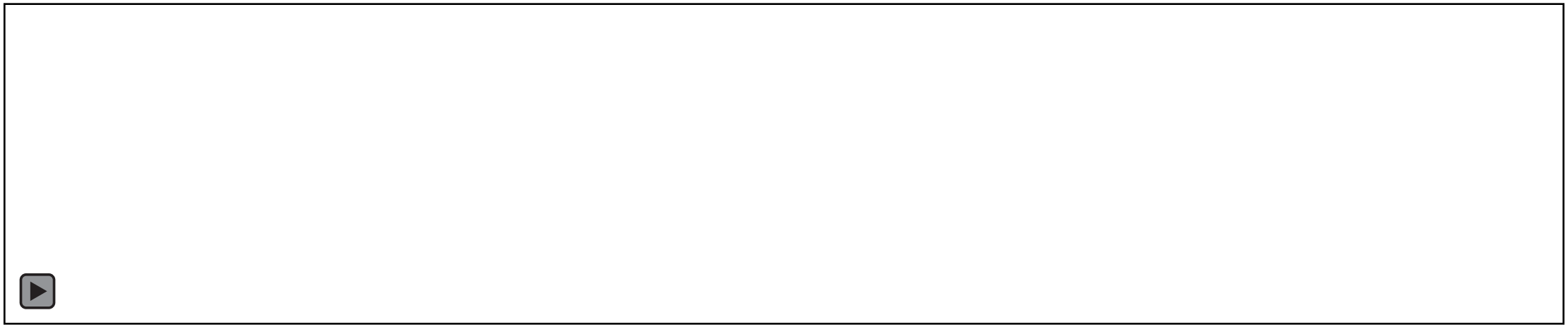
Force Prompting:
Video Generation Models Can Learn and Generalize Physics-based Control Signals

Nate Gillman[1], Charles Herrmann[2*], Michael Freeman[1], Daksh Aggarwal[1], Evan Luo[1], Deqing Sun[2], Chen Sun[1*]

[1]Brown University, [2]Google DeepMind

Train with Limited Synthetic Data

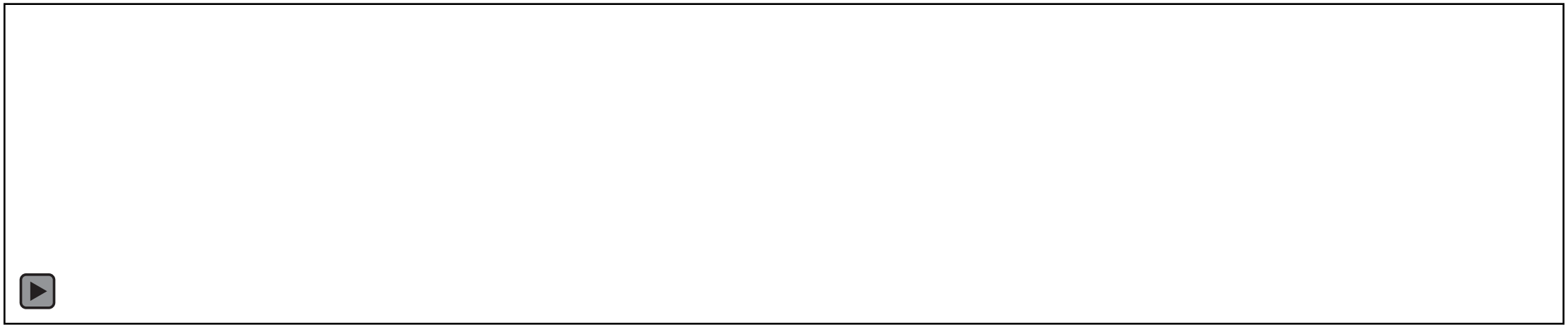
Train with Limited Synthetic Data



Local Force Model (Poke)

Train with Limited Synthetic Data

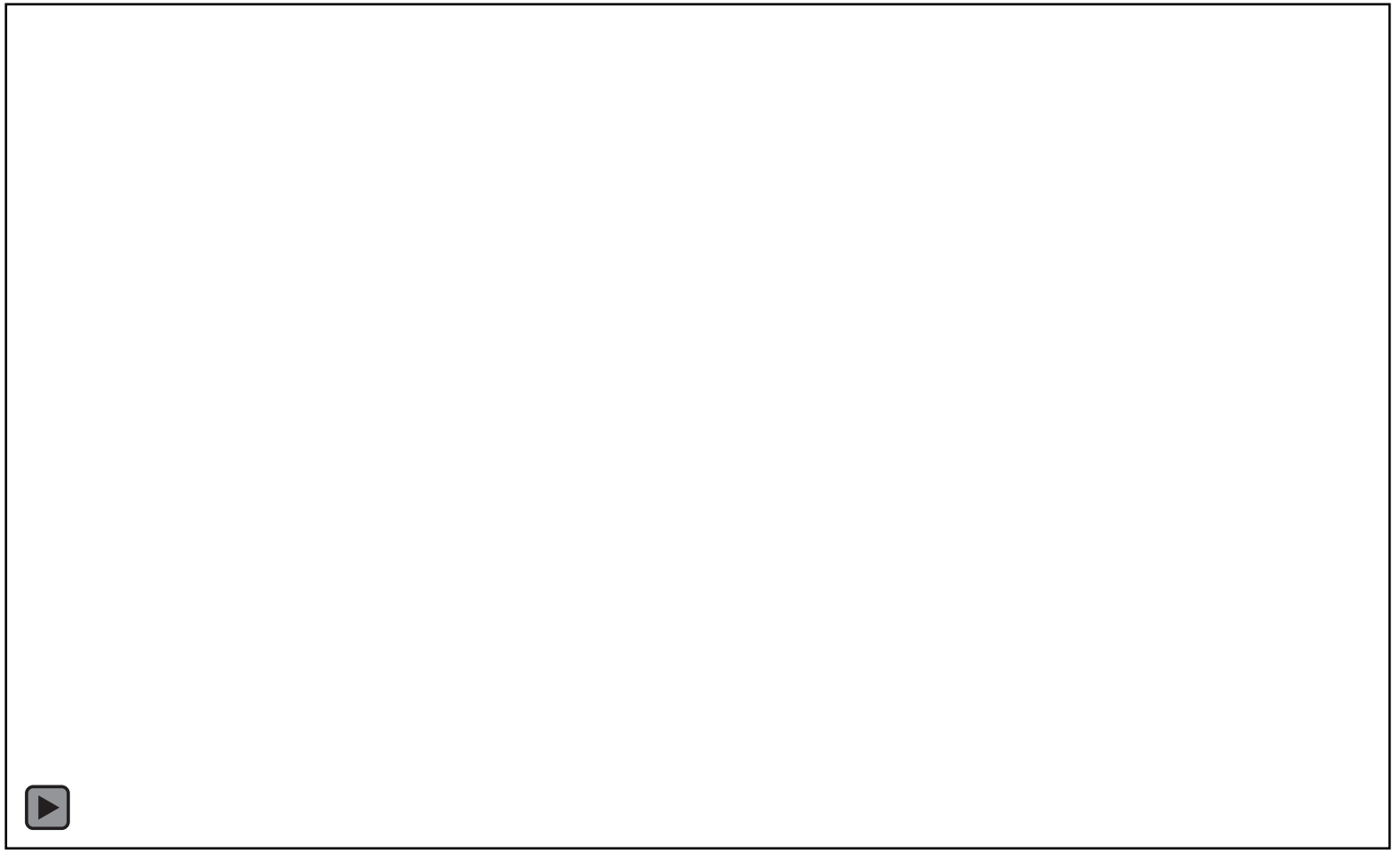
Train with Limited Synthetic Data



Global Force Model (Wind)

Video Model Generalizes Force Conditioning

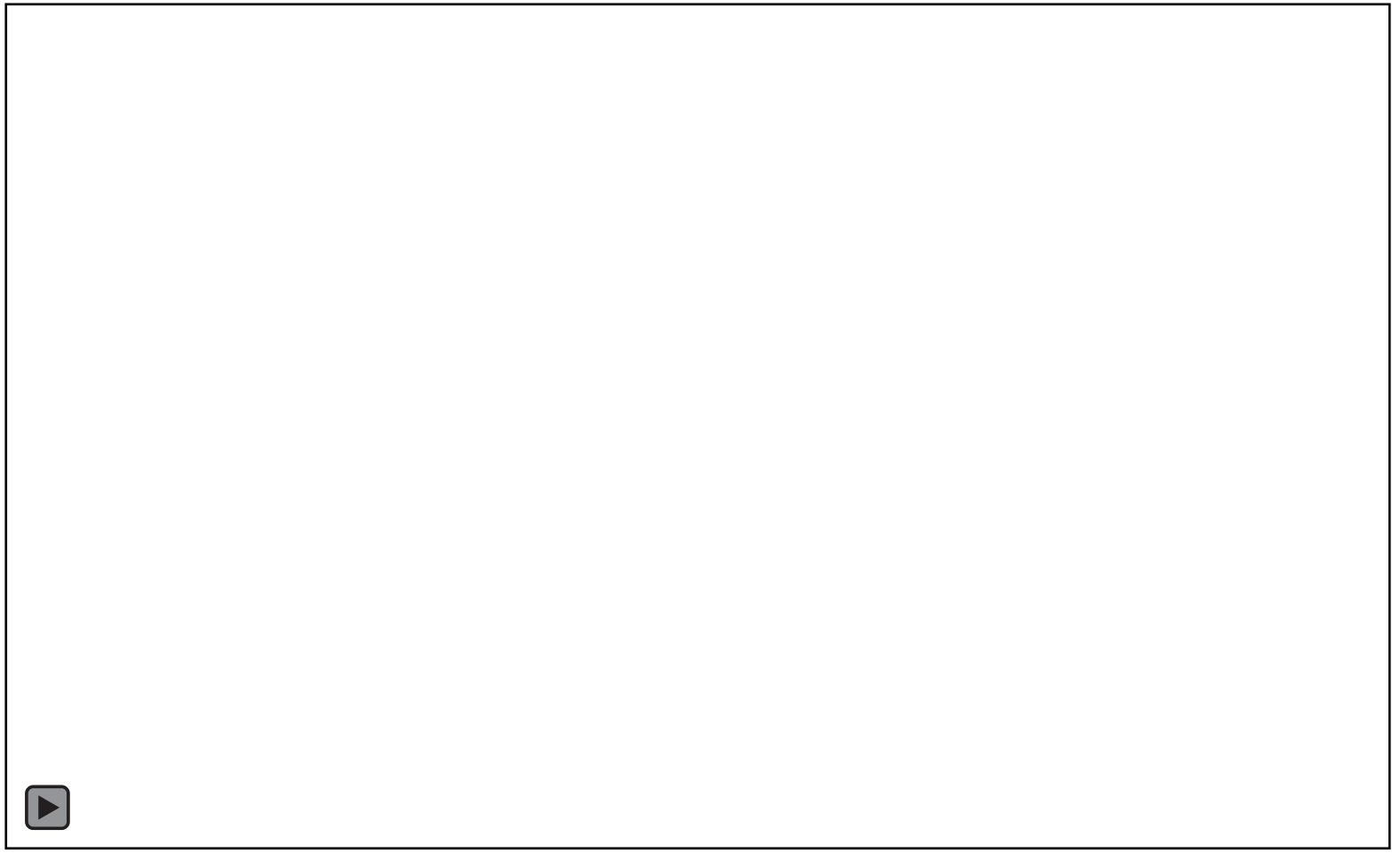
Video Model Generalizes Force Conditioning



Generalizes to Different Objects and Geometries

Video Model Generalizes Force Conditioning

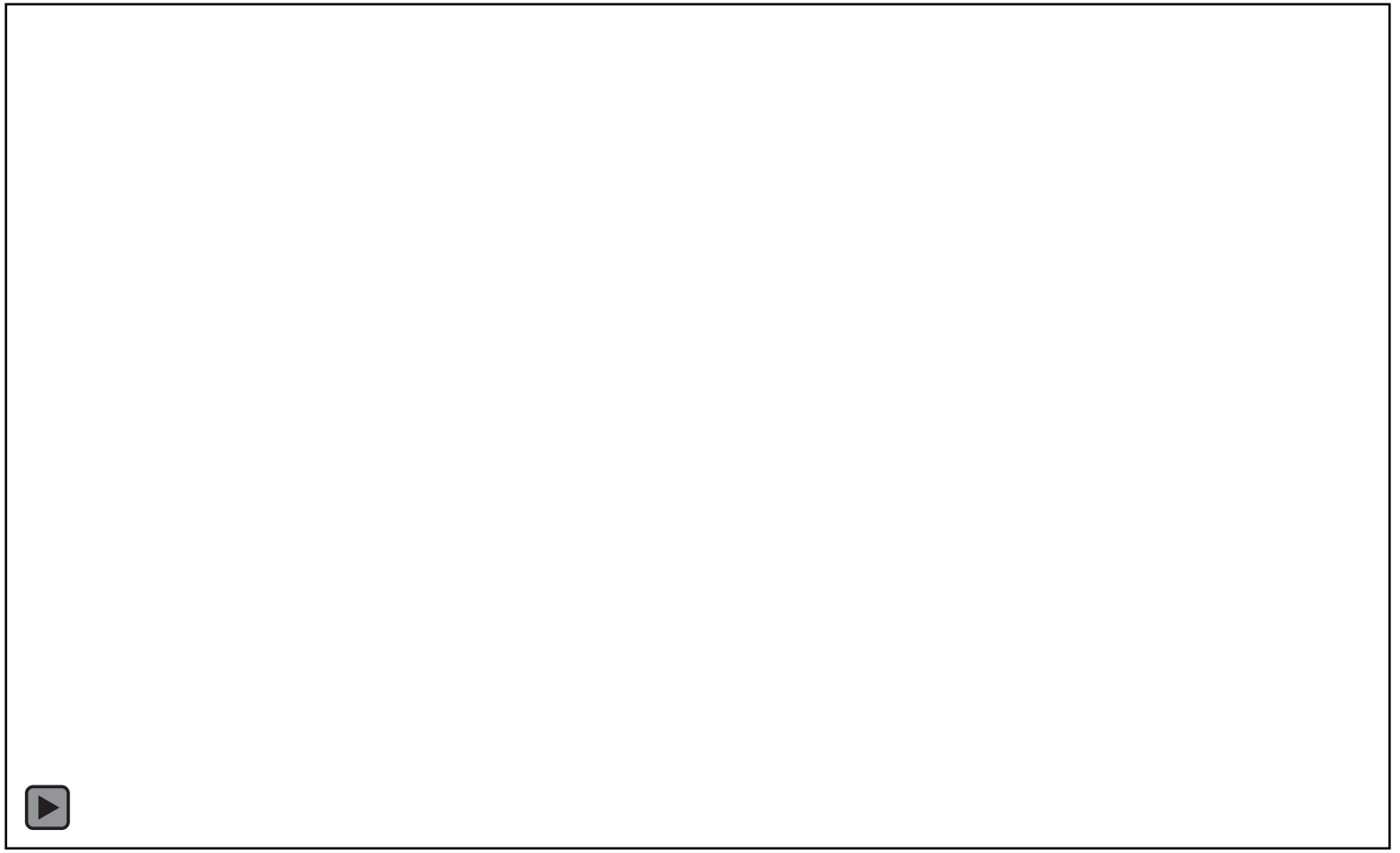
Video Model Generalizes Force Conditioning



Generalizes to Different Settings and Materials

Video Model Generalizes Force Conditioning

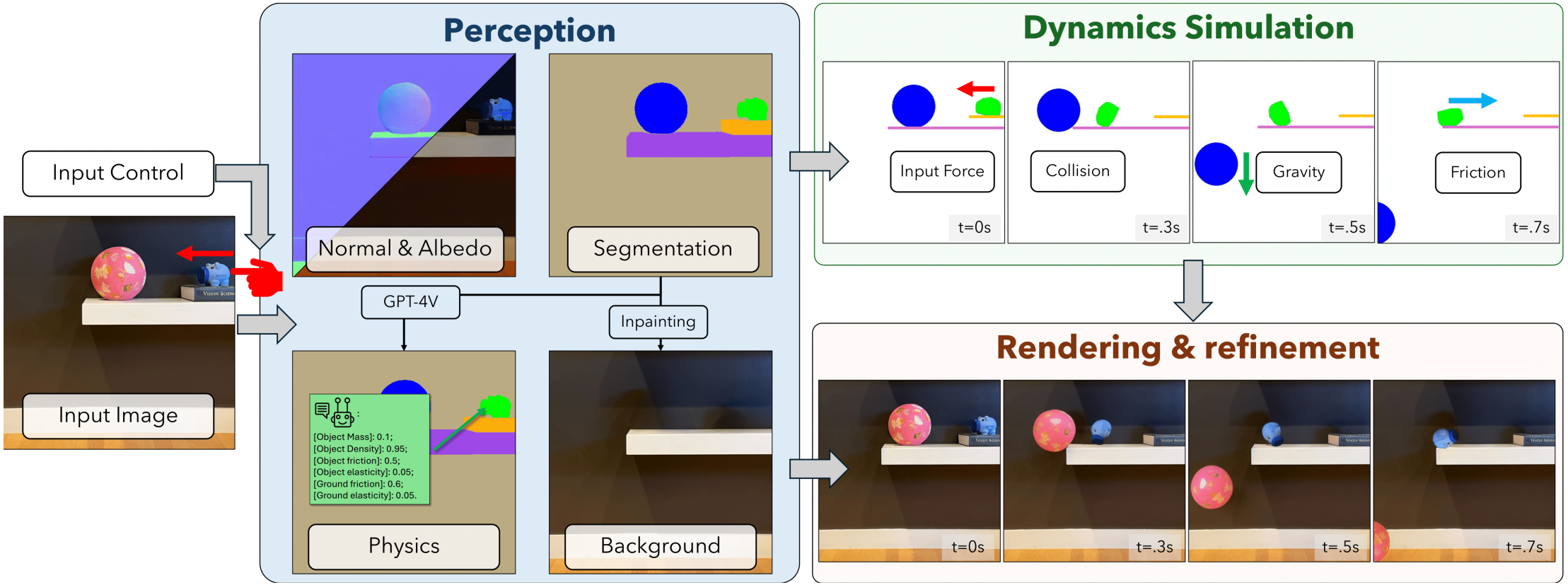
Video Model Generalizes Force Conditioning



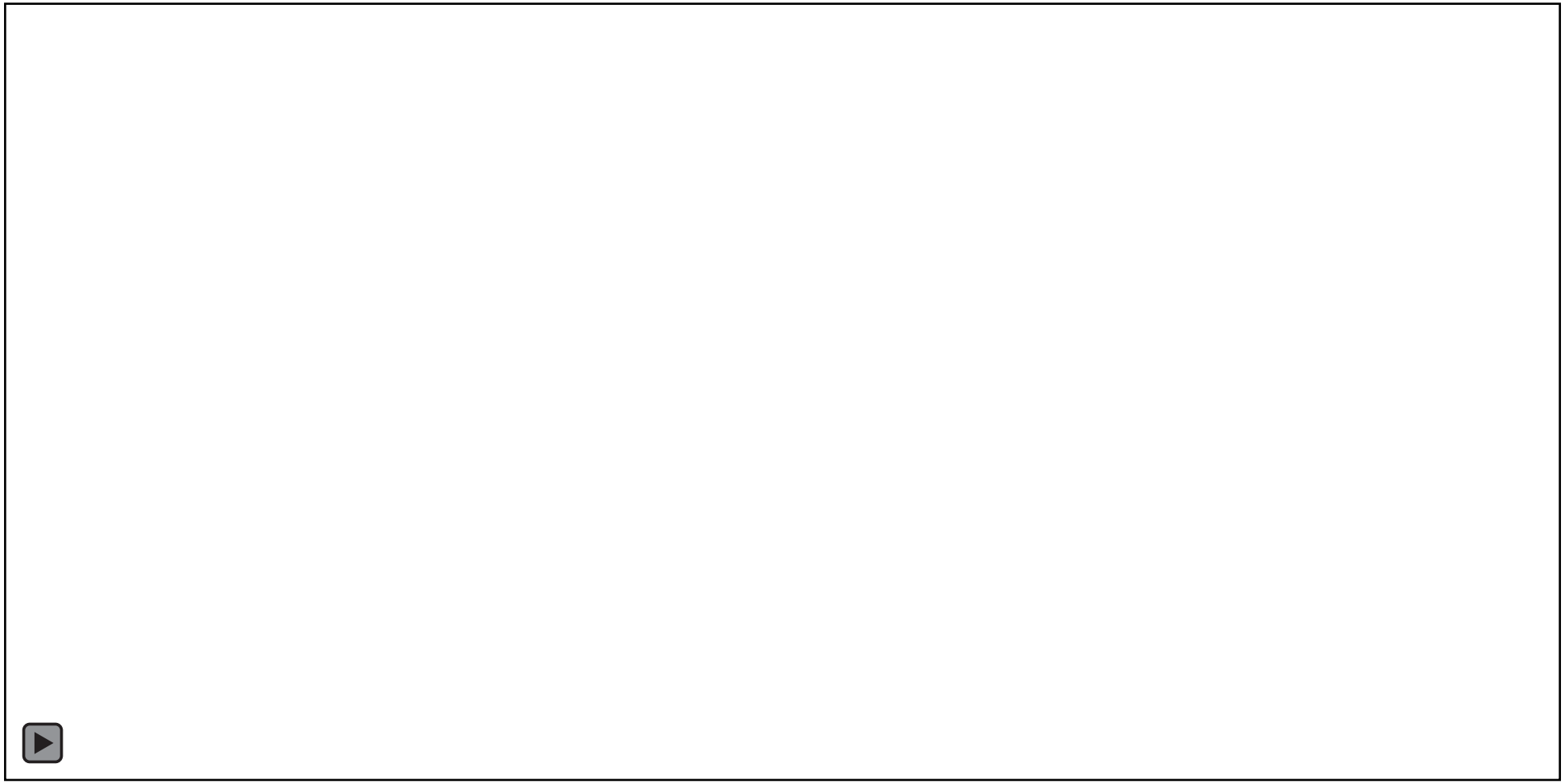
Generalizes to Different Affordances

Related Works

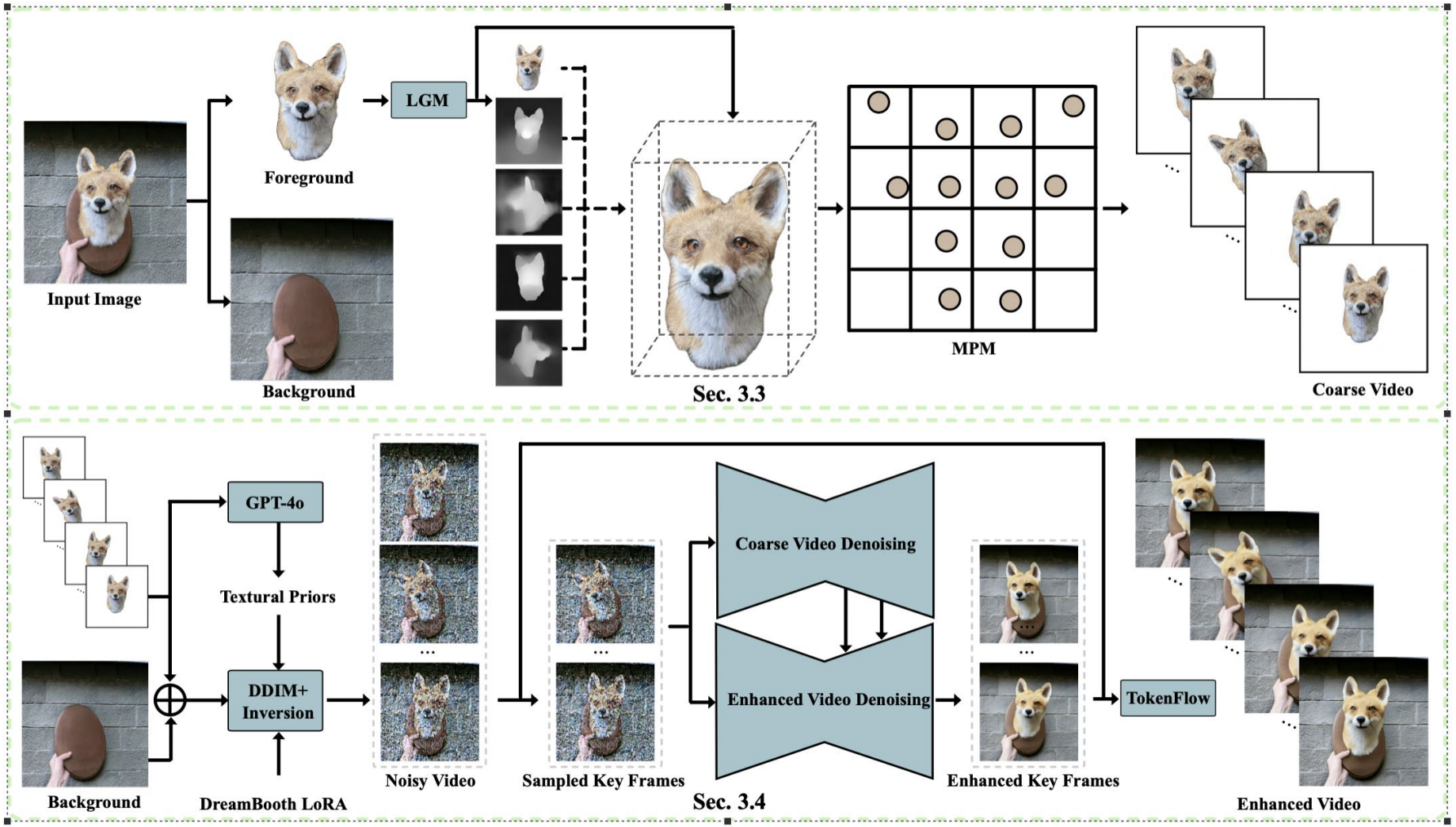
PhysGen


Input Control
Input Image
Perception
Normal & Albedo
Segmentation
GPT-4V
Inpainting
[Object Mass]: 0.1;
[Object Density]: 0.95;
[Object friction]: 0.5;
[Object elasticity]: 0.05;
[Ground friction]: 0.6;
[Ground elasticity]: 0.05.
Physics
Background
Dynamics Simulation
Input Force
t=0s
Collision
t=.3s
Gravity
t=.5s
Friction
t=.7s
Rendering & refinement
t=0s
t=.3s
t=.5s
t=.7s

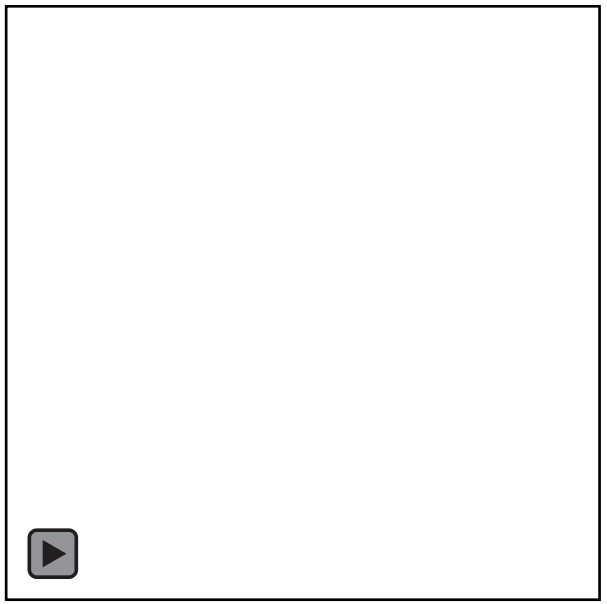
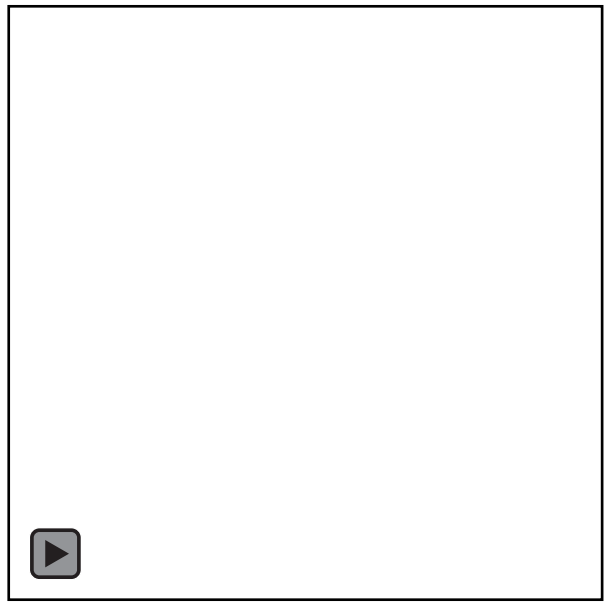
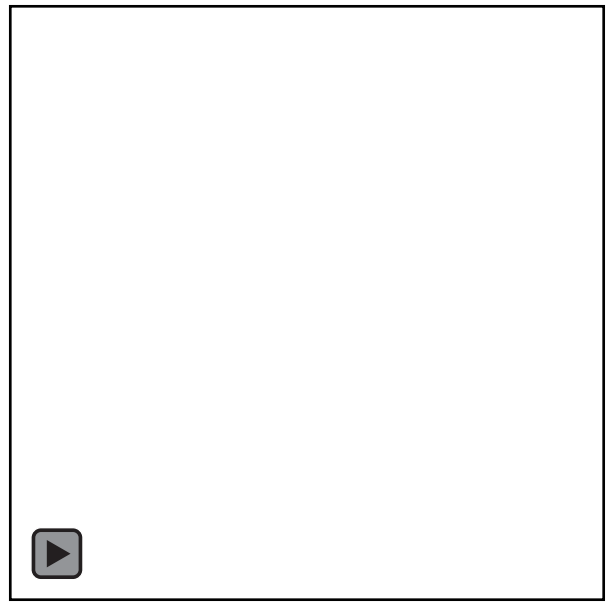
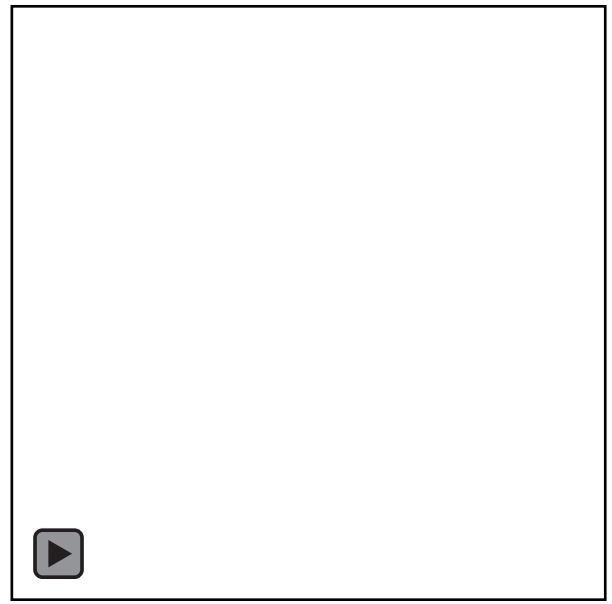
PhysGen



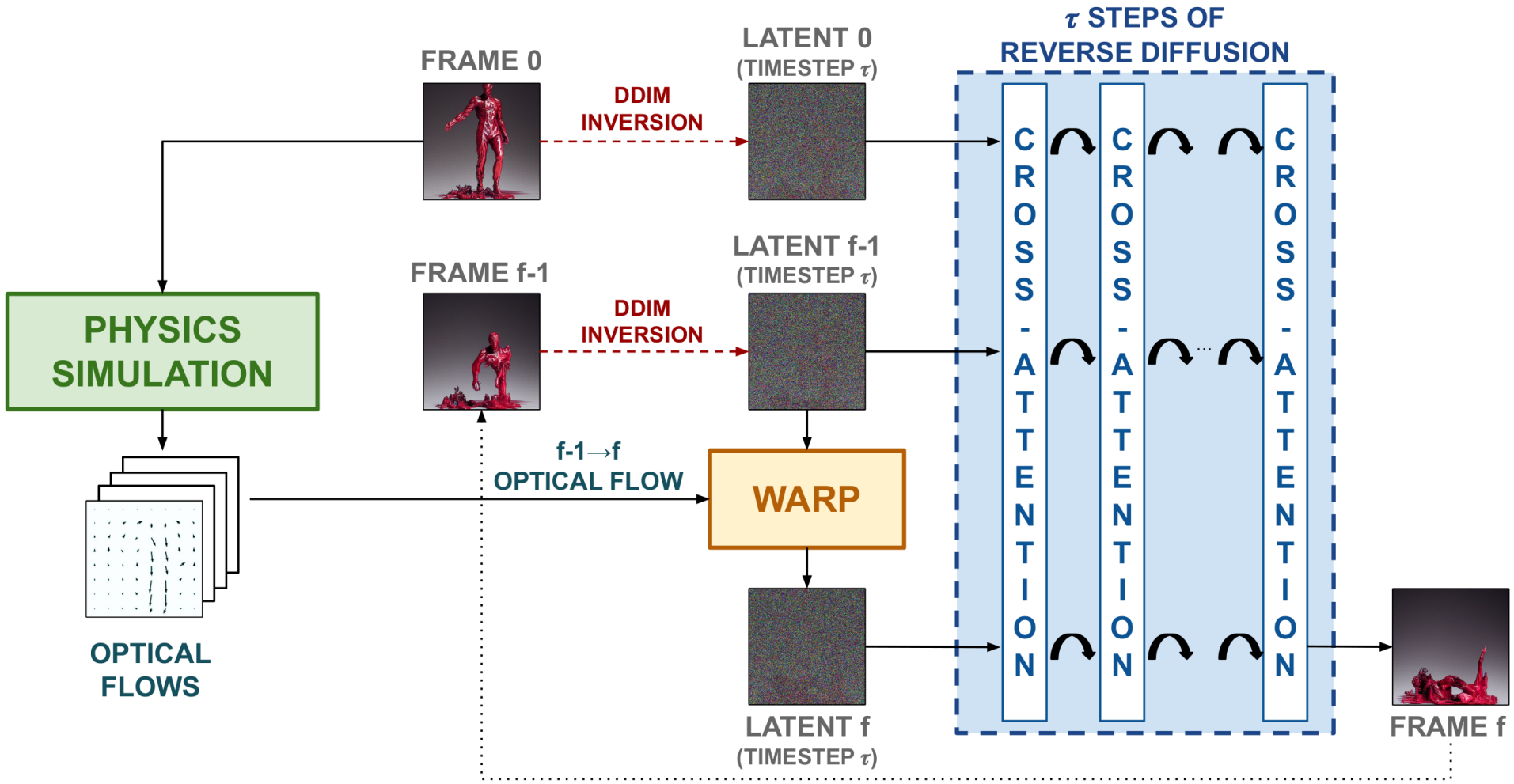
PhysMotion


Input Image
Foreground
Background
LGM
Sec. 3.3
MPM
Coarse Video
GPT-4o
Textural Priors
DDIM+ Inversion
DreamBooth LoRA
Noisy Video
Sampled Key Frames
Coarse Video Denoising
Enhanced Video Denoising
Enhanced Key Frames
Sec. 3.4
TokenFlow
Enhanced Video

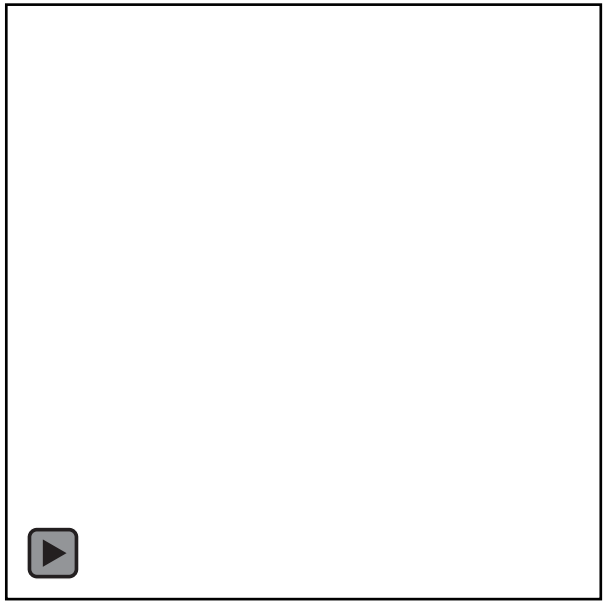
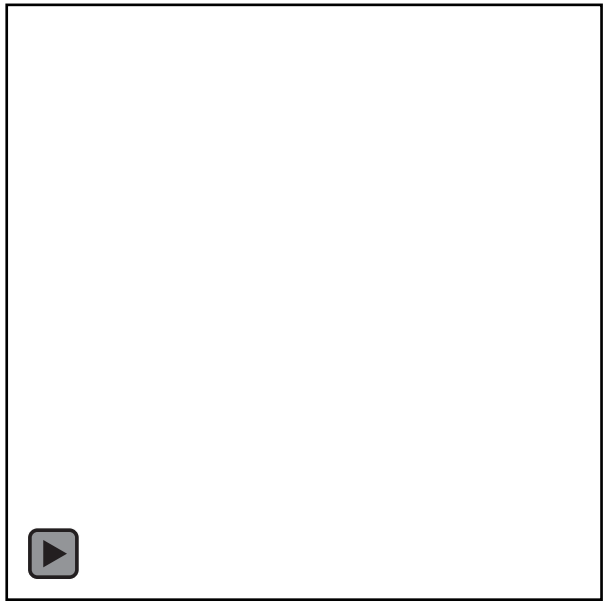
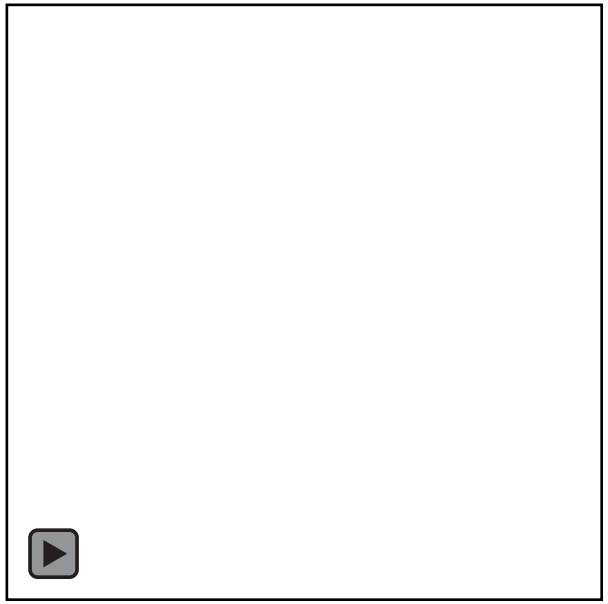
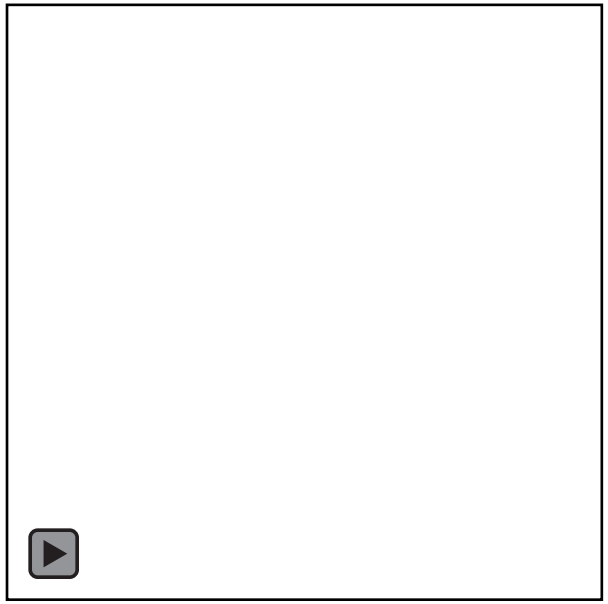
PhysMotion

MotionCraft



MotionCraft

Method

Synthetic Training Data (Global Force)

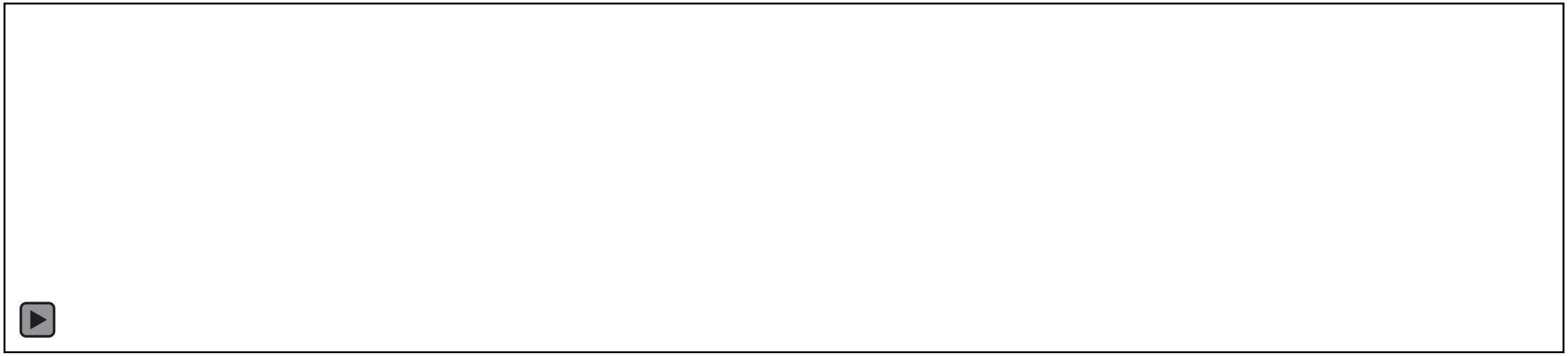
Built using Blender to simulate flags reacting to wind.

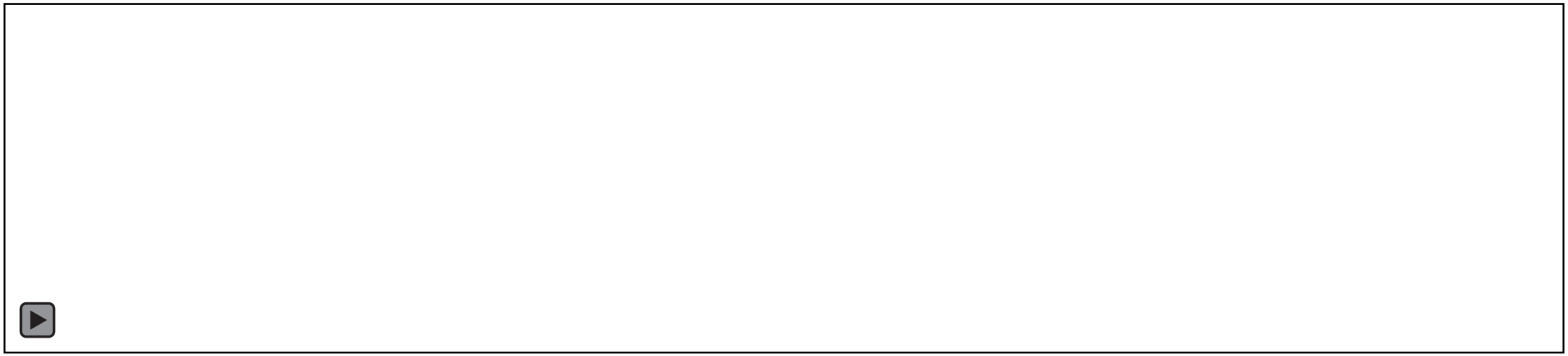
15,000 videos total (3D simulation of wind-affected flag motion)

Randomized parameters include:

- Flag quantity: Uniform between 1 and 64
- Flag colors and positions
- Camera placement
- Wind direction: Uniform in [0°, 360°)
- Wind speed: Float in [0, 1] (0 = no wind, 1 = strong wind))

Synthetic Training Data (Global Force)





Synthetic Training Data (Local Force)

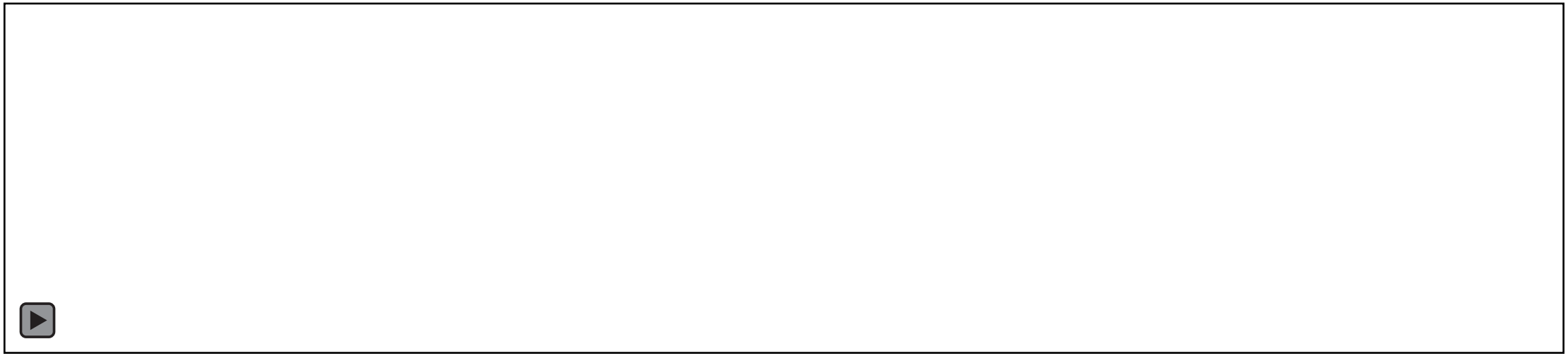
Scenario 1: Ball Pushing (12k Videos)

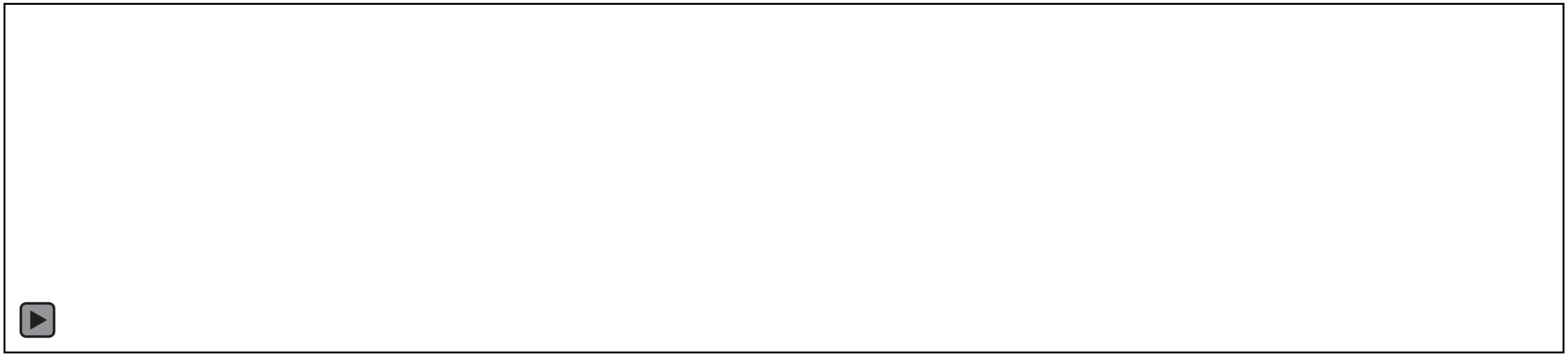
One ball rolls from a hidden force while the others stay still (in Blender)

Randomized parameters include:

- Ball quantity: Uniform {2, 3, 4}
- Ball type: Soccer ball (2/3) or heavier bowling ball (1/3, 4× mass)
- Ball color (from 108), positions, and camera placement
- Ground textures: 42 Polyhaven HDRIs
- Force angle: [0°, 360°), force magnitude: [0, 1]

Synthetic Training Data (Ball Pushing)





Synthetic Training Data (Local Force)

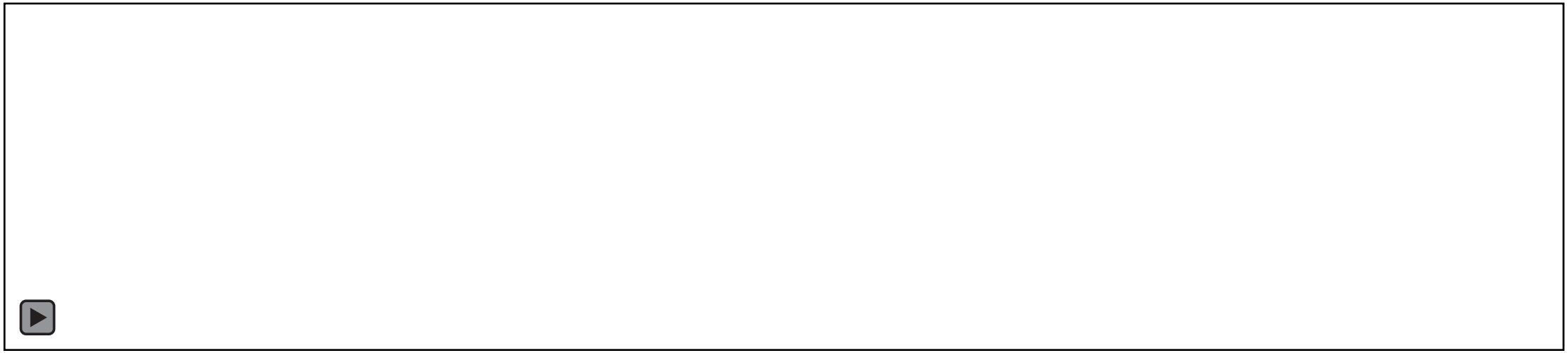
Scenario 2: Uses PhysDreamer (11k Videos)

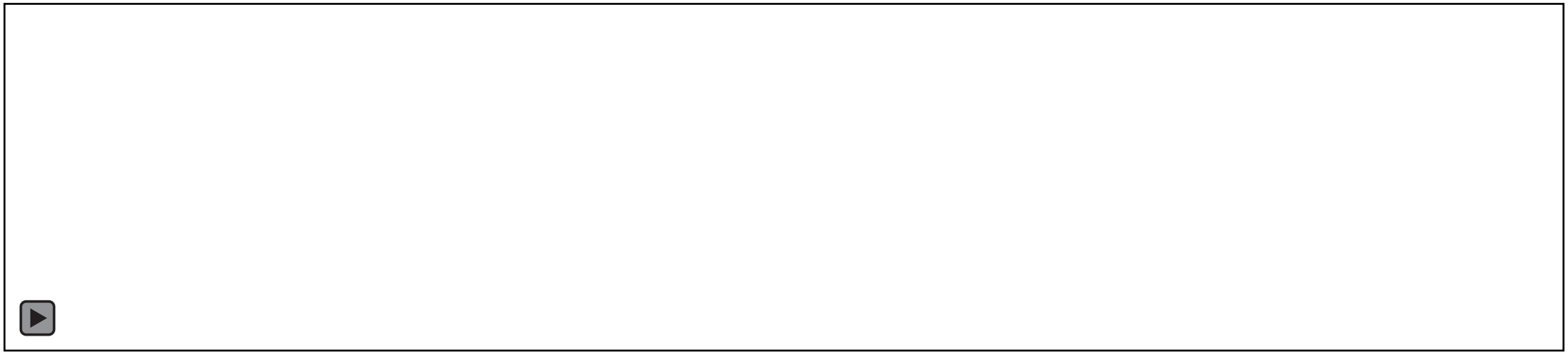
A carnation sways after being poked by an unseen point-force

Randomized parameters include:

- Camera view
- Contact points
- Force angle/magnitude

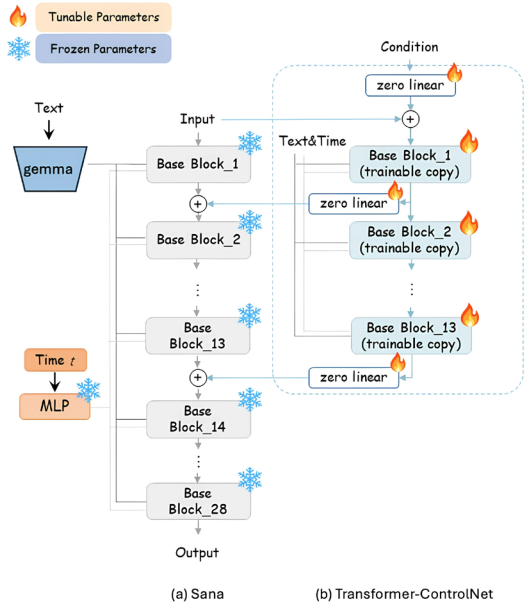
Synthetic Training Data (PhysDreamer)

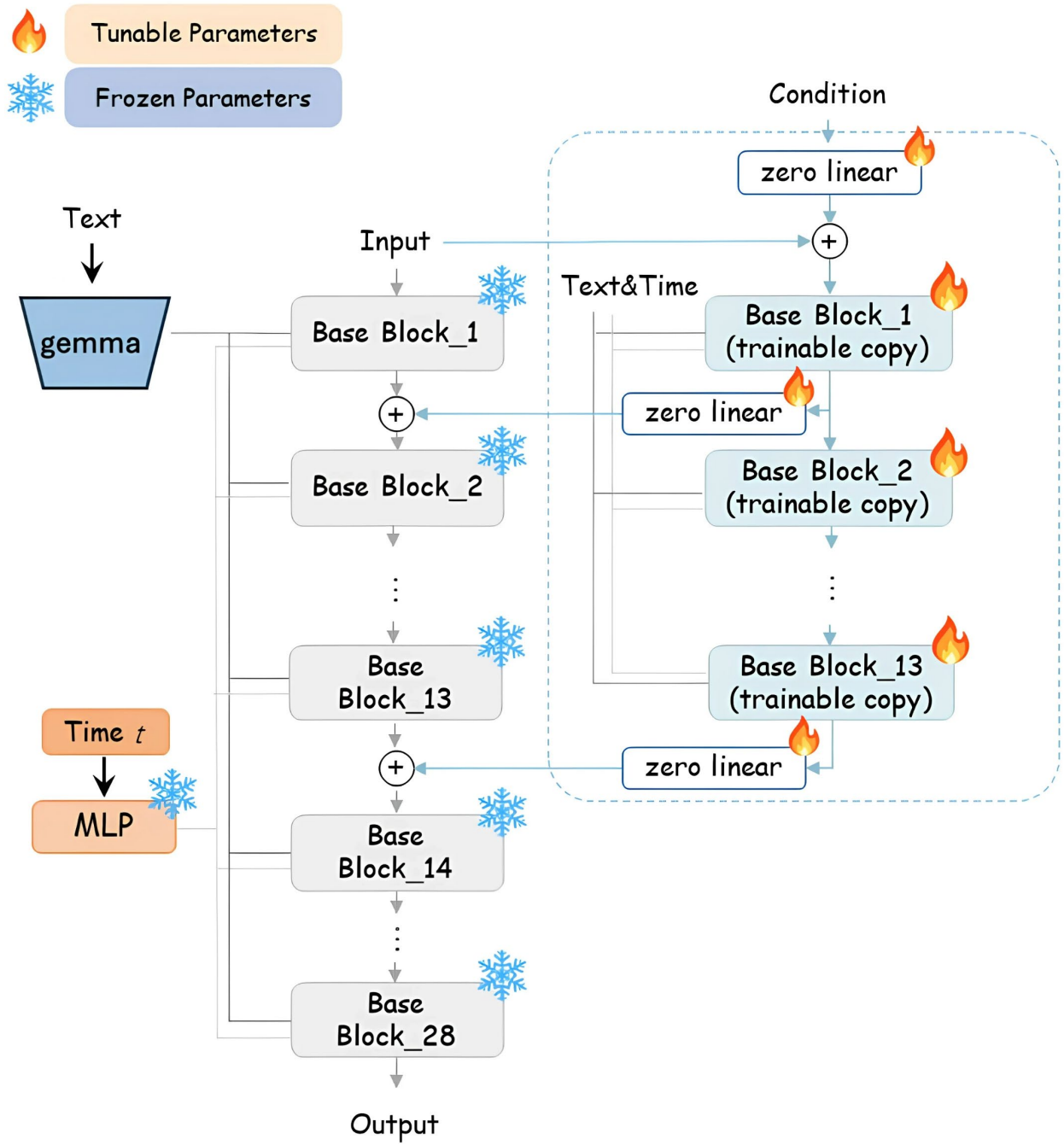




Architecture & Training

- Based on CogVideoX-5B-I2V: 49-frame videos at 8 fps using text + initial frame
- Added ControlNet to inject physics prompts (π) via downscaling & zero conv
- The first 6 transformer layers are fine-tuned; base model frozen
- Trained for 5000 steps on 4×A100 GPUs (80GB), ~1 day





(a) Sana (b) Transformer-ControlNet

Global Force Encoding

Inputs: Force magnitude $F \in [0,1]$, angle $\theta \in [0°,360°)$

Encoded as a 3-channel tensor:

- Channel 1: $-1+2F \in [-1,1]$
- Channel 2: $\cos(\theta)$
- Channel 3: $\sin(\theta)$

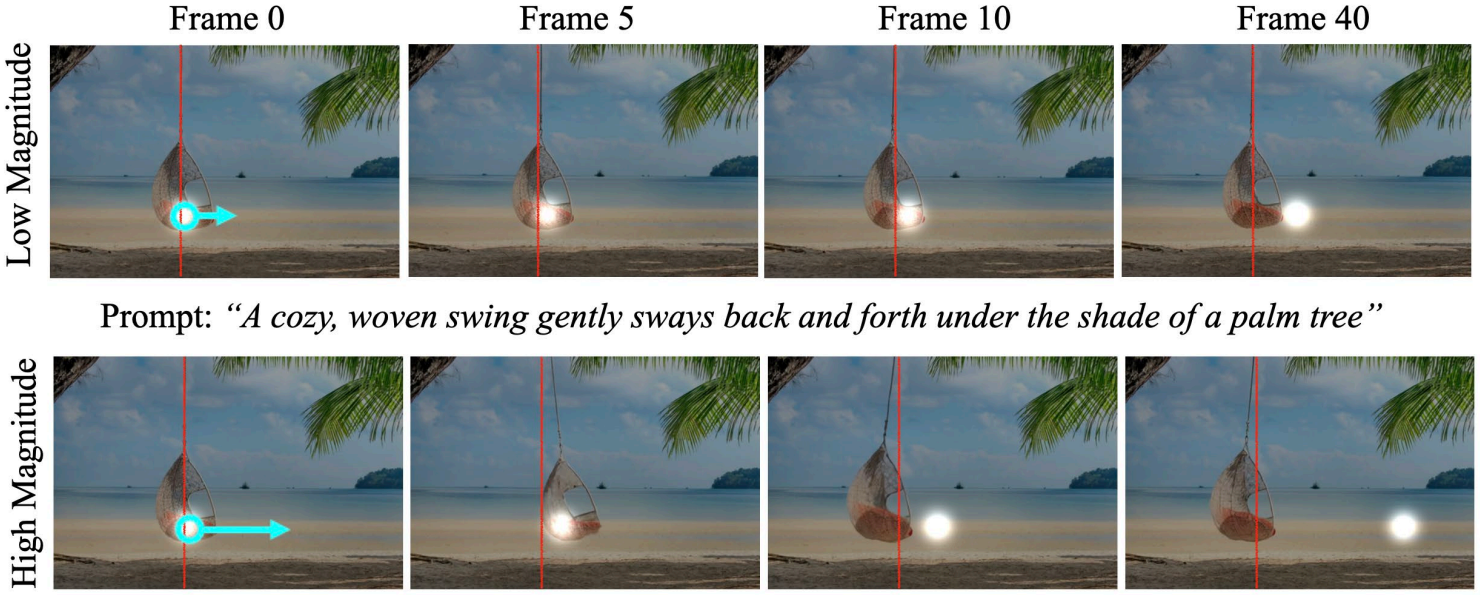
Resulting tensor: $\pi \in R^{f \times 3 \times h \times w}$ (same shape as video)

Local Force Encoding

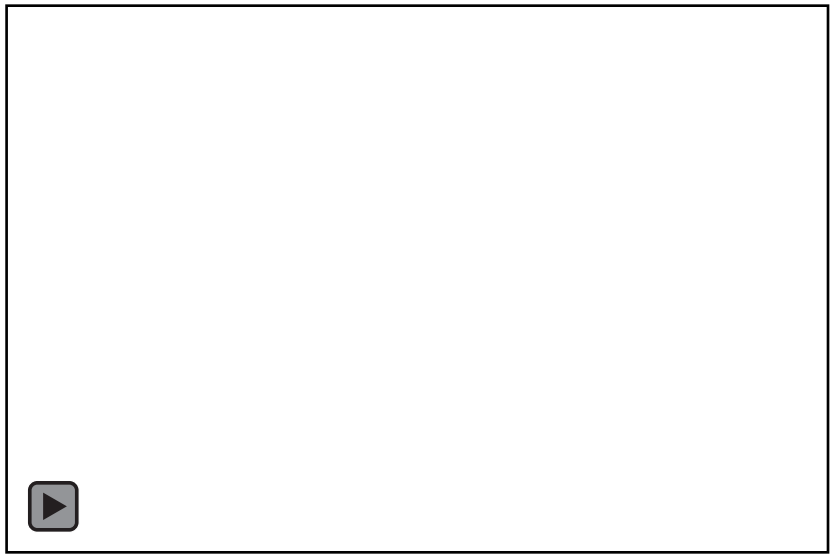
Inputs: Pixel location (x,y), force magnitude F, angle θ

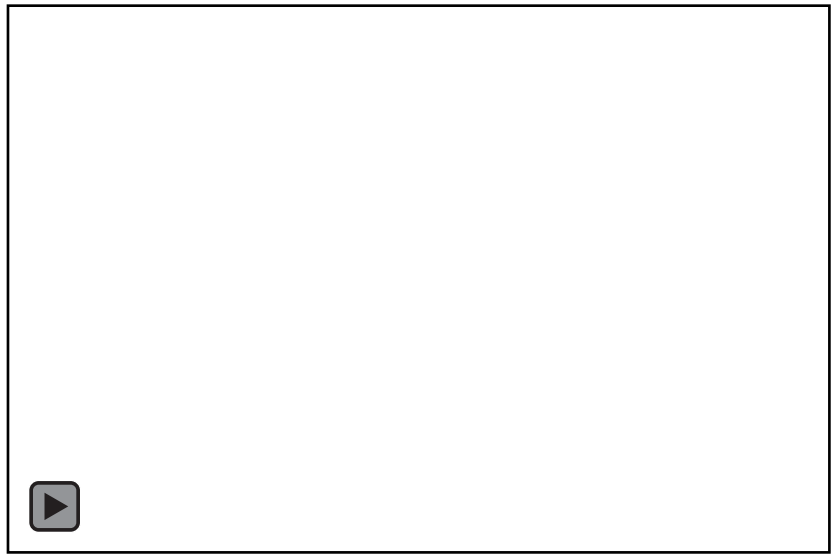
Encoded as a moving Gaussian blob:

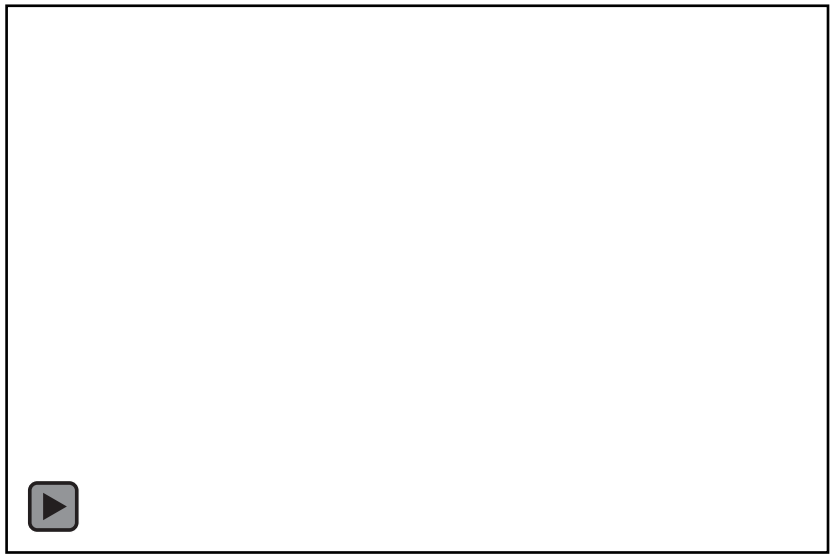
- Starts at (x,y), moves along θ direction
- Displacement proportional to F
- Creates dynamic spatio-temporal tensor representing the force vector

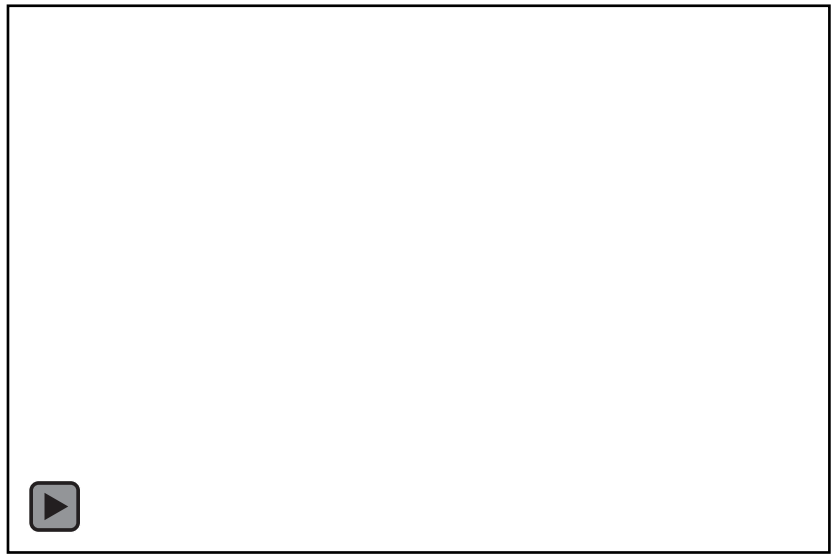


Qualitative Results (Global Force)

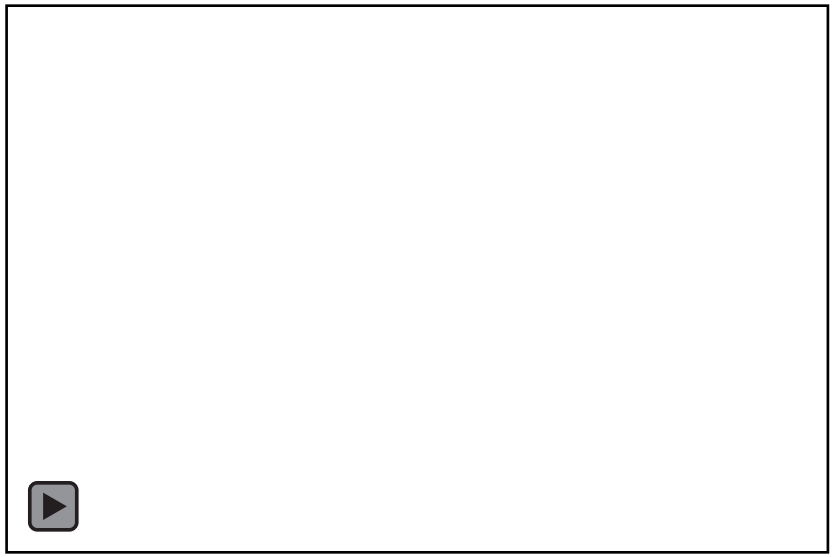


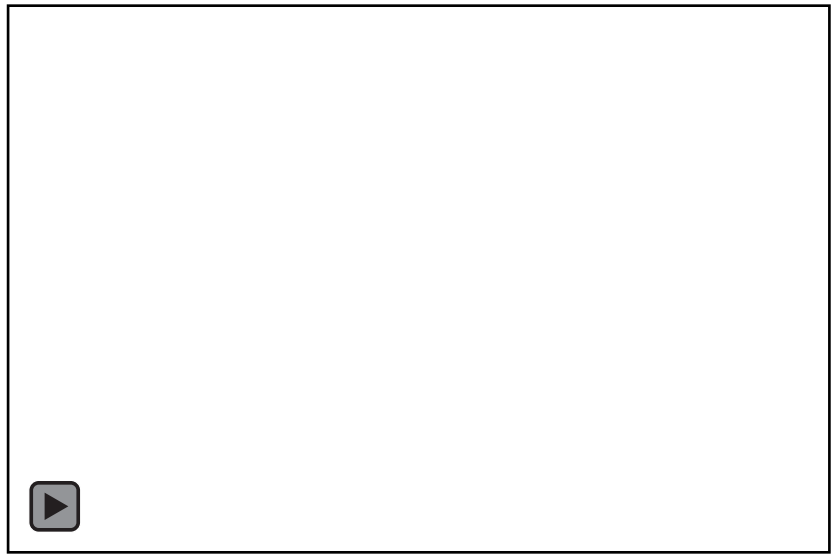


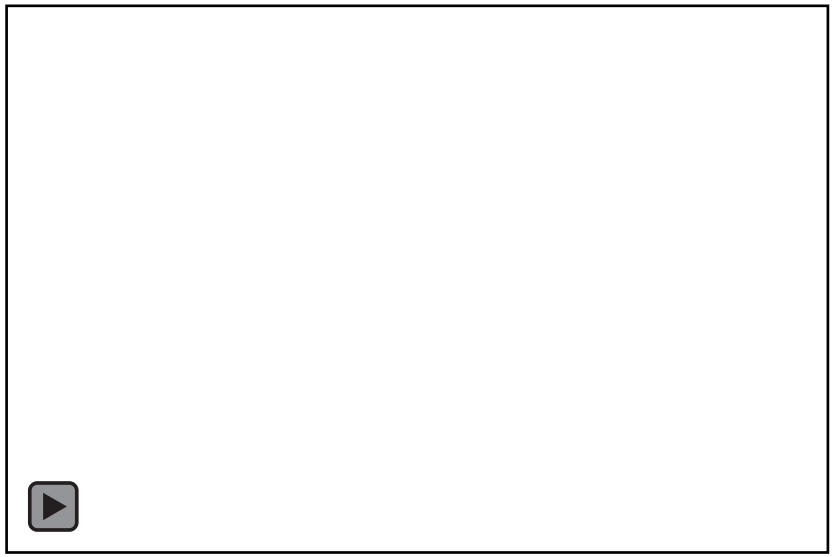


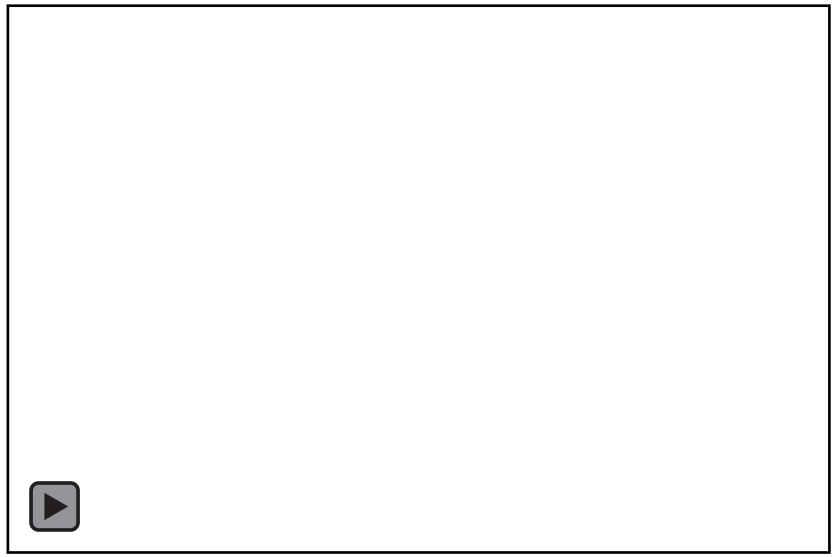


Qualitative Results (Local Force)

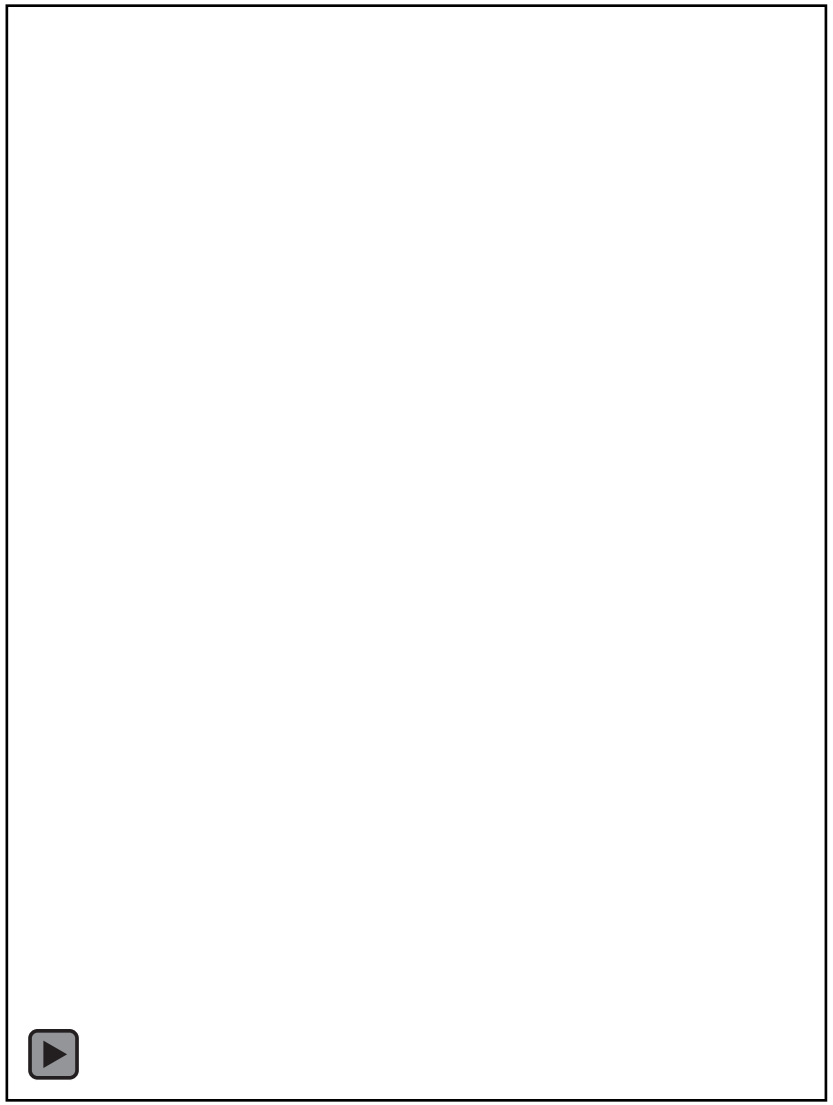


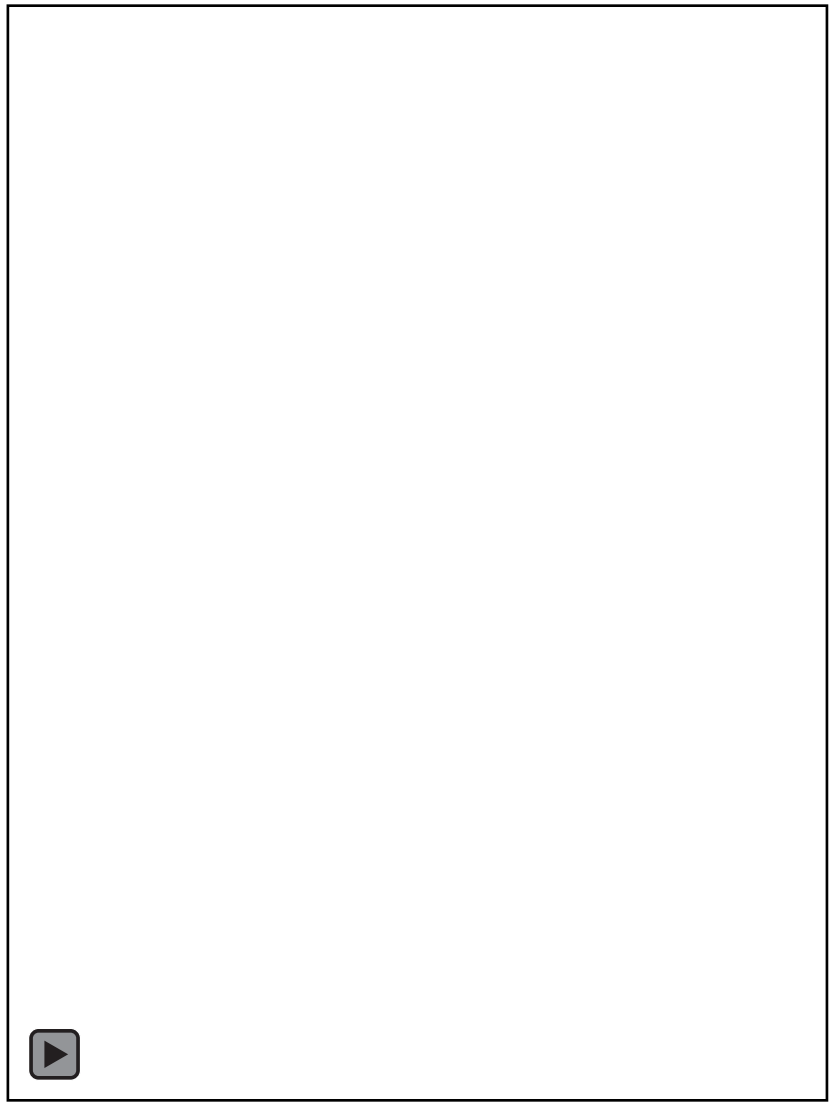




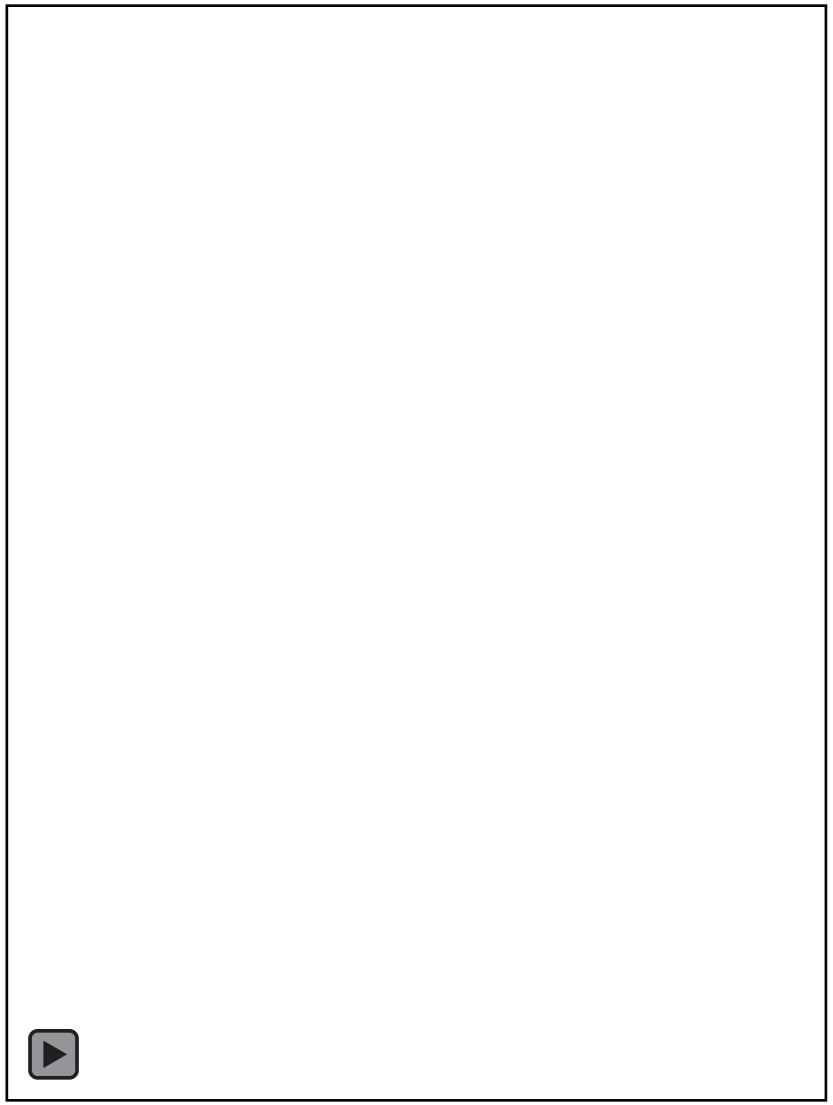


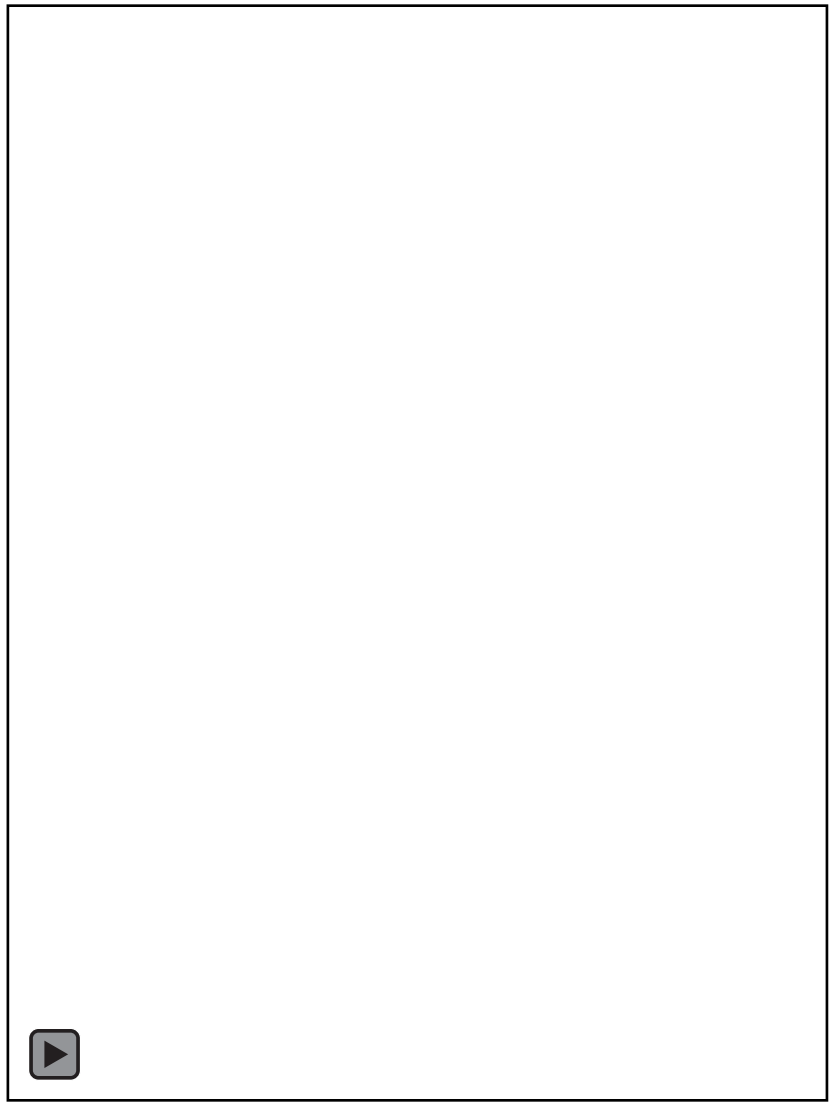
Qualitative Results (Mass Understanding)





Qualitative Results (Mass Understanding)





Limitations

Failure Case #1: The Physics is Out-of-Domain For the Base Model

The egg rolls in the prompted direction, but the base video model has difficulty rolling non-spherical objects, so the egg appears to float

The kite is blown in the prompted direction, but the base video model has difficulty generating a physically plausible video of a kite dragging a person

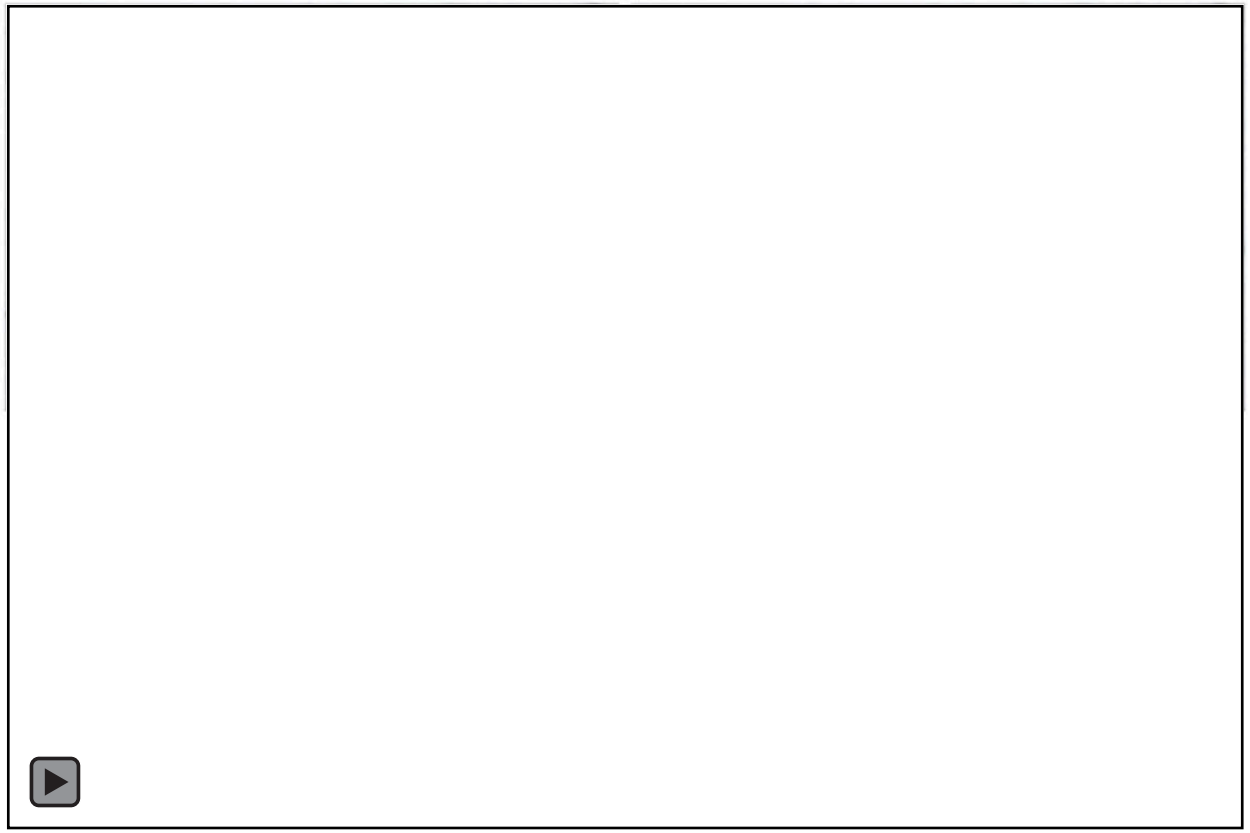
Limitations

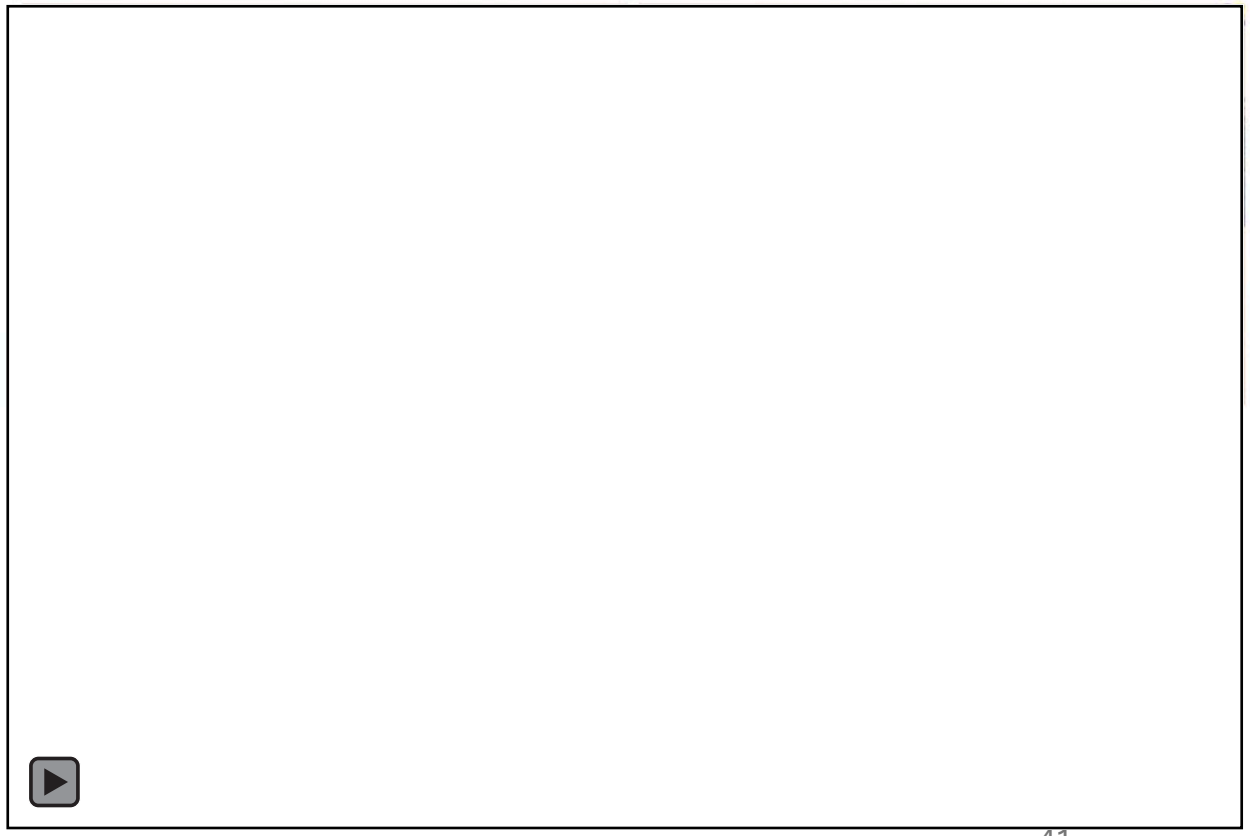
Failure Case #1: The Physics is Out-of-Domain For the Base Model

The kite is blown in the prompted direction, but the base video model has difficulty generating a physically plausible video of a kite dragging a person

Limitations

Failure Case #1: The Physics is Out-of-Domain For the Base Model





Limitations

Failure Case #2: The Base Video Model's Prior Competes with the Force Prompt

The rocking chair moves in the prompted direction, but the base video model has trouble distinguishing between foreground and background objects

The confetti moves in the prompted direction, but the base video model forces the scene to conjure extra confetti

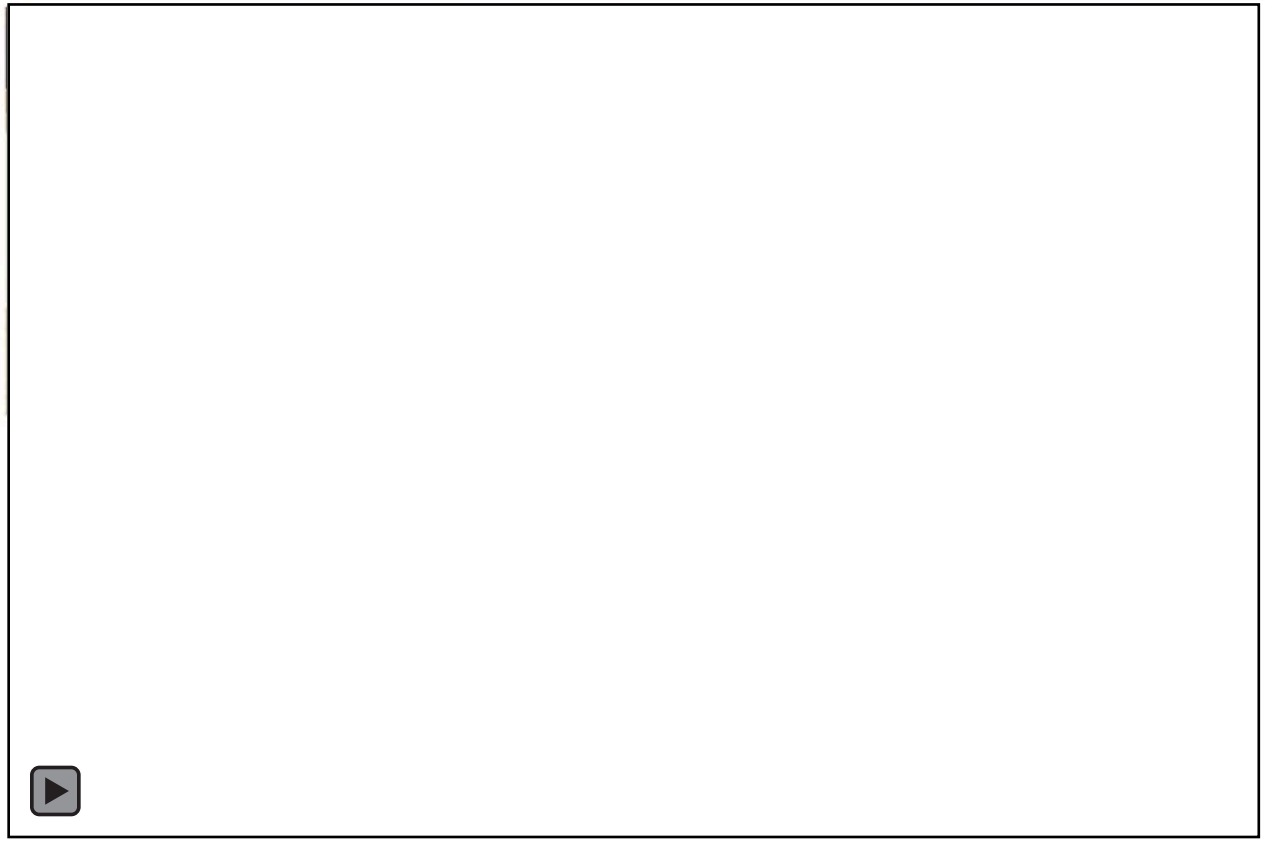
Limitations

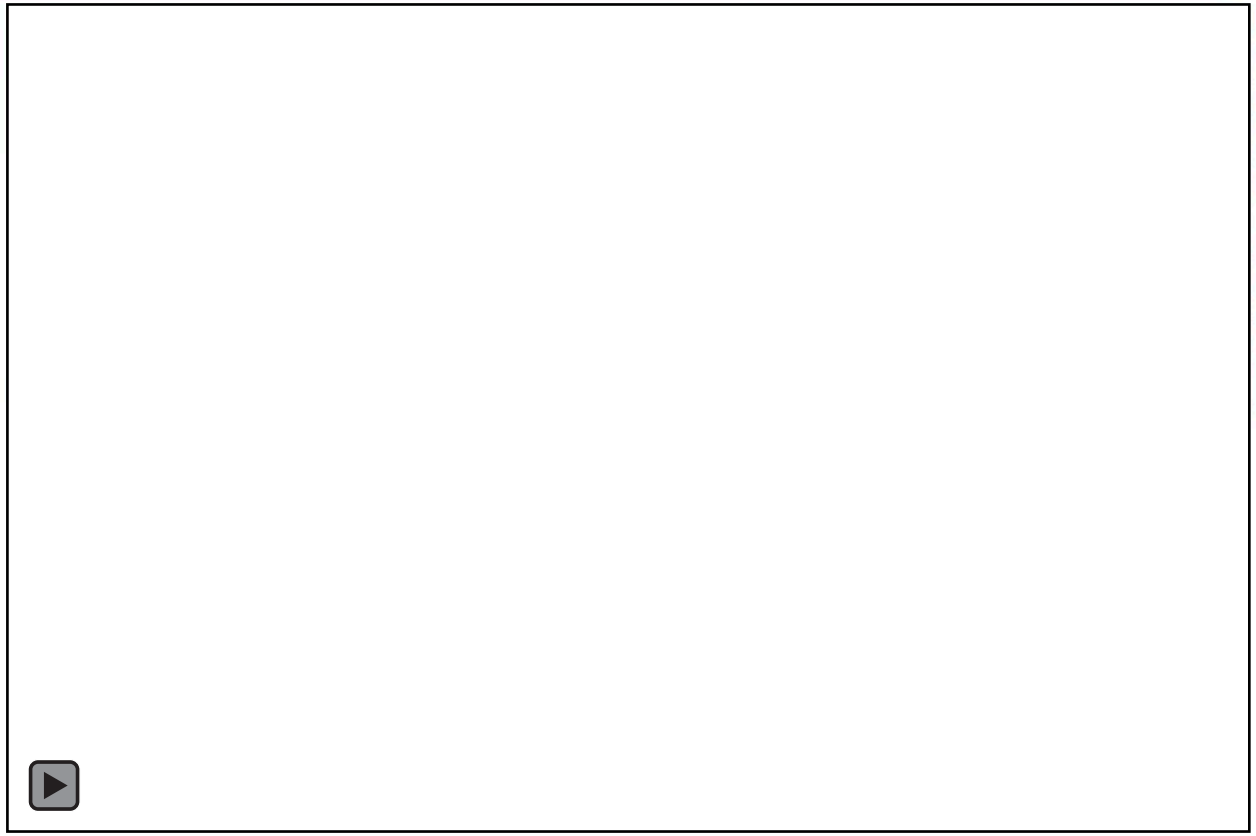
Failure Case #2: The Base Video Model's Prior Competes with the Force Prompt

The confetti moves in the prompted direction, but the base video model forces the scene to conjure extra confetti

Limitations

Failure Case #2: The Base Video Model's Prior Competes with the Force Prompt





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