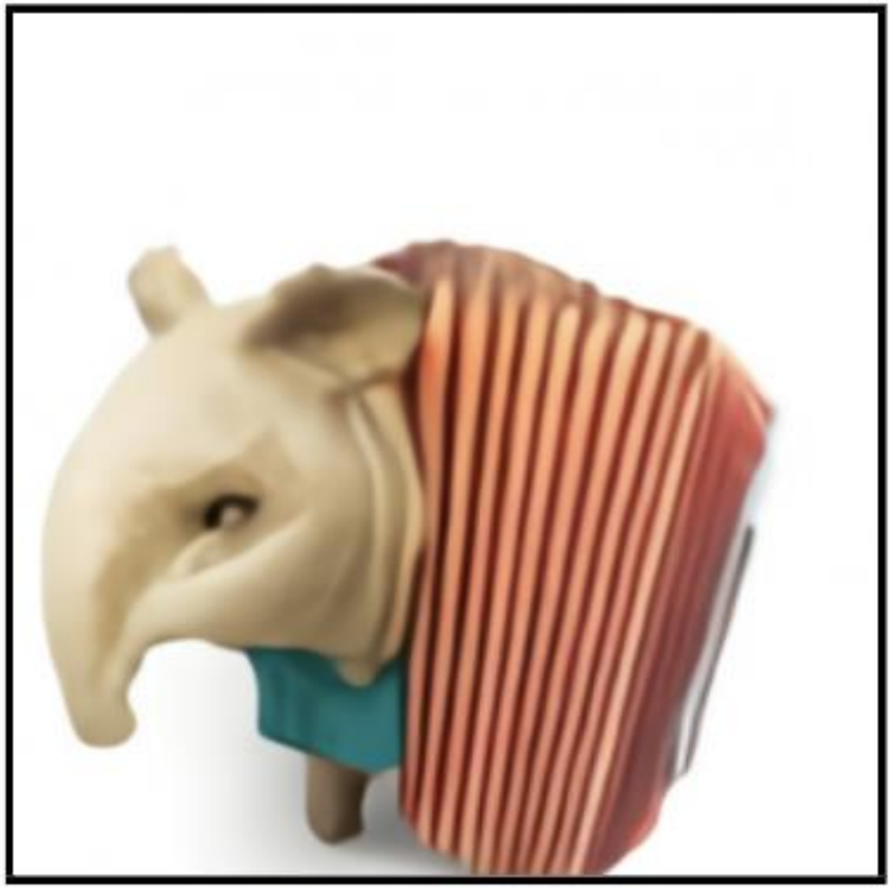
Controllable 3D Generation

Vova Kim | Senior Research Scientist, Adobe Research

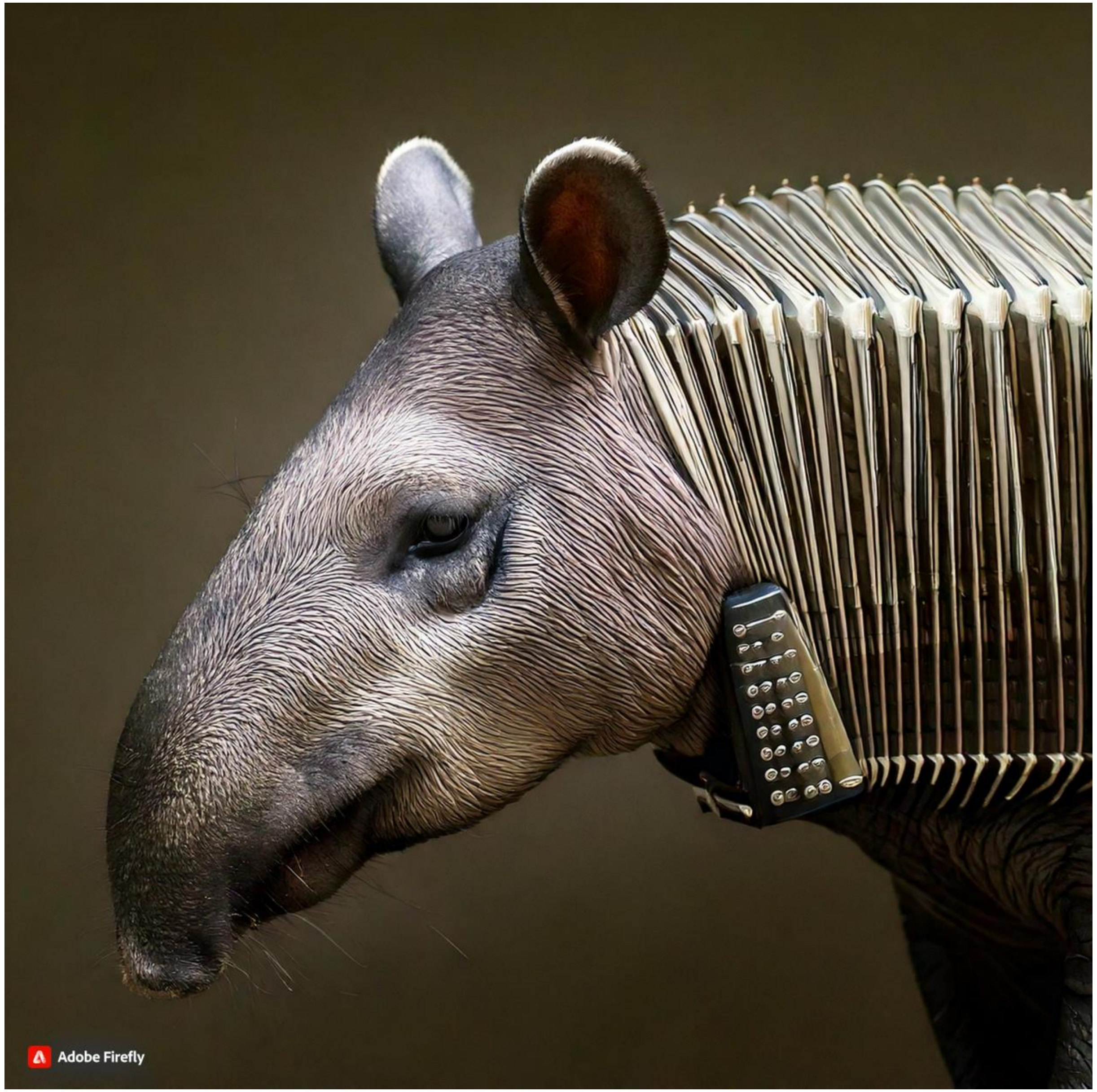
Adobe

Motivation

- Text-to-Image ~ amazing progress



2021

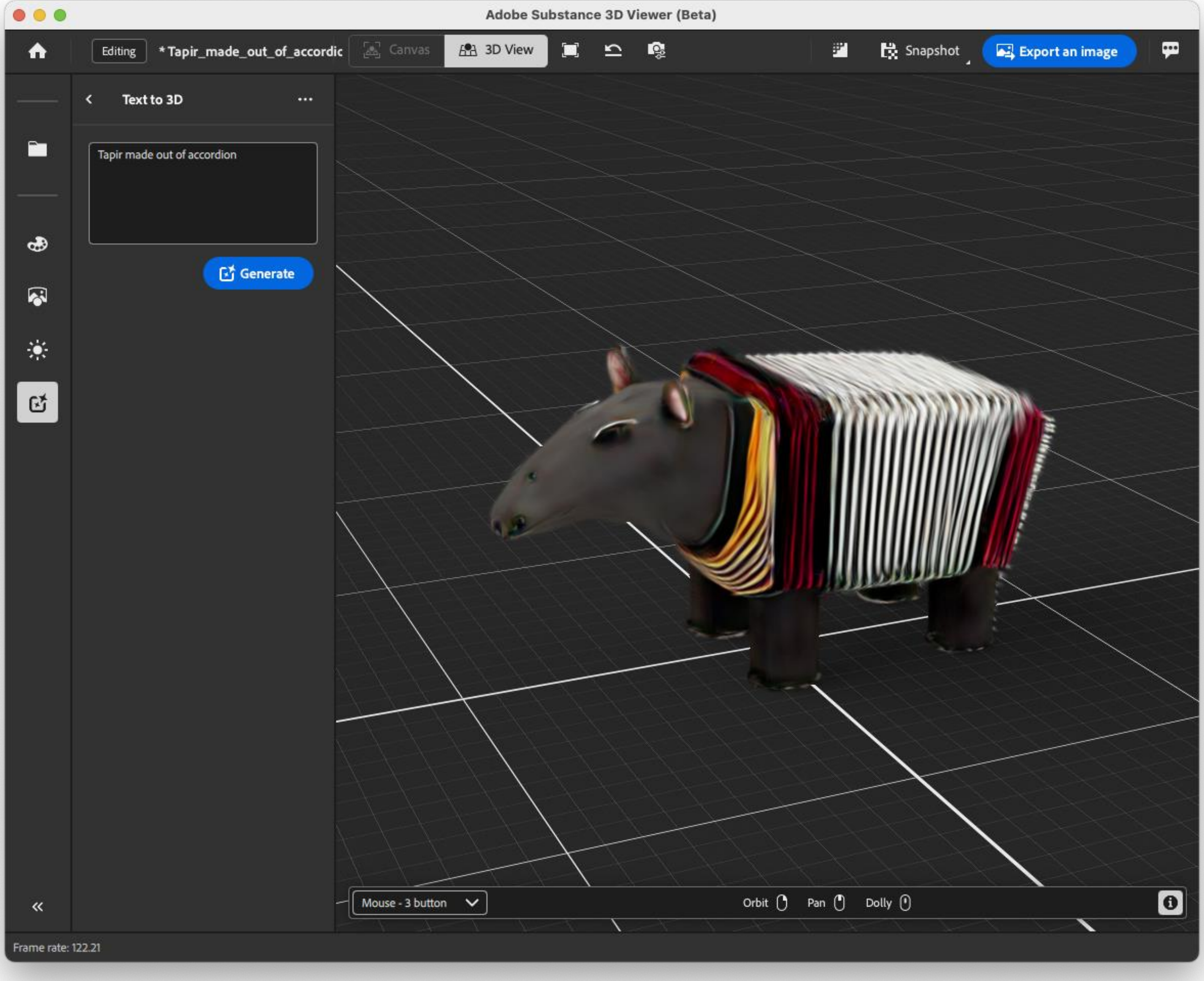


2024

“Tapir made out of accordion”

Motivation

- Text-to-Image
- Text-to-3D



Tapir made out of accordion

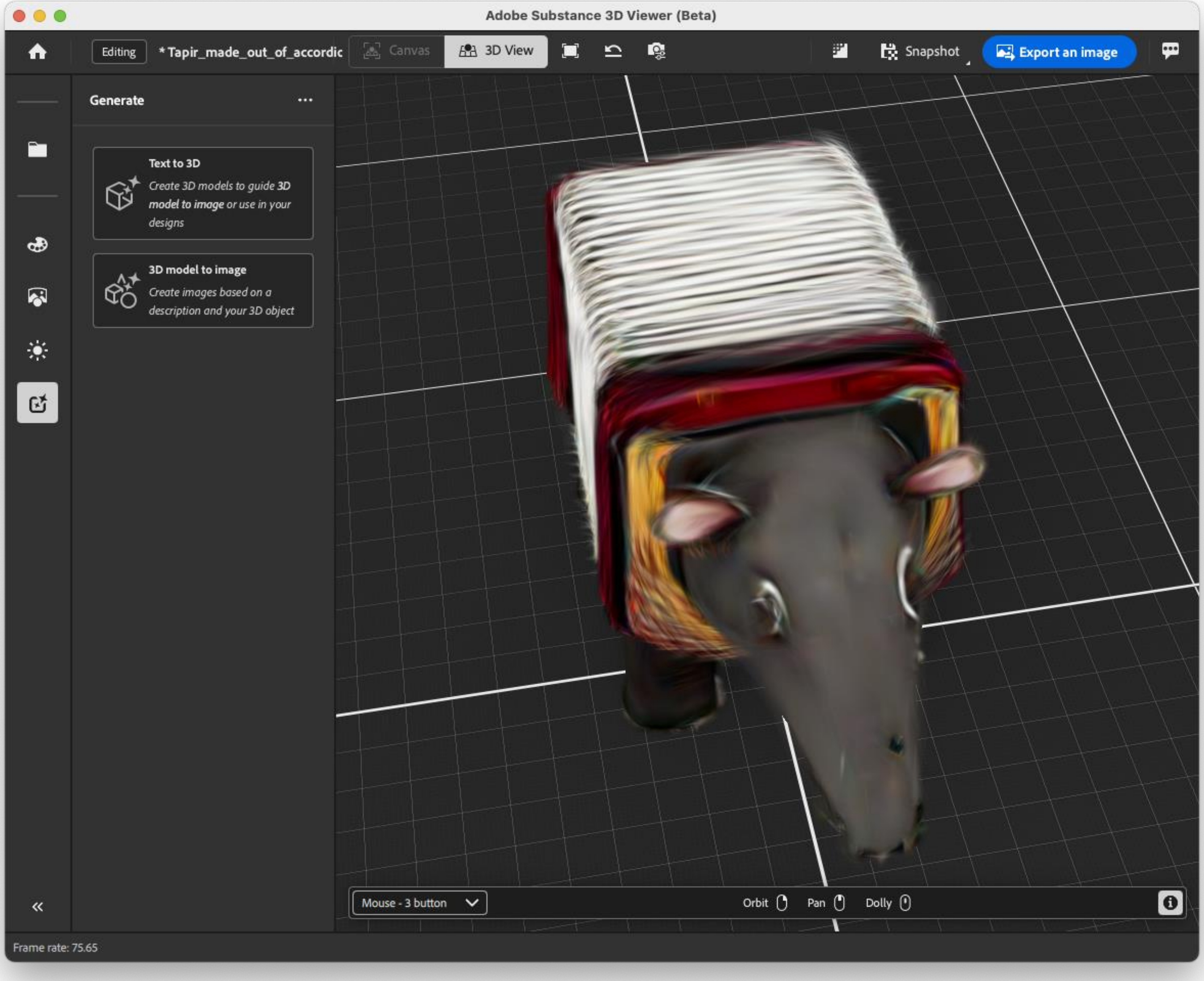
Frame rate: 122.21

“Tapir made out of accordion”

Adobe

Motivation

- Text-to-Image
- Text-to-3D



Text to 3D

*Create 3D models to guide 3D **model to image** or use in your designs*

3D model to image

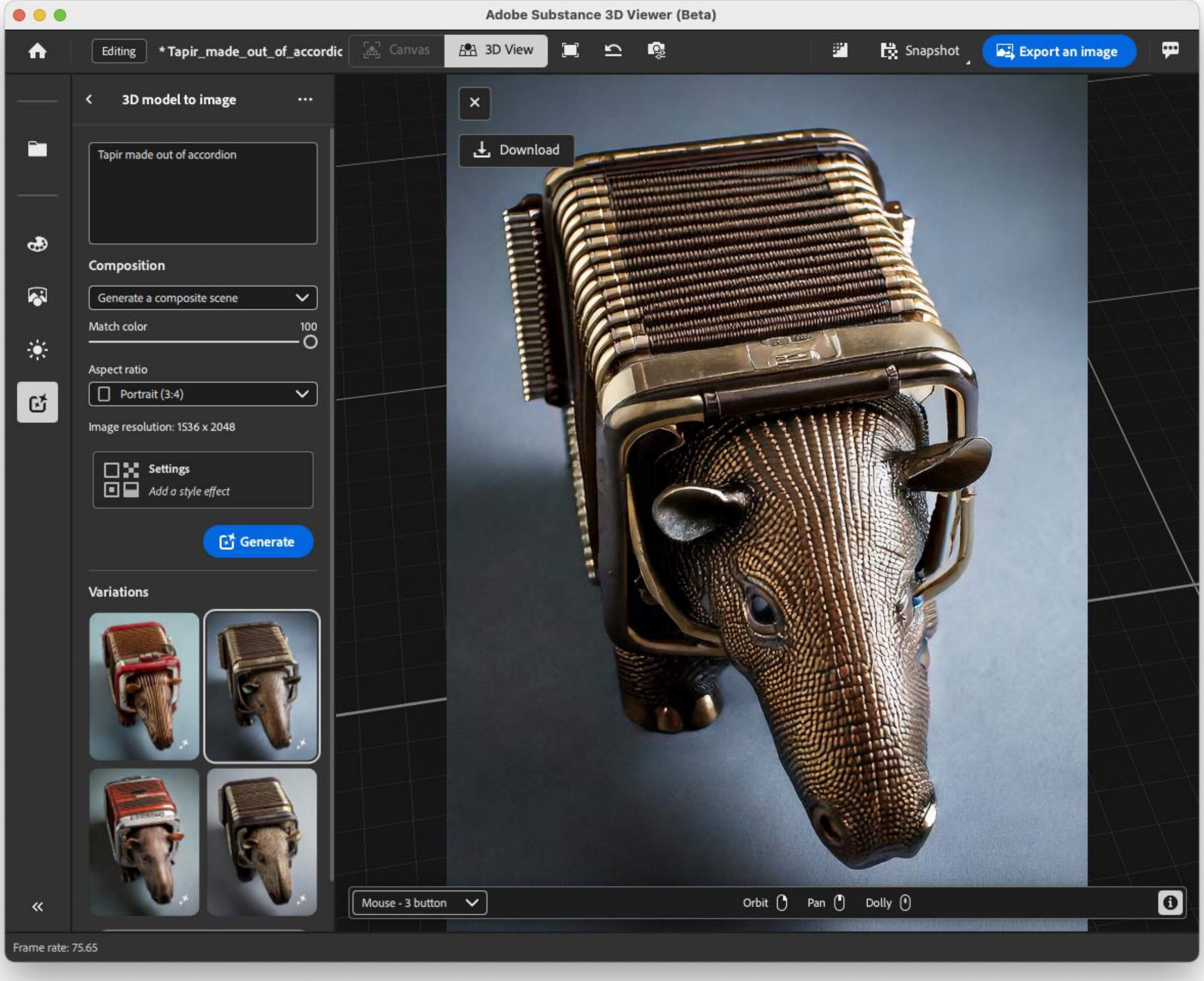
Create images based on a description and your 3D object

Frame rate: 75.65

"Tapir made out of accordion"

Motivation

- Text-to-Image
- Text-to-3D
- 3D-to-Image



3D model to image

Tapir made out of accordion

Composition

Generate a composite scene

Match color 100

Aspect ratio

Portrait (3:4)

Image resolution: 1536 x 2048

Settings

Add a style effect

Generate

Variations

Frame rate: 75.65

“Tapir made out of accordion”

Motivation

- Text-to-Image
- Text-to-3D
- 3D-to-Image

How do we customize it further? E.g.

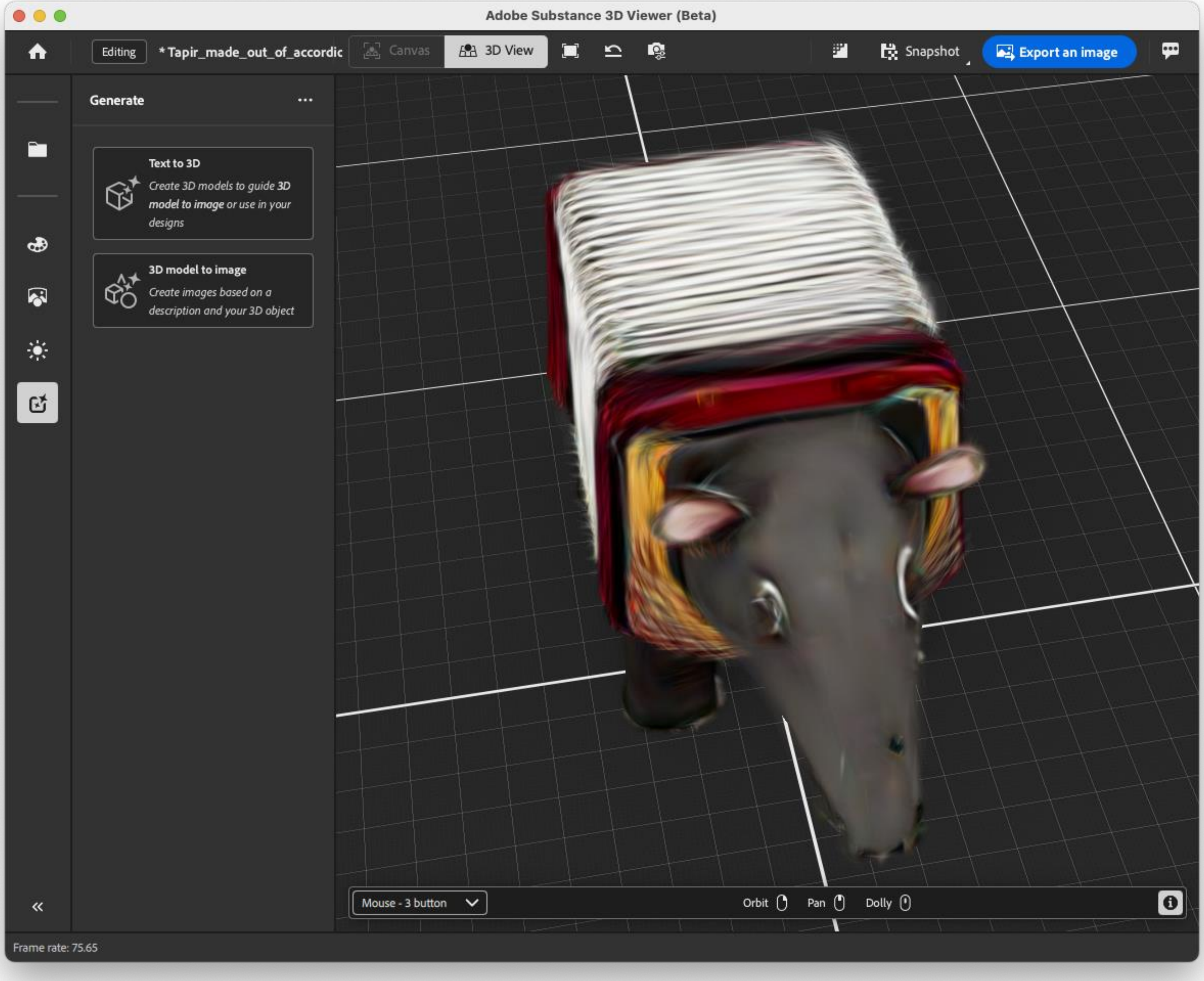
Tilt head (deform)

Make it look more like hippo (deform)

Add unicorn horn (sculpt)

Make skin look more natural (detail)

Move eye to another location (detail)



Text to 3D

Create 3D models to guide 3D model to image or use in your designs

3D model to image

Create images based on a description and your 3D object

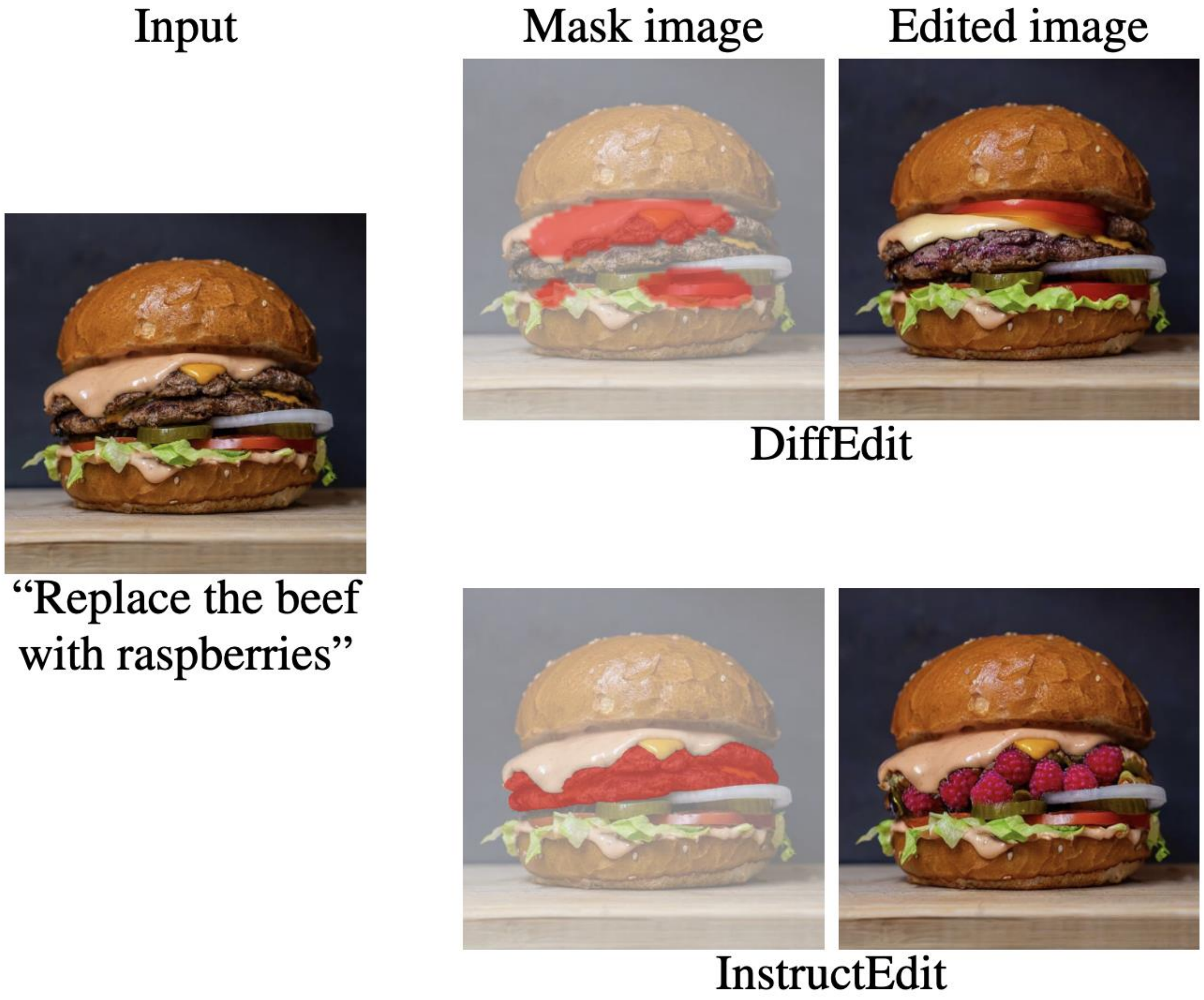
Frame rate: 75.65

“Tapir made out of accordion”

Adobe

Prior Work: Image Editing

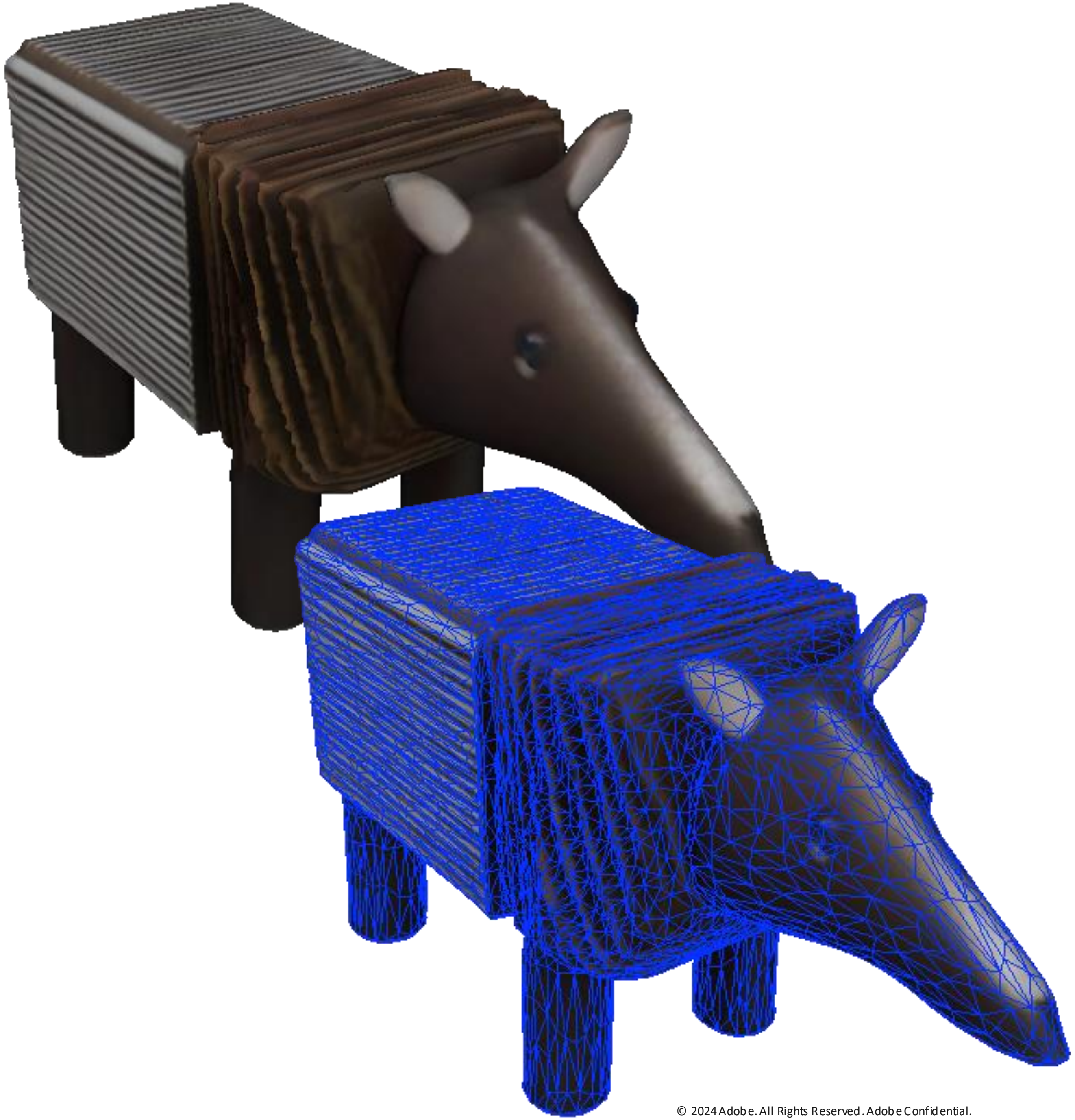
- Entangled control:
  - Geometry
  - Materials
  - Lighting
  - Camera Intrinsics
  - Camera Extrinsics
  - Composition
- Text alone is not always well-defined



InstructEdit, Wang et al, 2023

Prior Work: Representations

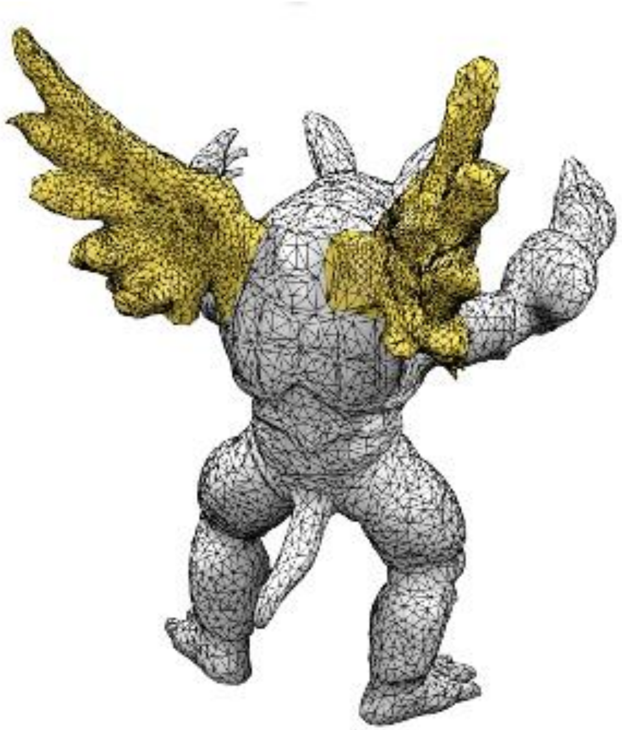
- NeRFs / Gaussian Splats
  - Geometry **(no surface priors)**
  - Materials **(not easy to disentangle from appearance)**
  - Lighting **(not easy to disentangle from appearance)**
  - Camera Intrinsics
  - Camera Extrinsics
  - Composition
- Meshes / Surfaces
  - Traditional CG models fully disentangle appearance
  - Easy to use with traditional tools that artists know well
  - **(hard to learn, hard to represent, poor gradients)**



Overview

- Support mesh outputs (but use other representations as needed)
- Inspired by traditional workflows

Neural **Deformation**

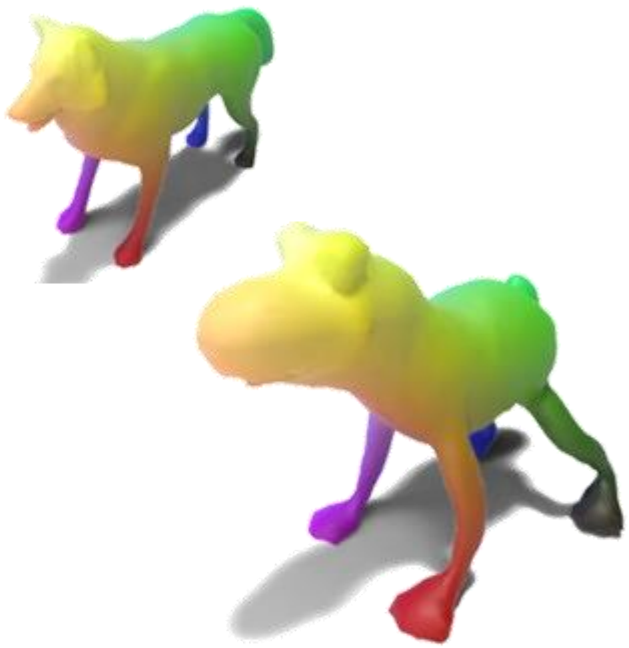


Generative **Scultping**

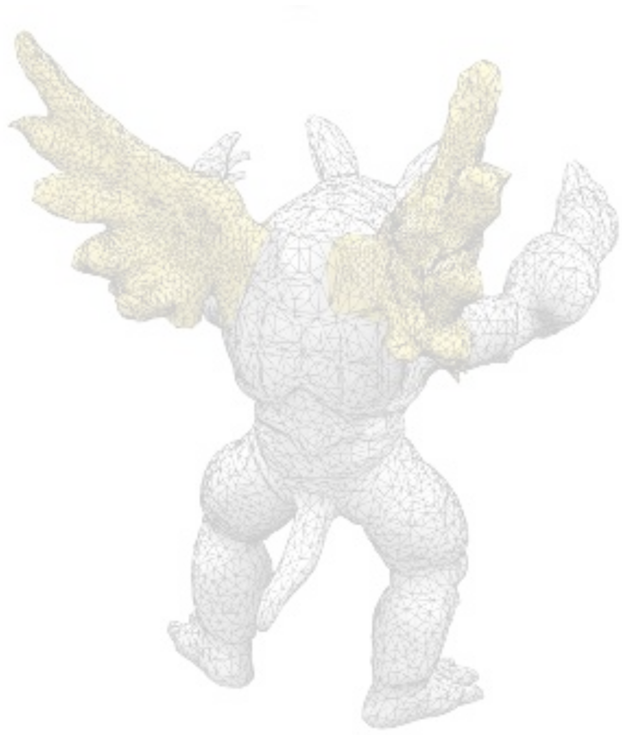
Generative **Detailization**

Overview

- Support mesh outputs (but use other representations as needed)
- Inspired by traditional workflows



Neural **Deformation**

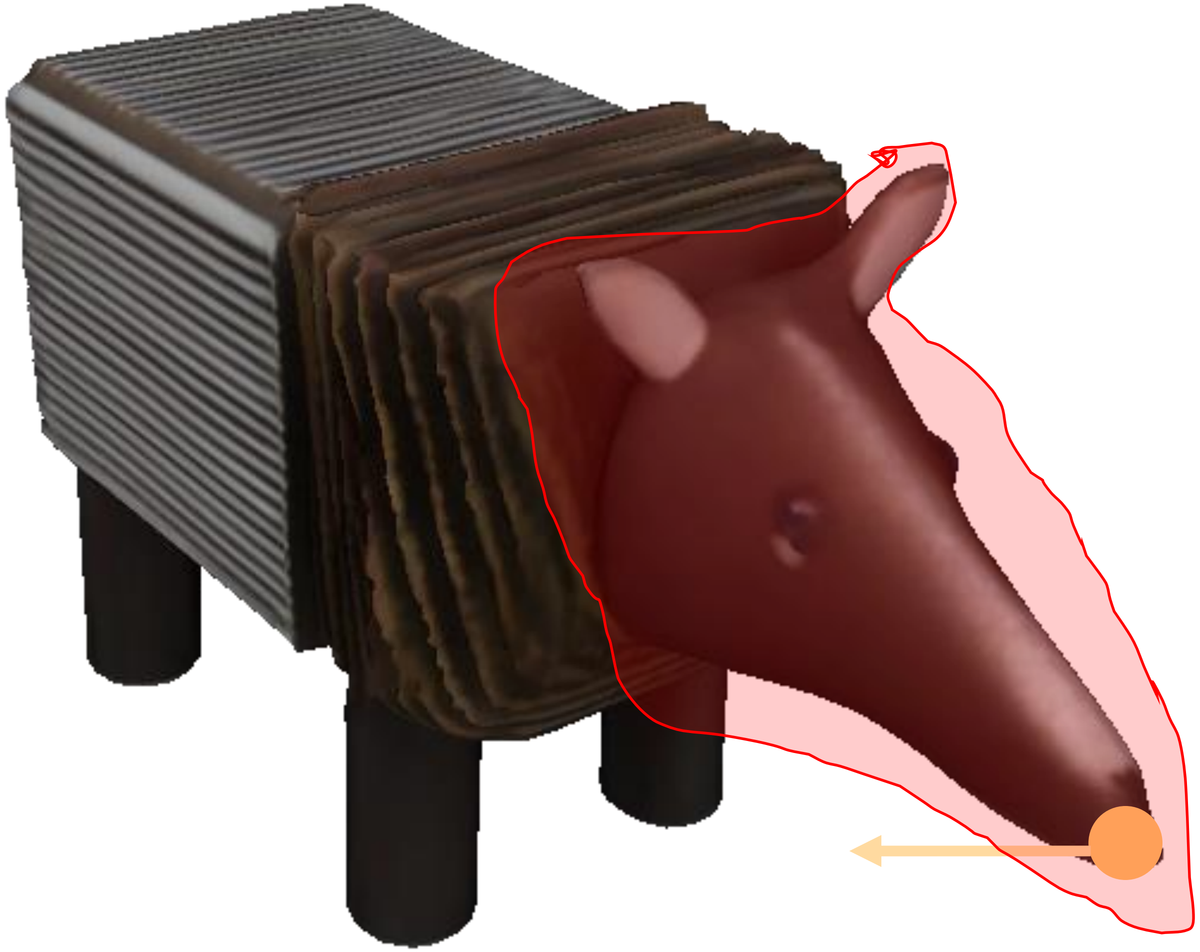


Generative **Scultping**

Generative **Detailization**

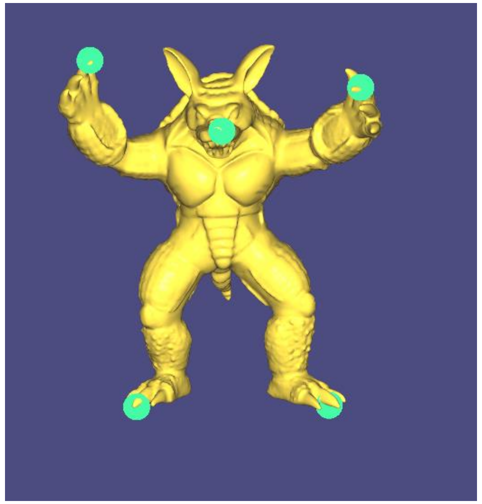
Deformation Examples

- “tilt tapir’s head” ●
- “make it’s head look more like a hippo”

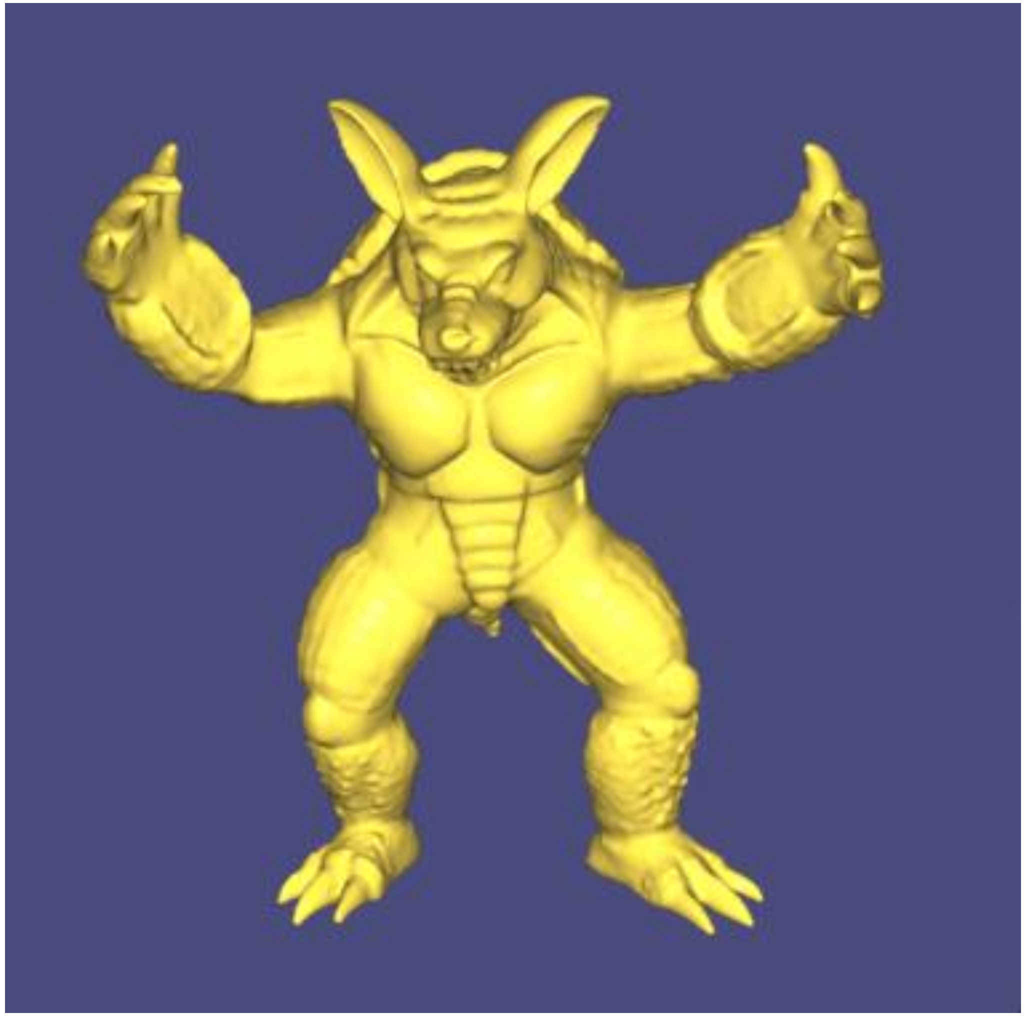


Prior Works: Deformation as Geometric Optimization

- Move control points AND preserve original geometry
- No semantics, e.g.:
  - Rubber-like behavior
  - Symmetry not preserved



Control points



E.g.: As-rigid-as-possible (ARAP)

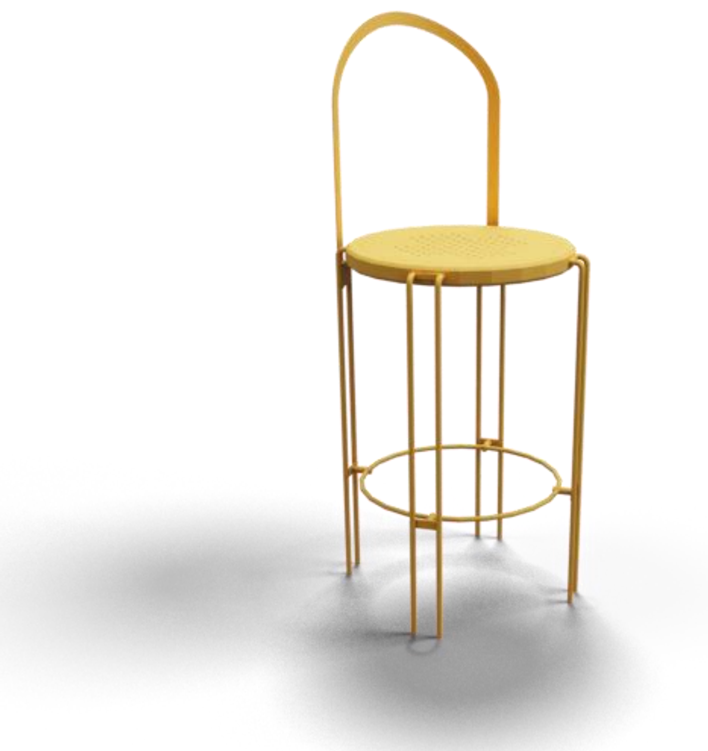
Adobe

As-Rigid-As-Possible Surface Modeling, Sorkine and Alexa, SGP 2007

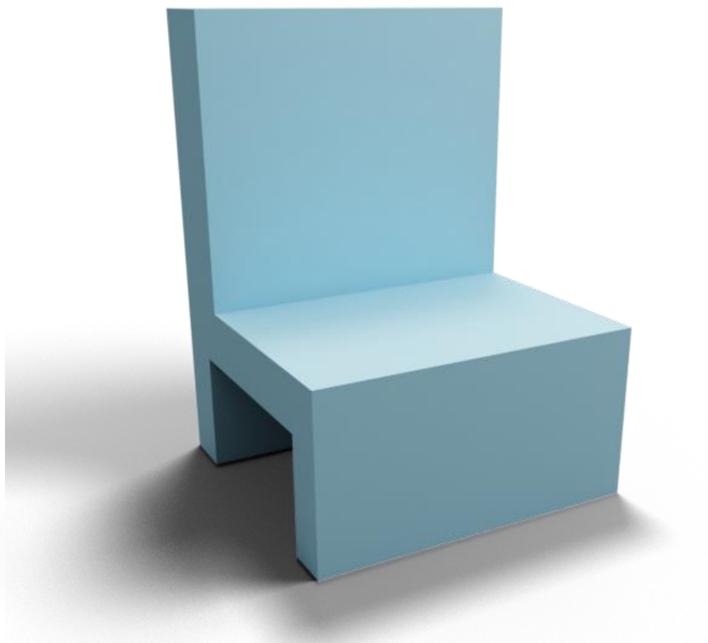
Prior Works: Deformation as a Learnable Map

- Naïve: learn direct map ~ hard to make it smooth enough

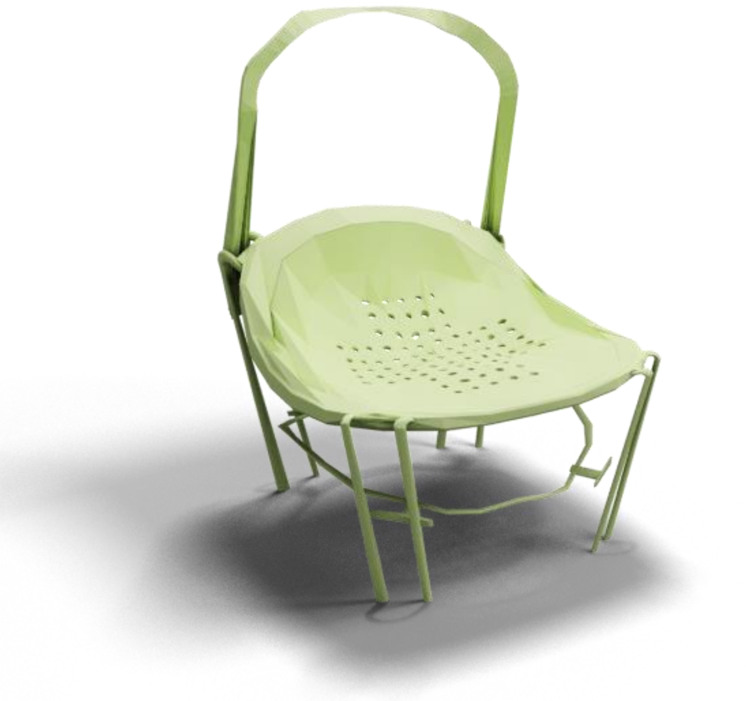
$$f_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}^3$$



Source



Target



Deep Deformation, Groueix et al. CGF 2019

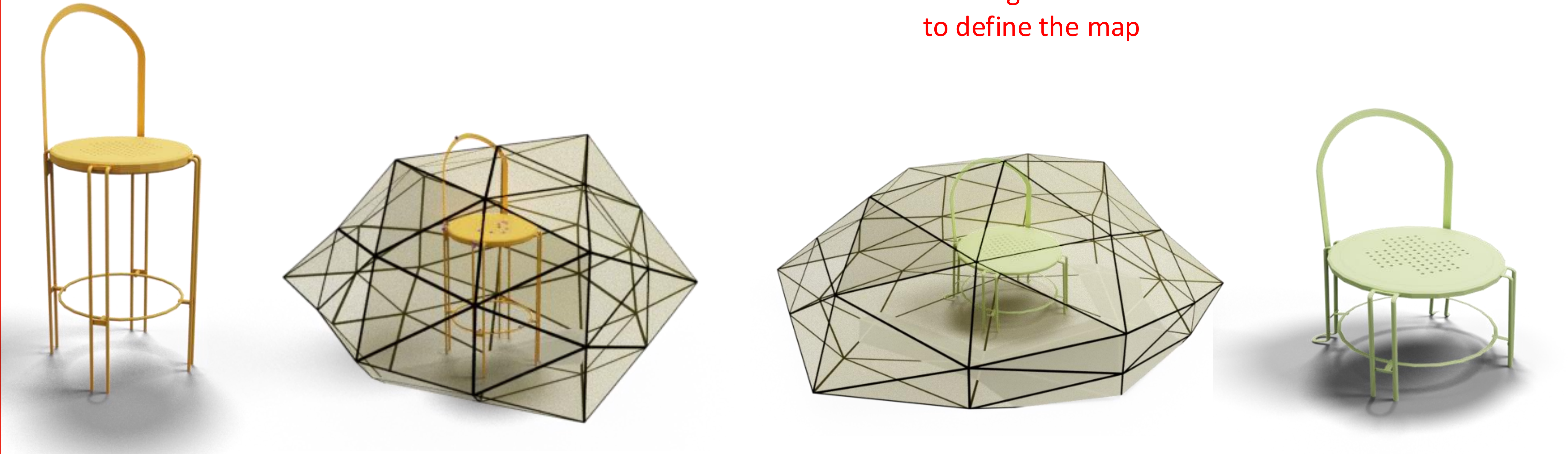
Prior Works: Deformation as a Learnable Map

- Cage-based: learn cage parameters ~ hard to predict expressive & accurate cages

$$f_\theta : \boxed{\mathcal{C}_{\text{init}} \times \mathcal{C}_{\text{deformed}}} \rightarrow \boxed{\text{MVC}} \rightarrow \mathbb{R}^3 \rightarrow \mathbb{R}^3$$

Predict cage parameters with a neural network

Use Cage-Based Deformation to define the map



Init Cage

Deformed Cage

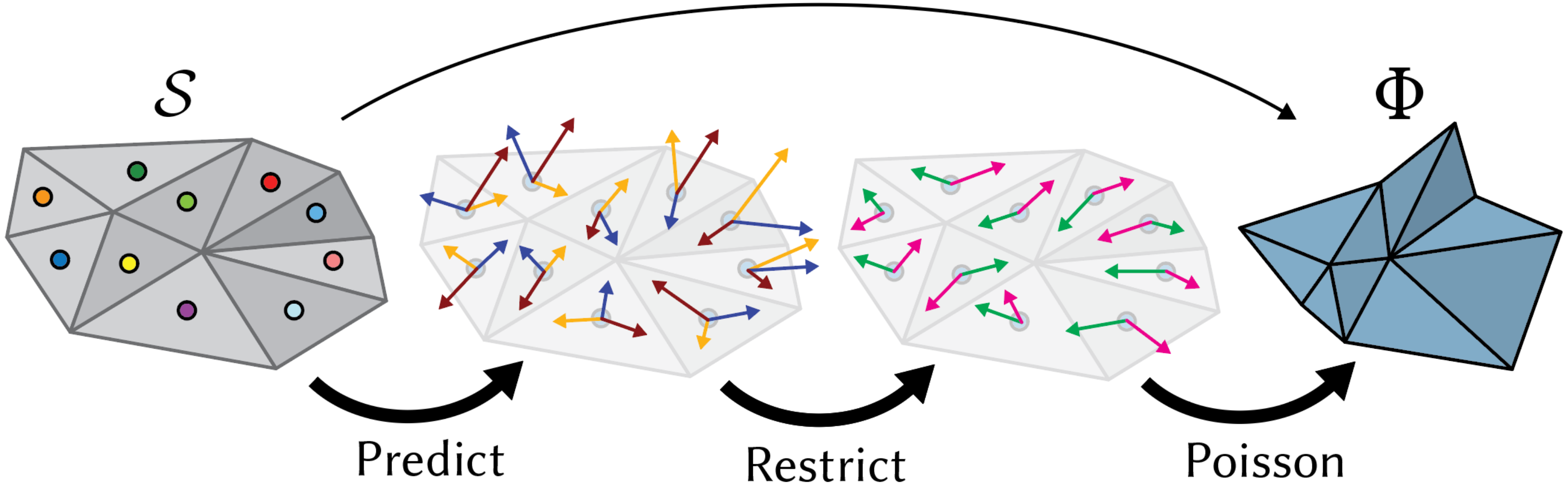
Neural Cages, Yifan et al., CVPR 2020

Prior Works: Deformation as a Learnable Map

- Neural Jacobian Fields
  - Smooth field (e.g., rotation is the same matrix)
  - Easy to maintain geometry details: shape-aware

$$\cancel{f_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}^3}$$

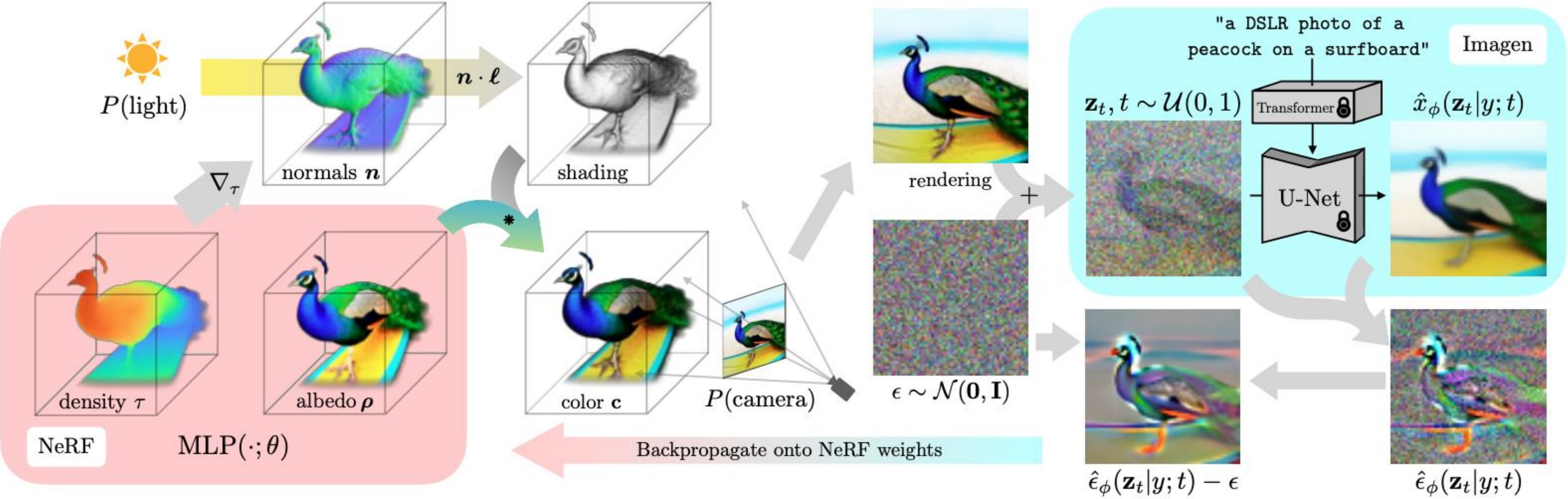
$$f_\theta : \underbrace{\mathcal{S}}_{\text{Surface features}} \rightarrow \underbrace{\mathbb{R}^{3\times 3}}_{\text{Deformation matrix}}$$



Neural Jacobian Fields, Aigerman et al., SIGGRAPH 2022

Prior Works: Score Distillation Sampling (SDS)

- Allows interjecting priors from pre-trained 2D model



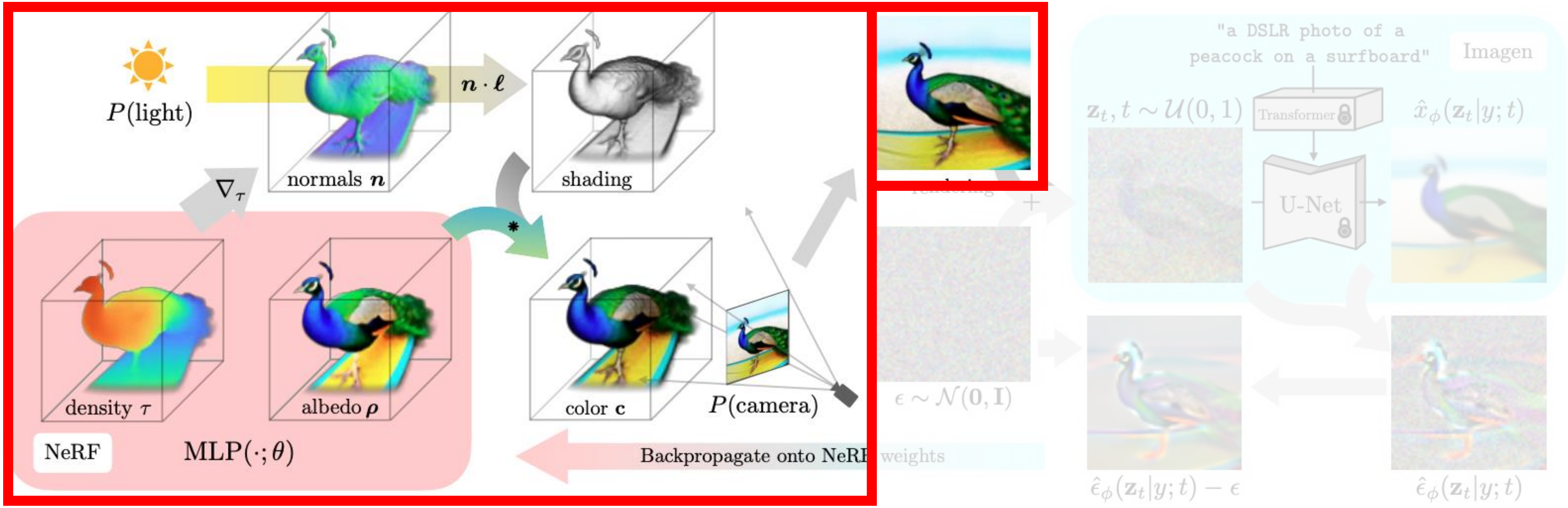
$P(\text{light})$
$\boldsymbol{n} \cdot \boldsymbol{\ell}$
normals $\boldsymbol{n}$
shading
∇_τ
density τ
albedo $\boldsymbol{\rho}$
color $\mathbf{c}$
NeRF
$\text{MLP}(\cdot;\theta)$
$P(\text{camera})$
rendering
+
$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
"a DSLR photo of a peacock on a surfboard"
Imagen
$\mathbf{z}_t, t \sim \mathcal{U}(0,1)$
Transformer
U-Net
$\hat{x}_\phi(\mathbf{z}_t|y;t)$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t) - \epsilon$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t)$
Backpropagate onto NeRF weights

Dreamfusion, Poole et al., 2022

Prior Works: Score Distillation Sampling (SDS)

- Allows interjecting priors from pre-trained 2D model

Differentiable Rendering

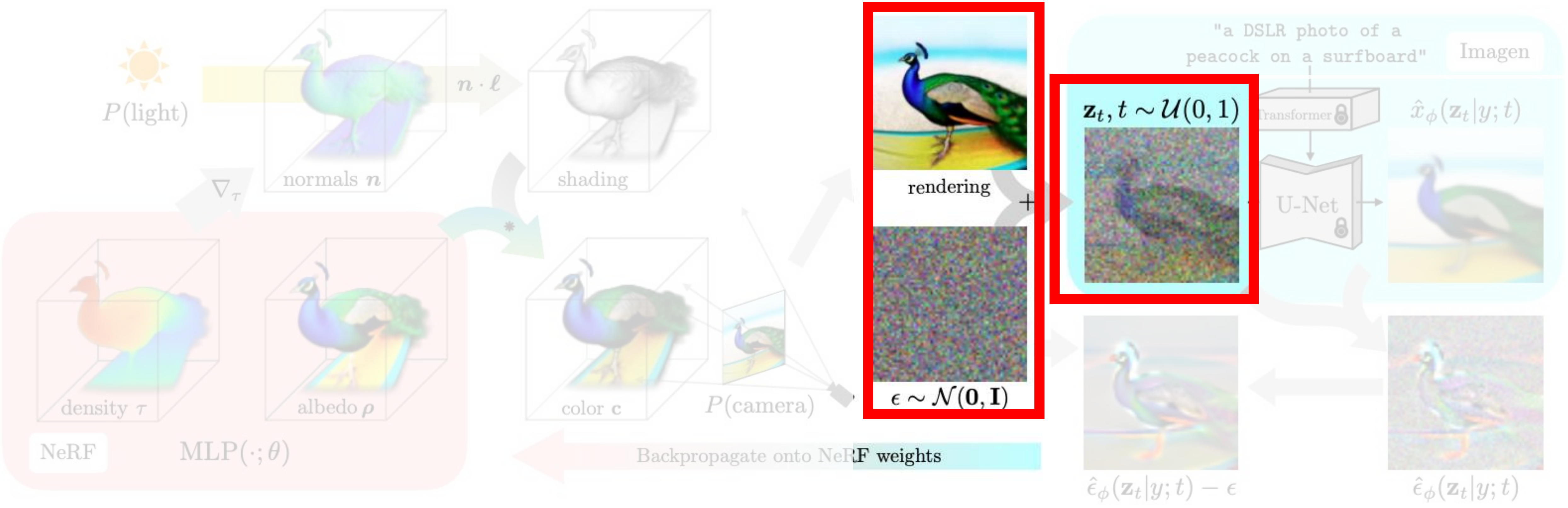

$P(\text{light})$
normals $\boldsymbol{n}$
$\boldsymbol{n} \cdot \boldsymbol{\ell}$
shading
∇_τ
density τ
albedo $\boldsymbol{\rho}$
color $\mathbf{c}$
$P(\text{camera})$
NeRF
$\text{MLP}(\cdot;\theta)$
Backpropagate onto NeRF weights
rendering
"a DSLR photo of a peacock on a surfboard"
Imagen
$\mathbf{z}_t, t \sim \mathcal{U}(0,1)$
Transformer
U-Net
$\hat{x}_\phi(\mathbf{z}_t|y;t)$
$\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t) - \epsilon$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t)$

Dreamfusion, Poole et al., 2022

Prior Works: Score Distillation Sampling (SDS)

- Allows interjecting priors from pre-trained 2D model

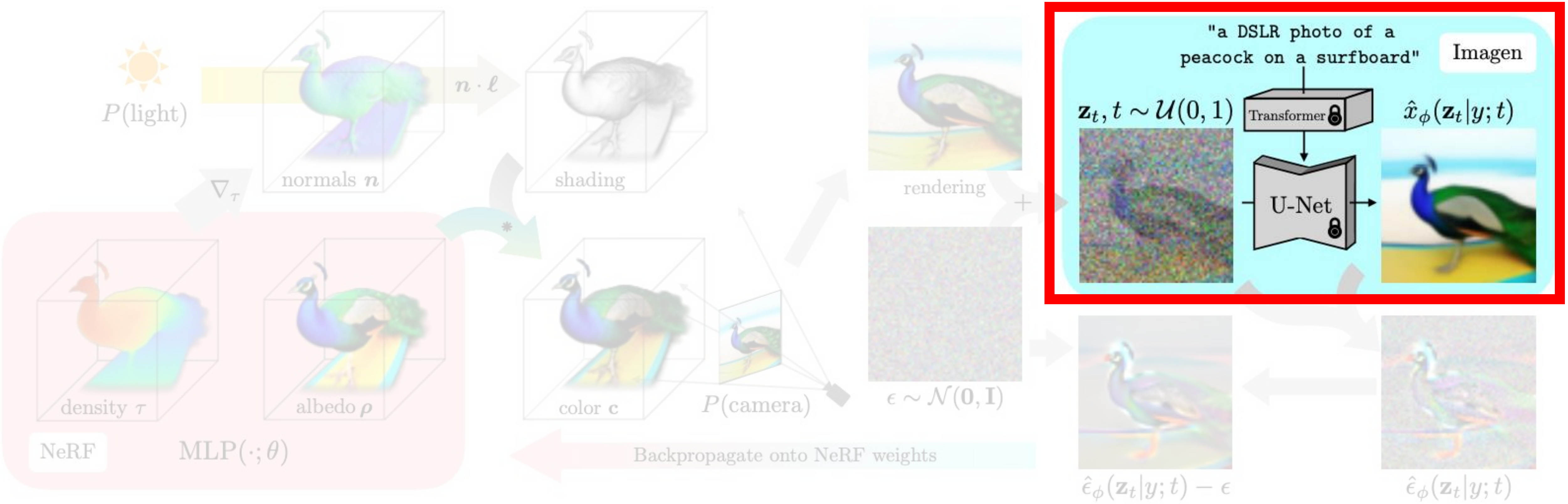
Add noise to a rendered image


P(light)
normals n
$n \cdot \ell$
shading
∇_τ
density τ
albedo ρ
color $\mathbf{c}$
P(camera)
NeRF
MLP$(\cdot;\theta)$
Backpropagate onto NeRF weights
rendering
+
$\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
$\mathbf{z}_t, t \sim \mathcal{U}(0,1)$
"a DSLR photo of a peacock on a surfboard"
Imagen
Transformer
U-Net
$\hat{x}_\phi(\mathbf{z}_t|y;t)$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t) - \epsilon$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t)$

Dreamfusion, Poole et al., 2022

Prior Works: Score Distillation Sampling (SDS)

- Allows interjecting priors from pre-trained 2D model



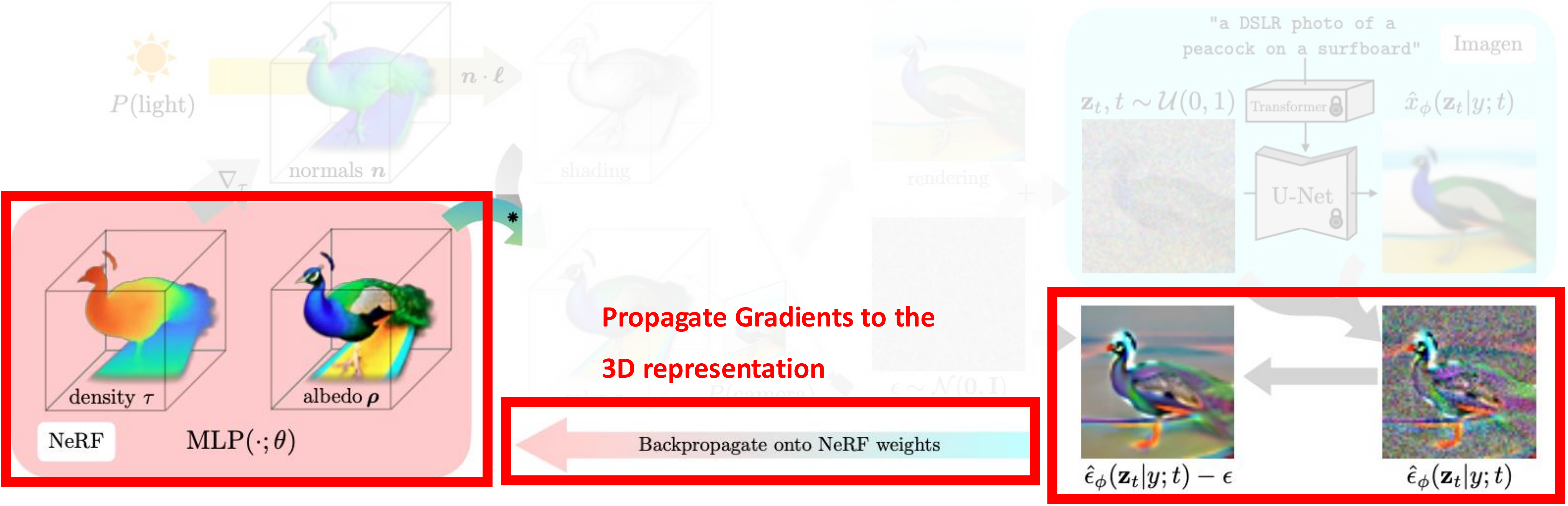
Denoise via a pre-trained Image Diffusion Model

P(light)
$n \cdot \ell$
normals n
shading
rendering
∇_τ
density τ
albedo ρ
color $\mathbf{c}$
P(camera)
NeRF
MLP$(\cdot;\theta)$
Backpropagate onto NeRF weights
$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
"a DSLR photo of a peacock on a surfboard"
Imagen
$\mathbf{z}_t, t \sim \mathcal{U}(0,1)$
Transformer
U-Net
$\hat{x}_\phi(\mathbf{z}_t|y;t)$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t) - \epsilon$
$\hat{\epsilon}_\phi(\mathbf{z}_t|y;t)$

Dreamfusion, Poole et al., 2022

Prior Works: Score Distillation Sampling (SDS)

- Allows interjecting priors from pre-trained 2D model



Propagate Gradients to the 3D representation

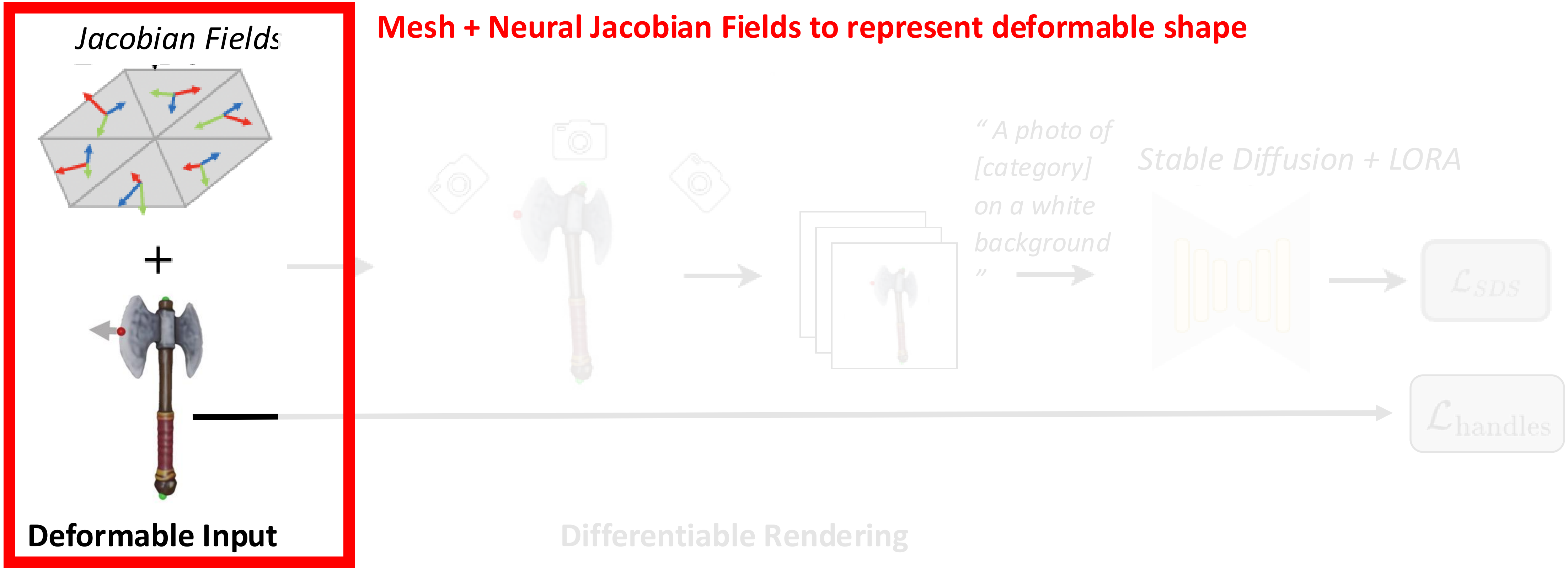
Dreamfusion, Poole et al., 2022

As-Plausible-As-Possible Deformation

Input

APAP, Yoo et al., CVPR 2024

As-Plausible-As-Possible Deformation



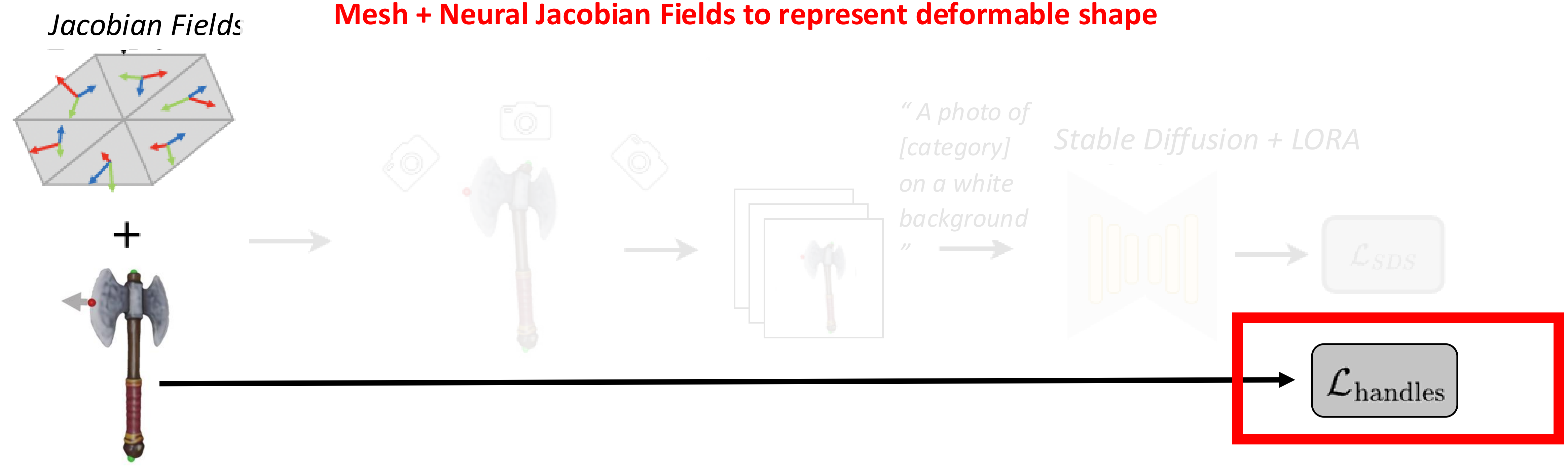
Mesh + Neural Jacobian Fields to represent deformable shape

APAP, Yoo et al., CVPR 2024

As-Plausible-As-Possible Deformation

- STAGE 1: Deform via Geometric Optimization only

Mesh + Neural Jacobian Fields to represent deformable shape


Jacobian Fields
+
" A photo of [category] on a white background "
Stable Diffusion + LORA
$\mathcal{L}_{SDS}$
$\mathcal{L}_{\text{handles}}$

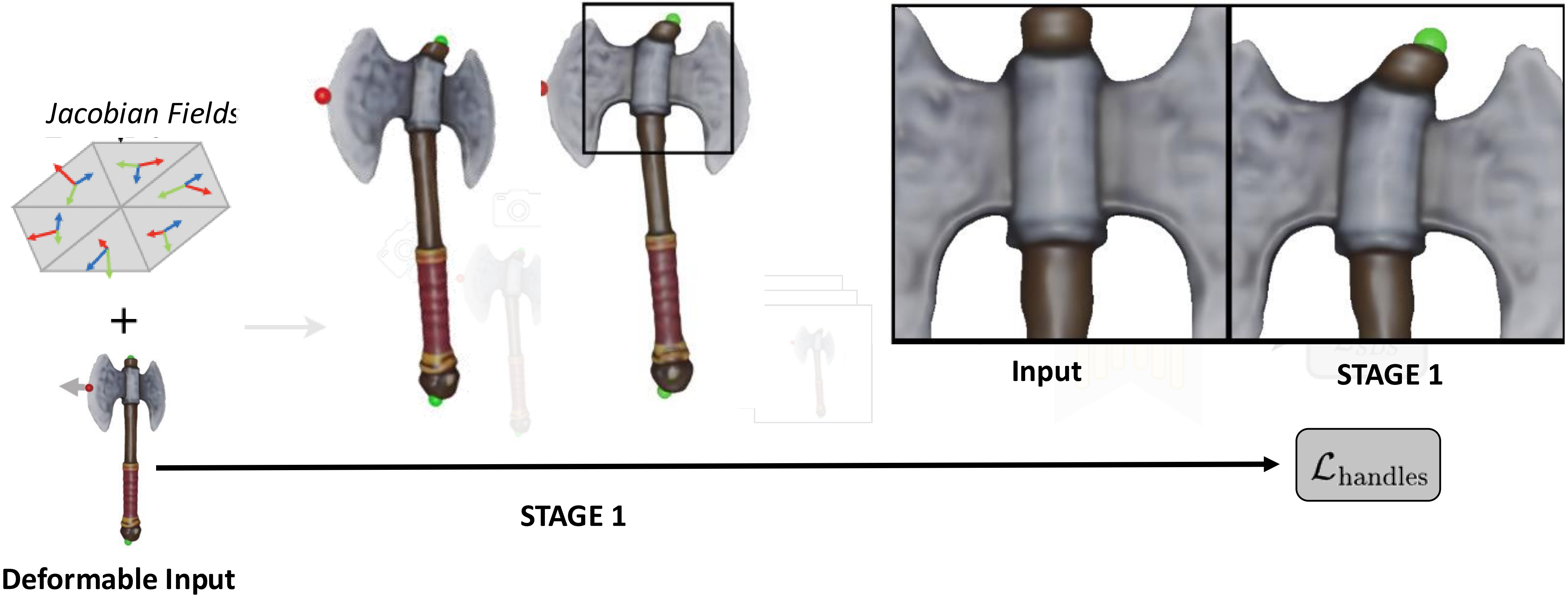
Deformable Input

Adobe

APAP, Yoo et al., CVPR 2024

As-Plausible-As-Possible Deformation

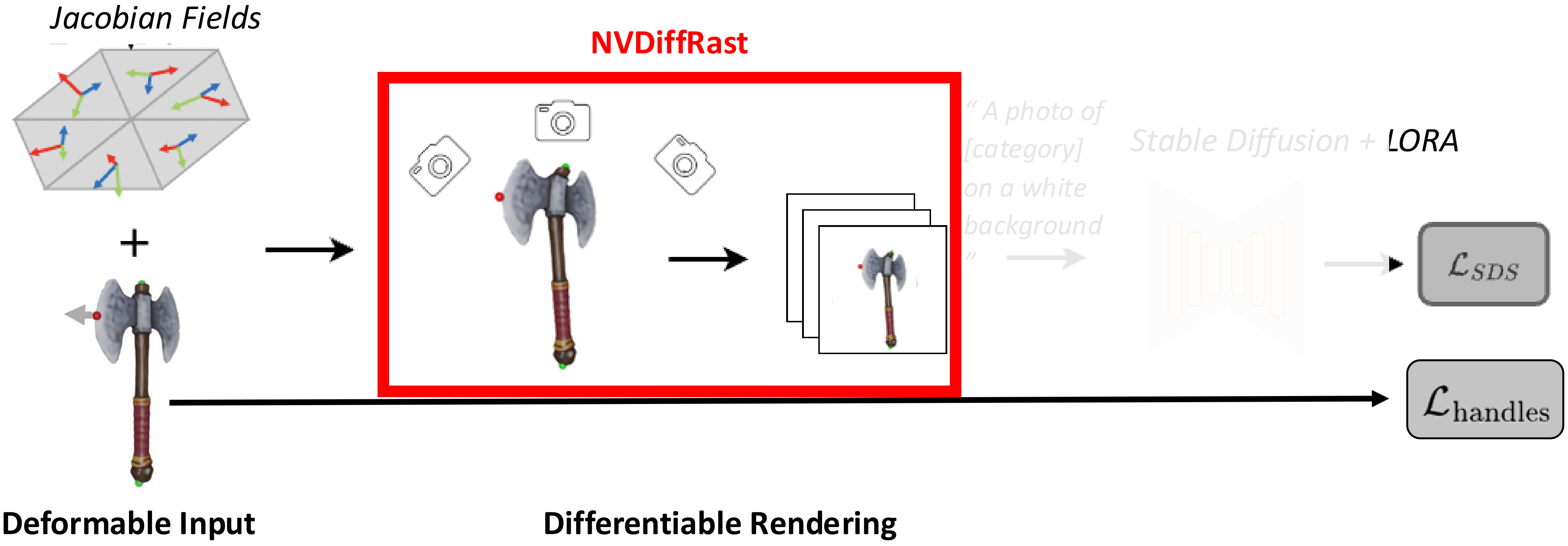
- STAGE 1: Deform via Geometric Optimization only



APAP, Yoo et al., CVPR 2024

As-Plausible-As-Possible Deformation

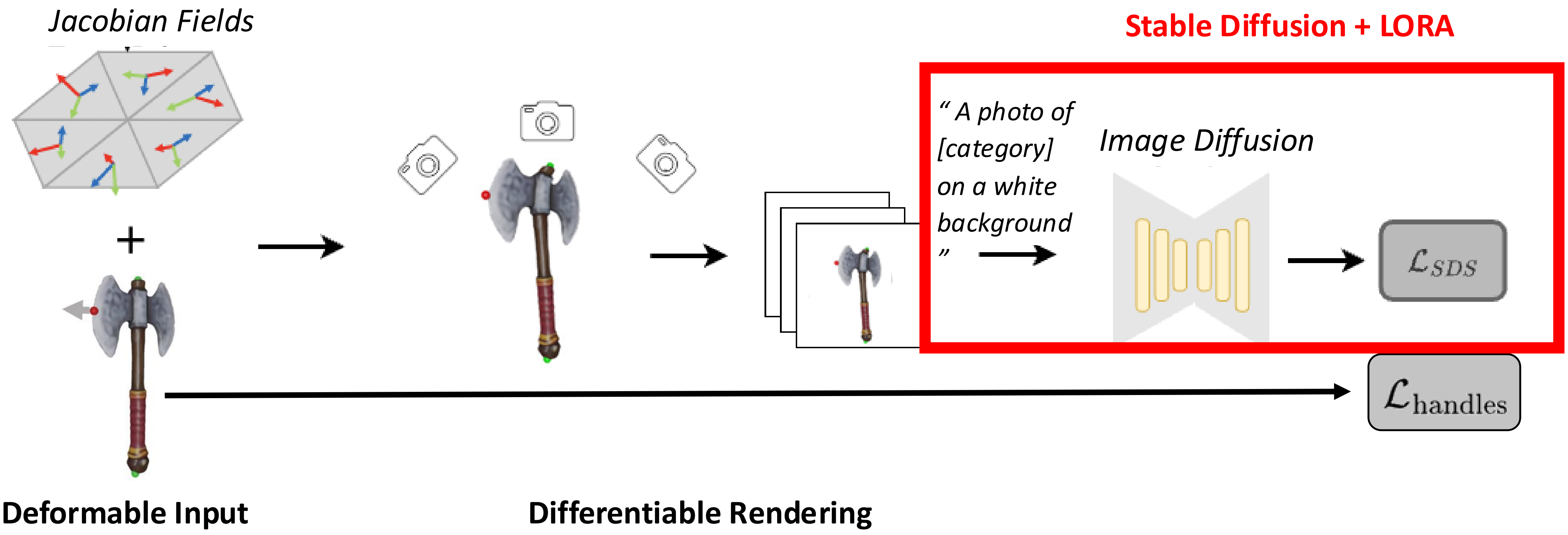
- STAGE 1: Deform via Geometric Optimization only
- STAGE 2: Project to “plausible” using SDS


Jacobian Fields
+
NVDiffRast
“ A photo of [category] on a white background ”
Stable Diffusion + LORA
$\mathcal{L}_{SDS}$
$\mathcal{L}_{\text{handles}}$
Deformable Input
Differentiable Rendering

APAP, Yoo et al., CVPR 2024

As-Plausible-As-Possible Deformation

- STAGE 1: Deform via Geometric Optimization only
- STAGE 2: Project to “plausible” using SDS



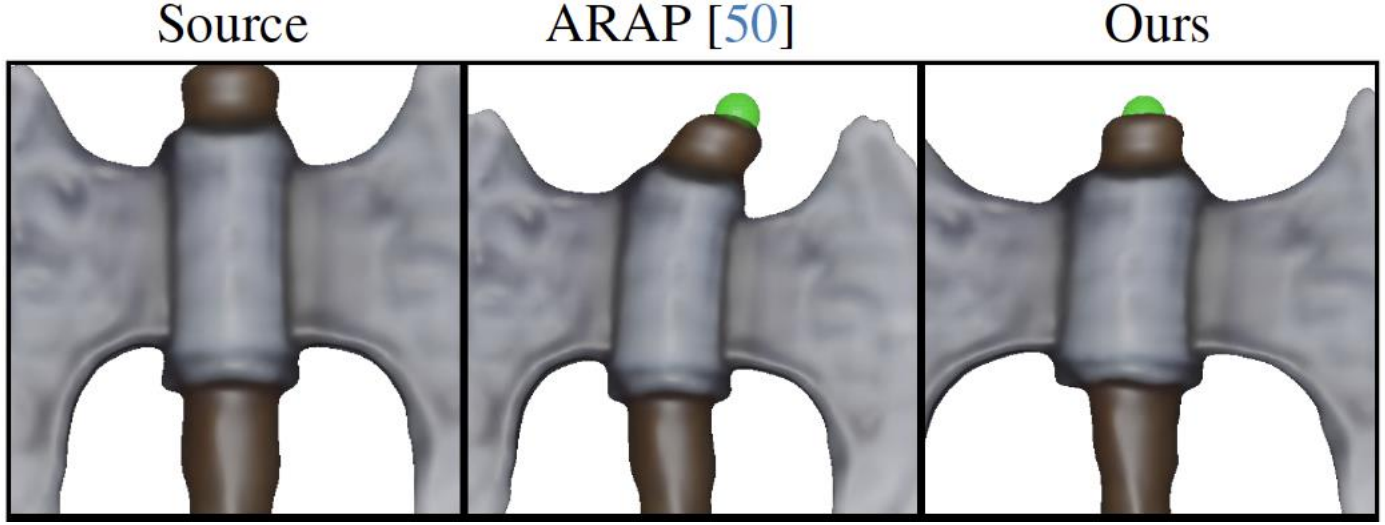
Jacobian Fields
+
Deformable Input
Differentiable Rendering
Stable Diffusion + LORA
“ A photo of [category] on a white background ”
Image Diffusion
$\mathcal{L}_{SDS}$
$\mathcal{L}_{\text{handles}}$

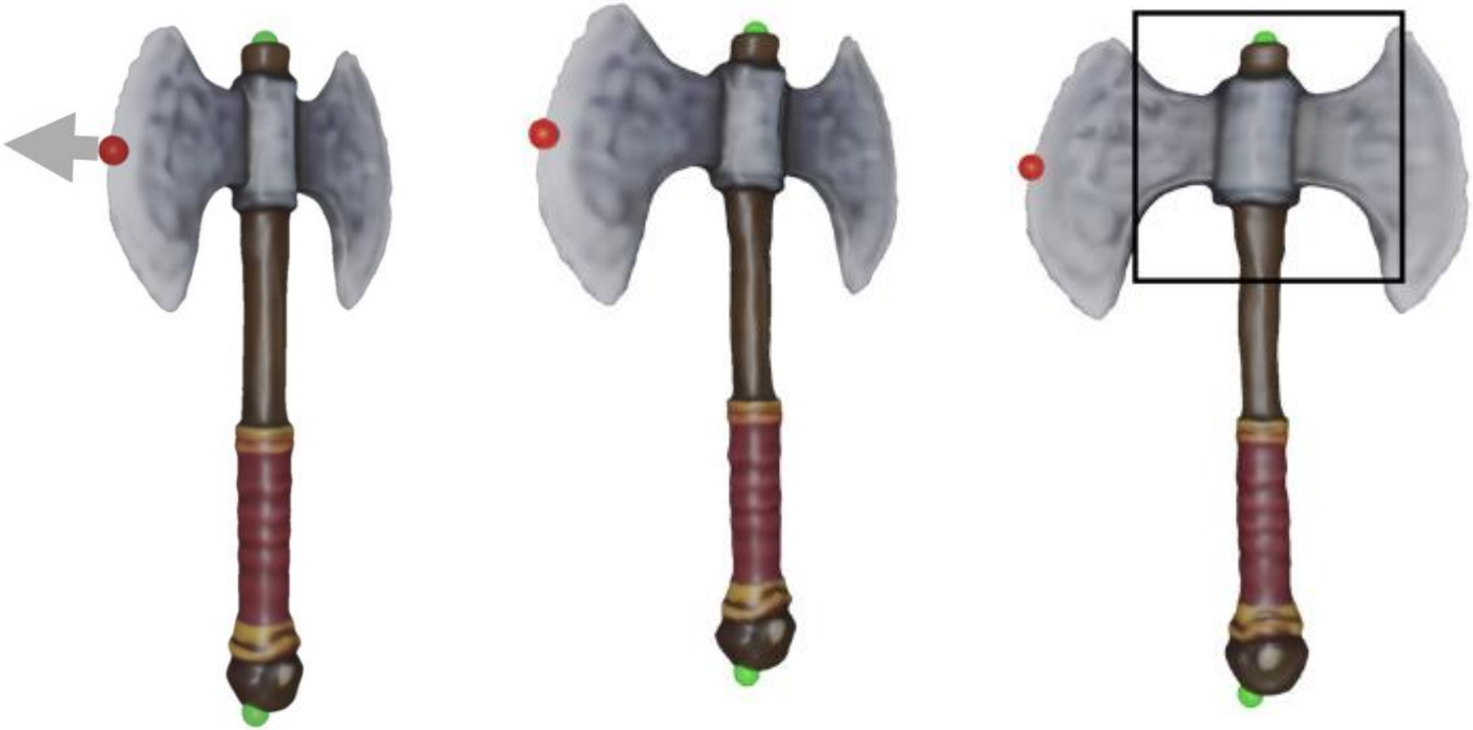
Adobe

APAP, Yoo et al., CVPR 2024

As-Plausible-As-Possible Deforn

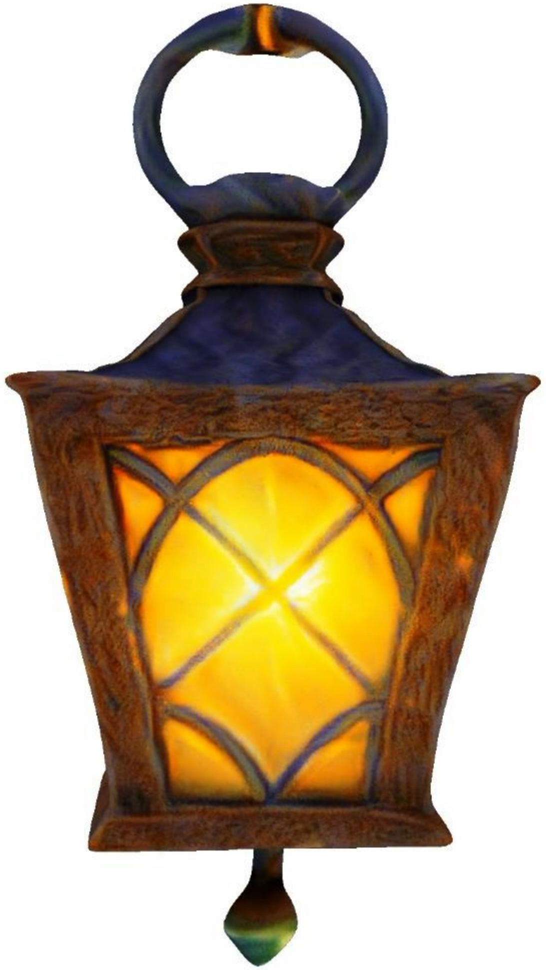
- STAGE 1: Deform via Geometric Optimization
- STAGE 2: Project to “plausible” using SDS





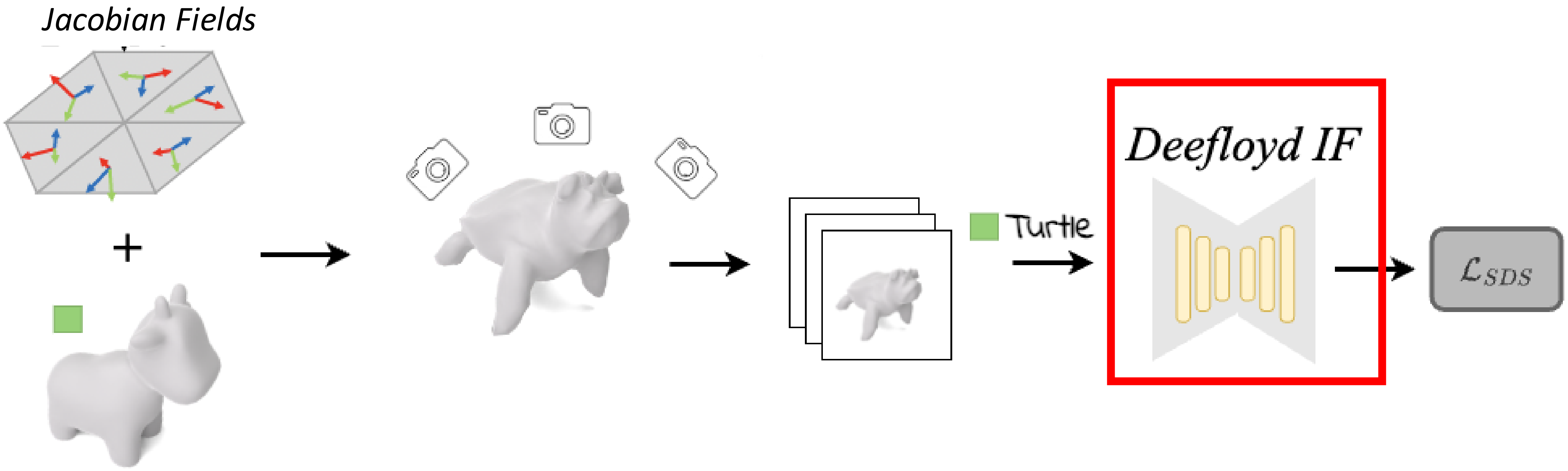
APAP, Yoo et al., CVPR 2024

Source Mesh



Deforming with Language Controls

- Condition on text



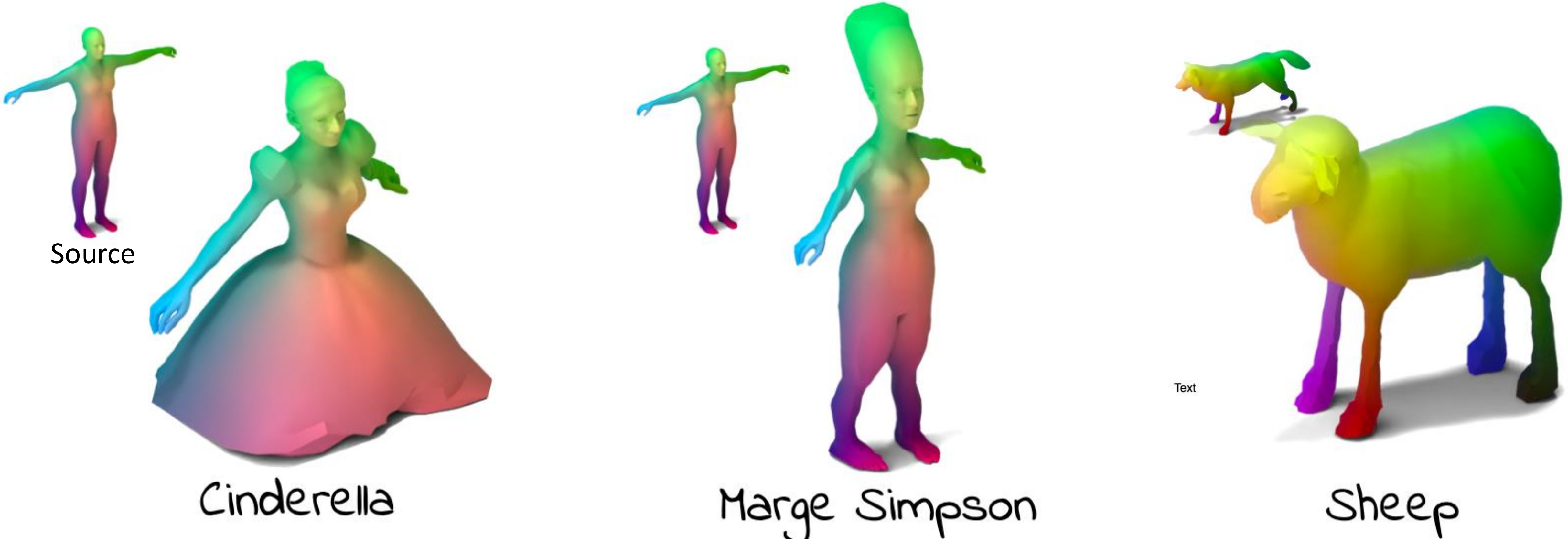
Deformable Input

Differentiable Rendering

SDS Loss

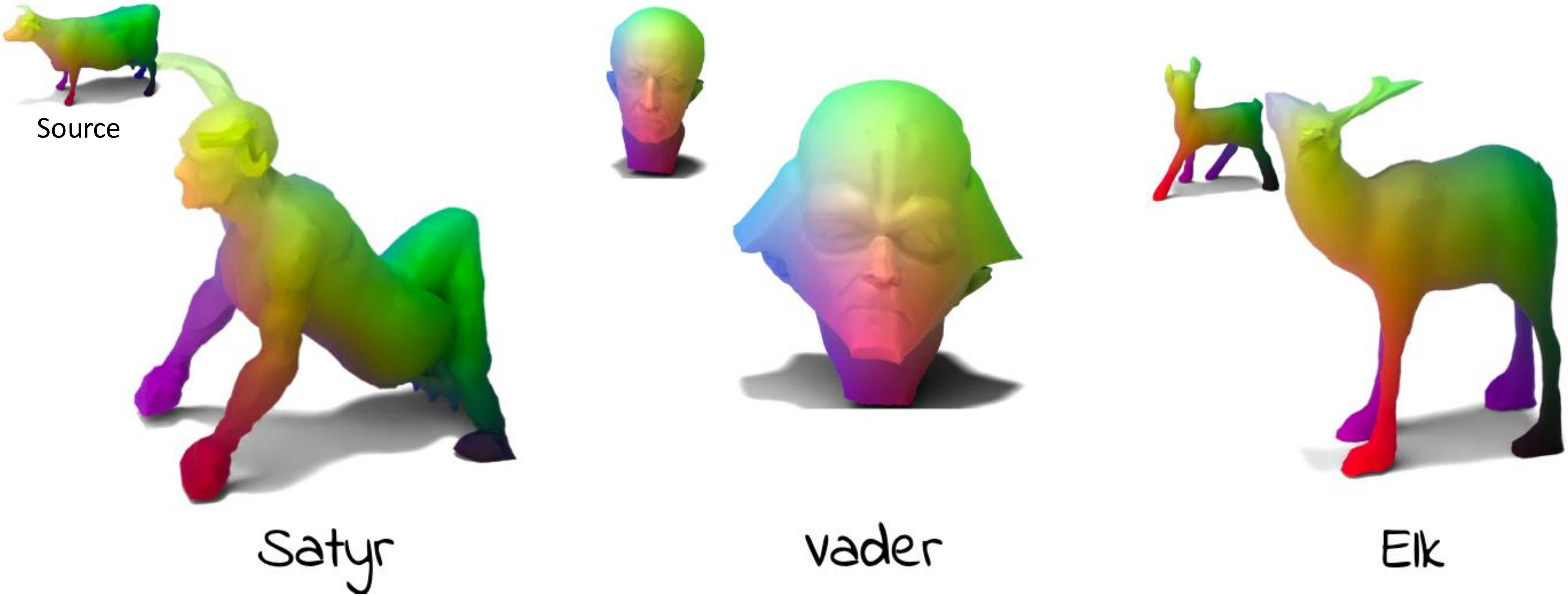
MeshUp, Kim et al., 3DV 2025

Deforming with Language Controls



MeshUp, Kim et al., 3DV 2025

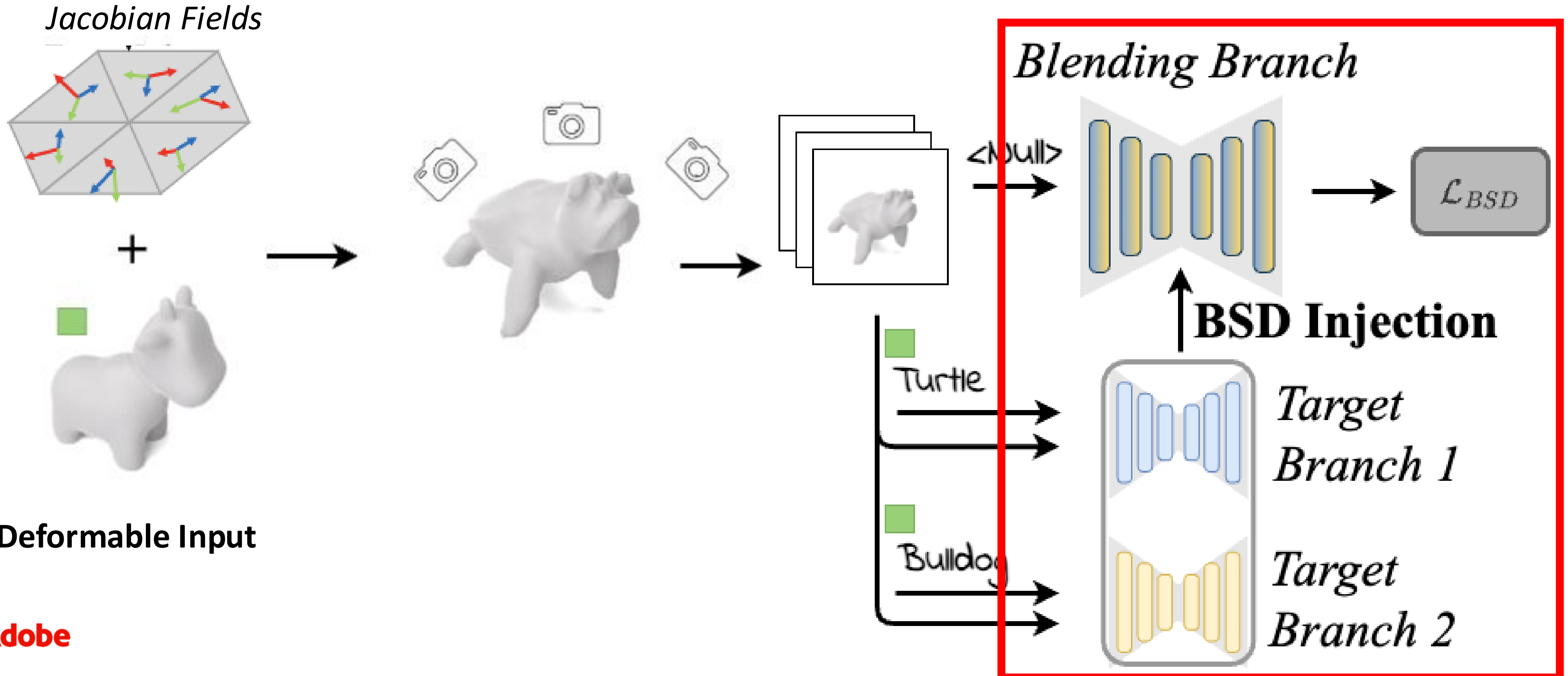
Deforming with Language Controls



MeshUp, Kim et al., 3DV 2025

Deforming with Language Controls: Multi-Target

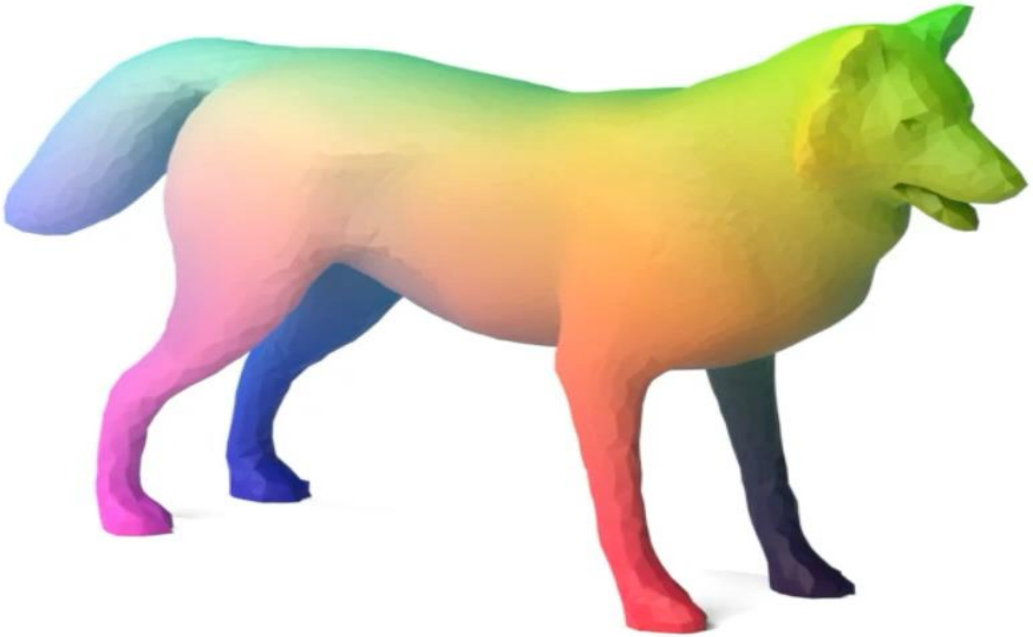
- Inject weighted features into attention module



et al., 3DV 2025

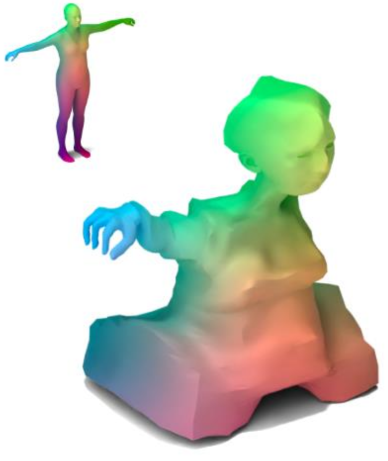
Adobe

Deforming with Language Controls: Multi-Target

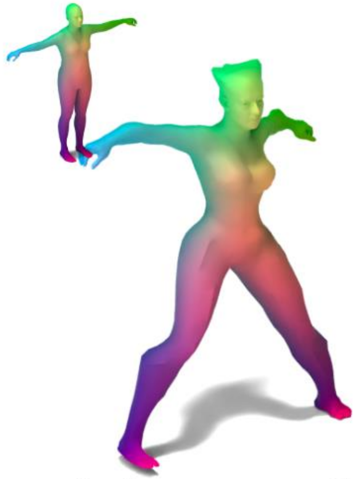


MeshUp, Kim et al., 3DV 2025

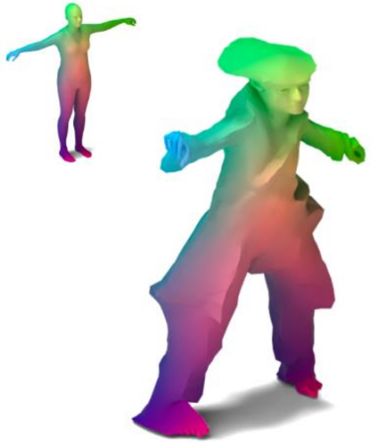
Deforming with Language Controls: Multi-Target



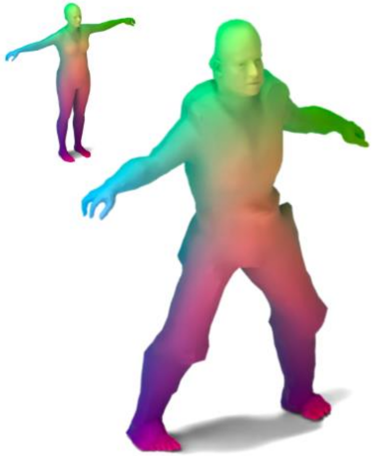
WALLE-E 60%
Eve 30%



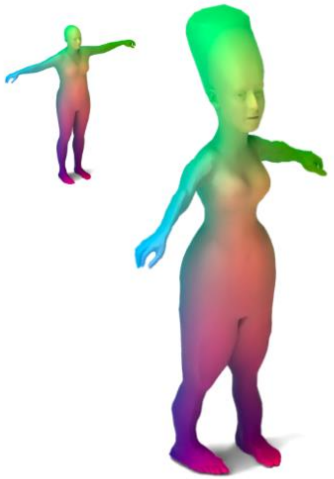
Cat Woman 60%
Wonder Woman 40%



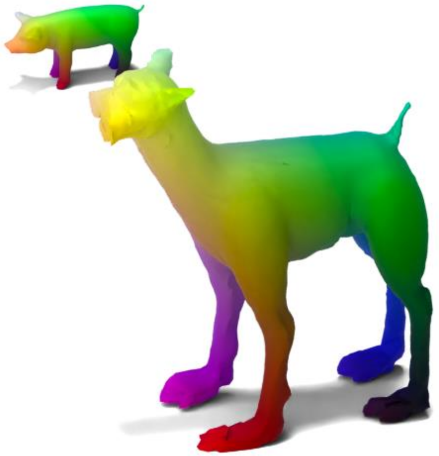
Jack Sparrow 100%



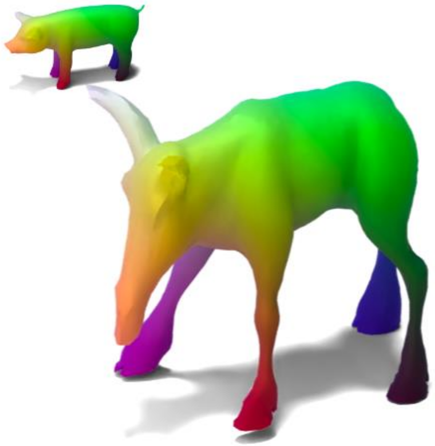
Luke Skywalker 60%
Dath Vader 40%



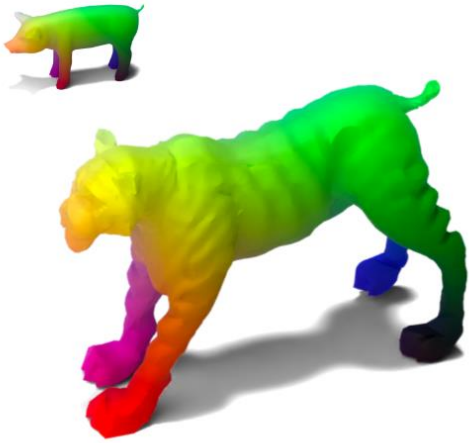
Marge Simpson 100%



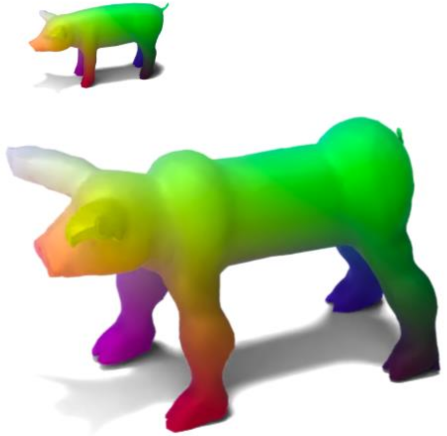
Baku 100%



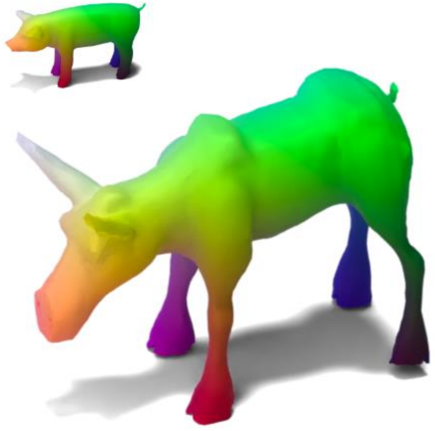
Gazelle 70%
Springbok 30%



Tiger 70%
Leopard 50%



Ra 70%
Anubis 30%

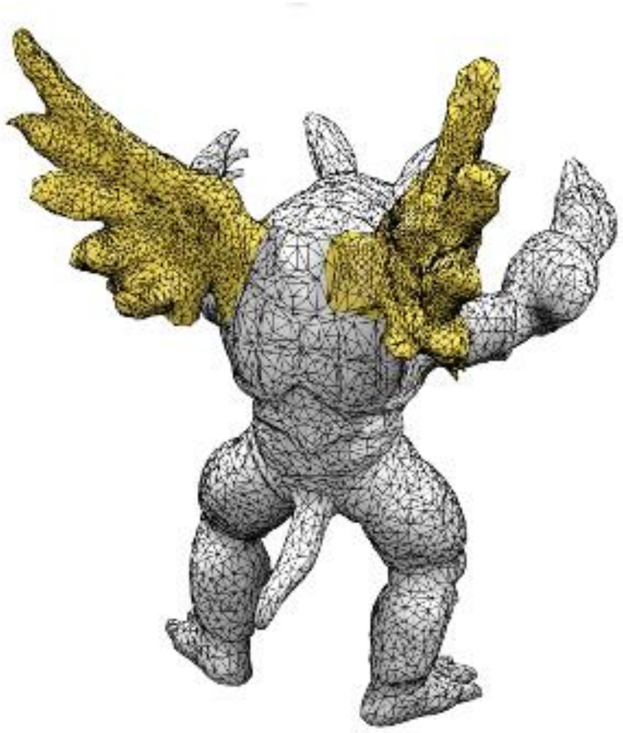


Warthog 70%
Wildebeast 30%

Overview

- Support mesh outputs (but use other representations as needed)
- Inspired by traditional workflows

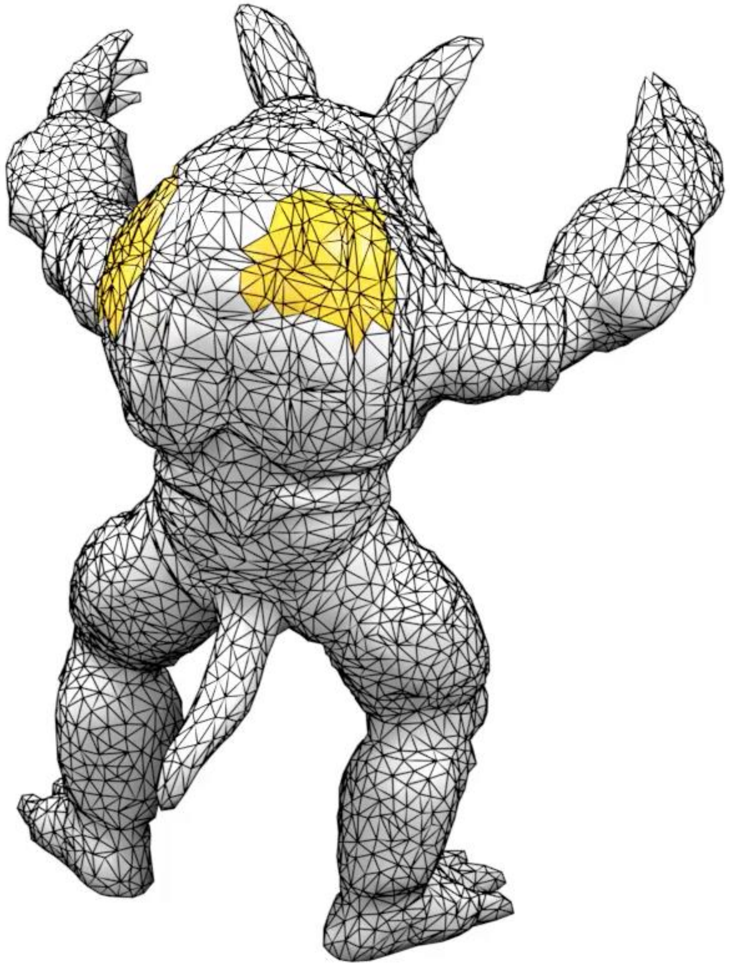
Neural **Deformation**



Generative **Scultping**

Generative **Detailization**

Generative Sculpting



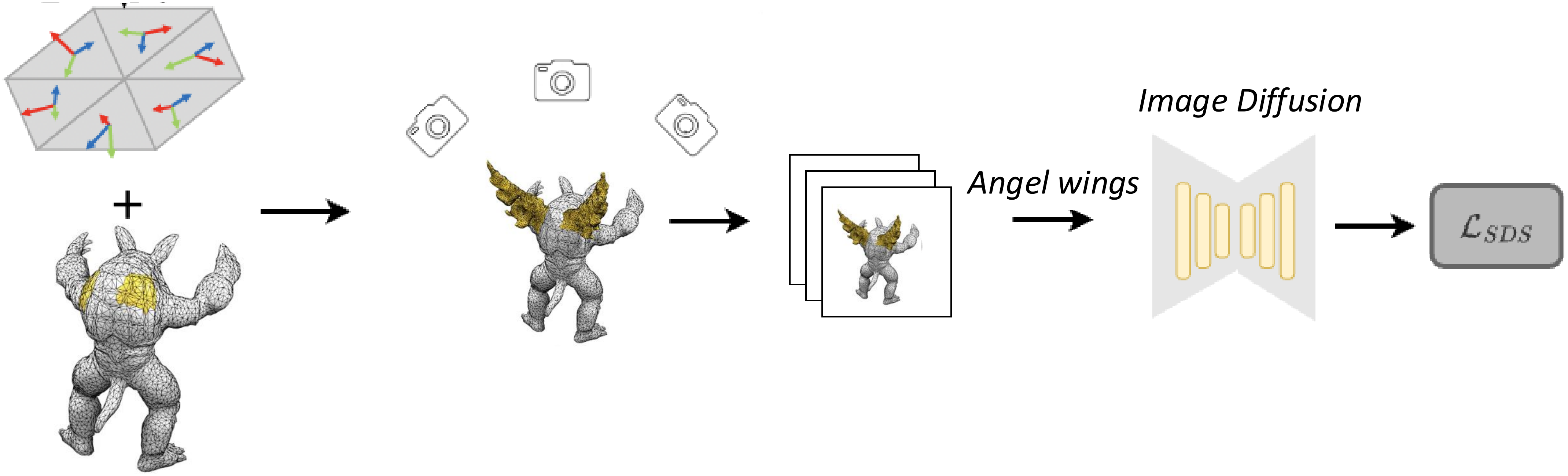
“Man with angel wings”

Adobe

MagicClay, Barda et al., SIGGRAPH A 2024

Generative Sculpting via Deformation

- Only deform highlighted area



Jacobian Fields

Image Diffusion

Angel wings

Deformable Input

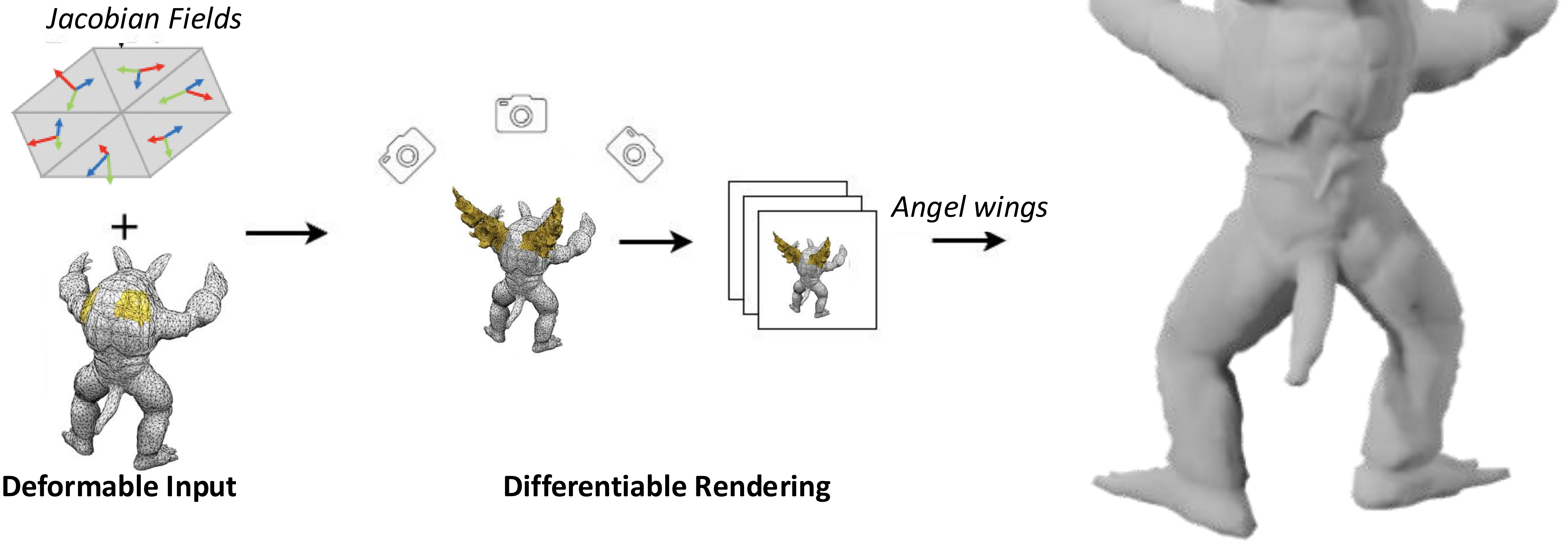
Differentiable Rendering

SDS Loss

MagicClay, Barda et al., SIGGRAPH A 2024

Generative Sculpting via Deformation

- Only deform highlighted area



Jacobian Fields

Deformable Input

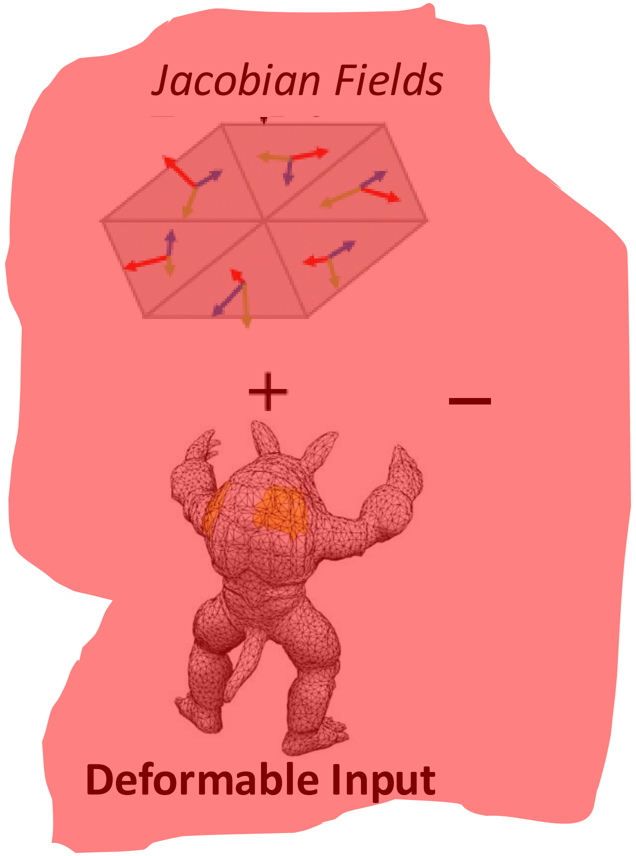
Differentiable Rendering

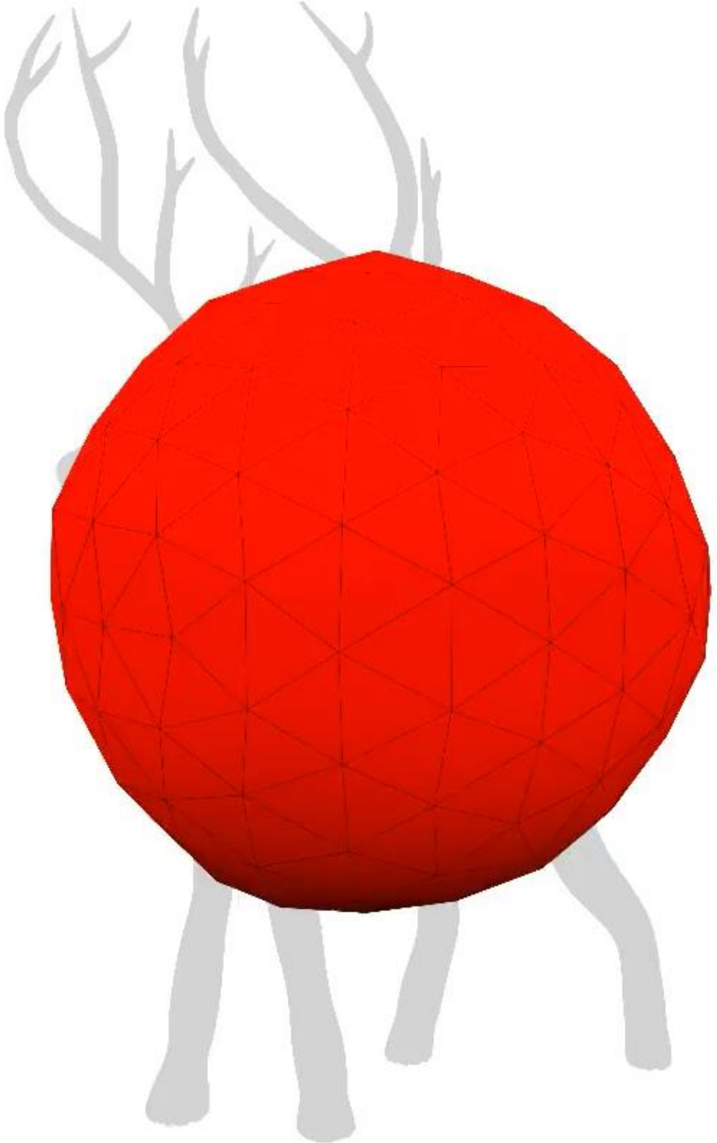
Angel wings

MagicClay, Barda et al., SIGGRAPH A 2024

Prior Work: Continuous Remeshing

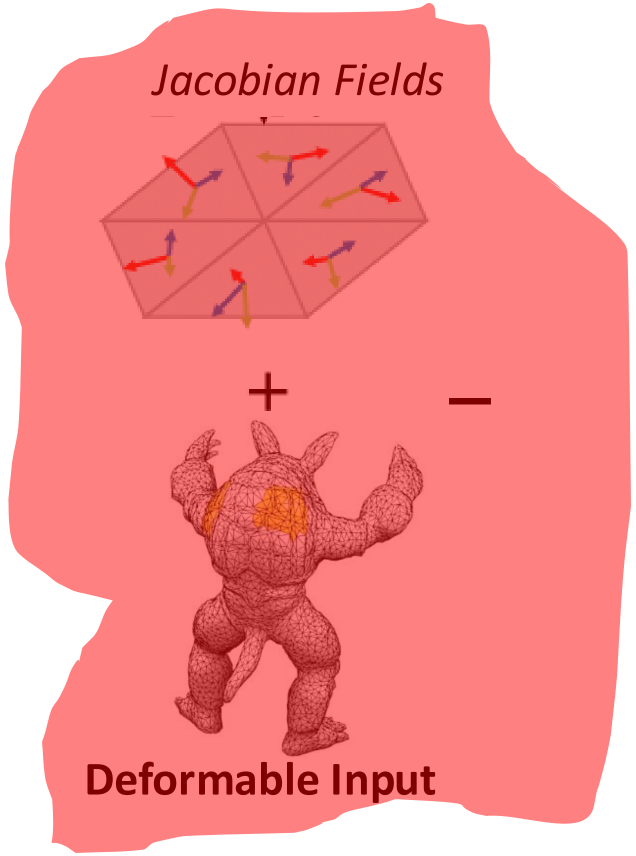
- Dynamic remeshing: allows to add more details



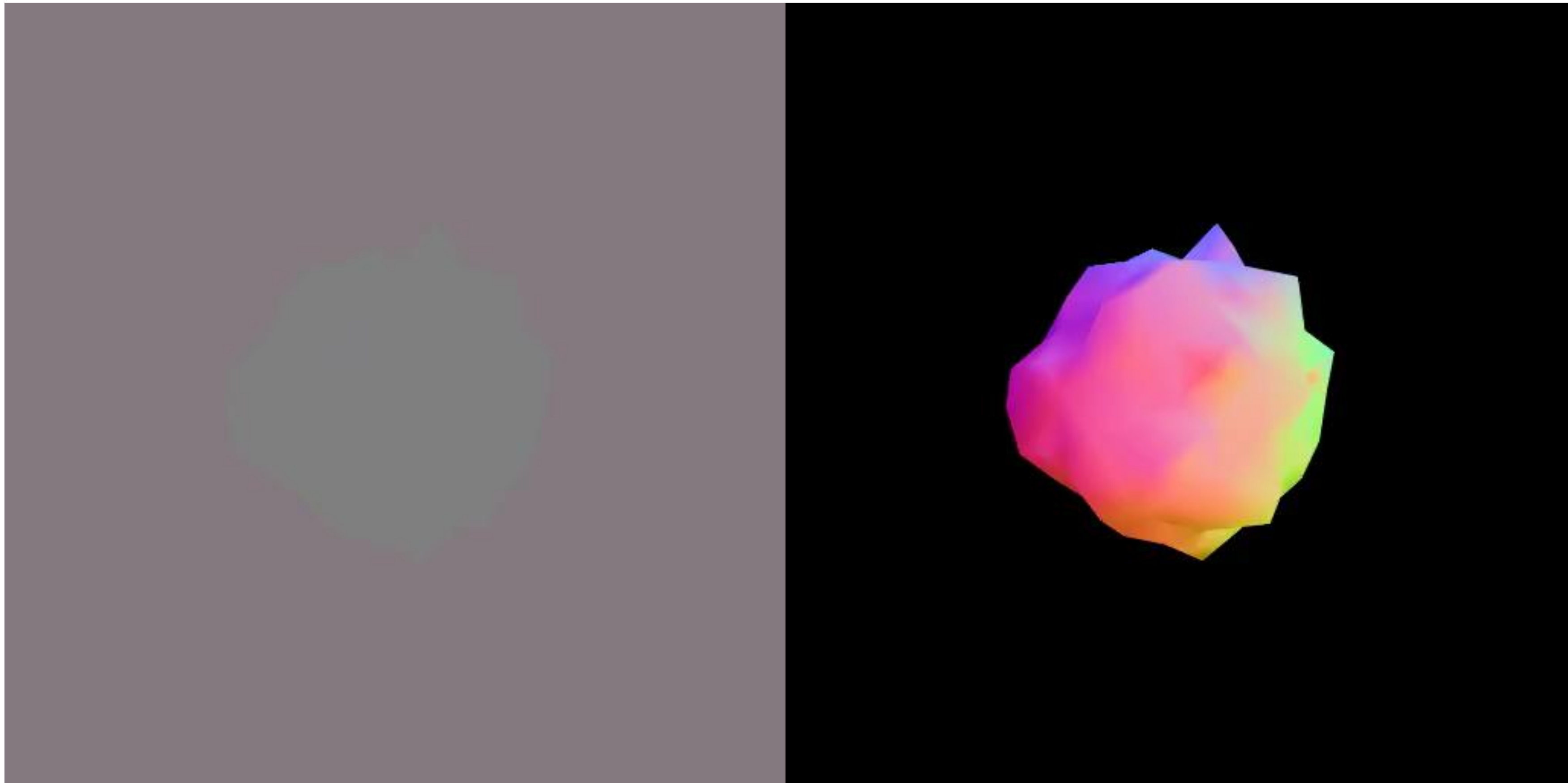


Prior Work: Continuous Remeshing

- Dynamic remeshing: allows to add more details
- Remeshing + SDS



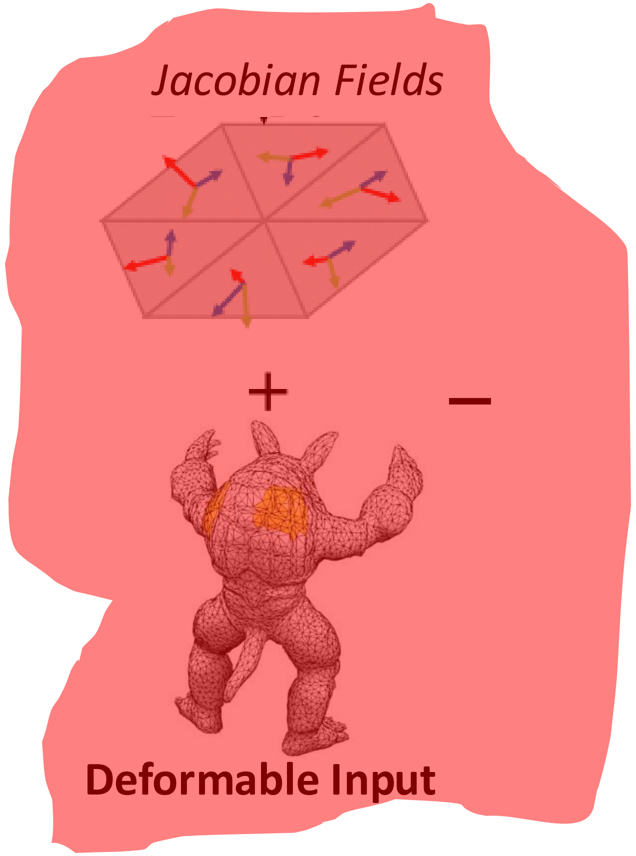
Prompt: “A deer”



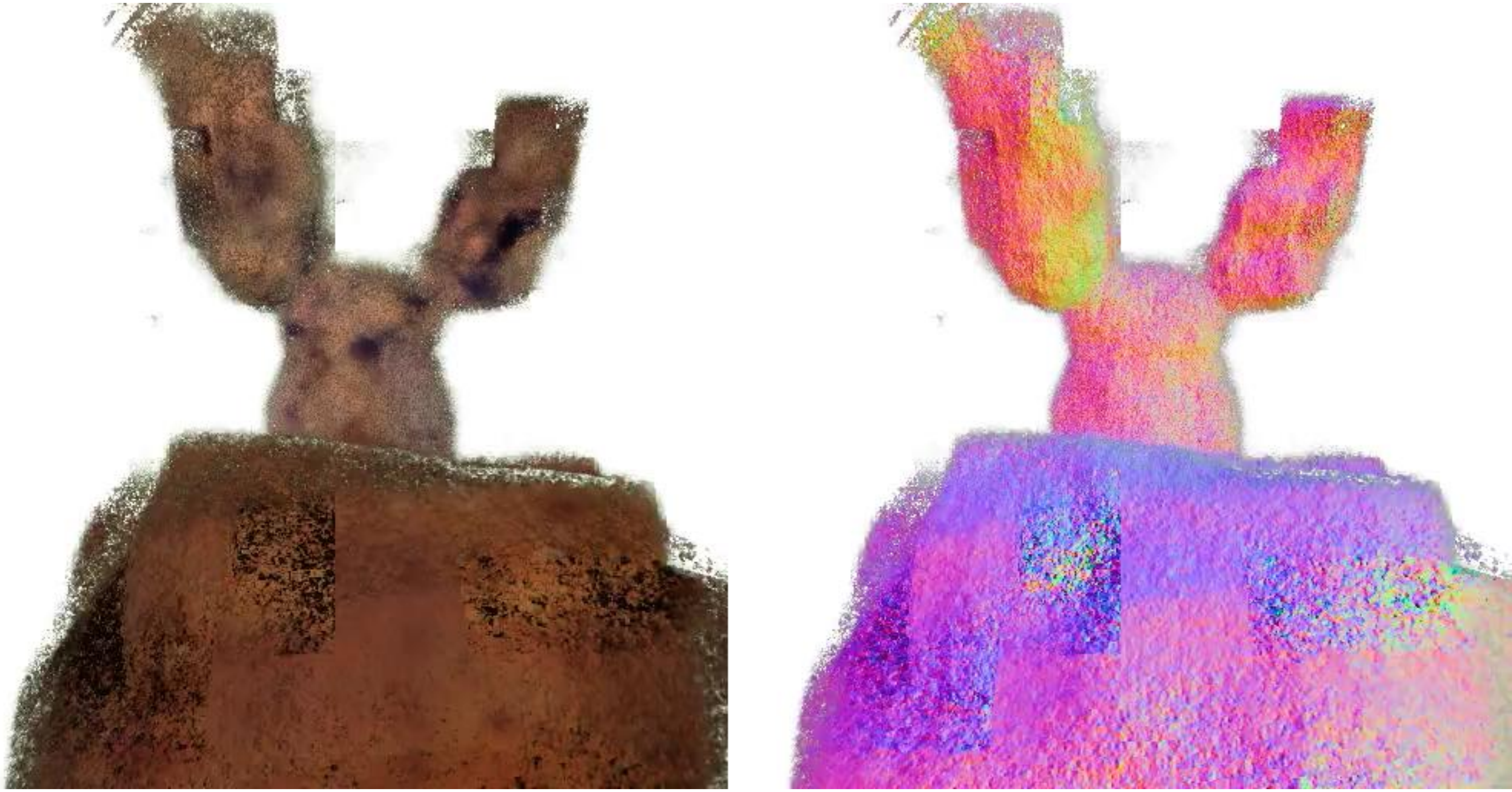
~~Prior Work: Continuous Remeshing~~

- Use continuous remeshing instead?

~~▪ Remeshing~~ + SDF + SDS

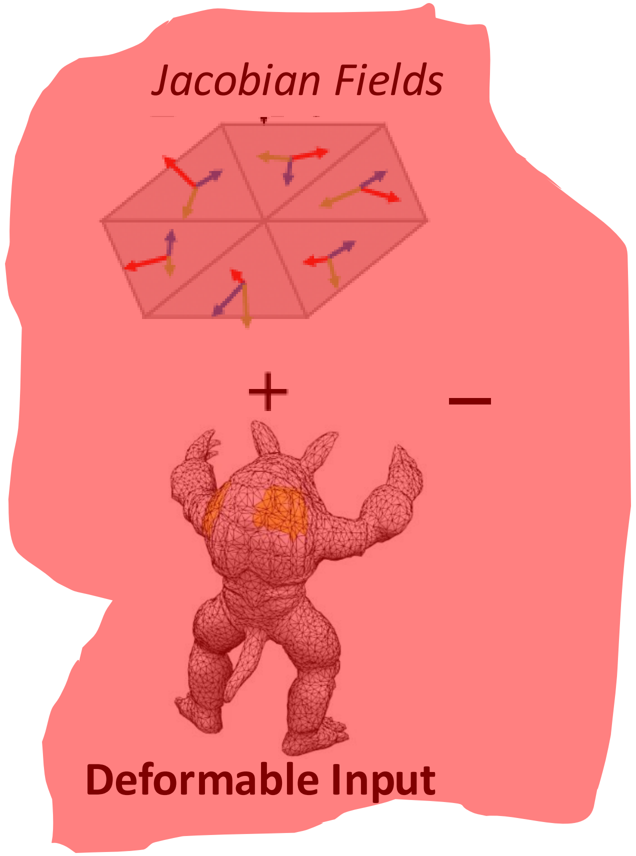


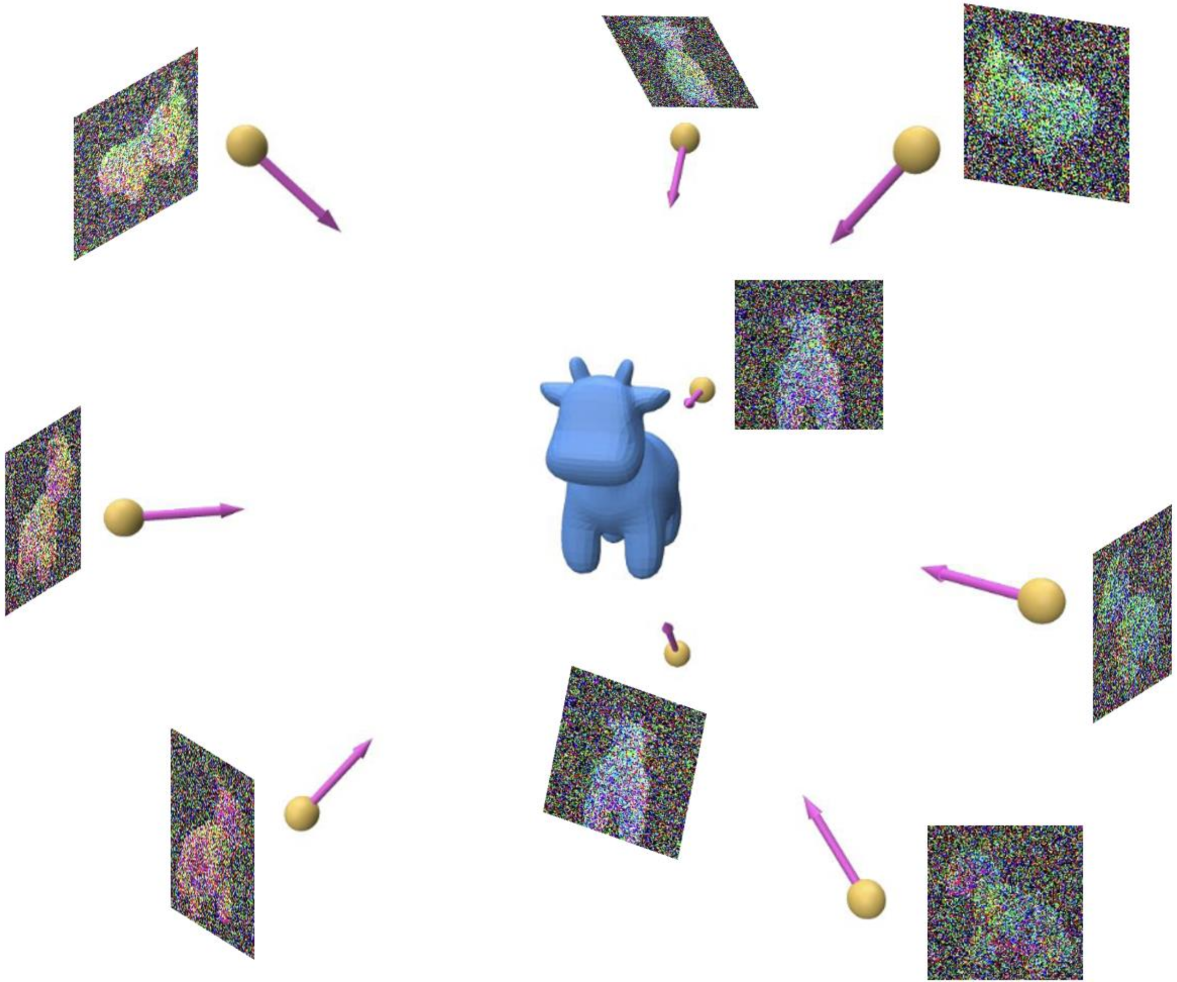
Prompt: “A deer”



Controlled Experiment: Continuous Remeshing vs SDF

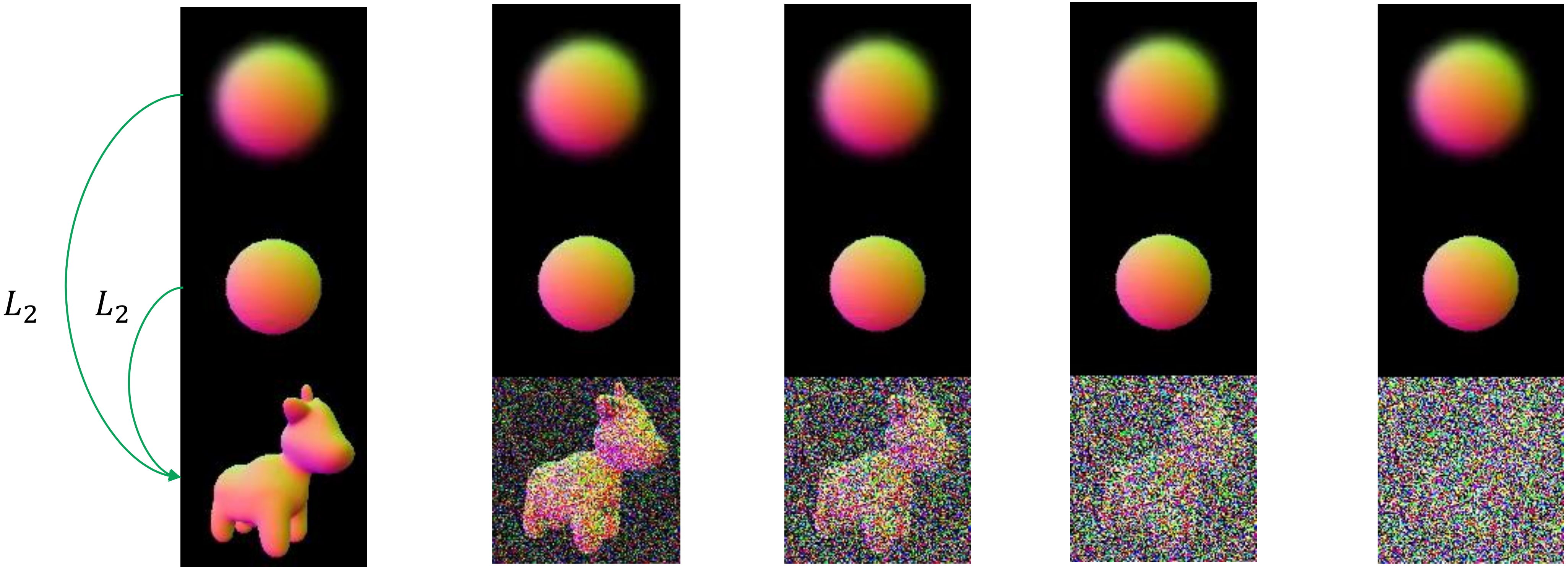
- Reconstruct from renderings with different levels of noise





Controlled Experiment: Continuous Remeshing vs SDF

- Brittle



SDF Fitting

Mesh Fitting with Remeshing

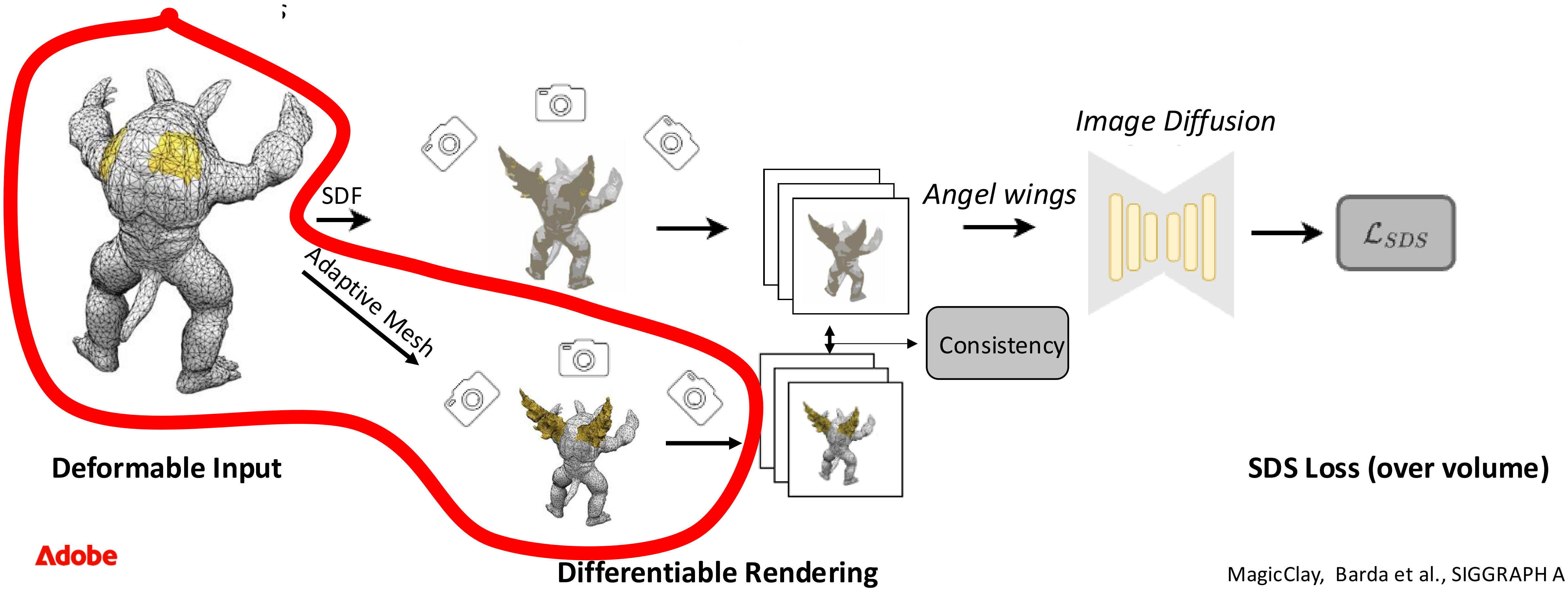
Noise Magnitude

Adobe

MagicClay, Barda et al., SIGGRAPH A 2024

Generative Sculpting via Deformation: MagicClay

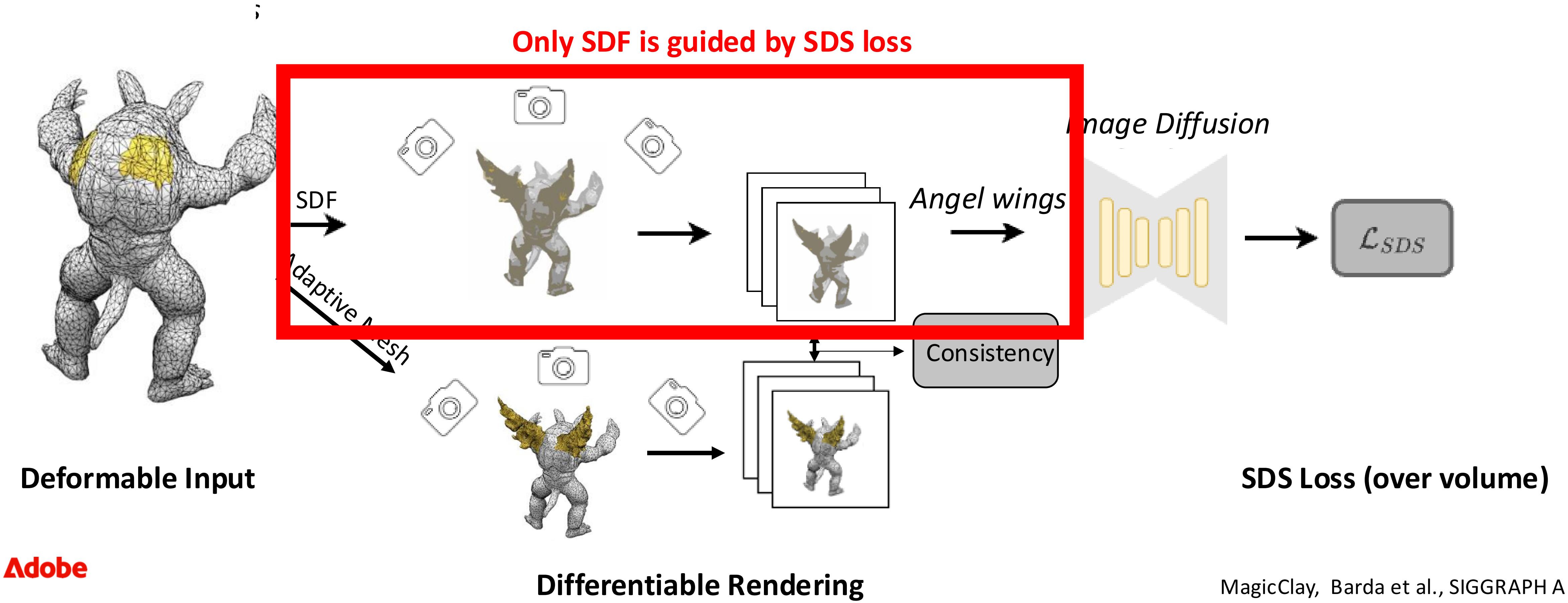
- **Dynamic remeshing: allows to add more details**
- Hybrid (SDF+Mesh) representation



MagicClay, Barda et al., SIGGRAPH A 2024

Generative Sculpting via Deformation: MagicClay

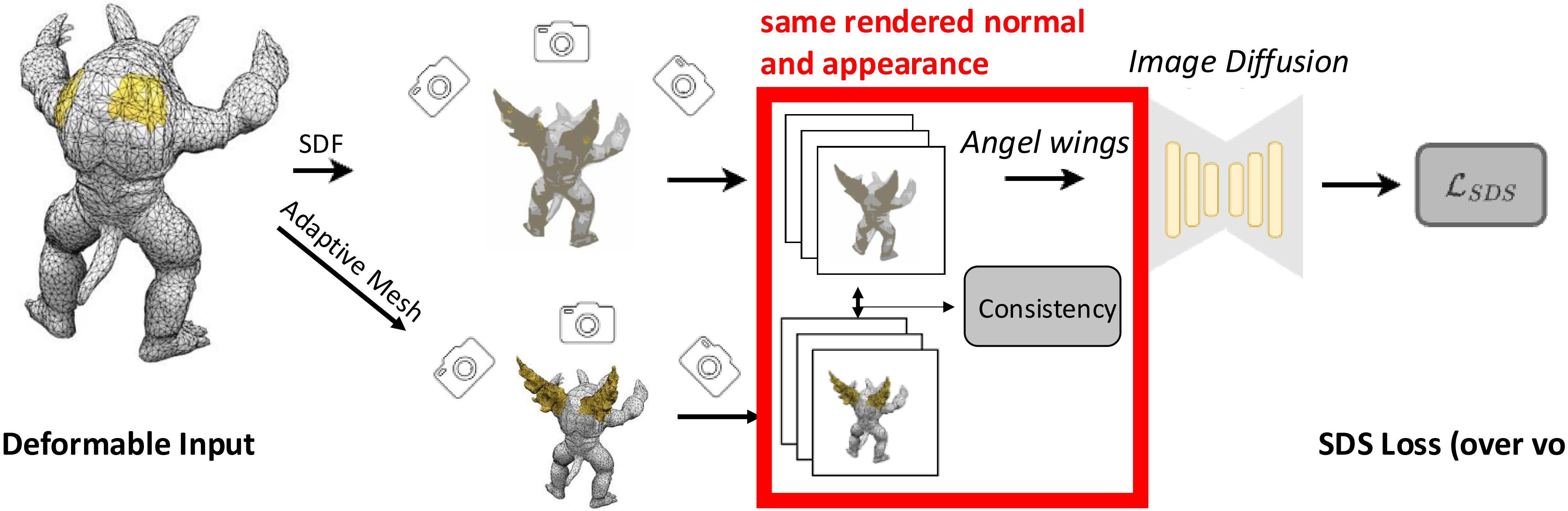
- Dynamic remeshing: allows to add more details
- **Hybrid (SDF+Mesh) representation**



MagicClay, Barda et al., SIGGRAPH A 2024

Generative Sculpting via Deformation: MagicClay

- Dynamic remeshing: allows to add more details
- **Hybrid (SDF+Mesh) representation**



Added consistency loss: same rendered normal and appearance

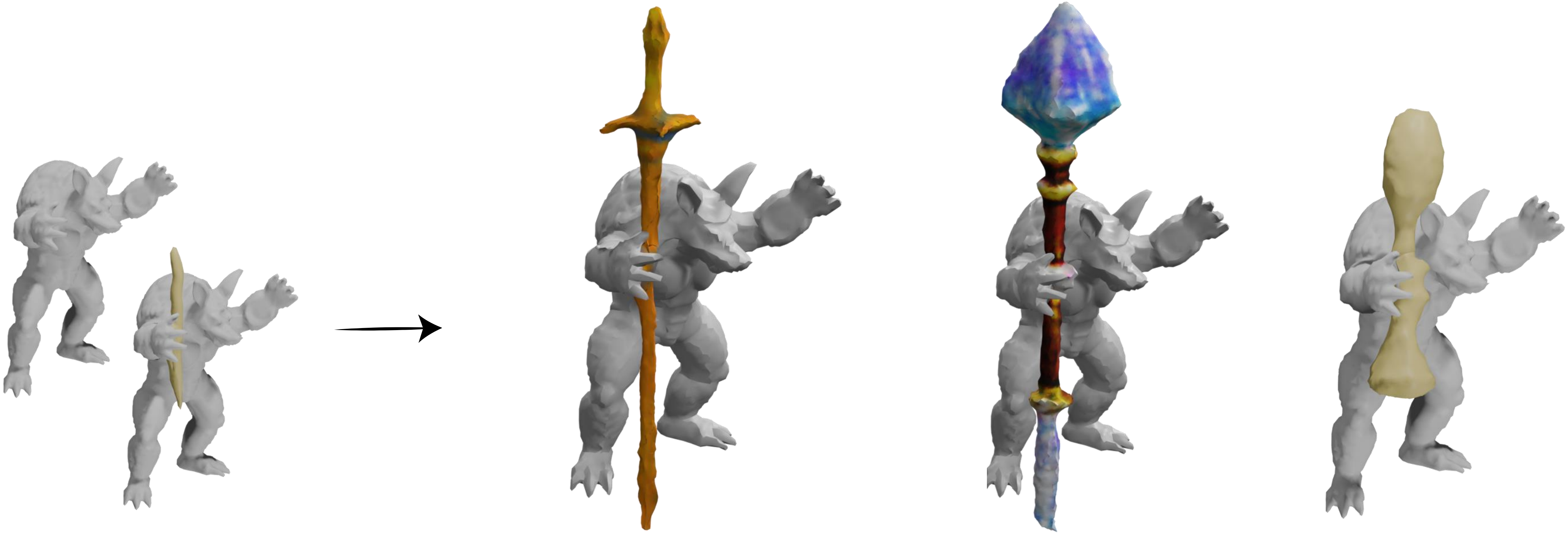
Deformable Input

Differentiable Rendering

SDS Loss (over volume)

MagicClay, Barda et al., SIGGRAPH A 2024

Generative Sculpting via Deformation: MagicClay



“Man holding a… …knight’s sword” …Wizard staff” …maraca”

MagicClay, Barda et al., SIGGRAPH A 2024

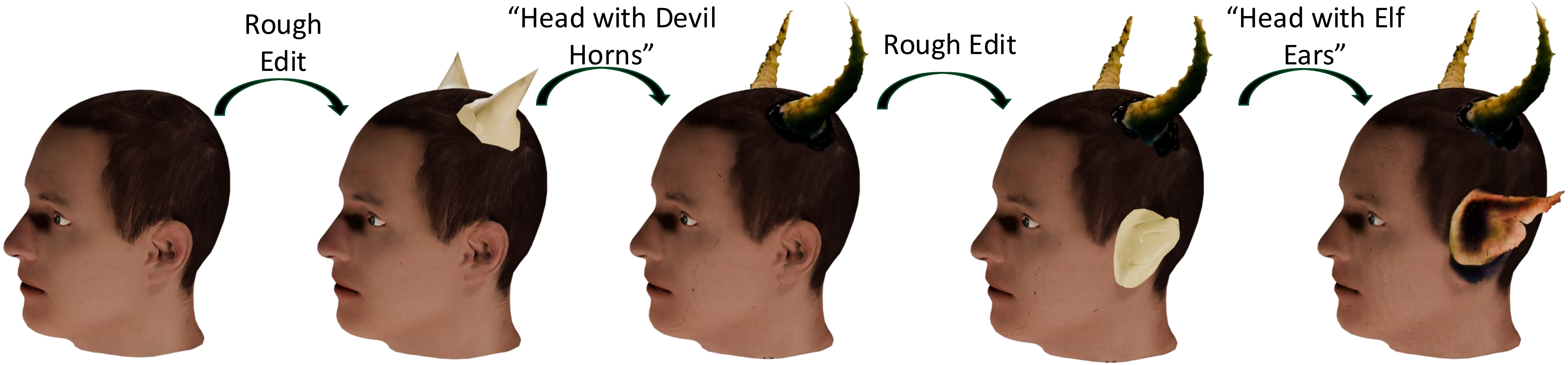
SDS Limitations

- Brittle (requires careful tuning of hyperparameters)
- **Inconsistent guidance from different views**

MagicClay, Barda et al., SIGGRAPH A 2024

SDS Limitations

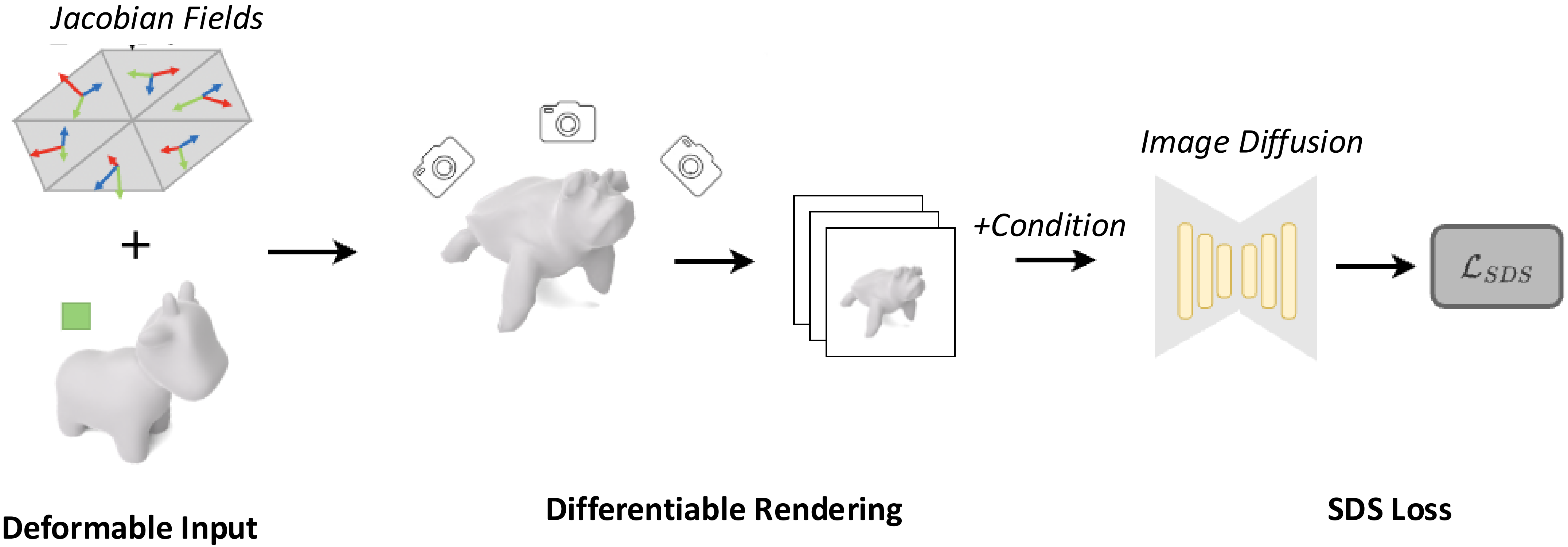
- Brittle (requires careful tuning of hyperparameters)
- **Inconsistent guidance from different views**
- **Slow (hours per iteration)**



MagicClay, Barda et al., SIGGRAPH A 2024

Getting Rid of SDS

- Reminder: SDS pipeline



Jacobian Fields

+

Deformable Input

Differentiable Rendering

+Condition

Image Diffusion

$\mathcal{L}_{SDS}$

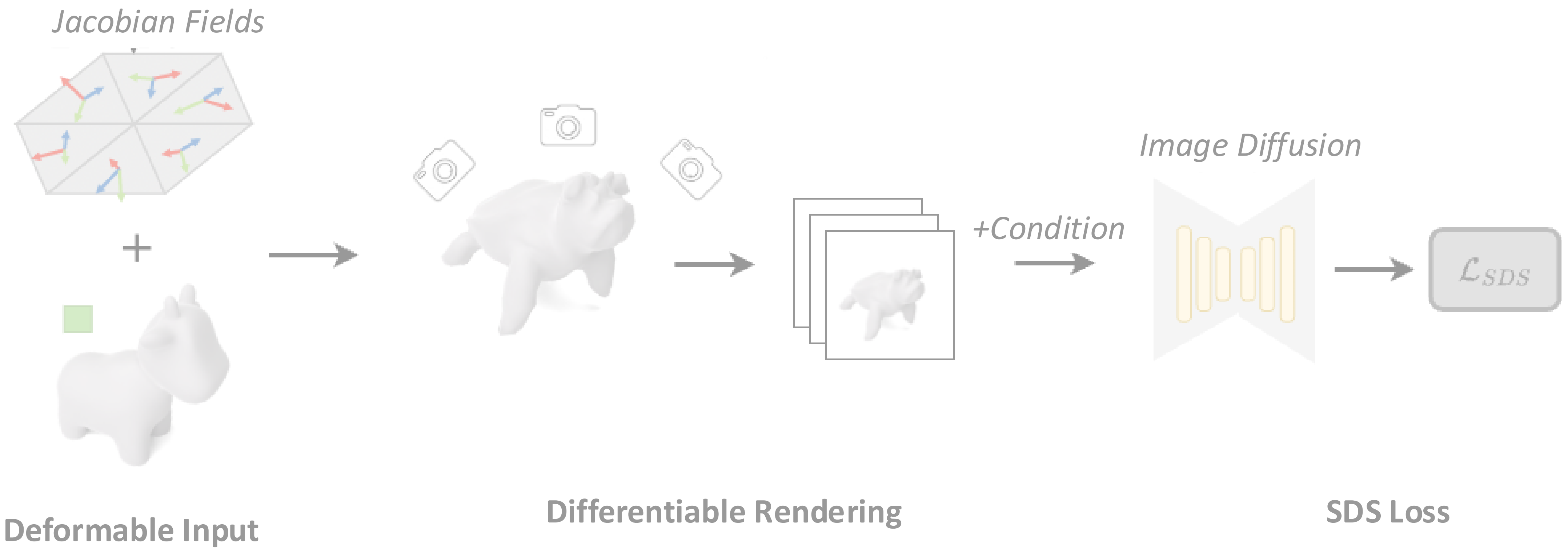
SDS Loss

Adobe

Instant3dit, Barda et al., In submission, 2025

Generative Sculpting via Multi-view Inpainting

- ~~Reminder: SDS pipeline~~

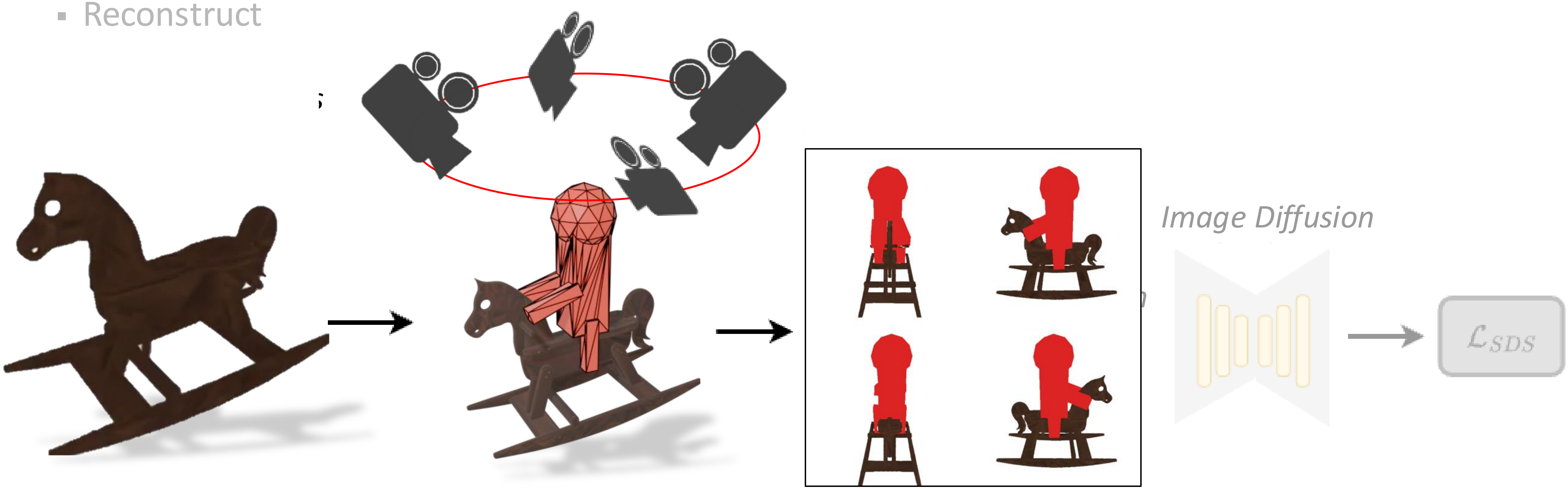


Adobe

Instant3dit, Barda et al., In submission, 2025

Generative Sculpting via Multi-view Inpainting

- In-paint multi-view
- Reconstruct

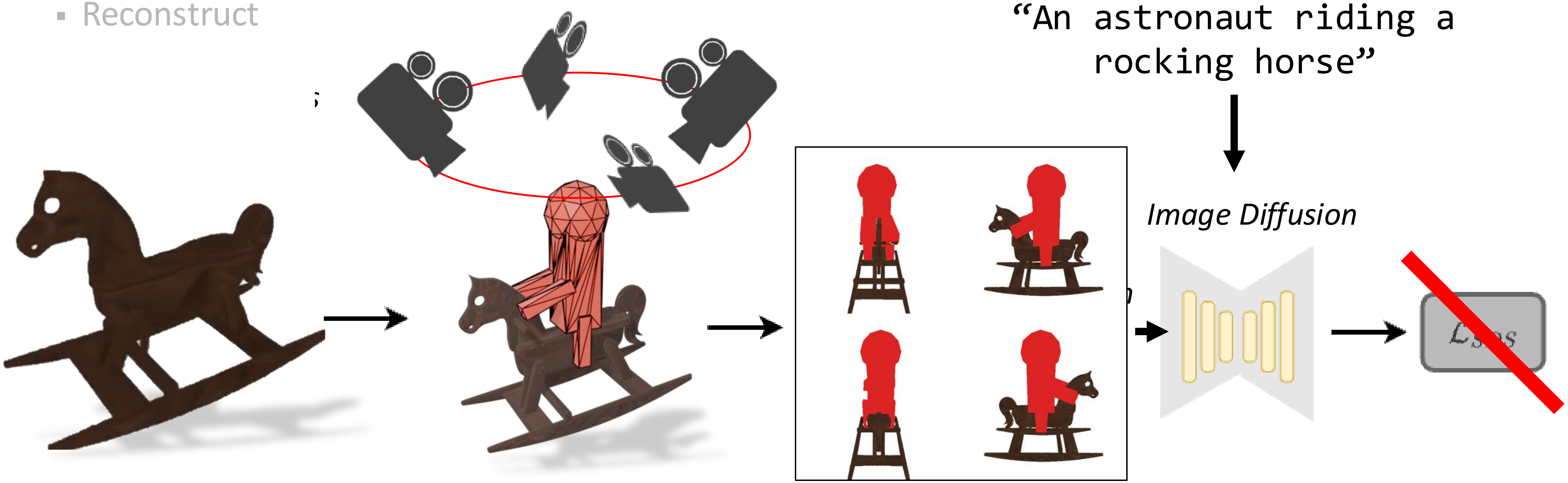


Input **~~Differentiable~~ Rendering (masks)** **SDS Loss**

Instant3dit, Barda et al., In submission, 2025

Generative Sculpting via Multi-view Inpainting

- In-paint multi-view
- Reconstruct



Input

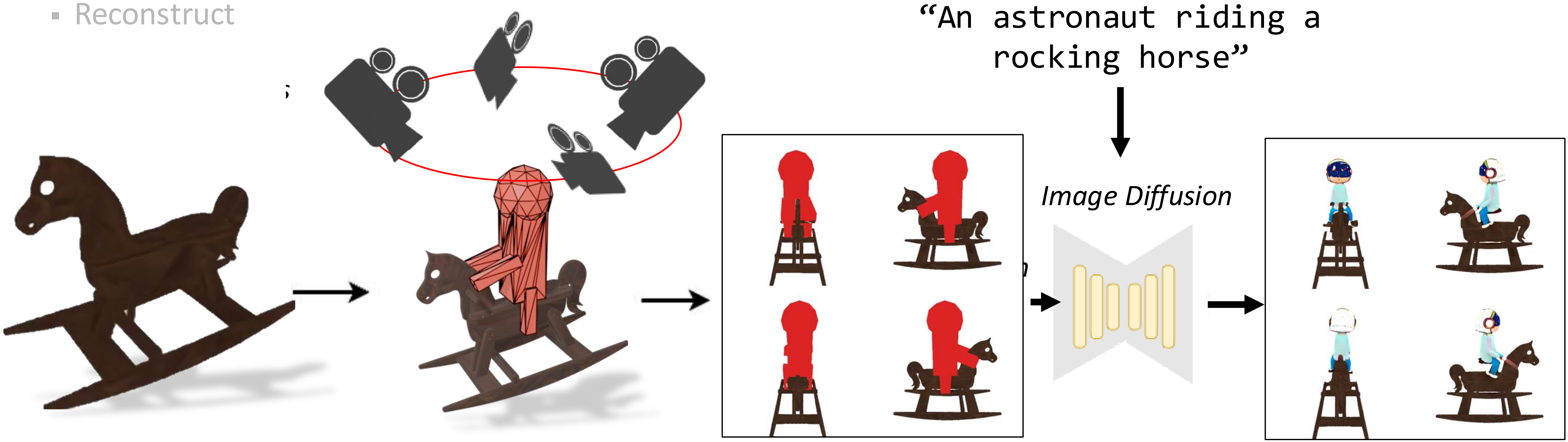
~~Differentiable~~ Rendering (masks)

~~SDS Loss~~

Instant3dit, Barda et al., In submission, 2025

Generative Sculpting via Multi-view Inpainting

- In-paint multi-view
- Reconstruct



“An astronaut riding a rocking horse”

Image Diffusion

Input

Rendering (masks)

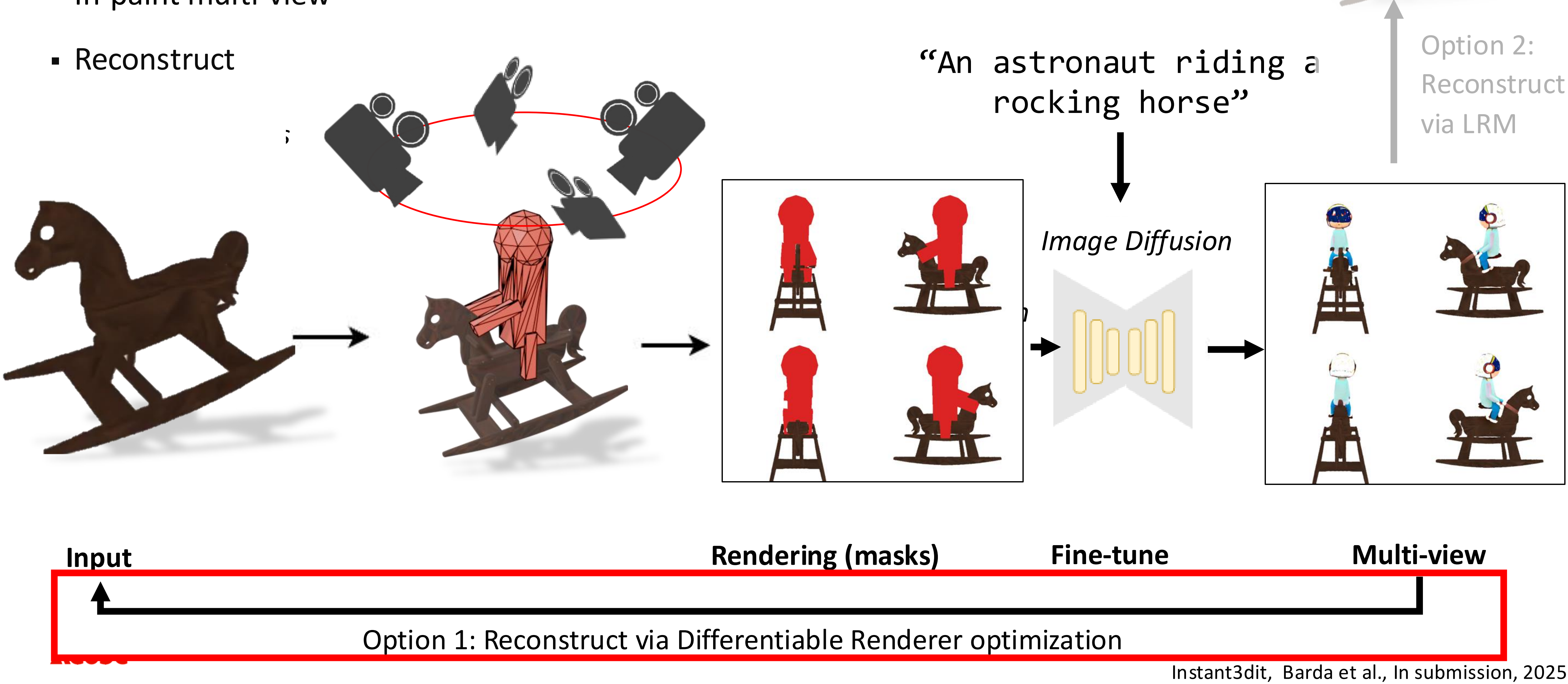
Fine-tune

Multi-view

Instant3dit, Barda et al., In submission, 2025

Generative Sculpting via Multi-view Inpainting

- In-paint multi-view
- Reconstruct



“An astronaut riding a rocking horse”

Image Diffusion

Option 2: Reconstruct via LRM

Input

Rendering (masks)

Fine-tune

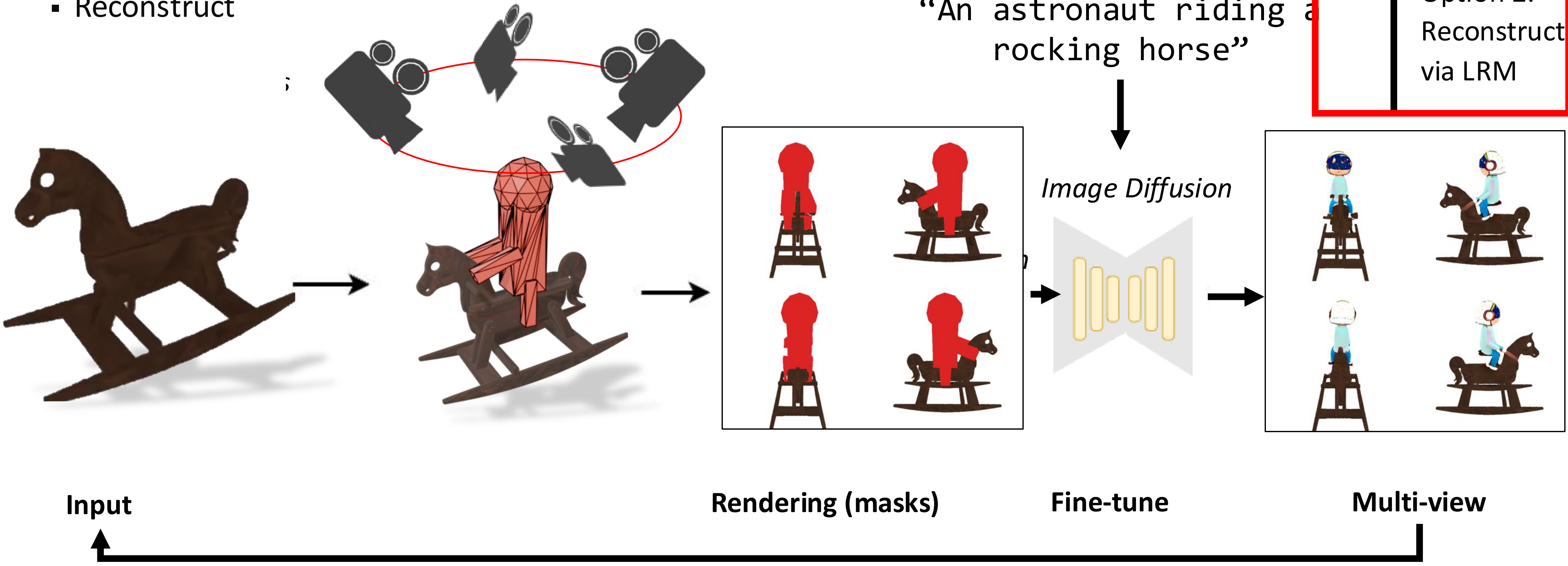
Multi-view

Option 1: Reconstruct via Differentiable Renderer optimization

Instant3dit, Barda et al., In submission, 2025

Generative Sculpting via Multi-view Inpainting

- In-paint multi-view
- Reconstruct



“An astronaut riding a rocking horse”

Image Diffusion

Option 2: Reconstruct via LRM

Input

Rendering (masks)

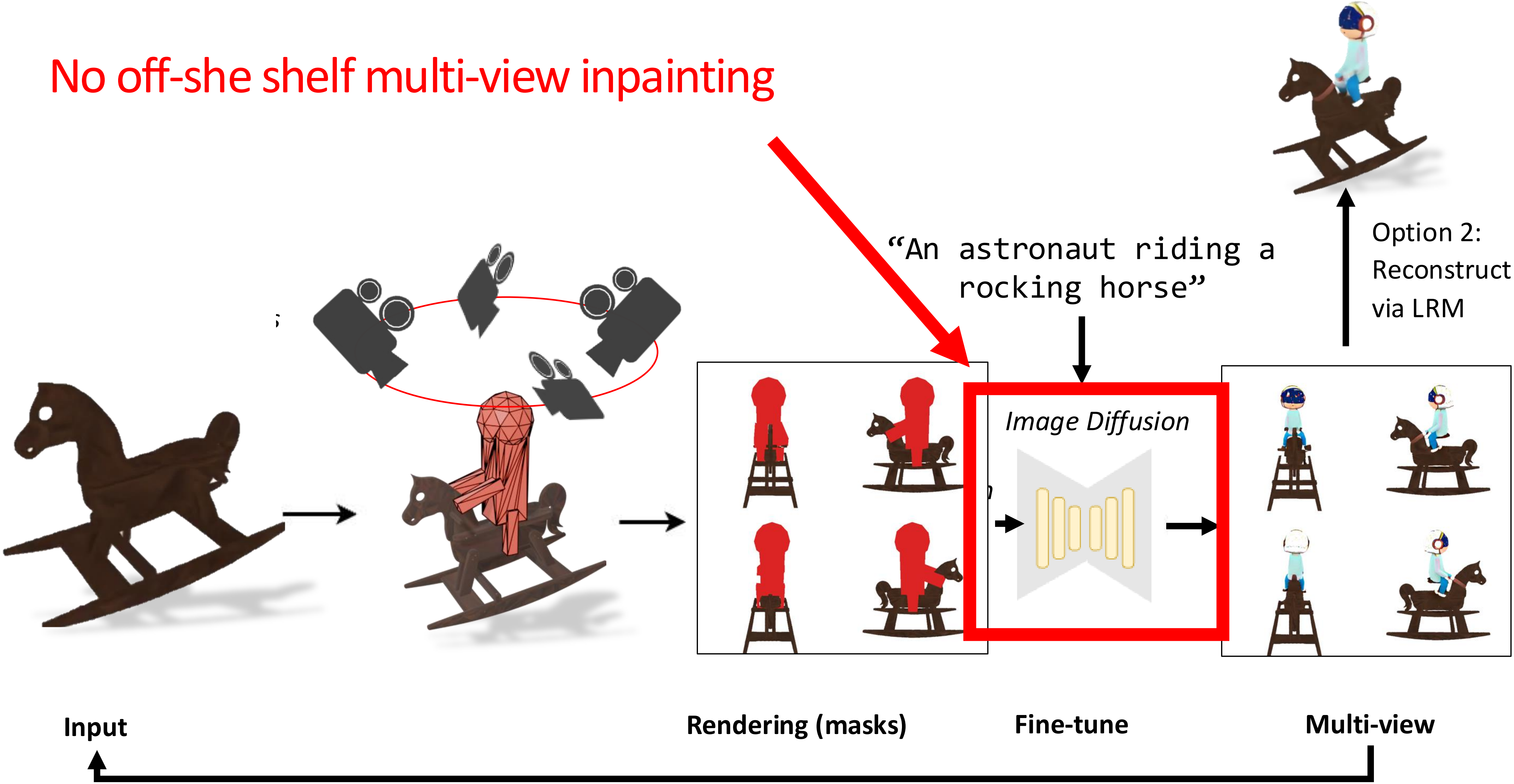
Fine-tune

Multi-view

Option 1: Reconstruct via Differentiable Renderer optimization

Instant3dit, Barda et al., In submission, 2025

No off-she shelf multi-view inpainting



“An astronaut riding a rocking horse”

Option 2: Reconstruct via LRM

Image Diffusion

Input

Rendering (masks)

Fine-tune

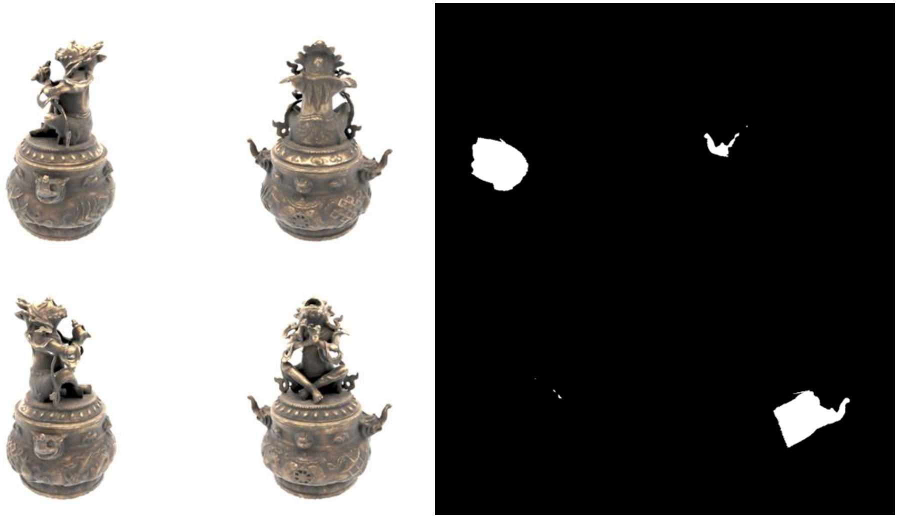
Multi-view

Option 1: Reconstruct via Differentiable Renderer optimization

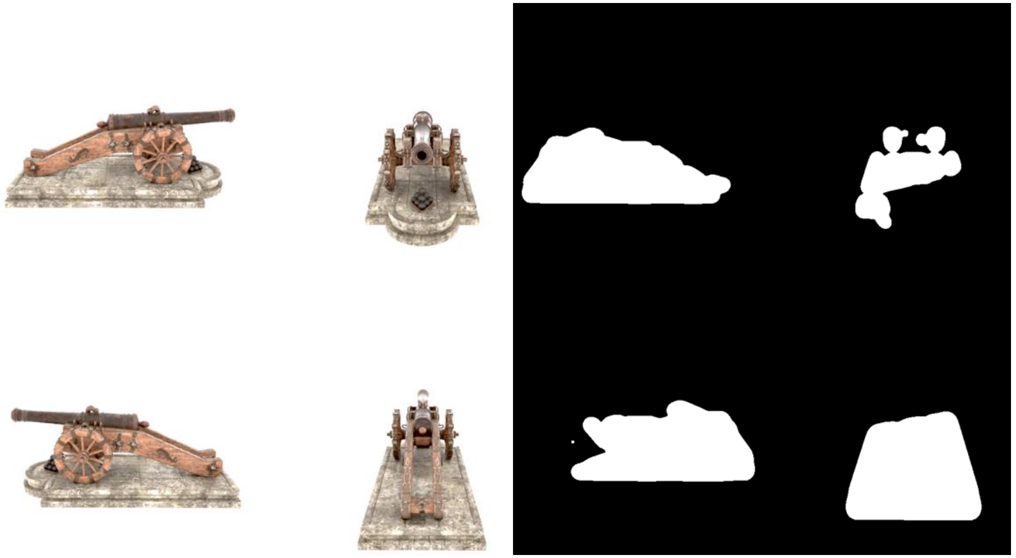
Instant3dit, Barda et al., In submission, 2025

Training Data

- 5000 high quality meshes from the Objaverse dataset
- Each rendered from 4 canonical views
- Detailed Captions created using LLaVa



"A bronze sculpture of a mythical creature with intricate detailing, featuring a dragon-like creature with scales and a flowing mane, perched atop a round base adorned with additional decorative elements. The surface is textured and ornate patterns embellish the pedestal."



"A low-poly cannon model with a textured surface, featuring a cylindrical barrel and a wooden carriage with large wheels. The cannon is mounted on a stone base, which has a small round platform in front of it. The overall design suggests an old-fashioned or historical style."

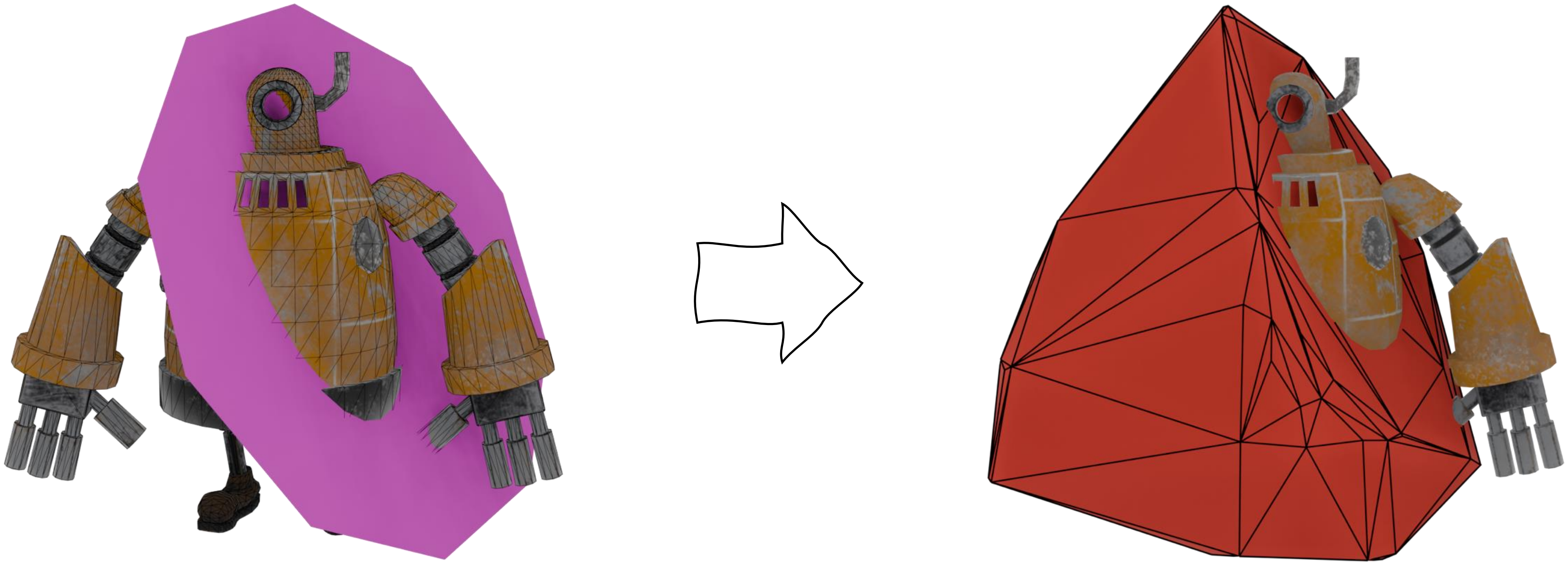
I: Coarse Edit

- Scenario: “user creates a large chunk with a coarse proxy”

Instant3dit, Barda et al., In submission, 2025

I: Coarse Edit

- Scenario: “user creates a large chunk with a coarse proxy”
- 3D mask is a polyhedron fully containing the region to be inpainted



Instant3dit, Barda et al., In submission, 2025

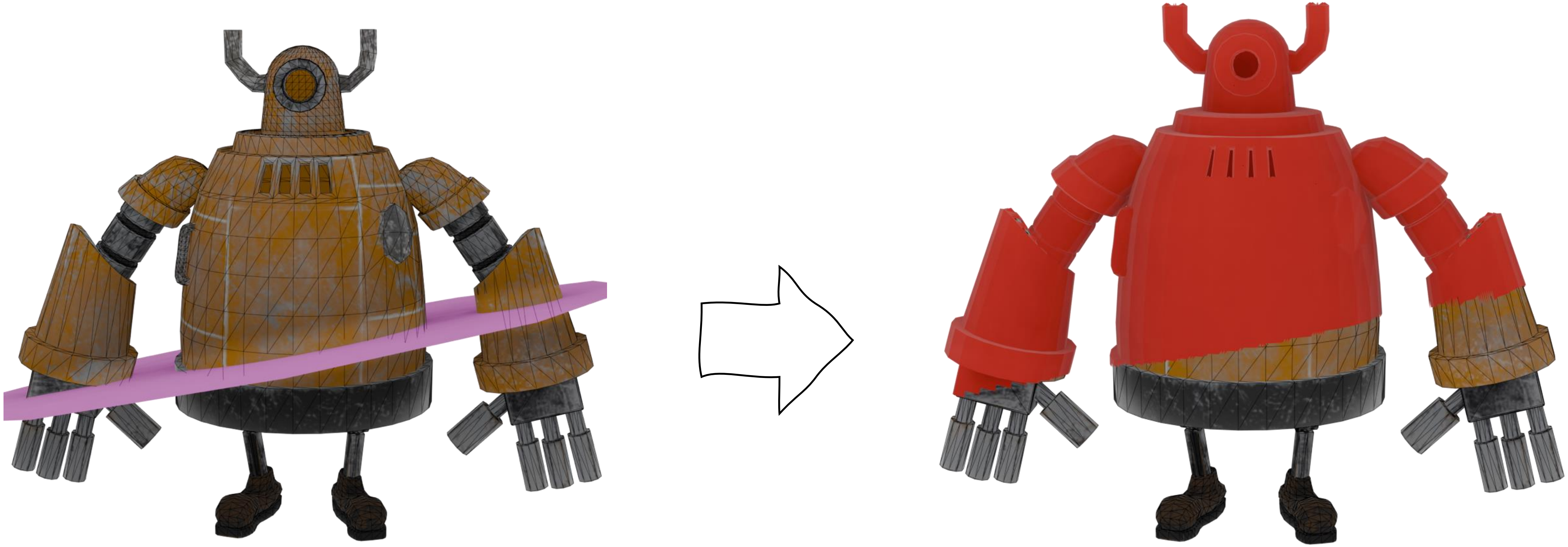
II: Mesh Sculpting

- Scenario: “user provides a more precise approximation to the target”

Instant3dit, Barda et al., In submission, 2025

II: Mesh Sculpting

- Scenario: “user provides a more precise approximation to the target”
- 3D mask is the exact mesh that needs to be inpainted



Instant3dit, Barda et al., In submission, 2025

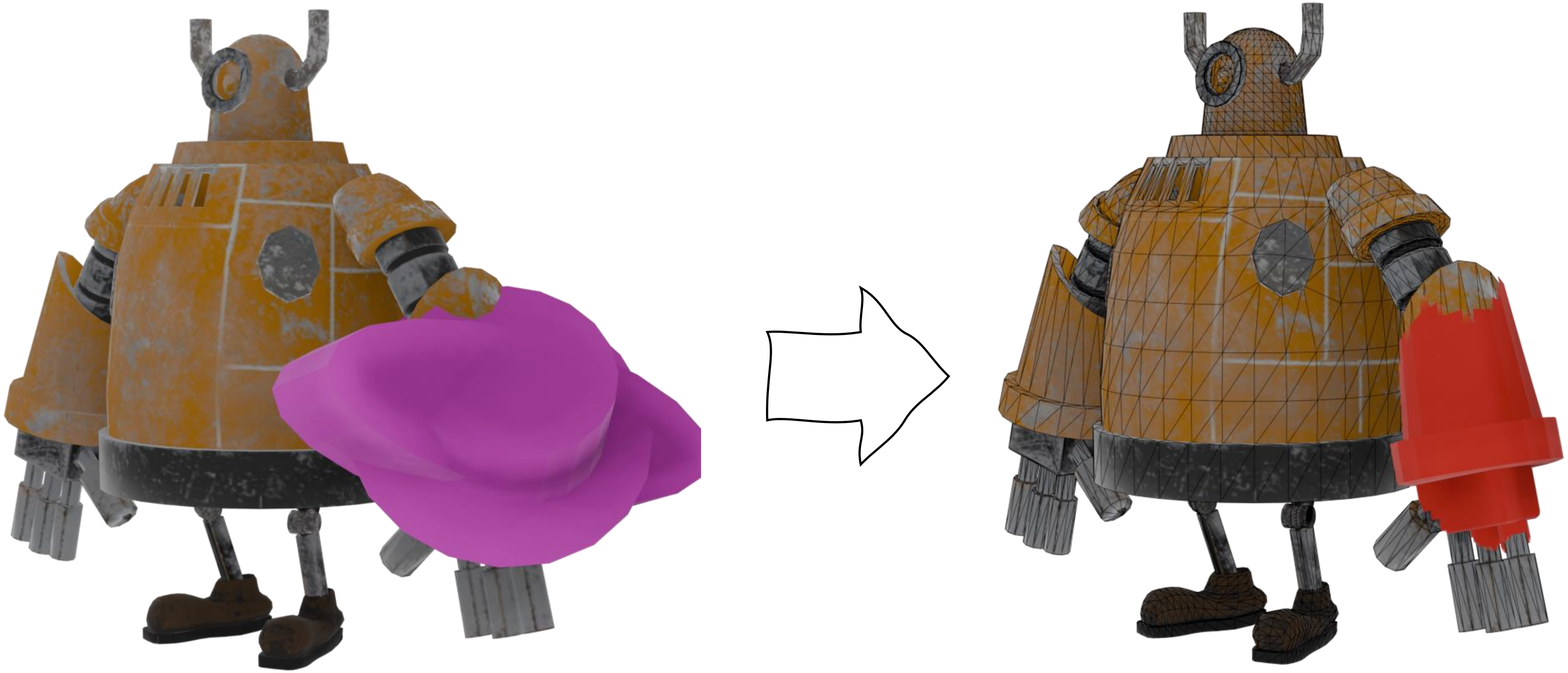
III: Surface Edit

- Scenario: “user provides wants local geometry and texture modifications”

Instant3dit, Barda et al., In submission, 2025

III: Surface Edit

- Scenario: “user provides wants local geometry and texture modifications”
- 3D mask is a small region of the exact mesh



Instant3dit, Barda et al., In submission, 2025

Training Multi-View Inpainting

- Train from scratch on 5B Multi-view inpainted images?

Instant3dit, Barda et al., In submission, 2025

Training Multi-View Inpainting

- ~~Train from scratch on 5B Multi-view inpainted images?~~
- Option 1: fine-tune multi-view image generation to teach it to inpaint

Opt 1: Instant3D

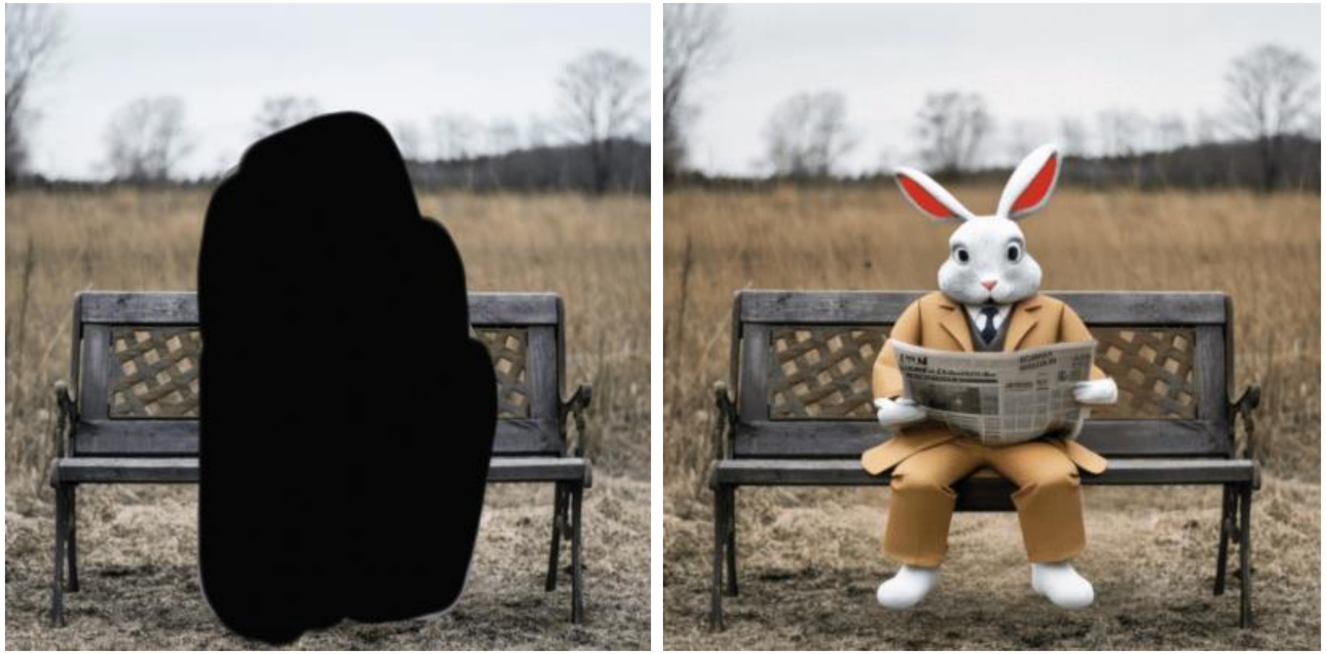
Adobe

Instant3dit, Barda et al., In submission, 2025

Training Multi-View Inpainting

- ~~Train from scratch on 5B Multi-view inpainted images?~~
- Option 1: fine-tune multi-view image generation to teach it to inpaint
- Option 2: fine-tune generative image inpainting to teach it to create multi-view

Opt 1: Instant3D



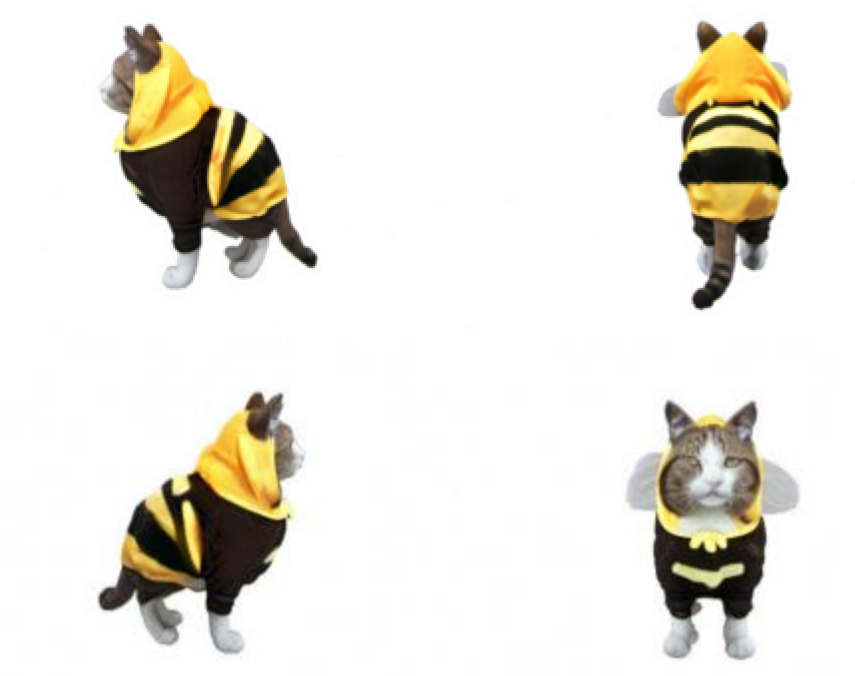
Opt 2: SDXL-Inpaint

Instant3dit, Barda et al., In submission, 2025

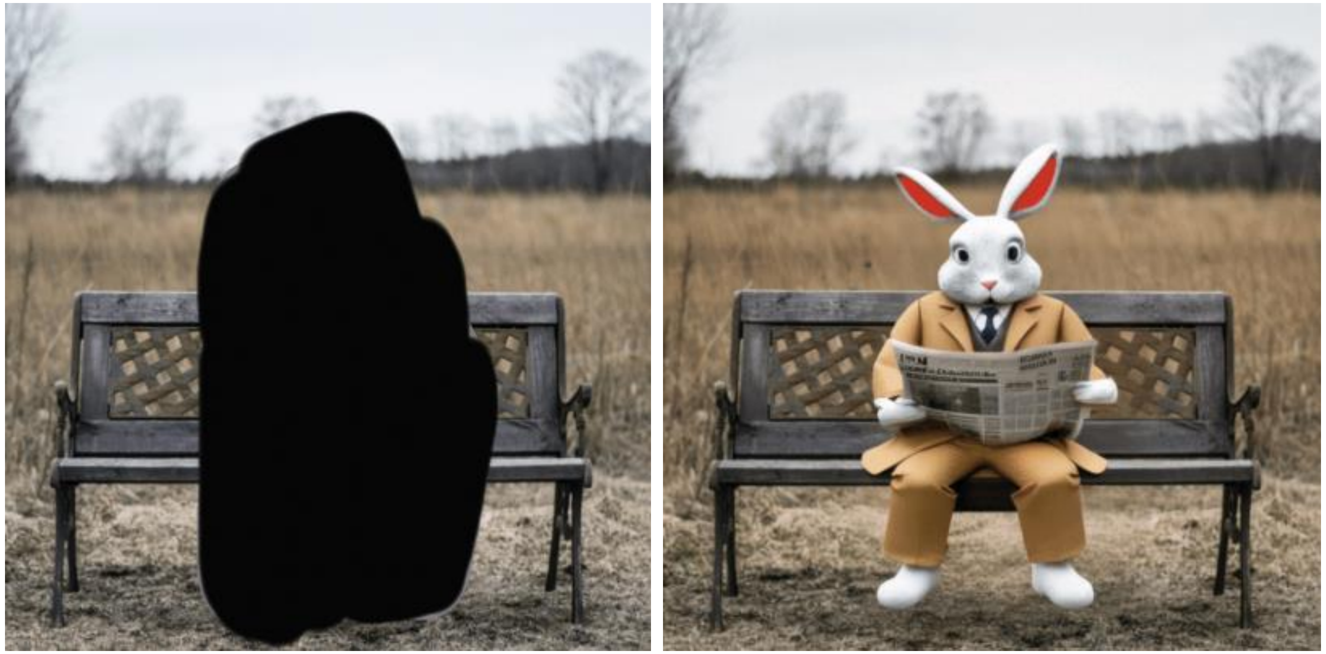
Training Multi-View Inpainting

Which one is a better base model?

- ~~Train from scratch on 5B Multi-view inpainted images?~~
- Option 1: fine-tune multi-view image generation to teach it to inpaint
- Option 2: fine-tune generative image inpainting to teach it to create multi-view



Opt 1: Instant3D



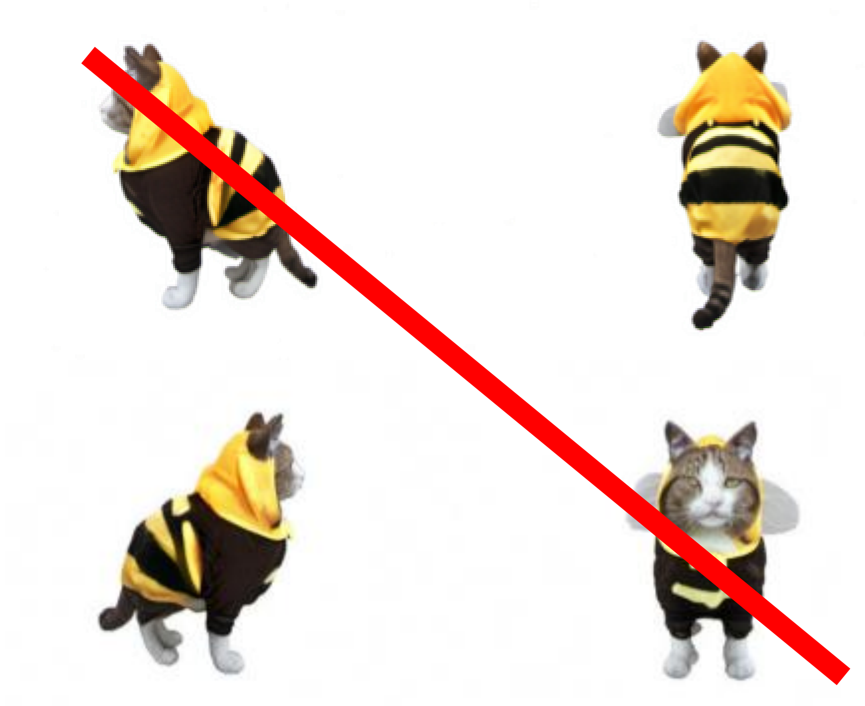
Opt 2: SDXL-Inpaint

Instant3dit, Barda et al., In submission, 2025

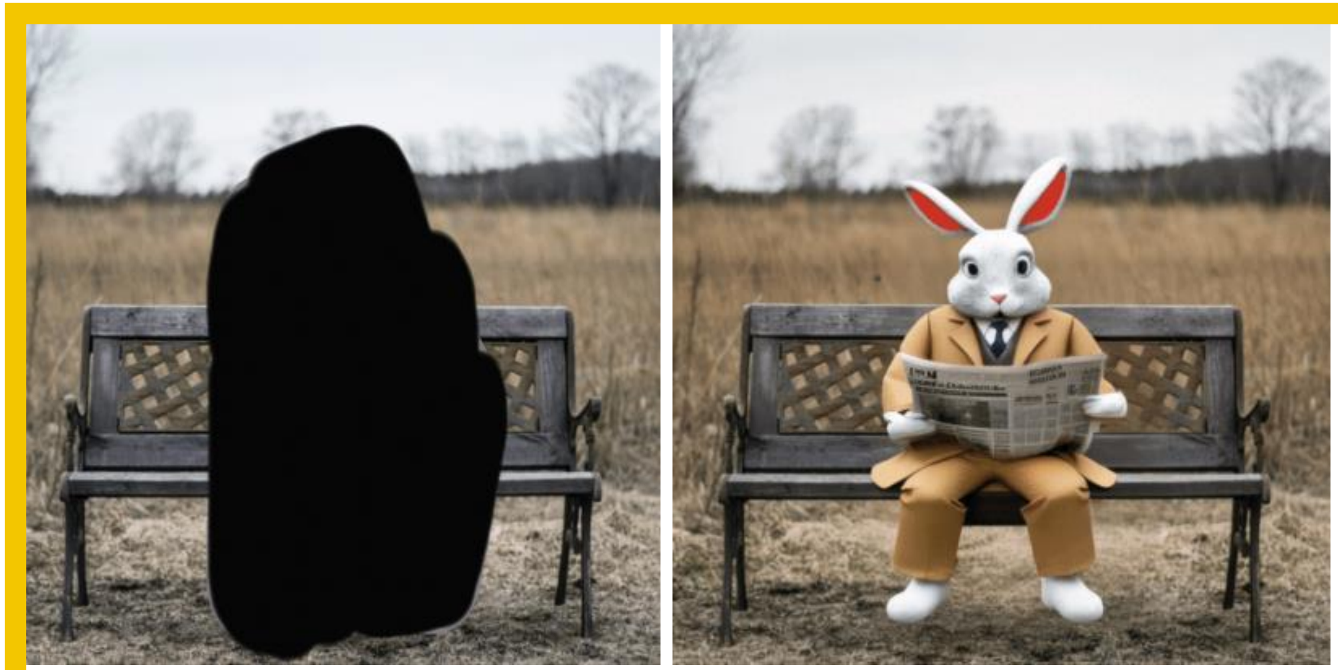
Training Multi-View Inpainting

Which one is a better base model?

- ~~Train from scratch on 5B Multi-view inpainted images?~~
- ~~Option 1: fine-tune multi-view image generation to teach it to inpaint~~
- Option 2: fine-tune generative image inpainting to teach it to create multi-view



Opt 1: Instant3D

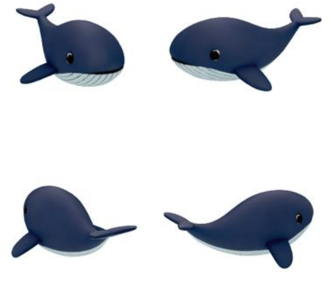


Opt 2: SDXL-Inpaint

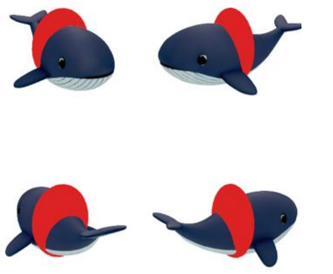
Instant3dit, Barda et al., In submission, 2025

Ablation: choose the right backbone

- What if we start with multi-view backbone instead of inpainting backbone?



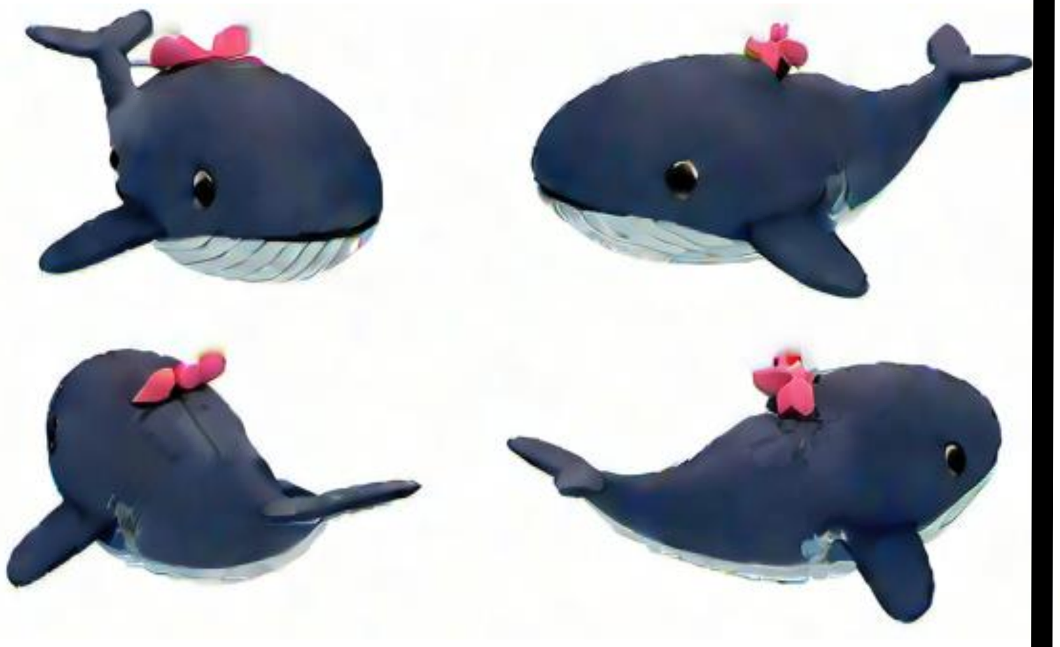
Input



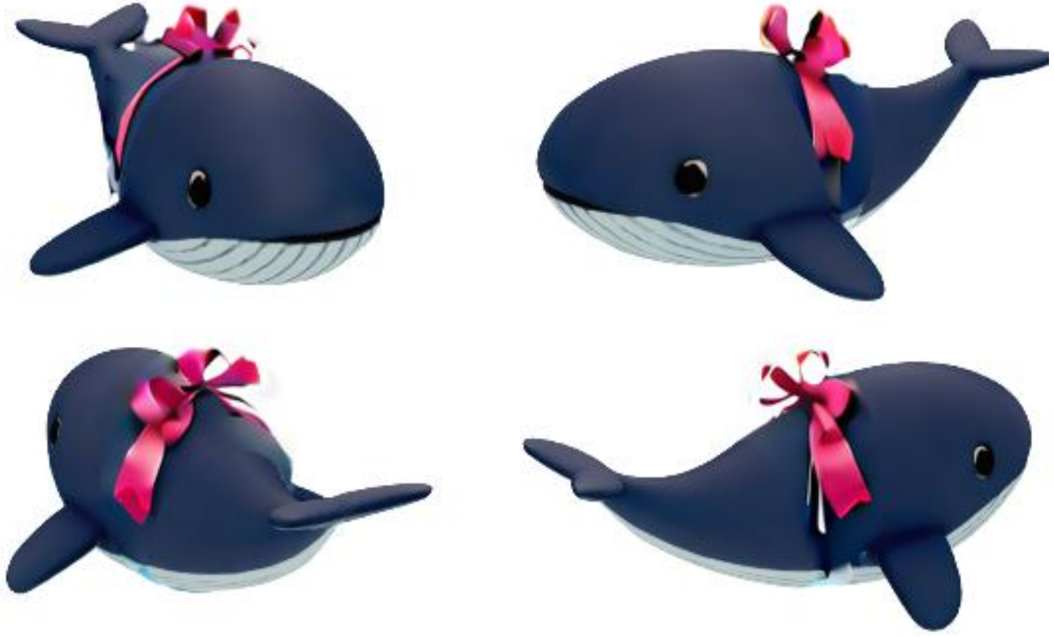
Mask

“A whale wrapped in a pink bow”

Prompt



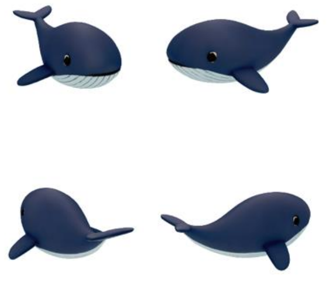
Instant3D Finetuned



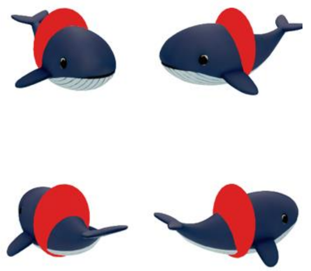
SDXL IP Finetuned (Ours)

Ablation: choose the right backbone

- What if we start with multi-view backbone instead of inpainting backbone?



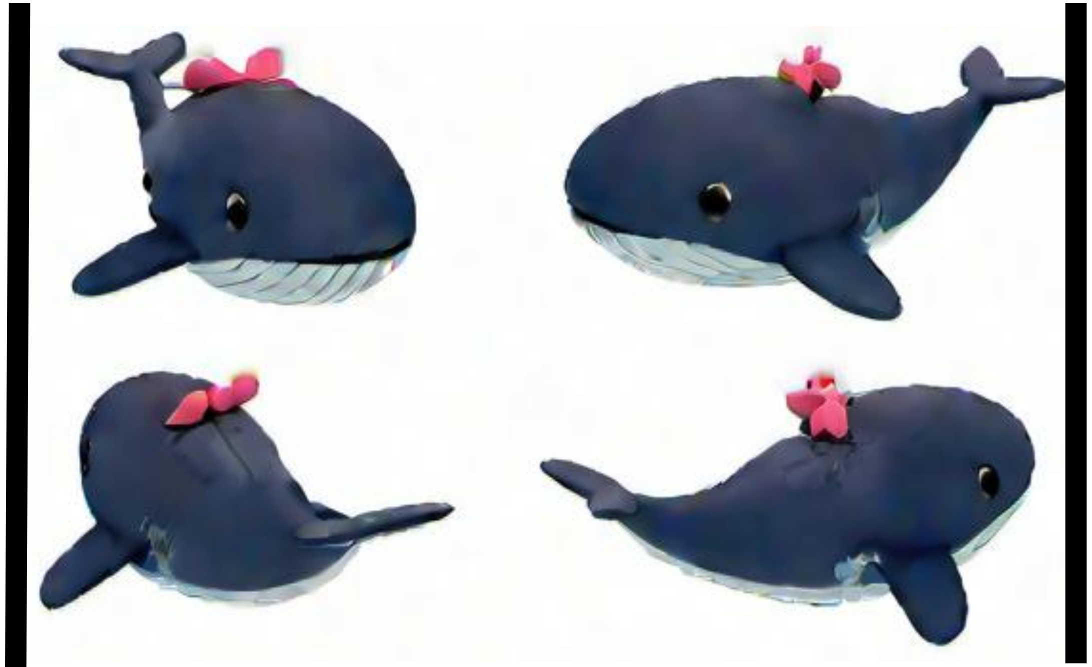
Input



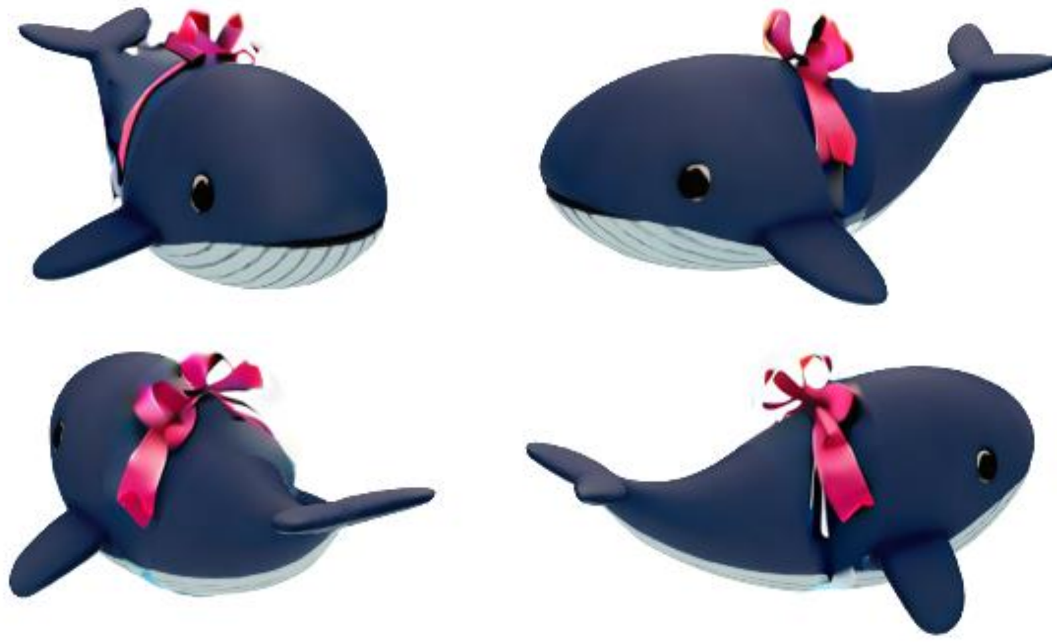
Mask

“A whale wrapped in a pink bow”

Prompt



Instant3D Finetuned



SDXL IP Finetuned (Ours)

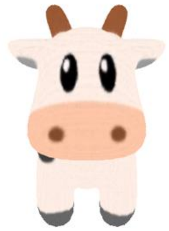
Benchmark Stats

	Instant3D Finetuned		SDXL IP Finetuned (Ours)
Prompt Adherance (CLIPL):	28.57	vs	29.01
Multi-View Consistency (DreamSim):	0.97	vs	0.100
Visual Quality (FID):	121.1	vs	118.4

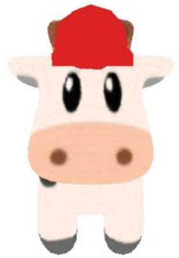
be Confidential.

Ablation: use multiple masks

- What if we train only on one type of mask?



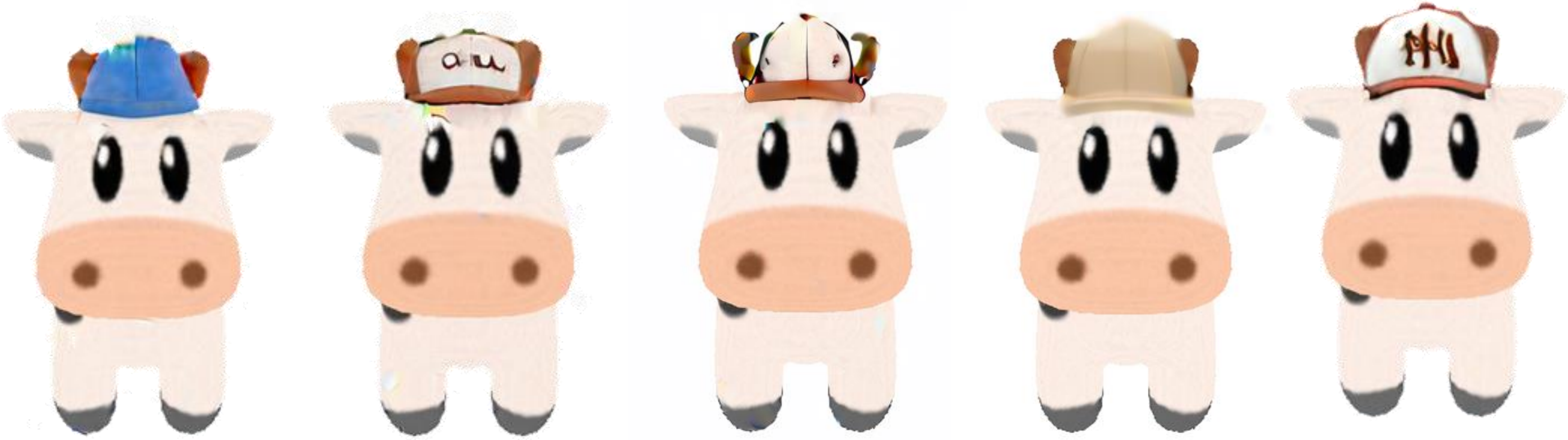
Input



Mask

“A cow with a baseball cap”

Prompt



Random 2d	Type I only	Type II only	Type III only	**All (ours)**

Benchmark Stats (FID)

Random 2d	Type I only	Type II only	Type III only	**All (ours)**
131.1	121.3	128.2	142.2	118.4

Instant3dit, Barda et al., In submission, 2025

Interactive Generative Sculpting

speed 5x

Instant3dit demo.

Adobe

Instant3dit, Barda et al., In submission, 2025

Results: support all representations

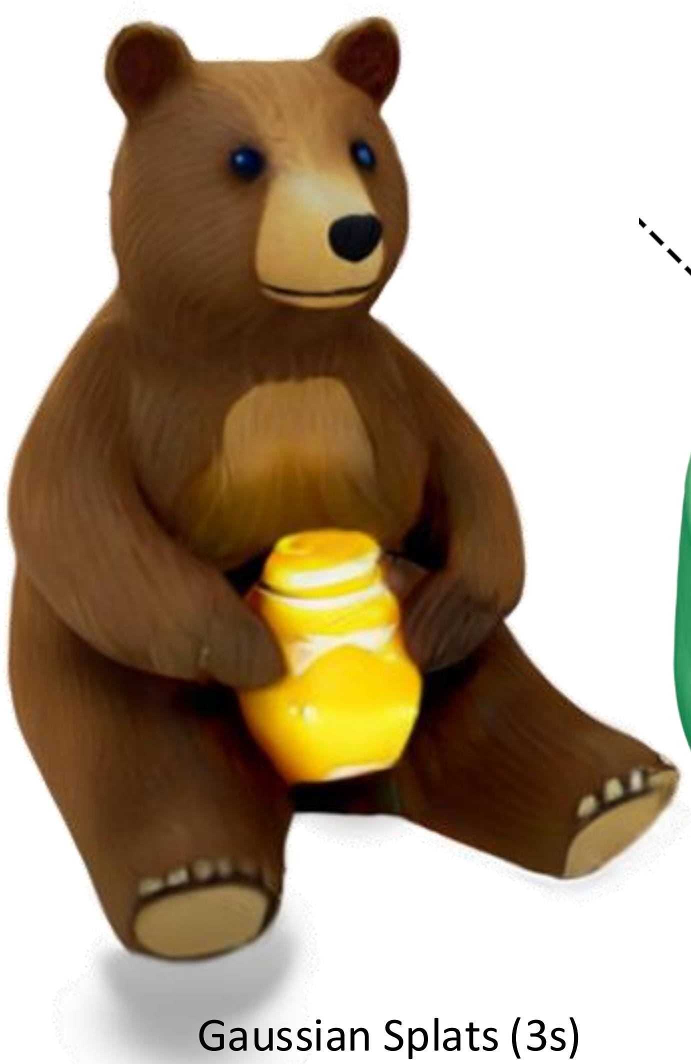
- Just swap different LRM models (or optimize mesh)

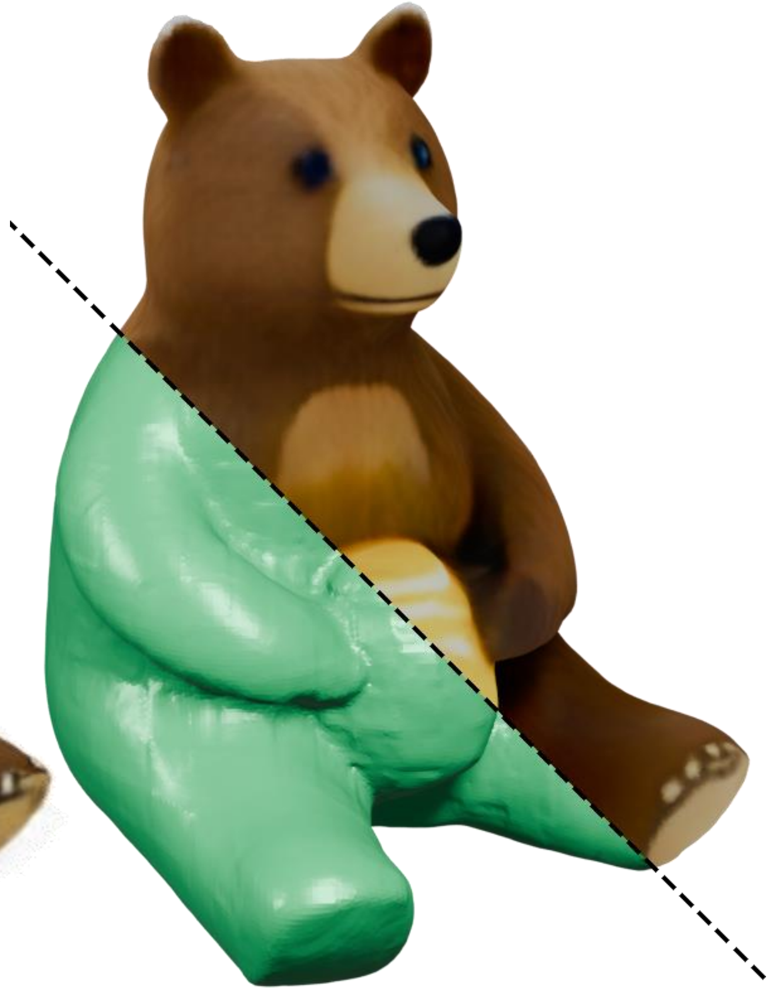
Input

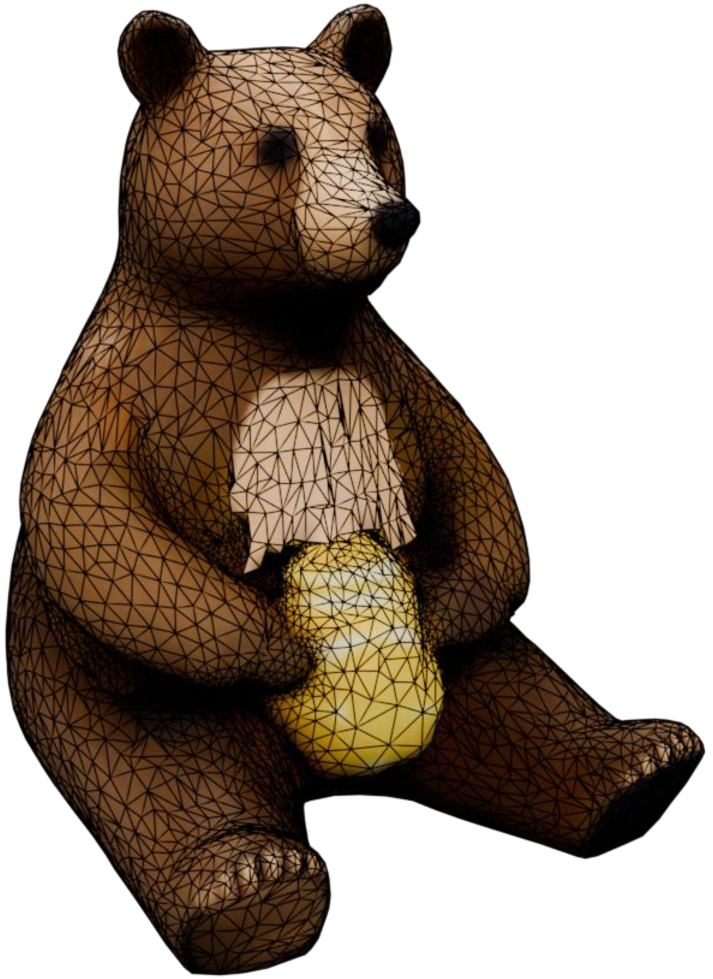
Mask

“A bear holding a honey pot”

Prompt

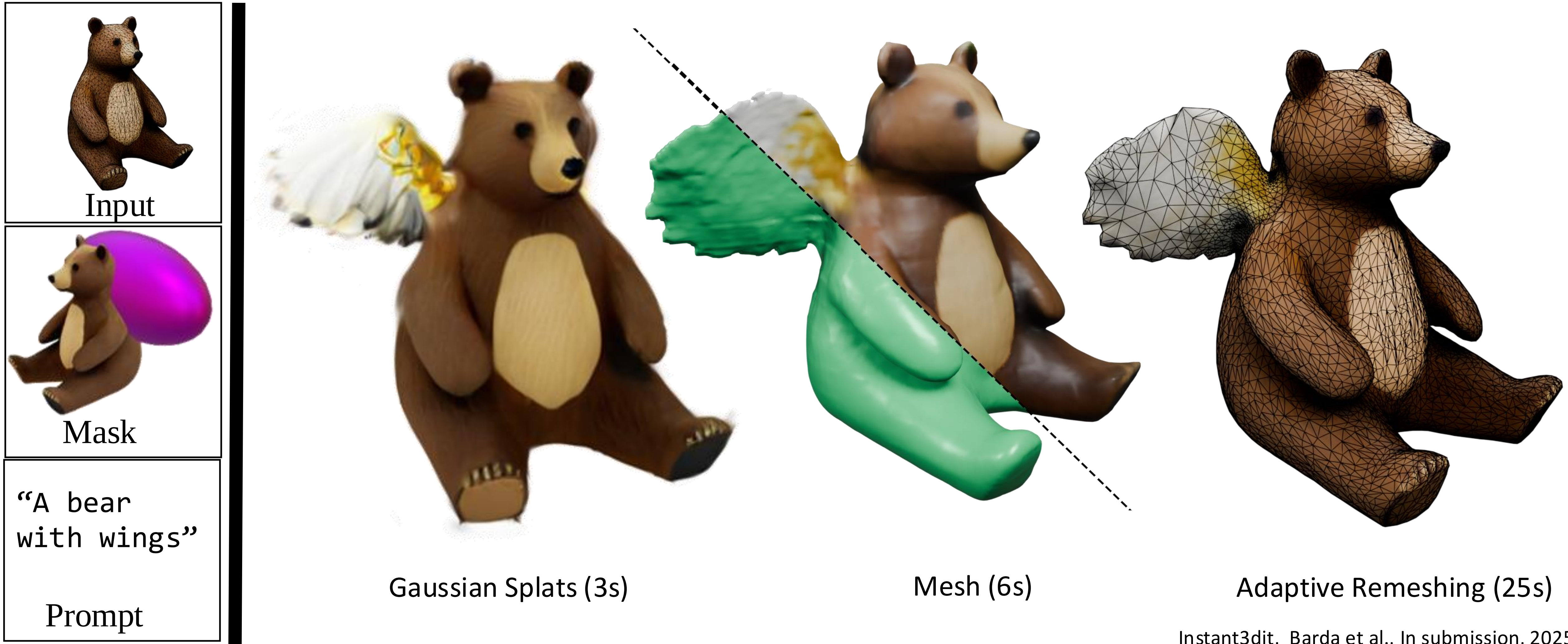

Gaussian Splats (3s)


Mesh (6s)


Adaptive Remeshing (25s)

Results: support all representations

- Just swap different LRM models (or optimize mesh)



Gaussian Splats (3s)

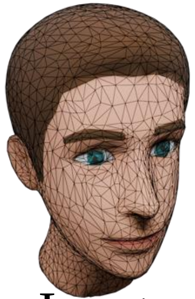
Mesh (6s)

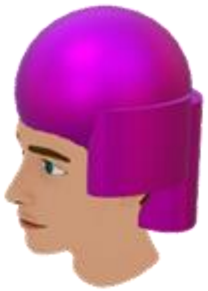
Adaptive Remeshing (25s)

Instant3dit, Barda et al., In submission, 2025

Results: support all representations

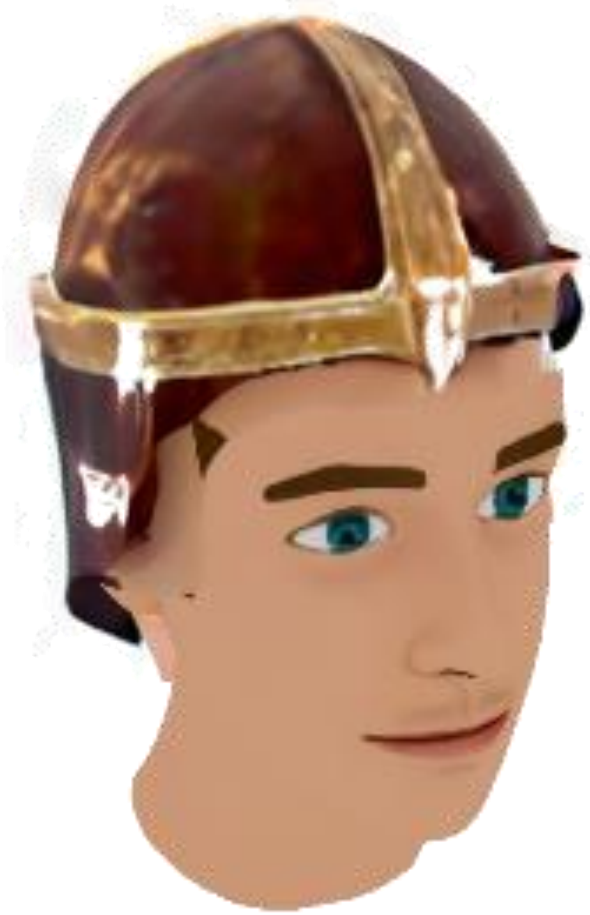
- Just swap different LRM models (or optimize mesh)

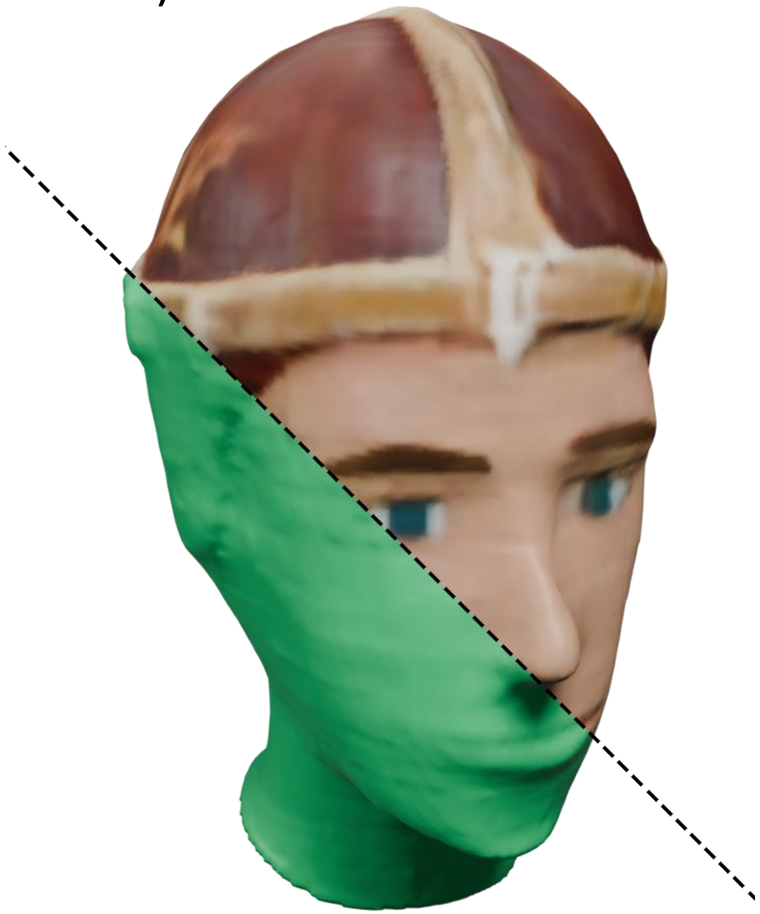

Input

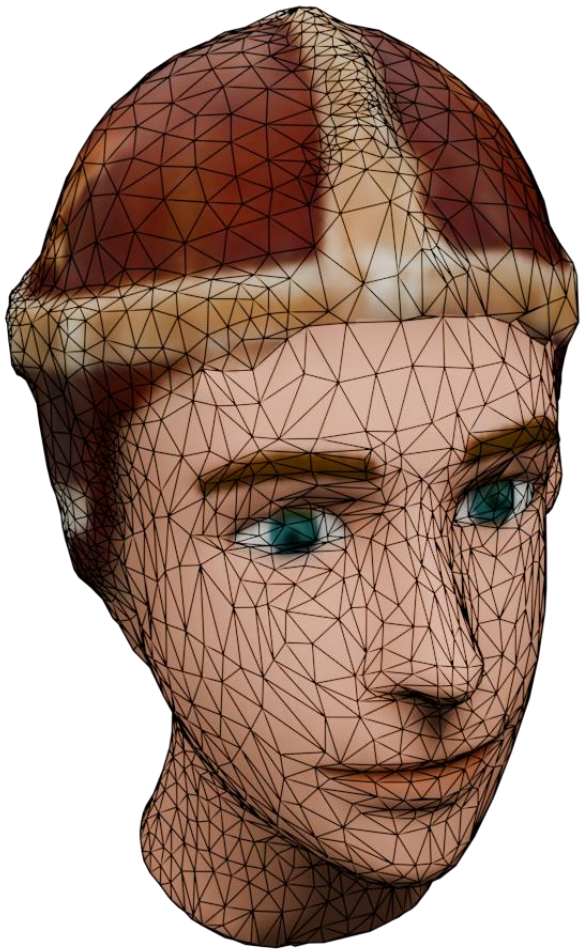

Mask

“Man wearing a medieval helmet”

Prompt


NeRF (3s)

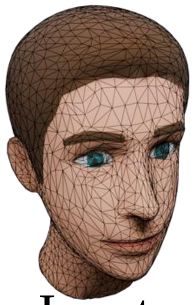

Mesh (6s)


Adaptive Remeshing (25s)

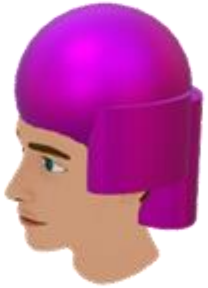
Instant3dit, Barda et al., In submission, 2025

Results: support all representations

- Just swap different LRM models (or optimize mesh)



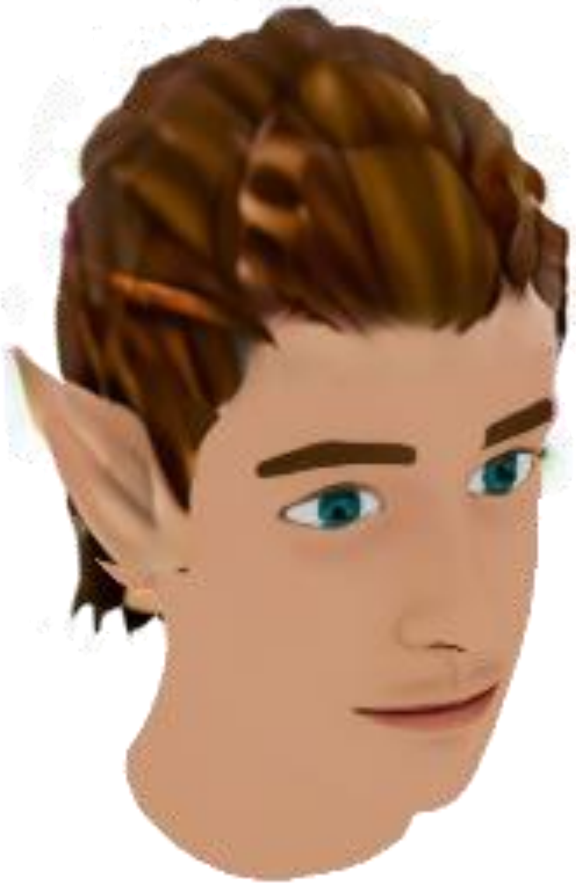
Input



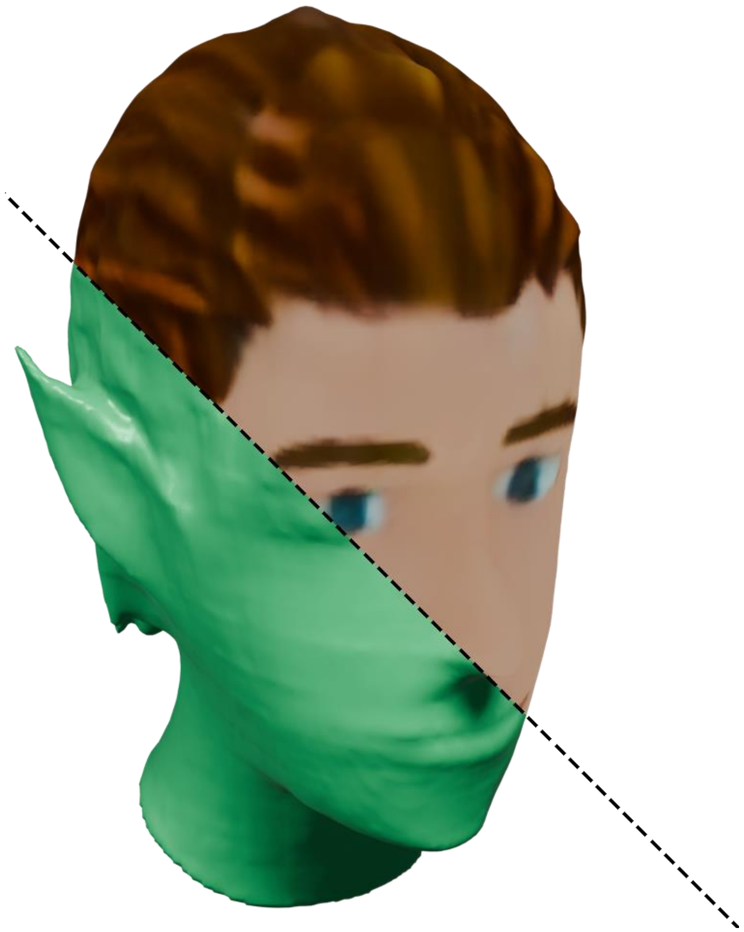
Mask

“An elven warrior”

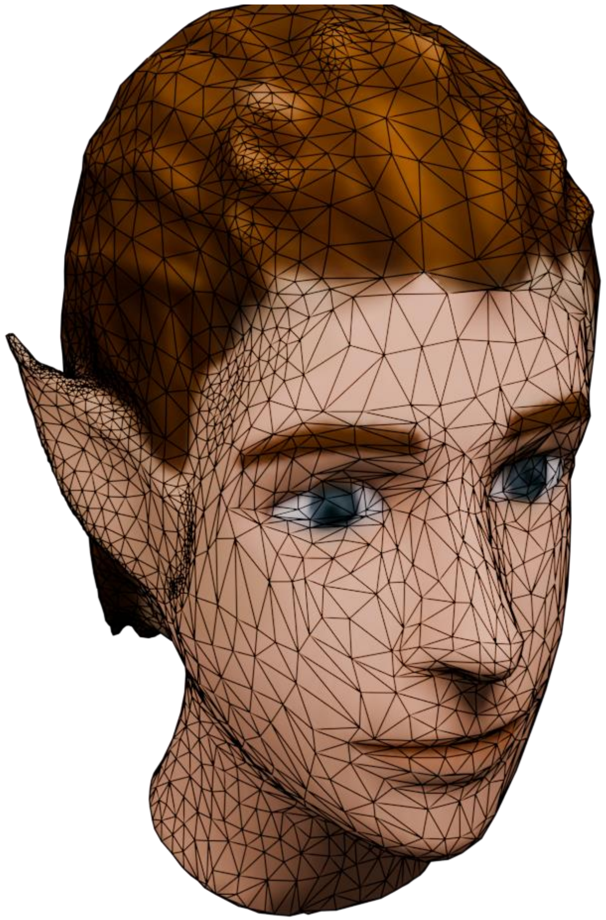
Prompt



NeRF (3s)



Mesh (6s)



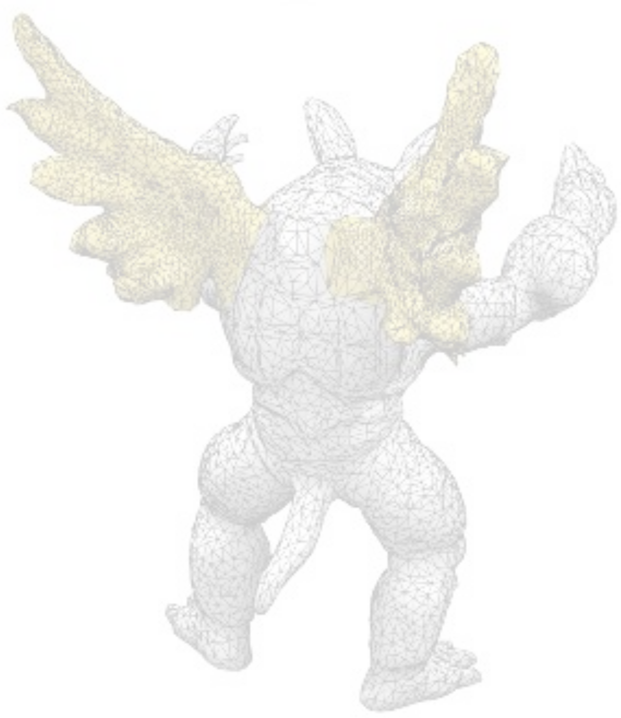
Adaptive Remeshing (25s)

Instant3dit, Barda et al., In submission, 2025

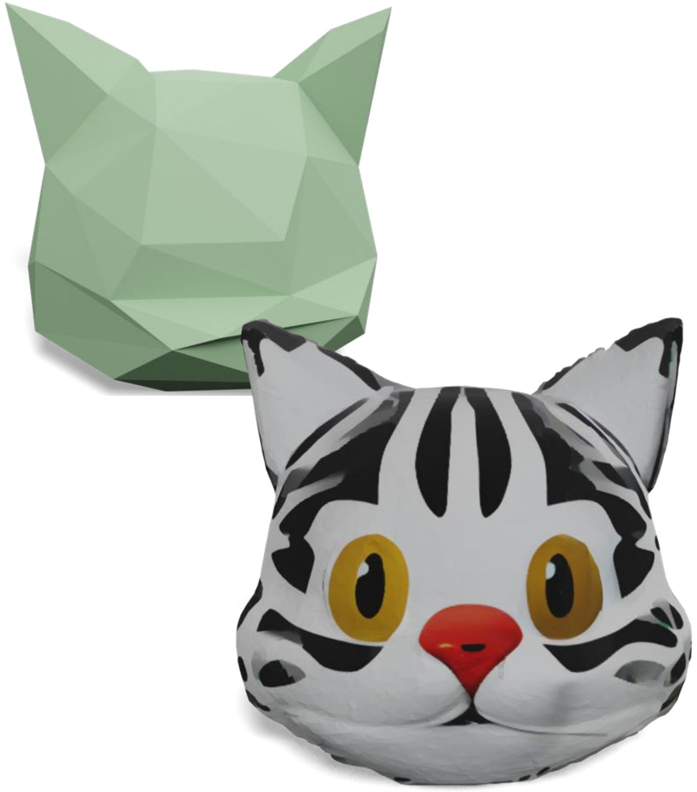
Overview

- Support mesh outputs (but use other representations as needed)
- Inspired by traditional workflows

Neural Deformation

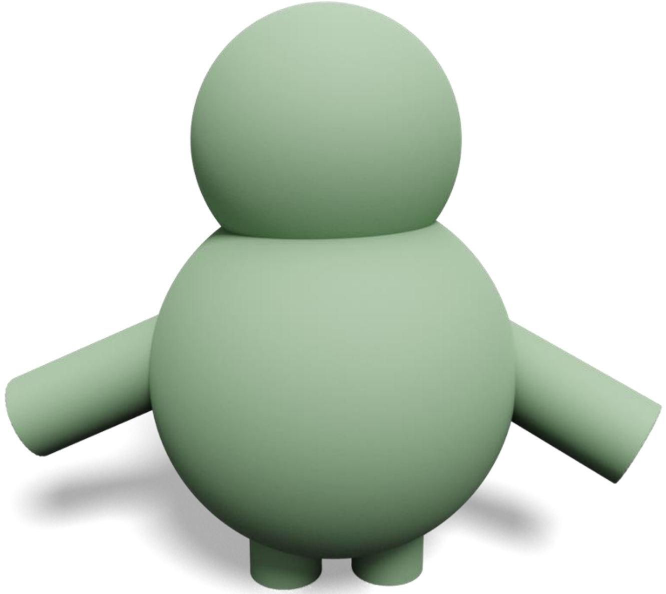


Generative Scultping

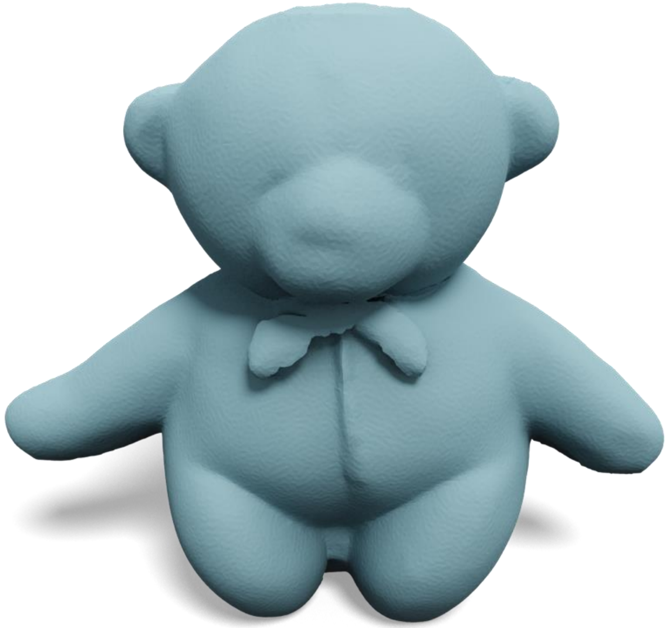


Generative Detailization

Generate or Change Details



>_: A teddy bear →



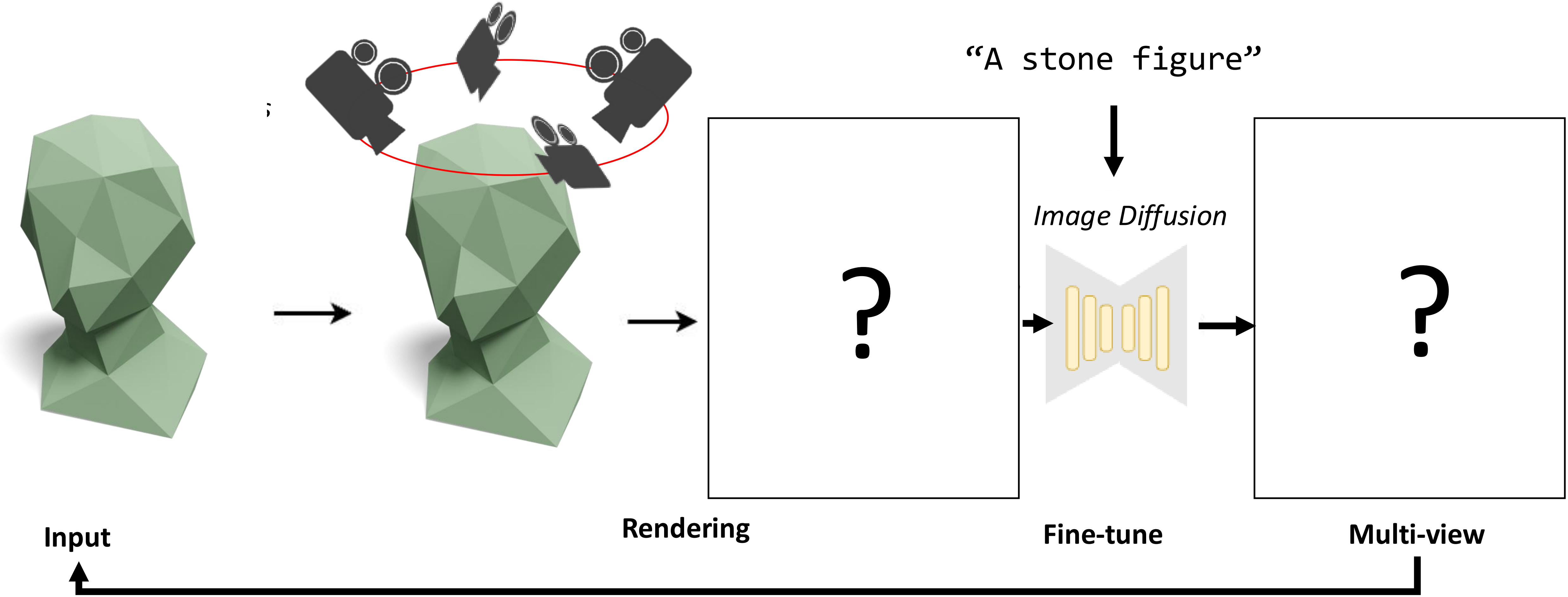
Input 3D mesh

Input text

Output 3D mesh

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

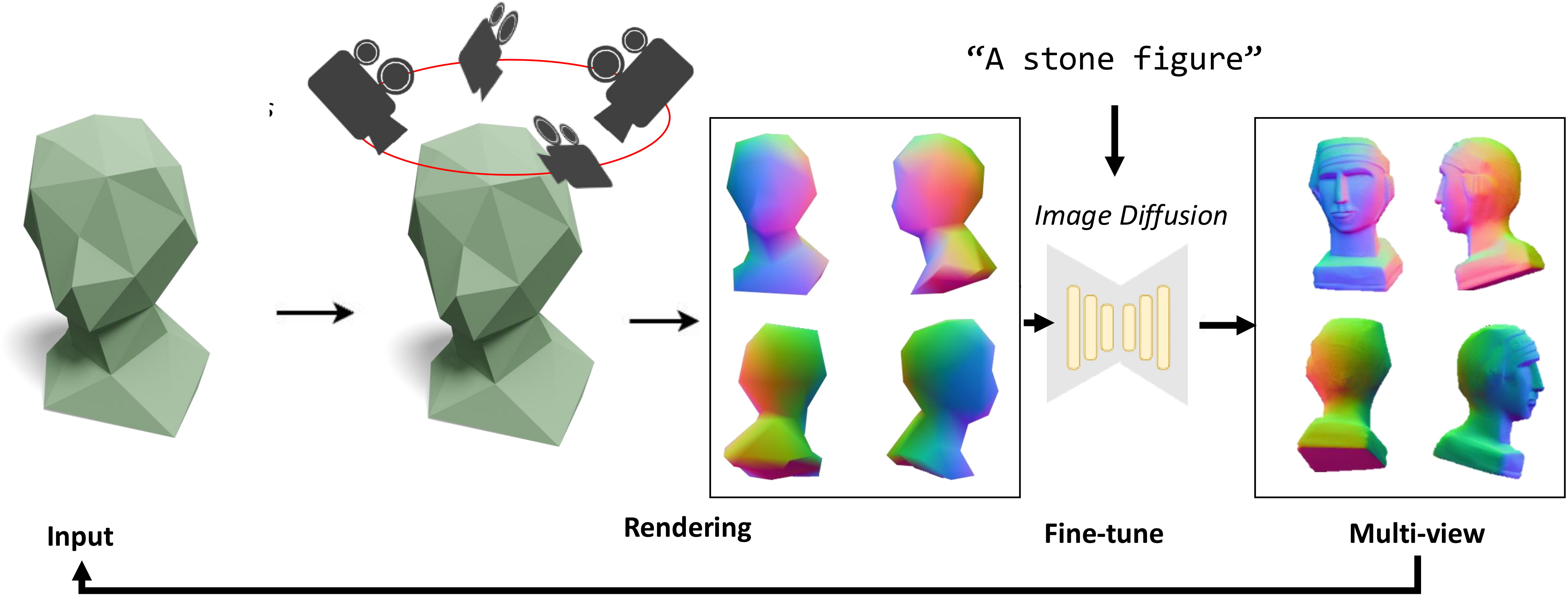
Multi-view Detail Generation



Option 1: Reconstruct via Differentiable Renderer optimization

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

Multi-view ~~Detail~~ Normal Generation



“A stone figure”

Image Diffusion

Input

Rendering

Fine-tune

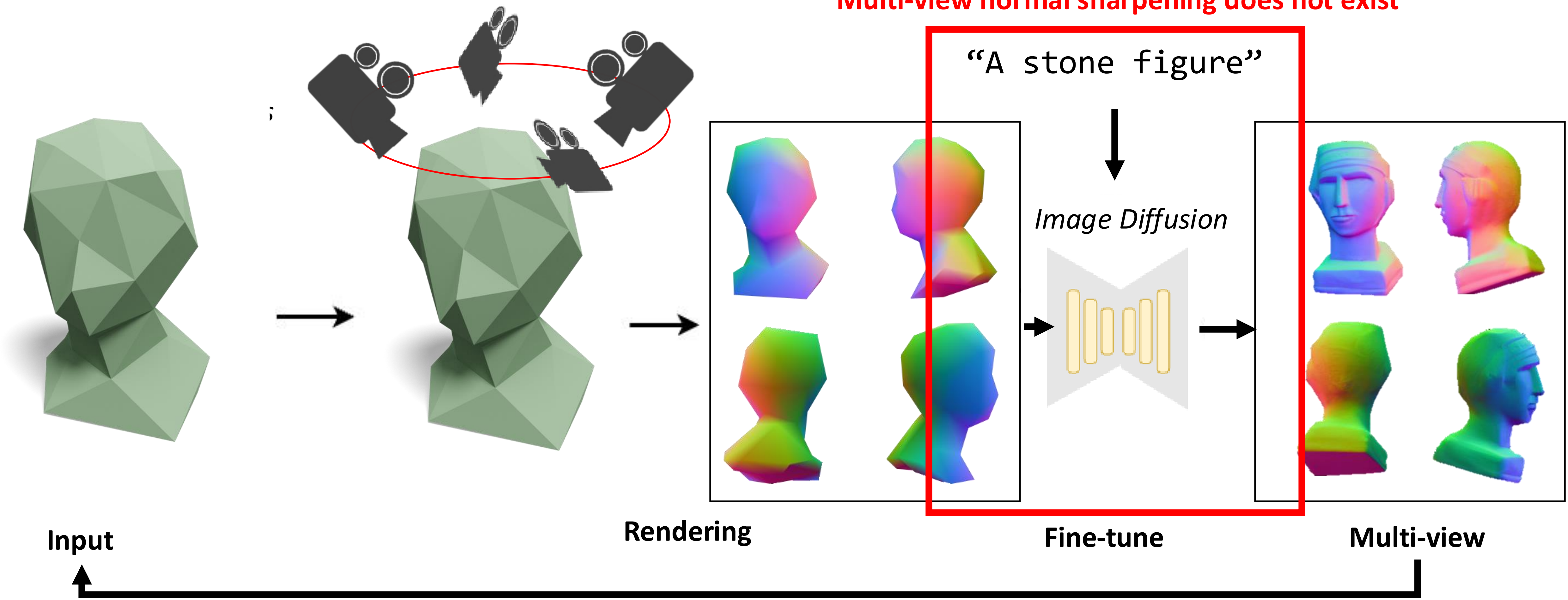
Multi-view

Option 1: Reconstruct via Differentiable Renderer optimization

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

Multi-view ~~Detail~~ Normal Generation

Multi-view normal sharpening does not exist

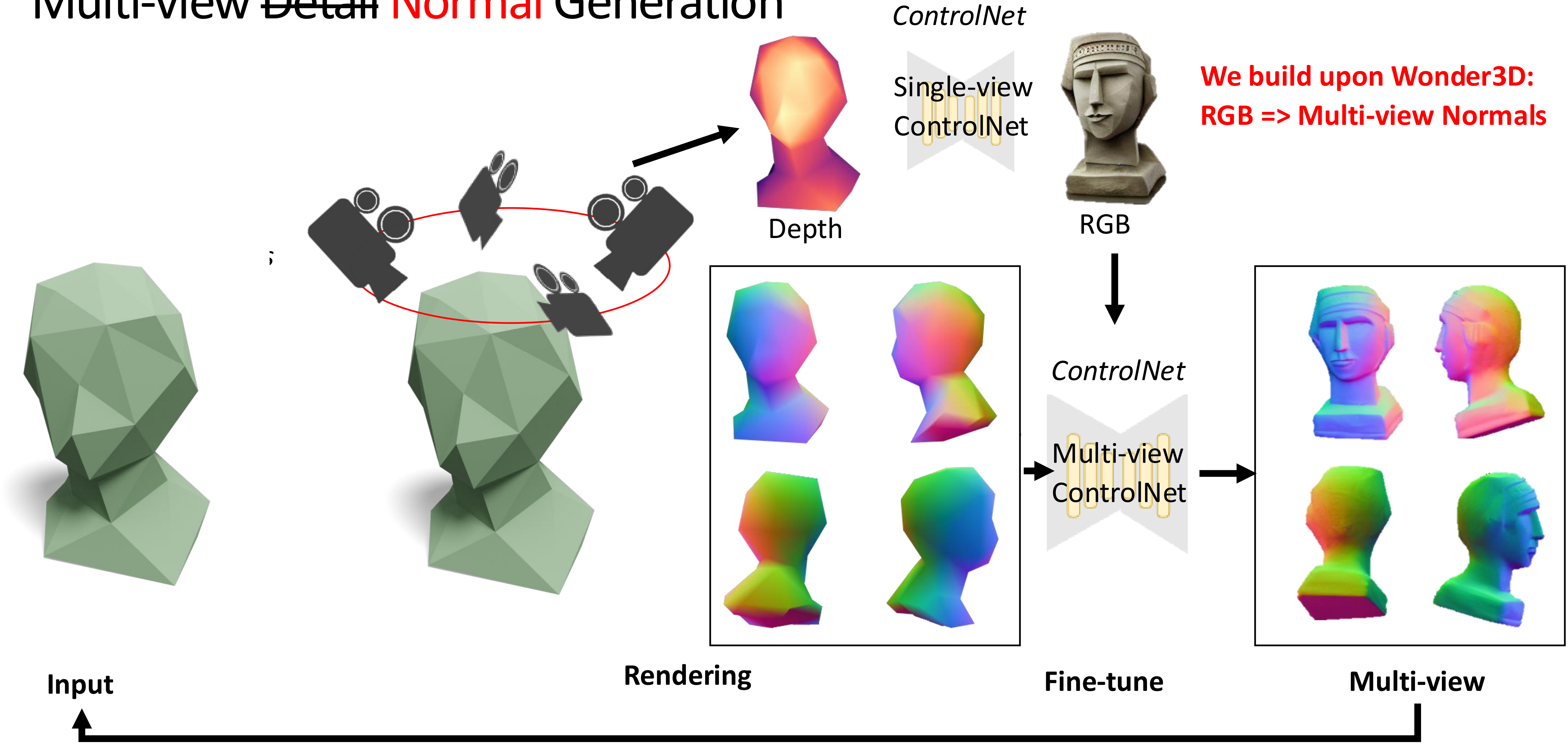


Option 1: Reconstruct via Differentiable Renderer optimization

Adobe

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

Multi-view ~~Detail~~ Normal Generation



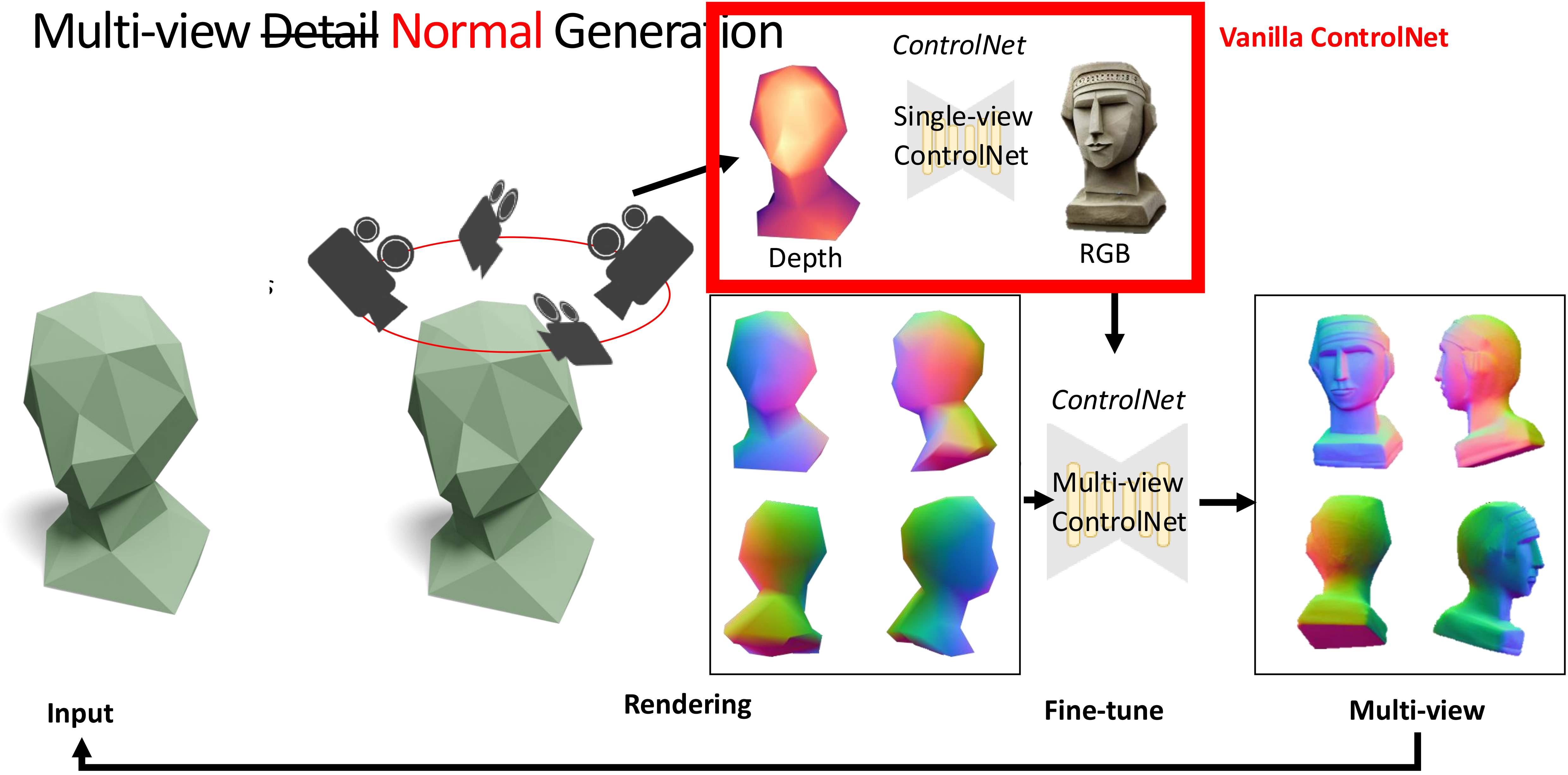
We build upon Wonder3D:
RGB => Multi-view Normals

Option 1: Reconstruct via Differentiable Renderer optimization

Adobe

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

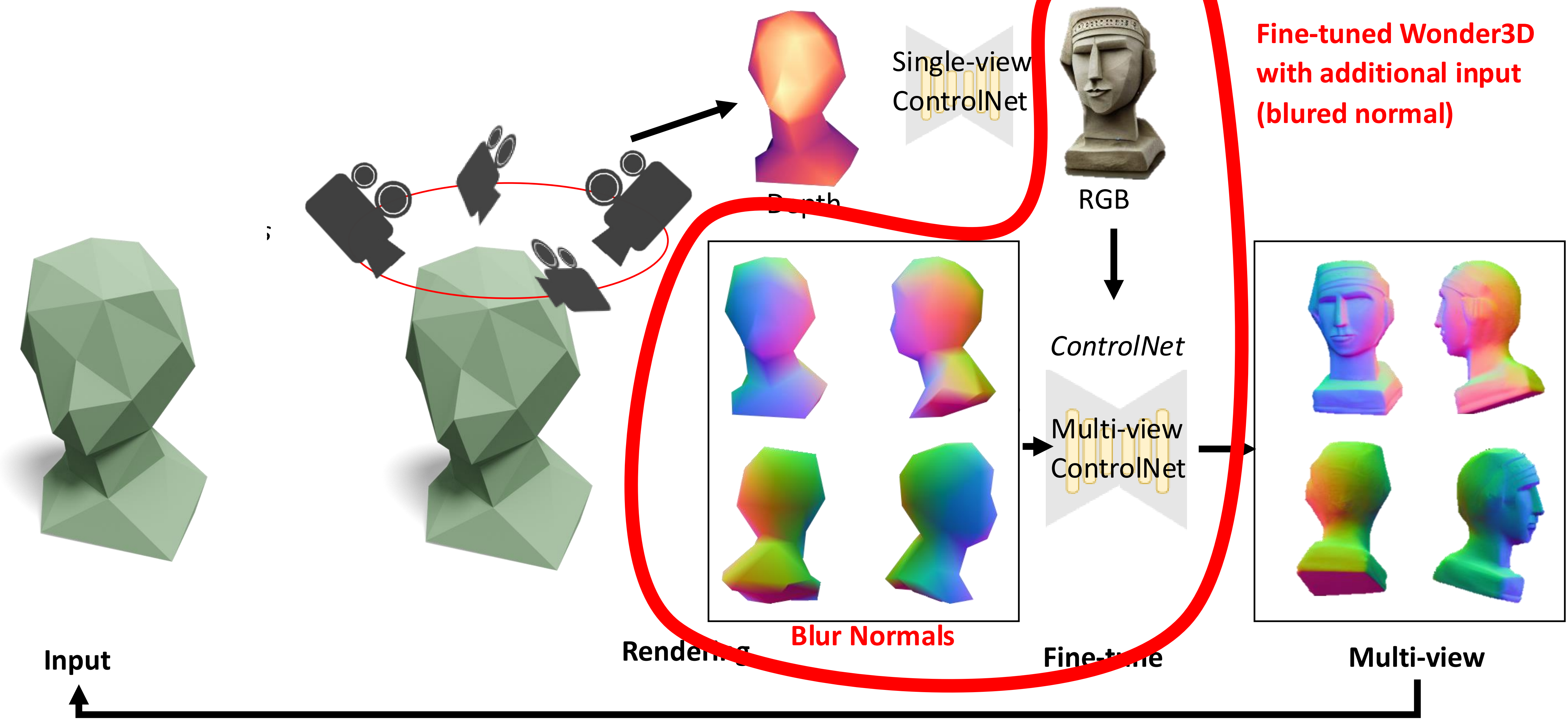
Multi-view ~~Detail~~ Normal Generation



Option 1: Reconstruct via Differentiable Renderer optimization

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

Multi-view ~~Detail~~ Normal Generation

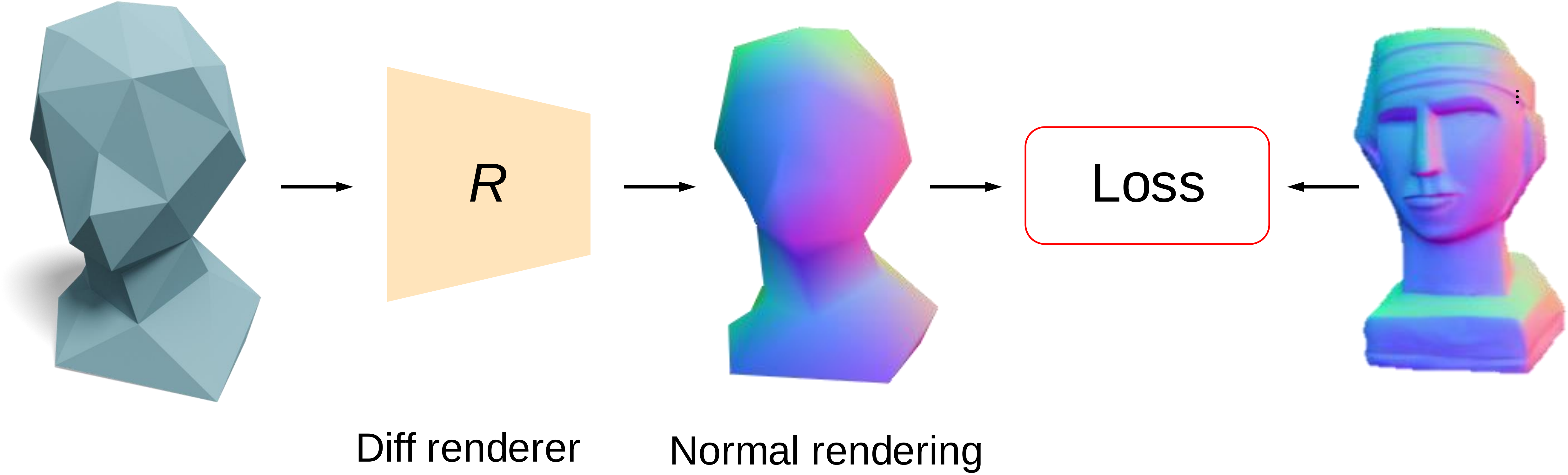


Fine-tuned Wonder3D with additional input (blured normal)

Option 1: Reconstruct via Differentiable Renderer optimization

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

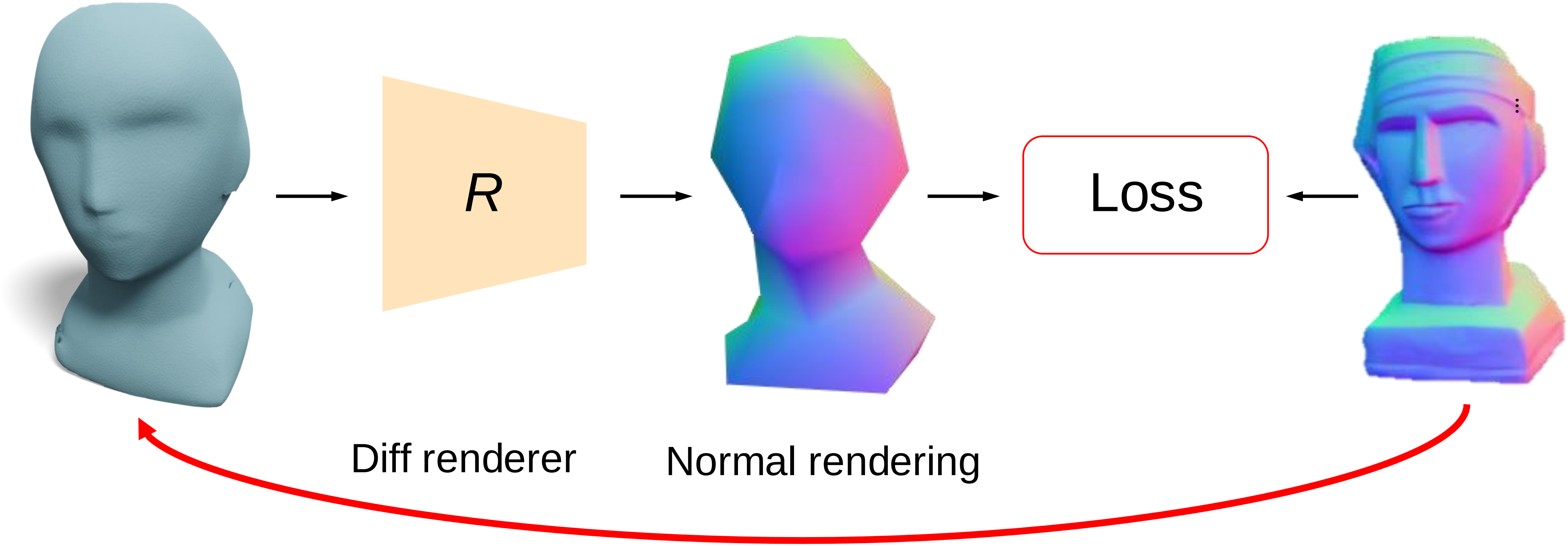
Mesh Optimization



Adobe

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

Mesh Optimization



R

Loss

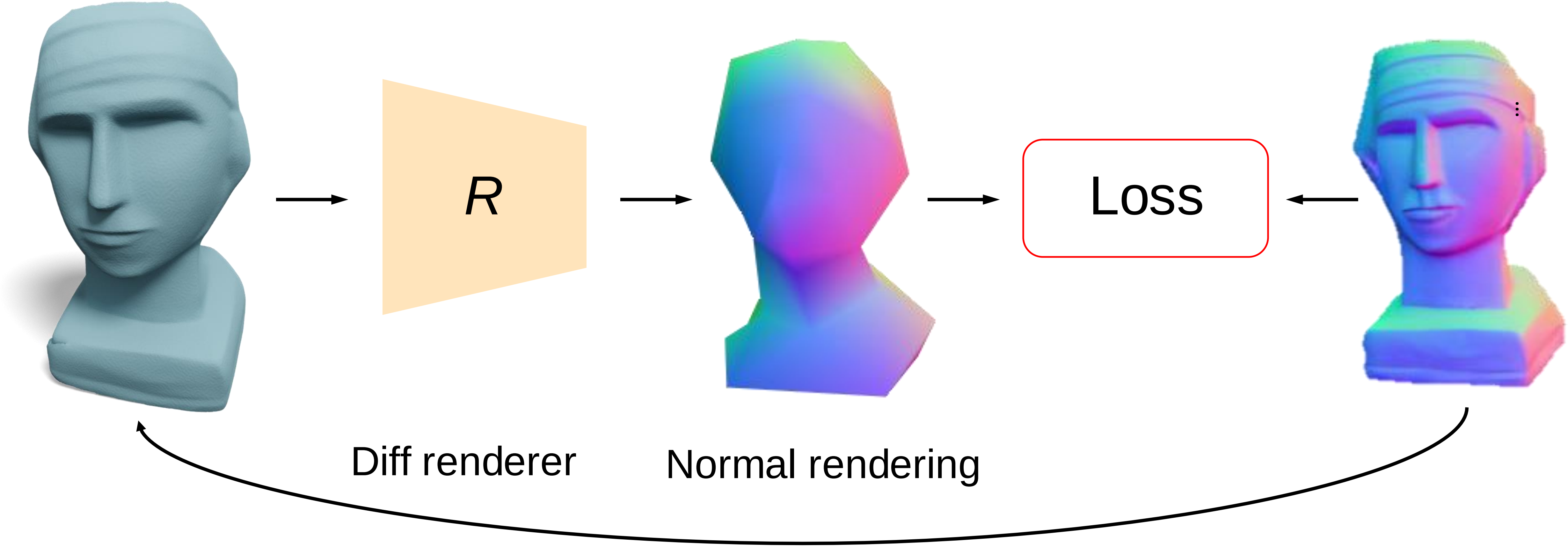
Diff renderer

Normal rendering

Gradient Descent

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

Mesh Optimization

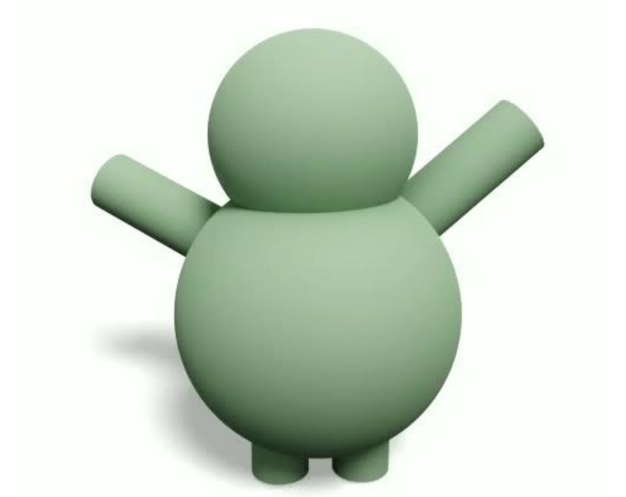


Gradient Descent

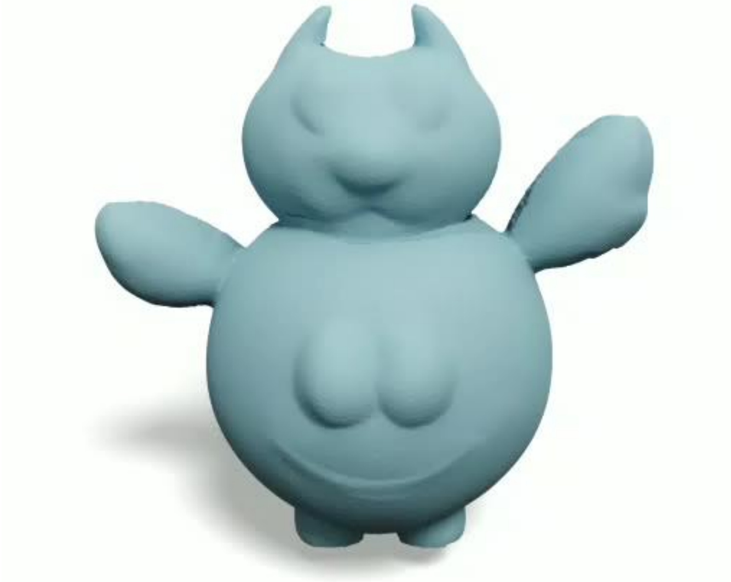
Adobe

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

>_: A cartoon figure

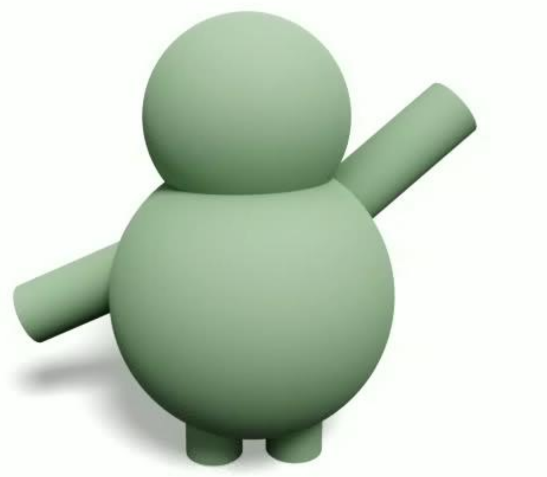


Input 3D mesh

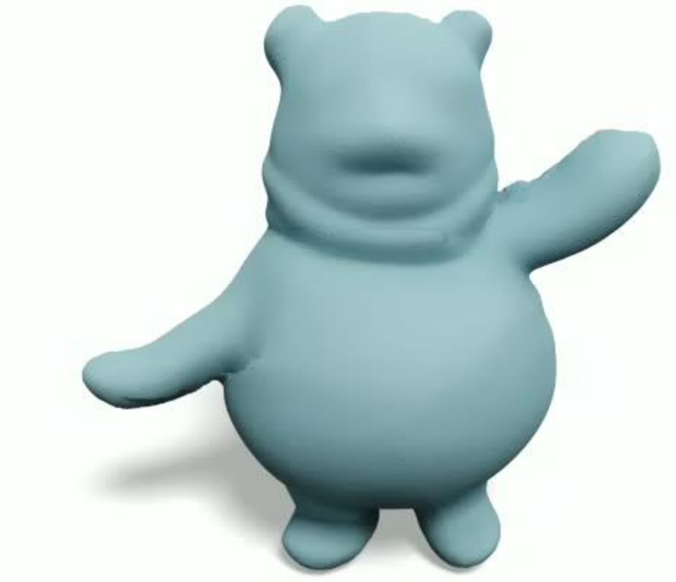


Output 3D mesh

>_: A teddy bear

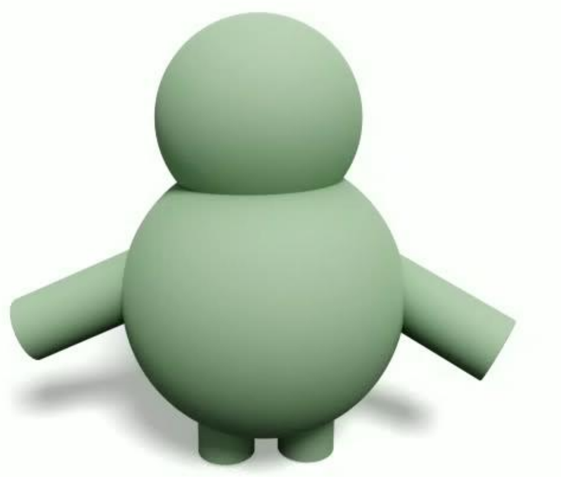


Input 3D mesh

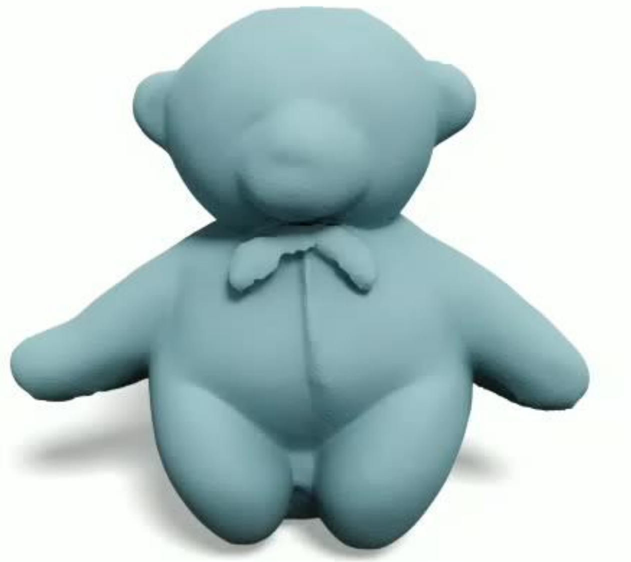


Output 3D mesh

>_: A teddy bear



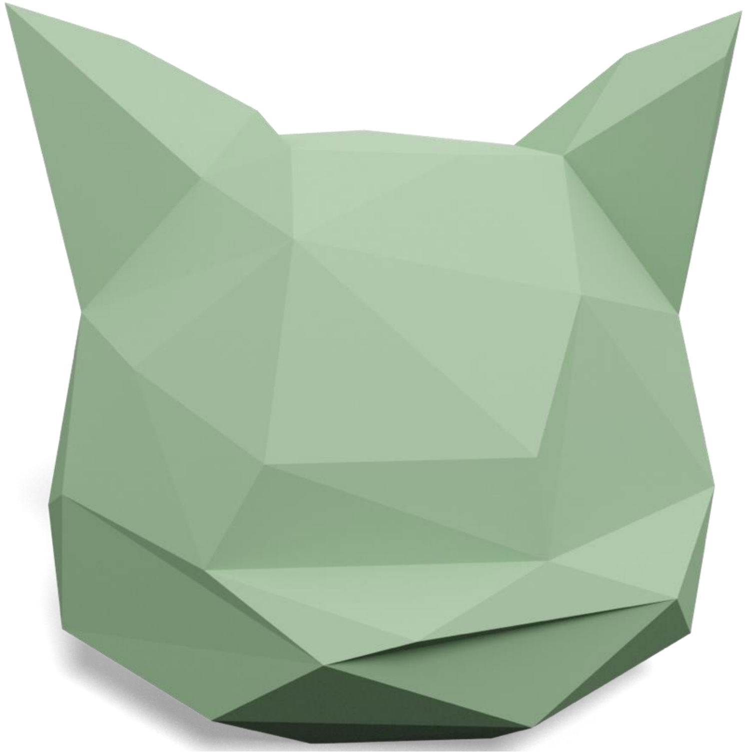
Input 3D mesh



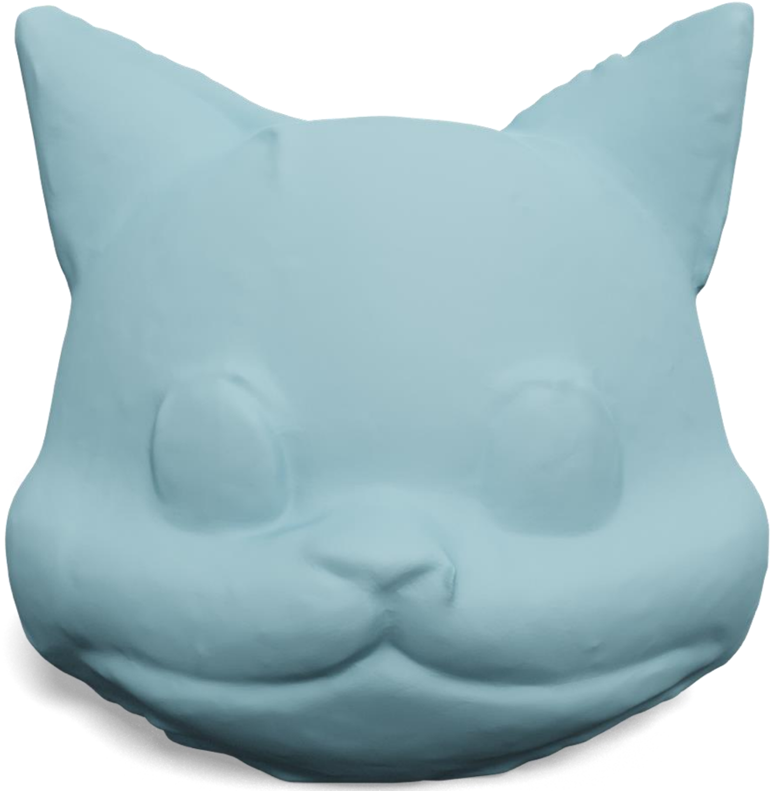
Output 3D mesh

Mesh Texturing

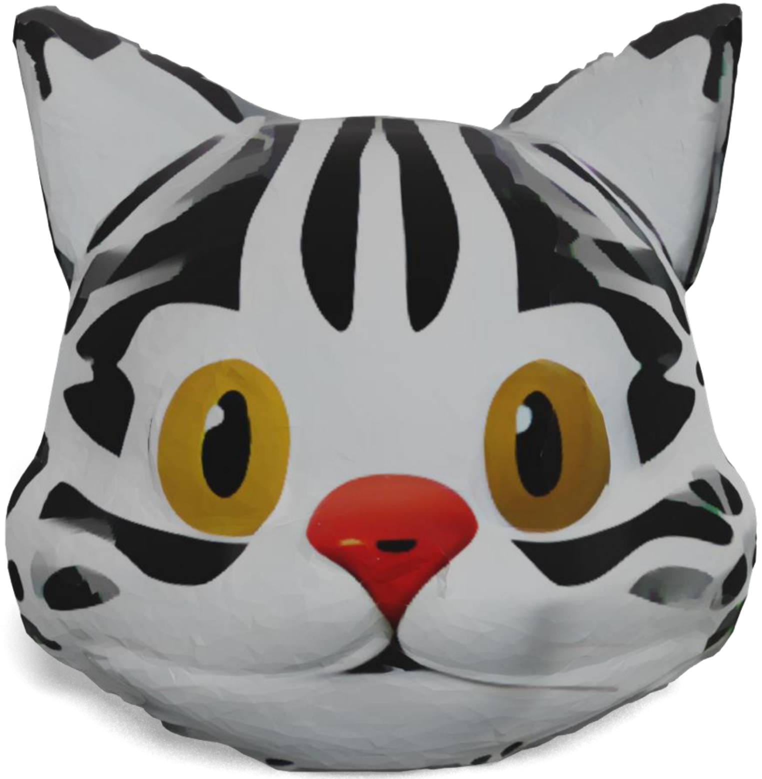
>_: A cartoon cat head



Input 3D mesh



Geometry

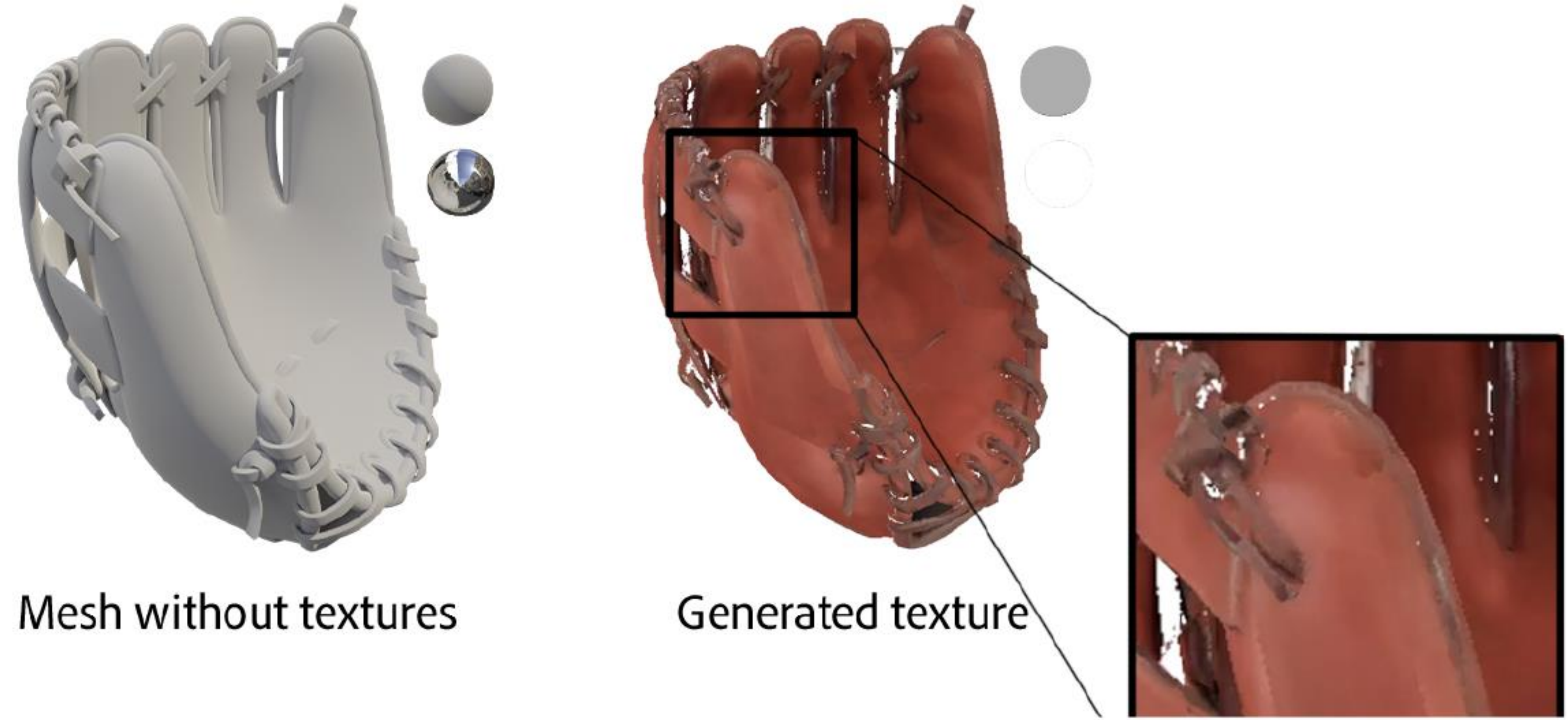


Texture

Text-Guided Refinement, Yun-Chun Chen et al., SIGGRAPH A, 2024

More on Texturing

- Textures are not just colors



Mesh without textures

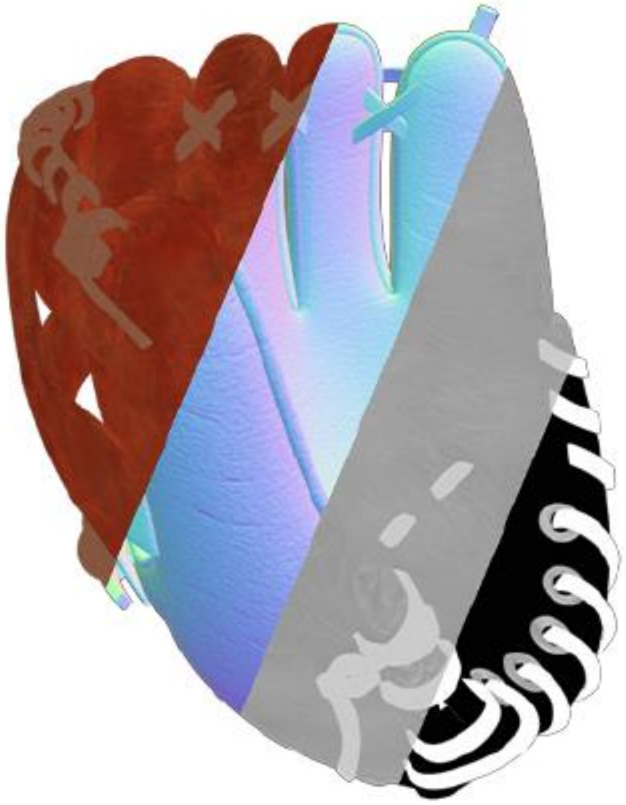
Generated texture

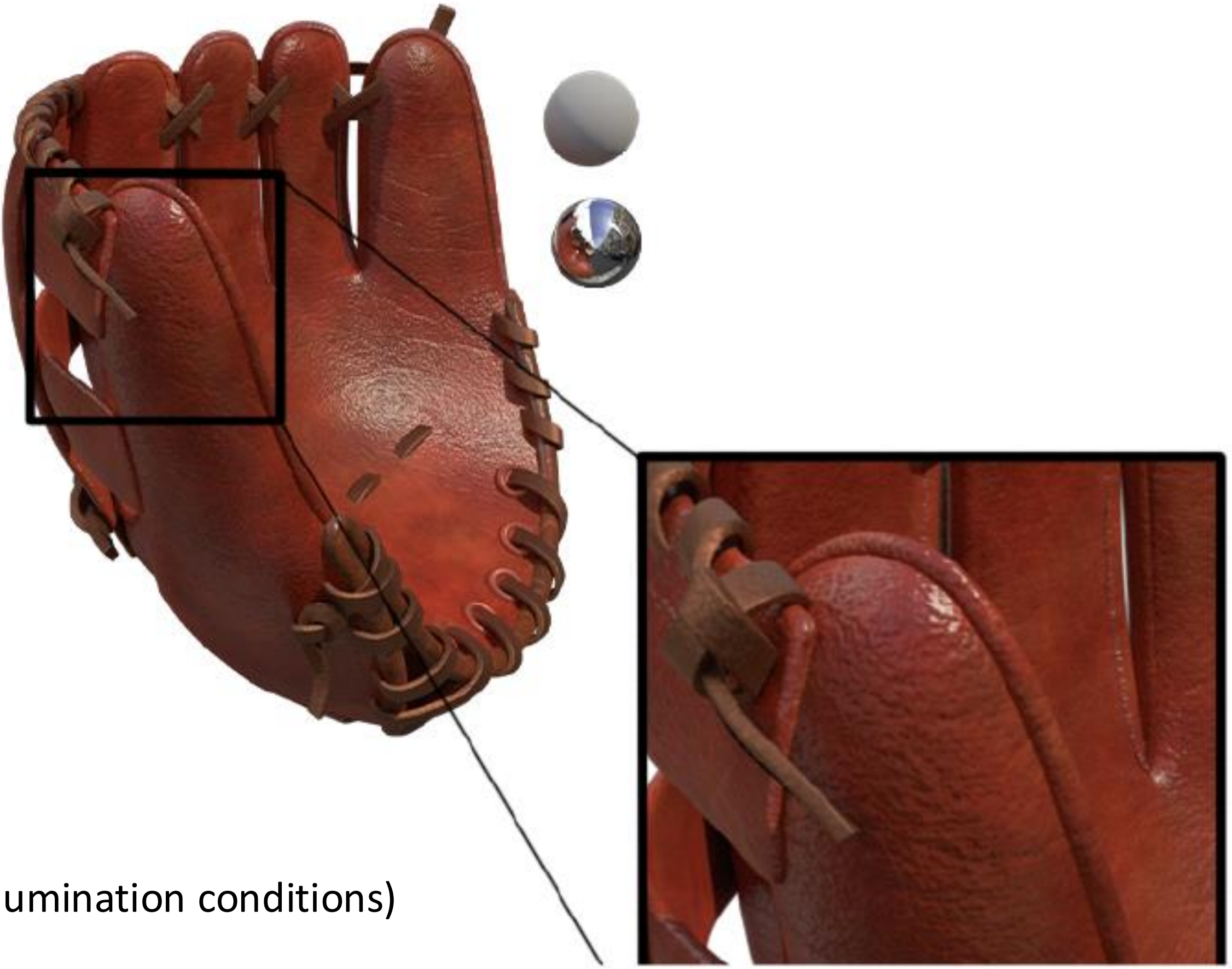
I.

More on Texturing

- Textures are not just colors

Mesh without textures



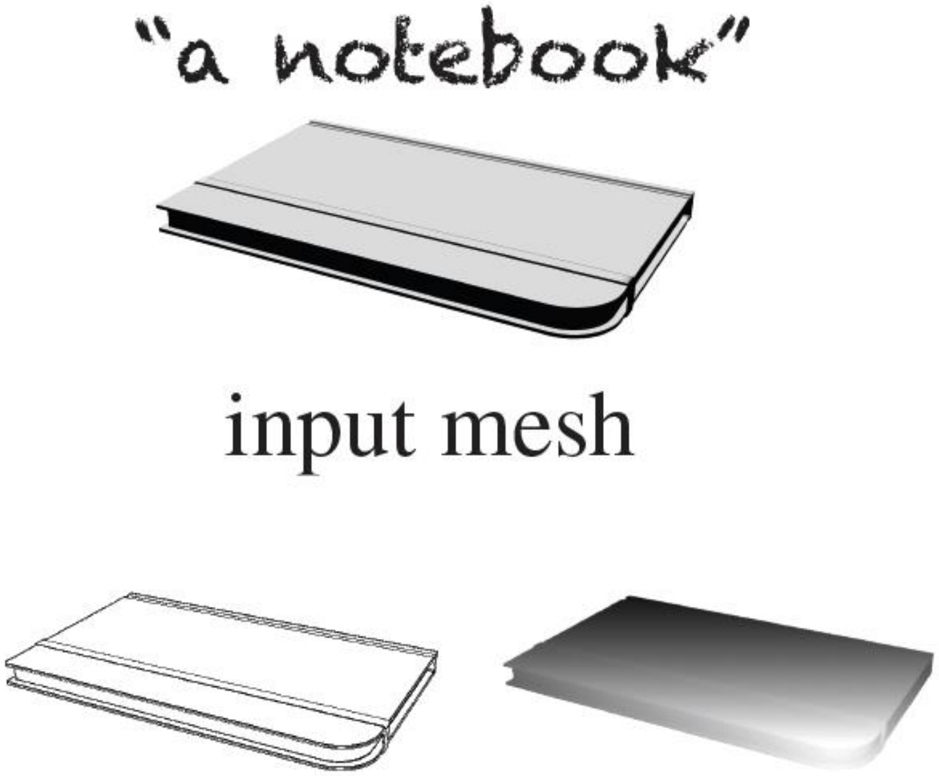


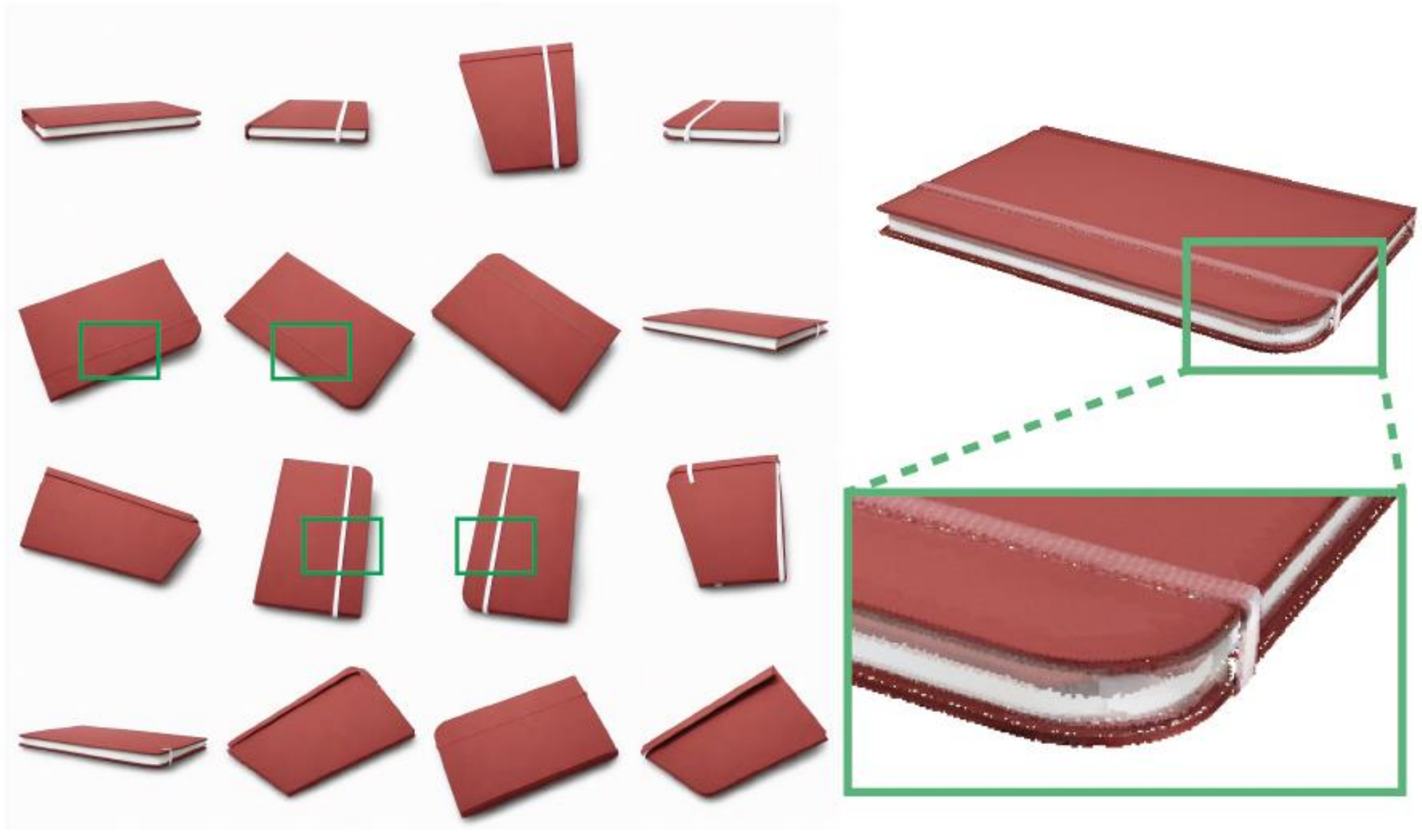
Material maps

(works with different illumination conditions)

More on Texturing

- Multi-view generation

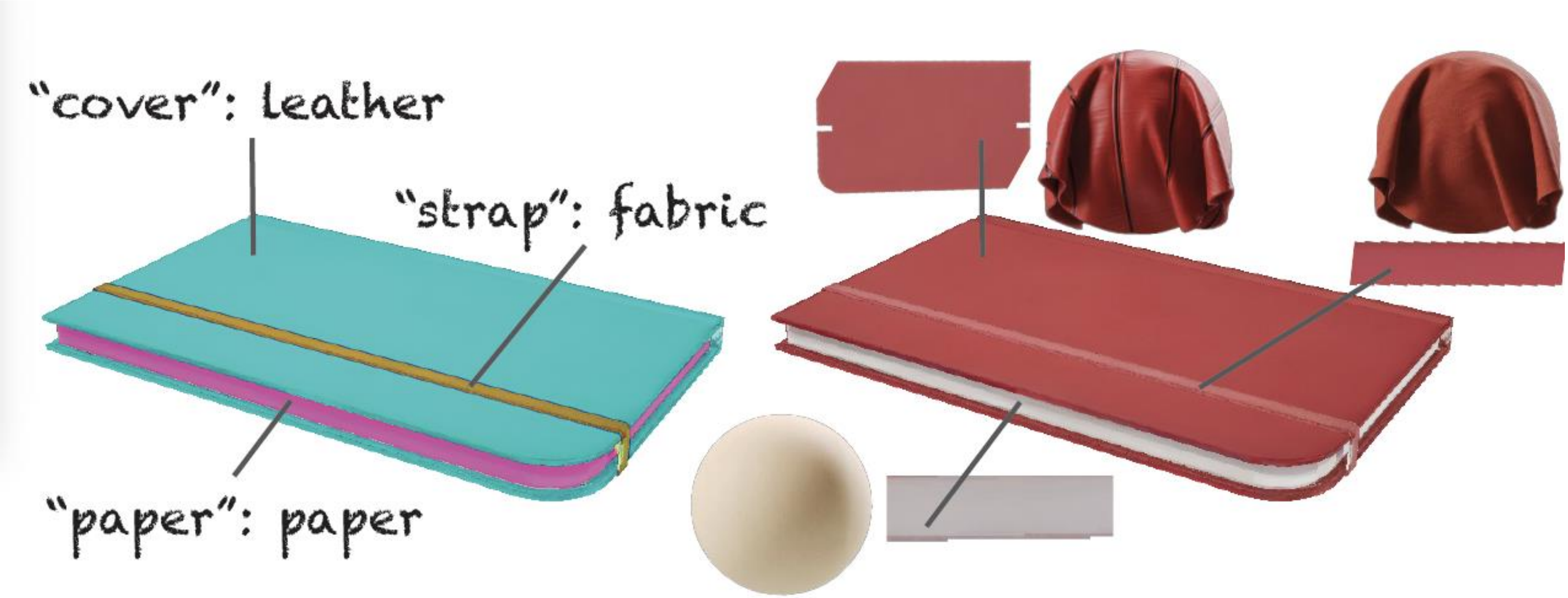




AutoTexturing, Ceylan et al. 2024

More on Texturing

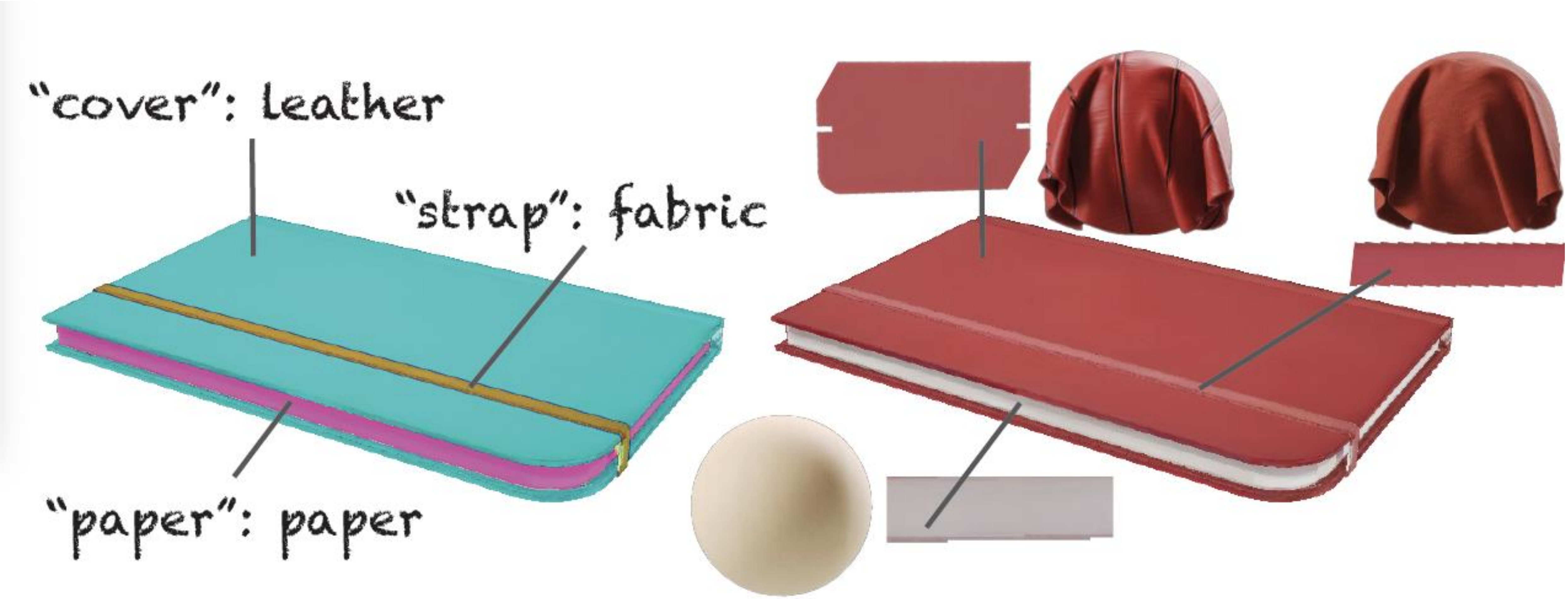
- VLLM-assisted retrieval



MatAtlas, Ceylan et al. 2024

More on Texturing – But How Do We Segment?

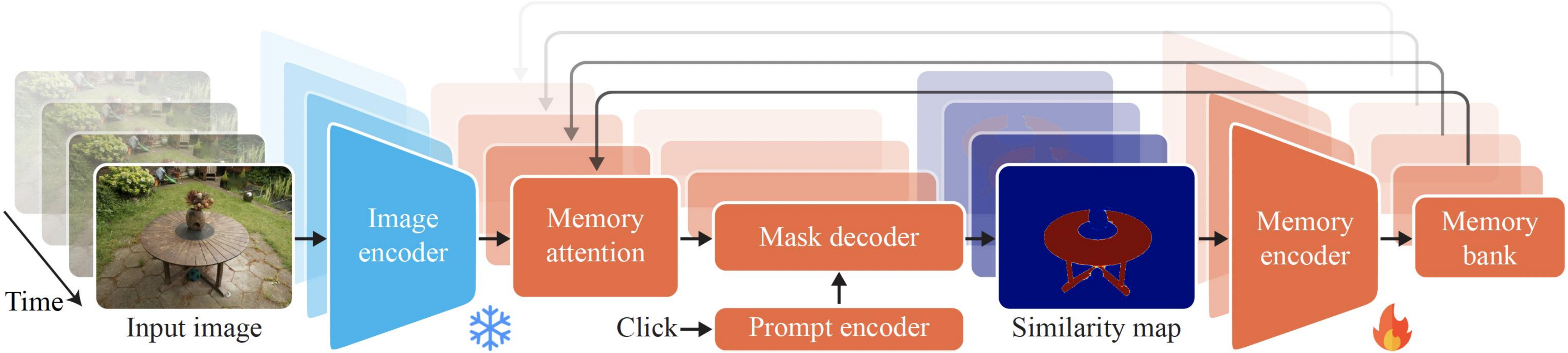
- VLLM-assisted retrieval



MatAtlas, Ceylan et al. 2024

Material-aware 3D Segmentation

- Fine-tune SAM2 (object segmentation in video) to segment materials

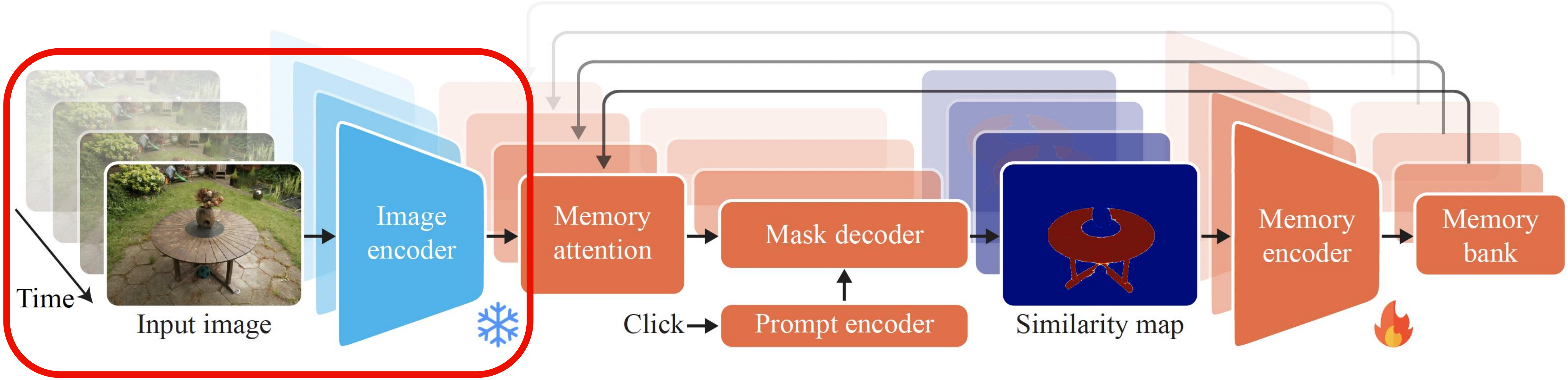


Adobe

SAMa, Fischer et al., In Submission, 2025

Material-aware 3D Segmentation

- Fine-tune SAM2 (object segmentation in video) to segment materials

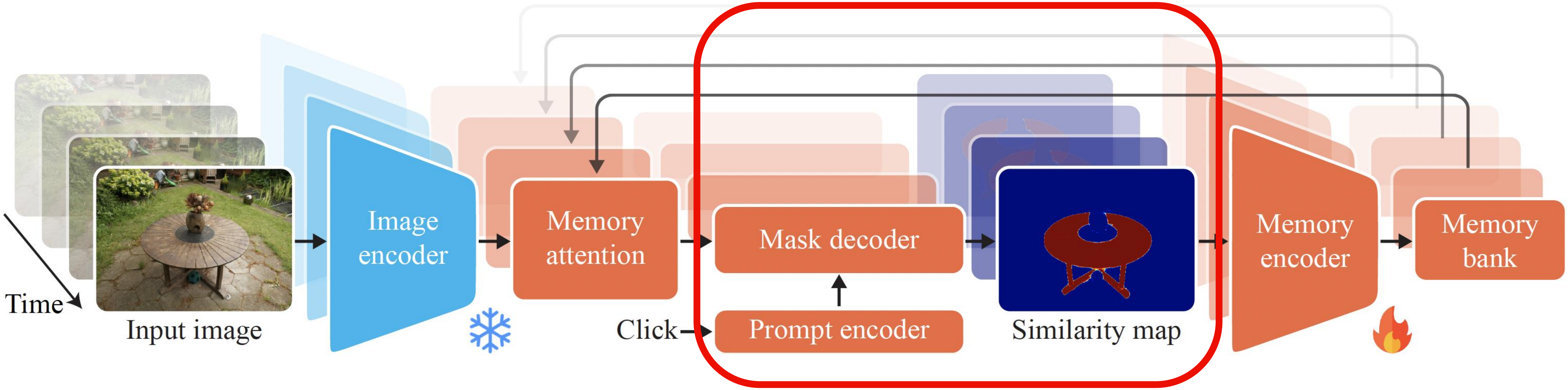


Pre-trained VIT Image Encoder (frozen)

SAMa, Fischer et al., In Submission, 2025

Material-aware 3D Segmentation

- Fine-tune SAM2 (object segmentation in video) to segment materials



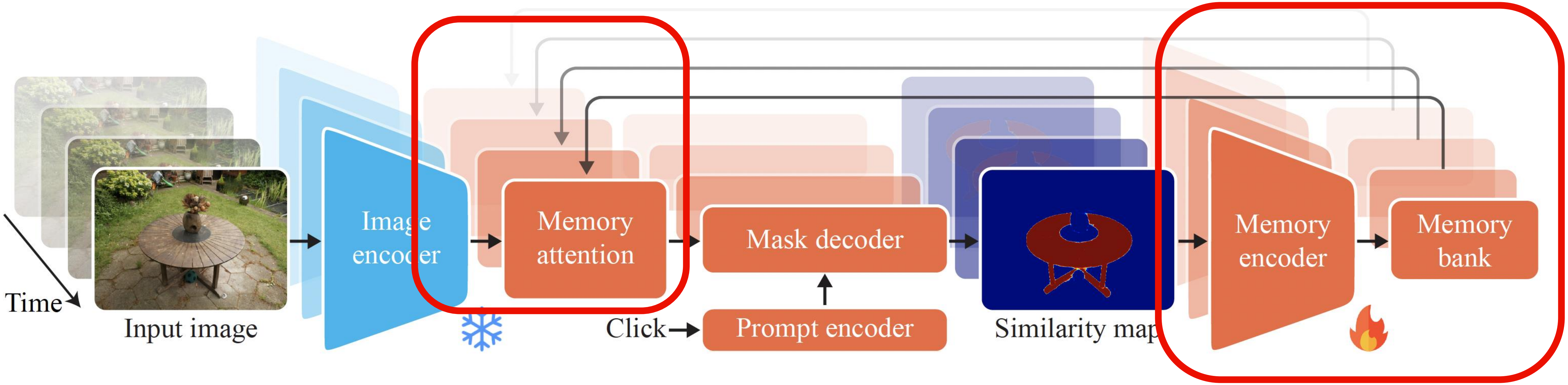
Fine-tune mask decoder with point selection conditioning (train on material data)

Adobe

SAMa, Fischer et al., In Submission, 2025

Material-aware 3D Segmentation

- Fine-tune SAM2 (object segmentation in video) to segment materials

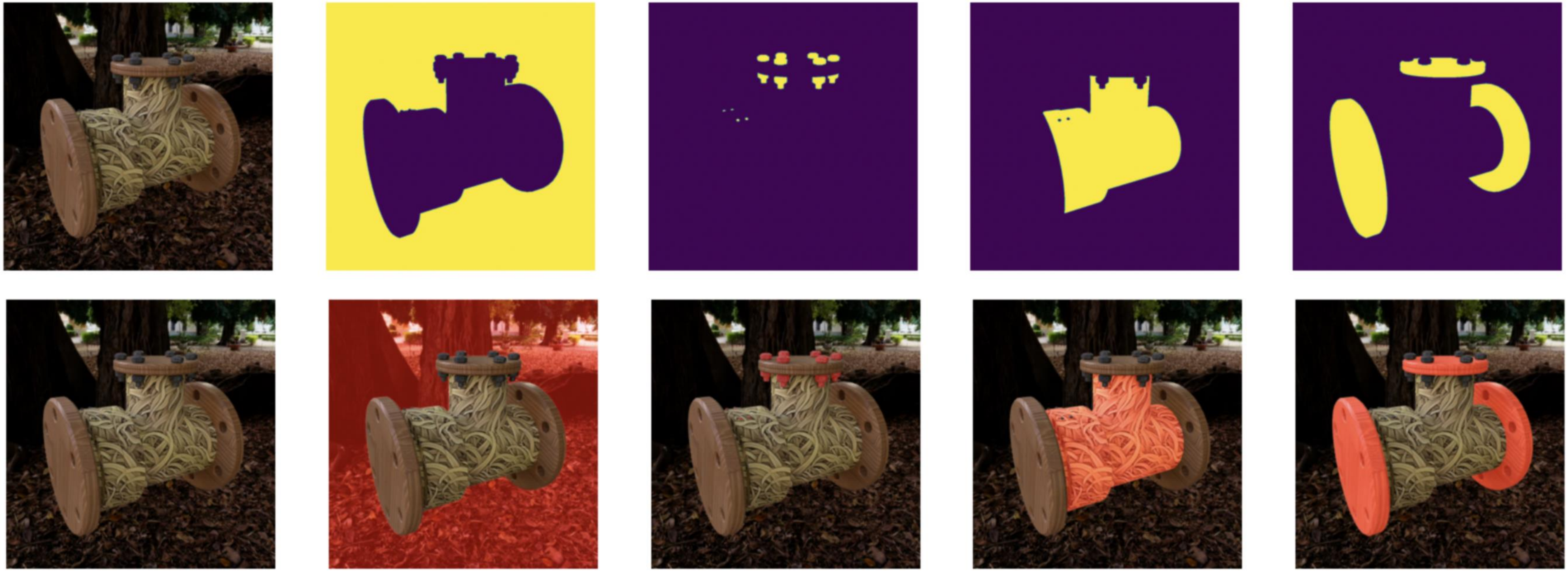


Fine-tune memory attention (improves consistency)

SAMa, Fischer et al., In Submission, 2025

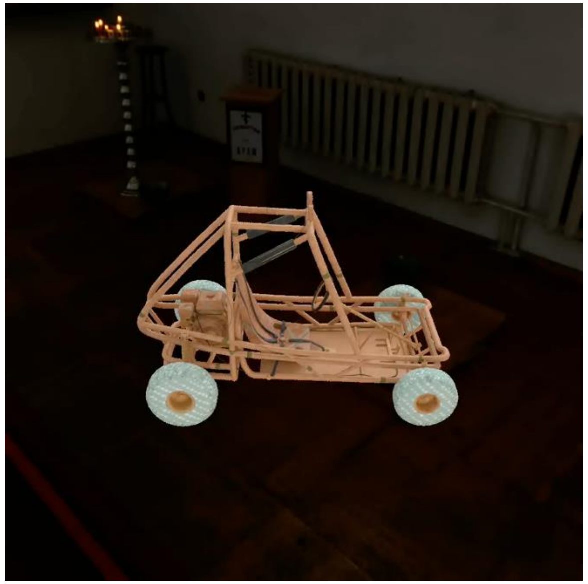
Synthetic Training Data for Material-aware Segmentation

- Per-frame masks



SAMa, Fischer et al., In Submission, 2025

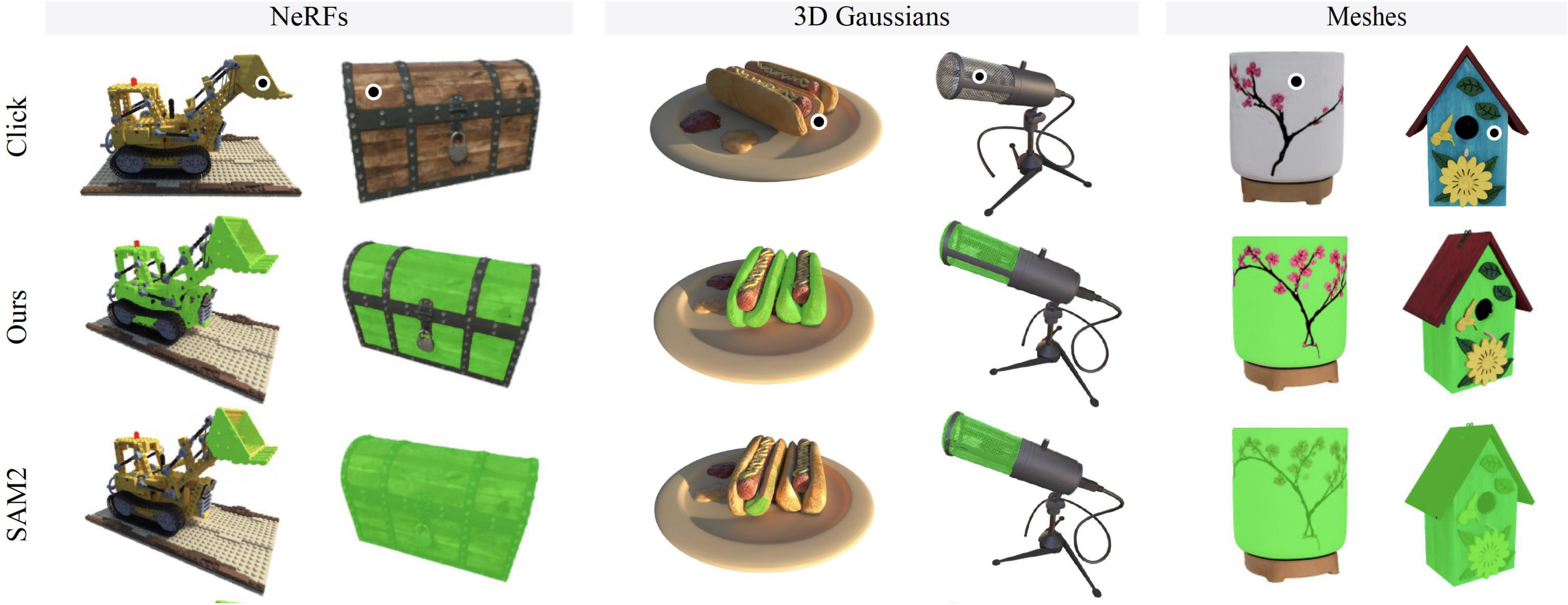
Synthetic Training Data for Material-aware Segmentation



SAMa, Fischer et al., In Submission, 2025

Object vs Material Segmentation

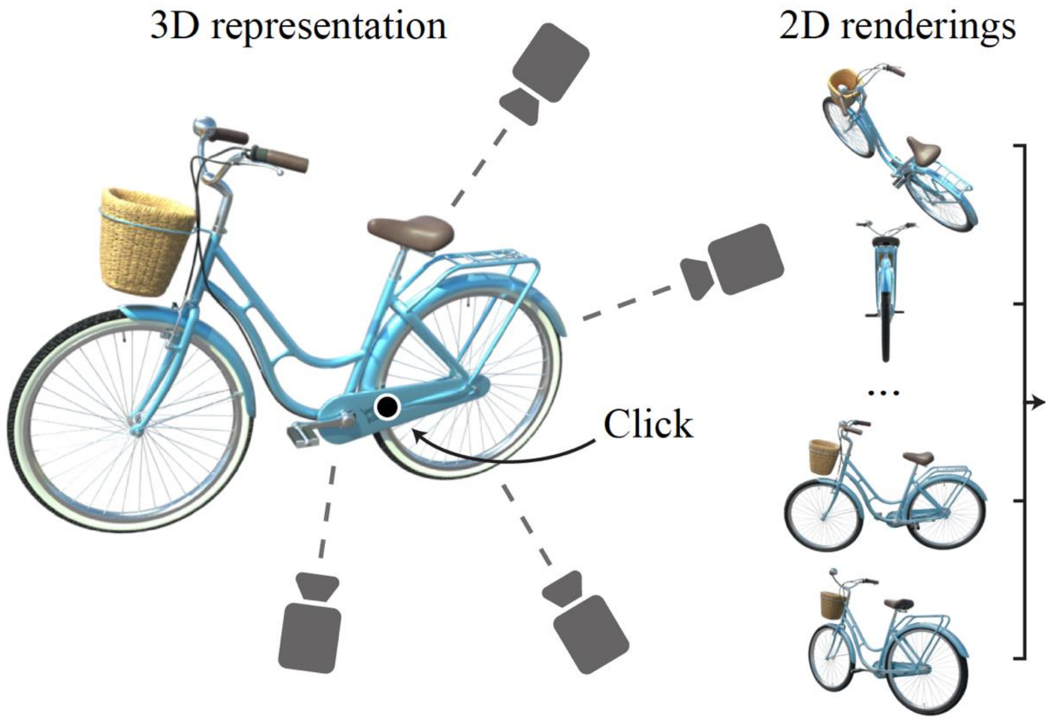
- SAM2 is not suitable for material-aware segmentation without fine-tuning



SAMa, Fischer et al., In Submission, 2025

Inferring 3D Segmentations

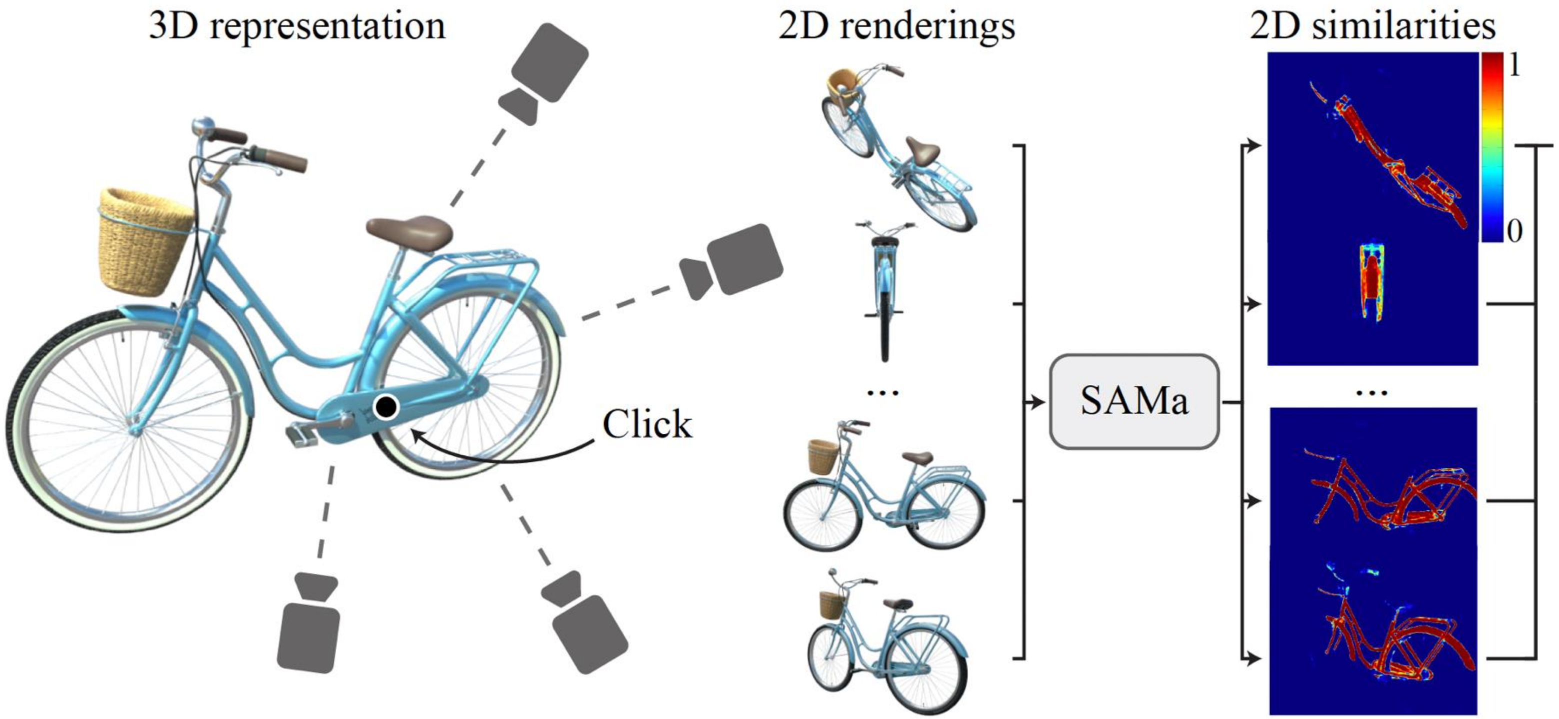
- Sample 25 views



SAMa, Fischer et al., In Submission, 2025

Inferring 3D Segmentations

- Sample 25 views => predict 2D features

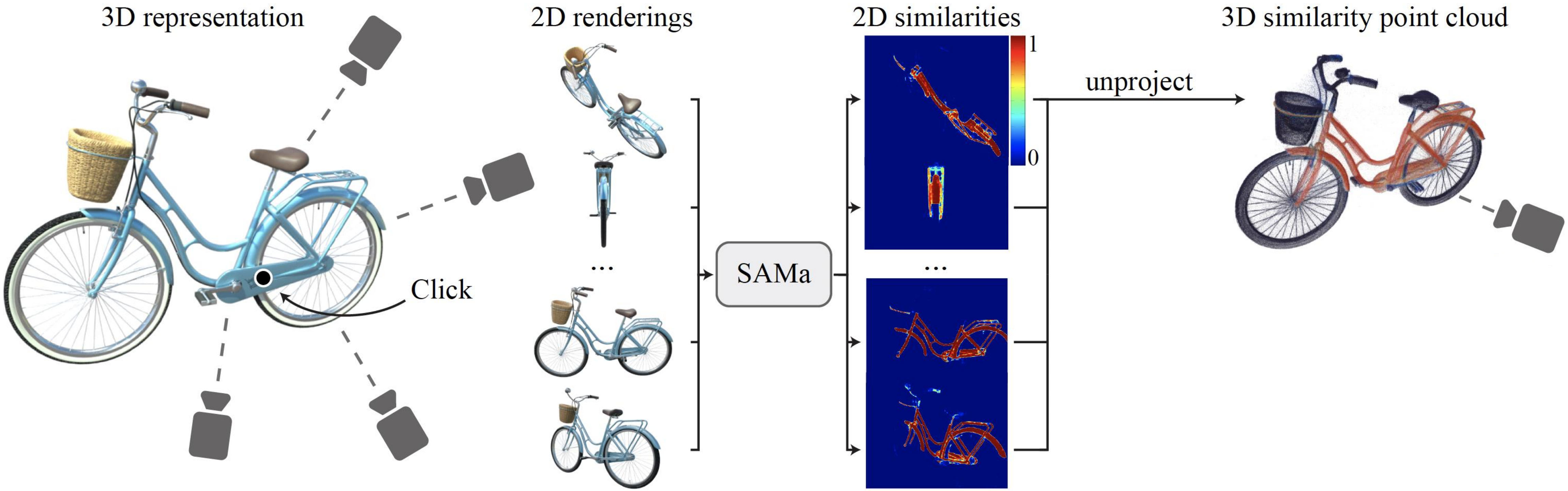


Adobe

SAMa, Fischer et al., In Submission, 2025

Inferring 3D Segmentations

- Sample 25 views => predict 2D features => store features in 3D point cloud

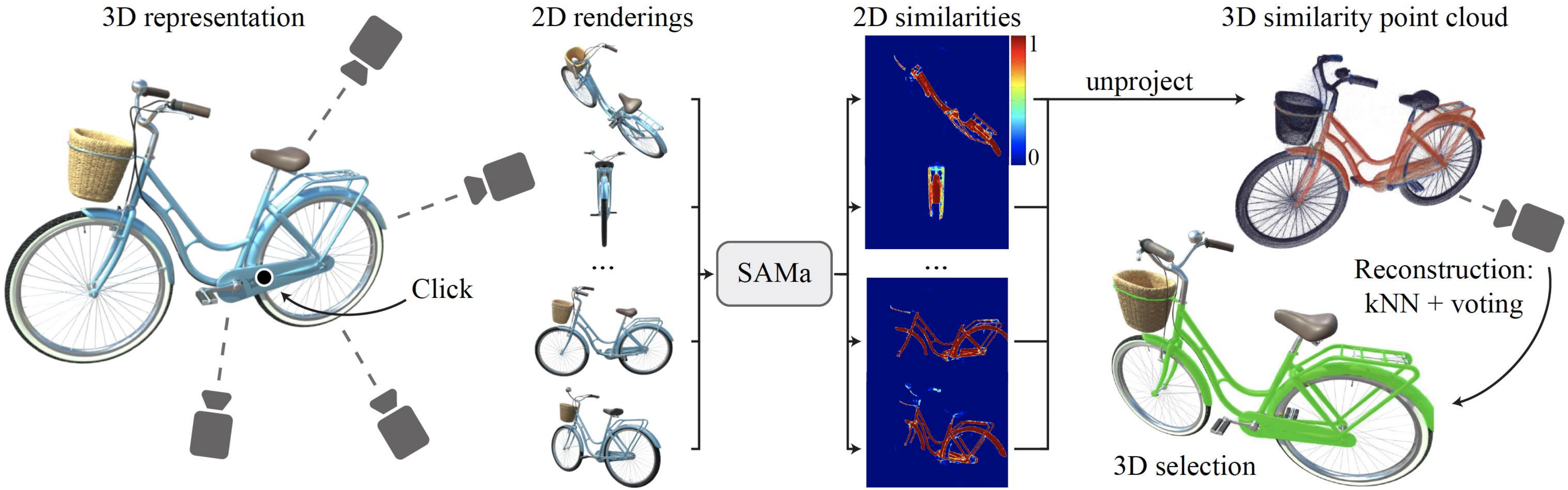


Adobe

SAMa, Fischer et al., In Submission, 2025

Inferring 3D Segmentations

- Sample 25 views => predict 2D features => store features in 3D point cloud
- Given user click => select via kNN voting

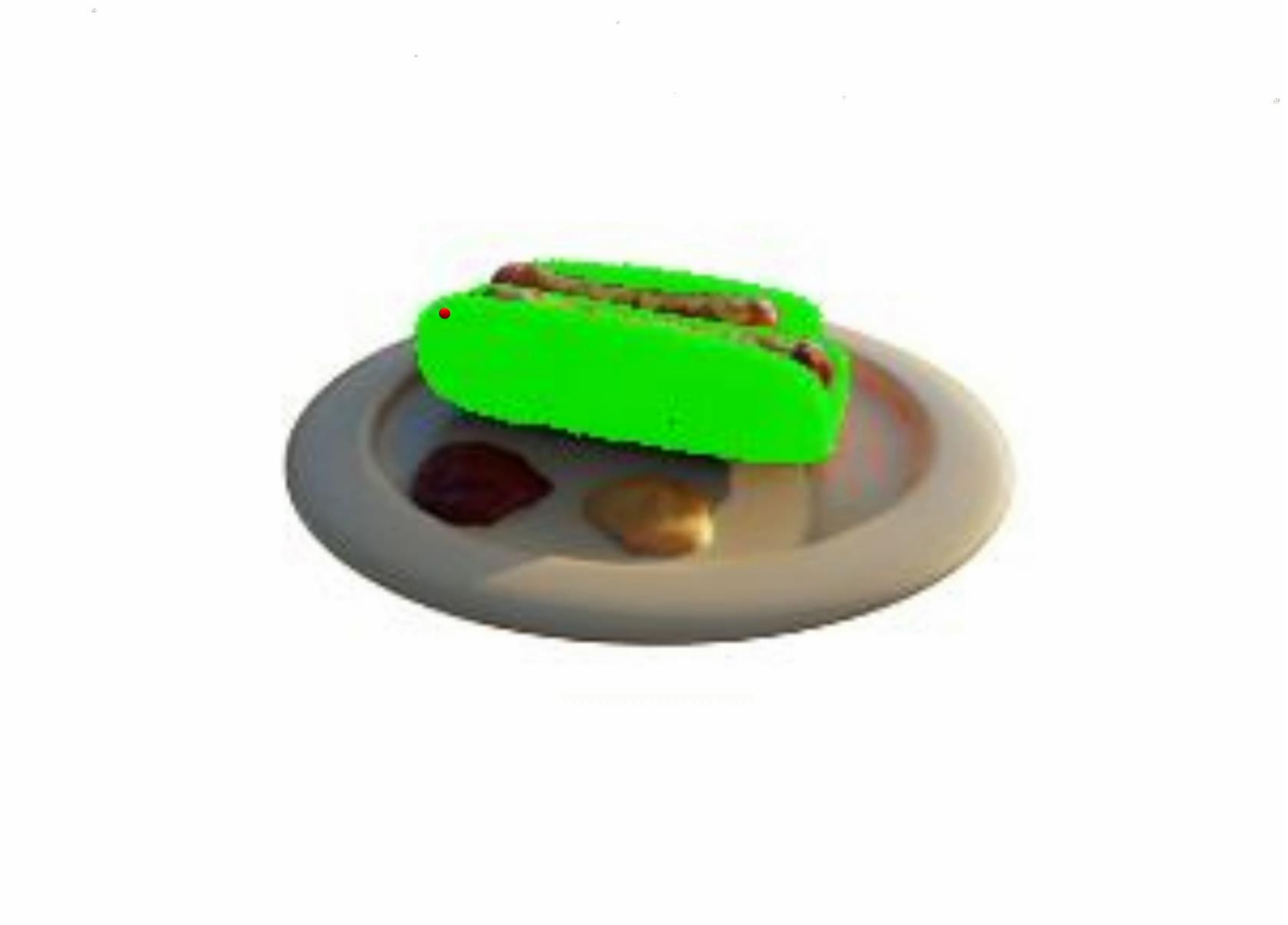


Adobe

SAMa, Fischer et al., In Submission, 2025

Interactive Segmentation

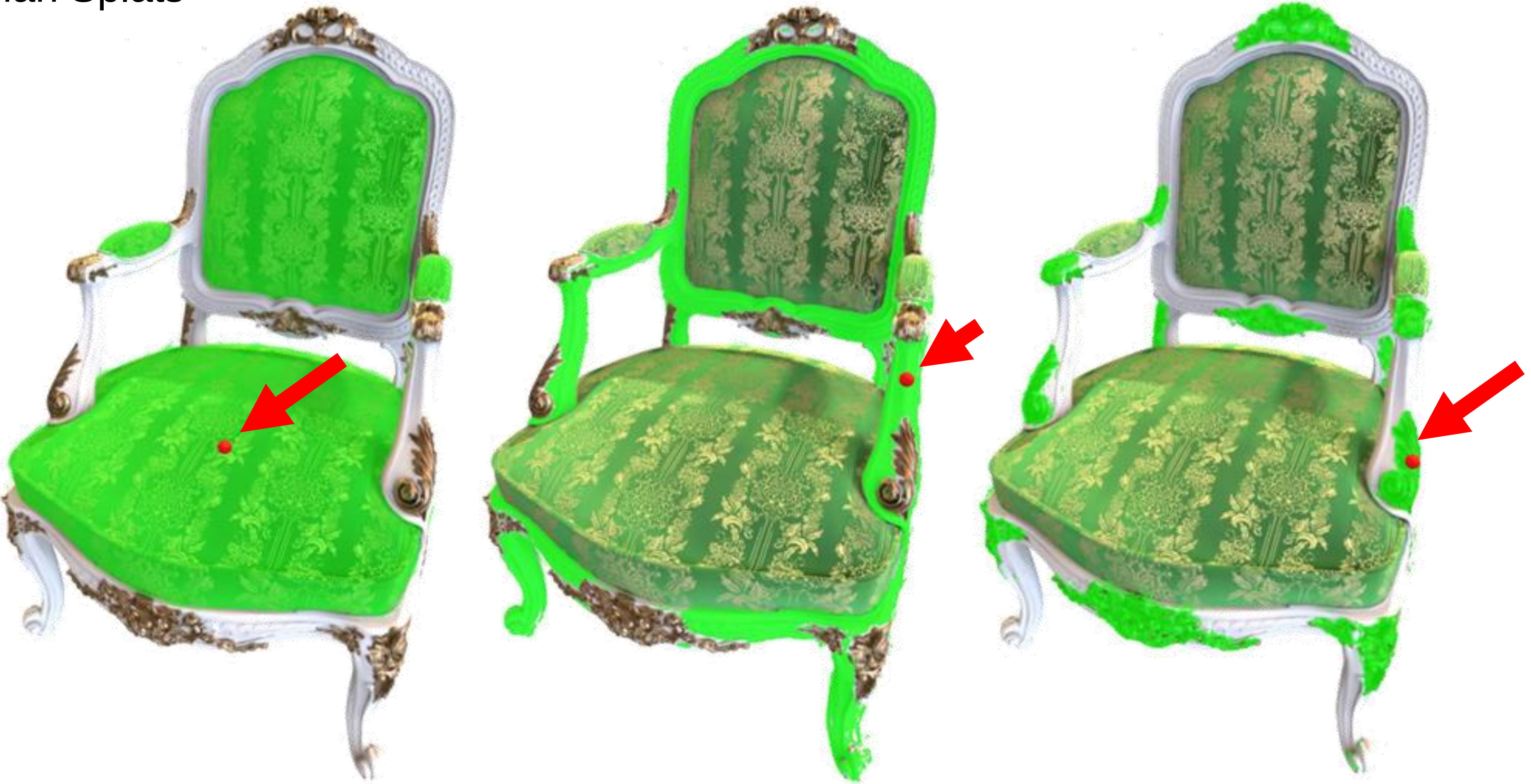
- At runtime selection can be done in 2ms per click



SAMa, Fischer et al., In Submission, 2025

Representation-Agnostic

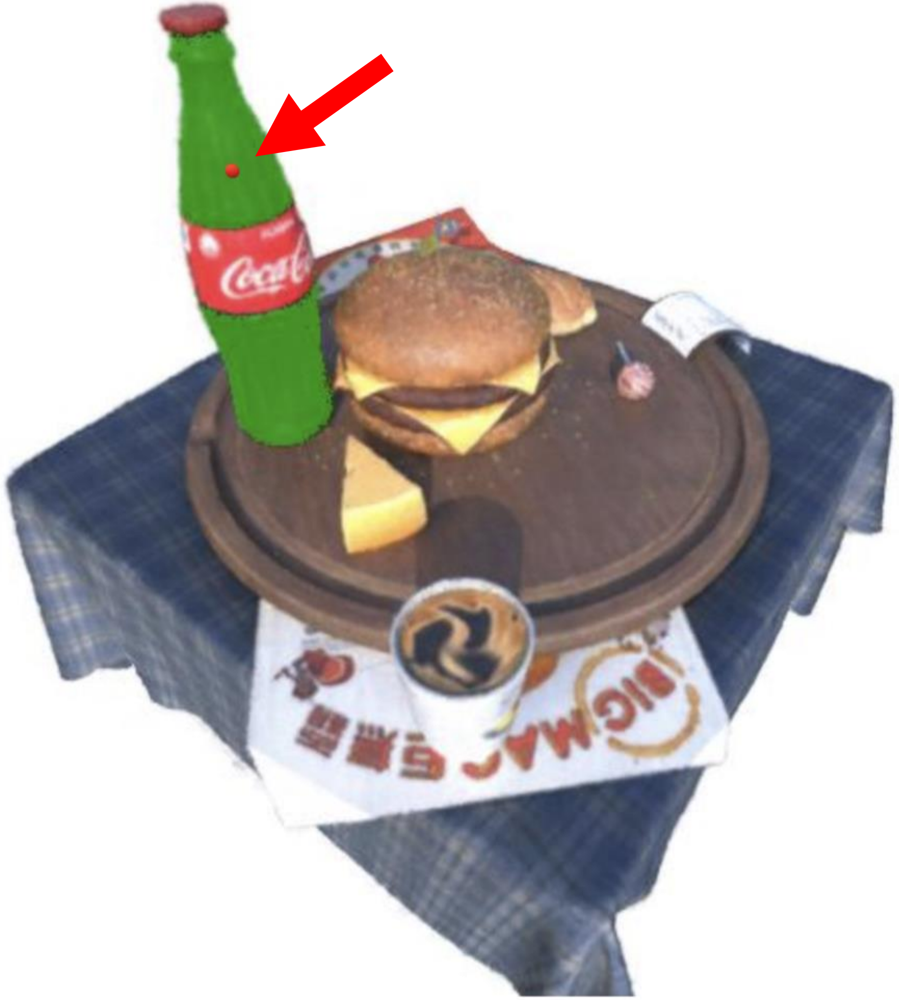
- Gaussian Splats

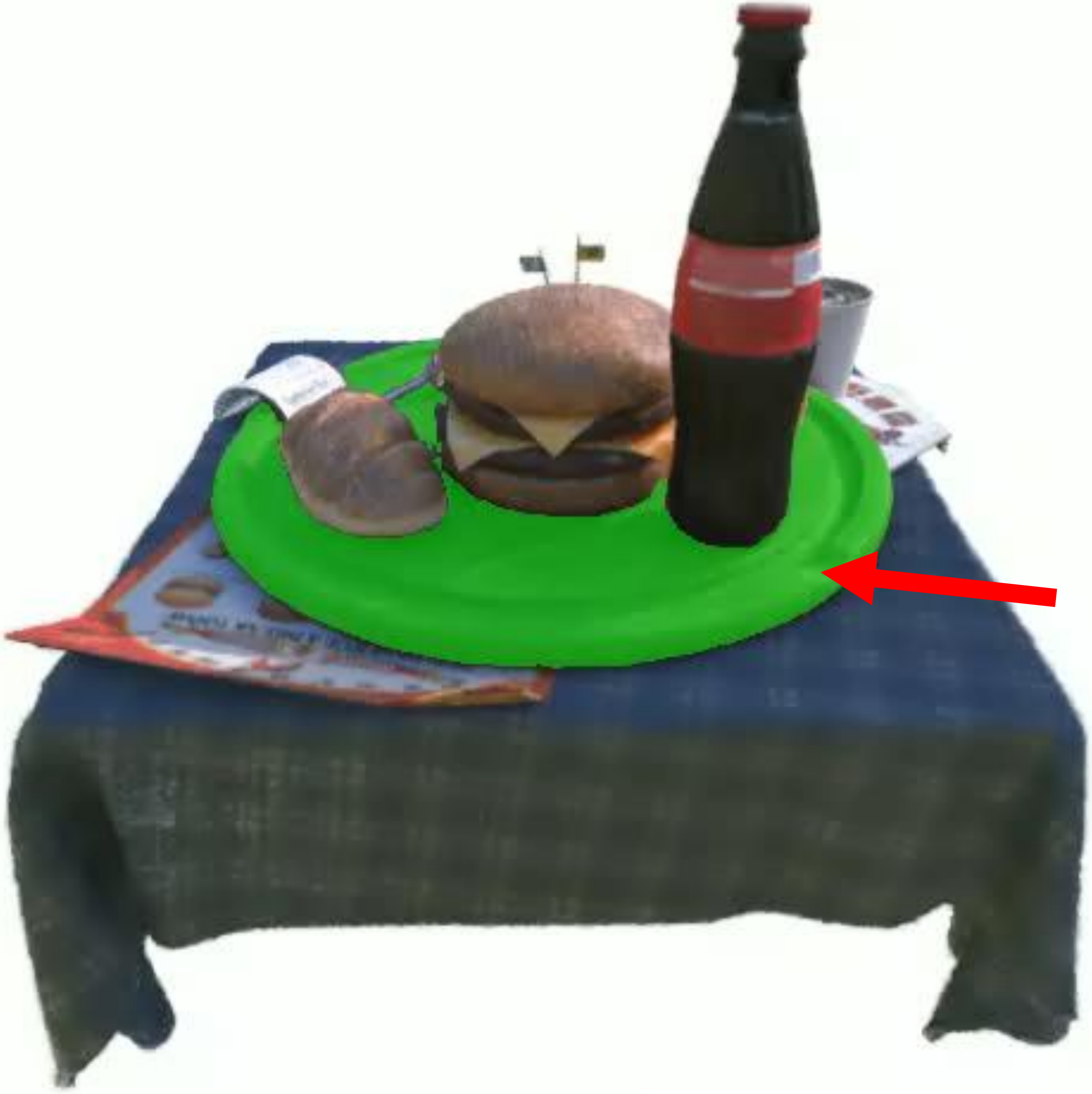


SAMa, Fischer et al., In Submission, 2025

Representation-Agnostic

- NeRFs

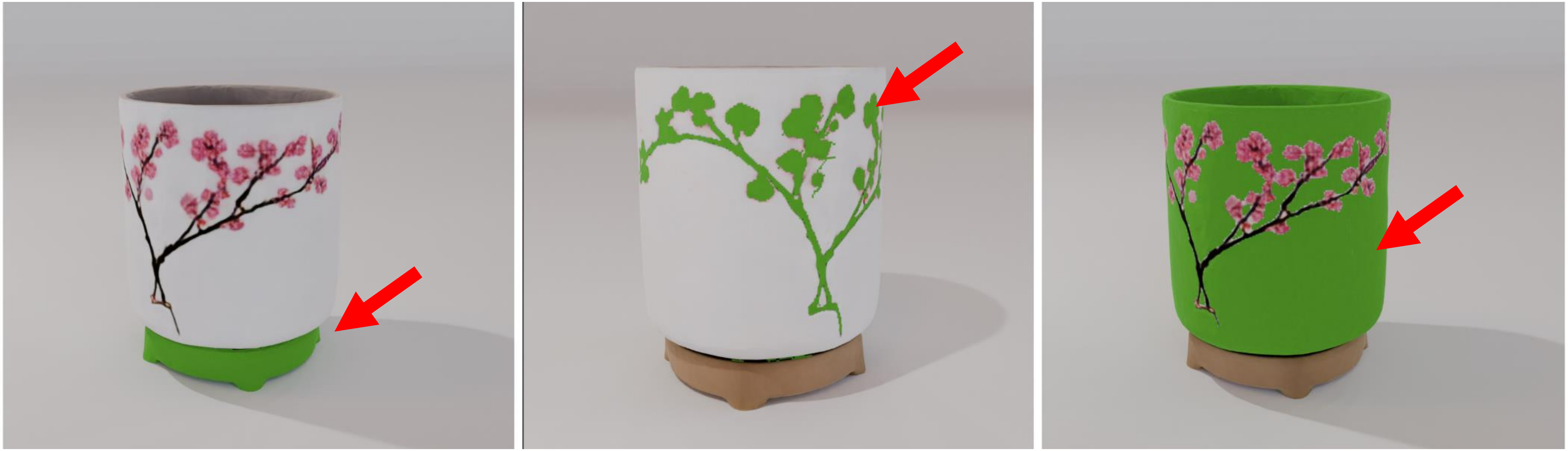




SAMa, Fischer et al., In Submission, 2025

Representation-Agnostic

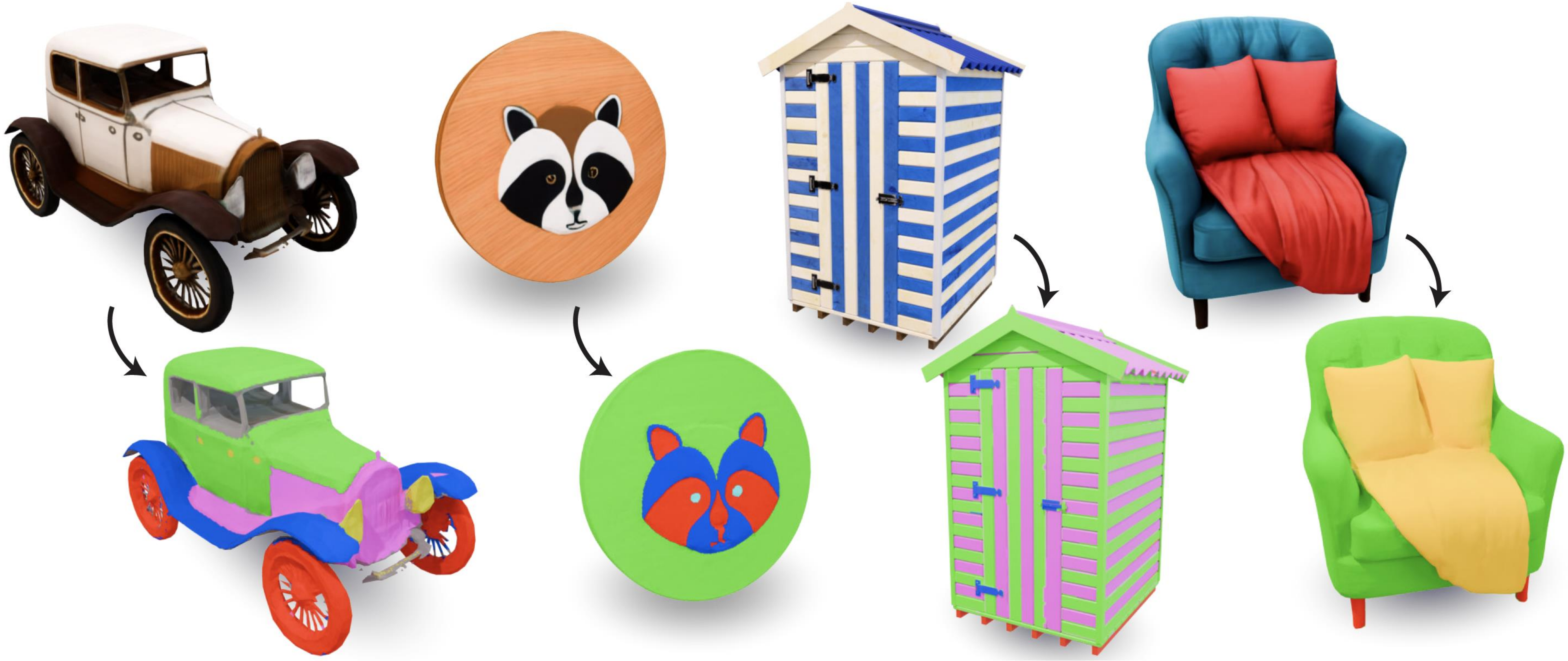
- Meshes



SAMa, Fischer et al., In Submission, 2025

Automatic Segmentation

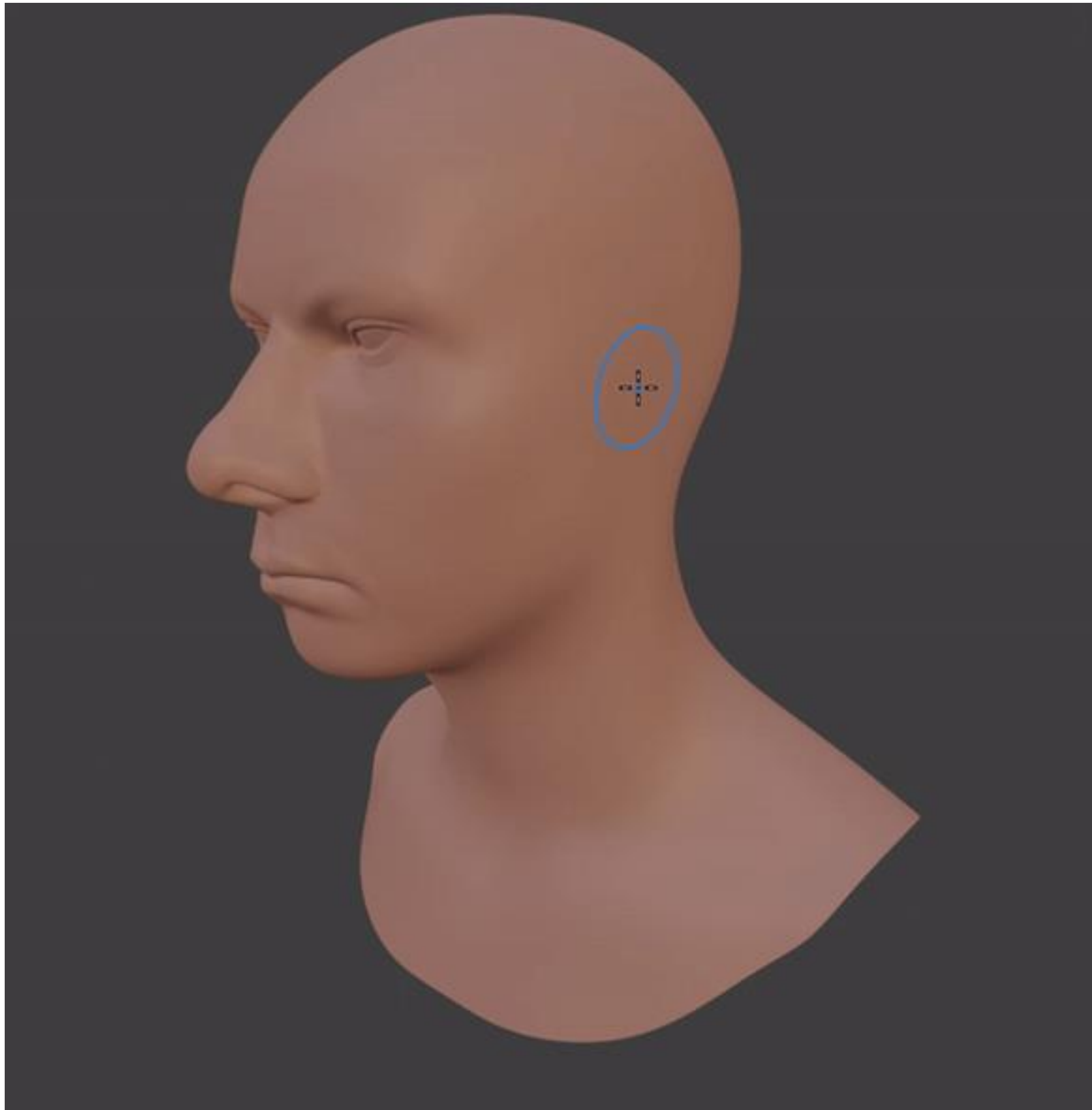
- Simulate many clicks => drop regions with high overlap



SAMa, Fischer et al., In Submission, 2025

Spatial Controlling the Details

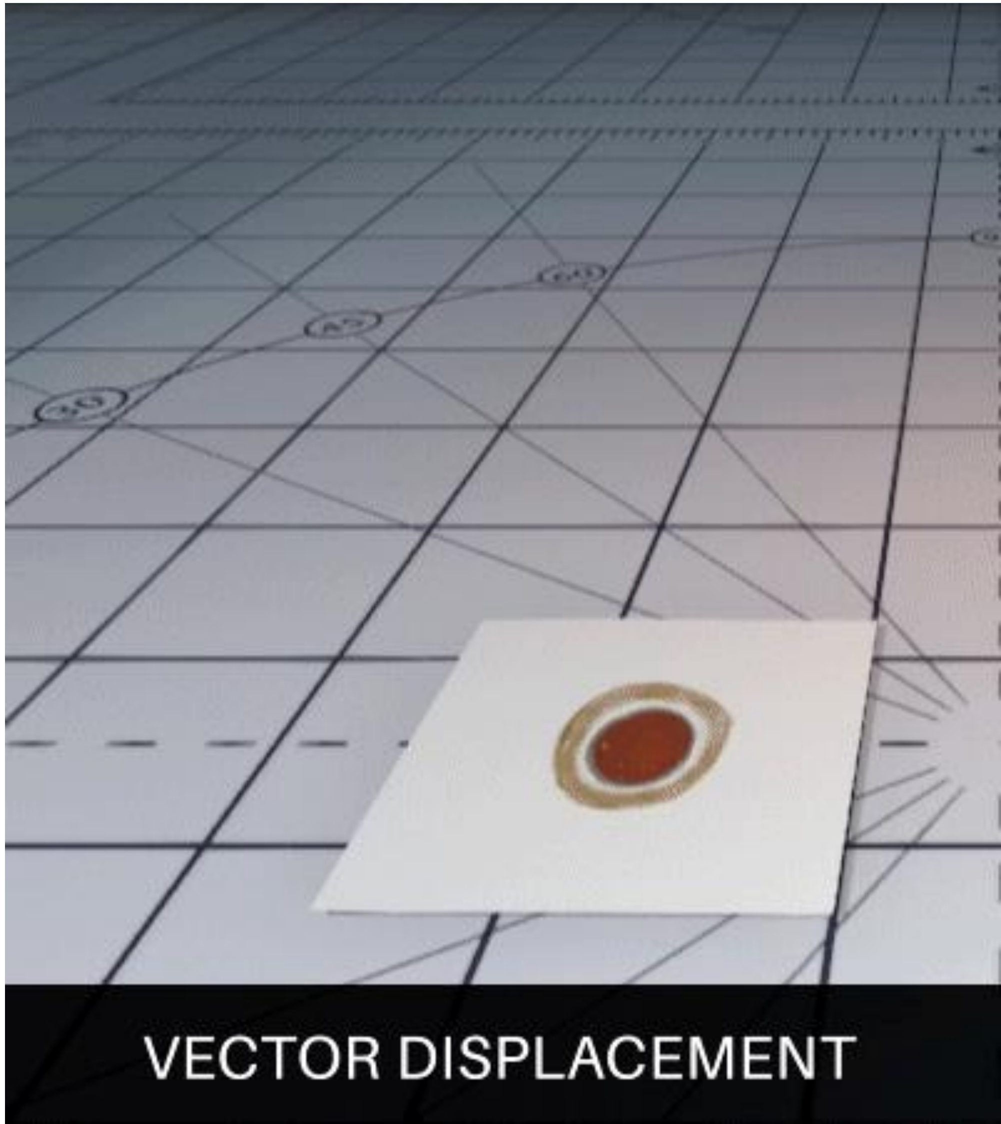
- Precisely position the ear on the head



Adobe

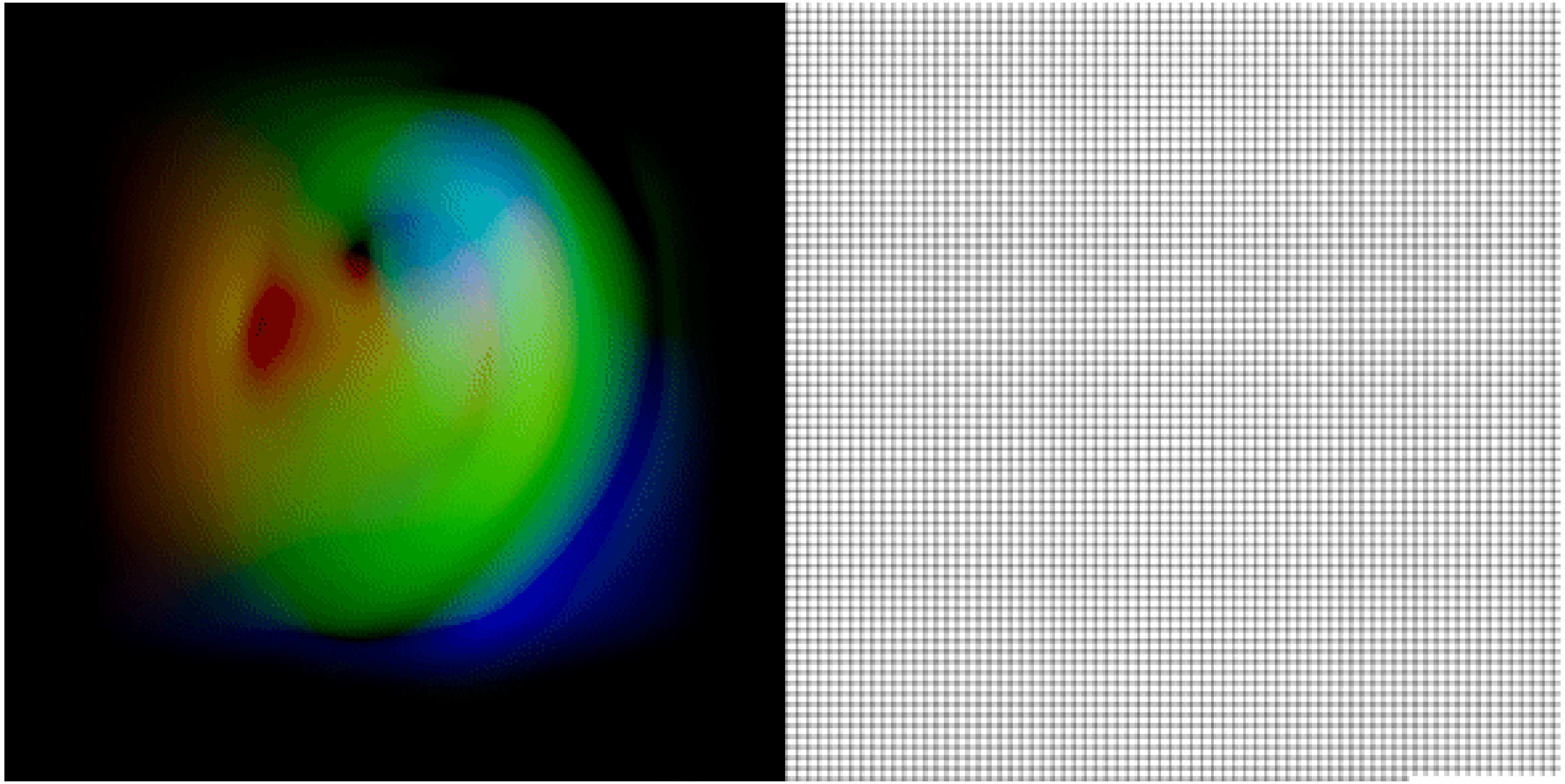
Vector Displacement Map (VDM)

- Map over 2D Plane: $f \ : [0,1]^2 \to \mathbb{R}^3$



Vector Displacement Map (VDM)

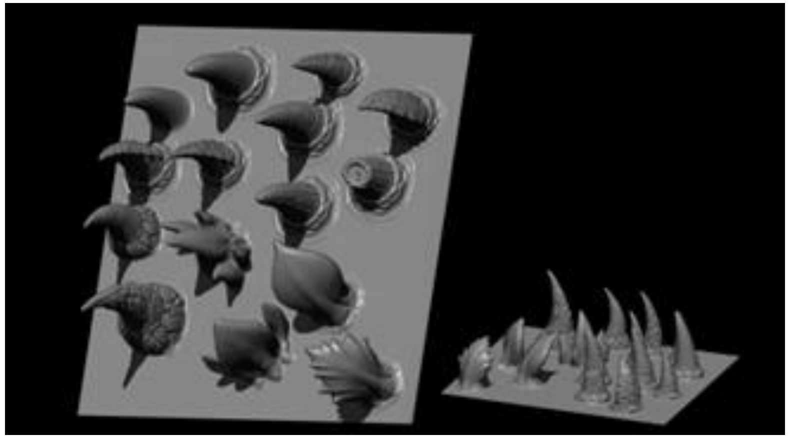
- Map over 2D Plane: $f \ : [0,1]^2 \rightarrow \mathbb{R}^3$
- Geometry Image



GenVDM, Yang et al., In Submission, 2025

Why do we care?

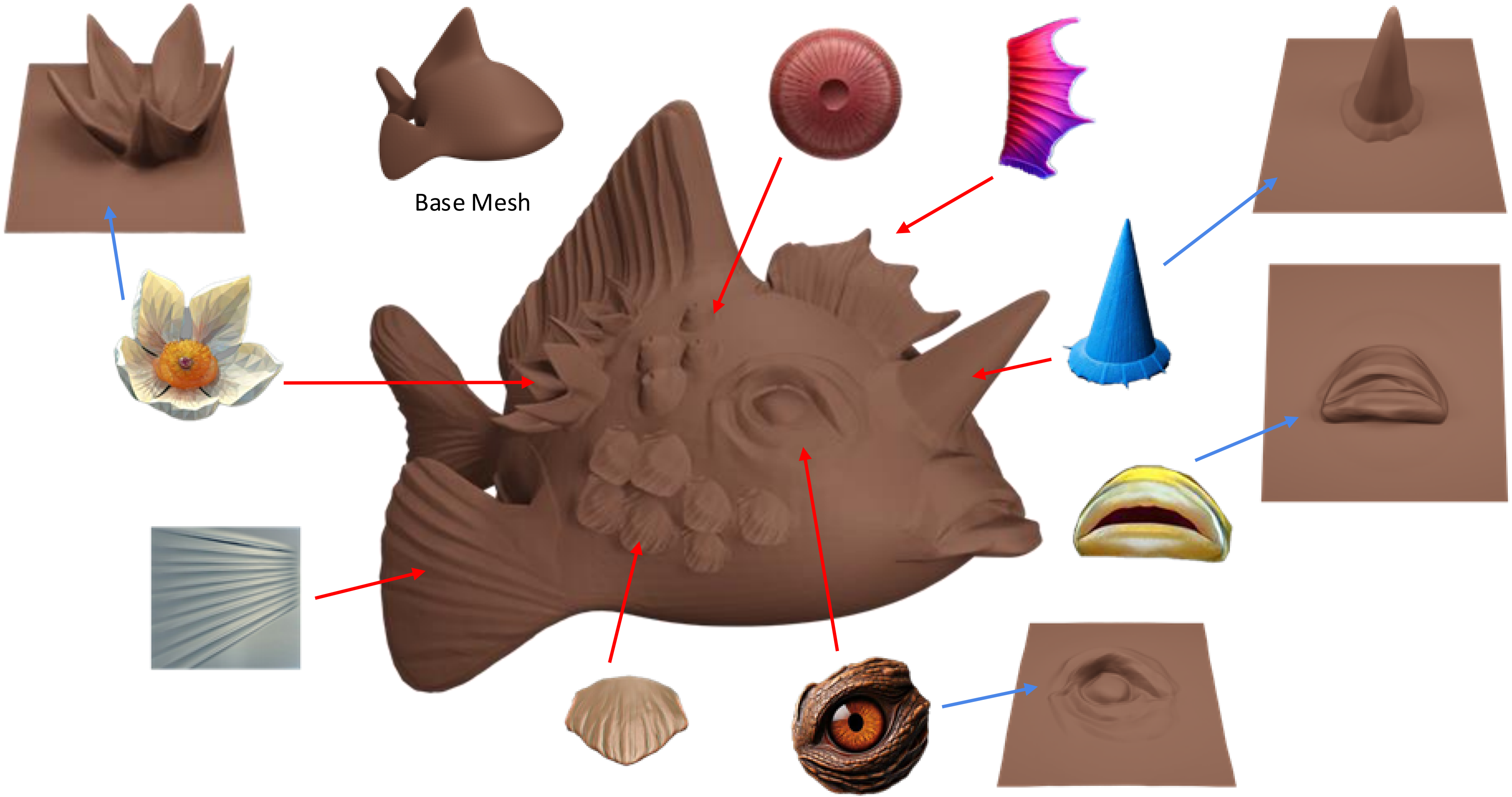
- Heavily used stock asset



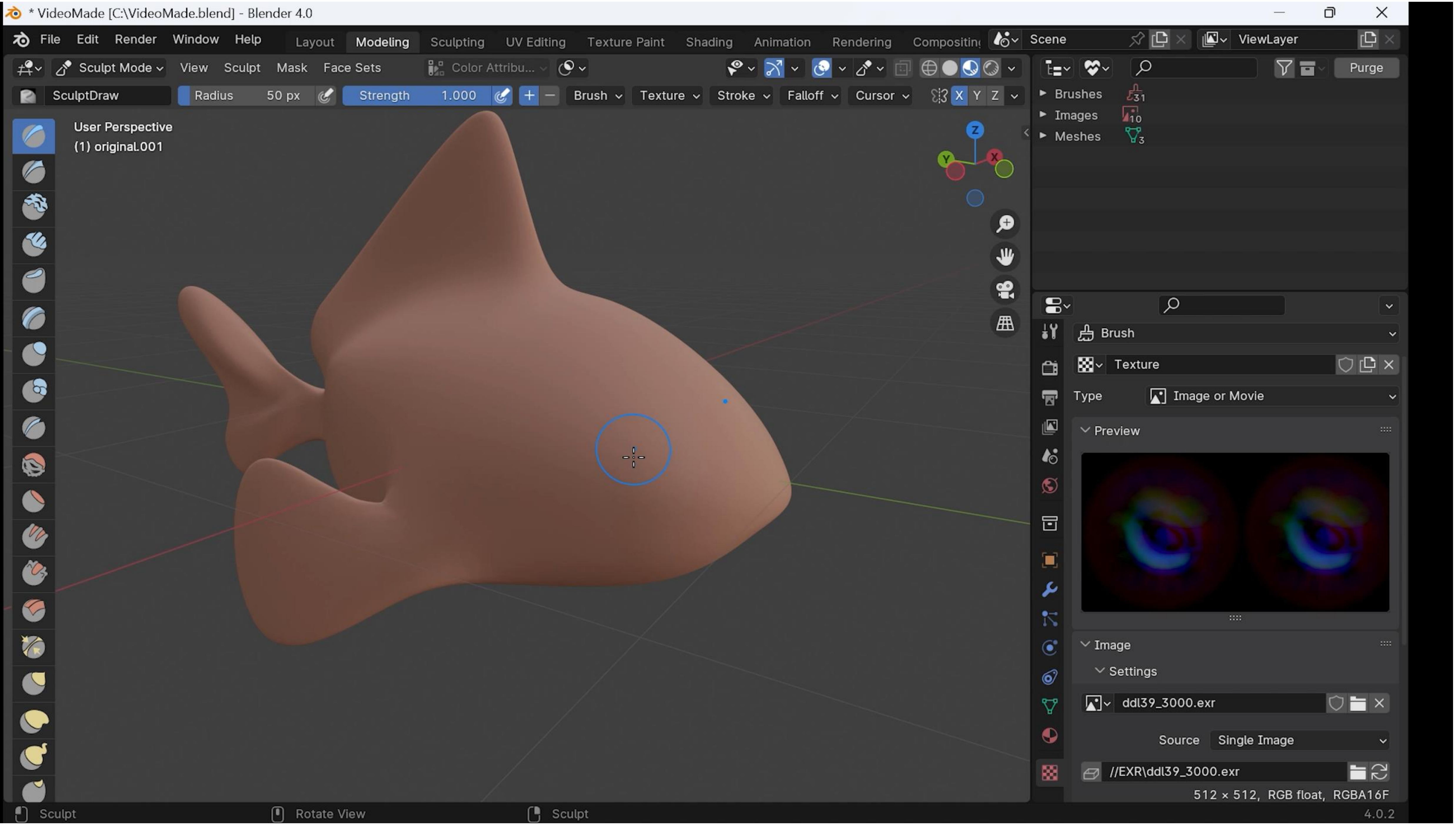
GenVDM, Yang et al., In Submission, 2025

Generative VDMs

- Create VDM from an image



GenVDM, Yang et al., In Submission, 2025



* VideoMade [C:\VideoMade.blend] - Blender 4.0

SculptDraw

Radius 50 px

Strength 1.000

User Perspective

(1) original.001

Brush

Texture

Type Image or Movie

ddl39_3000.exr

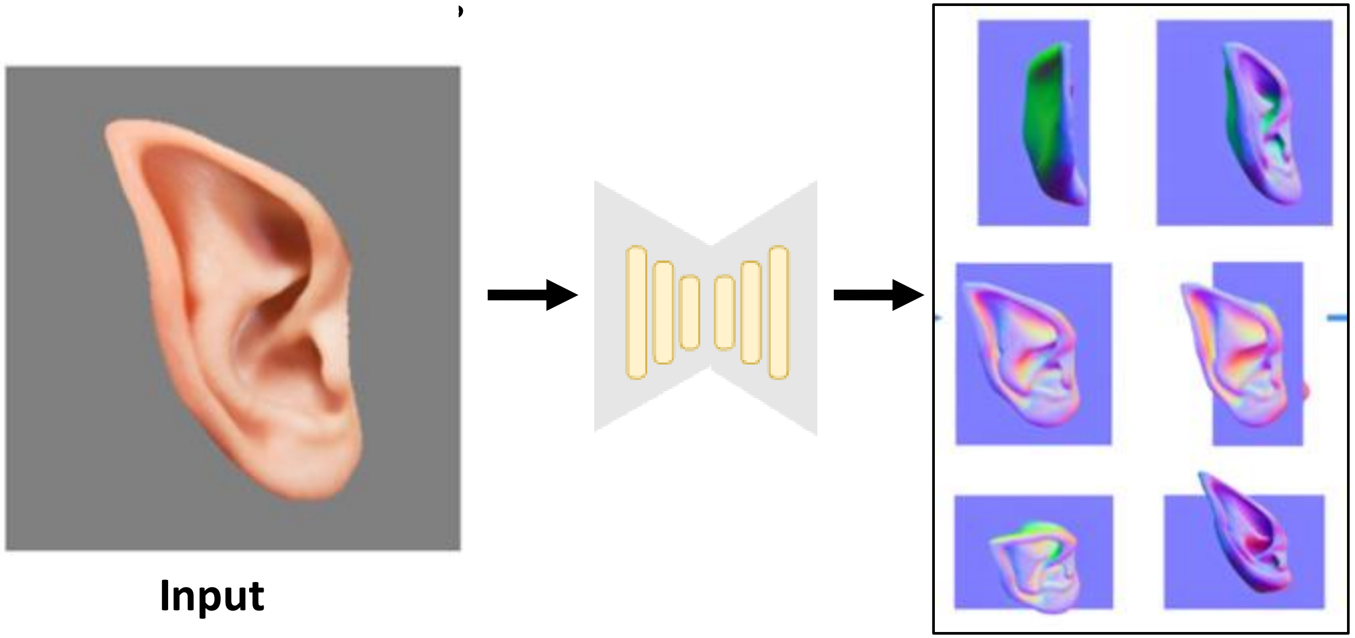
Source Single Image

//EXR\ddl39_3000.exr

512 × 512, RGB float, RGBA16F

Multi-view VDM Generation

- Start from an image



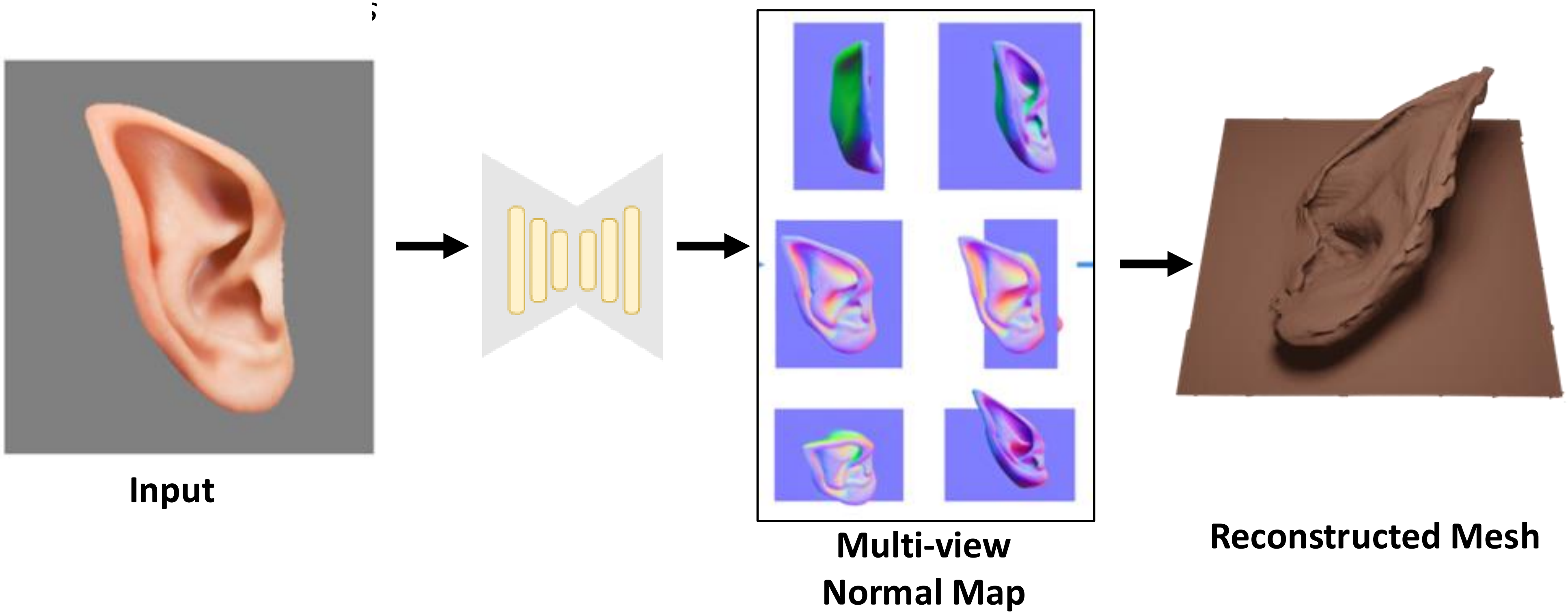
Input

Multi-view Normal Map

GenVDM, Yang et al., In Submission, 2025

Multi-view VDM Generation

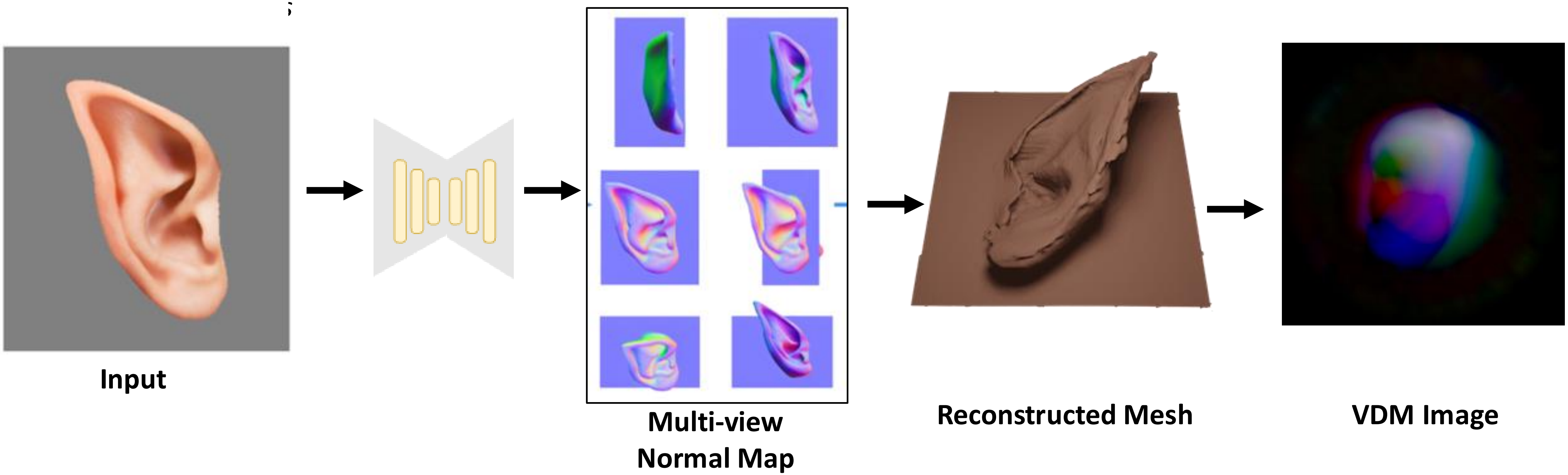
- Start from an image, reconstruct



GenVDM, Yang et al., In Submission, 2025

Multi-view VDM Generation

- Start from an image, reconstruct, optimize to get VDM



Input

Multi-view Normal Map

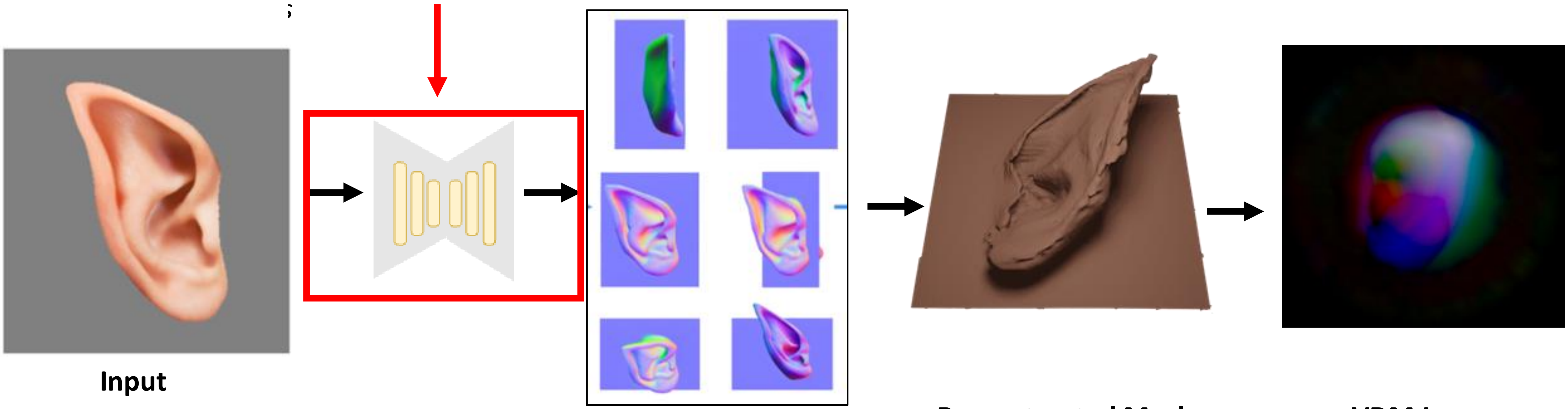
Reconstructed Mesh

VDM Image

GenVDM, Yang et al., In Submission, 2025

Multi-view VDM Generation

- Start from an image, reconstruct, optimize to get VDM



Multi-View Stable Diffusion

Input

Multi-view Normal Map

Reconstructed Mesh

VDM Image

GenVDM, Yang et al., In Submission, 2025

Training Data

- Can we take a model pre-trained on full objects?

Input

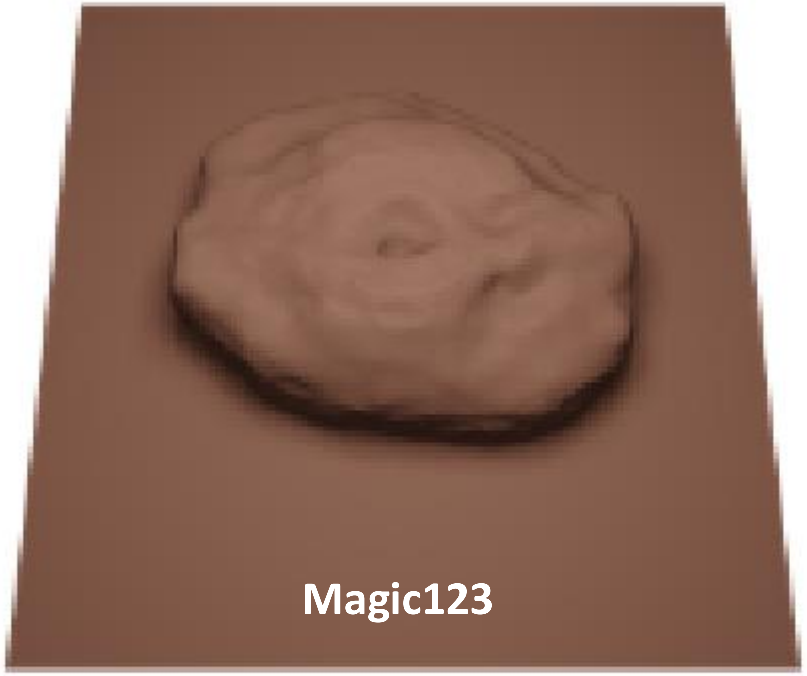
Adobe

GenVDM, Yang et al., In Submission, 2025

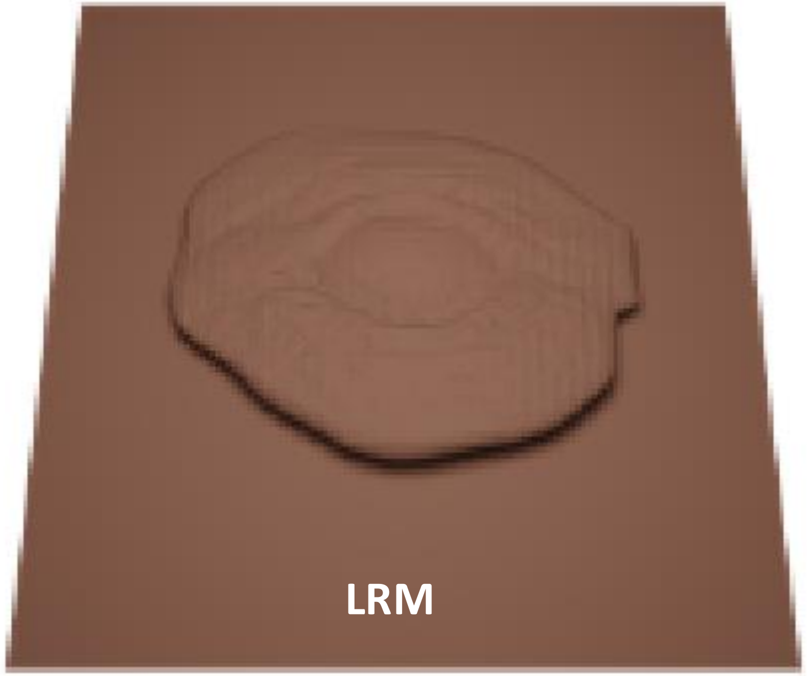
Training Data

- Can we take a model pre-trained on full objects? **NO!**

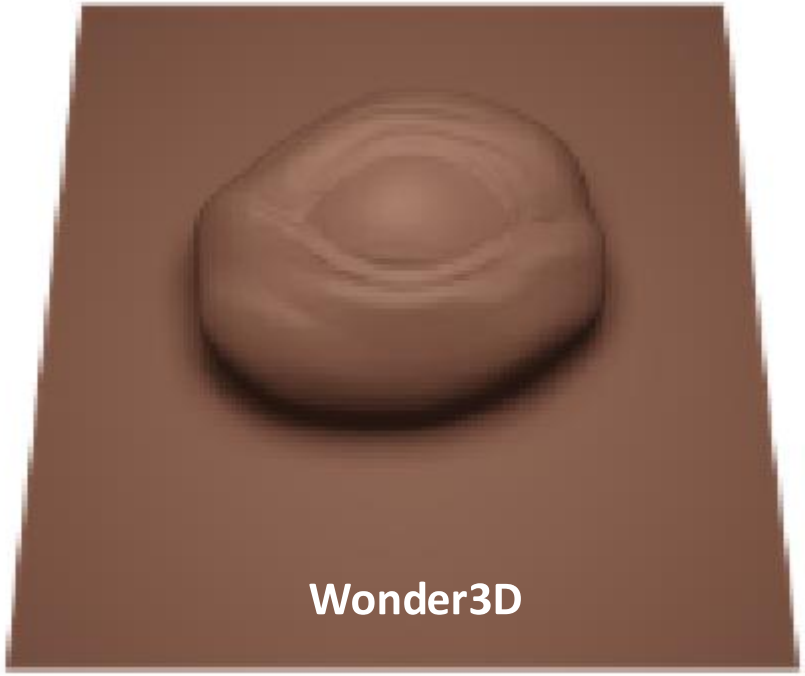
Input



Magic123



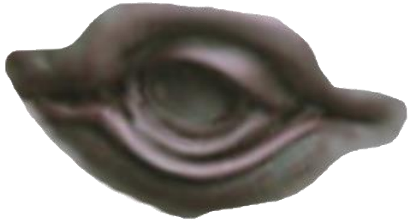
LRM



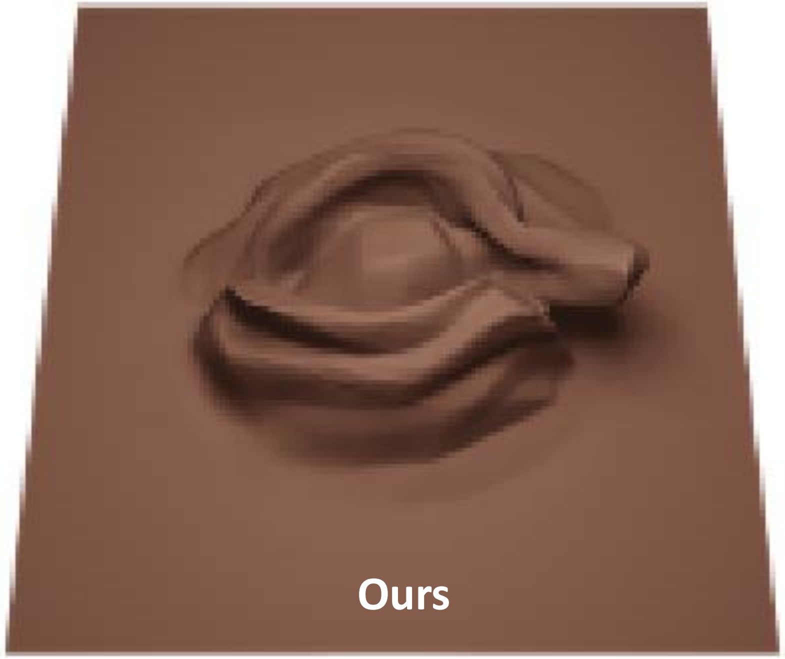
Wonder3D

Training Data

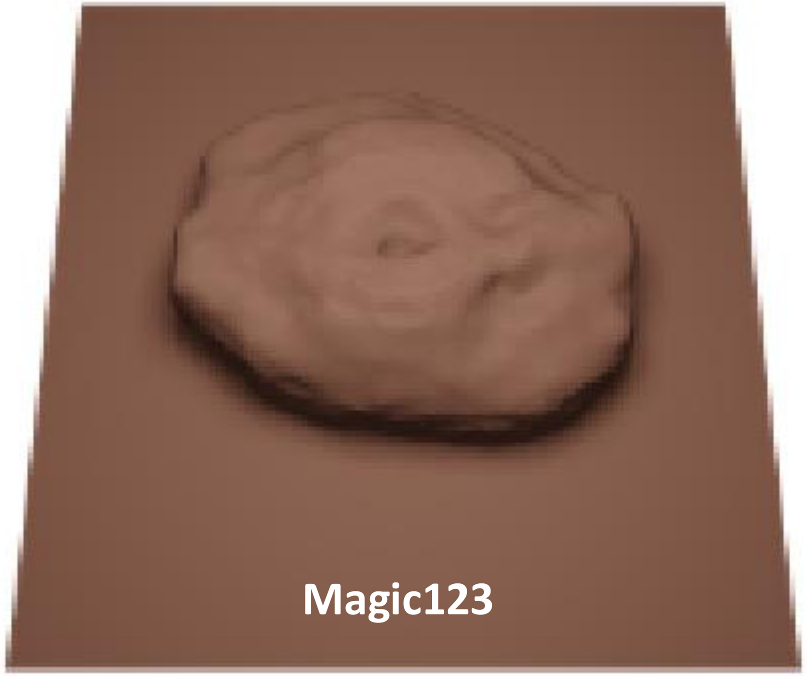
- Can we take a model pre-trained on full objects? **NO!**



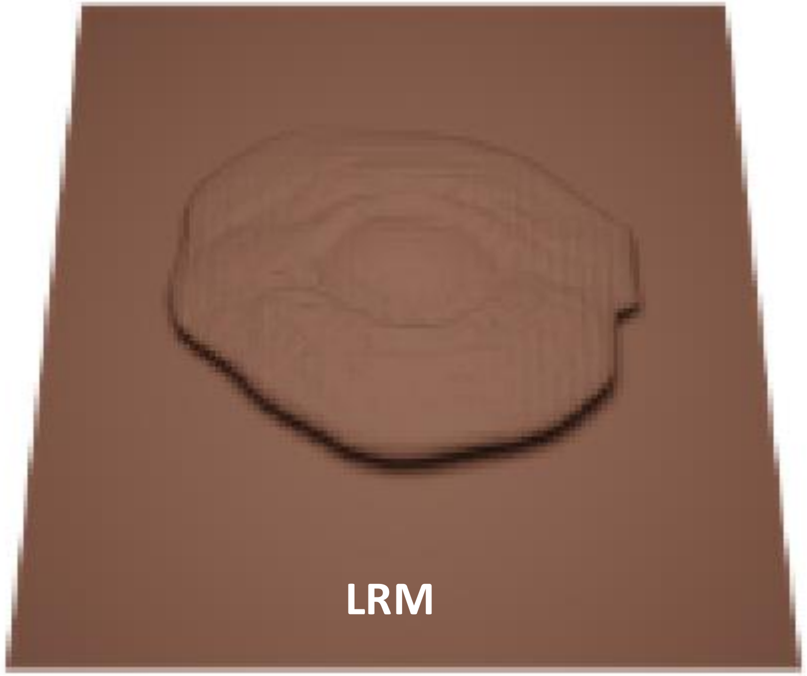
Input



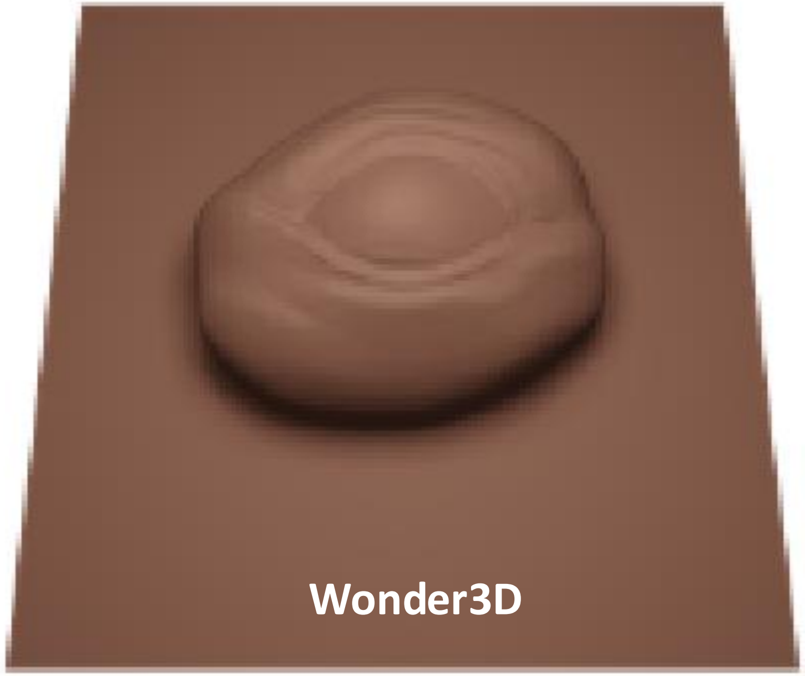
Ours



Magic123



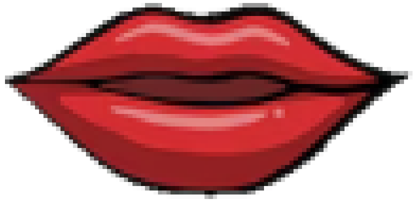
LRM



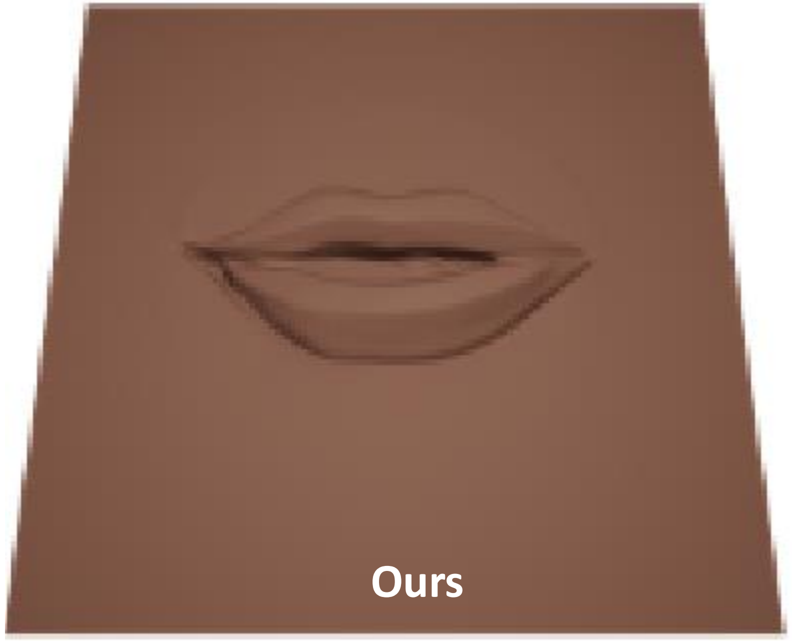
Wonder3D

Training Data

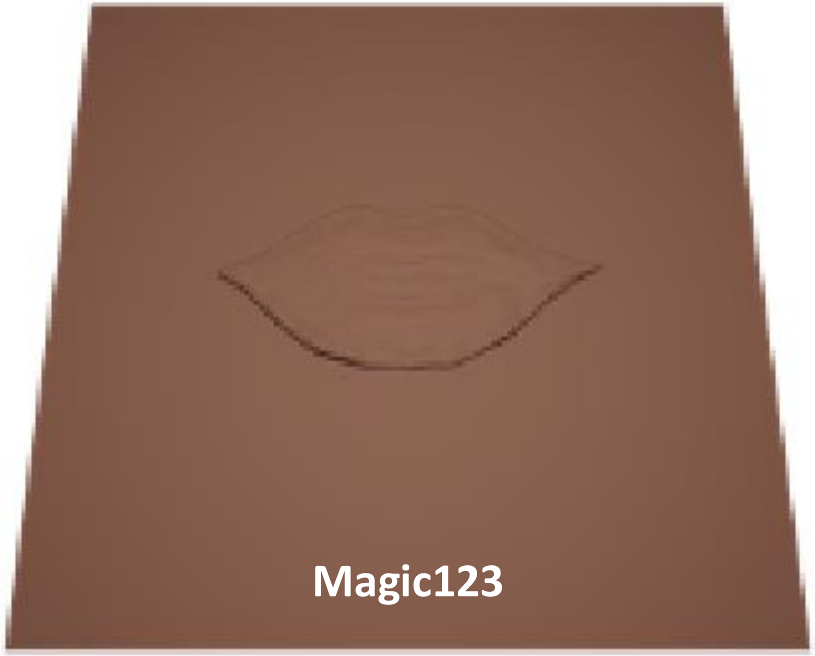
- Can we take a model pre-trained on full objects? **NO!**



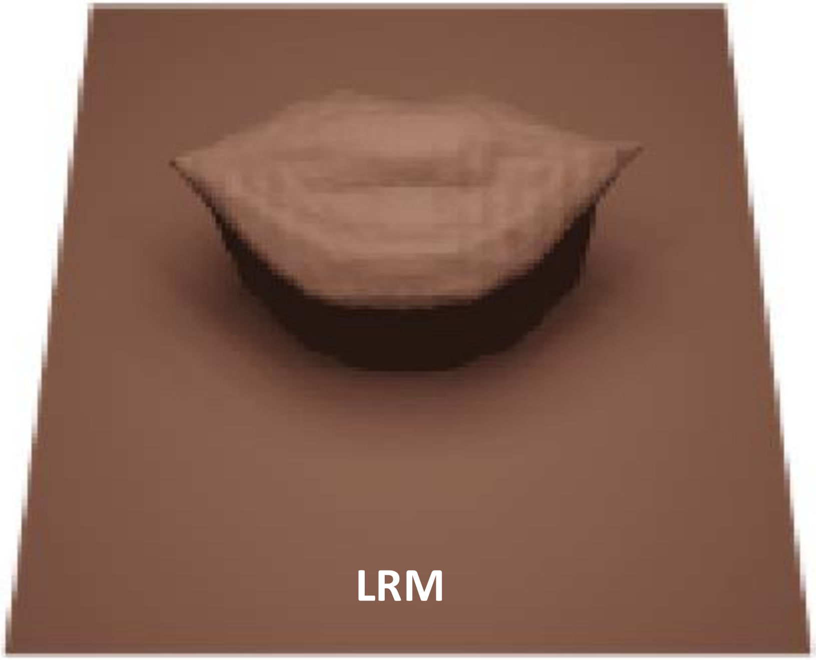
Input



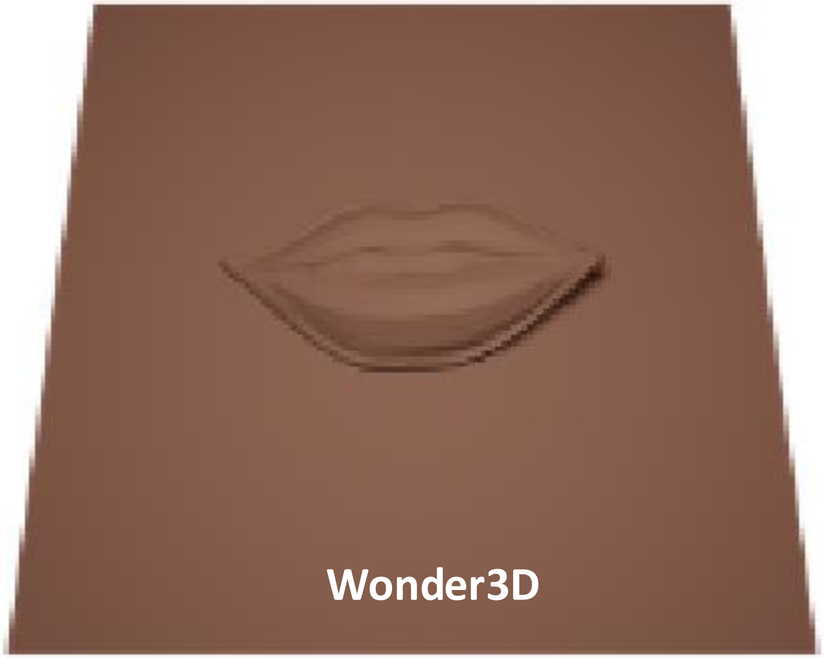
Ours



Magic123



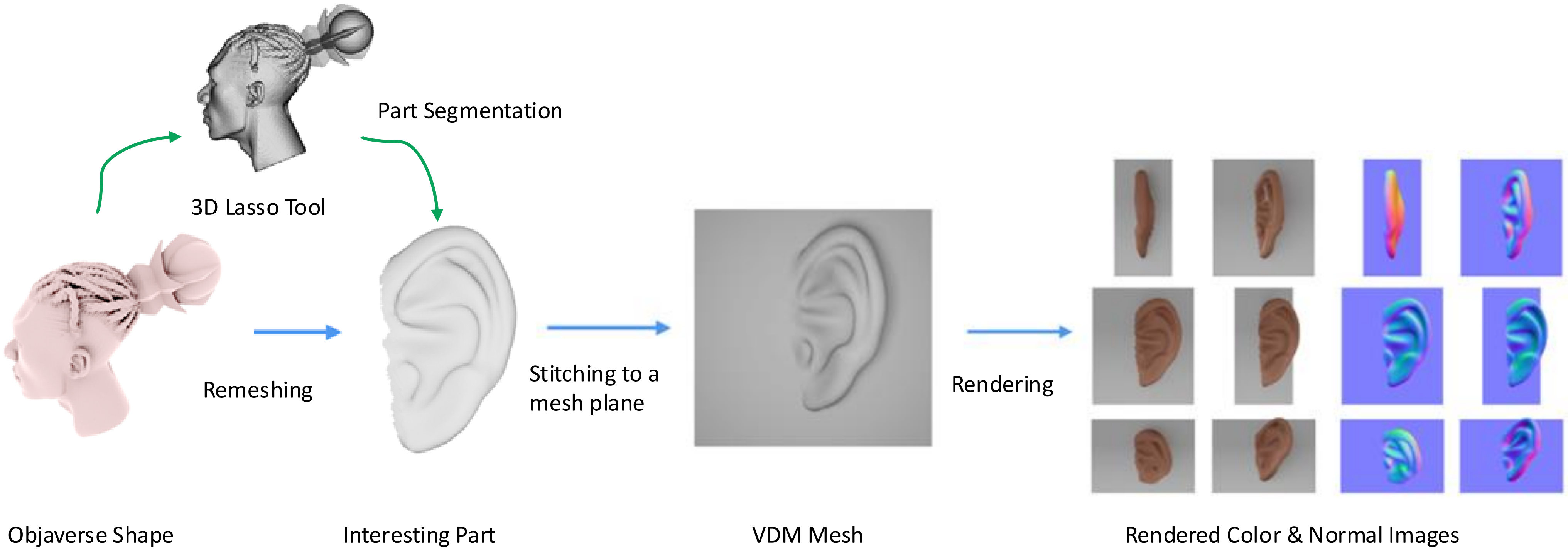
LRM



Wonder3D

Training Data

- Can we take a model pre-trained on full objects? **NO!**
- VDM data pipeline using Objaverse

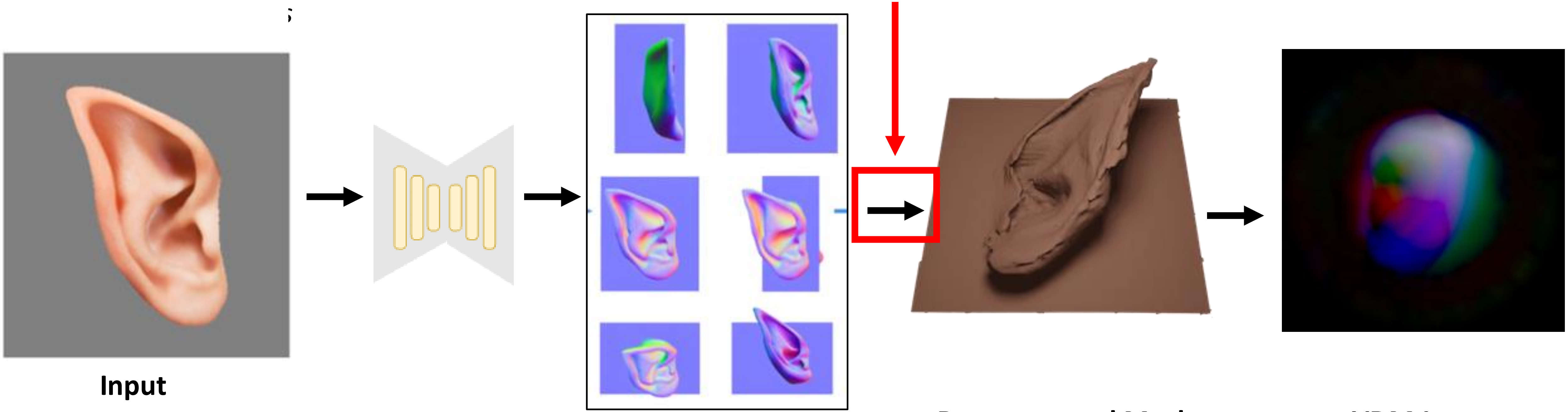


We create a dataset of 1200 examples!

GenVDM, Yang et al., In Submission, 2025

Multi-view VDM Generation

- Start from an image, reconstruct, optimize to get VDM



NeuS-based Reconstruction

Input

Multi-view Normal Map

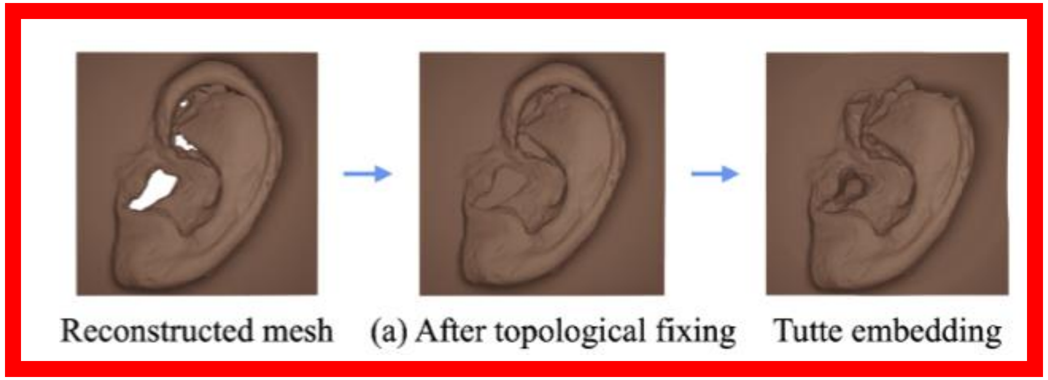
Reconstructed Mesh

VDM Image

GenVDM, Yang et al., In Submission, 2025

Multi-view VDM Generation

- Start from an image, reconstruct, optimize to get VDM

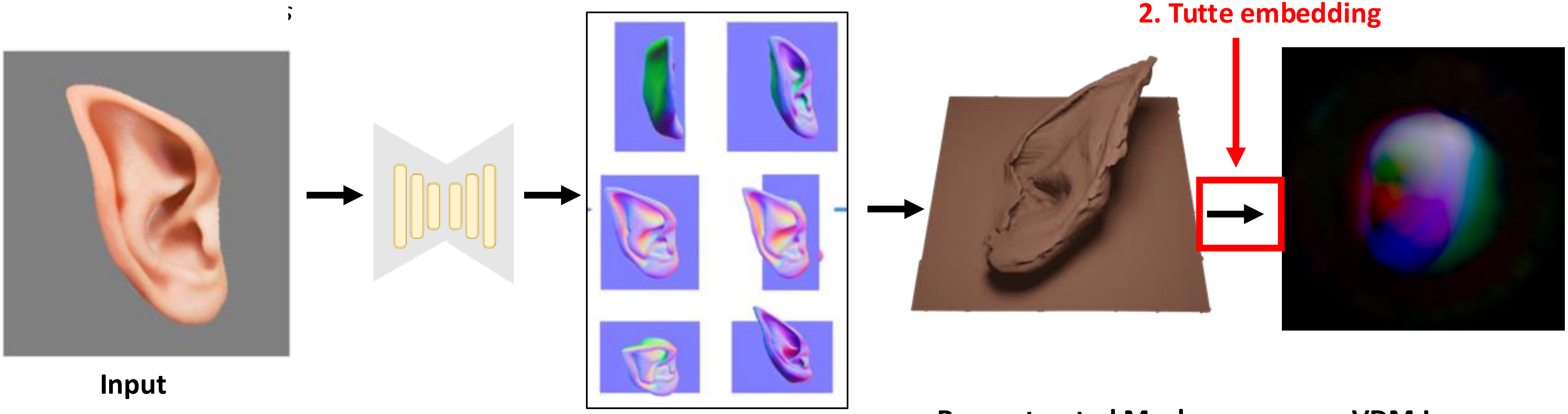


Reconstructed mesh → (a) After topological fixing → Tutte embedding

Naïve Baseline:

1. fill holes

2. Tutte embedding



Input

Multi-view Normal Map

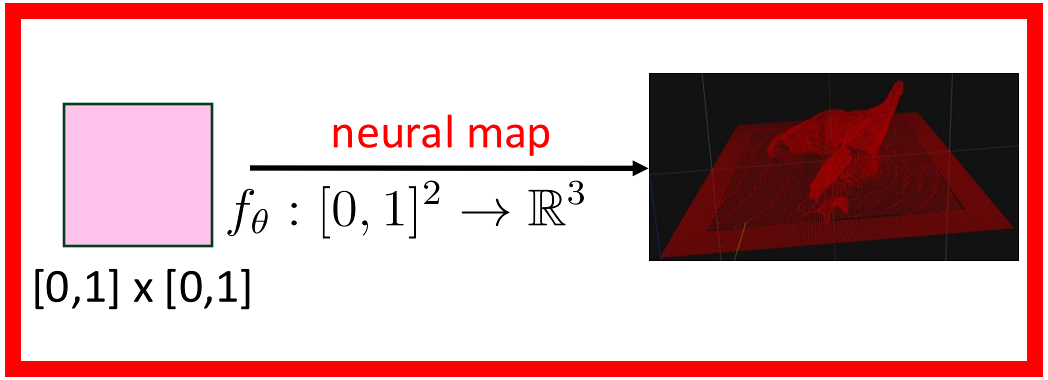
Reconstructed Mesh

VDM Image

GenVDM, Yang et al., In Submission, 2025

Multi-view VDM Generation

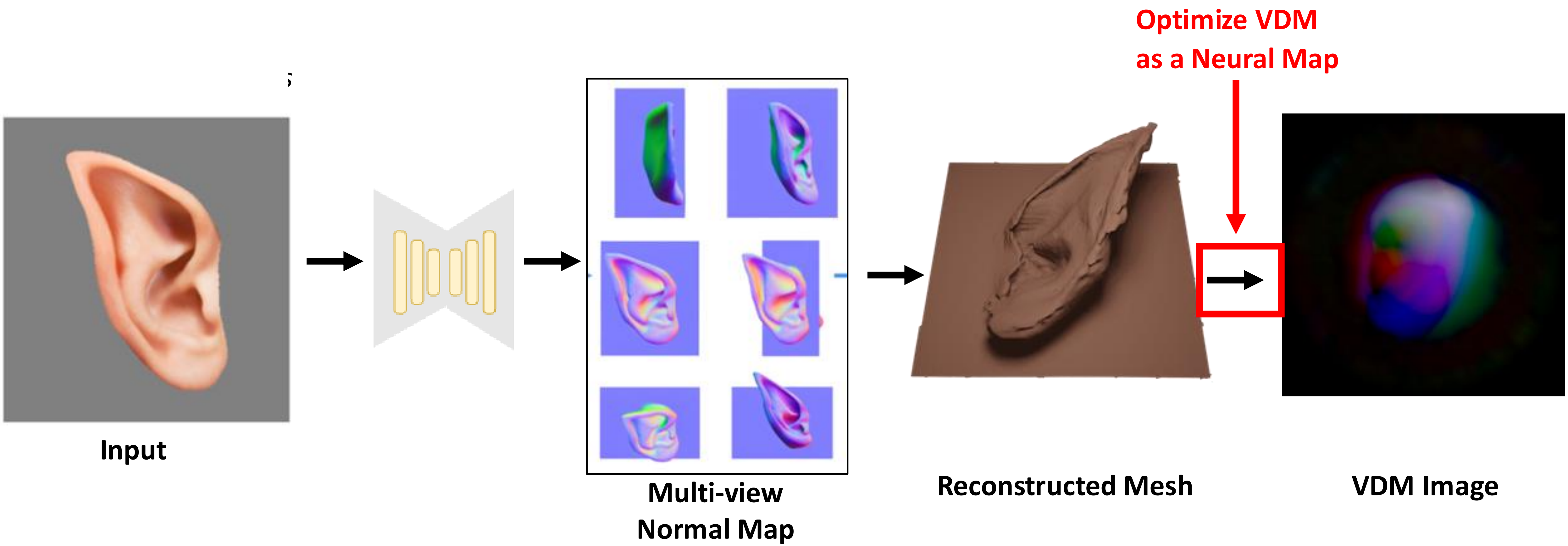
- Start from an image, reconstruct, optimize to get VDM



neural map

$f_\theta : [0,1]^2 \to \mathbb{R}^3$

[0,1] x [0,1]



Optimize VDM as a Neural Map

Input

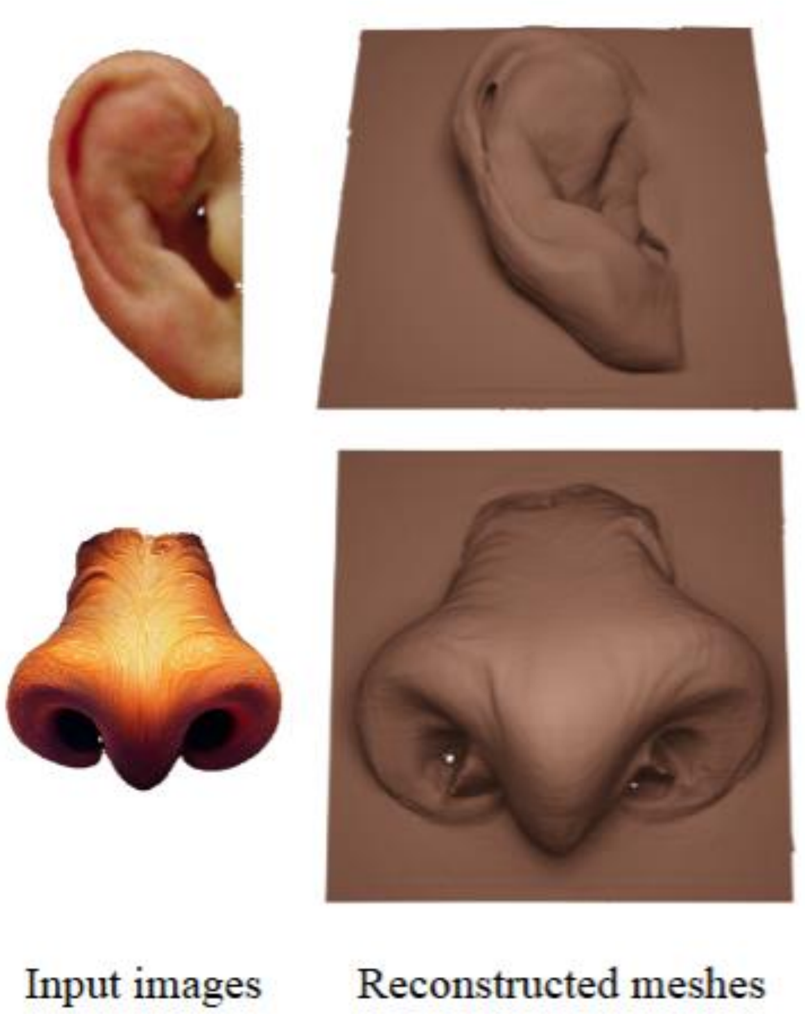
Multi-view Normal Map

Reconstructed Mesh

VDM Image

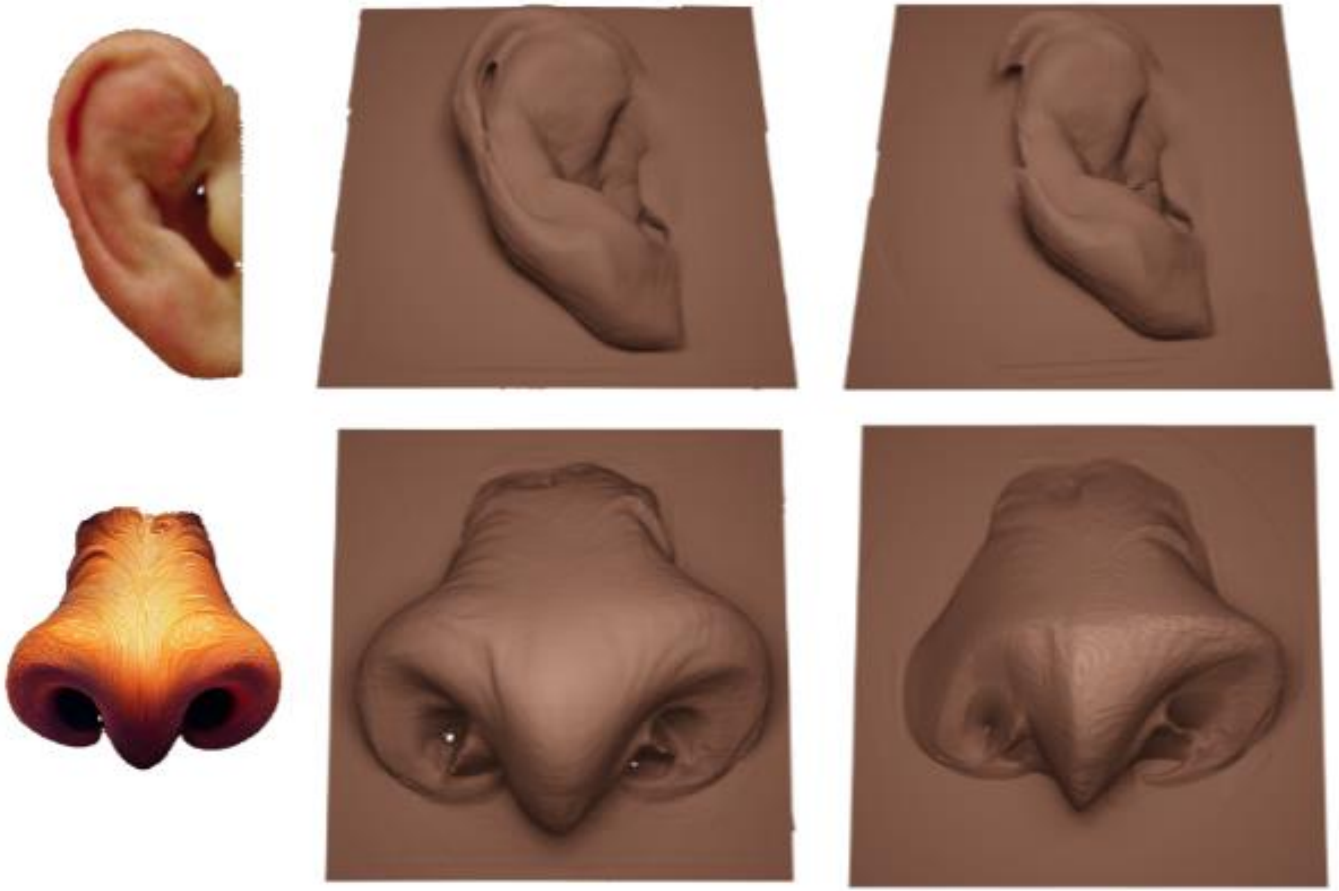
GenVDM, Yang et al., In Submission, 2025

Ablation: Reconstruction Alternatives



GenVDM, Yang et al., In Submission, 2025

Ablation: Reconstruction Alternatives



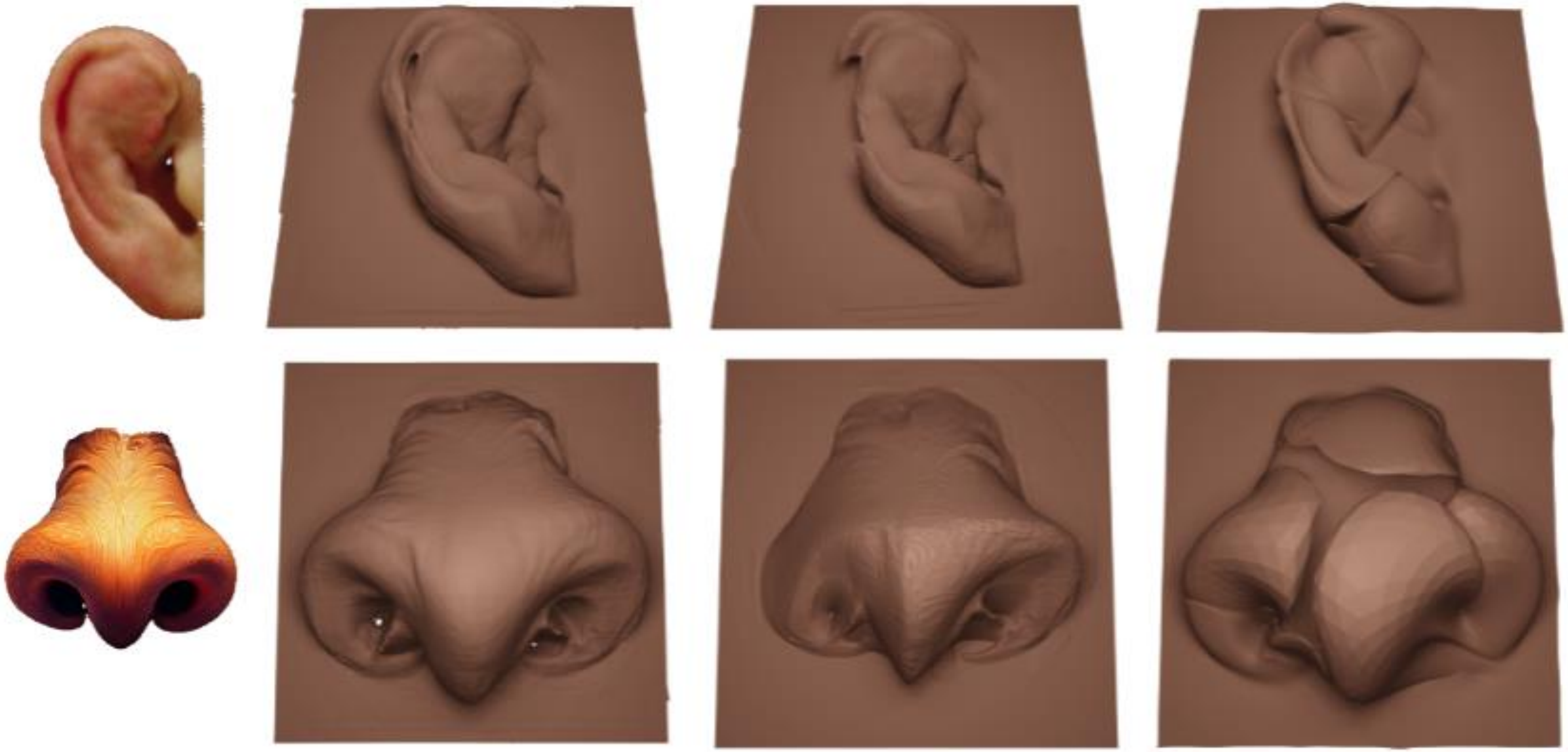
Input images

Reconstructed meshes

(a) Topological fixing & Tutte embedding

GenVDM, Yang et al., In Submission, 2025

Ablation: Reconstruction Alternatives



Input images

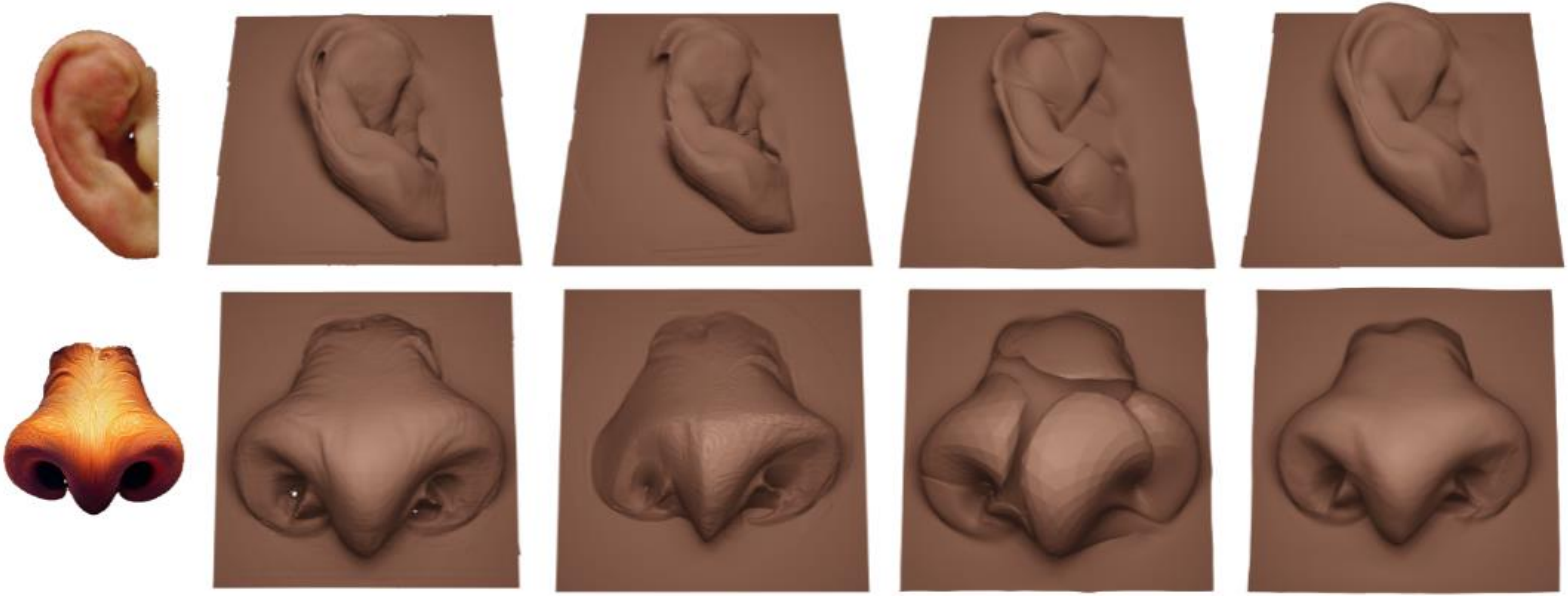
Reconstructed meshes

(a) Topological fixing & Tutte embedding

(b) Mesh optimization

GenVDM, Yang et al., In Submission, 2025

Ablation: Reconstruction Alternatives



Input images

Reconstructed meshes

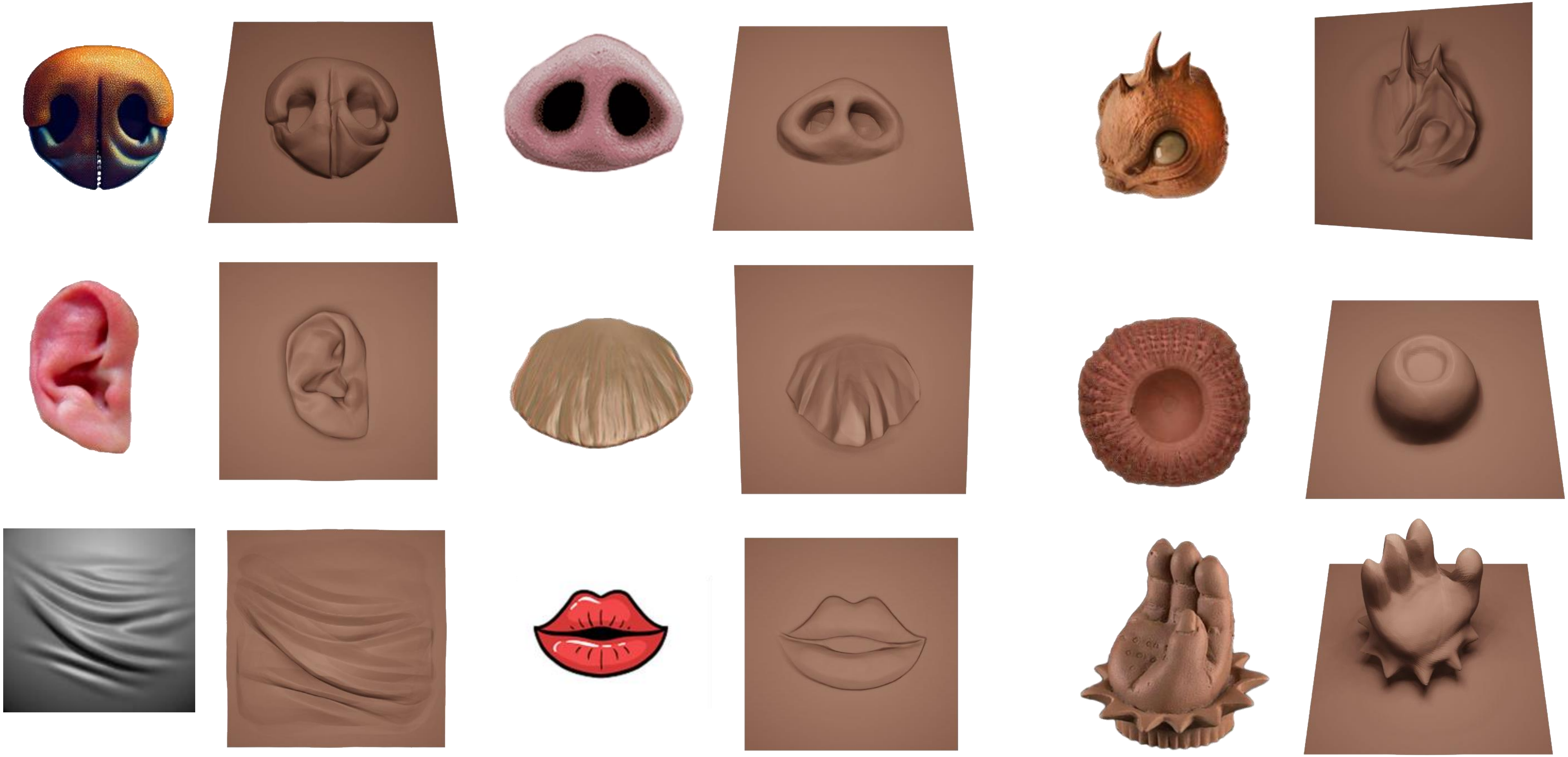
(a) Topological fixing & Tutte embedding

(b) Mesh optimization

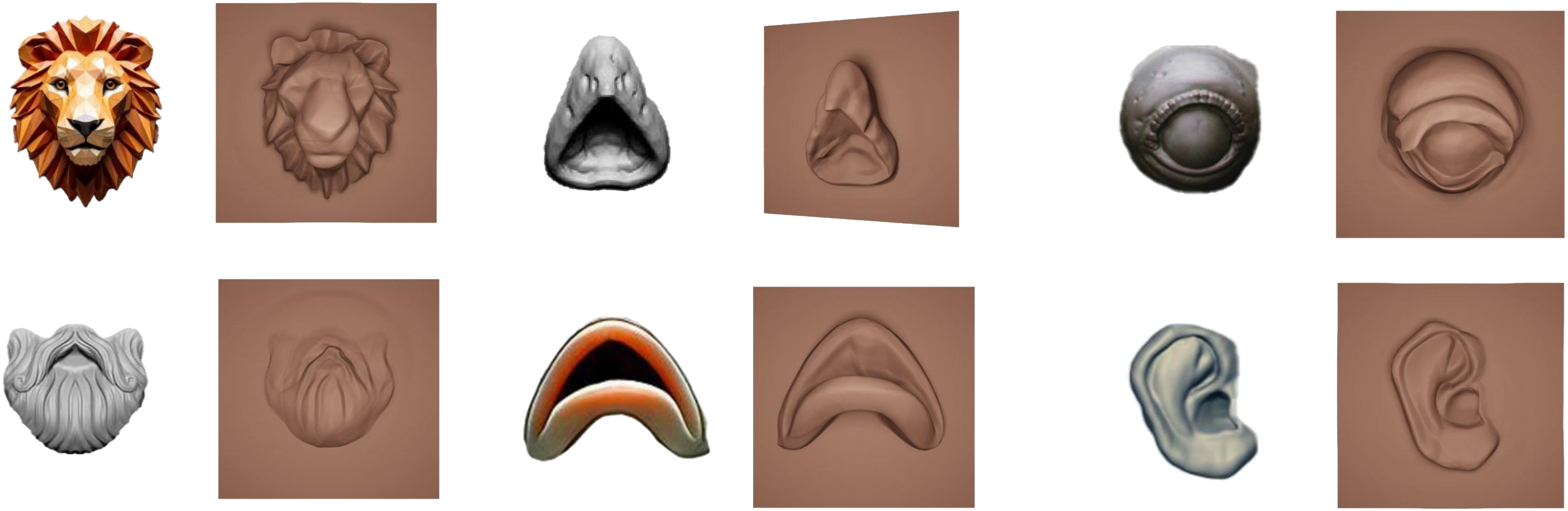
(c) Our approach (Neural Deformation Field)

GenVDM, Yang et al., In Submission, 2025

Generated VDMs

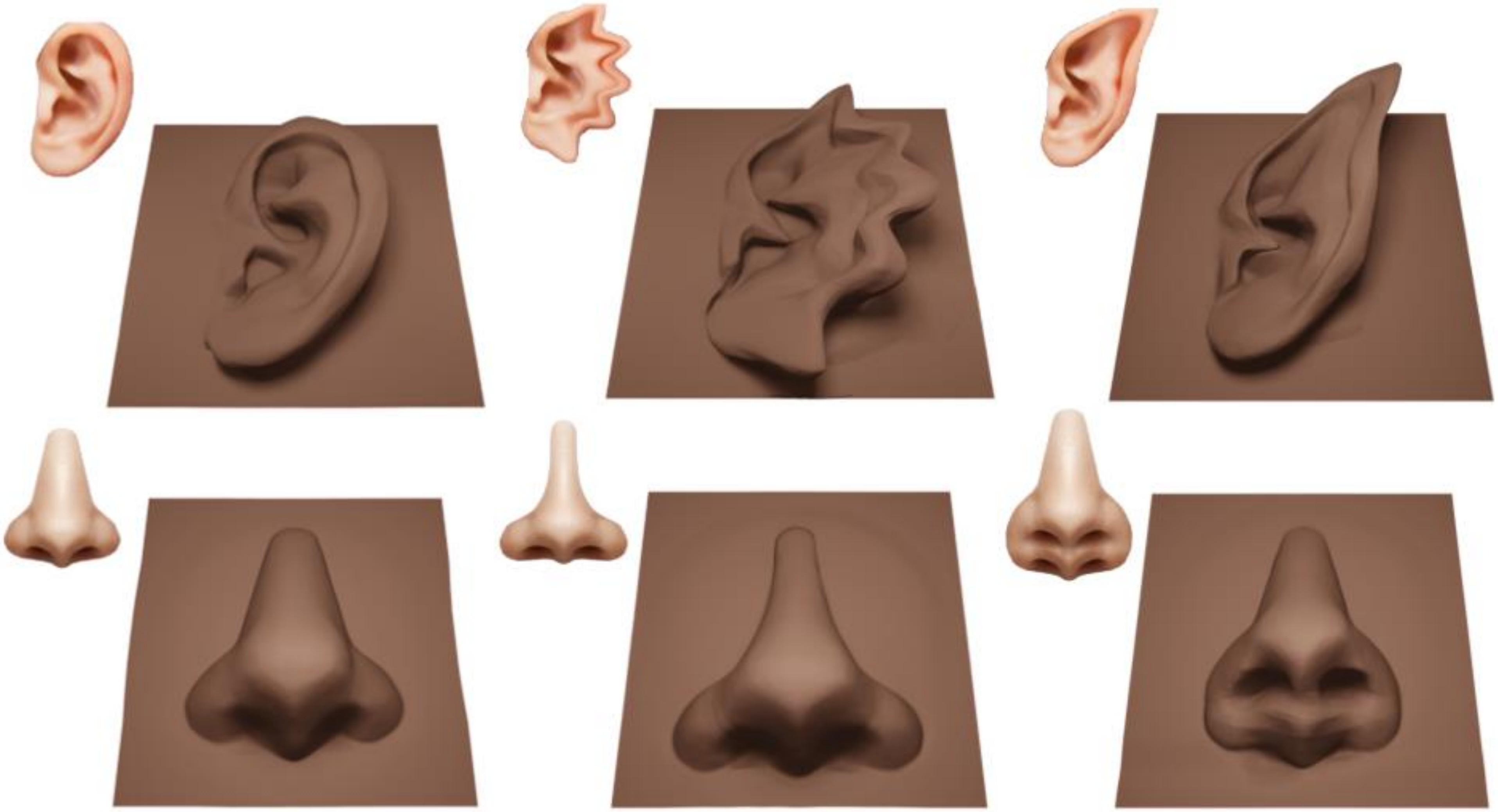


Generated VDMs



GenVDM, Yang et al., In Submission, 2025

Image Editing for VDM



Original

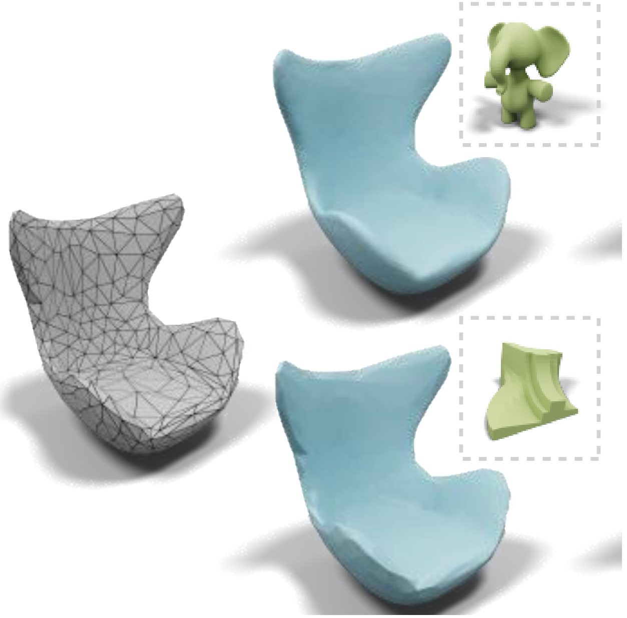
Edition 1

Edition 2

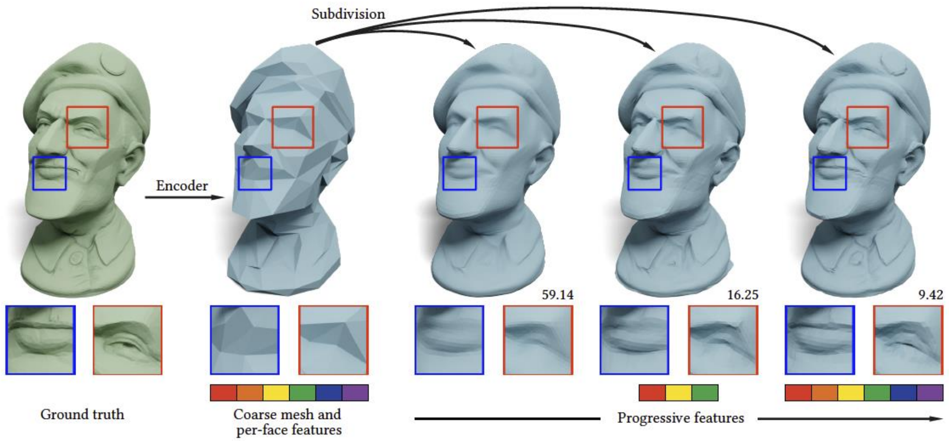
Other Work on Detailization

"a classic furniture piece made of polished wood with subtle details"

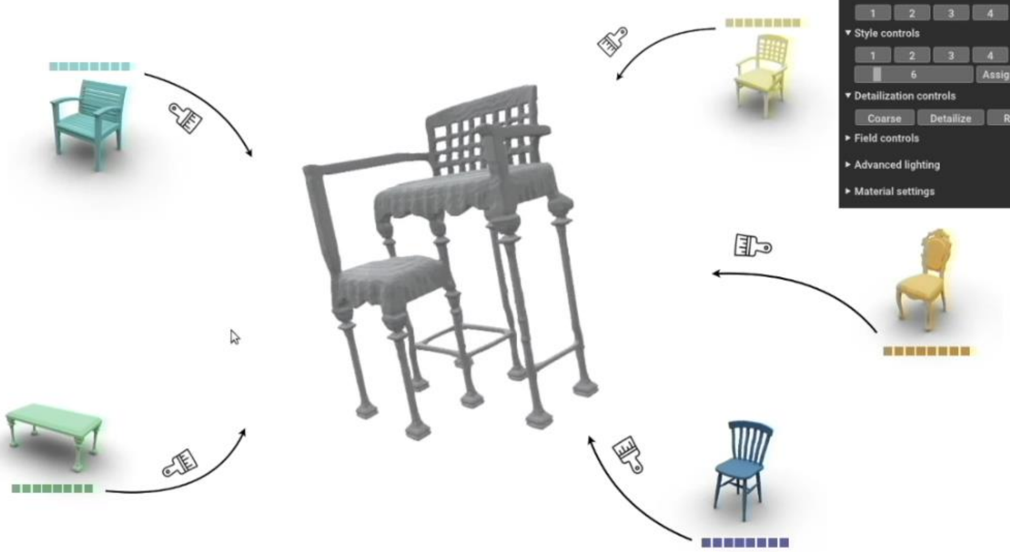
ART-DECO (Qimin Chen et al.), under review, 2025



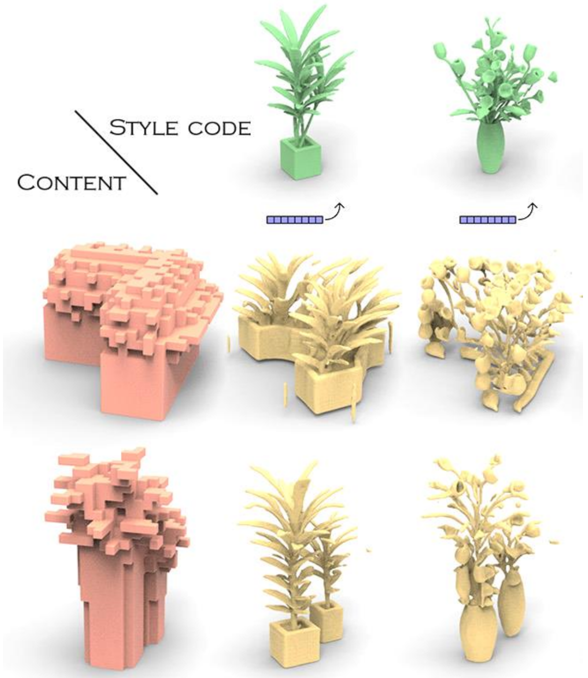
Neural Progressive Meshes (Hsueh-Ti Liu et al.), SIGGRAPH 2020



Neural Progressive Meshes (Yun-Chun Chen et al.), SIGGRAPH 2023



DECOLLAGE (Qimin Chen et al.), ECCV 2024



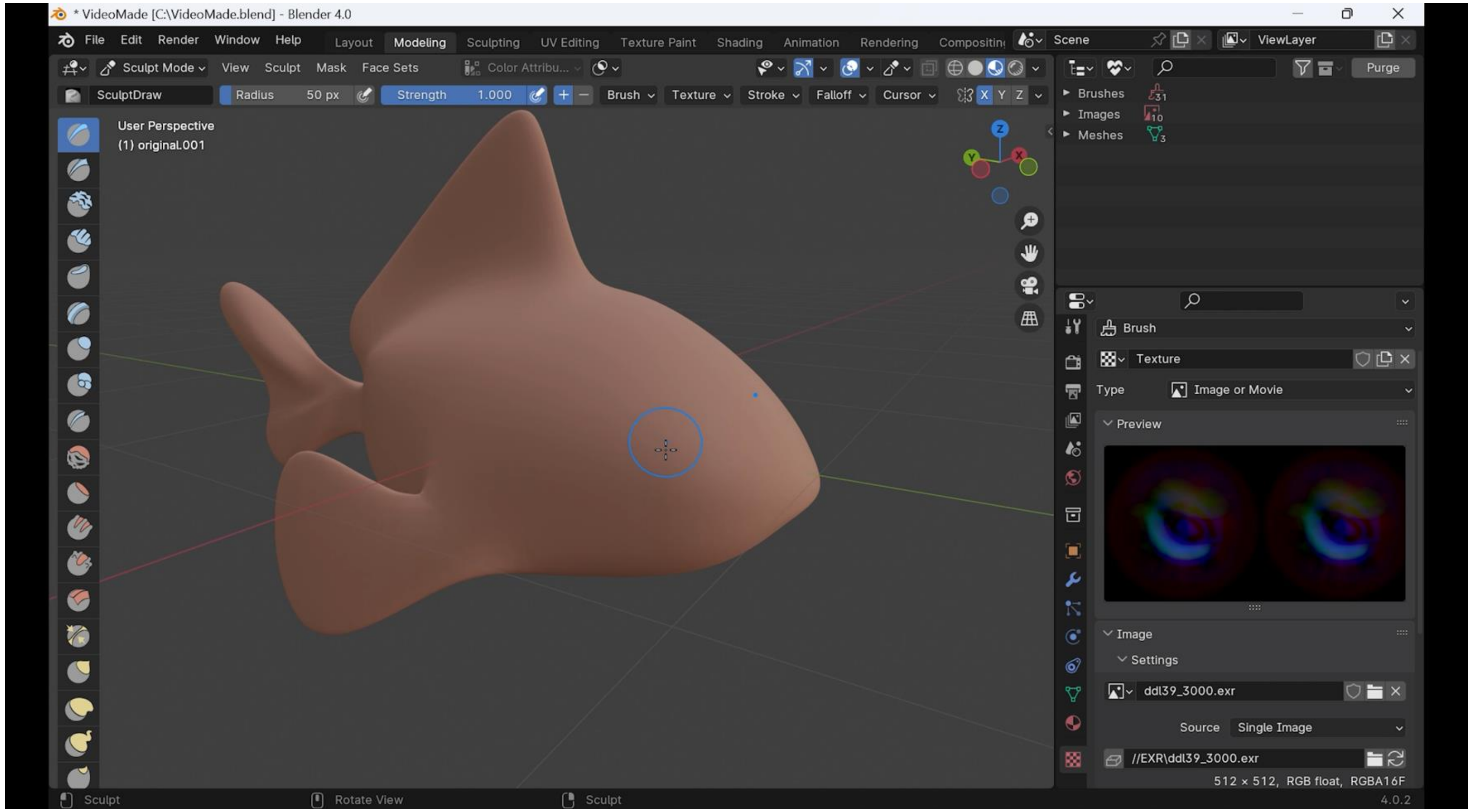
DecorGAN (Zhiqin Chen et al.), CVPR 2021

Summary

- Workflows
- Representations
- Multi-view predictions

Summary

- **Workflows**
  - Coarse deformation
  - Coarse sculpting
  - Detailize via normal maps
  - Detailize via VDMs
  - …
- Representations
- Multi-view predictions



* VideoMade [C:\VideoMade.blend] - Blender 4.0

Sculpt Mode

SculptDraw

Radius 50 px

Strength 1.000

User Perspective
(1) original.001

Texture

Type Image or Movie

ddl39_3000.exr

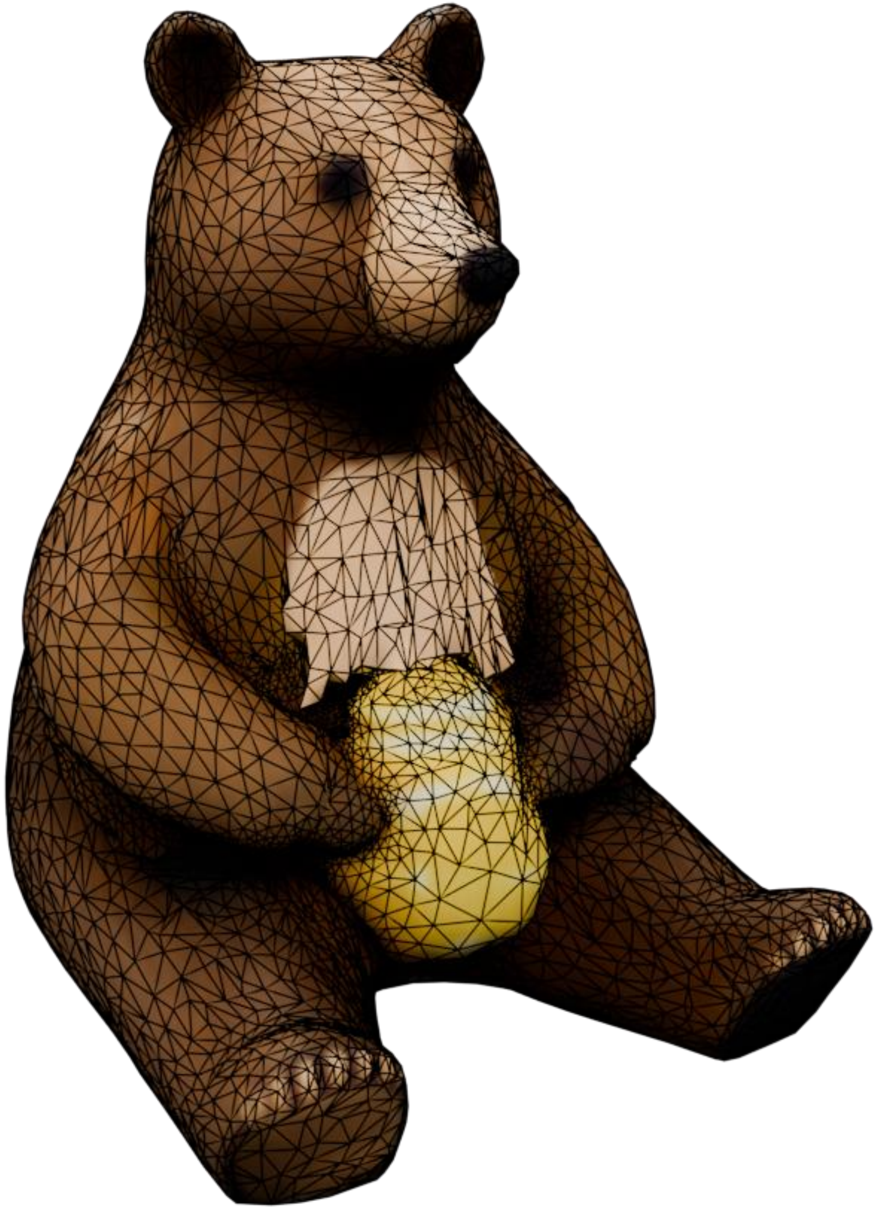
Source Single Image

//EXR\ddl39_3000.exr

512 x 512, RGB float, RGBA16F

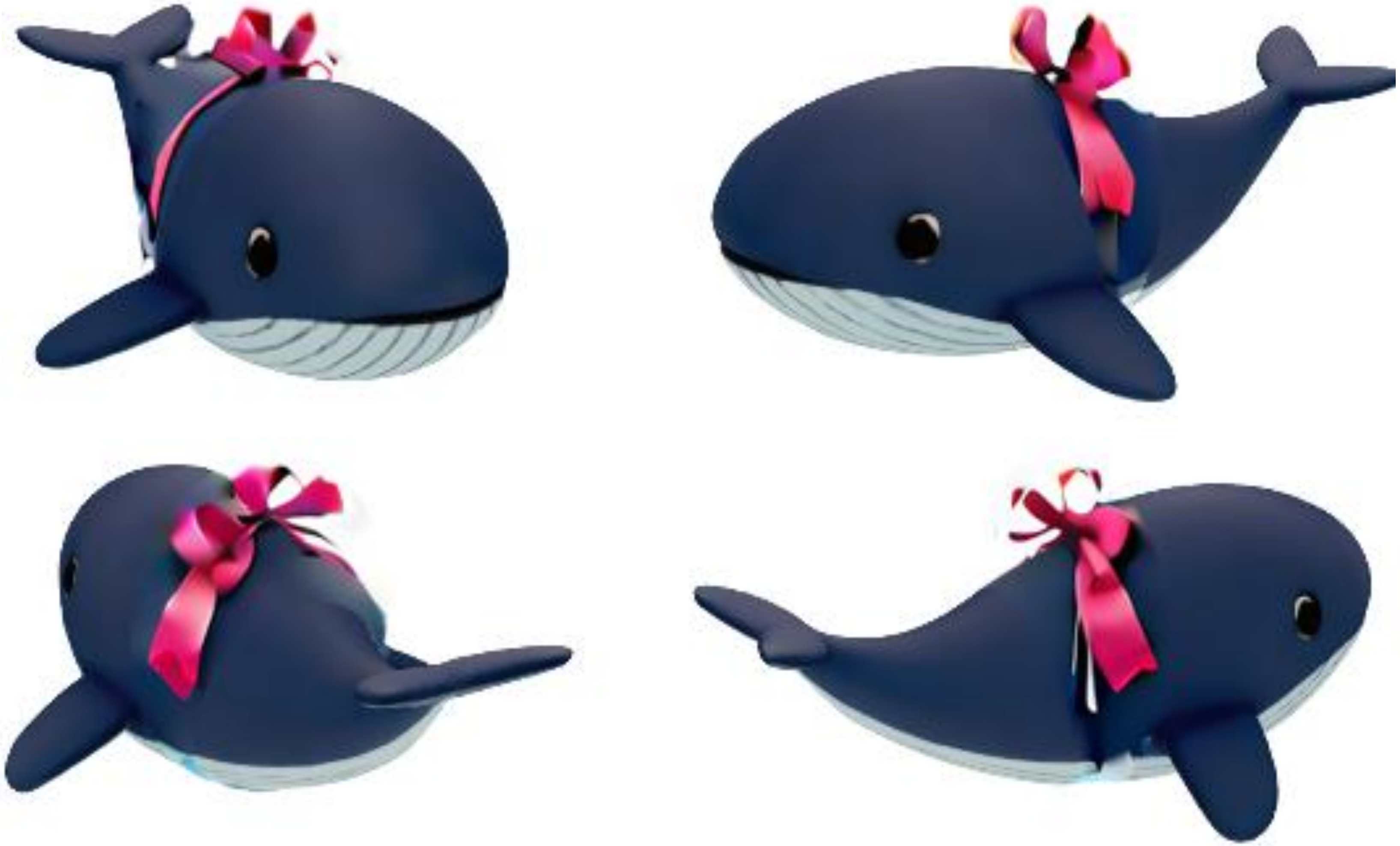
Summary

- Workflows
- **Representations**
  - Meshes are essential for many existing workflows and pipelines
  - Hybrid representations allows to get the best of both worlds
  - Representation-agnostic methods via multi-view + LRM / Optimization
- Multi-view predictions



Summary

- Workflows
- Representations
- Multi-view predictions
  - Leverage pre-trained 2D priors
  - Work with different LRMs
  - Often need to be fine-tuned



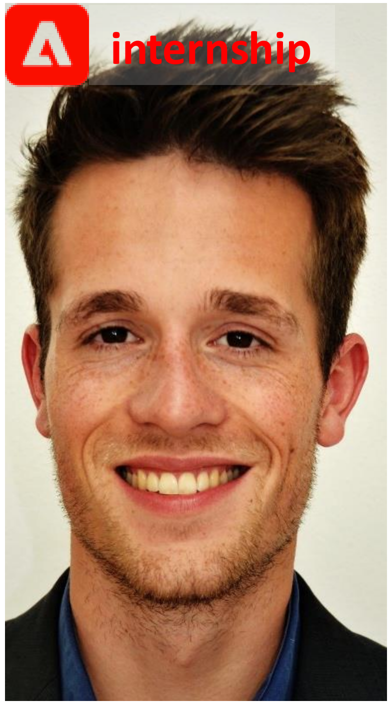
Collaborators

APAP
Seungwoo Yoo

MeshUp
Brian Kim

1. MagicClay
2. Instant3dit
Amir Barda

Text-guided refinement
Yun-Chun Chen



SAMa
Michael Fischer

VDM
Yuezhi Yang

Adobe: Matheus Gadelha, Thibault Groueix, Zhiqin Chen, Siddhartha Chaudhuri, Iliyan Giorgiev, Valentin Deschaintre, Alec Jacobson

Academia: Noam Aigerman, Amir Barda, Rana Hanocka, Qixing Huang, Kunho Kim, Itai Lang, Minhyuk Sung, Hao Zhang

Adobe