JOG3R: Towards 3D-Consistent Video Generators

Chun-Hao Huang[1], Niloy J. Mitra[1,2], Hyeonho Jeong[3], Jae Shin Yoon[1], Duygu Ceylan[1]

[1]Adobe Research, [2]University College London, [3]KAIST

BMVC 2025

DiT-based Video diffusion models

✓ Generate photoreal videos

× Videos often lack 3D consistency

× Do not estimate 3D point maps

ViT-based 3D foundation models

✓ Estimate 3D point maps

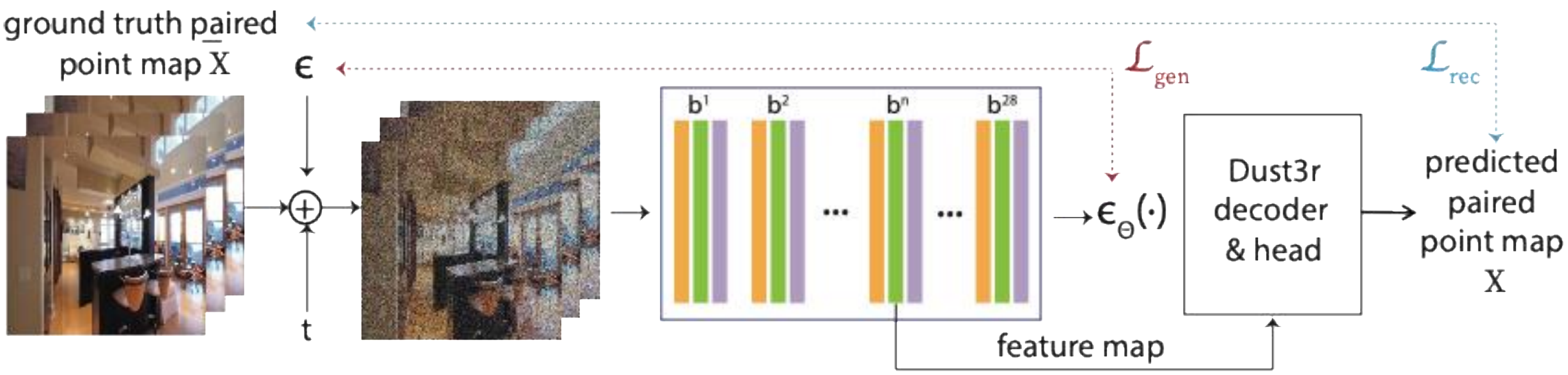
✓ Estimate camera poses

× Do not generate videos

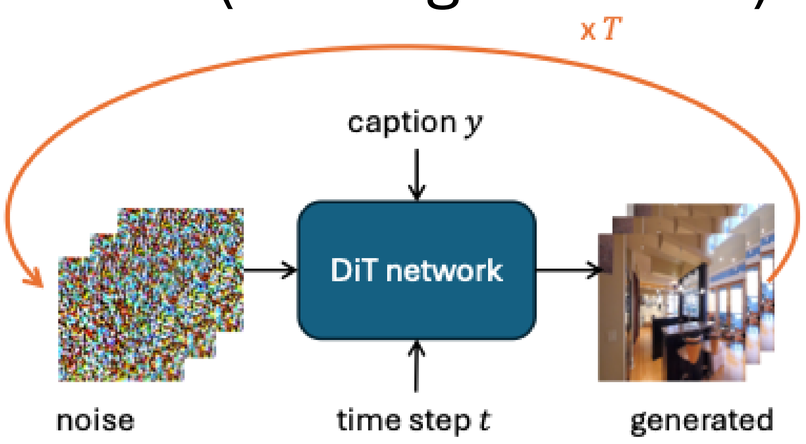
JOG3R == JOint Generation and 3d Reconstruction

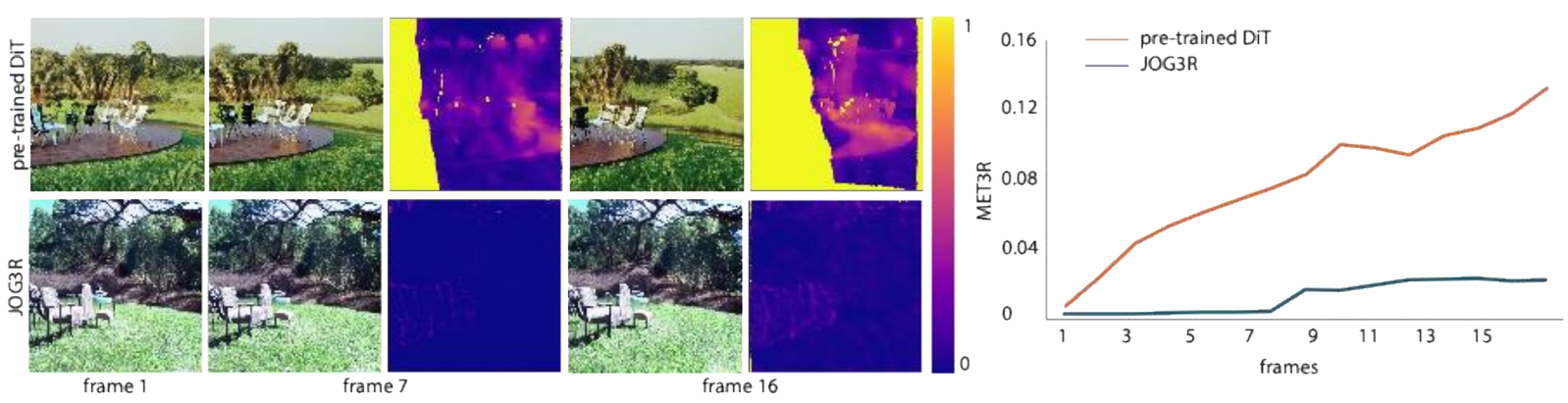
✓ Generate photoreal videos ✓ Estimate camera poses ✓ Estimate 3D point maps

Architecture

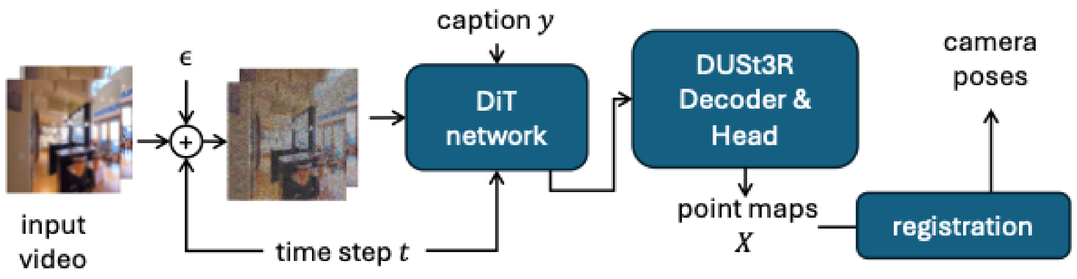


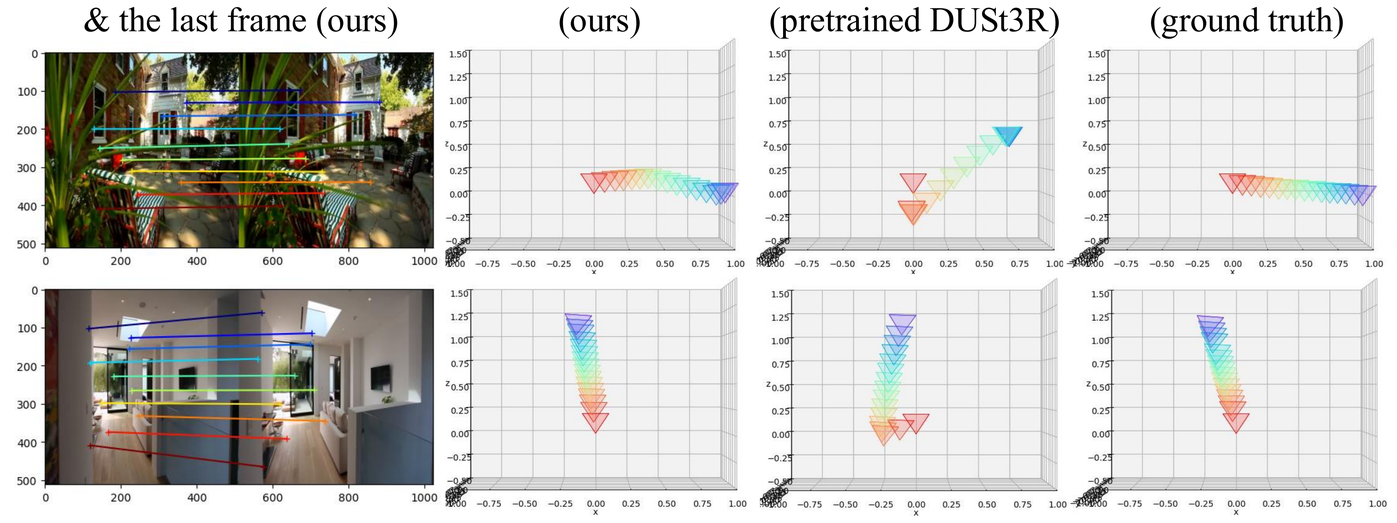
1. T2V (Video generation)





2. V2C (Camera pose estimation)





3a. T2V → V2C

3b. T2V+C (Joint generation)