LIVE: Long-horizon Interactive Video World Modeling

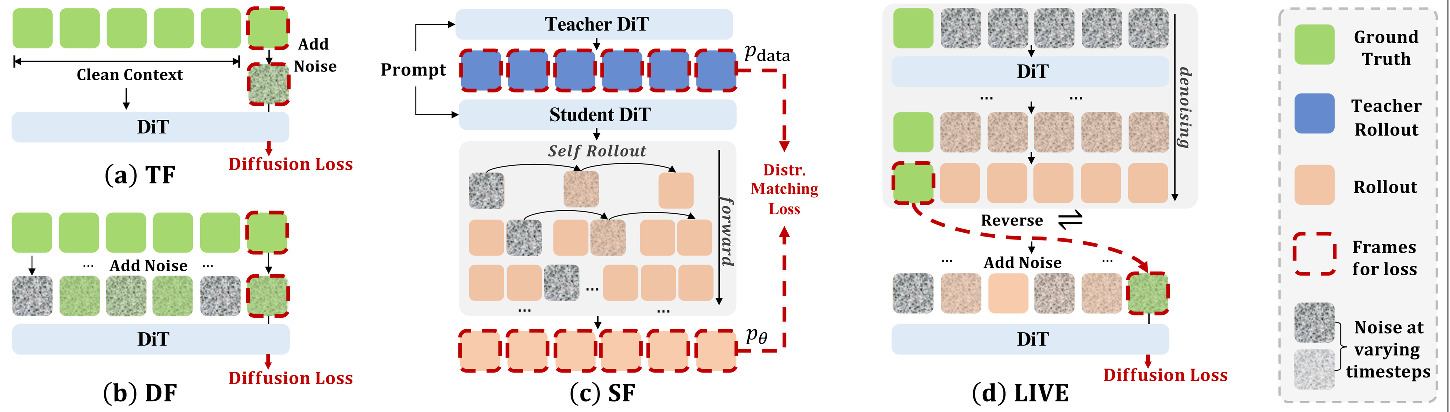
Junchao Huang[1,2,3], Ziyang Ye[1], Xinting Hu[4]

Tianyu He[3], Guiyu Zhang[1], Shaoshuai Shi[5], Jiang Bian[3], Li Jiang[1,2]

[1] CUHK-SZ [2] SLAI [3] Microsoft Research [4] HKU [5] DiDi Chuxing

Microsoft DiDi ICML International Conference On Machine Learning

Introduction



Teacher Forcing (TF)

- Training: use 100% GT as context.
- Inference: use the model's own predicted frames as context.

—> train–test distribution mismatch.

Diffusion Forcing (DF)

- Training: add noise to the GT context to simulate imperfect conditioning.
- noisy GT ≠ real rollout

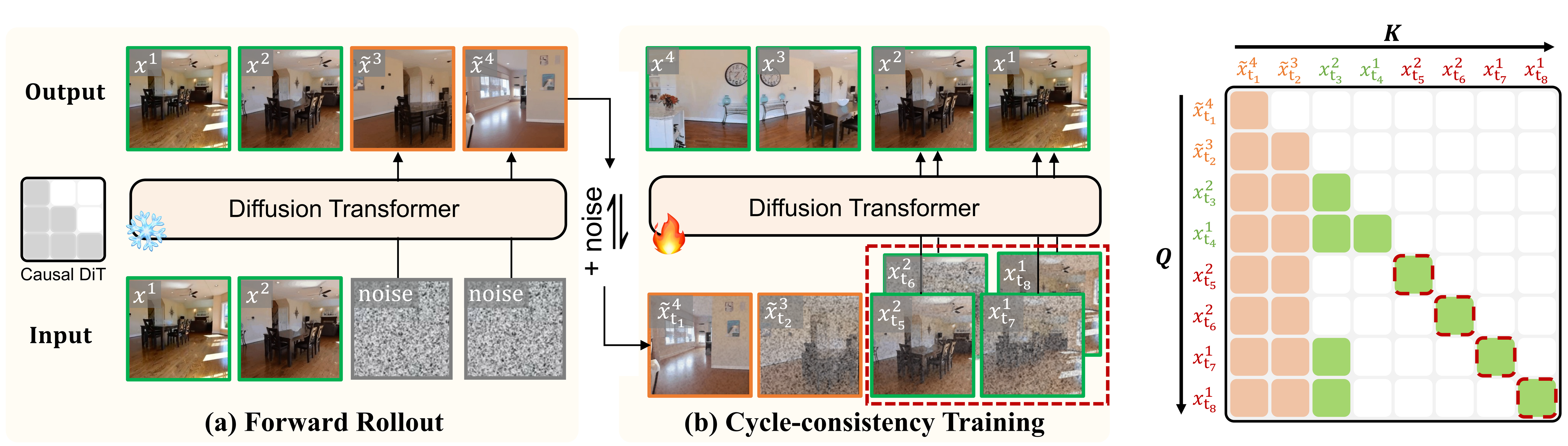
Self Forcing (SF)

- Training: use a teacher model to distill sequence-level distribution matching.
- This requires a pretrained interactive teacher model.

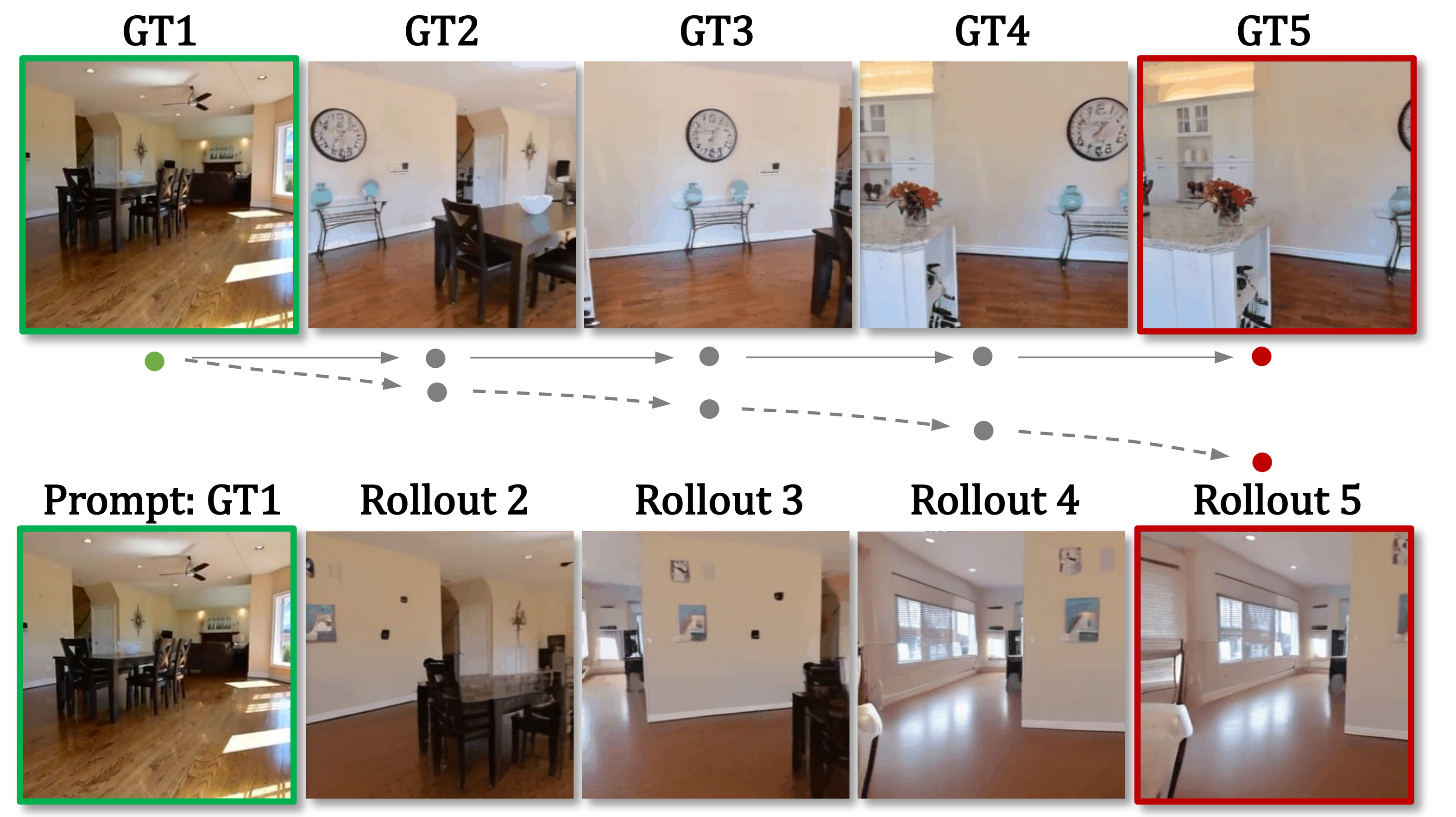
LIVE Key Insight: Cycle Consistency

- We require the model to use its own real rollout as context, and then recover the initial frame through reverse actions. In this way, the model learns a process in which the quality gradually improves, while the context distribution is consistent with that at inference time.
- Minimizing the recovery loss → implicitly constrains the error bound of the forward rollout.

Overview of LIVE



Challenges



Experiments

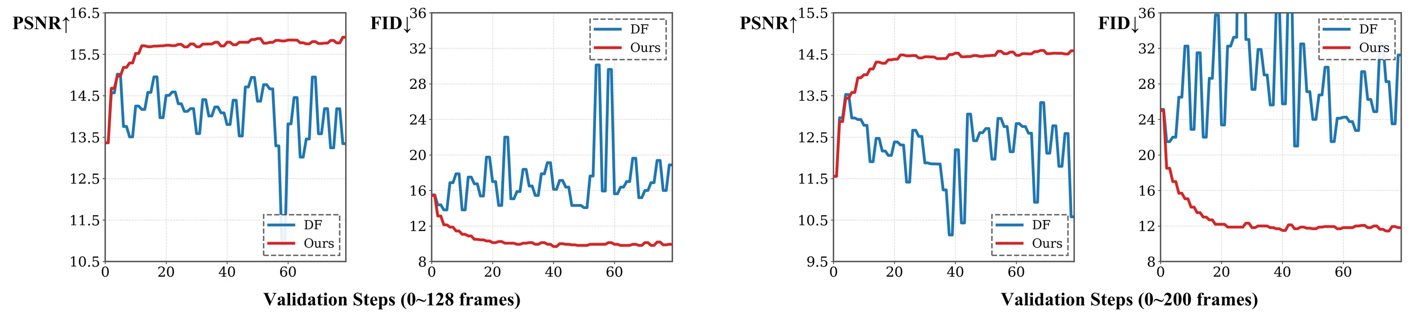


Figure 6. Post-training performance from a converged DF checkpoint. Continued DF training stagnates with oscillating metrics, while LIVE achieves substantial improvements that amplify at longer horizons. LIVE converges to comparable FID across 128-frame and 200-frame generation, demonstrating uniform quality regardless of rollout length.

Table 1. RealEstate10K full test set results across different rollout lengths. LIVE achieves state-of-the-art performance, with particularly large gains at longer sequences demonstrating superior long-horizon generation capability.

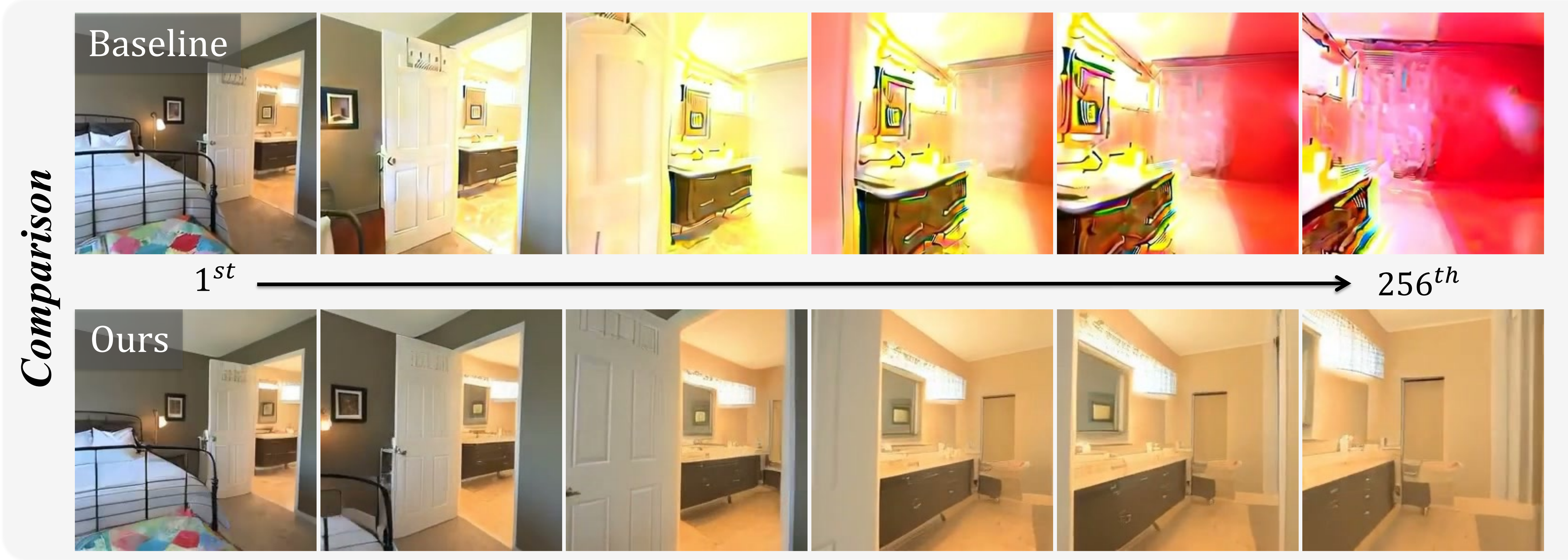
Method	Real-time	0~64 frames			0~128 frames			0~200 frames			≥256 frames		
		PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
CameraCtrl (He et al., 2024)	×	14.09	0.3829	0.4366	11.69	0.5224	0.3651	10.25	0.6115	0.3181	9.48	0.6585	0.2886
DFoT (Song et al., 2025)	×	15.65	0.3053	0.4989	12.55	0.4601	0.3936	10.86	0.5613	0.3287	10.02	0.6128	0.2921
GF (Wu et al., 2025)	×	16.37	0.2450	0.5567	12.69	0.4190	0.4534	10.59	0.5400	0.3969	9.91	0.5936	0.3796
NFD-TF (Cheng et al., 2025)	✓	16.87	0.2571	0.5503	13.59	0.4302	0.4448	11.63	0.5526	0.3724	10.58	0.6222	0.3281
NFD-DF (Cheng et al., 2025)	✓	16.59	0.2558	0.5723	13.82	0.3922	0.5015	12.21	0.4956	0.4598	11.51	0.5506	0.4397
LIVE	✓	**18.11**	**0.2215**	**0.5810**	**15.91**	**0.3298**	**0.5096**	**14.57**	**0.4163**	**0.4630**	**13.89**	**0.4682**	**0.4400**

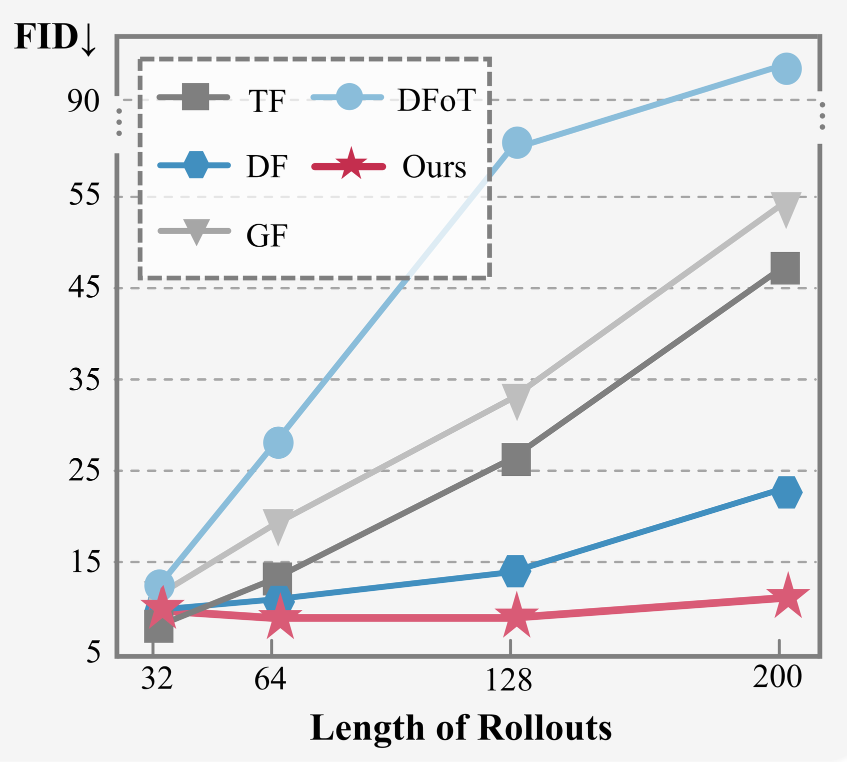
Table 2. Results on interactive game environments. LIVE achieves consistent improvements over baselines on both realistic game engine videos (UE Engine) and interactive gameplay (Minecraft), demonstrating strong performance for interactive world modeling.

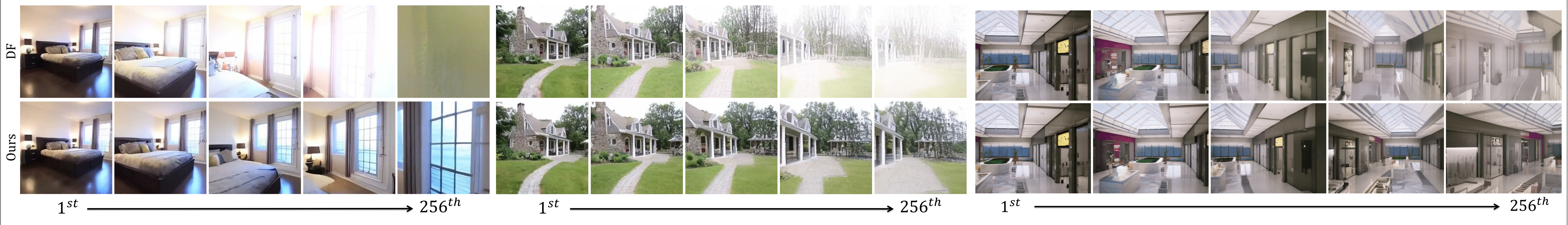
Method	0~64 frames			0~128 frames			0~256 frames			≥400 frames		
	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
UE Engine (Realistic Game Engine)												
NFD-TF (Cheng et al., 2025)	17.16	0.3387	0.4953	14.95	0.4597	0.4245	12.97	0.5702	0.3625	11.80	0.6318	0.3286
NFD-DF (Cheng et al., 2025)	17.15	0.3357	0.5062	14.71	0.4586	0.4441	12.27	0.5799	0.3956	11.02	0.6456	0.3760
LIVE	**17.83**	**0.3145**	**0.5204**	**15.85**	**0.4210**	**0.4600**	**14.04**	**0.5214**	**0.4085**	**12.96**	**0.5794**	**0.3834**

Method	0~32 frames			0~64 frames			0~128 frames			0~200 frames		
	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
Minecraft (Interactive Gameplay)												
NFD-TF (Cheng et al., 2025)	16.09	0.3474	0.6224	14.62	0.4067	0.5930	13.06	0.4781	0.5560	12.10	0.5255	0.5311
NFD-DF (Cheng et al., 2025)	17.39	0.2888	0.6401	15.54	0.3586	0.6036	13.52	0.4469	0.5594	12.34	0.5091	0.5332
LIVE	**17.87**	**0.2698**	**0.6558**	**16.31**	**0.3271**	**0.6291**	**14.90**	**0.3877**	**0.6037**	**14.02**	**0.4299**	**0.5885**

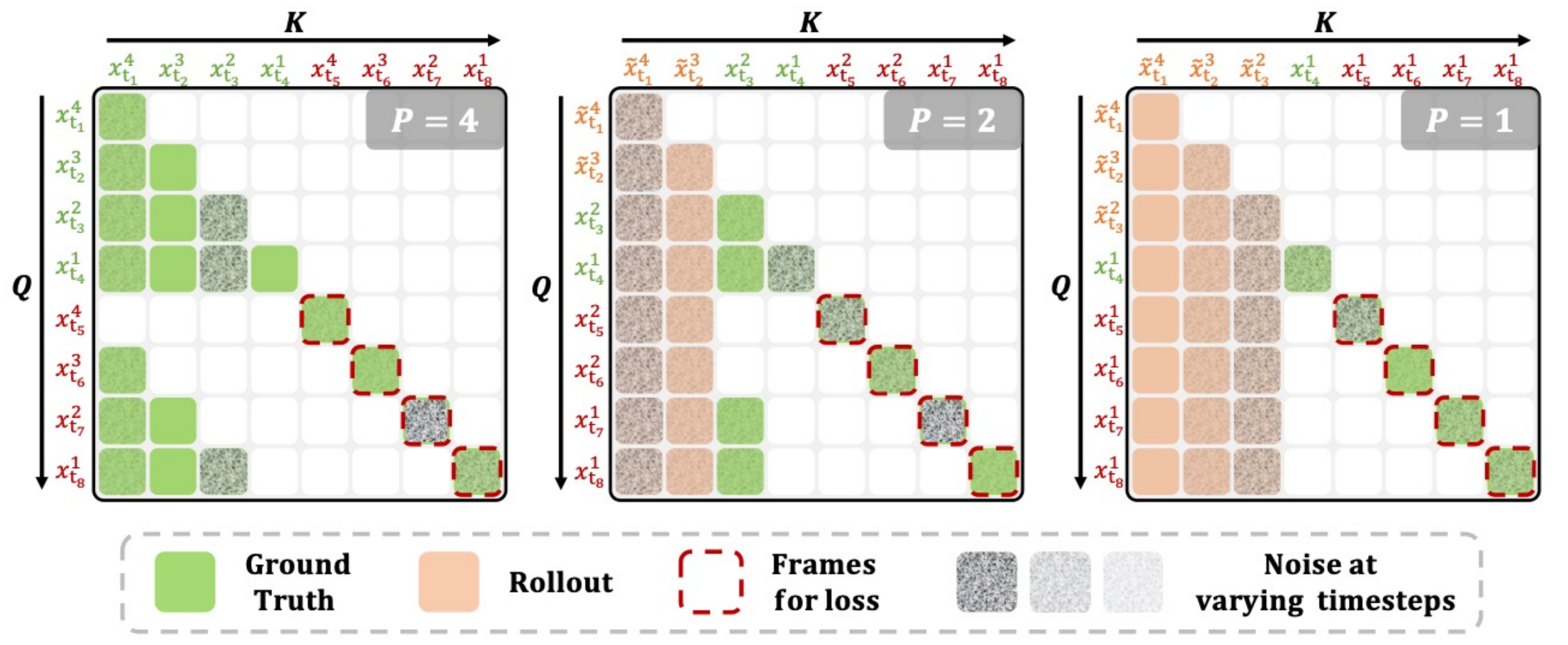
Stable Long-Horizon Rollouts







Progressive Curriculum



Website & Contact

Project Web HomePage Wechat