4Real-Video-V2:
Fused View-Time Attention and Feedforward Reconstruction for 4D Scene Generation

Chaoyang Wang*, Ashkan Mirzaei*, Vidit Goel, Willi Menapace, Aliaksandr Siarohin, Avalon Vinella, Michael Vasilkovsky, Ivan Skorokhodov, Vladislav Shakhrai, Sergey Korolev, Sergey Tulyakov, Peter Wonka

Snap Inc., KAUST, * Denotes equal contribution

Problem Statement

1. Text-to-video models are improving, but 4D generation is still limited
2. Need synchronized multi-view videos to build dynamic 3D content
3. Existing 4D methods suffer from view inconsistency or heavy optimization

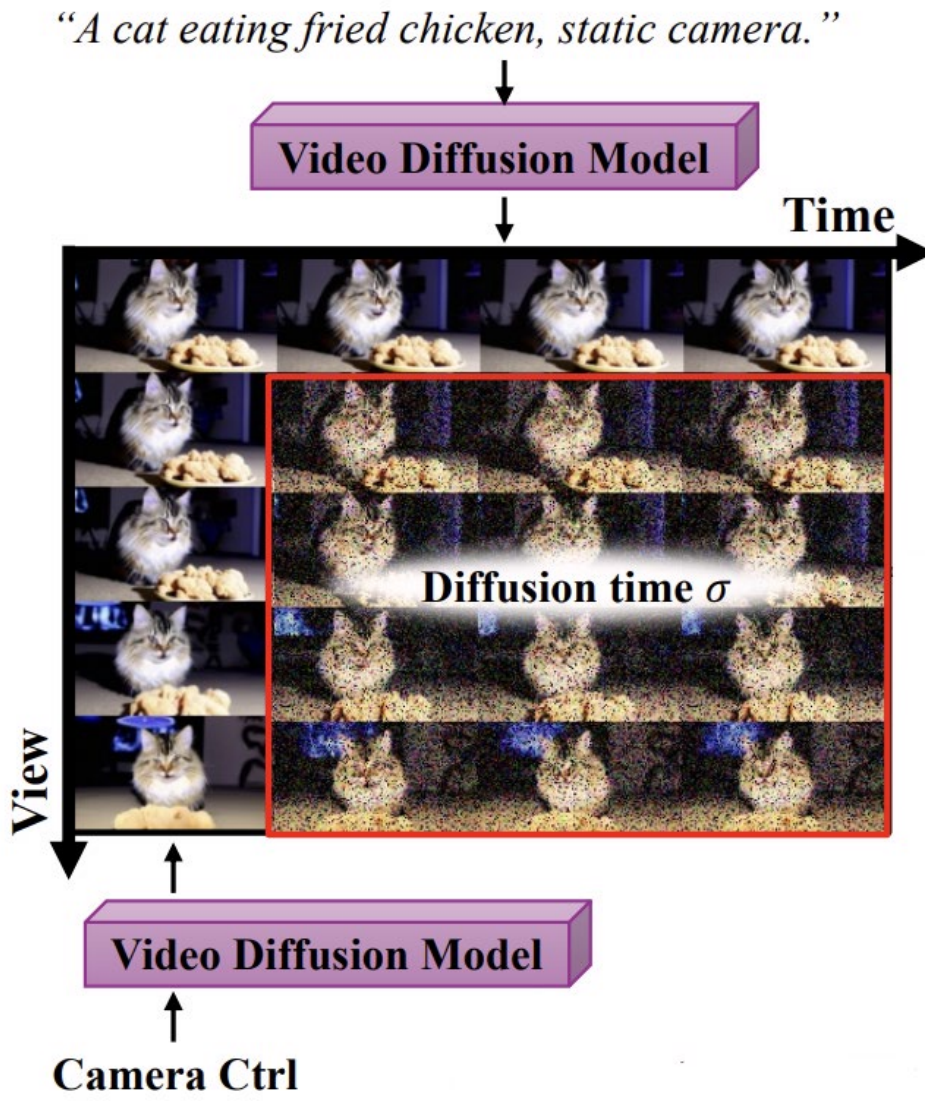
Key Challenges in 4D Generation

- Data Scarcity
  - Limited availability of high-quality 3D and 4D data
- Architectural Complexity
  - Handle multiple viewpoints and temporal dynamics
  - Synchronization across views and time
- Reconstruction Challenges
  - Traditional methods are slow and iterative

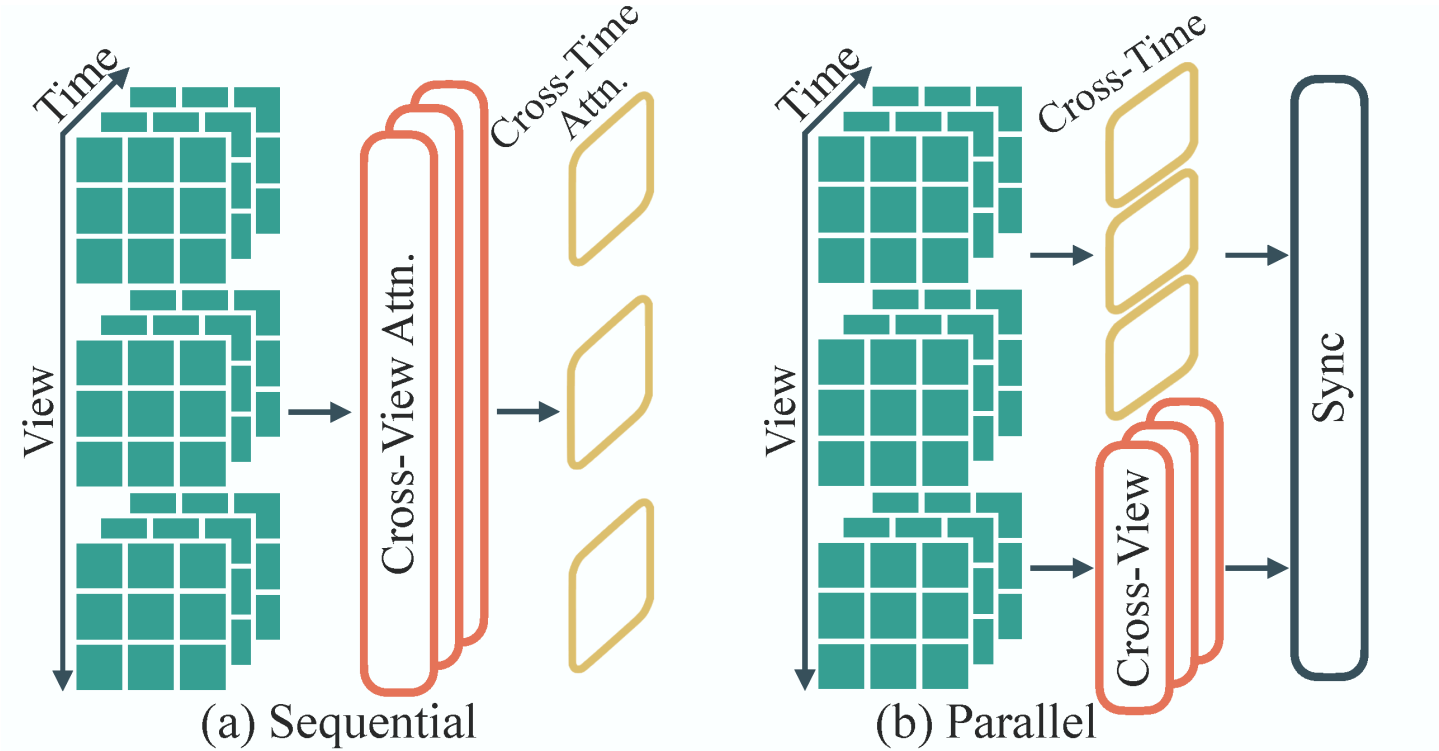
Key Contributions

- Parameter-efficient: no extra weights beyond pre-trained video model
- Pose-free Gaussian-splat reconstructor with camera-token replacement
- End-to-end runtime: 4 min for 8 views × 29 frames on 1× A100
- State-of-the-art quality on Objaverse & NVIDIA Dynamic benchmarks

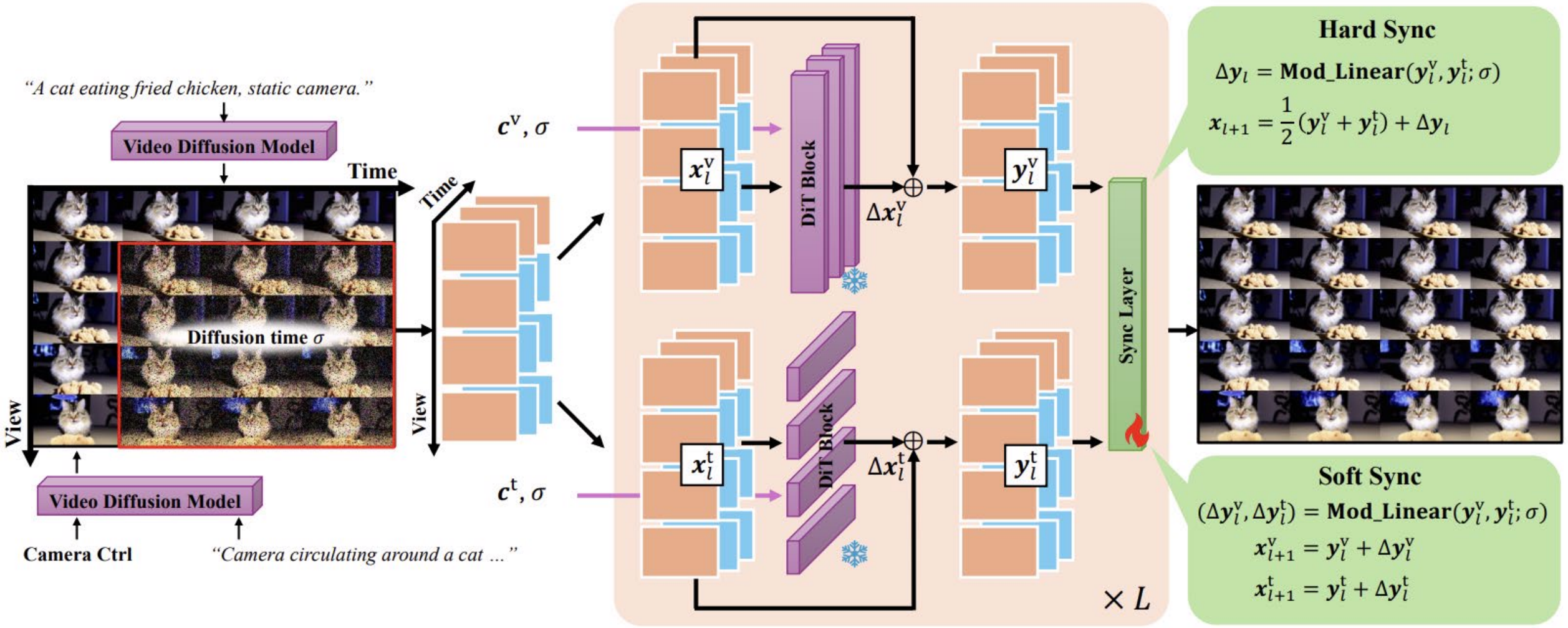
Prior Work: 4Real-Video



Prior Work: 4Real-Video



Prior Work: 4Real-Video



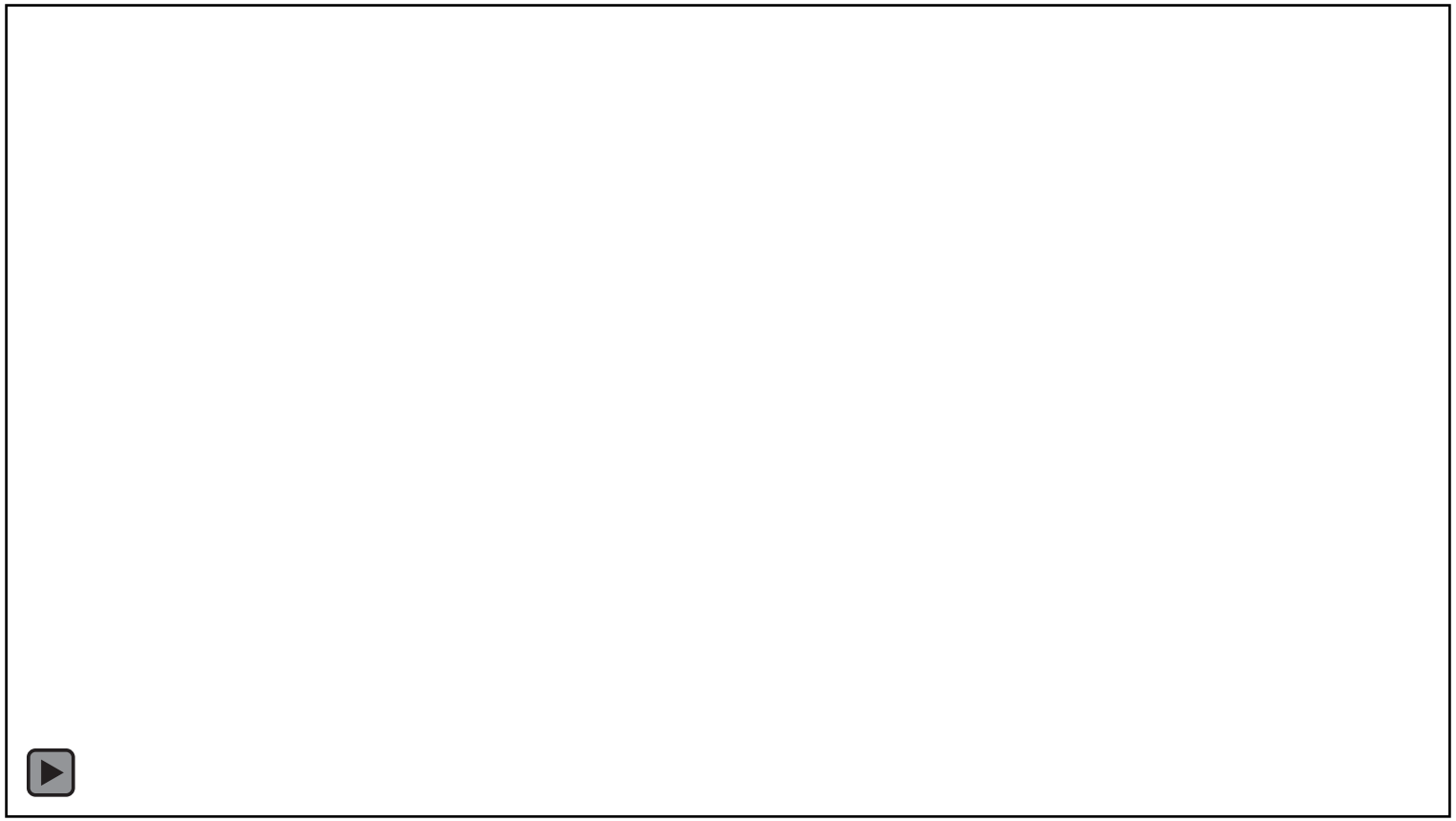
Ablation Study

Method	FID ↓	CLIP ↑	FVD ↓		FVD-Test ↓		Visual Quality↑		Temporal Consist.↑		Factual Consist. ↑	
			Time	View	Time	View	Time	View	Time	View	Time	View
SV4D [44]	204.81	19.46	1053.10	1245.42	814.50	323.99	2.26	2.02	2.03	1.68	2.12	1.99
MotionCtrl [42]	87.10	20.20	1556.36	1509.76	1170.04	302.18	2.36	2.30	2.38	2.25	2.38	2.33
Sequential	96.64	28.16	1662.54	1797.15	897.08	597.19	2.30	2.28	2.21	2.15	2.23	2.20
Soft w/o Obj	80.17	28.11	1392.48	1720.47	318.18	302.18	2.41	2.39	2.37	2.31	2.35	2.33
Hard Sync	79.92	28.16	972.87	1045.35	316.14	**251.44**	2.42	2.40	2.40	2.33	2.37	2.34
Soft Sync	**78.36**	**28.22**	**906.16**	**1036.00**	**308.15**	261.02	**2.43**	**2.42**	**2.41**	**2.36**	**2.38**	**2.36**

Limitations of 4Real-Video

- Architecture Bottleneck:
  - Two-stream design with a synchronization layer, restricting model capacity and efficiency.
- Explicit 3D Output:
  - Needed slow, optimization-based methods (like NeRF fitting) for 3D geometry extraction.
- Freeze-time Robustness:
  - Struggled to keep dynamic content stable and consistent across multiple views in freeze-time scenarios.

How Can We Predict 3D Geometry Directly?



VGG Transformer

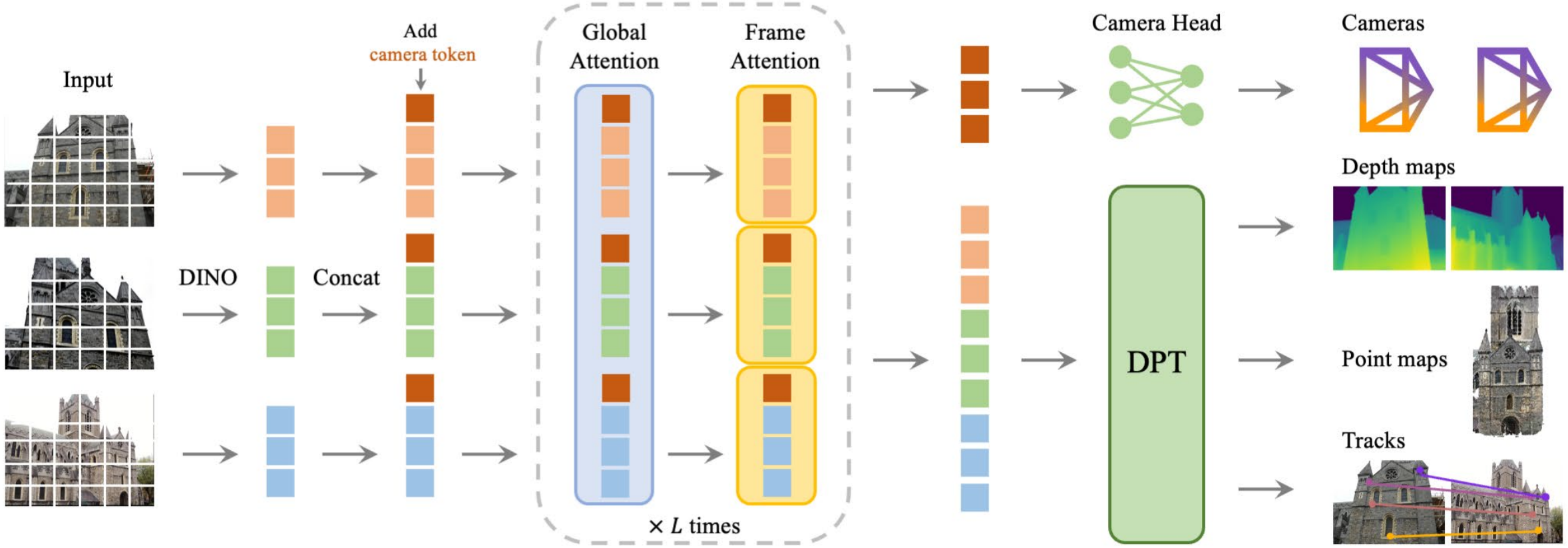
Images

VGGT

Reconstruction

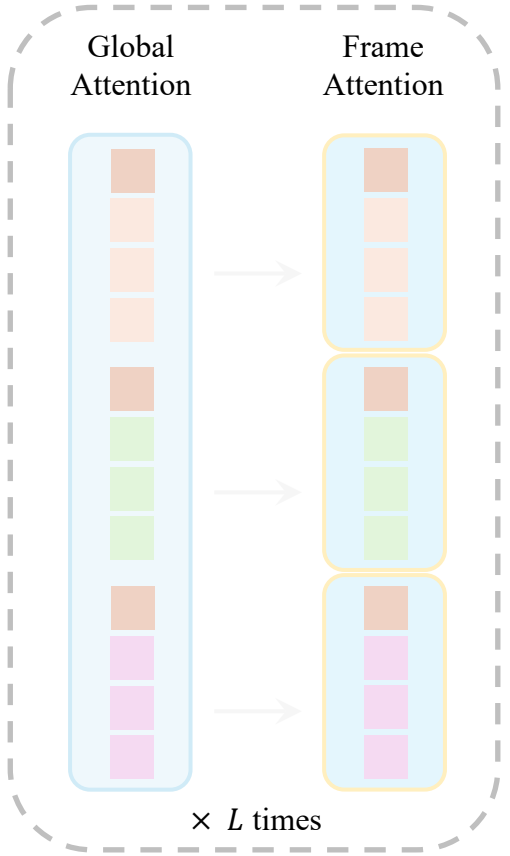
Cameras, Depths, Points, and Correspondences

VGG Transformer



Why Alternating-Attention?

- Global Attention
  - Ensures scene-level coherence
- Frame-wise Attention
  - Eliminates frame index embedding
    - For permutation equivariance
    - For flexible input length

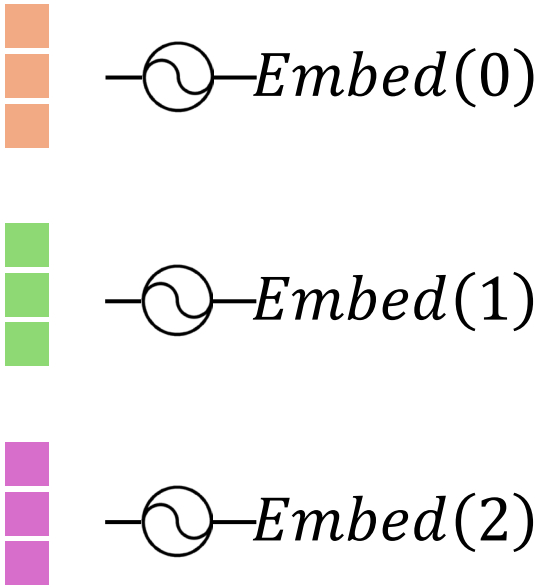


Why Alternating-Attention?

Frame 0 

Frame 1 

Frame 2 



Model()

$\neq$

Model()

Not permutation equivariant

Why Alternating-Attention?

Frame 0  $Embed(0)$

Frame 1  $Embed(1)$

Frame 2  $Embed(2)$

⋮

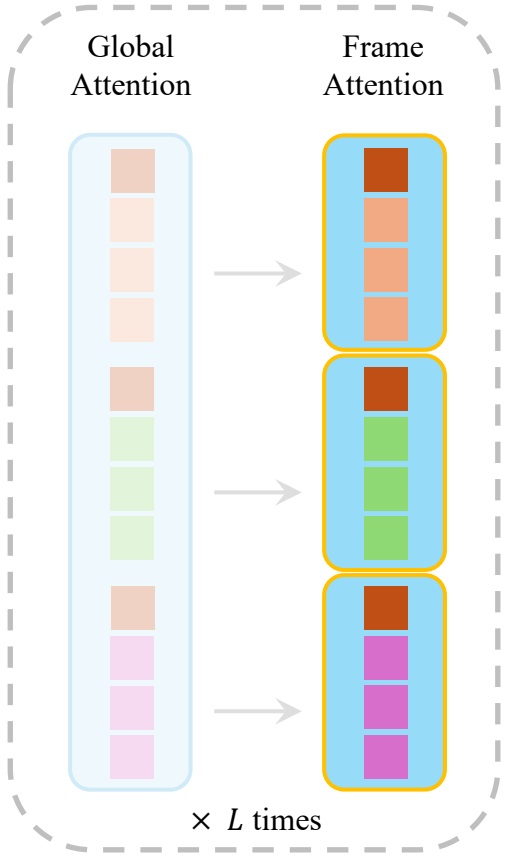
Frame 842 But model never sees $Embed(842)$ during training

Why Alternating-Attention?

Frame 0  $Embed(0)$

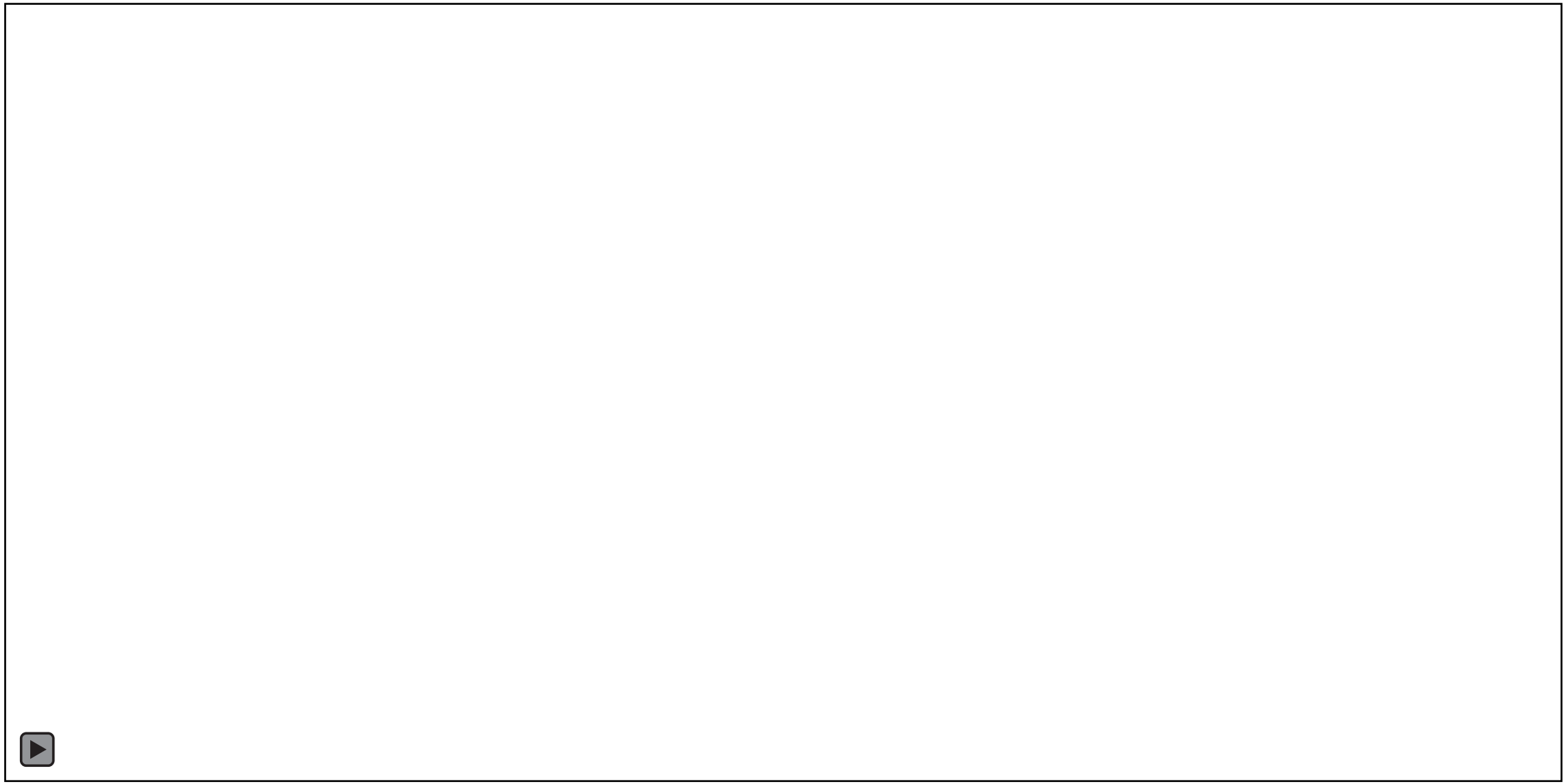
Frame 1  $Embed(1)$

Frame 2  $Embed(2)$



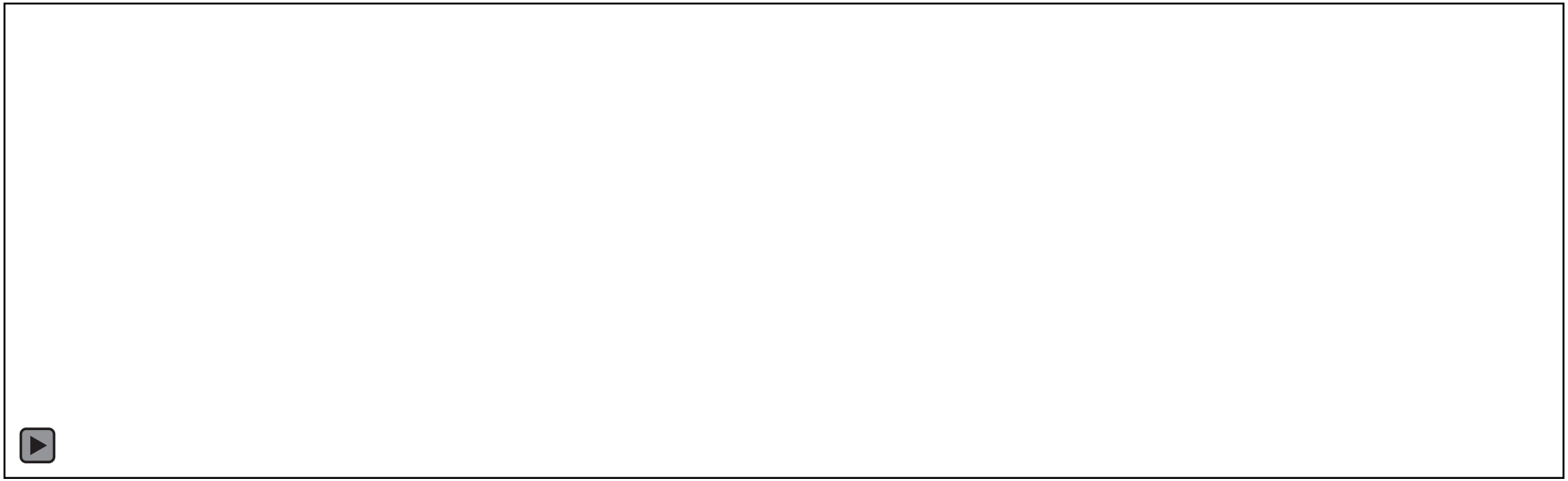
Replaces frame index embedding by Frame-wise Attention

VGGT Examples

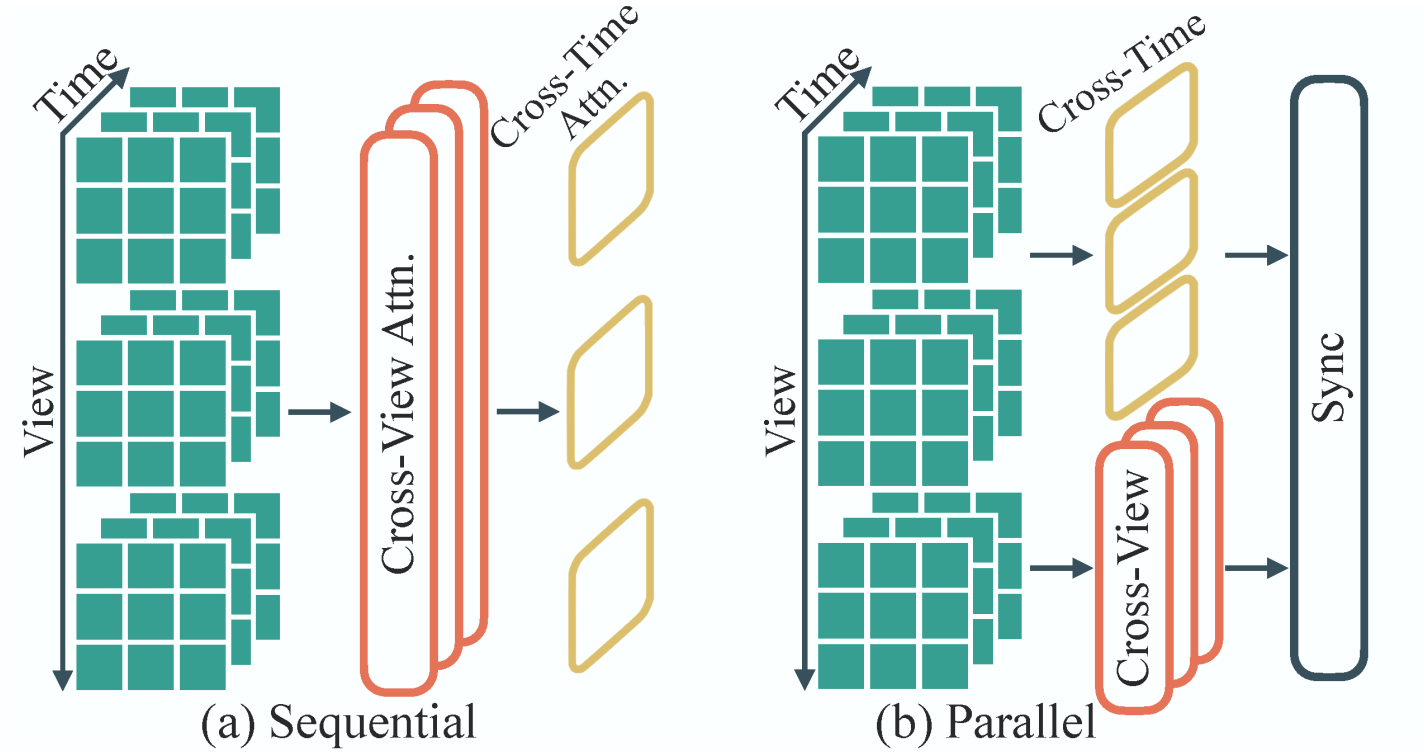


Let's Dive into the Main Method

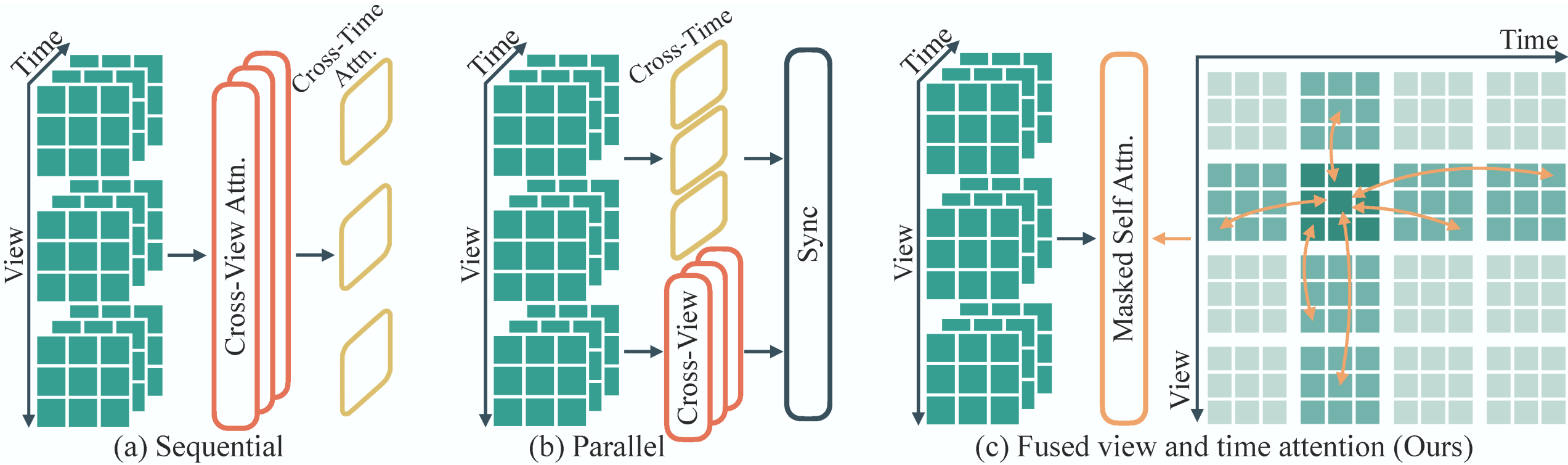
4Real-Video-V2



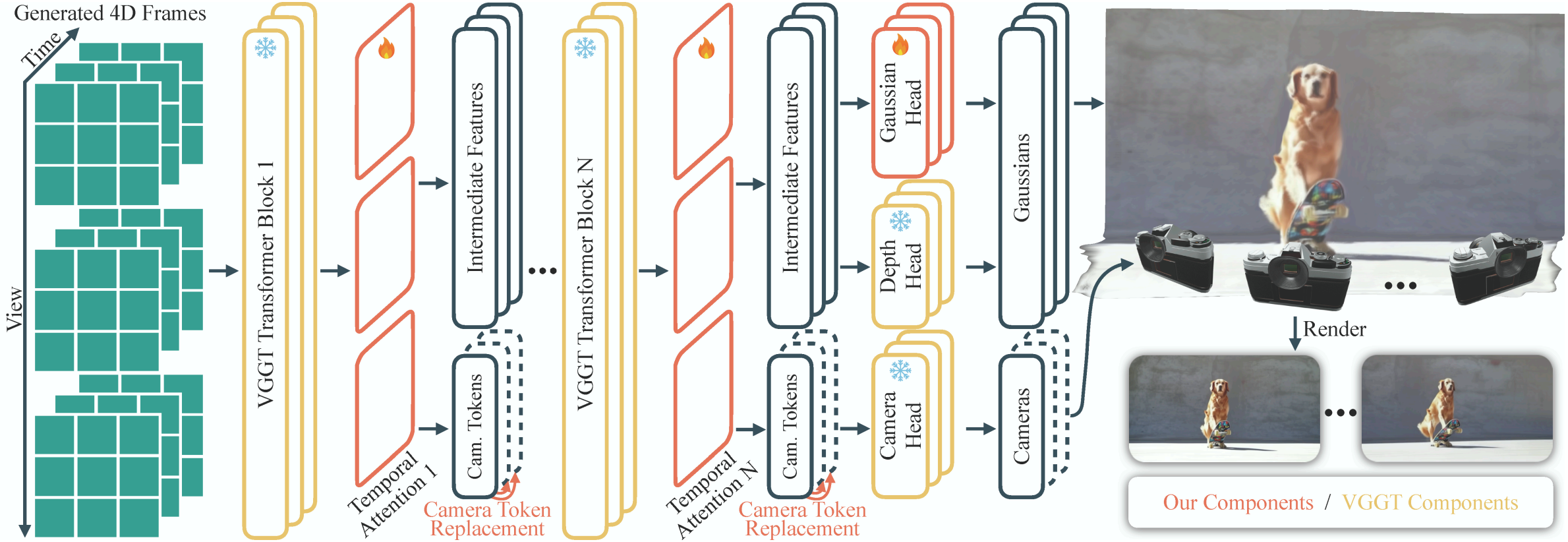
Architectures for 4D Video Generation



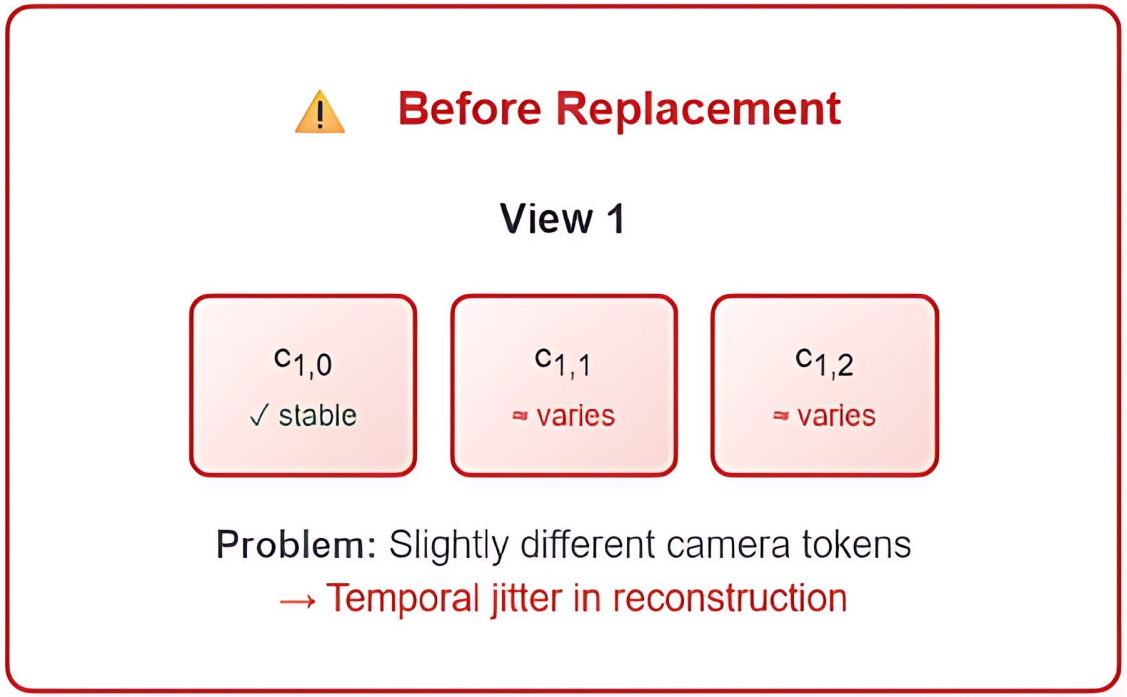
Architectures for 4D Video Generation



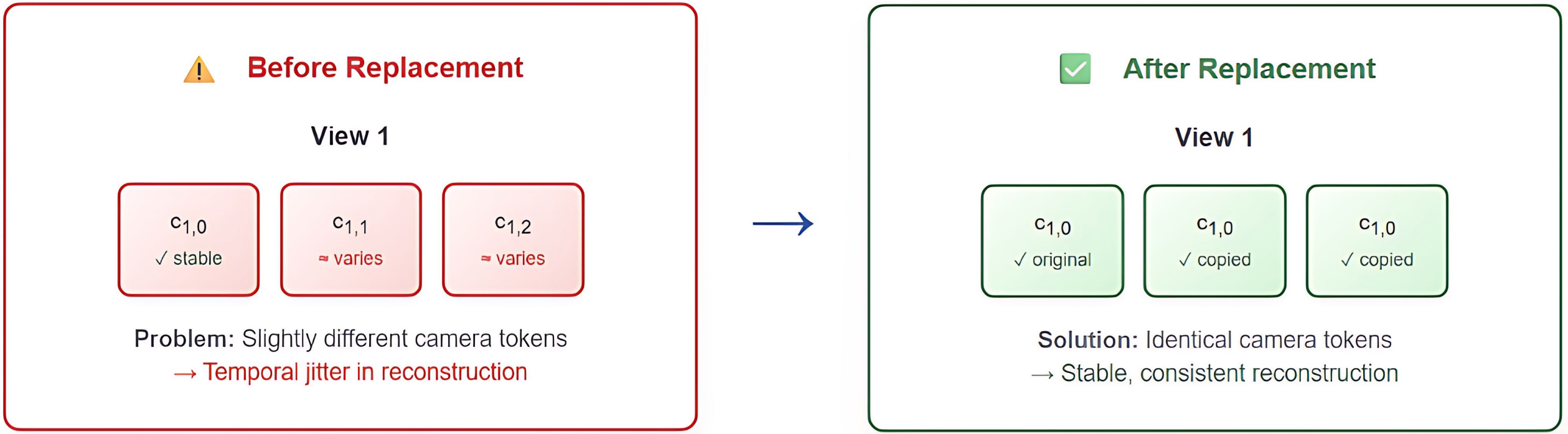
Feedforward Reconstruction Model



Camera Token Replacement



Camera Token Replacement



3D Gaussian Parameters

📍 Position

$$\mu = (x, y, z)$$

3D spatial coordinates of the Gaussian center in world space

🔄 Rotation

$$q = (w, x, y, z)$$

Quaternion representing 3D orientation of the Gaussian ellipsoid

📏 Scale

$$s = (sx, sy, sz)$$

3D scaling factors along each axis of the Gaussian ellipsoid

👁️ Opacity

$$\alpha \in [0, 1]$$

Transparency value controlling the Gaussian's contribution to rendering

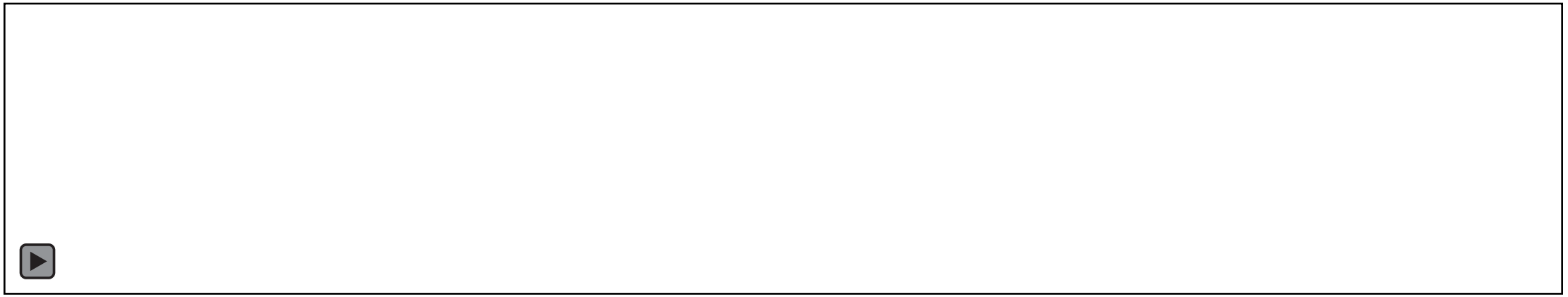
Ablation Study

Method	PSNR
w/o cam. token replacement	22.48
w/o temporal attention	22.60
Full model	**23.39**

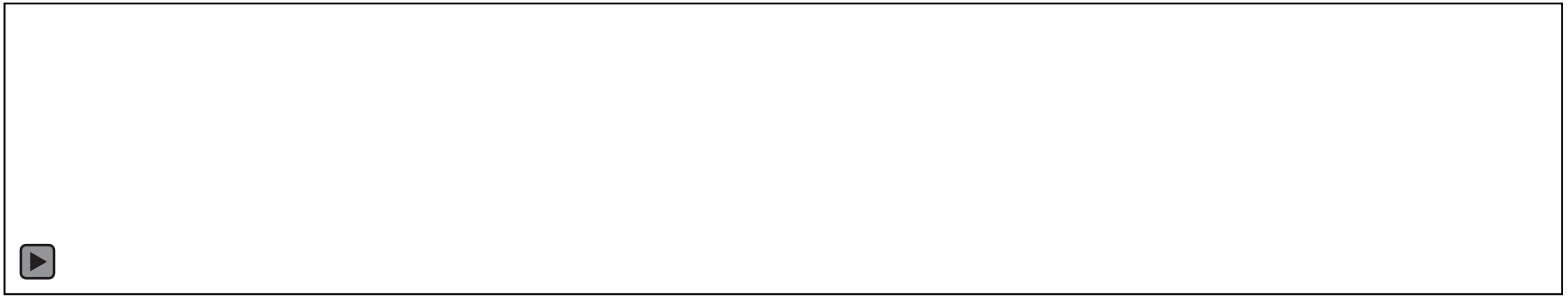
Ablation of the feedforward reconstruction model on the test set of the dynamic Kubric dataset.

Generate 4D Videos from Text

Frozen Time:

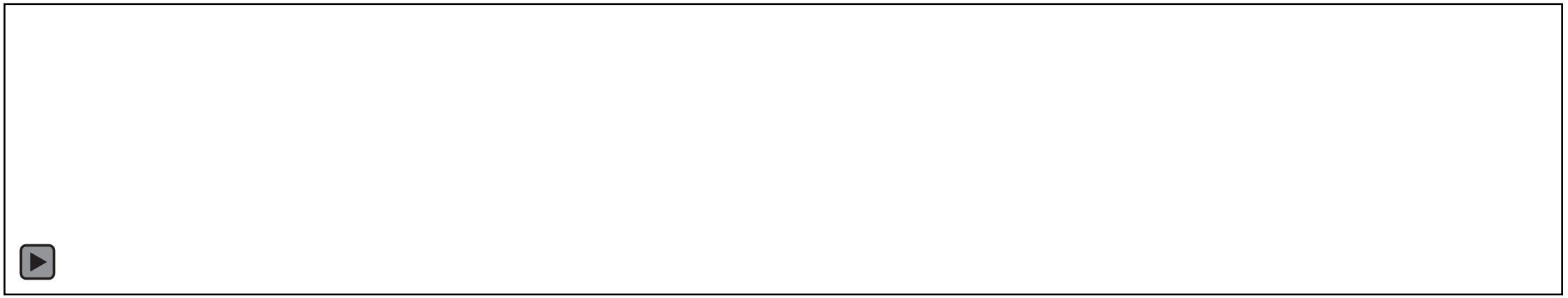


Fixed view:

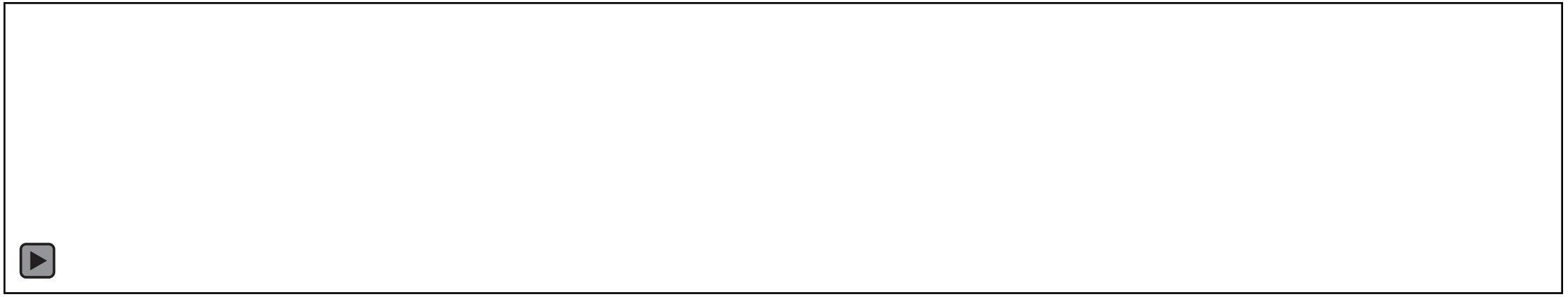


Generate 4D Videos from Text

Frozen Time:

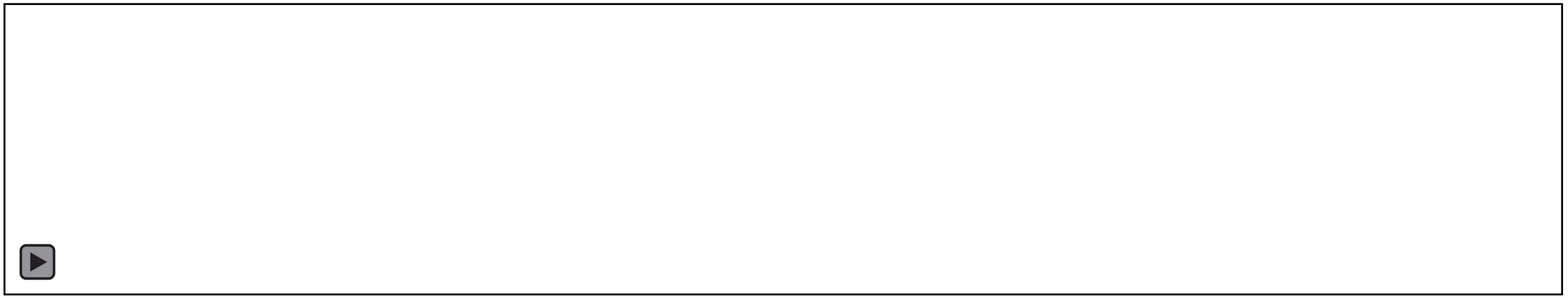


Fixed view:

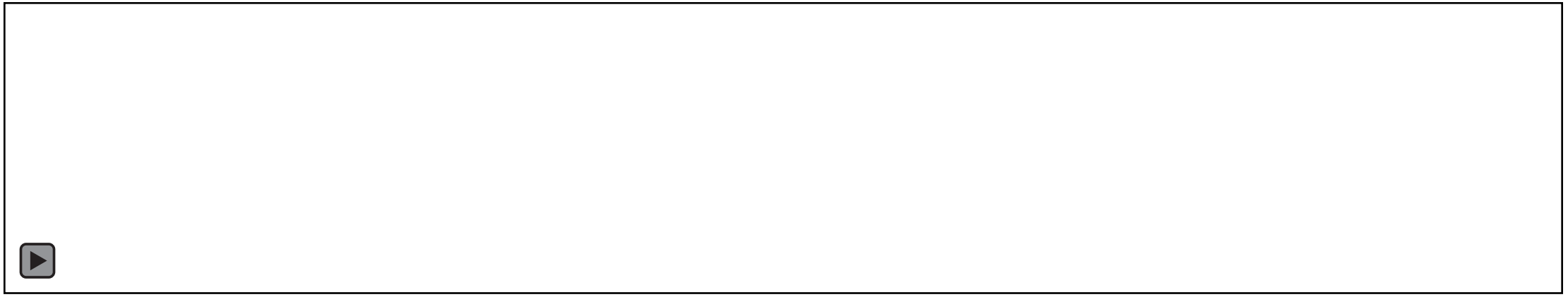


Animating Real 3D Scenes

Frozen Time:

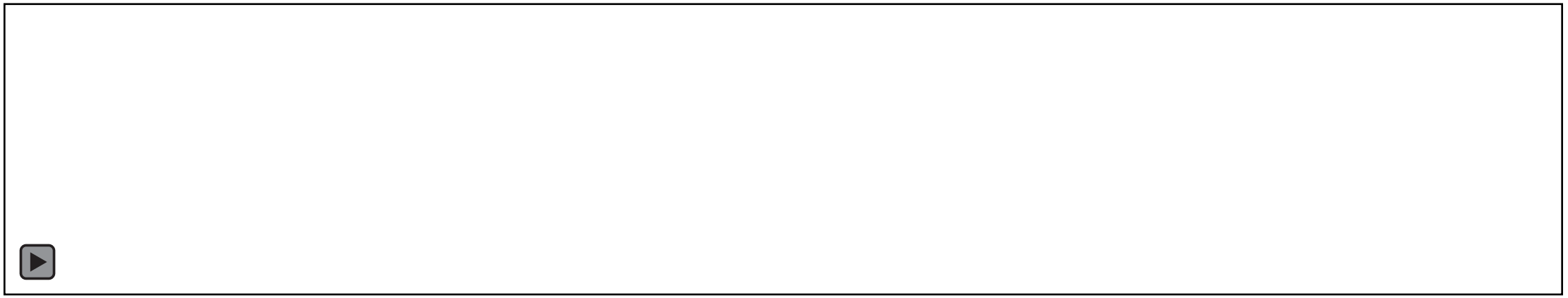


Fixed view:

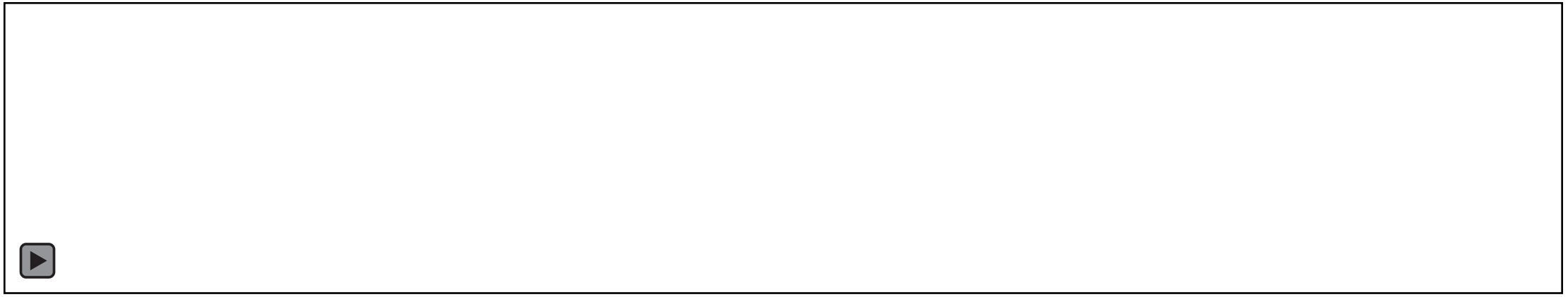


Animating Real 3D Scenes

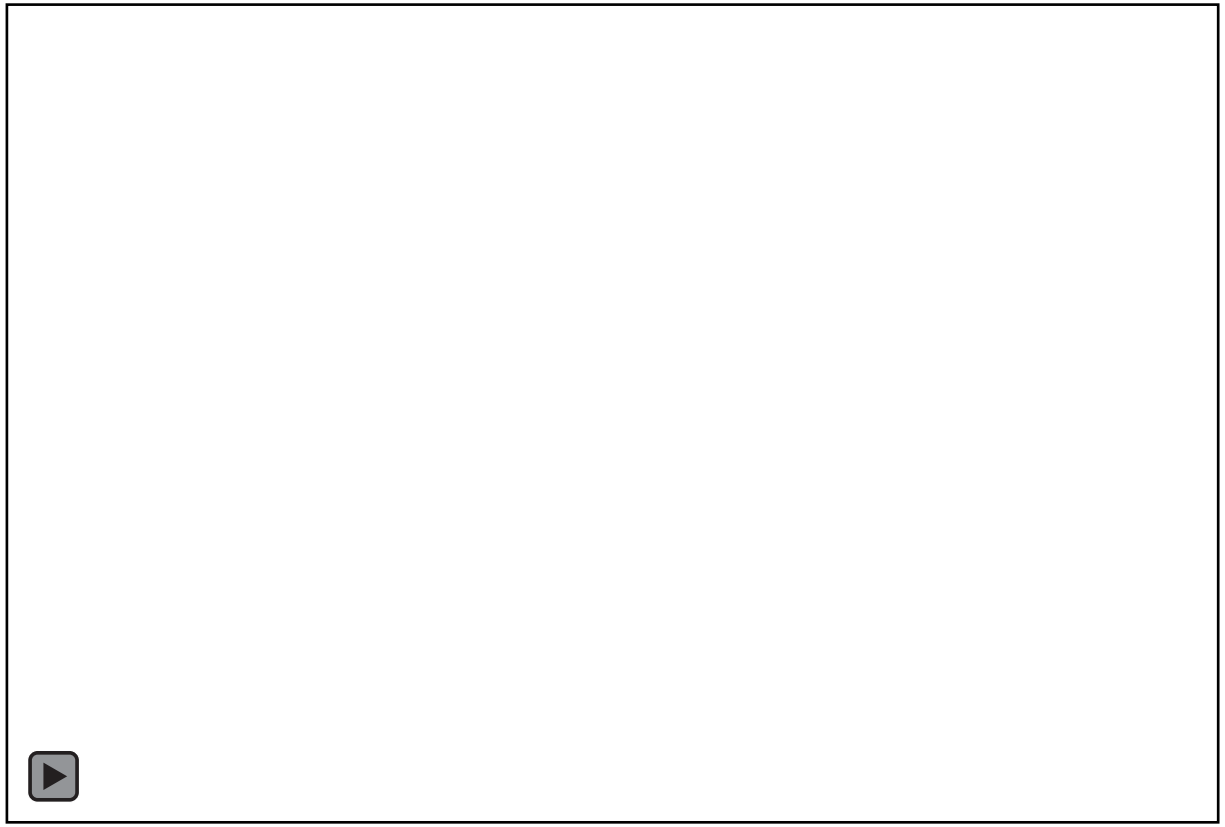
Frozen Time:

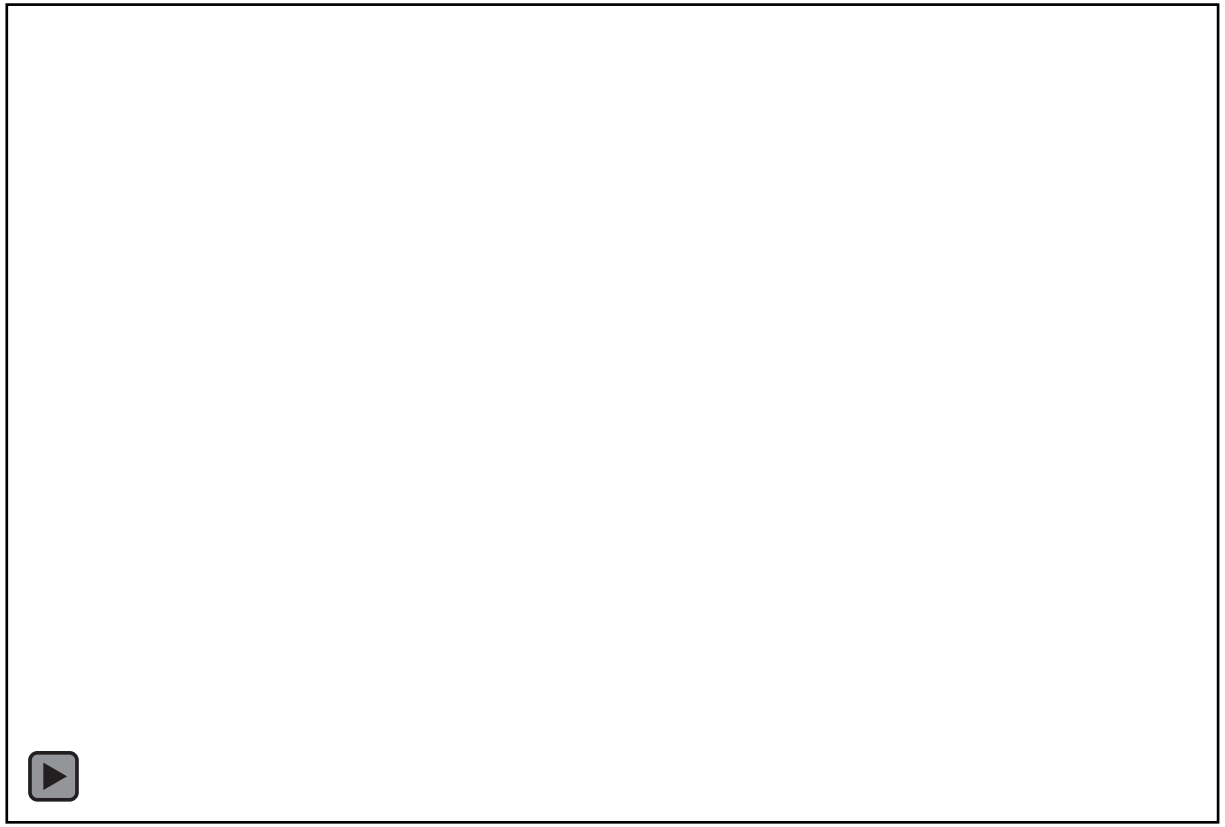


Fixed view:

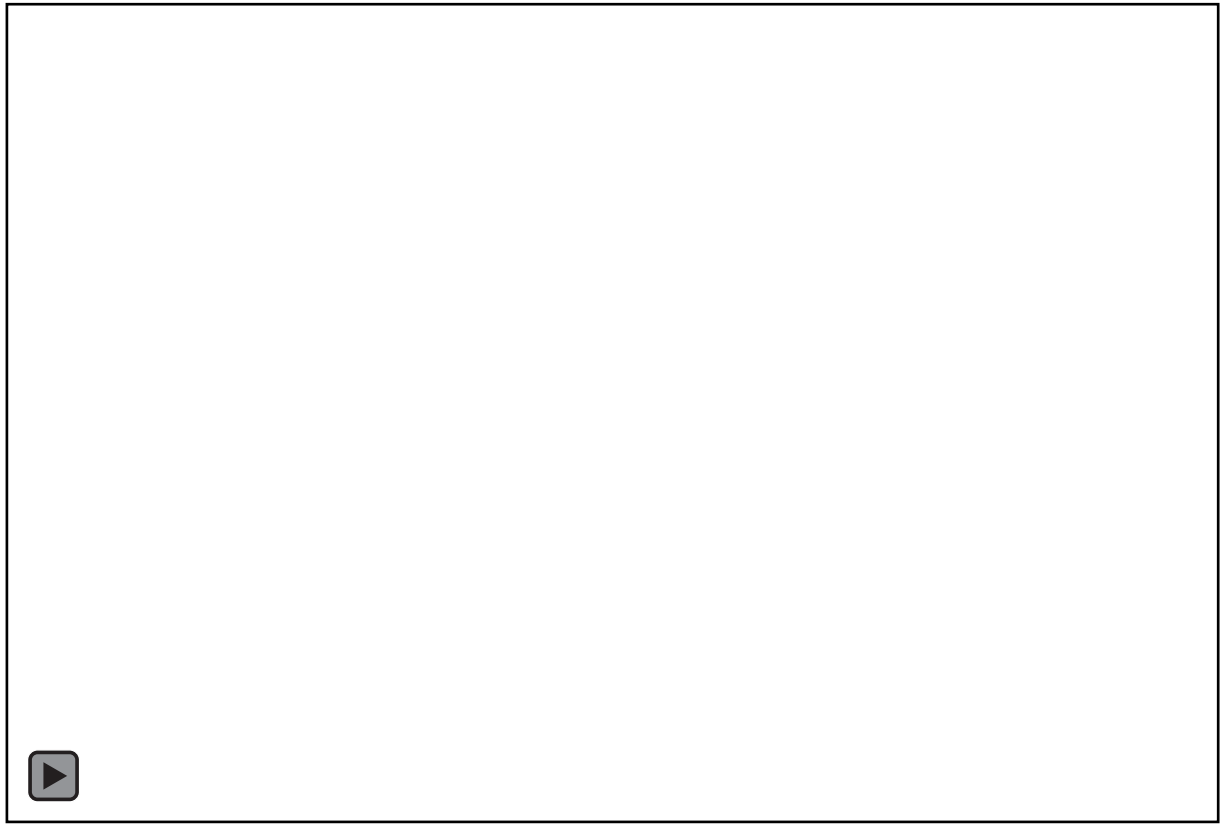


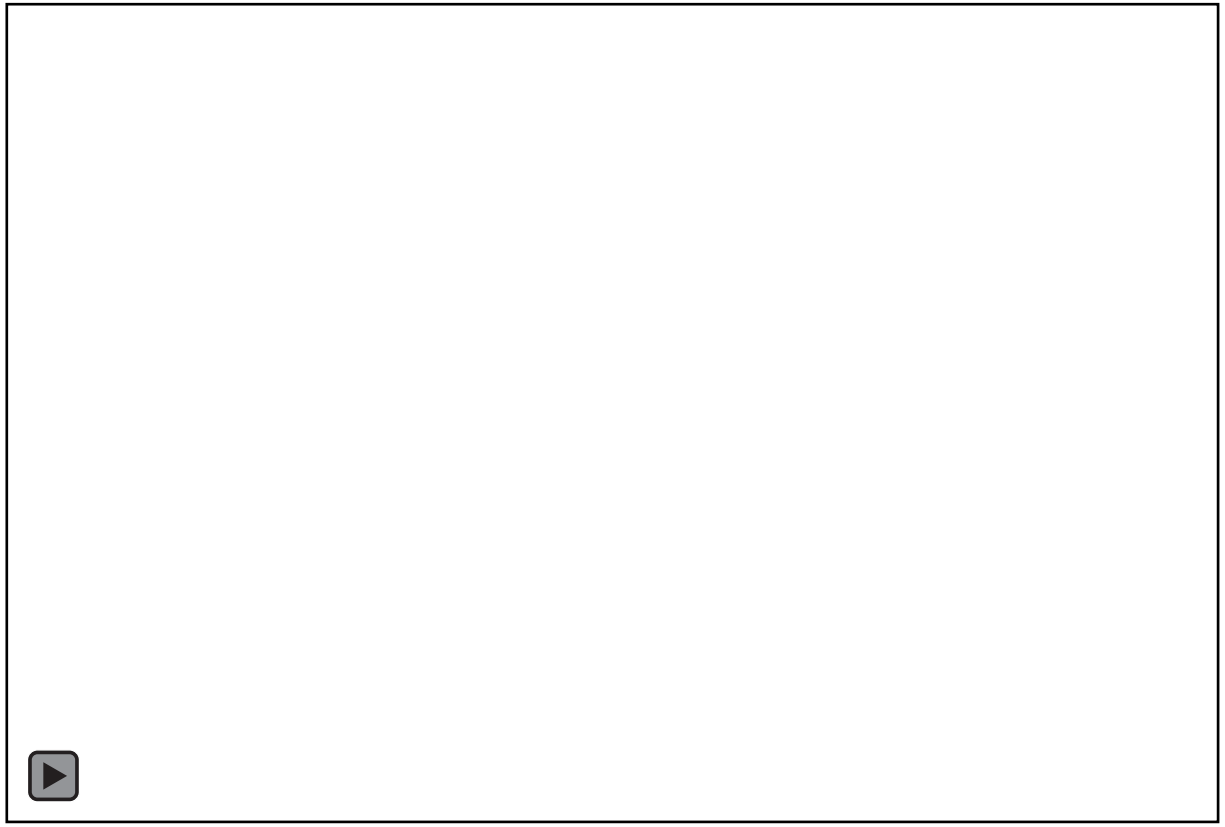
Interactive Renderings of Dynamic Gaussians





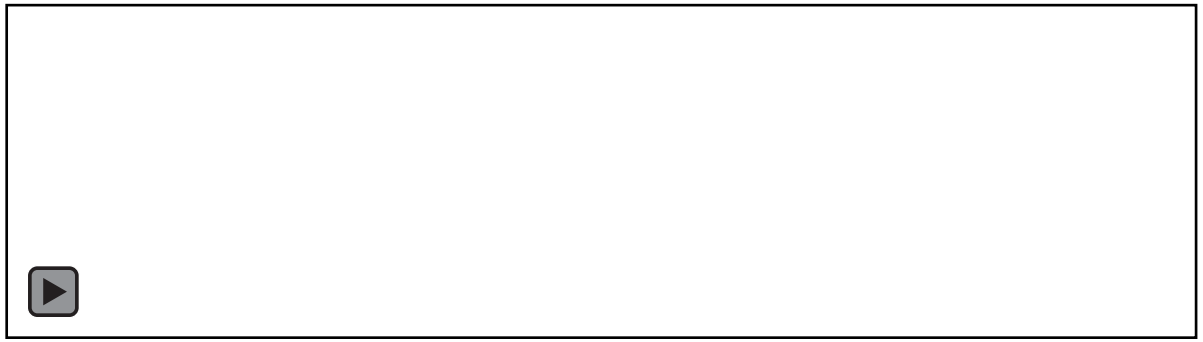
Interactive Renderings of Dynamic Gaussians



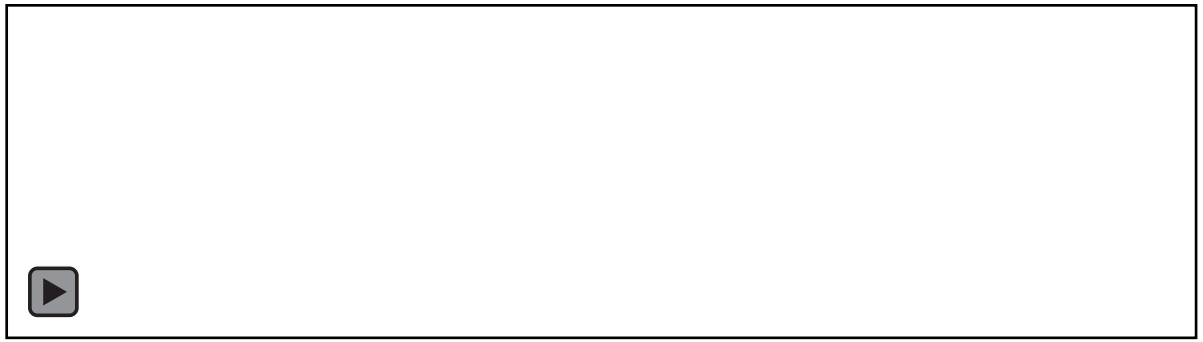


Comparison

Fixed view:

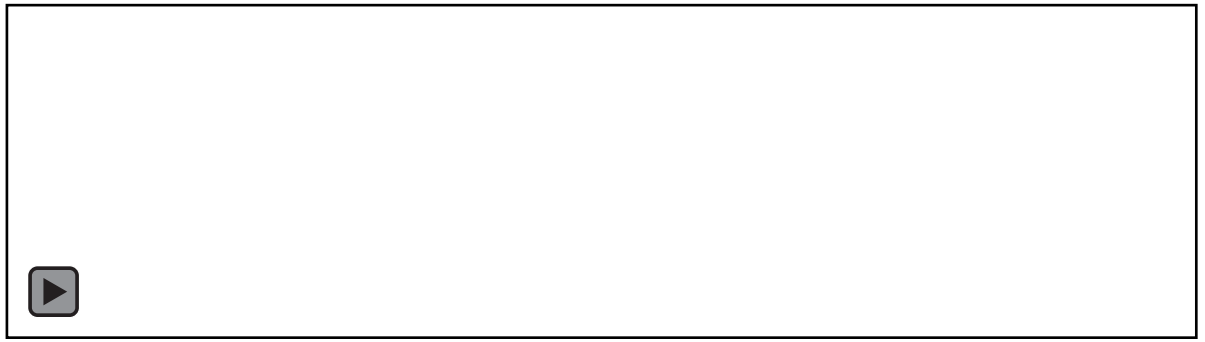


4Real-Video-V2

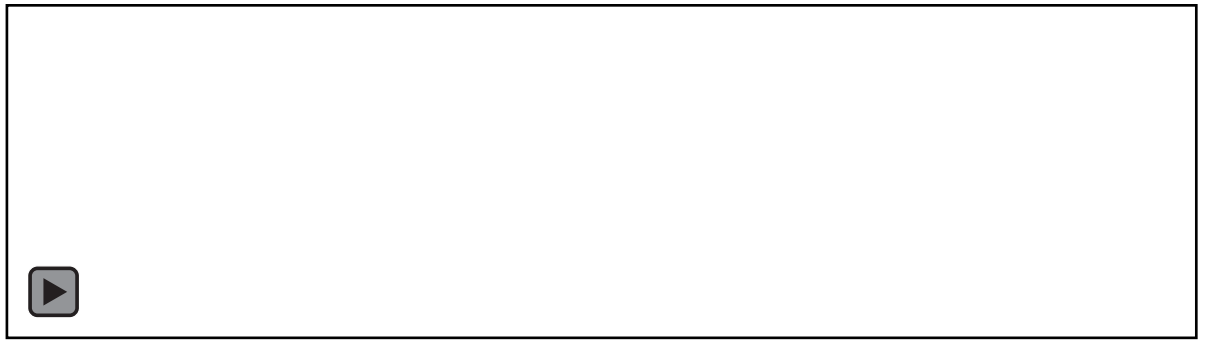


4Real-Video

Frozen Time:



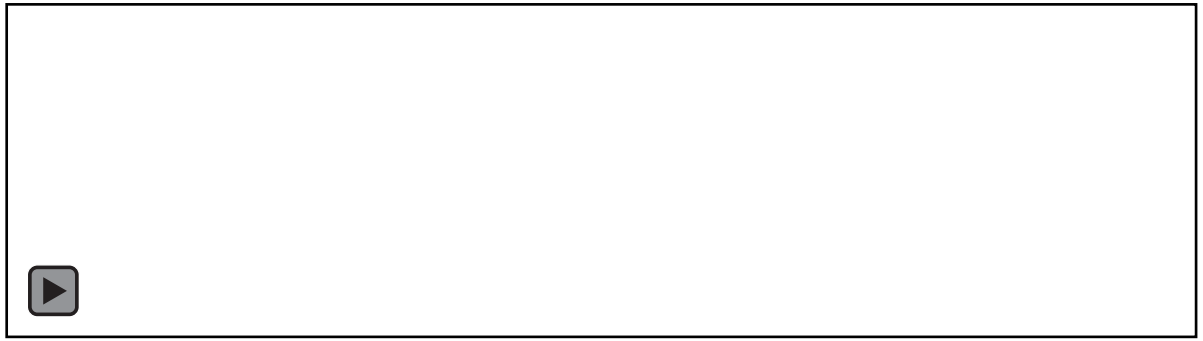
4Real-Video-V2



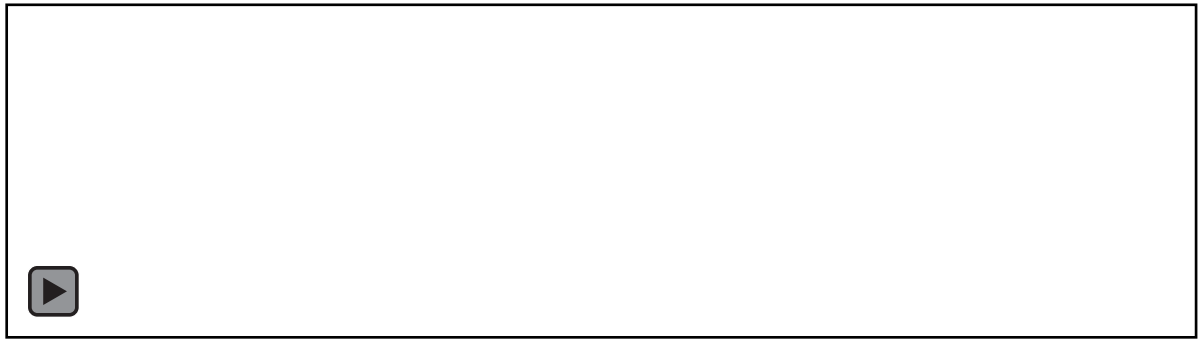
4Real-Video

Comparison

Fixed view:

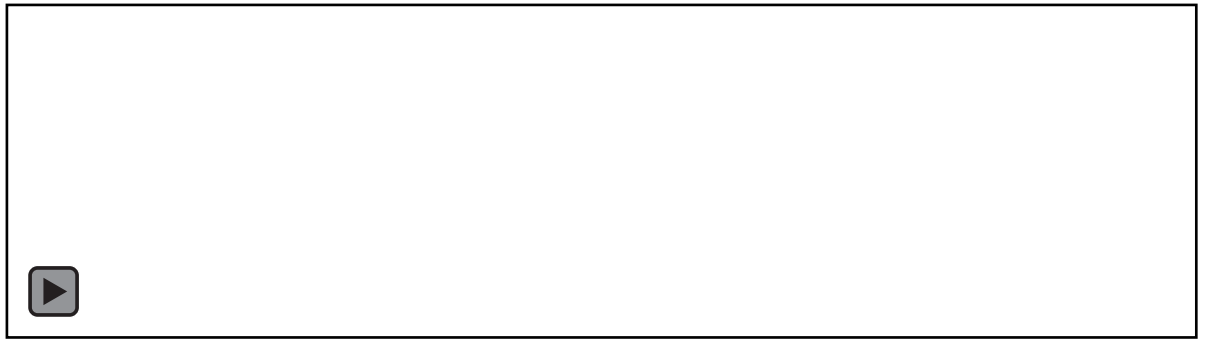


4Real-Video-V2

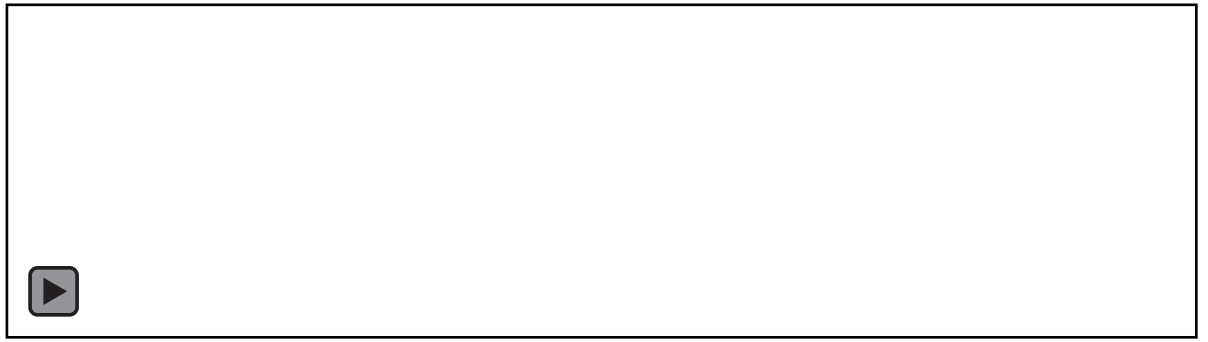


4Real-Video

Frozen Time:



4Real-Video-V2



4Real-Video

Limitations

- No full 360° coverage yet (partial sphere)
- Occasional layering artefacts in splats
- 4 min generation time

Thank you!

- Thank you for your attention!
- I appreciate your time and interest.
- If you have any questions, please feel free to ask.
- Contact information: alimohammadiamirhossein@gmail.com