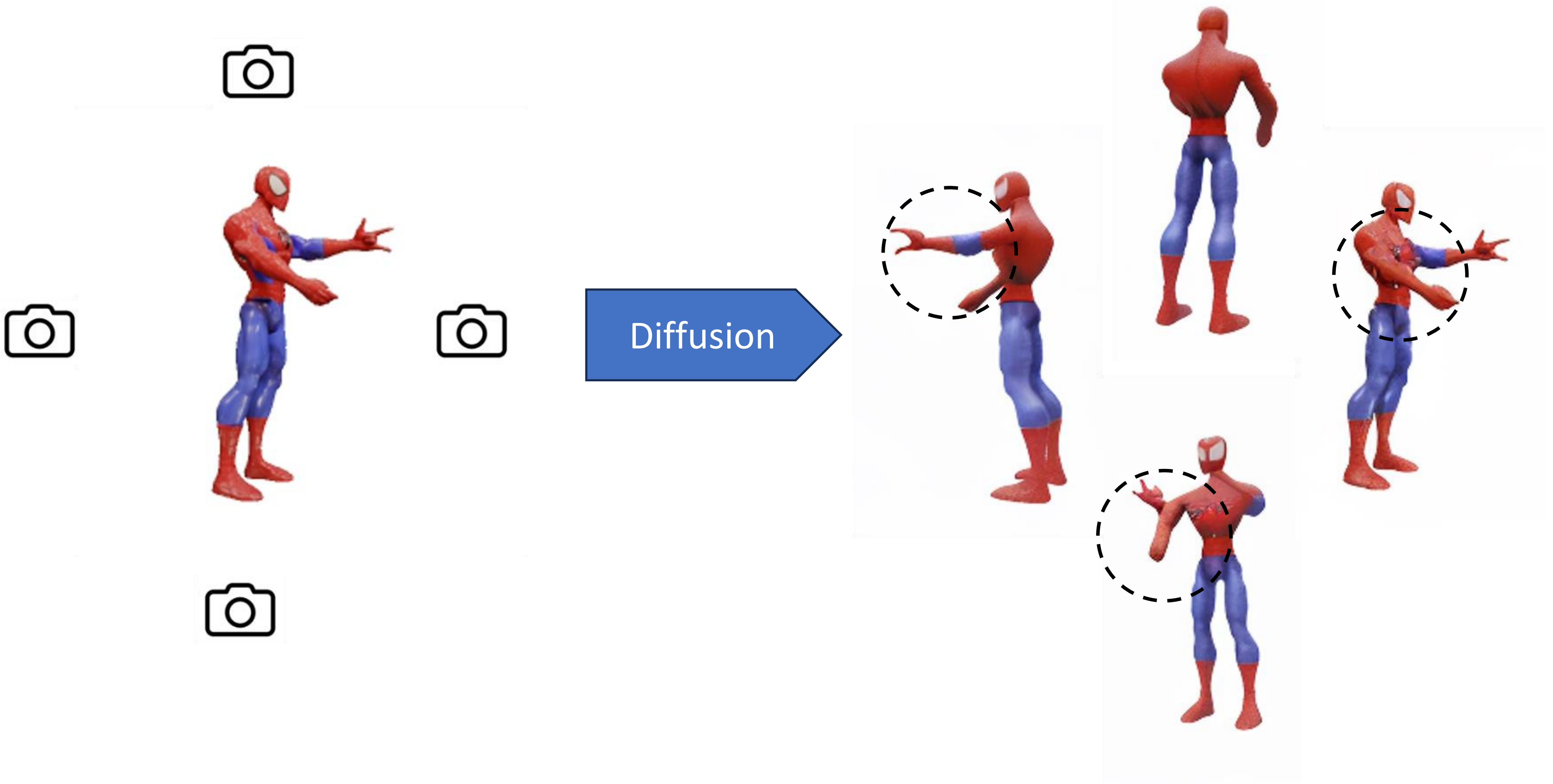
Input Image

Single-view Diffusion

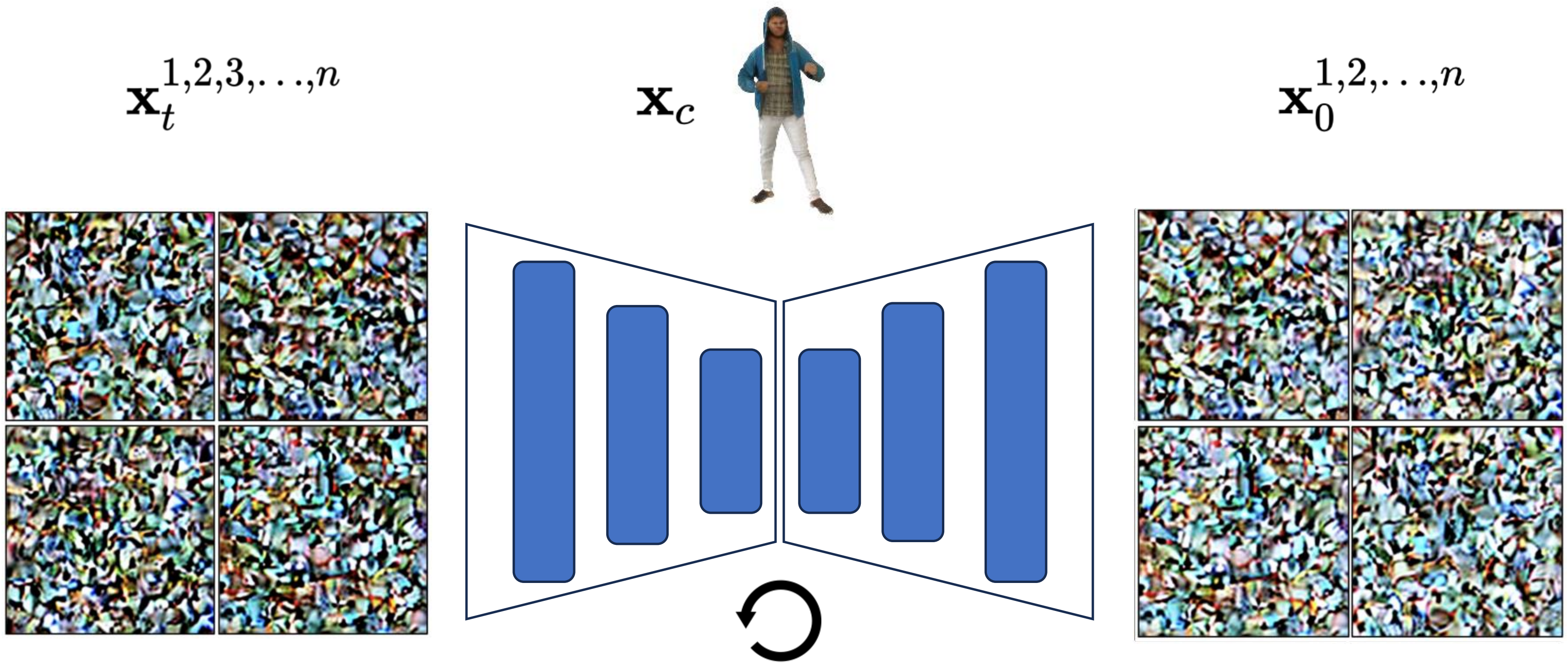
Multi-view Diffusion

3D Consistency in 2D Multi-view Diffusion Model

- 2D Multi-view Diffusion has no explicit 3D representation (e.g. NeRF or 3D-GS). Thus, the 3D consistency of the generated images are not constrained.

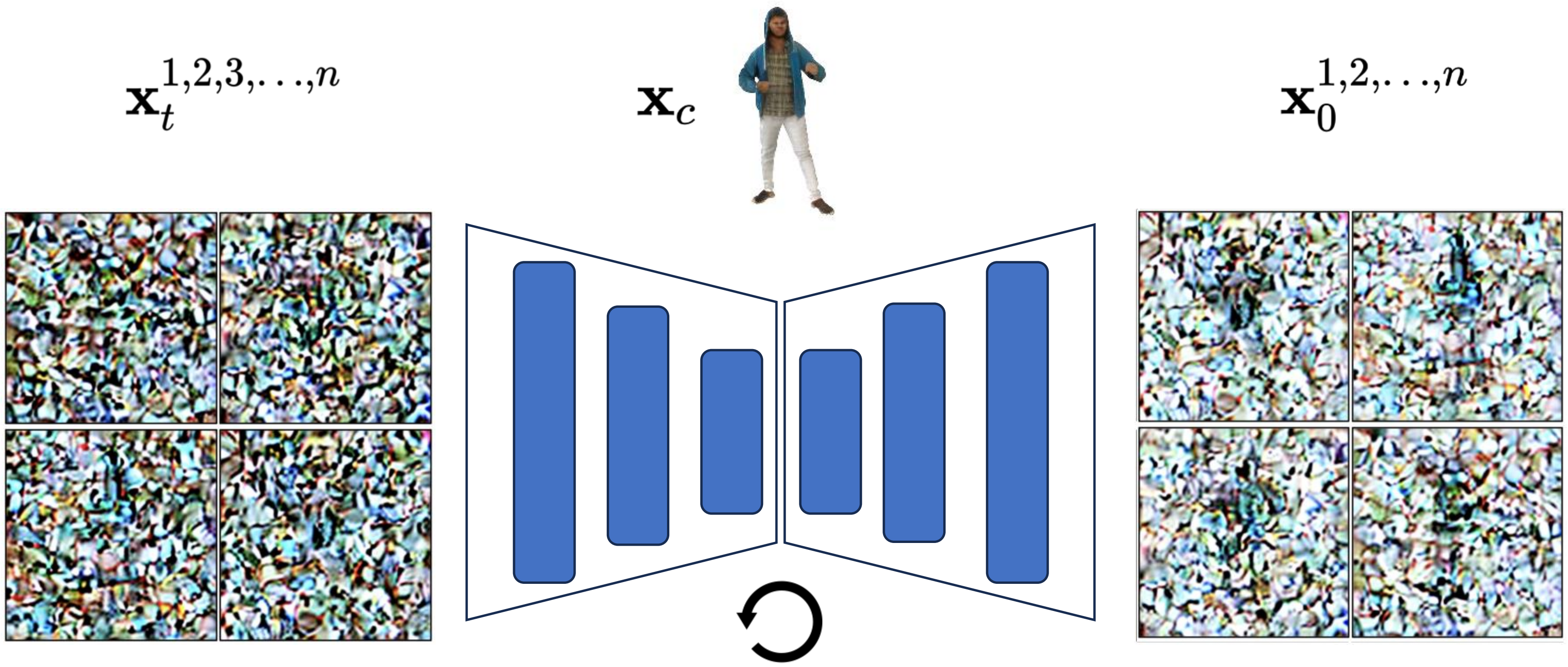


Multi-view Diffusion Model



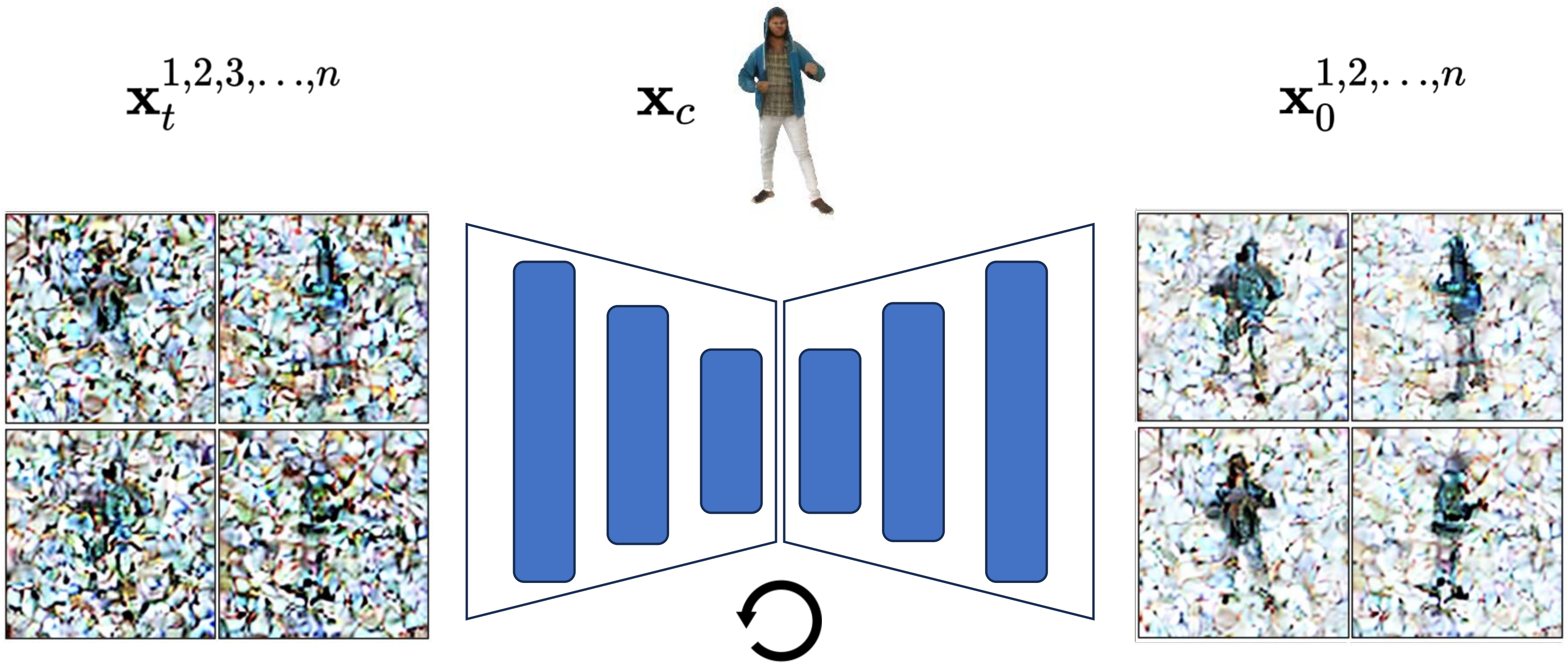
Pre-trained on **5B** 2D images and **800K** 3D objects

Multi-view Diffusion Model



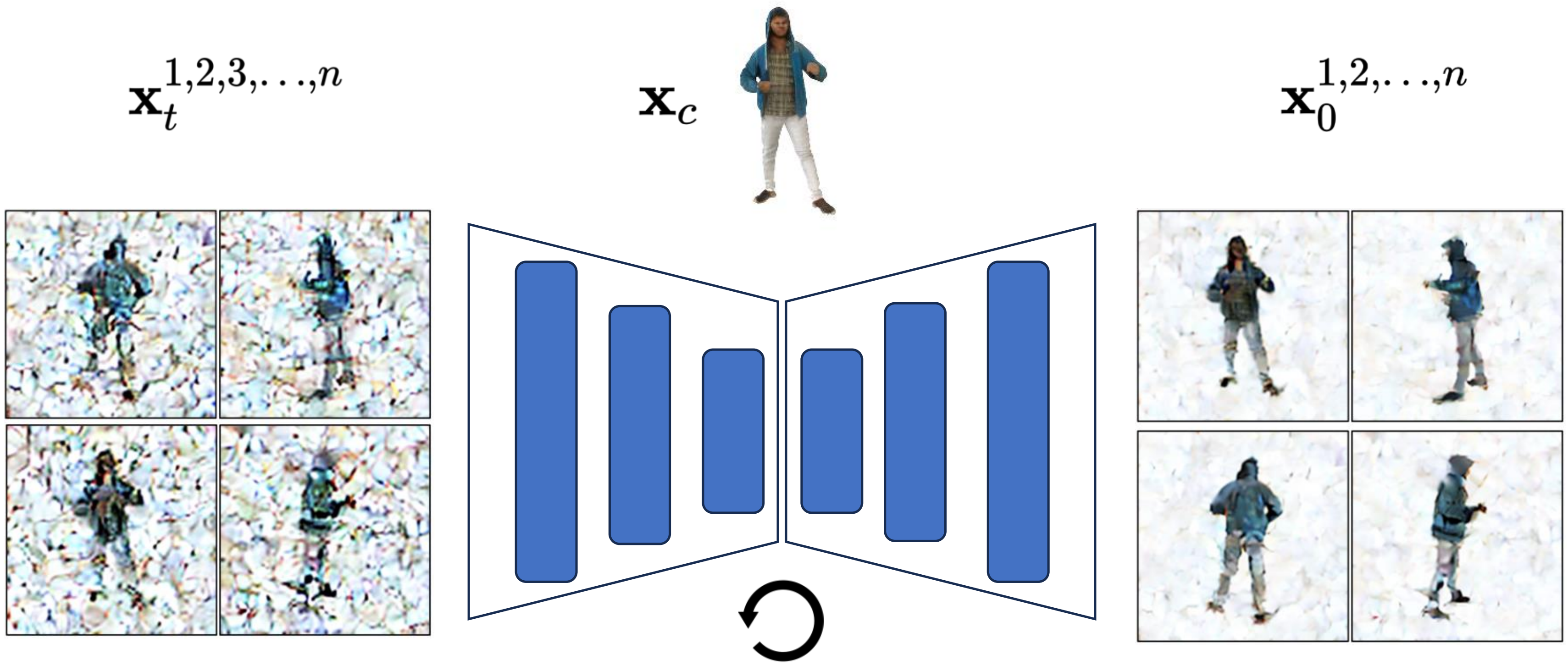
Pre-trained on **5B** 2D images and **800K** 3D objects

Multi-view Diffusion Model



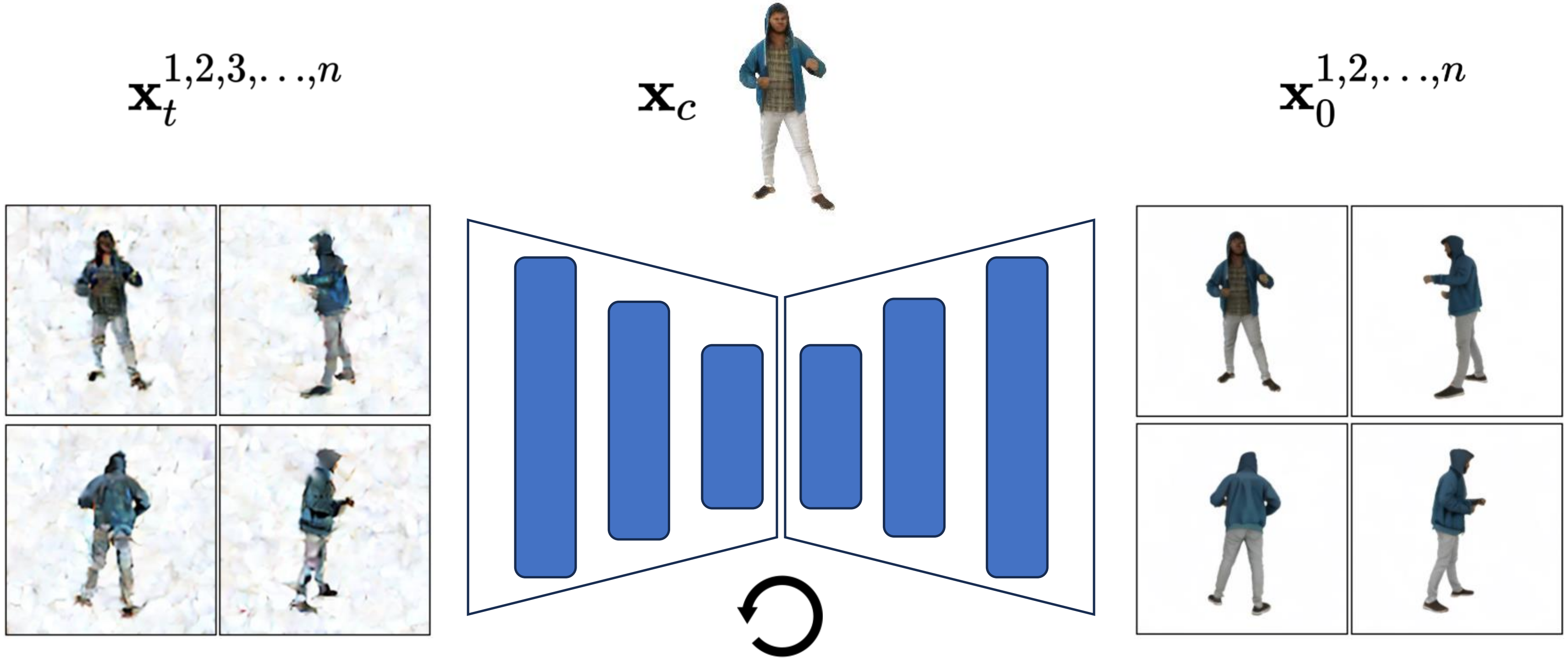
Pre-trained on **5B** 2D images and **800K** 3D objects

Multi-view Diffusion Model



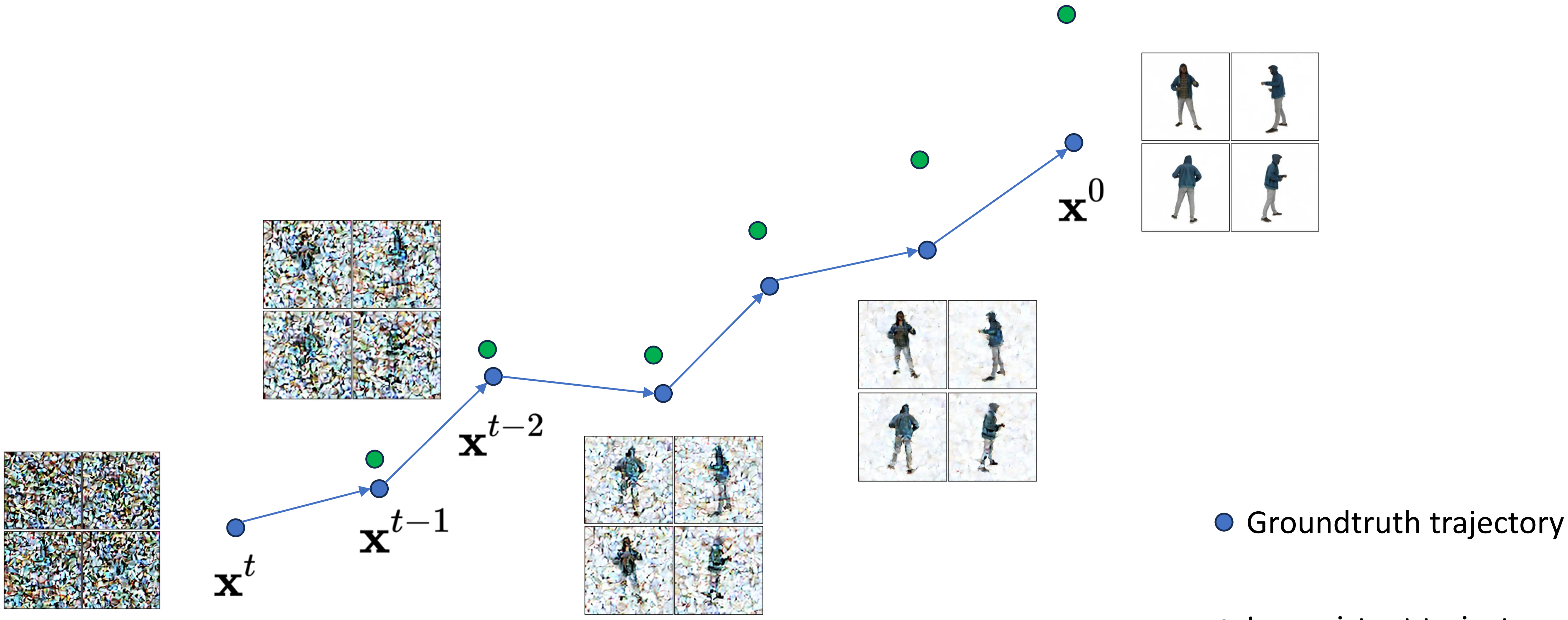
Pre-trained on **5B** 2D images and **800K** 3D objects

Multi-view Diffusion Model



Pre-trained on **5B** 2D images and **800K** 3D objects

Inconsistency accumulates along trajectory

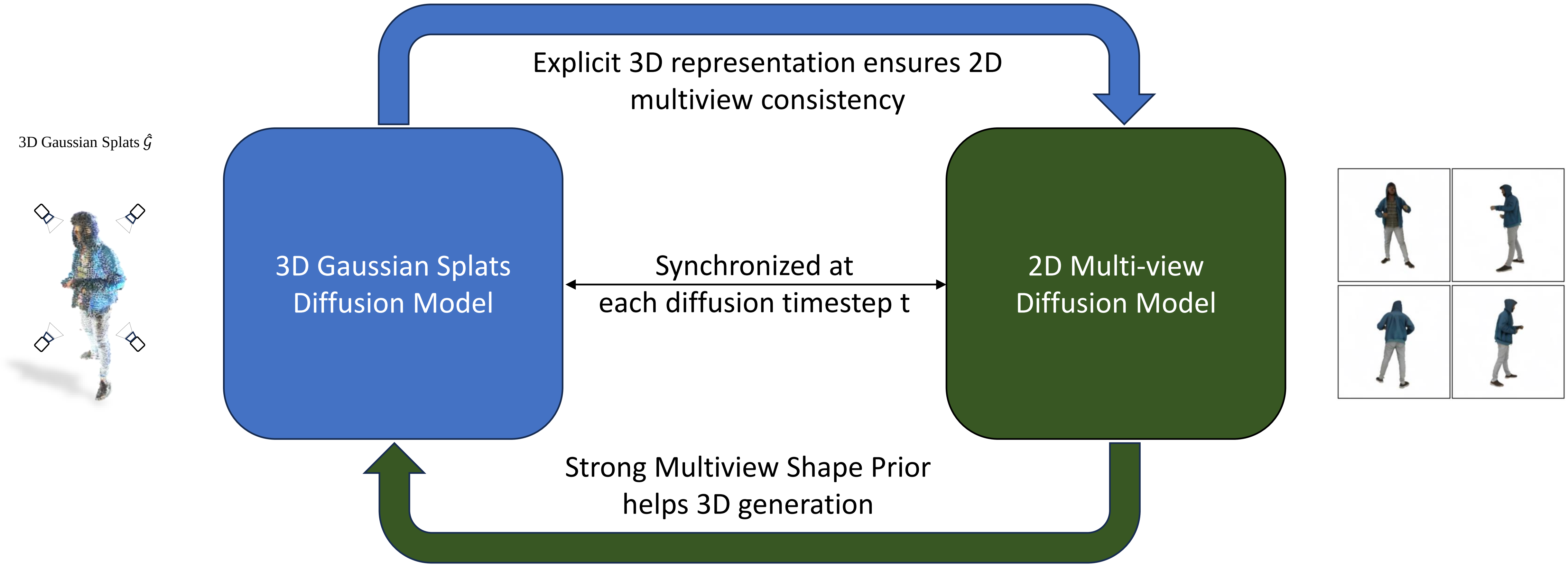


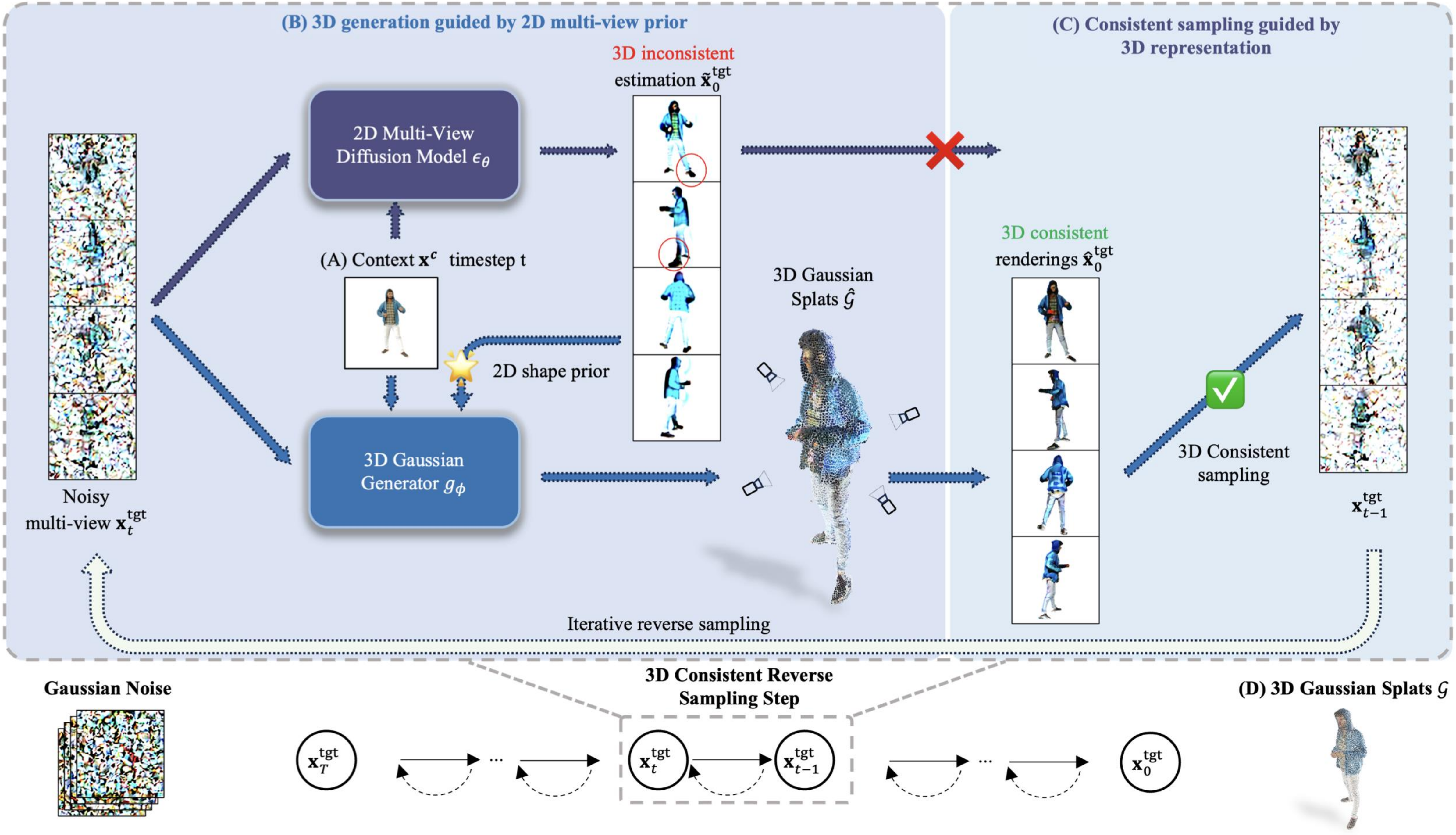
Main contents

- 3D Reconstruction from single Image
- 2D Diffusion Model for Novel-view Synthesis
  - Novel-view Diffusion Models
  - Multi-view Image Diffusion Models
- **Sync 2D Diffusion & 3D Reconstruction**

Gen-3Diffusion: Sync 2D Diffusion & 3D Recon

- 2D Difffusion leverages Image Diffusion Prior
- 3D Reconstructor provides 3D representation




(B) 3D generation guided by 2D multi-view prior
(C) Consistent sampling guided by 3D representation
3D inconsistent
estimation $\tilde{\mathbf{x}}_0^{\text{tgt}}$
2D Multi-View Diffusion Model ϵ_θ
(A) Context $\mathbf{x}^c$ timestep t
2D shape prior
3D Gaussian Generator g_ϕ
Noisy multi-view $\mathbf{x}_t^{\text{tgt}}$
3D Gaussian Splats $\hat{\mathcal{G}}$
3D consistent
renderings $\hat{\mathbf{x}}_0^{\text{tgt}}$
3D Consistent sampling
$\mathbf{x}_{t-1}^{\text{tgt}}$
Iterative reverse sampling
Gaussian Noise
3D Consistent Reverse Sampling Step
$\mathbf{x}_T^{\text{tgt}}$
$\mathbf{x}_t^{\text{tgt}}$
$\mathbf{x}_{t-1}^{\text{tgt}}$
$\mathbf{x}_0^{\text{tgt}}$
(D) 3D Gaussian Splats $\mathcal{G}$

Algorithm

Algorithm 1 Joint 2D & 3D Diffusion Training

Input: Dataset of posed multi-view images $\mathbf{x}_0^{\text{tgt}}$, π^{tgt}, $\mathbf{x}_0^{\text{novel}}$, π^{novel}, a context image $\mathbf{x}^{\text{c}}$, text description y
Output: Optimized 2D multi-view diffusion model ϵ_θ and 3D-GS generative model g_ϕ

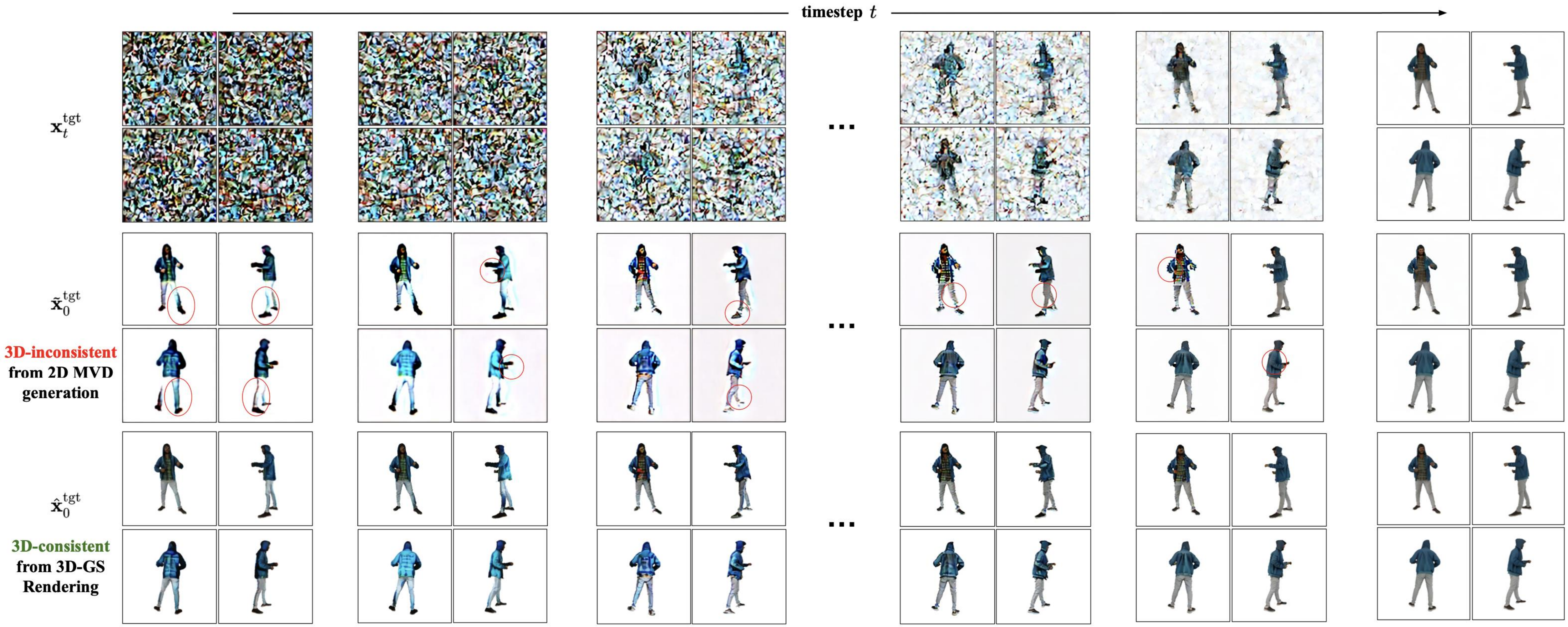
repeat
$\quad \{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}, \mathbf{x}^{\text{c}}, y\} \sim q(\{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}, \mathbf{x}^{\text{c}}, y\})$
$\quad t \sim \text{Uniform}(\{1, \ldots, T\})$; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\quad \mathbf{x}_t^{\text{tgt}} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
$\quad \tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t))$
$\quad \hat{\mathcal{G}} = g_\phi\left(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^{\text{c}}, \tilde{\mathbf{x}}_0^{\text{tgt}}\right)$ // Enhance conditional 3D generation with 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from ϵ_θ
$\quad \{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\} = \texttt{renderer}\left(\hat{\mathcal{G}}, \{\pi^{\text{tgt}}, \pi^{\text{novel}}\}\right)$
$\quad$ Compute loss $\mathcal{L}_{total}$ (Eq. (9))
$\quad$ Gradient step to update ϵ_θ, g_ϕ
until converged

Algorithm 2 3D Consistent Guided Sampling

Input: A context image $\mathbf{x}^c$ and text y; Converged 2D diffusion model ϵ_θ and 3D generative model g_ϕ
Output: 3D Gaussian Splats $\mathcal{G}$ of the 2D image $\mathbf{x}^c$

$\mathbf{x}_T^{\text{tgt}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
$\quad \tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t))$
$\quad \hat{\mathcal{G}} = g_\phi\left(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^{\text{c}}, \tilde{\mathbf{x}}_0^{\text{tgt}}\right)$
$\quad \hat{\mathbf{x}}_0^{\text{tgt}} = \texttt{renderer}\left(\hat{\mathcal{G}}, \pi^{\text{tgt}}\right)$
$\quad \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t\text{-}1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t\text{-}1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0^{\text{tgt}}$ // Guide 2D sampling with 3D consistent multi-view renderings
$\quad \mathbf{x}_{t-1}^{\text{tgt}} \sim \mathcal{N}\left(\mathbf{x}_{t-1}^{\text{tgt}}; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}\right), \tilde{\beta}_{t-1}\mathbf{I})\right)$
end for
return $\mathcal{G} = g_\phi\left(\mathbf{x}_0^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}, \mathbf{x}^{\text{c}}, t=0\right)$

Explicit 3D-GS helps 2D Diffusion



Results: Object Reconstruction

Reconstruction w/ novel view diffusion

Input Image

Gen-3Diffusion

Single-view Diffusion

Multi-view Diffusion

Reconstruction w/ novel view diffusion

Input Image **Gen-3Diffusion** **Single-view Diffusion** **Multi-view Diffusion**

Reconstruction w/ novel view diffusion

Input Image **Gen-3Diffusion** **Single-view Diffusion** **Multi-view Diffusion**

Reconstruction w/ novel view diffusion

Input Image

Gen-3Diffusion

Single-view Diffusion

Multi-view Diffusion

Reconstruction w/ direct 3d reconstruction

Input Image

Gen-3Diffusion

TripoSR

LGM

Reconstruction w/ direct 3d reconstruction

Input Image | **Gen-3Diffusion** | **TripoSR** | **LGM**

Reconstruction w/ direct 3d reconstruction

Input Image

Gen-3Diffusion

TripoSR

LGM

Reconstruction w/ direct 3d reconstruction

Input Image

Gen-3Diffusion

TripoSR

LGM

Results: Avatar Reconstruction

Reconstruction avatar appearance

Input Image **Gen-3Diffusion** **SiTH** **SiFU**

Reconstruction avatar appearance

Input Image **Gen-3Diffusion** **SiTH** **SiFU**

Children

Input Image

Gen-3Diffusion

SiTH

SiFU

Reconstruction avatar geometry

Input Image

Gen-3Diffusion

ICON

ECON

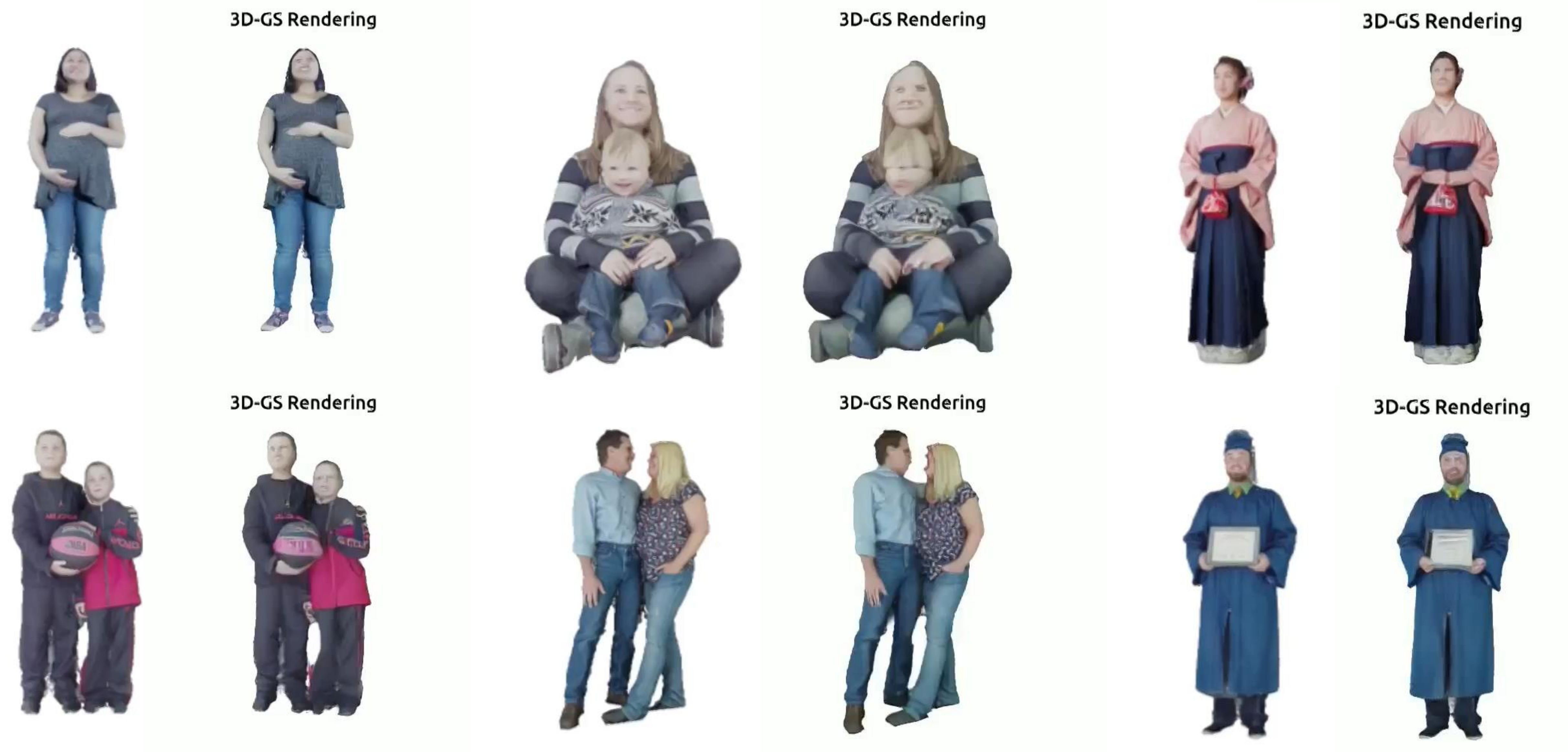
Reconstruction avatar geometry

Input Image **Gen-3Diffusion** **ICON** **ECON**

Strong generalization

Input Image **Gen-3Diffusion** **SiTH** **SiFU** **ECON**

Strong generalization



Take away messages

- With training on massive data, image diffusion models achieve superior image generation quality
- The image diffusion prior can be applied to 3D tasks, e.g. generate novel views
- Novel view diffusion models lack 3D consistency because they don't have an explicit 3D representation
- 2D Diffusion Models and 3D Reconstruction Models can be combined to achieve excellent reconstruction capability