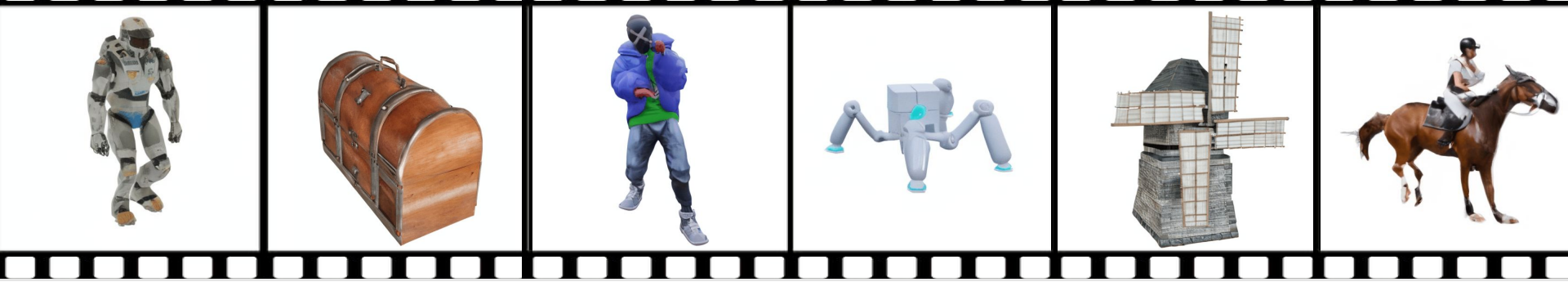


SV4D 2.0:
Enhancing Spatio-Temporal Consistency in Multi-View Video Diffusion for High-Quality 4D Generation

Chun-Han Yao*, Yiming Xie*, Vikram Voleti, Huaizu Jiang^, Varun Jampani^

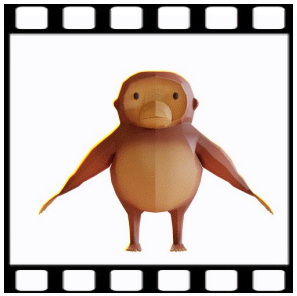
* Equal Contribution ^ Equal Advising

stability.ai

Background

Background

Problem Setting



4D Generation

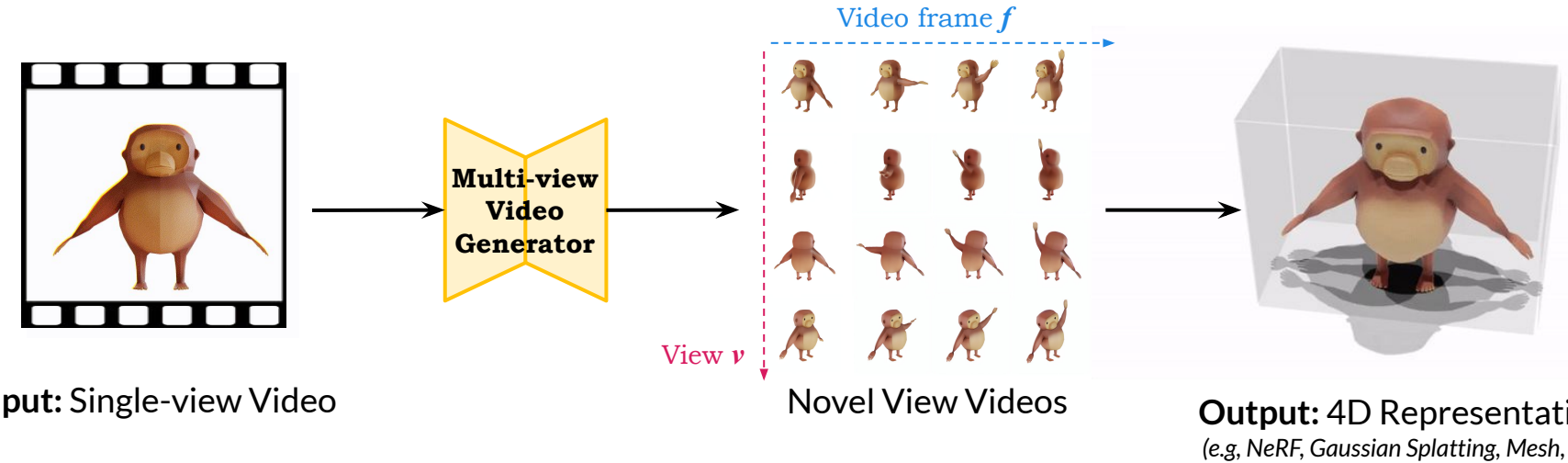
Input: Single-view Video

Output: 4D Representation
(e.g, NeRF, Gaussian Splatting, Mesh, etc)

Figure Credit: SV4D [Xie & Yao, et al.]

Background

Common Pipeline



Input: Single-view Video

Novel View Videos

Output: 4D Representation
(e.g, NeRF, Gaussian Splatting, Mesh, etc)

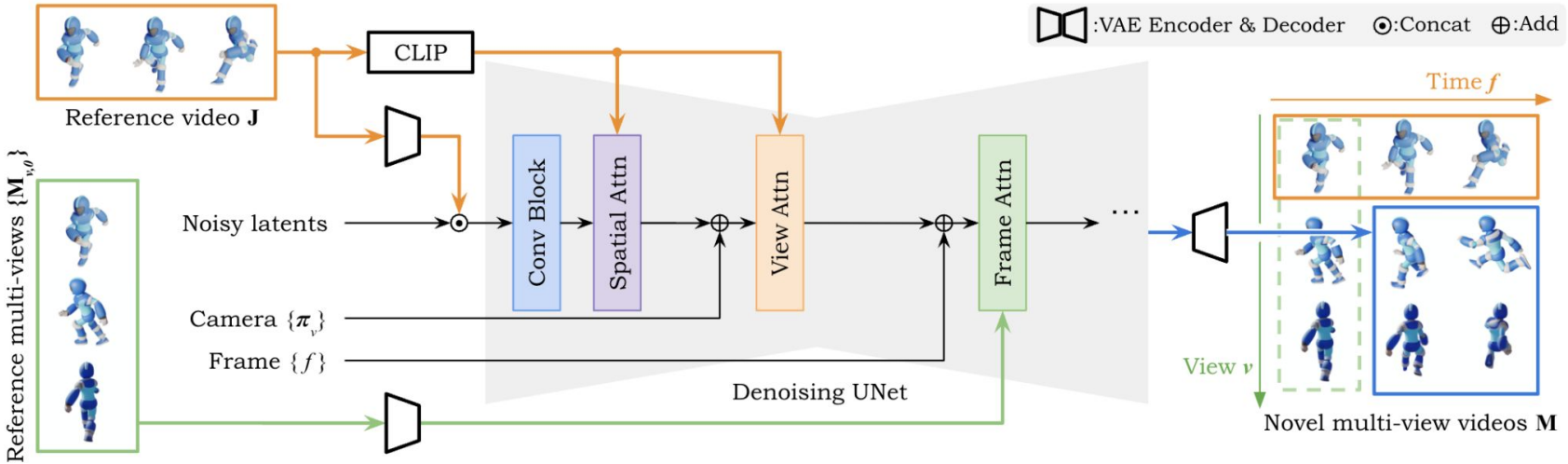
Novel View Video Synthesis

4D Optimization

Figure Credit: SV4D [Xie & Yao, et al.]

Background

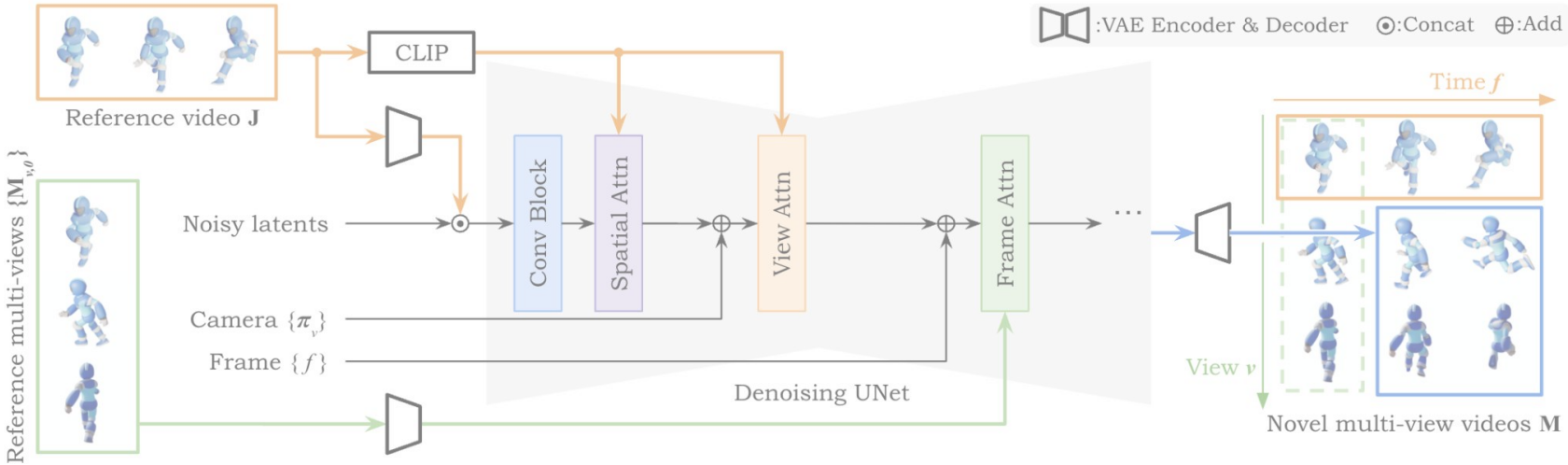
Multi-view Video Generator – SV4D [1]



[1] Xie Y*, Yao CH*, Voleti V, Jiang H, Jampani V. “Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency”. arXiv preprint arXiv:2407.17470.

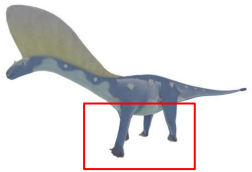
Background

Multi-view Video Generator – SV4D [1]



Limitations:

- **Dependent on the reference multi-views**
  - *Not robust to self-occlusion in the first frame*



Self-occlusion!

First frame of input video

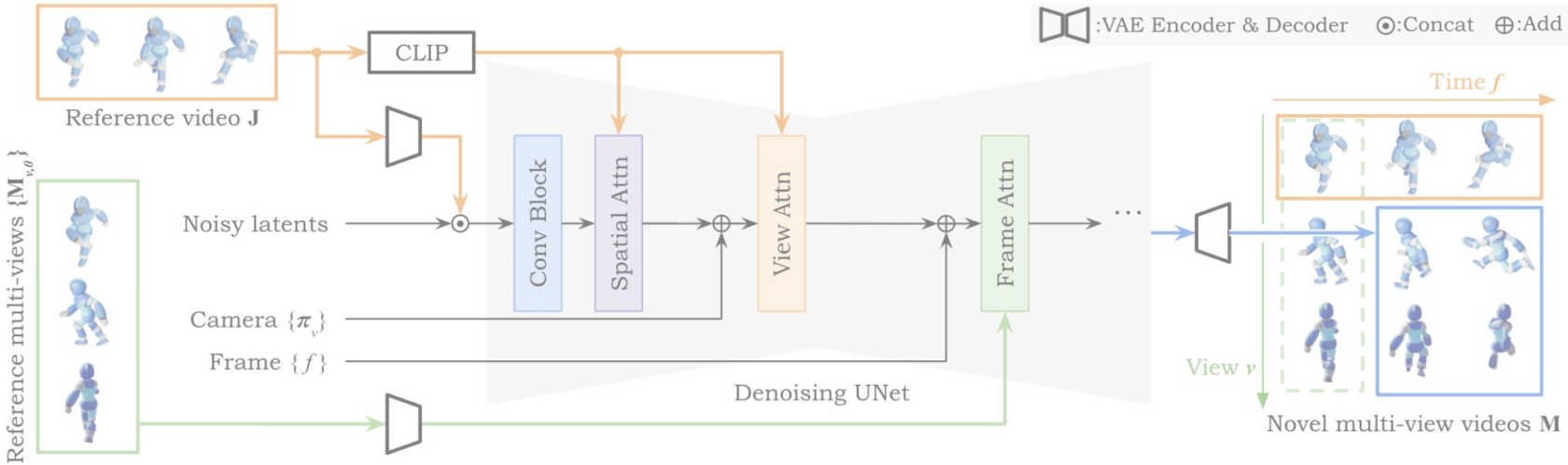
Missing one leg!

Reference multi-views from **SV3D**

SV4D generated novel-view video

Background

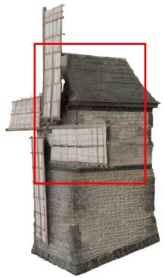
Multi-view Video Generator – SV4D [1]



Limitations:

- Dependent on the reference multi-views
  - *Not robust to self-occlusion in the first frame*
- **Often produces blurry details**

Input video

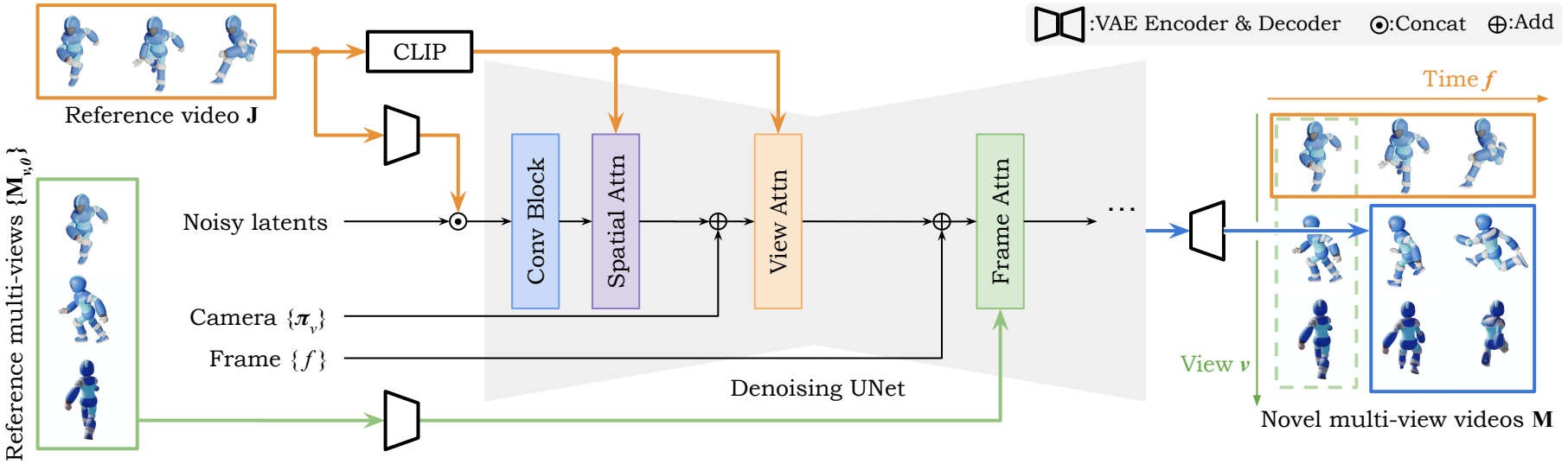


SV4D generated novel-view video

Stable Video 4D 2.0 (SV4D 2.0)

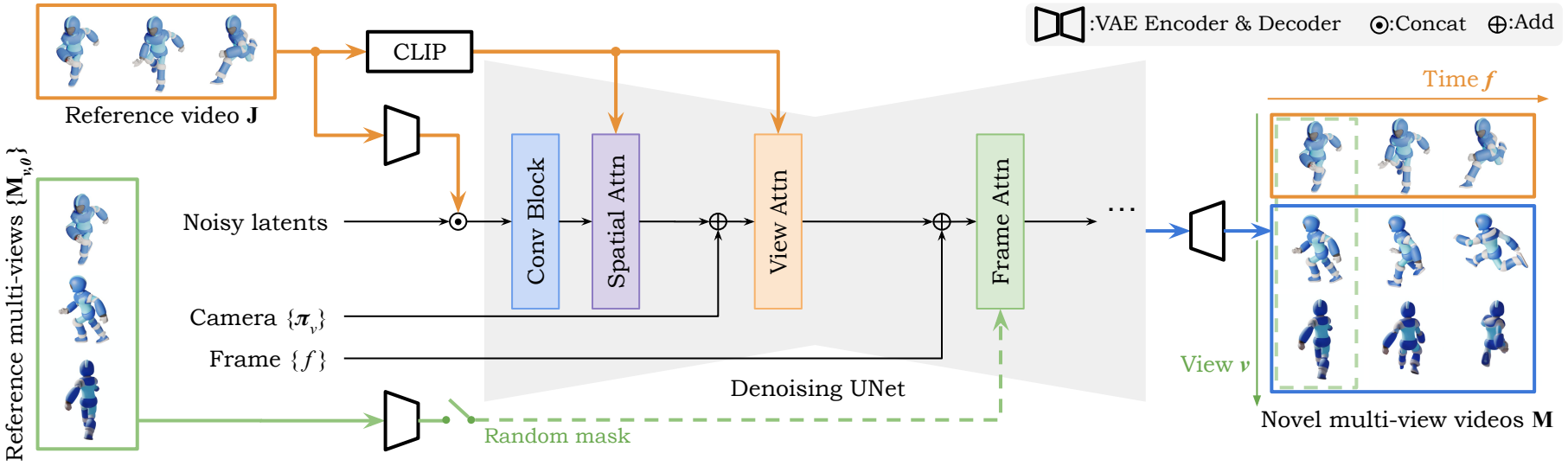
SV4D 2.0

Build upon SV4D



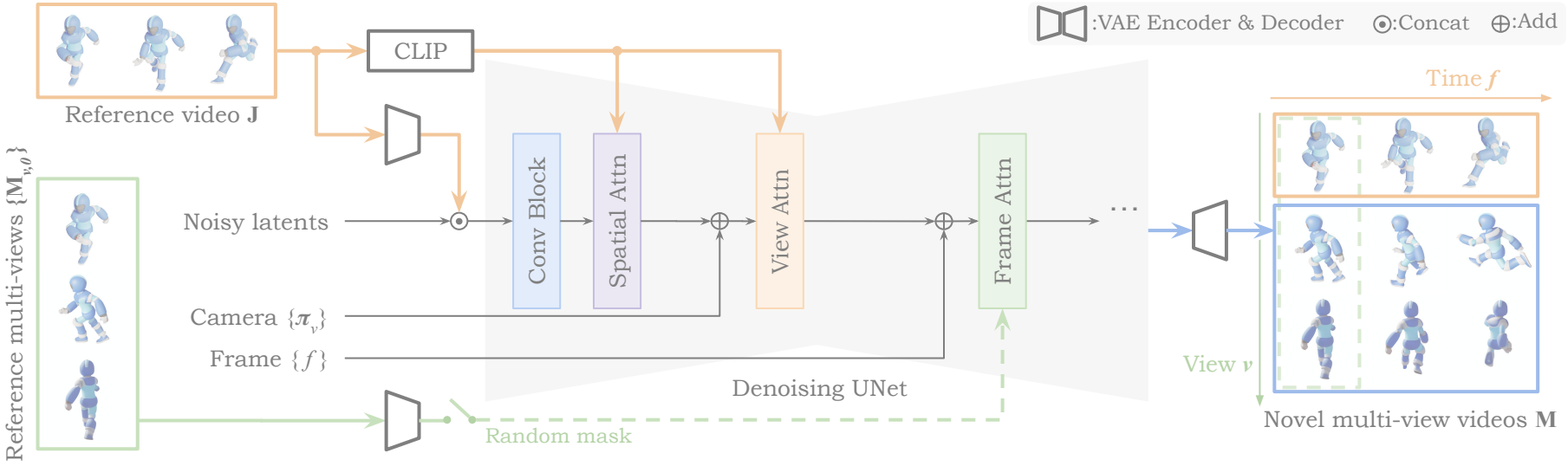
SV4D 2.0

Key Modification 1 – Random mask reference multi-view



SV4D 2.0

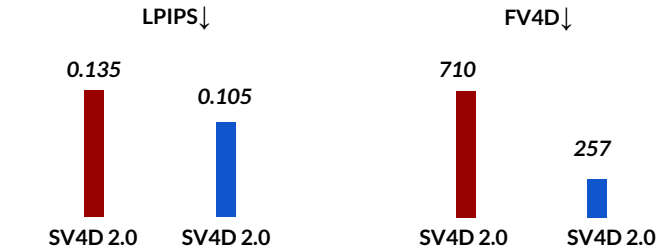
Key Modification 1 – Random mask reference multi-view



Input video

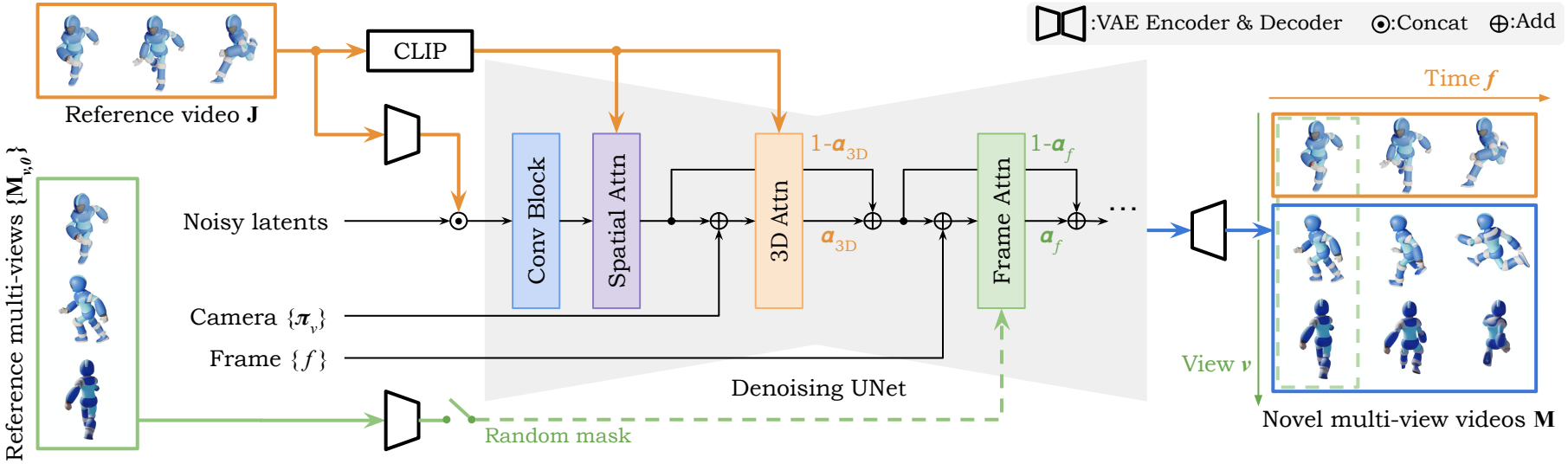
SV4D 2.0 w/o ref. multi-view

SV4D 2.0



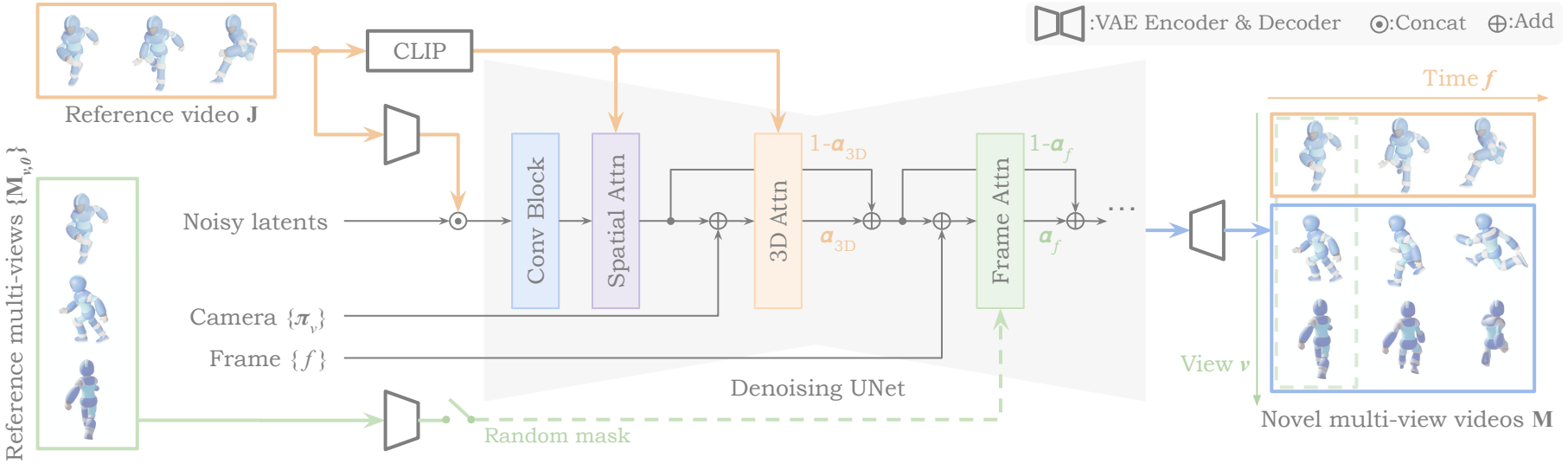
SV4D 2.0

Key Modification 2 – 3D Attention



SV4D 2.0

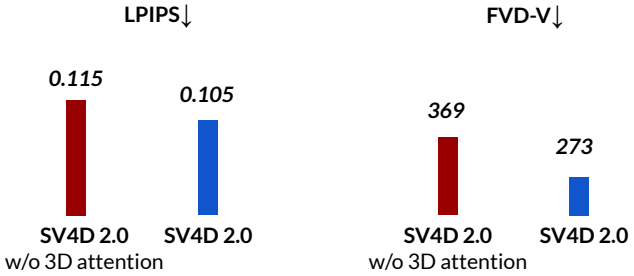
Key Modification 2 – 3D Attention



Input video

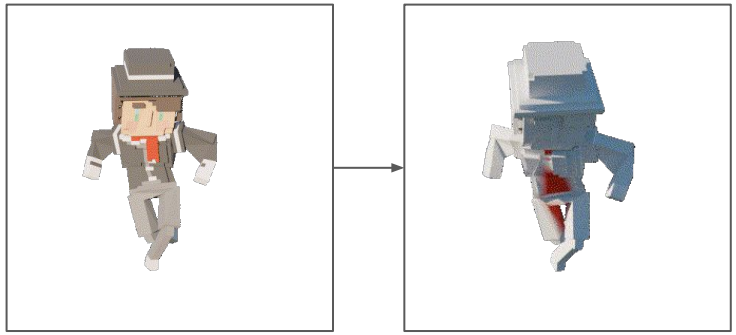
SV4D 2.0 w/o 3D attention

SV4D 2.0



SV4D 2.0

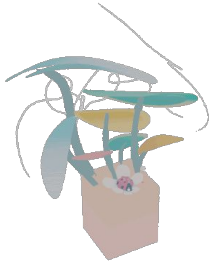
Key Modification 3 – Improving data quality



*Highlighted the most static surface with **RED***

Rectifying off-center objects

Inconsistent scaling



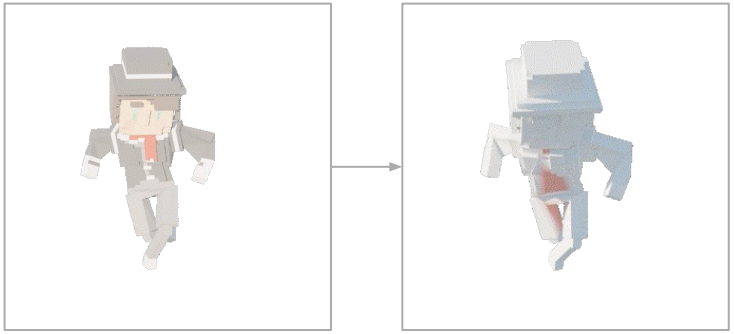
Minimal motion

Dark lighting

Filter out objects/lighting

SV4D 2.0

Key Modification 3 – Improving data quality



Highlighted the most static surface with RED

Rectifying off-center objects

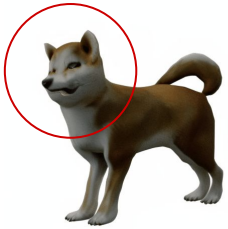
Inconsistent scaling

Small motion

Dark lighting

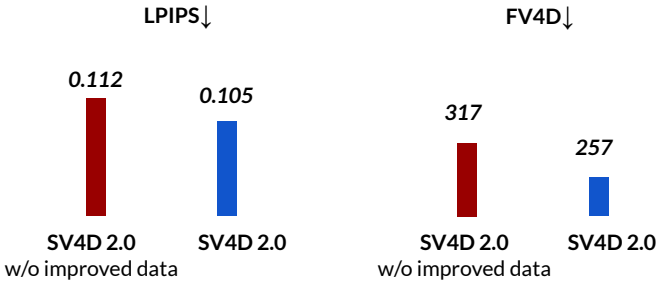
Filter out objects/lighting

Input video



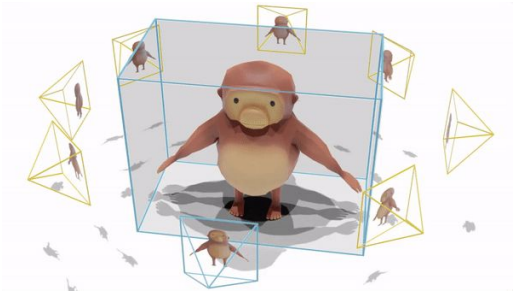
SV4D 2.0 w/o improved data

SV4D 2.0



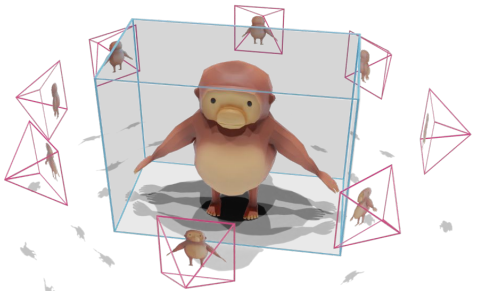
SV4D 2.0

Key Modification 4 – Progressive 3D-to-4D training



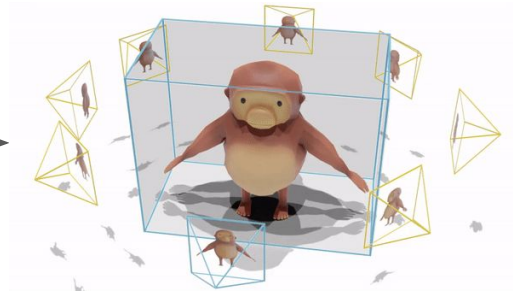
Training with **4D data**

SV4D training



Step 1: Training with **static 3D data**

$a_f = 0$

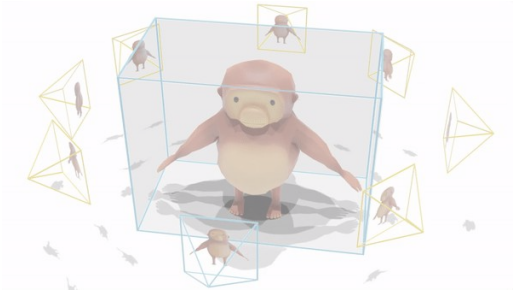


Step 2: Training with **4D data**

SV4D 2.0 training

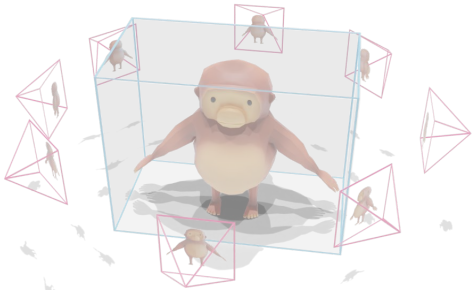
SV4D 2.0

Key Modification 4 – Progressive 3D-to-4D training



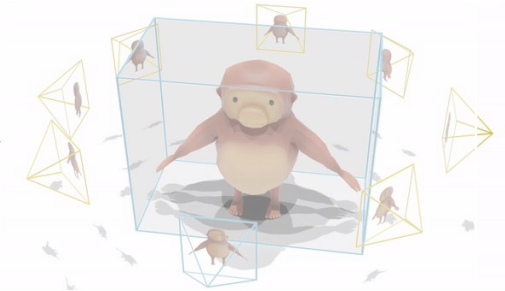
Training with **4D data**

SV4D training



Step 1: Training with **static 3D data**

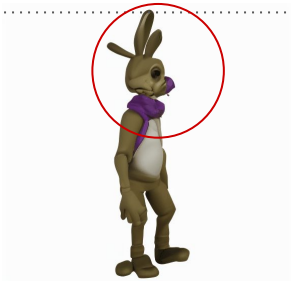
$a_f = 0$



Step 2: Training with **4D data**

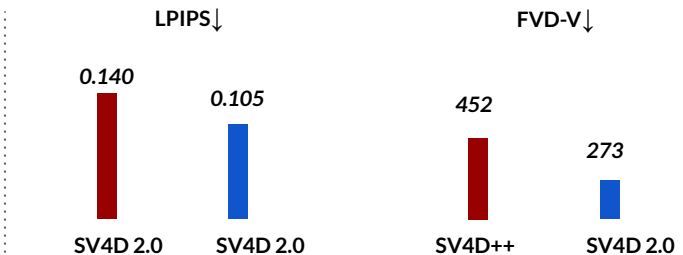
SV4D 2.0 training

Input video



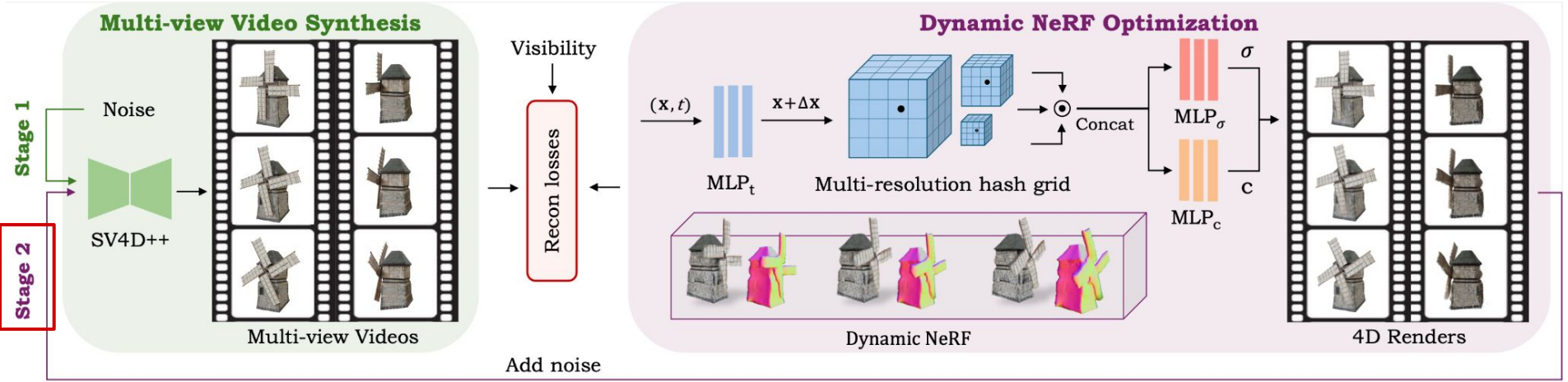
SV4D 2.0 w/o progressive training

SV4D 2.0



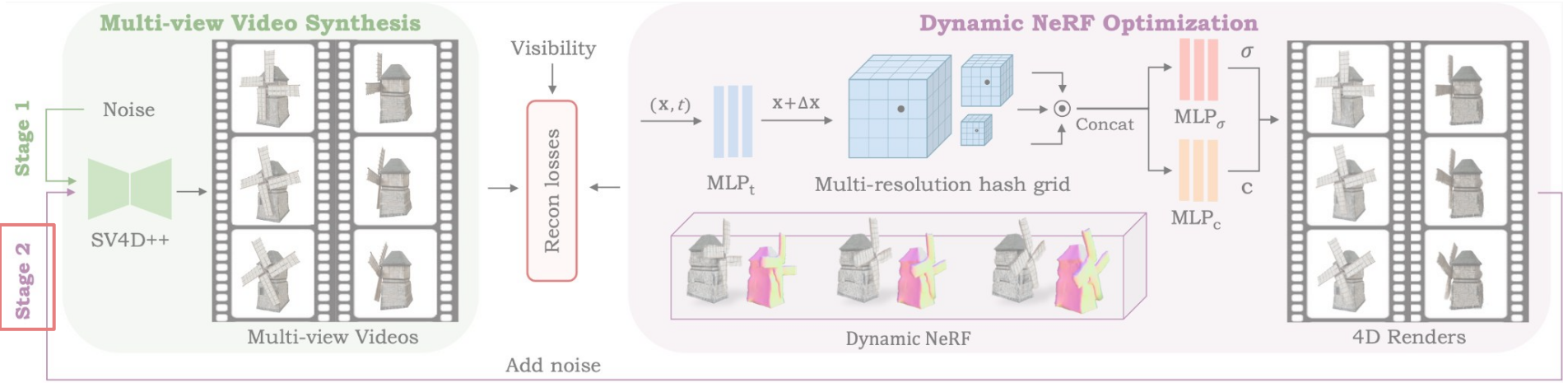
SV4D 2.0

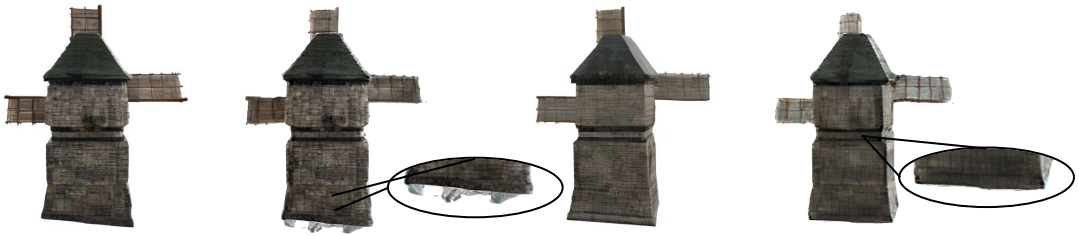
Key Modification 5 – Stage-2 refinement

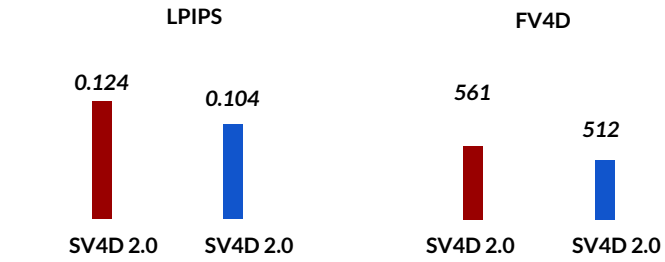


SV4D 2.0

Key Modification 5 – Stage-2 refinement







Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

(The visualization results are also available on the website provided in the Supplementary Material.)

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2








SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2








SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2








SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

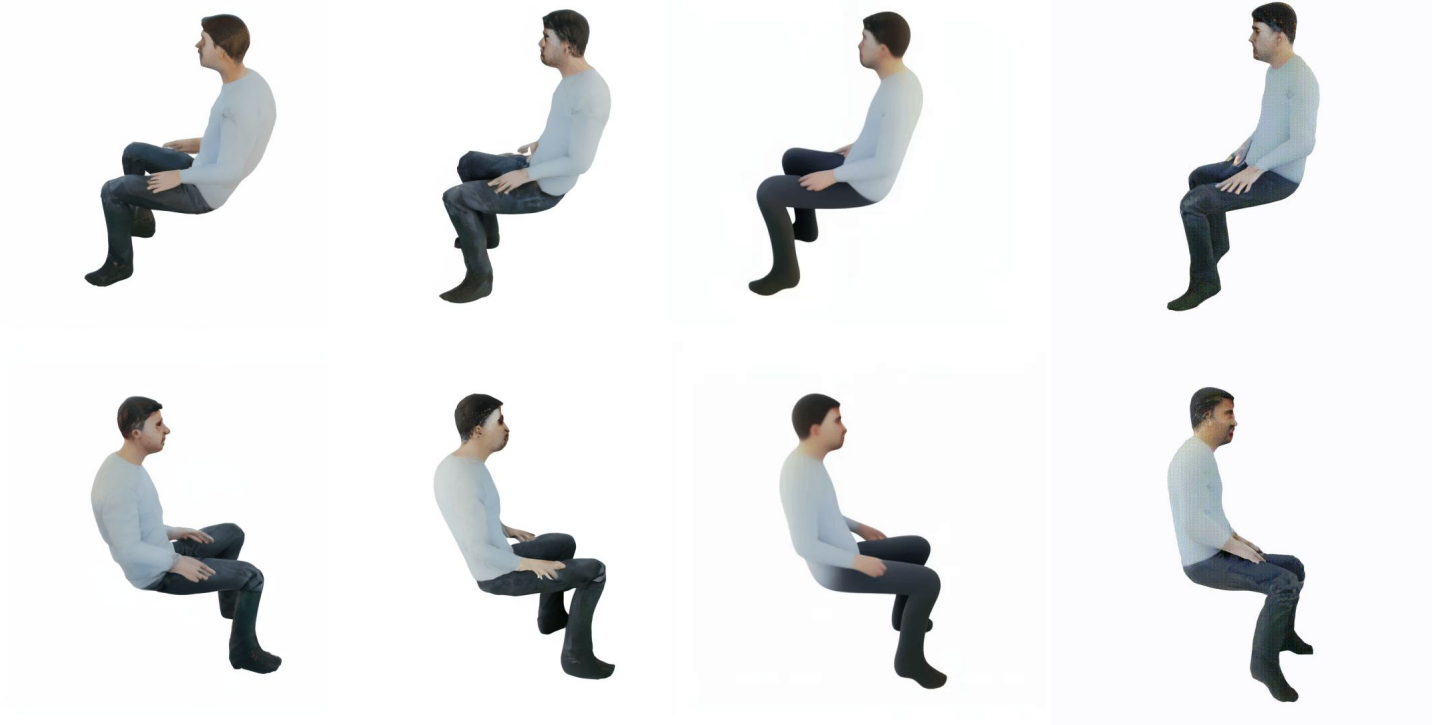
Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2



SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

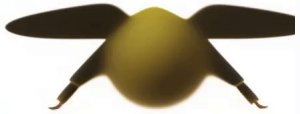
Novel View 1

Novel View 2










SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2

SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

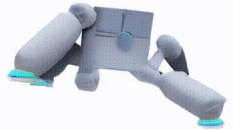
Visual Comparison - Novel-view video synthesis

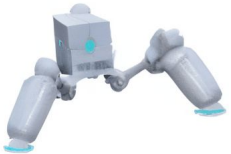
Input Video

Novel View 1

Novel View 2





SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2

SV4D 2.0 (Ours)

SV4D

Diffusion^2

SV3D

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2








SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2

SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

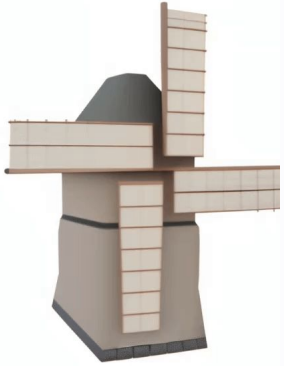
Qualitative Evaluation

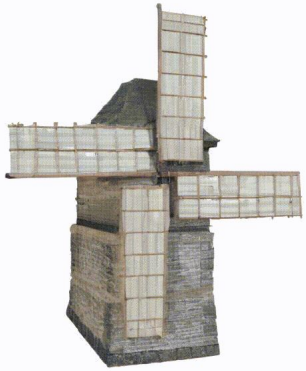
Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2





SV4D 2.0 (Ours) **SV4D** **Diffusion^2** **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

Novel View 1

Novel View 2

SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Qualitative Evaluation

Visual Comparison - Novel-view video synthesis

Input Video

SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video

SV4D 2.0 (Ours)

SV4D

Diffusion^2

SV3D

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video

SV4D 2.0 (Ours)

SV4D

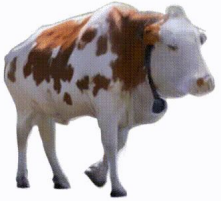
Diffusion^2

SV3D

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video

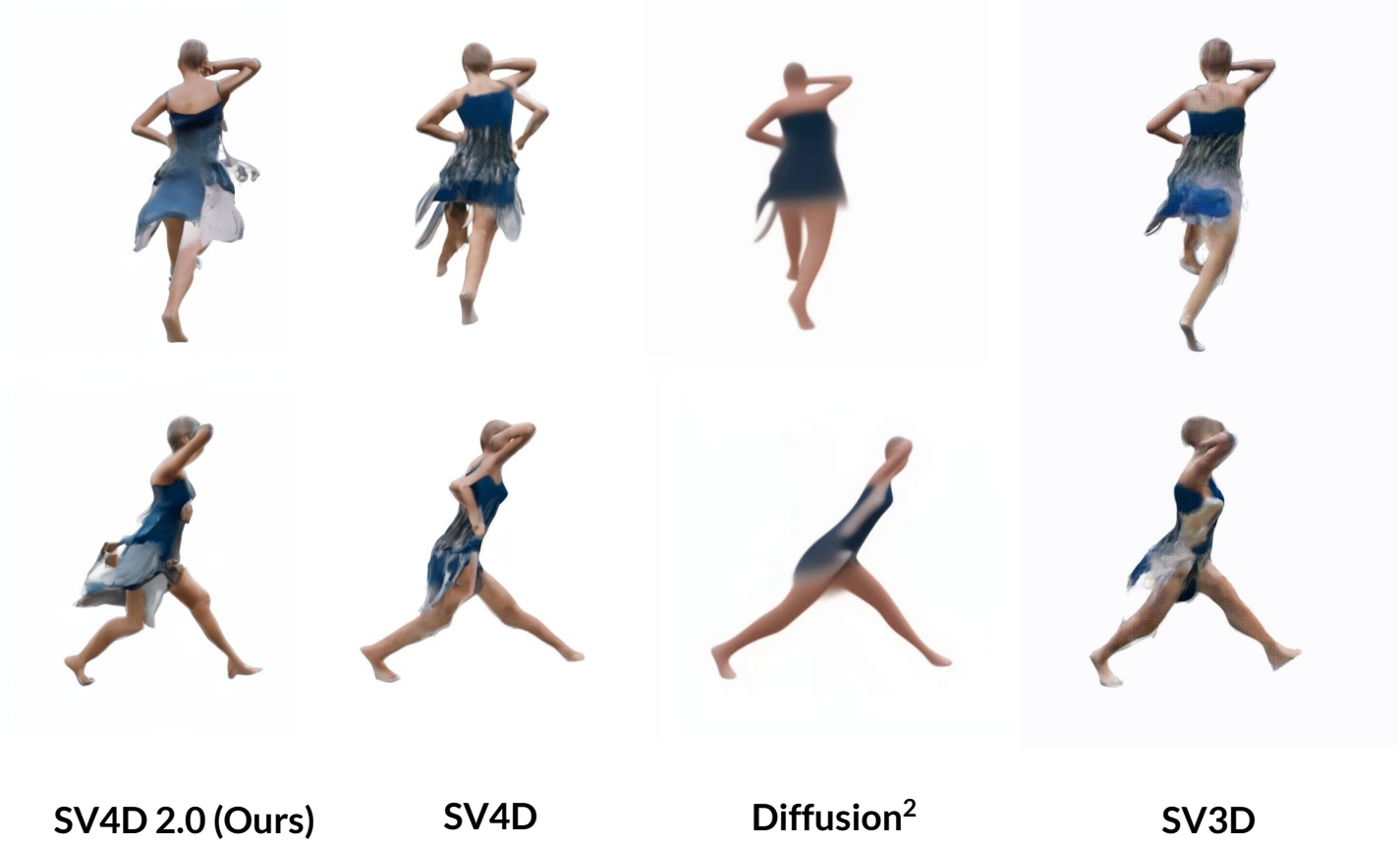


SV4D 2.0 (Ours) **SV4D** **Diffusion^2** **SV3D**

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video



SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video

SV4D 2.0 (Ours)

SV4D

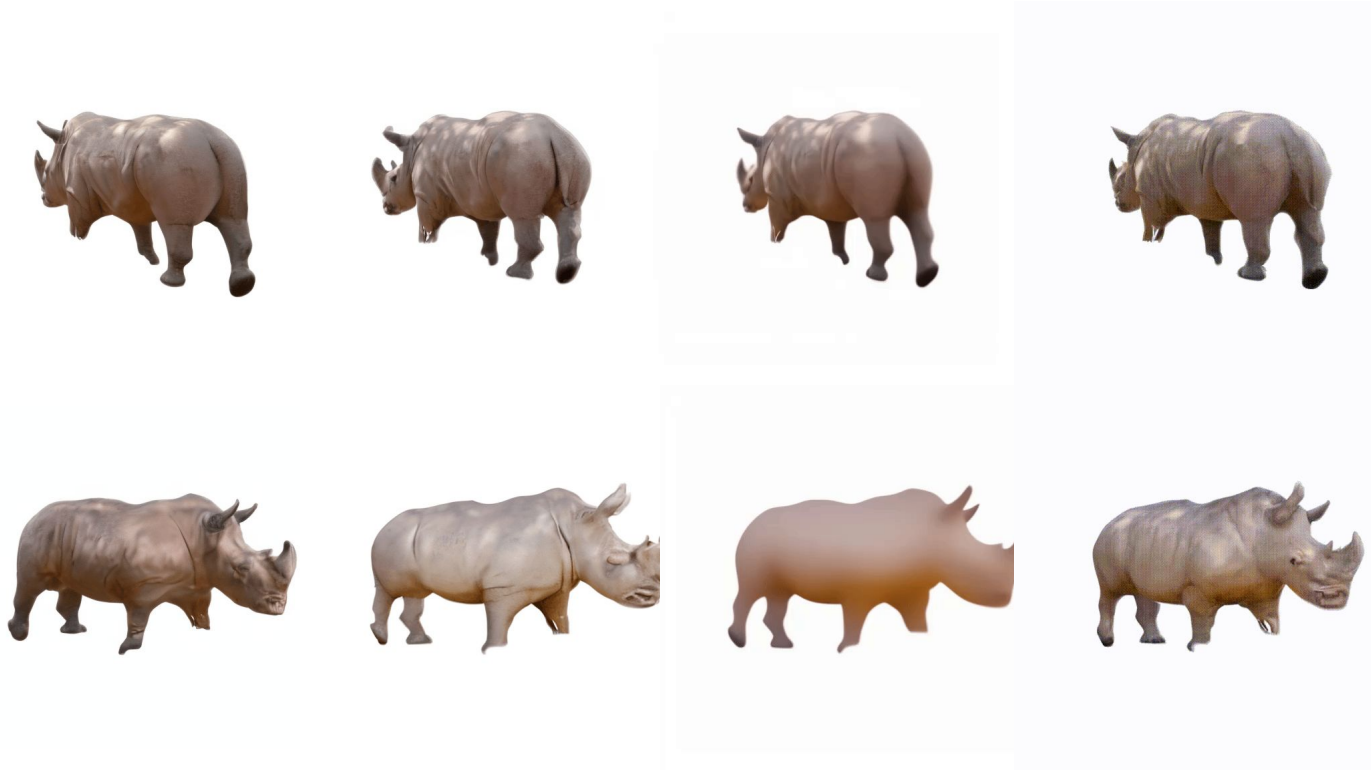
Diffusion^2

SV3D

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video



SV4D 2.0 (Ours) | **SV4D** | **Diffusion^2** | **SV3D**

Visual Comparison

Novel-view video synthesis – Real-world data

Input Video

SV4D 2.0 (Ours) **SV4D** **Diffusion^2** **SV3D**

Qualitative Evaluation

Visual Comparison - 4D Optimization

(The visualization results are also available on the website provided in the Supplementary Material.)

Visual Comparison

4D Optimization – Synthetic data

L4GM struggles with videos at non-zero elevation (training data primarily at 0° elevation).

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Synthetic data

L4GM struggles with videos at non-zero elevation (training data primarily at 0° elevation).

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Synthetic data

L4GM struggles with videos at non-zero elevation (training data primarily at 0° elevation).

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Real-world data

L4GM does not generalize well on real-world data (no video prior like ours)

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Real-world data

L4GM does not generalize well on real-world data (no video prior like ours)

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Real-world data

L4GM does not generalize well on real-world data (no video prior like ours)

Input Video

SV4D 2.0 (Ours) | **SV4D** | **STAG4D** | **L4GM** | **Consistent4D**

Visual Comparison

4D Optimization – Real-world data

L4GM does not generalize well on real-world data (no video prior like ours)

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Real-world data

L4GM does not generalize well on real-world data (no video prior like ours)

Input Video

SV4D 2.0 (Ours)

SV4D

STAG4D

L4GM

Consistent4D

Visual Comparison

4D Optimization – Real-world data

L4GM does not generalize well on real-world data (no video prior like ours)

Input Video

SV4D 2.0 (Ours) | **SV4D** | **STAG4D** | **L4GM** | **Consistent4D**

More Results on Real-world Video

(The visualization results are also available on the website provided in the Supplementary Material.)

More Real-world Results

Novel-view video synthesis

Raw Input

Masked Input

Input Video (real data)

Novel-view Video Synthesis

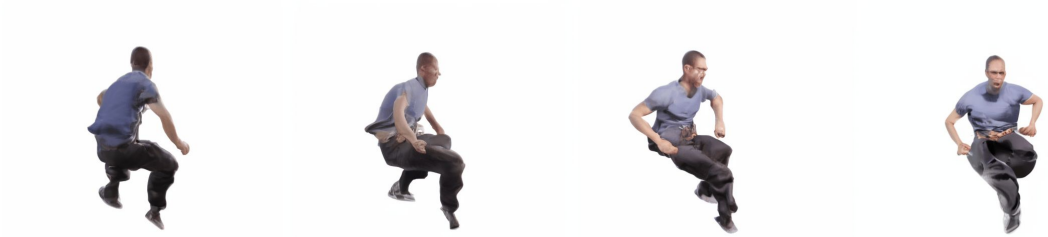
More Real-world Results

Novel-view video synthesis

Raw Input

Masked Input

Input Video (real data)



Novel-view Video Synthesis

More Real-world Results

Novel-view video synthesis

Raw Input

Masked Input

Input Video (real data)

Novel-view Video Synthesis

More Real-world Results

Novel-view video synthesis

Raw Input

Masked Input

Input Video (real data)

Novel-view Video Synthesis

More Real-world Results

Novel-view video synthesis

Raw Input

Masked Input

Input Video (real data)



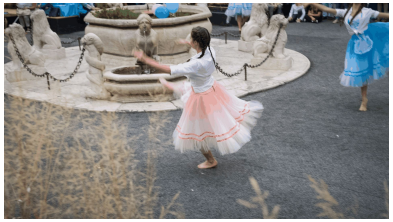





Novel-view Video Synthesis

More Real-world Results

Novel-view video synthesis



Raw Input

Masked Input

Input Video (real data)

Novel-view Video Synthesis

4D Optimization Results with Continuous View and Time Changes

(The visualization results are also available on the website provided in the Supplementary Material.)

4D Optimization Results

Continuous View and Time Changes

Input Video

4D Optimization

Input Video

4D Optimization

Input Video

4D Optimization

Input Video

4D Optimization

4D Optimization Results

Continuous View and Time Changes

Input Video

4D Optimization

Input Video

4D Optimization

Input Video

4D Optimization

Input Video

4D Optimization

SV4D 2.0 with DyNeRF *vs* 4D Gaussians

(The visualization results are also available on the website provided in the Supplementary Material.)

4D Optimization

DyNeRF vs 4D Gaussians

In our sparse-view setting:

- 4D Gaussians suffer from **temporal flickering and floater artifacts** due to its discrete nature
- DyNeRF interpolates better across sparse views and fast motion

4D Gaussians

DyNeRF

4D Gaussians

DyNeRF