Training-free Camera Control for Video Generation

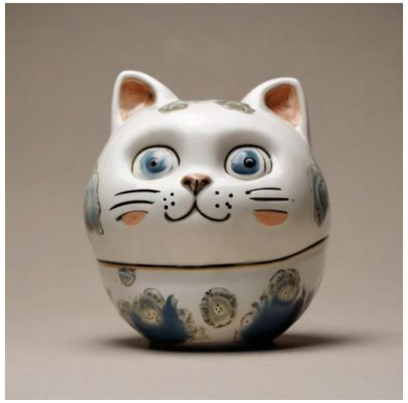
Chen Hou, Zhibo Chen
University of Science and Technology of China

Motivation

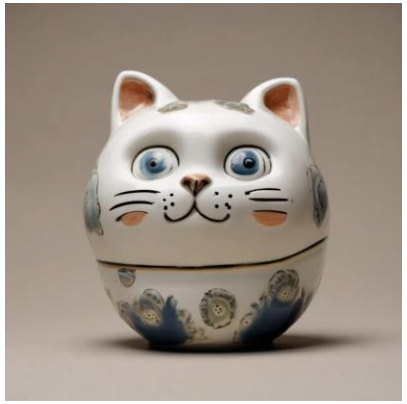
Discovery 1:
Text cannot control video's camera motion.

Prompt:

A ceramic cat.



A ceramic cat,
camera rotates around it.



(what we expect)

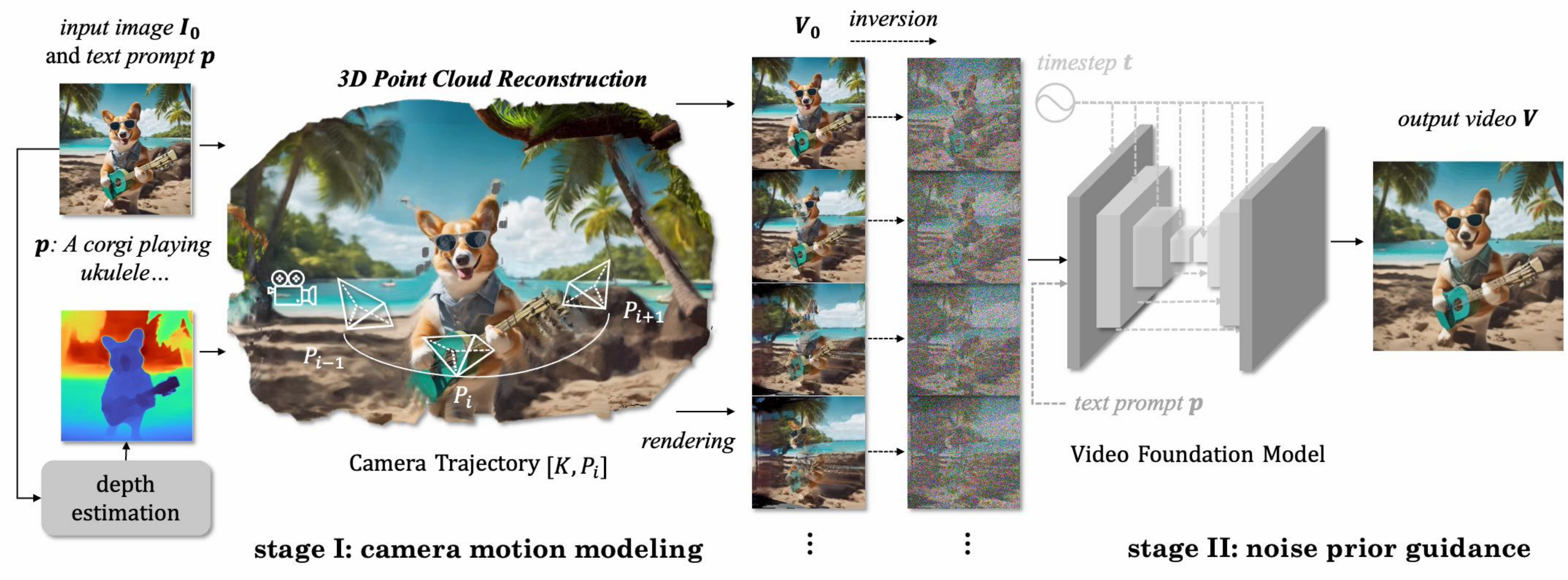
Discovery 2:
Finetuning can harm dynamics and diversity.

MotionCtrl

CameraCtrl

How to fully preserve large models' prior while controlling camera motion?

Inspiration: diffusion's latent noise could affect the layout of output


input image I_0 and text prompt p
p: A corgi playing ukulele...
depth estimation
3D Point Cloud Reconstruction
P_{i-1}
P_i
P_{i+1}
Camera Trajectory $[K, P_i]$
rendering
V_0
inversion
timestep t
text prompt p
Video Foundation Model
output video V
stage I: camera motion modeling
stage II: noise prior guidance

a. camera moves
b. render -> 2D inpainting -> 3D lifting
c. view consistency optimization

Experiments

Table 1: **Quantitative comparisons.** Our method attains comparable performance with finetuned methods in both video generation quality and camera motion alignment.

Method	Video Quality				Motion Accuracy		
	FVD ↓	FID ↓	IS ↑	CLIP-SIM ↑	ATE ↓	RPE-T ↓	RPE-R ↓
SVD	1107.93	68.51	7.21	0.3095	4.23	1.79	0.021
MotionCtrl+SVD	810.59	69.03	**7.17**	0.3076	4.19	**1.07**	0.012
CameraCtrl+SVD	951.80	**67.59**	6.82	**0.3138**	4.22	1.17	0.013
CamTrol+SVD	**778.46**	68.06	7.05	0.3110	**4.17**	**1.07**	**0.010**
Reference	-	-	-	-	*3.60*	*0.89*	*0.008*

Table 2: **Computational analysis of inference process,** evaluated under unified settings.

		SVD	MotionCtrl	CameraCtrl	CamTrol ($t_0 = 10$)
Max GPU memory(MB)		11542	31702	26208	11542
Time (s)	pre-process	-	-	-	56
	inference	11	32	42	8

Table 3: **Quantitative effect of t_0.**

t_0	Video Quality				Motion Accuracy		
	FVD ↓	FID ↓	IS ↑	CLIP-SIM ↑	ATE ↓	RPE-T ↓	RPE-R ↓
$t_0 = 5$	1079.88	68.52	7.14	0.3100	4.17	1.09	0.012
$t_0 = 10$	778.46	68.06	7.05	0.3110	4.17	1.07	0.010
$t_0 = 15$	754.14	67.98	7.00	0.3107	4.13	1.02	0.008

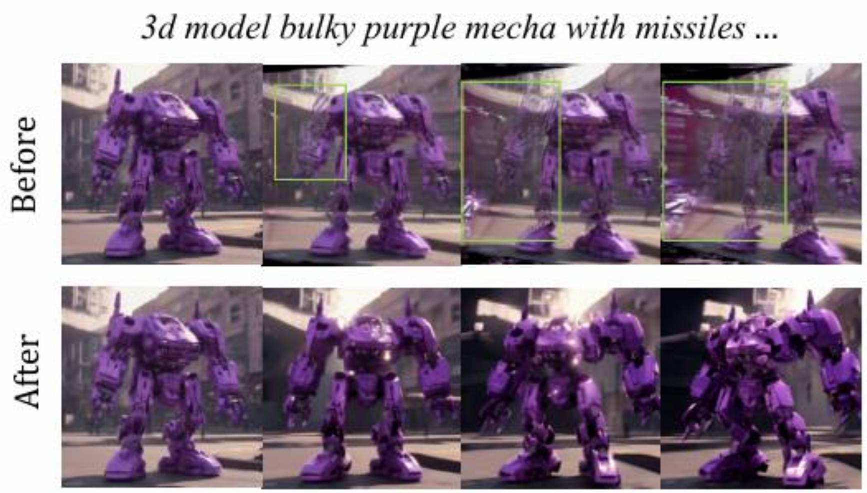


Figure 6: **Effectiveness of layout prior.**

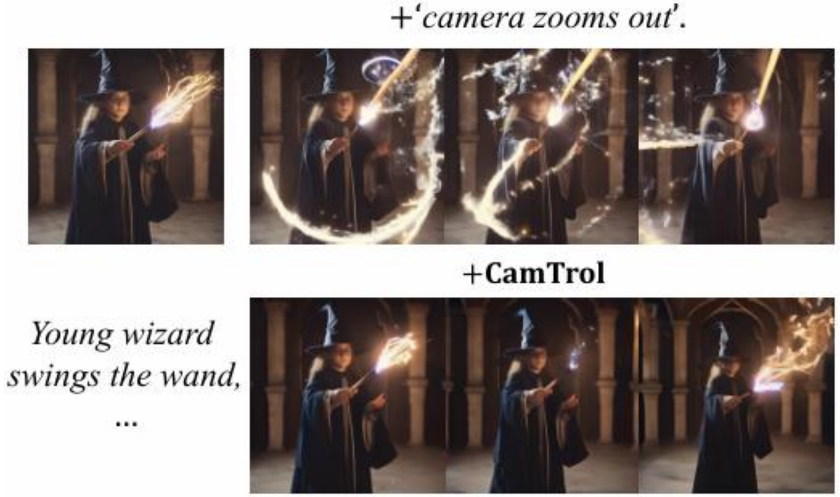
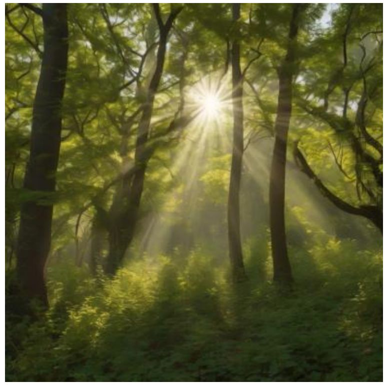


Figure 5: **Comparison with base model.**

Results – Basic Camera Movements

Zoom Out

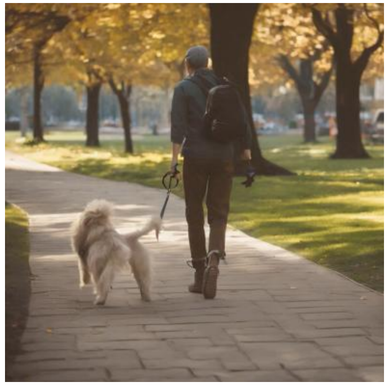


Pan Left

Tilt Up

Truck Right

Roll CW



Zoom In

Pan Right

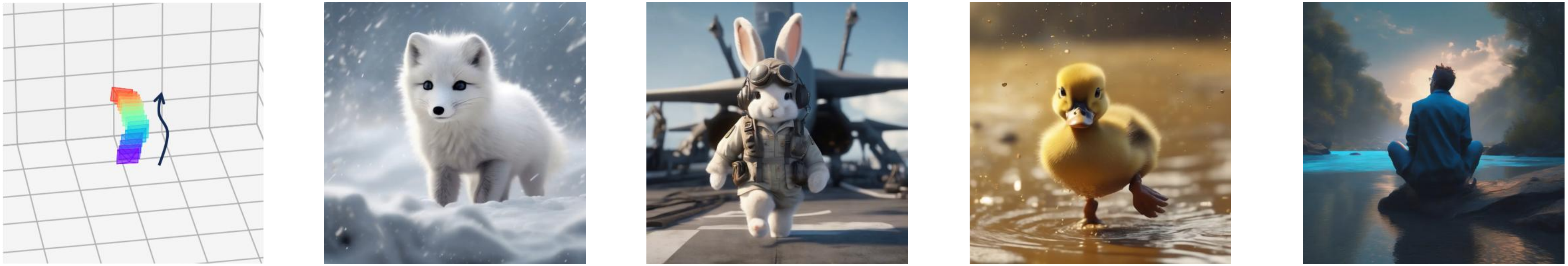
Tilt Down

Truck Left

Roll ACW

Results – Hybrid & Complex Motions

Hybrid: Zoom In first, then Pedestal Up.



Complex Trajectory

Results – Multi-Trajectory Generation

Zoom Out

Tilt Up

Pan Left

Zoom In

Tilt Down

Pan Right

Results – Different Motion Scales

Scale I

Scale II

Scale III

Results – Unsupervised 3D Video Generation

dynamic 3D rotation video

3D object video

Results – Plug-and-Play with Most Video Diffusion Models

Combined with CogVideoX-t2v:

Tilt Down

Hybrid: Zoom Out + Pedestal Up + Truck Left + Tilt Down + Pan Right

Zoom In

Rotate Clockwise

Thanks for listening!

For more video results,

please refer to our website: https://lifedecoder.github.io/CamTrol/.