AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers

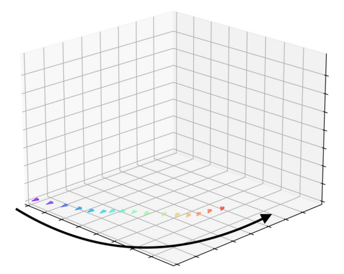
Sherwin Bahmani*[1,2,3] Ivan Skorokhodov*[3] Guocheng Qian[3] Aliaksandr Siarohin[3]
Willi Menapace[3] Andrea Tagliasacchi[1,4] David B. Lindell[1,2] Sergey Tulyakov[3]

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

Website

Motivation

- Camera-controlled video generation methods (e.g., **CameraCtrl** or **VD3D**) often degrade visual and motion quality
- Currently, there is no analysis of what pre-trained video diffusion transformers understand about 3D control

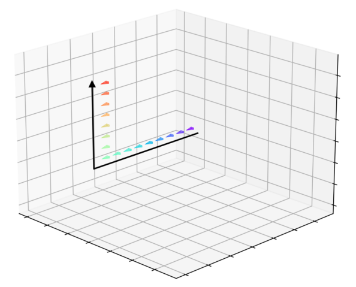


Task: We analyze and improve camera-controllable text-to-video generation

Inputs:

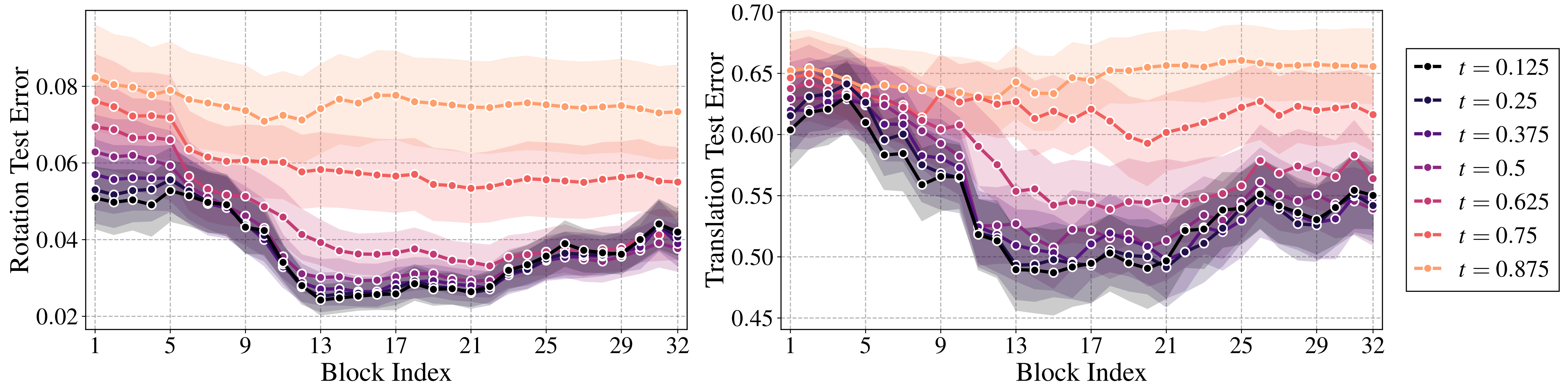
- Text and/or image describing the scene content
- Sequence of cameras (extrinsics and intrinsics) describing the camera motion

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



Video Models as Camera Pose Estimators

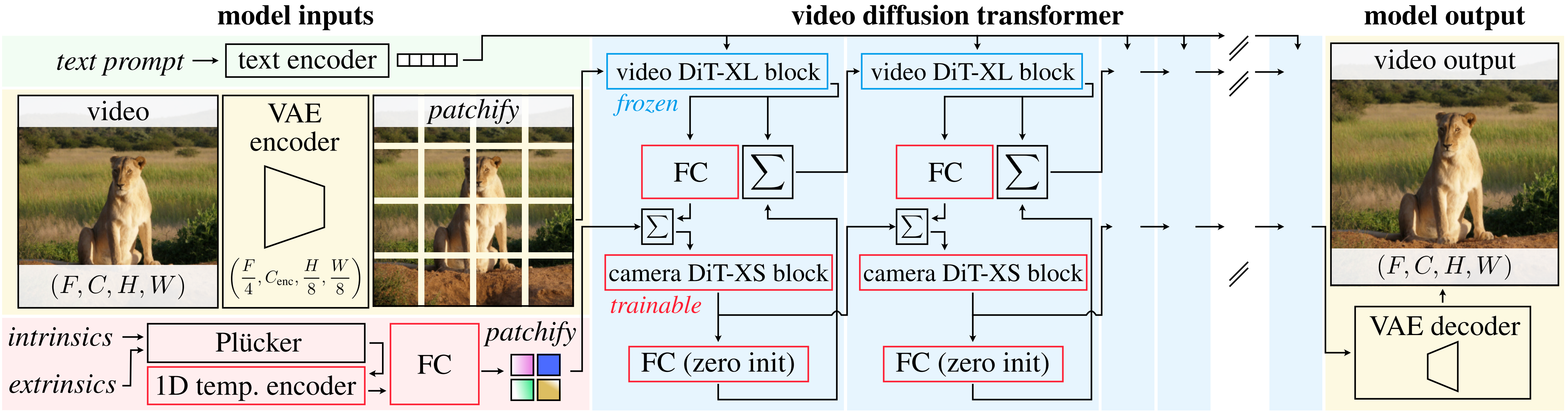
We conduct linear probing experiments based on the features of each DiT block for various noise levels



Result: Middle blocks carry most accurate information, i.e., camera emerges in early layers to help middle/late blocks

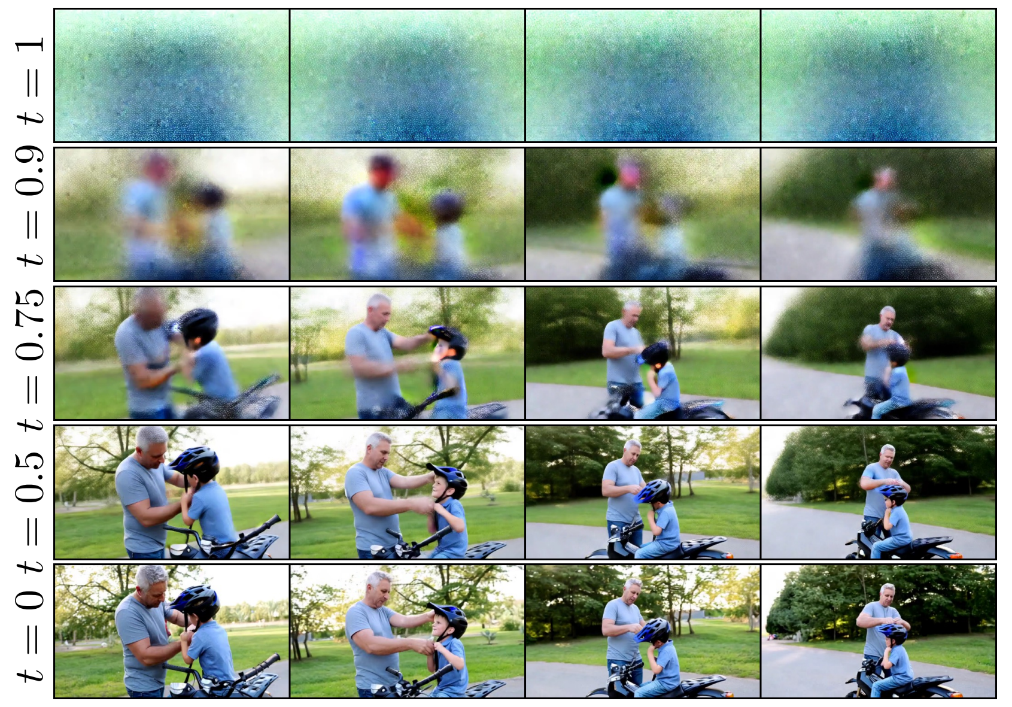
Method

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

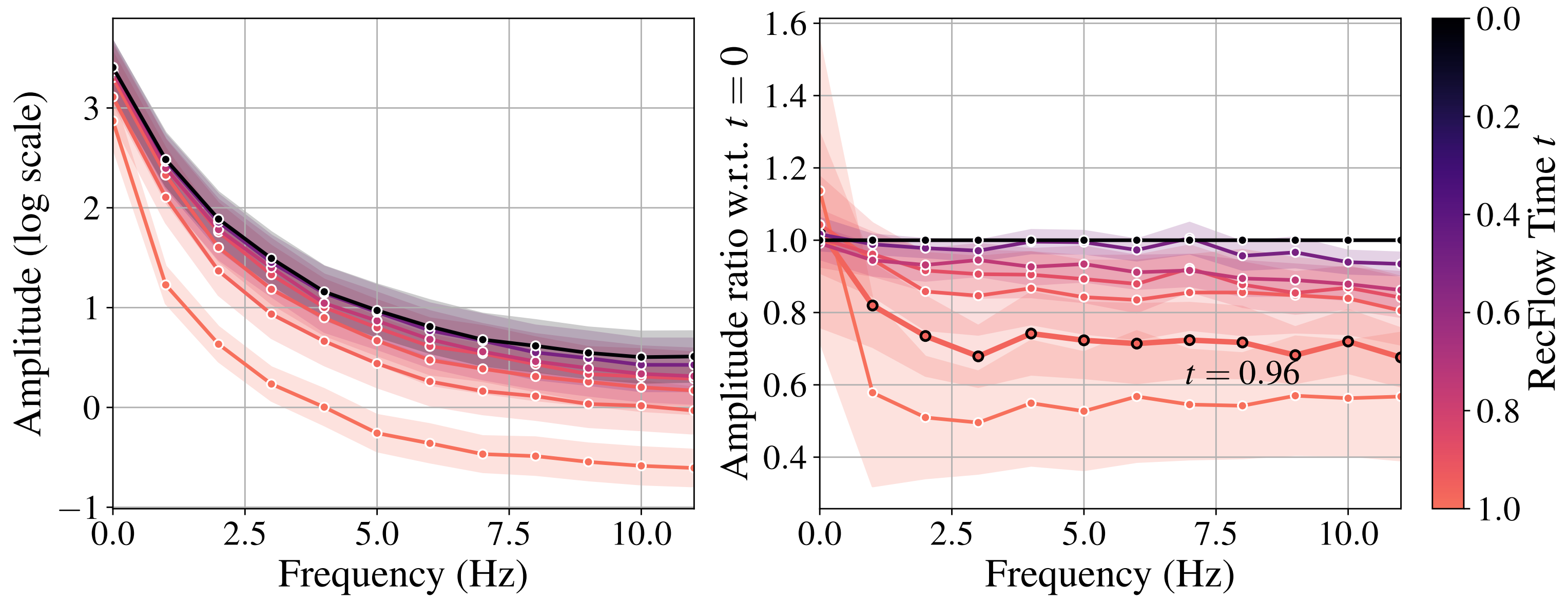


Motion Analysis for Diffusion Timesteps

Video at different diffusion timesteps



Motion spectral volumes for different diffusion timesteps and their ratio



Result: Even at t=0.96 (first ≈4% of the steps), the low-frequency motion components have already been created

Dataset Curation

Multi-view data (RealEstate10K)

+ Camera motion − Scene motion

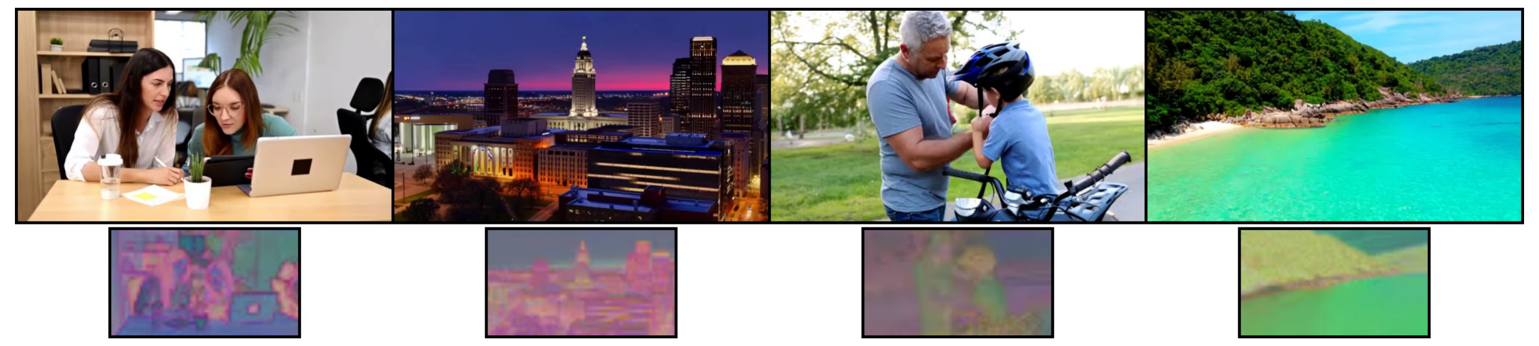
Single-view data (curated video data)

− Camera motion + Scene motion

We jointly train on static multi-view data (left) and dynamic single-view data (right)

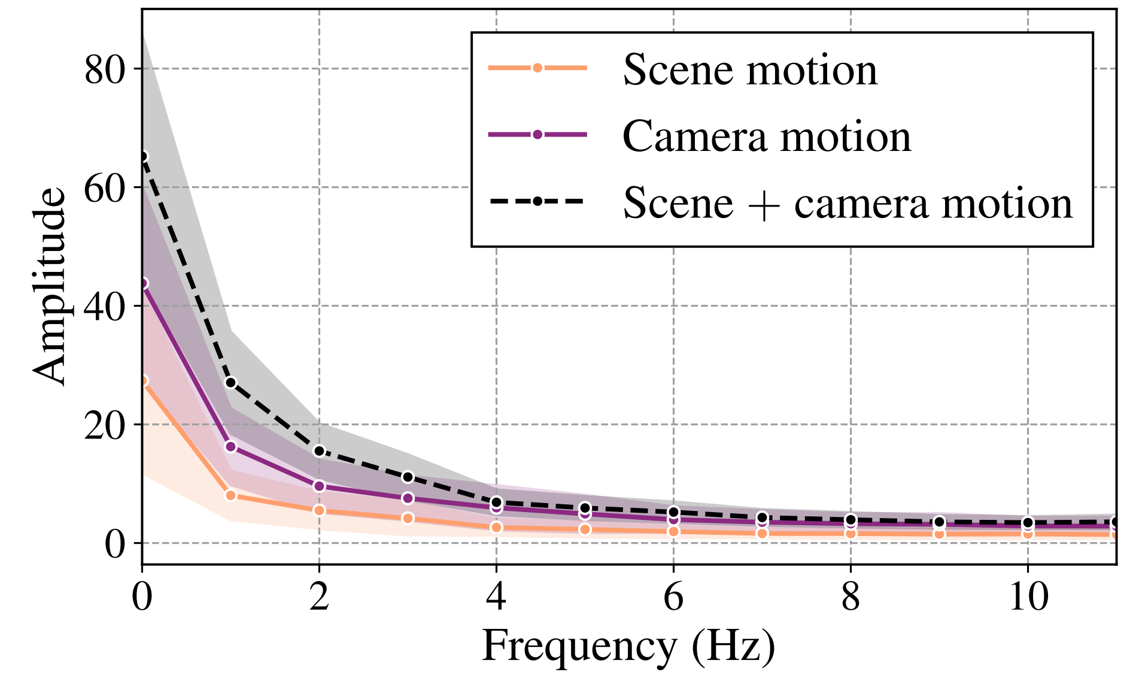
Motion Type Analysis

1. Estimate optical flow in the PCA space of video latents



2. Compute average magnitude of motion spectral volumes

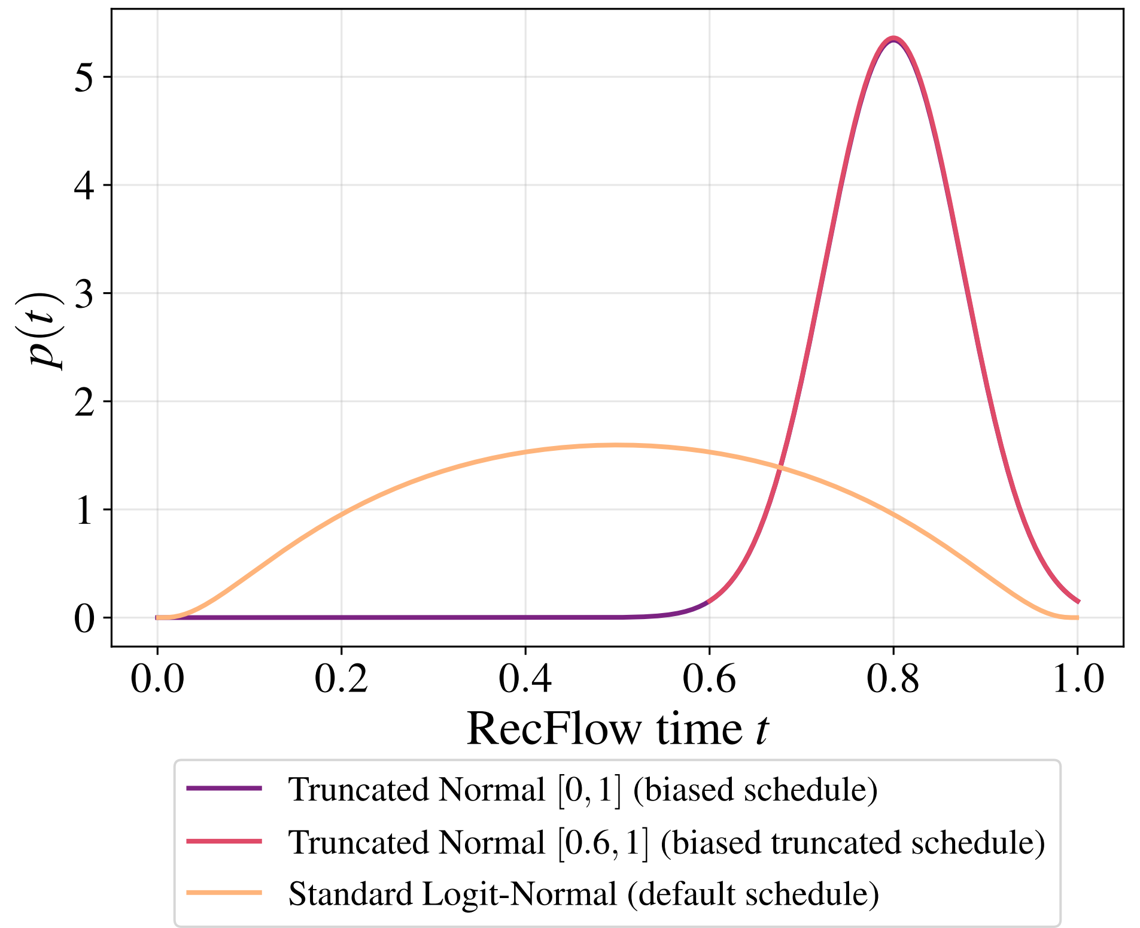
- Sliding temporal window creates multiple shorter sub-videos
- Fast Fourier Transform (FFT) independent for each pixel
- Average amplitudes along spatial, temporal offset, and batch



Result: Camera motion higher than scene motion at low frequencies

Diffusion Timestep Scheduling

We bias train and inference time scheduling towards low frequency


$p(t)$
RecFlow time t
Truncated Normal $[0, 1]$ (biased schedule)
Truncated Normal $[0.6, 1]$ (biased truncated schedule)
Standard Logit-Normal (default schedule)