



University
of Glasgow

Pixel-to-4D: Camera-Controlled Image-to-Video Generation with Dynamic 3D Gaussians

Melonie de Almeida, Daniela Ivanova, Tong Shi,
John Williamson, Paul Henderson

University of Glasgow





cameras



What are the existing approaches?

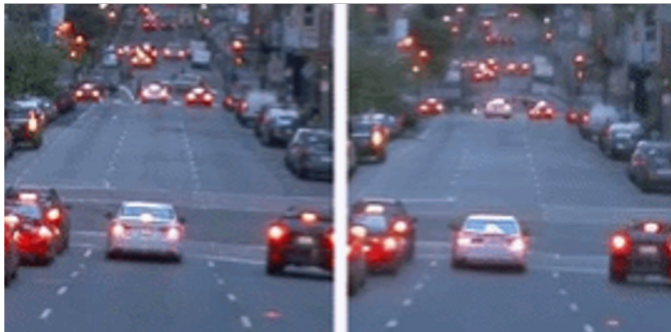
Conditioning **video diffusion models** on camera motion to generate controllable novel-view videos

Ex:-CameraCtrl, MotionCtrl, Caml2V, RealCami2V

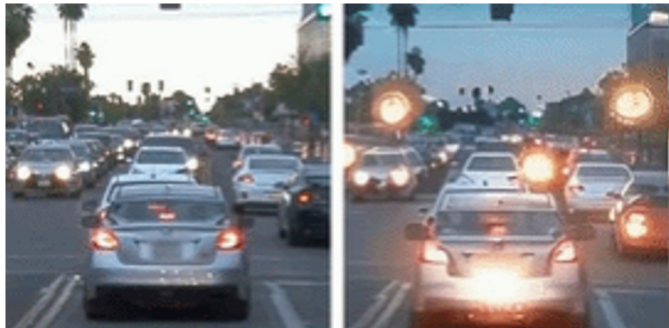
Limitations of Video Diffusion-Based Methods

Limitations of Video Diffusion-Based Methods

Failure to follow challenging trajectories



Unnecessary lighting changes



Object shape deformation



Style and appearance inconsistency



A better solution?



What if we could infer a dynamic 3D representation from a single image?

Predict a temporally evolving 3D representation, freely move the camera through the scene, and render novel viewpoints as the world unfolds.

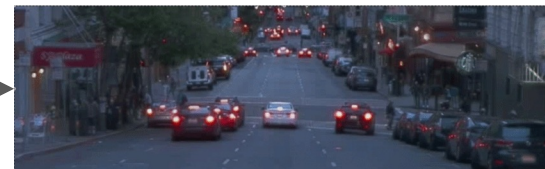
Overview



input



dynamic 3D representation



rendered video

Predicting a 4D scene for an input image

Step 1: predicting 3D scene for the input image – well studied

ex:- *Splatter-image* Szymanowicz et al., *CVPR 2024*

Step2: Predicting object dynamics

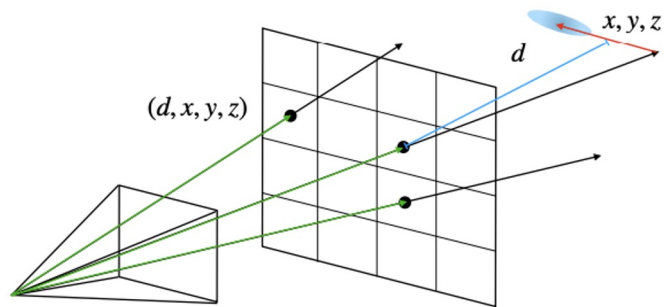
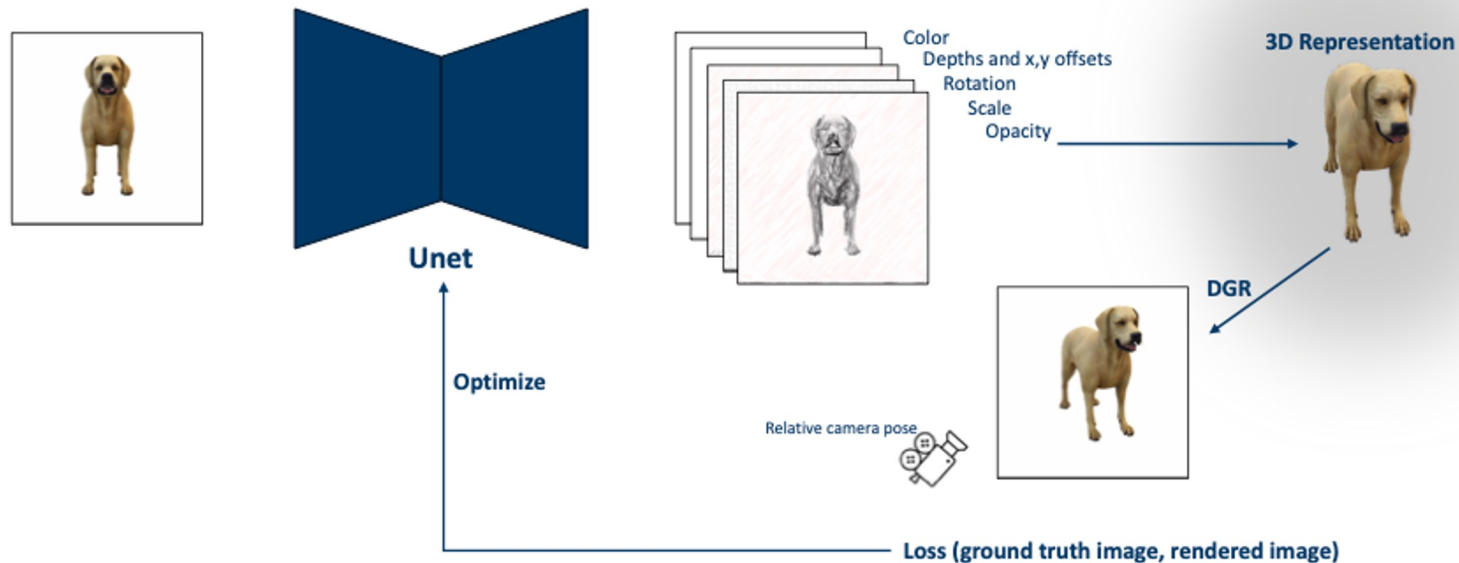


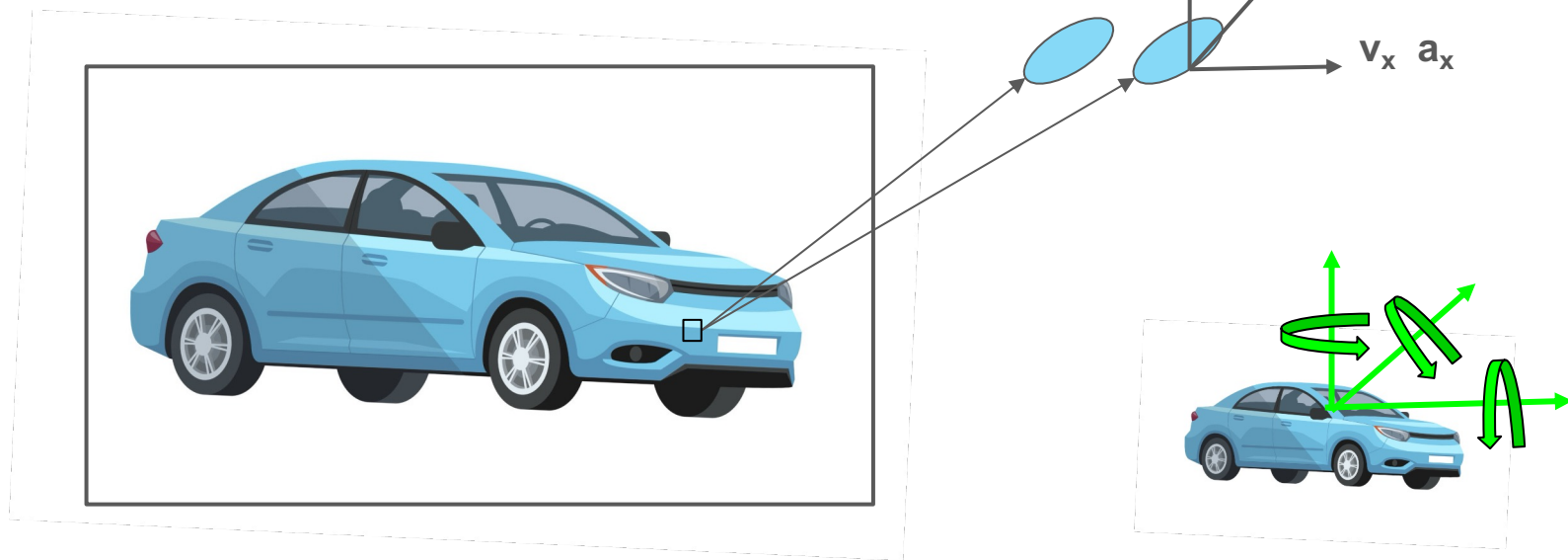
Image-to-3D: Splatter-Image

Szymanowicz et al., CVPR 2024



How to convert 3D into 4D

Pixel-aligned velocities and accelerations



Modeling motion uncertainty

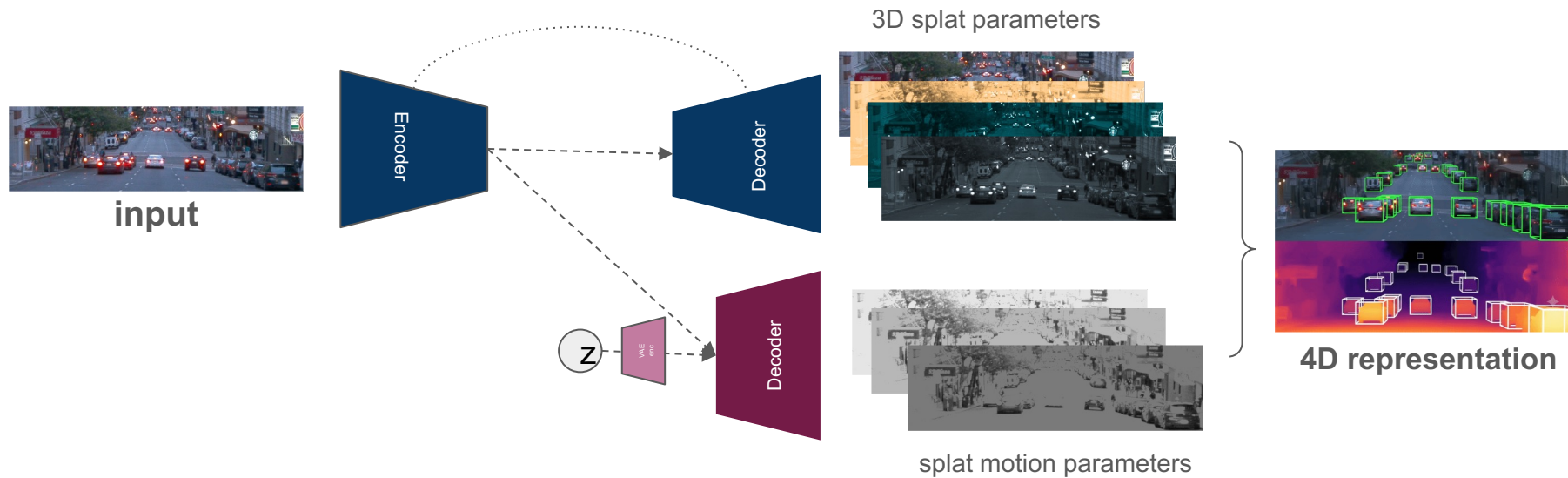
Learn a Bayesian posterior over motion

→ Sample plausible velocities and accelerations

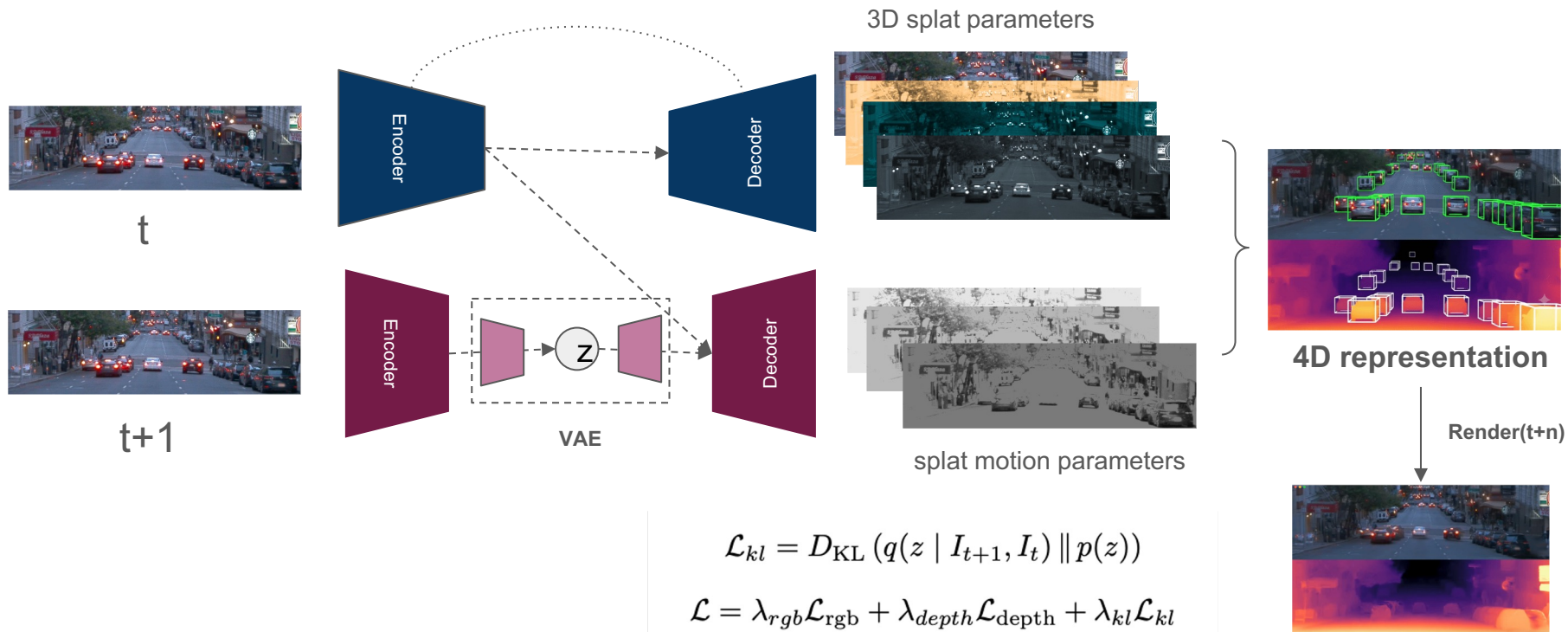
→ Amortized Variational Inference



Method



Training



KITTI dataset

gt

motionctrl

cameractrl

camv2v

realcamv2v

ours



LPIPS	0.507	0.490	0.488	0.447	0.387
FID	50.8	75.7	70.5	58.5	33.8

Waymo dataset

gt

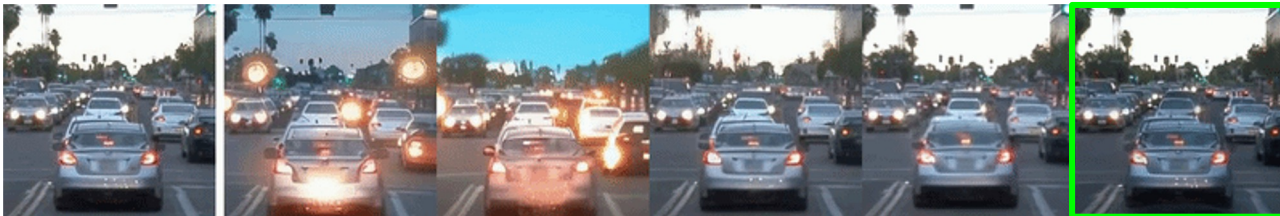
motionctrl

cameractrl

camv2v

realcamv2v

ours



LPIPS	0.491	0.501	0.499	0.449	0.352
FID	43.3	82.2	50.3	47.8	30.9

RealEstate10k dataset

gt

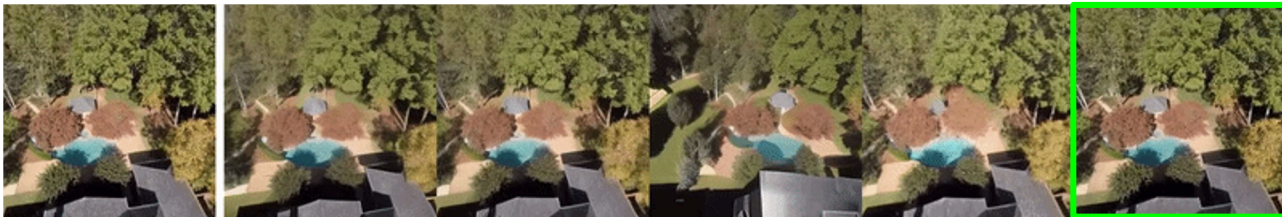
motionctrl

cameractrl

caml2v

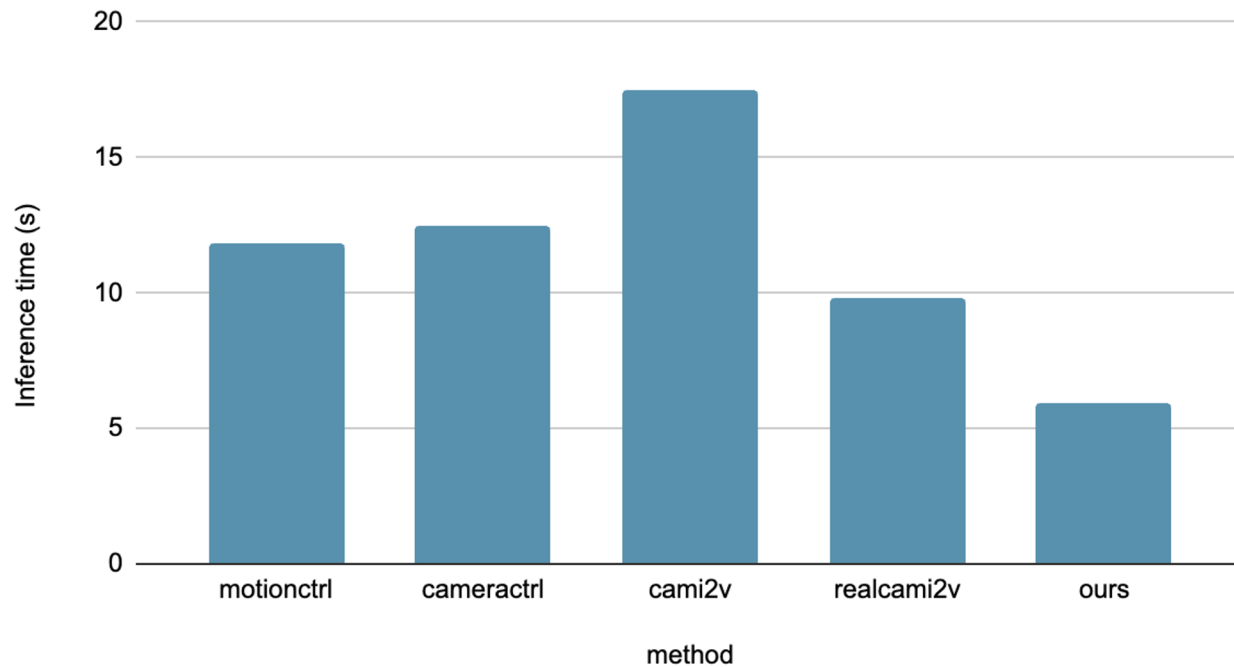
realcaml2v

ours

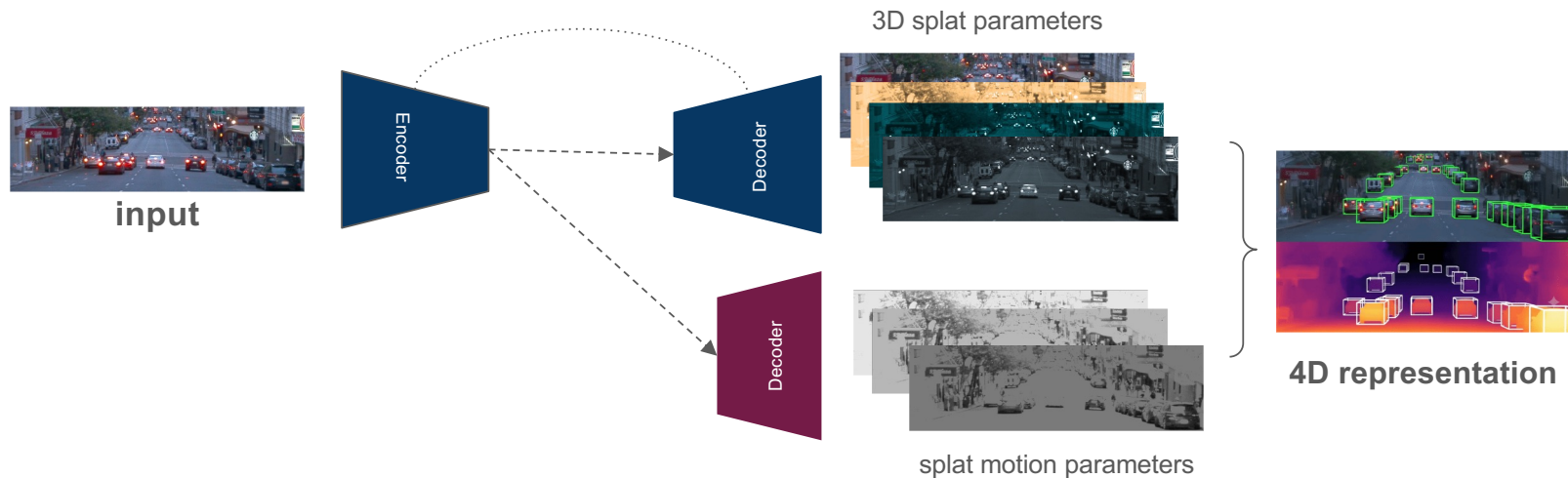


LPIPS	0.554	0.529	0.538	0.483	0.314
FID	57.5	62.8	56.7	49.1	24.2

Sampling time for 16 future frames (s)



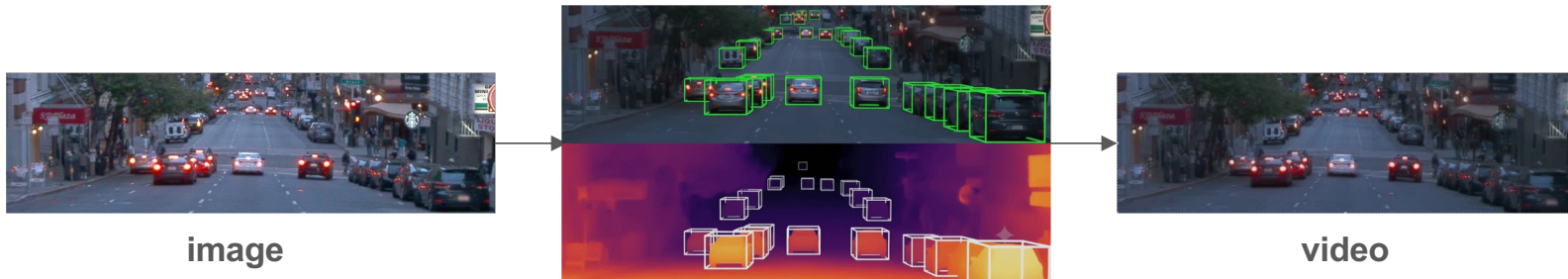
The Benefit of Probabilistic Modeling



	deterministic	probabilistic
LPIPS	0.293	0.288
FID	14.5	14.1

Conclusion

- Pixel-to-4D: Ultra-fast camera-controlled image-to-video generation
- One forward pass — no diffusion priors or test-time optimization.
- Accurate camera control with temporally consistent synthesis





University
of Glasgow

Thanks!

Project page:

