VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control

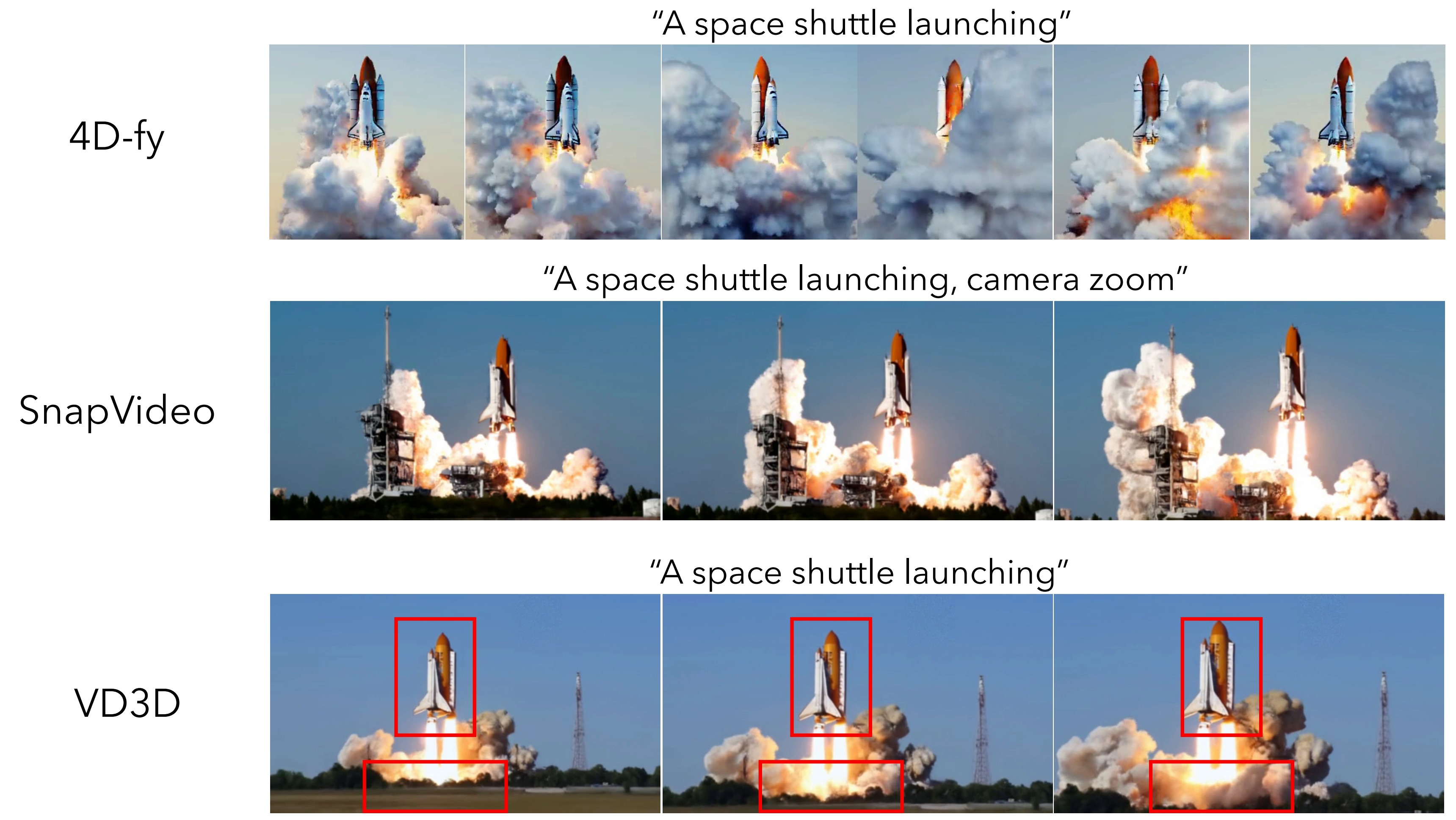
Sherwin Bahmani[1,2,3] Ivan Skorokhodov[3] Aliaksandr Siarohin[3] Willi Menapace[3] Guocheng Qian[3] Michael Vasilkovsky[3]
Hsin-Ying Lee[3] Chaoyang Wang[3] Jiaxu Zou[3] Andrea Tagliasacchi[1,4] David B. Lindell[1,2] Sergey Tulyakov[3]

[1] University of Toronto [2] Vector Institute [3] Snap Inc. [4] SFU

Website

Motivation

- Text-to-4D generation approaches (e.g., **4D-fy**) lack photorealism
- Previous text-to-video generation methods (e.g., **SnapVideo**) can only control camera with text


"A space shuttle launching"
4D-fy
"A space shuttle launching, camera zoom"
SnapVideo
"A space shuttle launching"
VD3D

Task: We tackle camera-controllability for text-to-video diffusion transformers

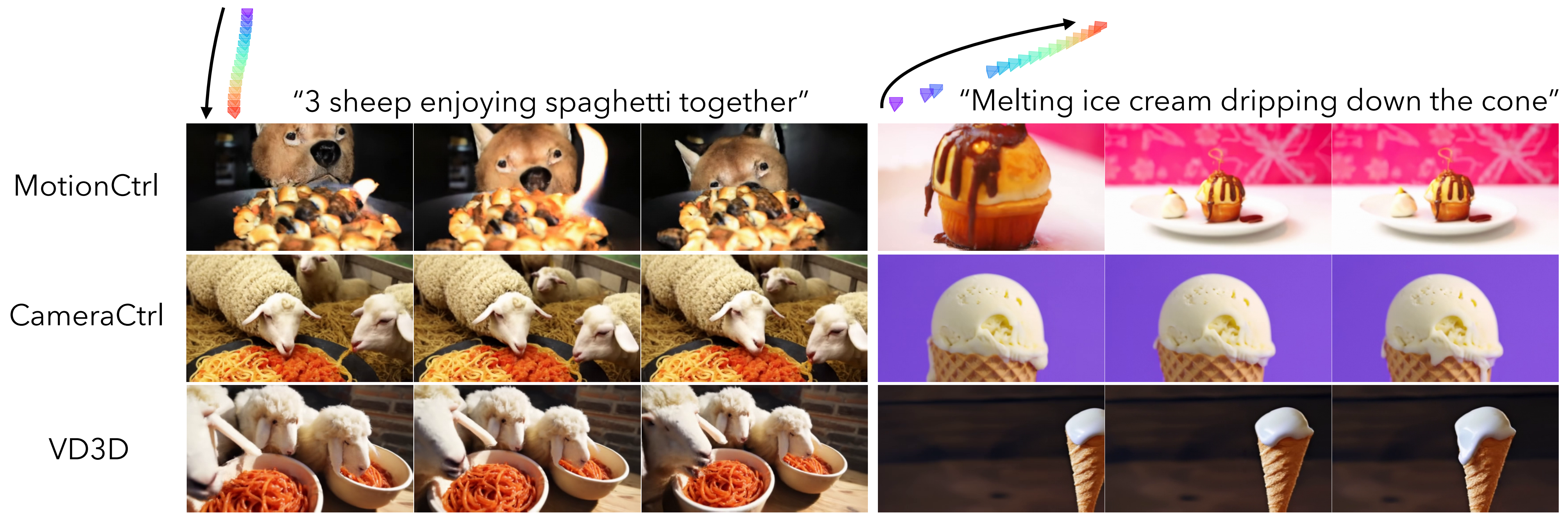
Inputs:

- Text and/or image describing the scene content
- Sequence of camera matrices (extrinsics and intrinsics) describing the camera motion

Output: Video following the text, image, and camera motion conditioning

Comparisons

MotionCtrl and CameraCtrl were designed for U-Net models and not transformers


"3 sheep enjoying spaghetti together"
"Melting ice cream dripping down the cone"
MotionCtrl
CameraCtrl
VD3D

Method

Our approach injects camera control through Plücker conditioning into a pre-trained video diffusion transformer

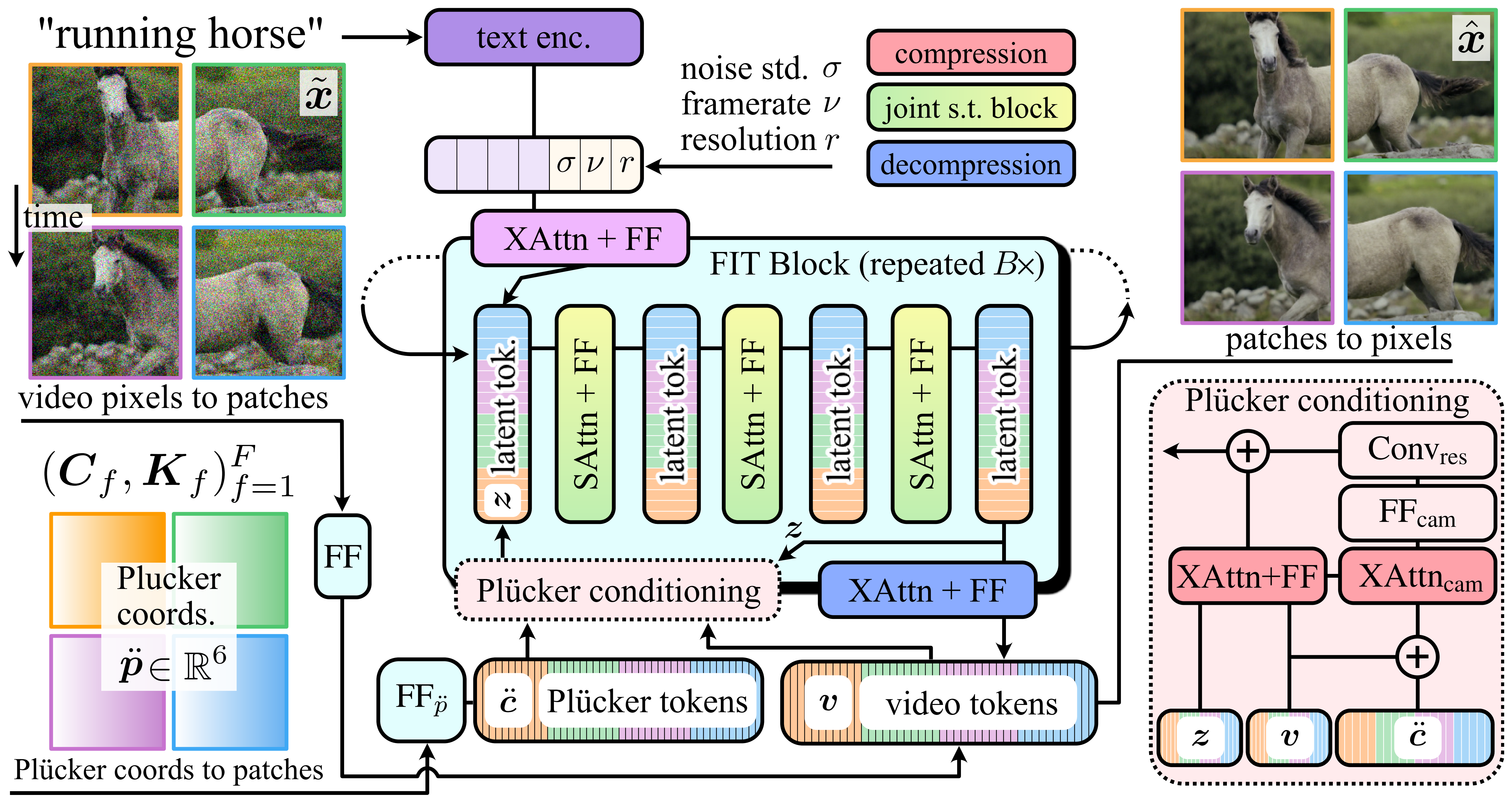
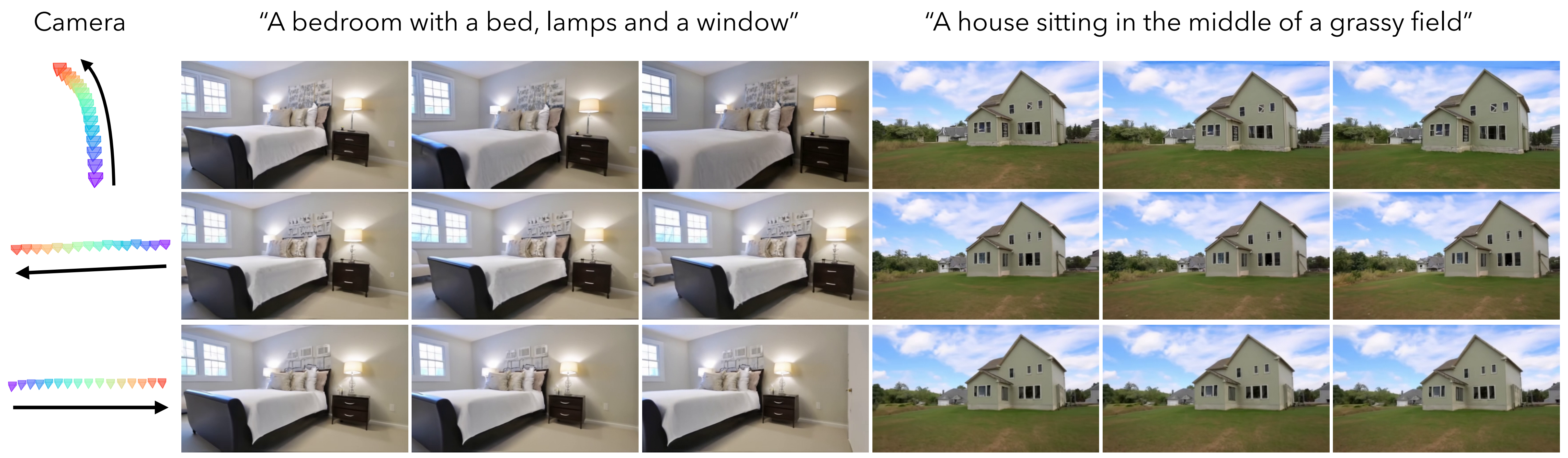


Image-to-Multiview Generation

VD3D can generate multiple viewpoints of a scene given a single real image


Camera
"A bedroom with a bed, lamps and a window"
"A house sitting in the middle of a grassy field"

Out-of-Distribution Cameras