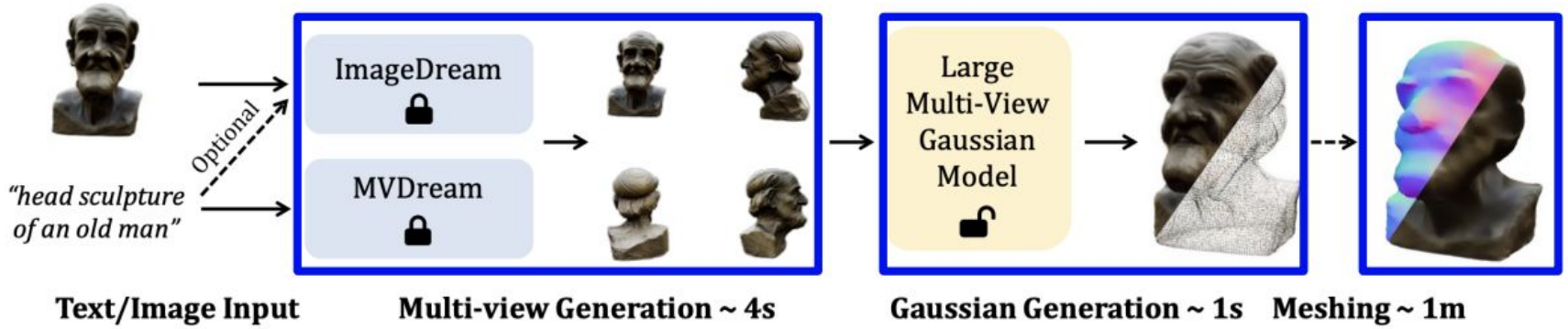
Team II
Paper Presentation II

MultiDiff: Consistent Novel View Synthesis from a Single Image (CVPR 2024)

Asiman Ziyaddinov, Jinhyuk Jang, Prin Phunyaphibarn

Previous talk: Team 3
LGM: Large Multi-View Gaussian Model for High -Resolution 3D Content Creation

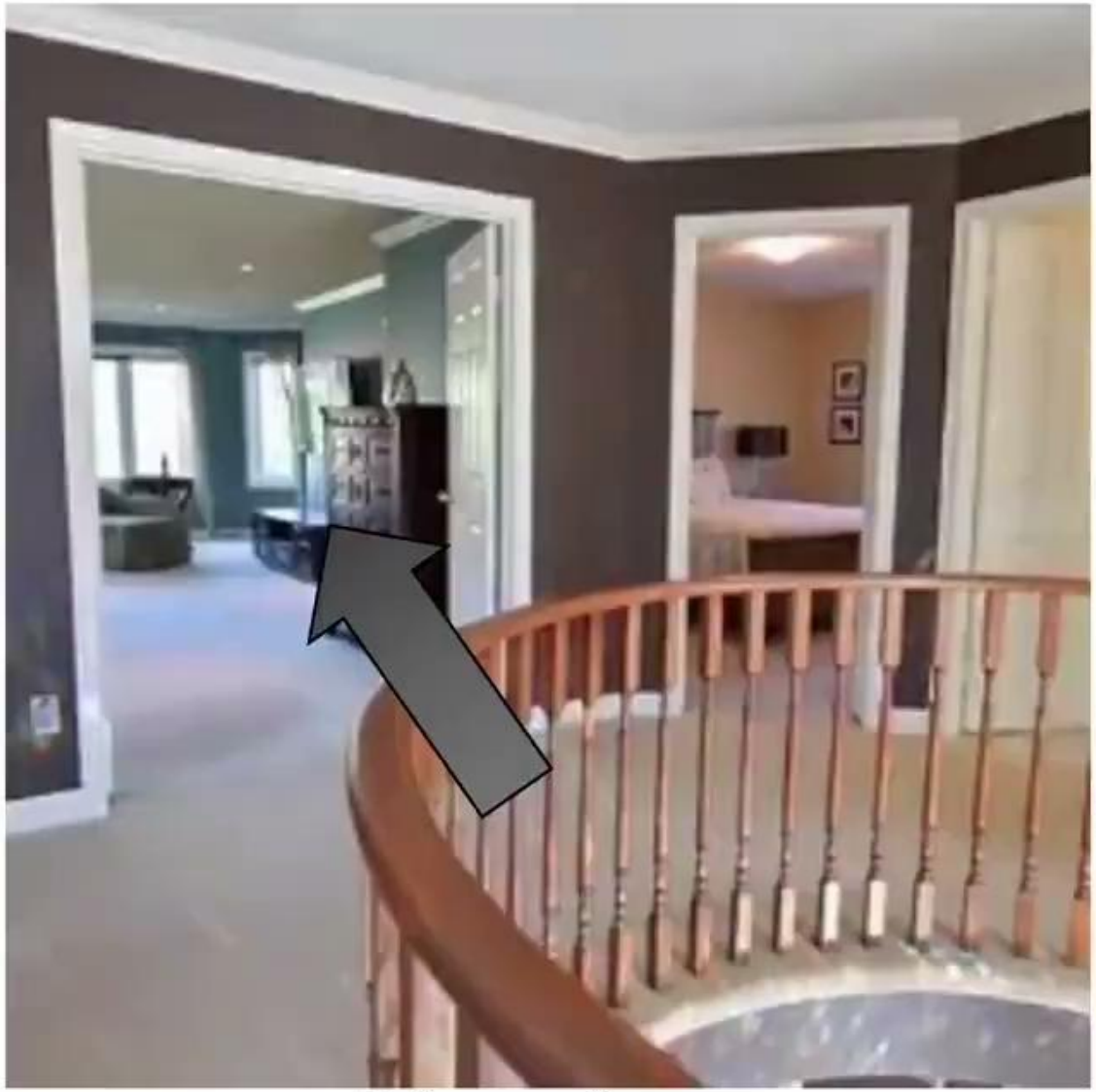


Multi-view Generation Creates images of multiple views

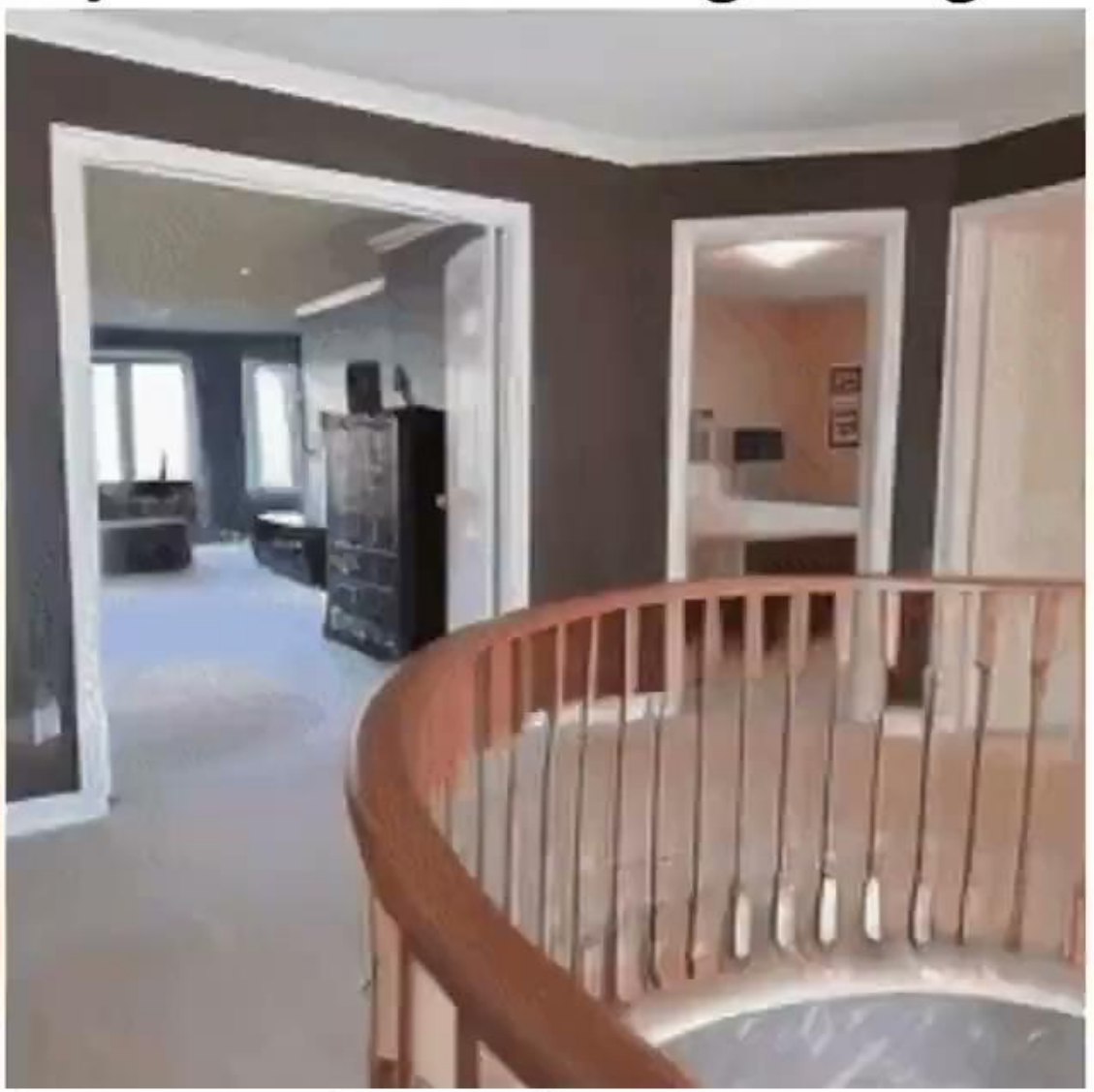
Gaussian Generation Create Gaussian from multiple view pixels

Mesh Extraction Convert 3D Gaussian into polygons

MultiDiff: Consistent Novel View Synthesis from a Single Image



Reference image



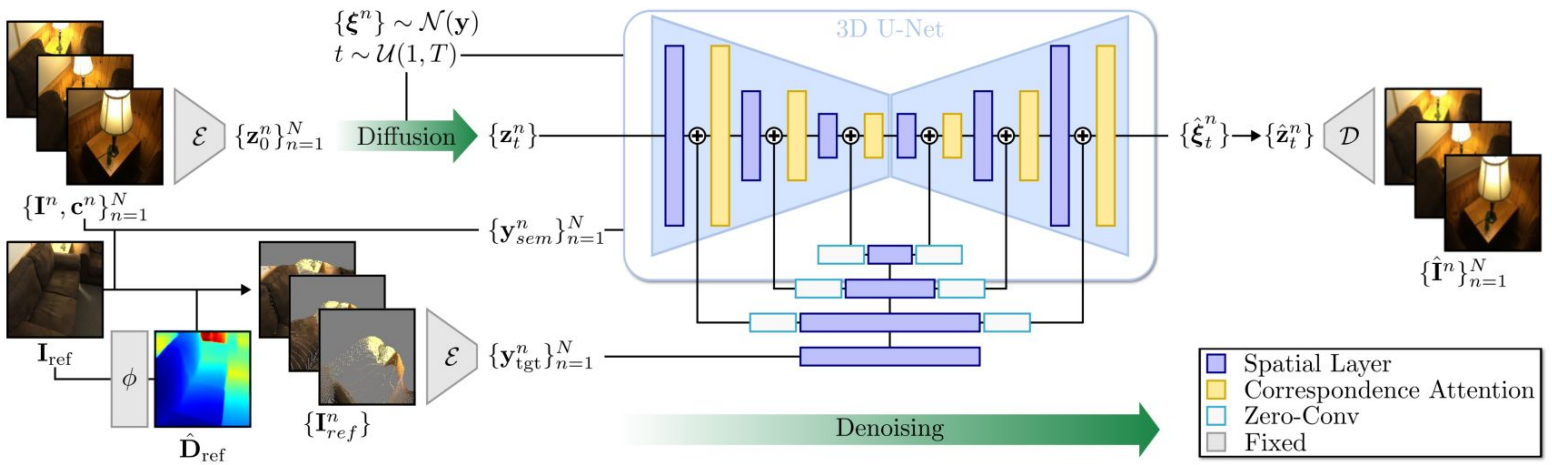
Generated sequence

The Problem: Novel View Synthesis

- What is novel view synthesis?
- Why is synthesizing consistent views from a single image challenging?
- Key challenges: Depth ambiguity, coherence across views, and limited input.

MultiDiff: The Proposed Approach

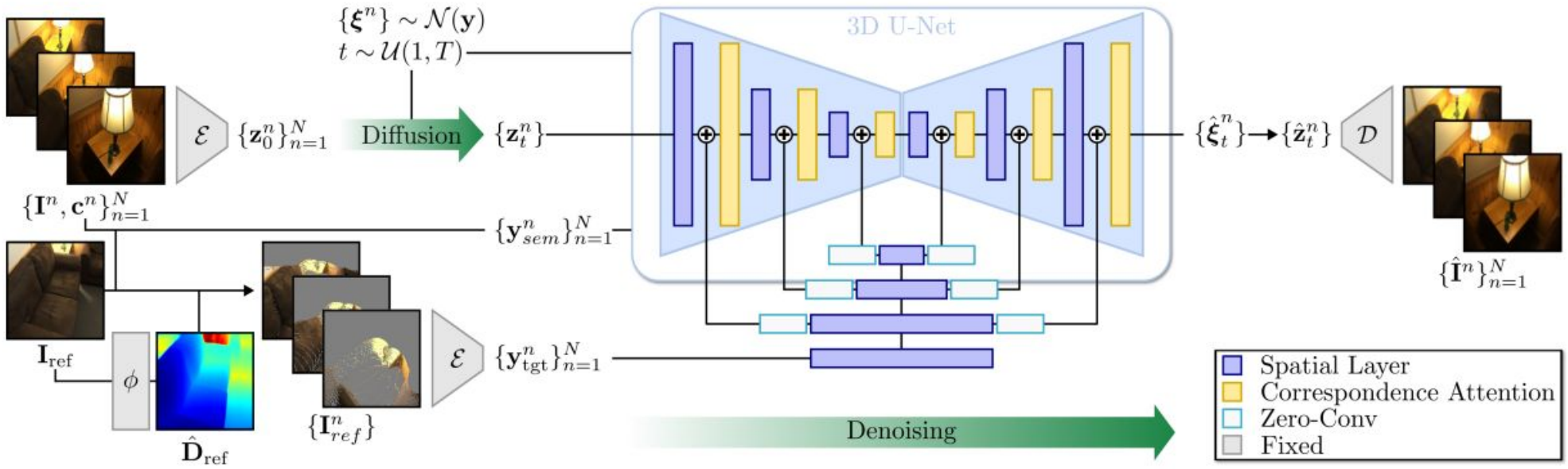
- Uses **diffusion models** to achieve consistent view synthesis.
- Training using **structured noise** to ensure coherence.
- Key advantages: Single-image input, consistent across viewpoints, and editable.



How Diffusion Models Enable View Synthesis

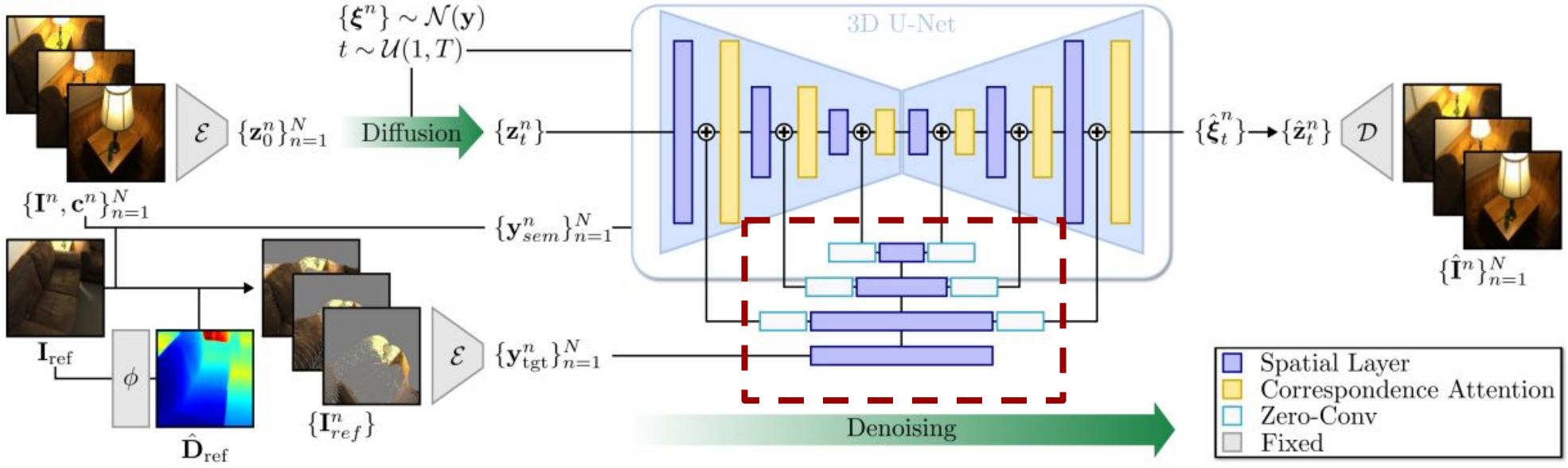
1. **Pretrained** Diffusion Model (Image Prior)
2. **Video** Diffusion Model (Temporal Prior)
3. **Geometric (depth) priors** maintain depth and perspective.
4. **Structured noise** is used to generate consistent views.

Fine-tuning Video Diffusion Models



Fine-tuning keeps strong video and image priors from the pretrained diffusion model

Fine-tuning Video Diffusion Models



ControlNet-like
Adapter

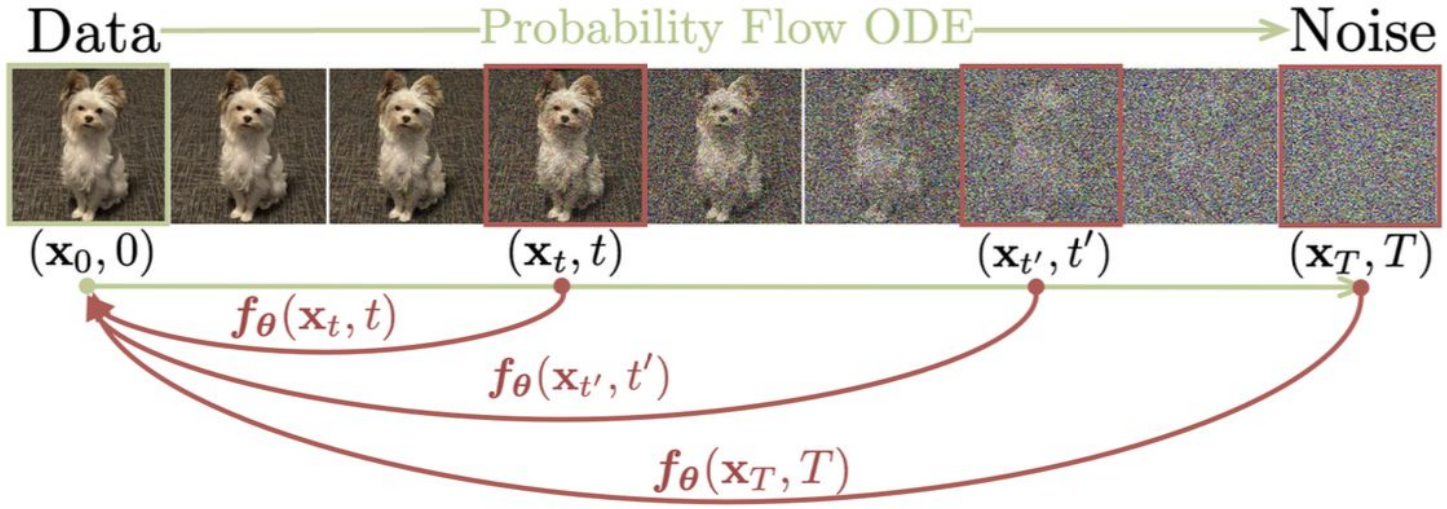
The ControlNet-like Adapter adds **<u>images</u> warped based on depth prediction** for stronger **geometric** guidance

Key Observation: Correlated Noise

Noise should be correlated across frames (warped based on depth map) to depict the 3D correspondences in the 3D scene

Chang, Pascal, et al. "How I Warped Your Noise: a Temporally-Correlated Noise Prior for Diffusion Models." *The Twelfth International Conference on Learning Representations*. 2024.

Diffusion Model Training



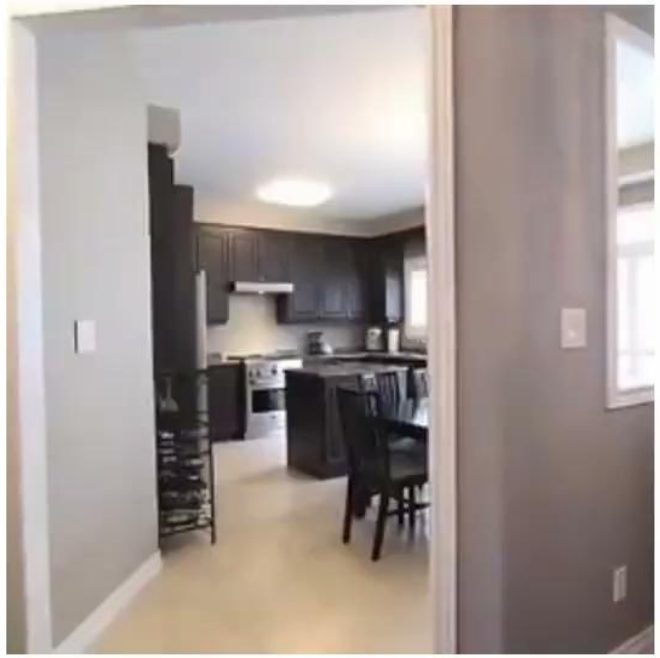
$$\mathbb{E}_{t\sim[1,T],\mathbf{x}_0,\boldsymbol{\epsilon}_t}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t)\|^2\right]$$

Train diffusion models by predicting structured noise

Song, Yang, et al. "Consistency Models." *International Conference on Machine Learning*. PMLR, 2023.
https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Structured Noise

Noise warped based on predicted depth



Reference image

Depth-based reference warp

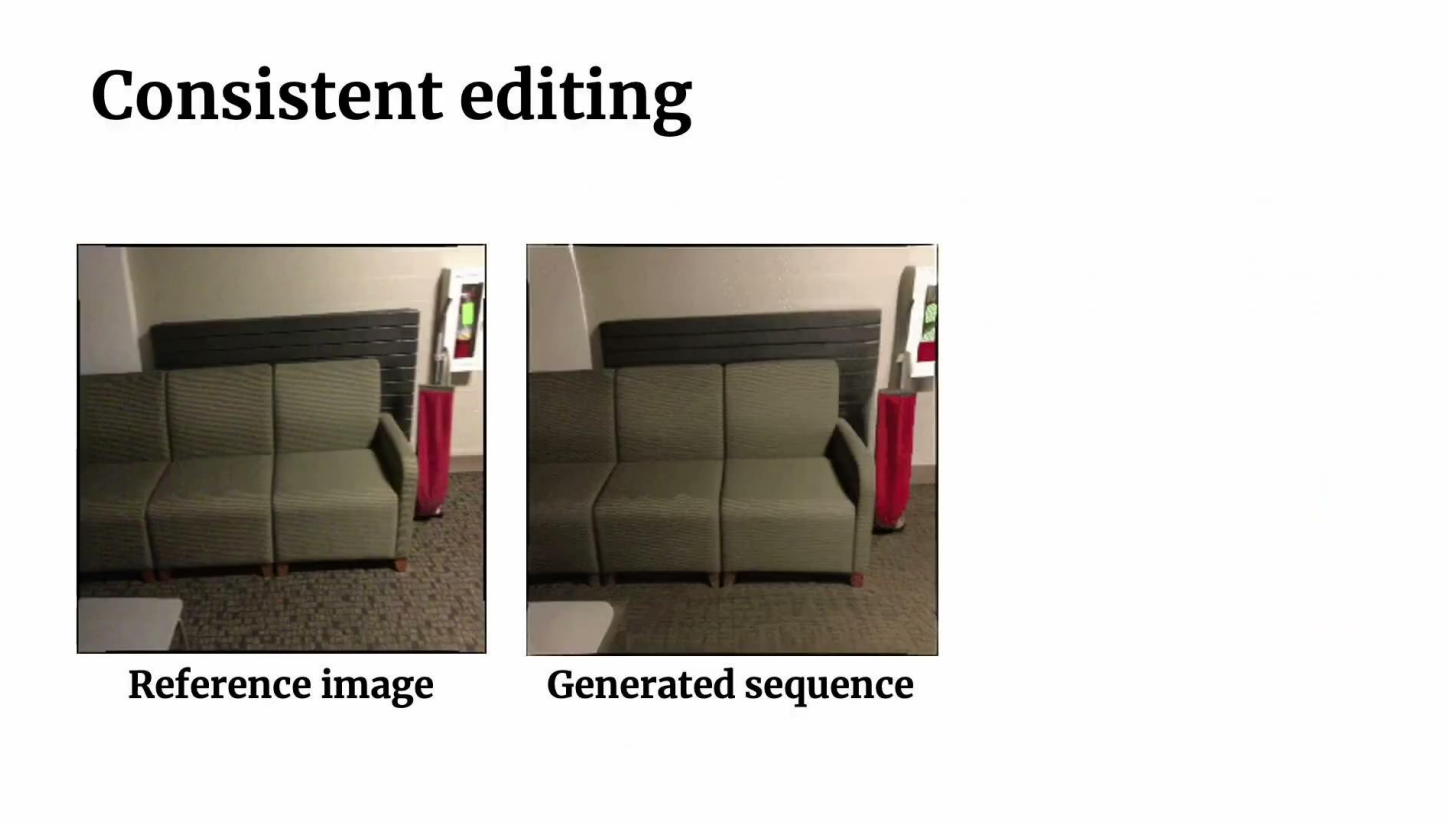
Structured noise

Importance of structured noise

Without structured noise (”MultiDiff w/o SN”), the color of the dining table is not maintained w.r.t. the reference image.

Why MultiDiff is Unique

- Generates coherent outputs from a single input image.
- Consistent view generation across trajectories.
- Editable outputs enable further applications like object manipulation.



Experiments

- Datasets used: **RealEstate10K** (Youtube) & **ScanNet** (1513 handheld captures).

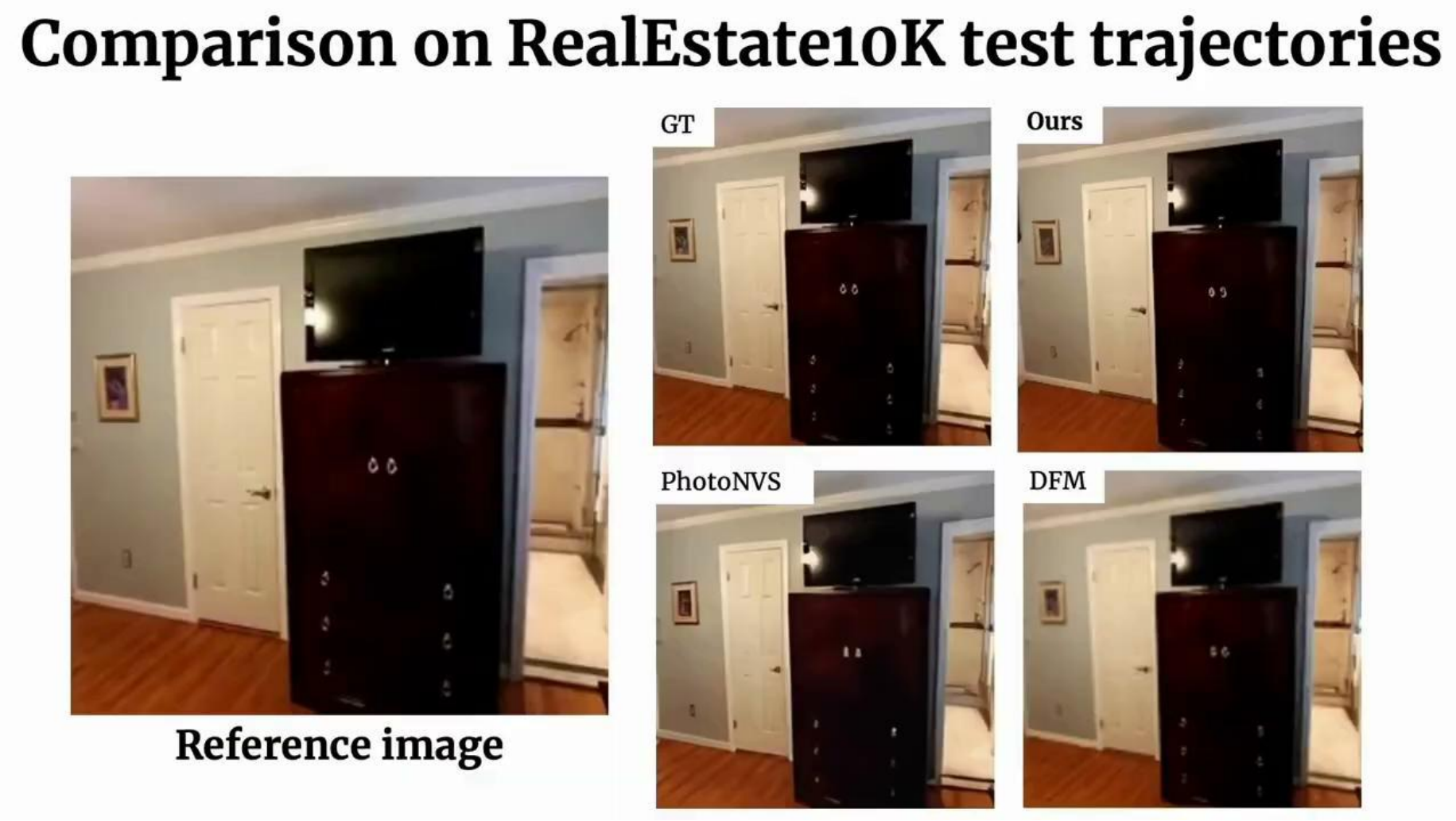
- Metrics evaluated: Fidelity, coherence, and perceptual quality.

Quantitative results

	Method	Short-term				Long-term			
		PSNR ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓	FVD ↓	mTSED ↑
128px	MVDiffusion [68]	13.14	0.439	43.28	0.013	43.58	0.013	186.6	0.506
	DFM [69]	**16.59**	0.444	75.19	0.036	111.9	0.069	167.2	**0.912**
	Text2Room [27]	15.01	0.452	39.87	0.008	82.44	0.0041	173.1	0.812
	PhotoNVS [82]	15.23	0.440	49.19	0.019	75.23	0.038	89.04	0.479
	MultiDiff (Ours) w/o SN	15.29	0.372	40.36	0.008	43.61	0.011	80.71	0.752
	MultiDiff (Ours)	15.50	**0.356**	**38.44**	**0.007**	**42.41**	**0.010**	**74.10**	0.776
256px	MVDiffusion [68]	12.88	0.502	50.18	0.017	51.60	0.018	230.1	0.361
	Text2Room [27]	14.32	0.514	46.69	0.014	93.09	0.058	201.1	**0.631**
	PhotoNVS [82]	14.61	0.542	63.21	0.033	96.85	0.059	134.2	0.263
	MultiDiff (Ours) w/o SN	14.80	0.445	47.10	0.013	50.84	0.016	119.3	0.529
	MultiDiff (Ours)	**15.00**	**0.431**	**43.84**	**0.010**	**47.11**	**0.013**	**114.9**	0.576

Qualitative evaluations

Qualitative evaluations



Conclusion

- **Achievements:** Coherent novel views from a single image, real-world applications.
- **Limitations:** Computationally intensive, dataset-dependent.
- **Future Work:** Faster inference, broader generalization.