Vid3D: Synthesis of Dynamic 3D Scenes using 2D Video Diffusion

We generate **3D videos** without explicitly enforcing **multiview consistency** over time

Rishab Parthasarathy [*1], Zachary Ankner [*12], Aaron Gokaslan [23]

[1] MIT, [2] Databricks Mosaic AI, [3] Cornell Tech, [*] equal contribution

We Study Dynamic 3D Scene Generation

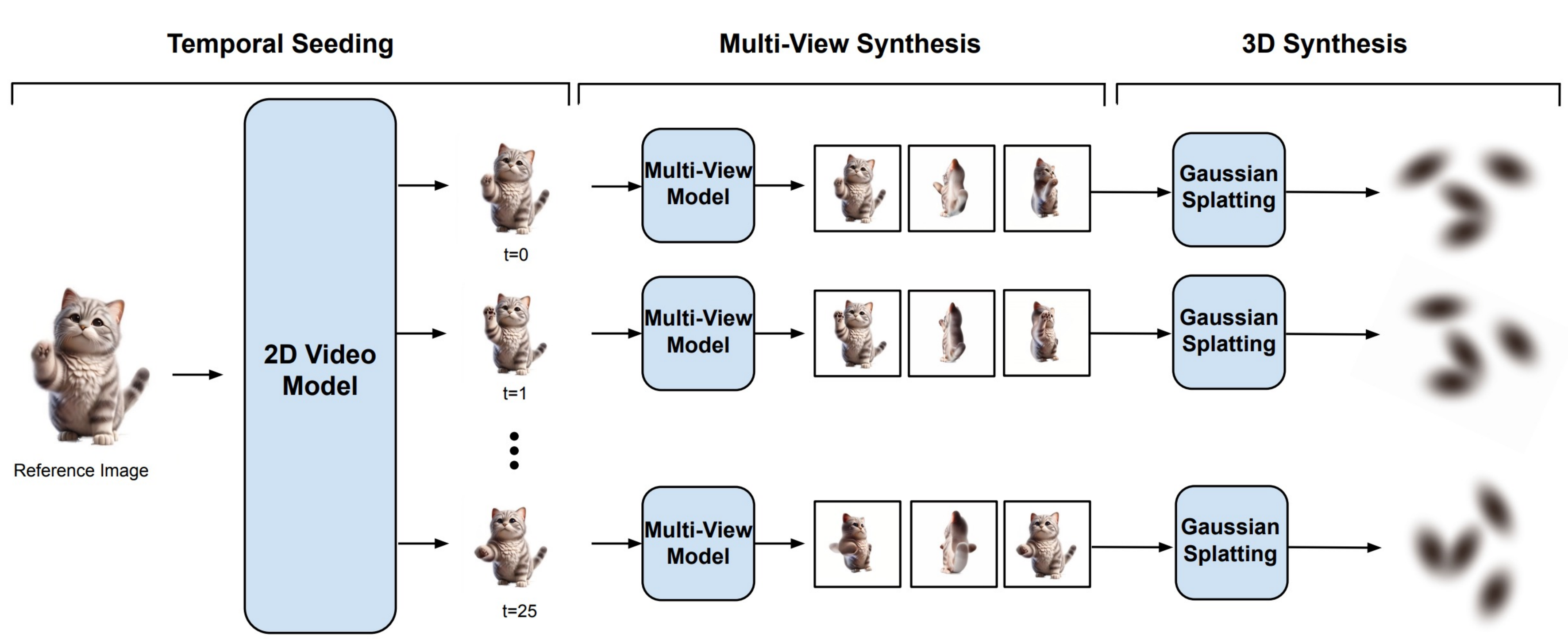
Recent work has modeled 3D temporal dynamics by jointly optimizing for consistency across both time and space. However, new work in 2D video models has suggested that strong temporal priors may be sufficient to learn correlation in space.

Research Questions

1. Can we generate 3D videos without 3D temporal priors?
2. What is the impact on performance of this change?
3. What is the impact of hyperparameters on 3D video quality?

Our Approach to 3D Video Generation

To create a model capable of generating 3D videos without modeling 3D temporal dynamics, we factorize the task into generating the 2D temporal dynamics of the scene **(temporal seeding)** and then generating 3D representations **(multi-view & 3D synthesis)** of each timestep in the 2D scene.



Methodology

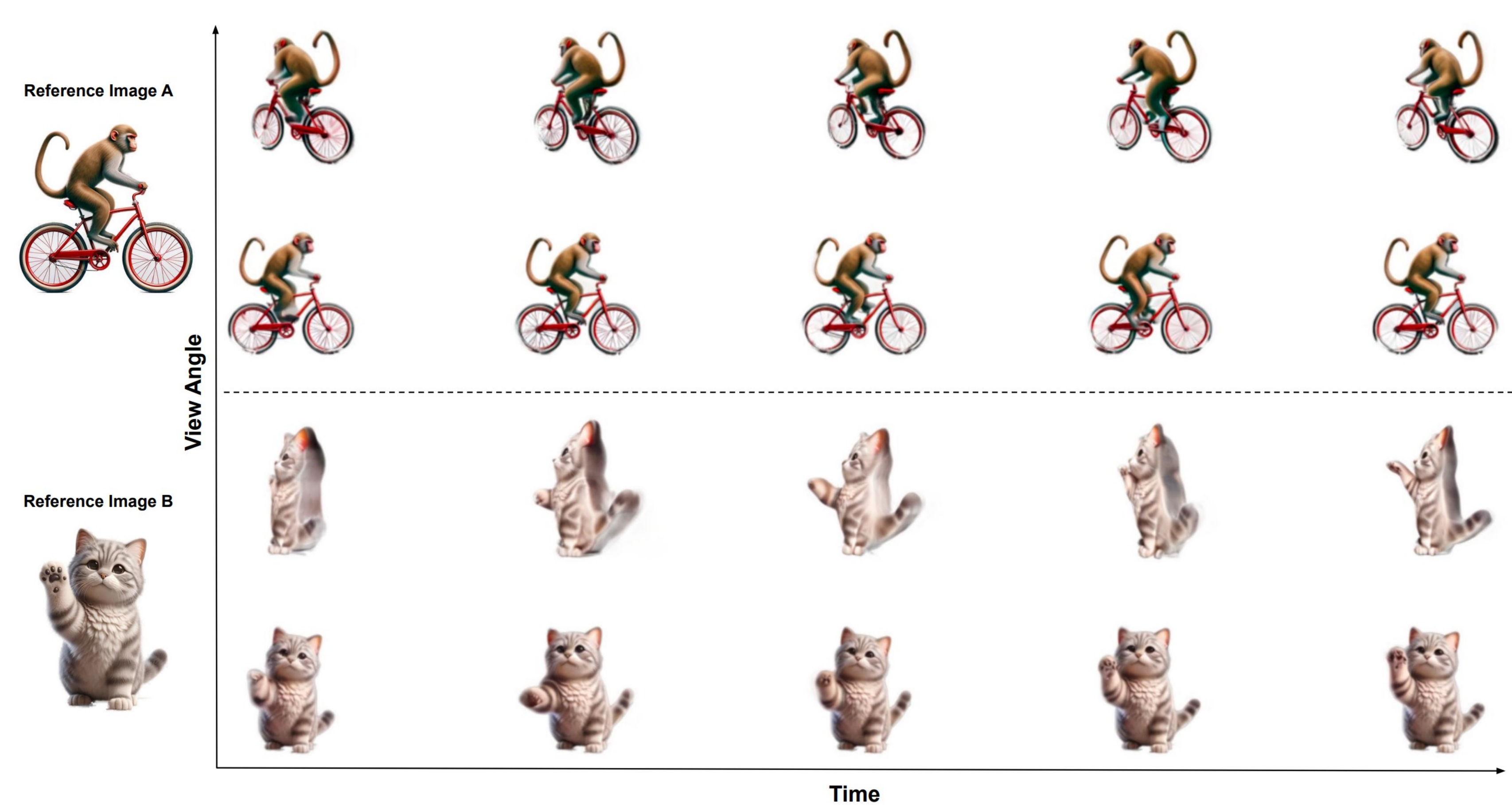
- We use Stable Video Diffusion for both **temporal seeding and multi-view synthesis** (finetuned on Objaverse). We use Gaussian Splatting for **3D synthesis**.
- We evaluate on the benchmark provided by Animate124.
- We evaluate using the CLIP-I score, which is defined as the average cosine similarity between the CLIP-features of the reference image and each frame in each 2D video rendering.

Evaluation of 3D Video Quality

Table 1. CLIP-I score for Vid3D compared to Animate124 and DreamGaussian4D, showing that our model does not need 3D temporal dynamics to yield competitive results.

Model	CLIP-I
Animate124	0.8544
DreamGaussian4D	0.9227
Vid3D (Ours)	0.8946

Sample Generations

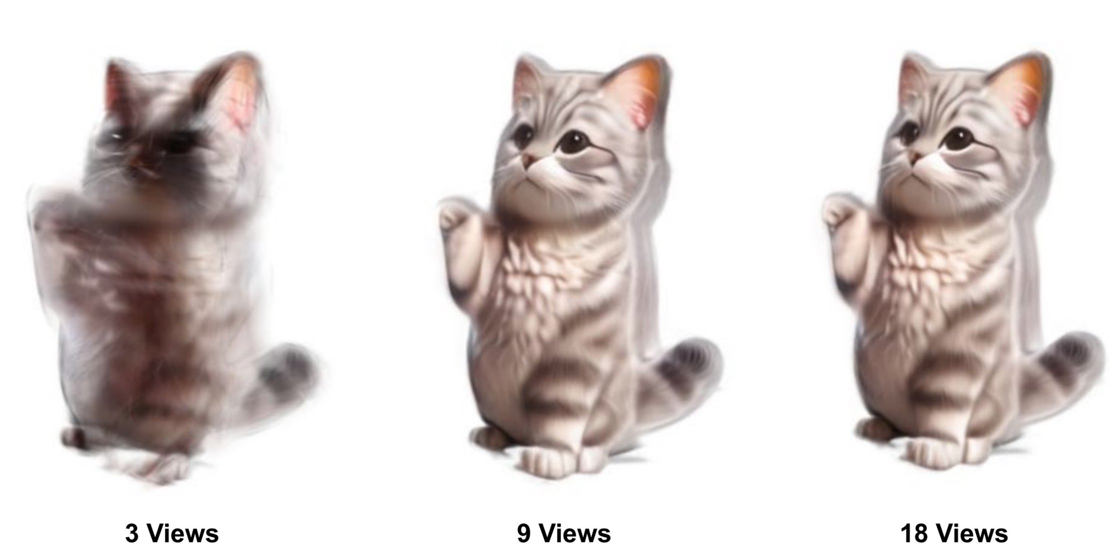


Evaluation of Varying Hyperparameters

We first evaluate reducing the **number of views generated.**

Table 2. CLIP-I score values for Vid3D for different numbers of views. This result shows that reducing the number of views from 18 to 9 does not significantly degrade performance, while further reduction does.

Number of views	CLIP-I
3	0.8532
9	0.8879
18 (Baseline)	0.8946



We then evaluate **modifying the motion score**.

Table 3. CLIP-I score values for Vid3D for different temporal seed motion scores. This result shows that there is a slight loss in quality for scenes with more motion.

Motion Score	CLIP-I
120 (Baseline)	0.8946
160	0.8893
200	0.8897