DIFFUSIONRENDERER: Neural Inverse and Forward Rendering with Video Diffusion Models

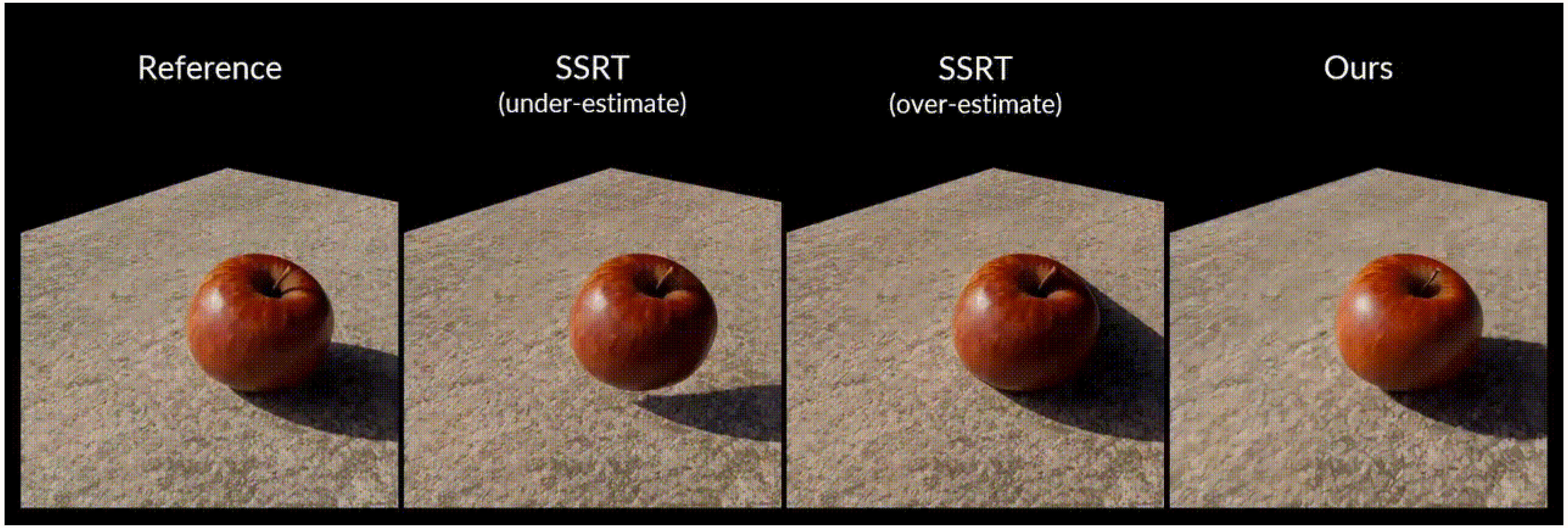
NVIDIA, University of Toronto, Vector Institute,
University of Illinois Urbana-Champaign

CVPR 2025

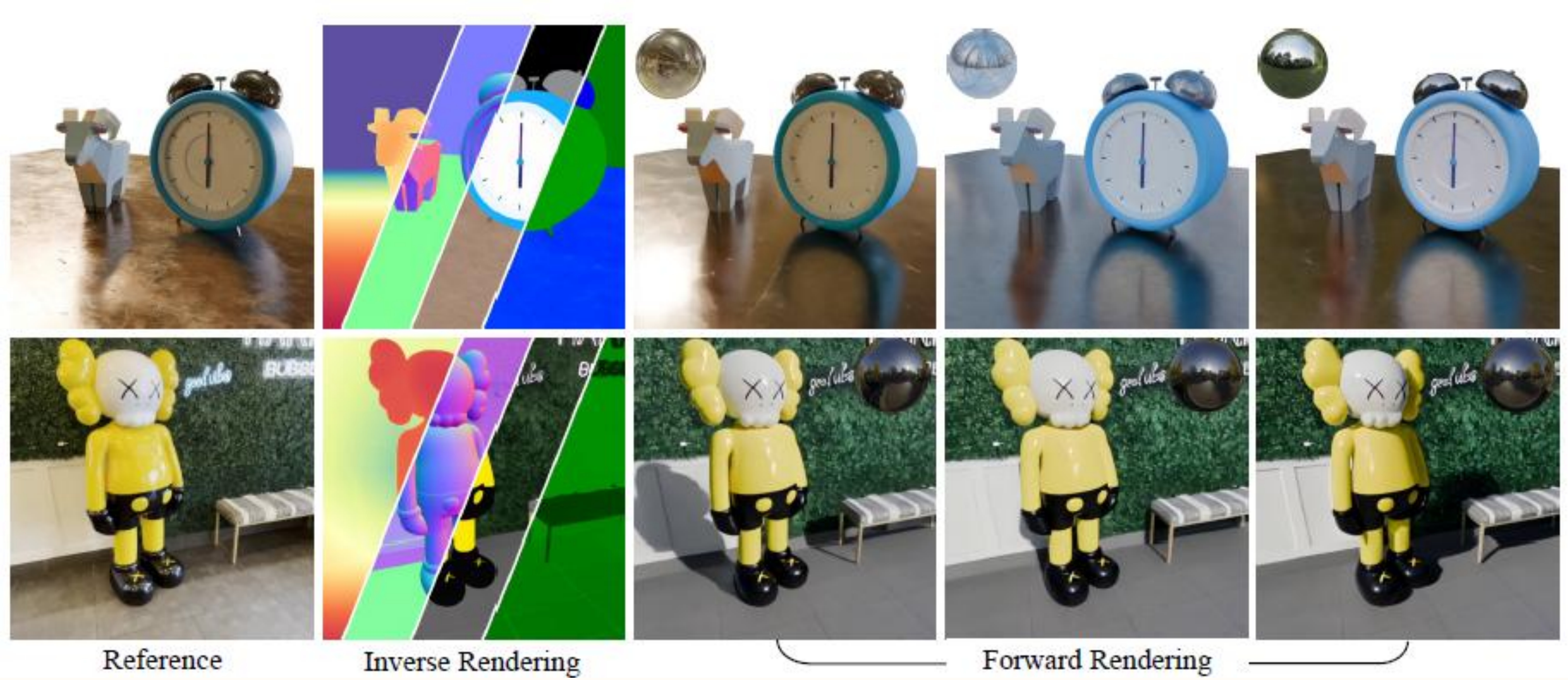
Contents

1. Introduction

- **Motivation**
  - **Classic physically-based rendering (PBR) relies on precise scene representations that are often impractical to obtain in real-world scenarios.**

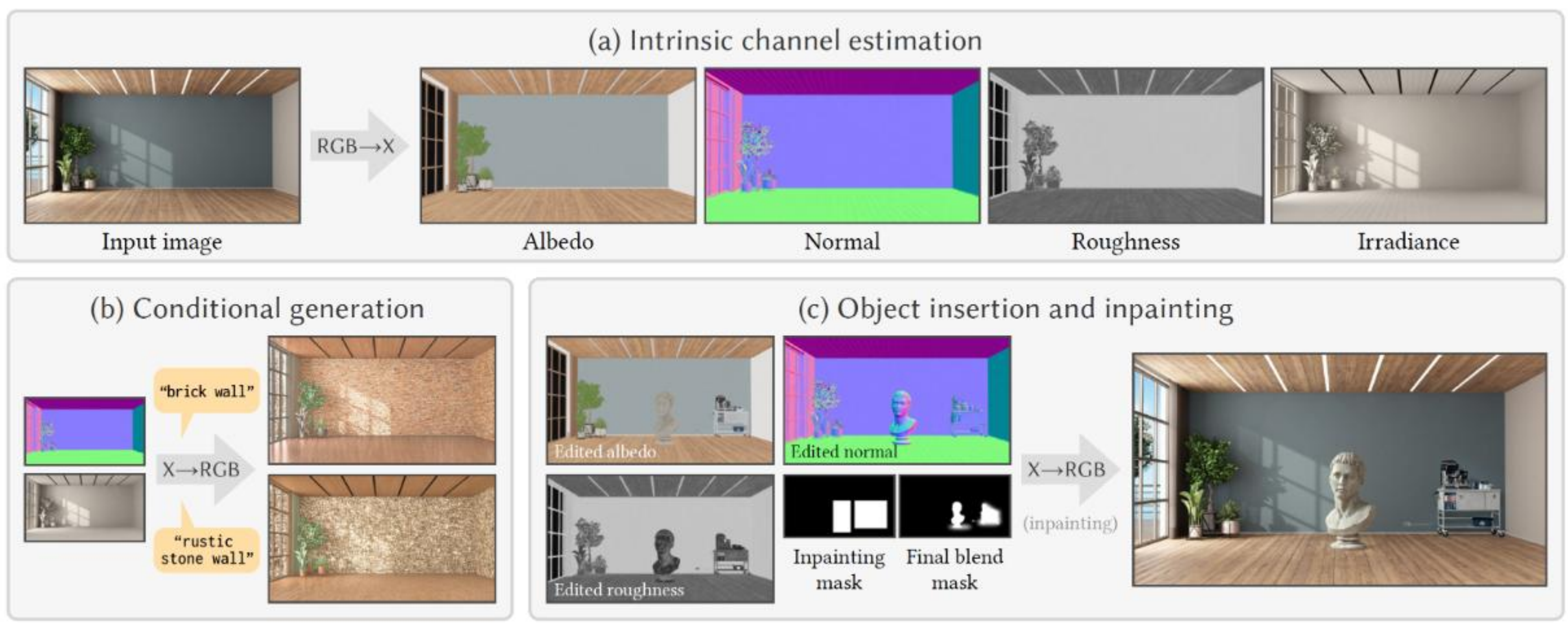


1. Introduction



2. Related Works

- **Neural Rendering (RGB↔X)**
  - **Train image diffusion models to both estimate a G-buffer from an image and to render an image from a G-buffer.**
  - **Limitation: Absence of temporal coherence**



2. Related Works

- **3D reconstruction-based relighting**
  - Relighting via 3D scene reconstruction from multi-view images and explicit inverse rendering to recover material properties.
  - Limitation :

    ① Need to optimize for each scene individually

    ② Quality may be affected by practical issues

2. Related Works

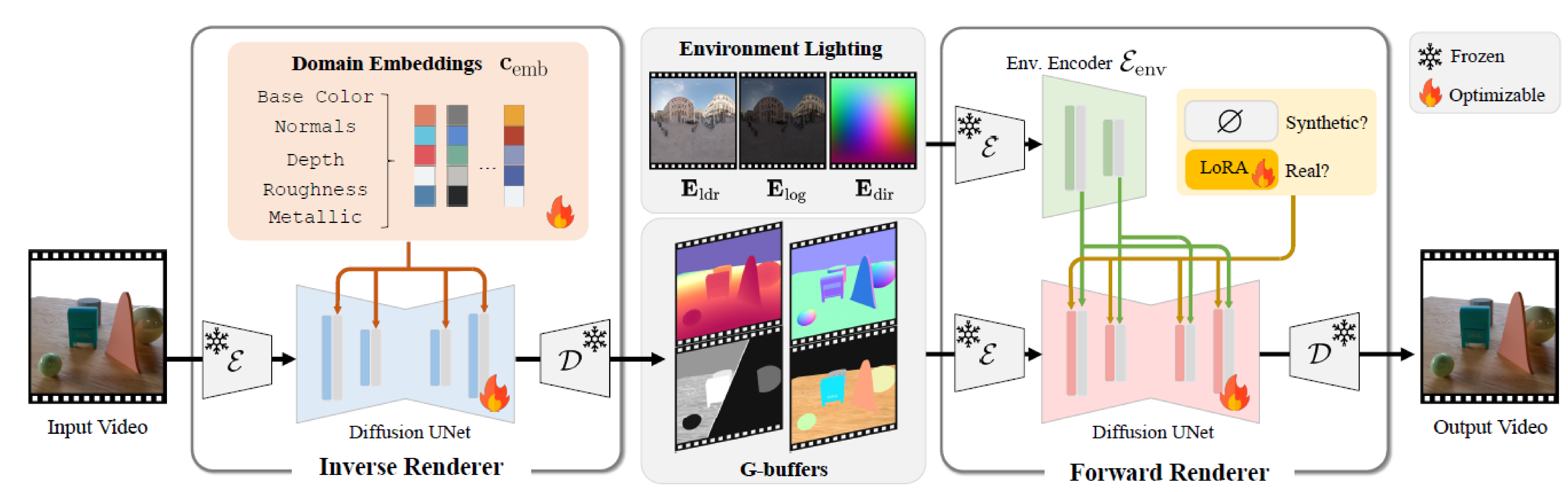
- **Diffusion Models**
  - Instead of calculating light rays like traditional PBR, the diffusion model treats relighting as a conditional image generation problem.
  - Limitation :

    ① Few multi-illumination datasets

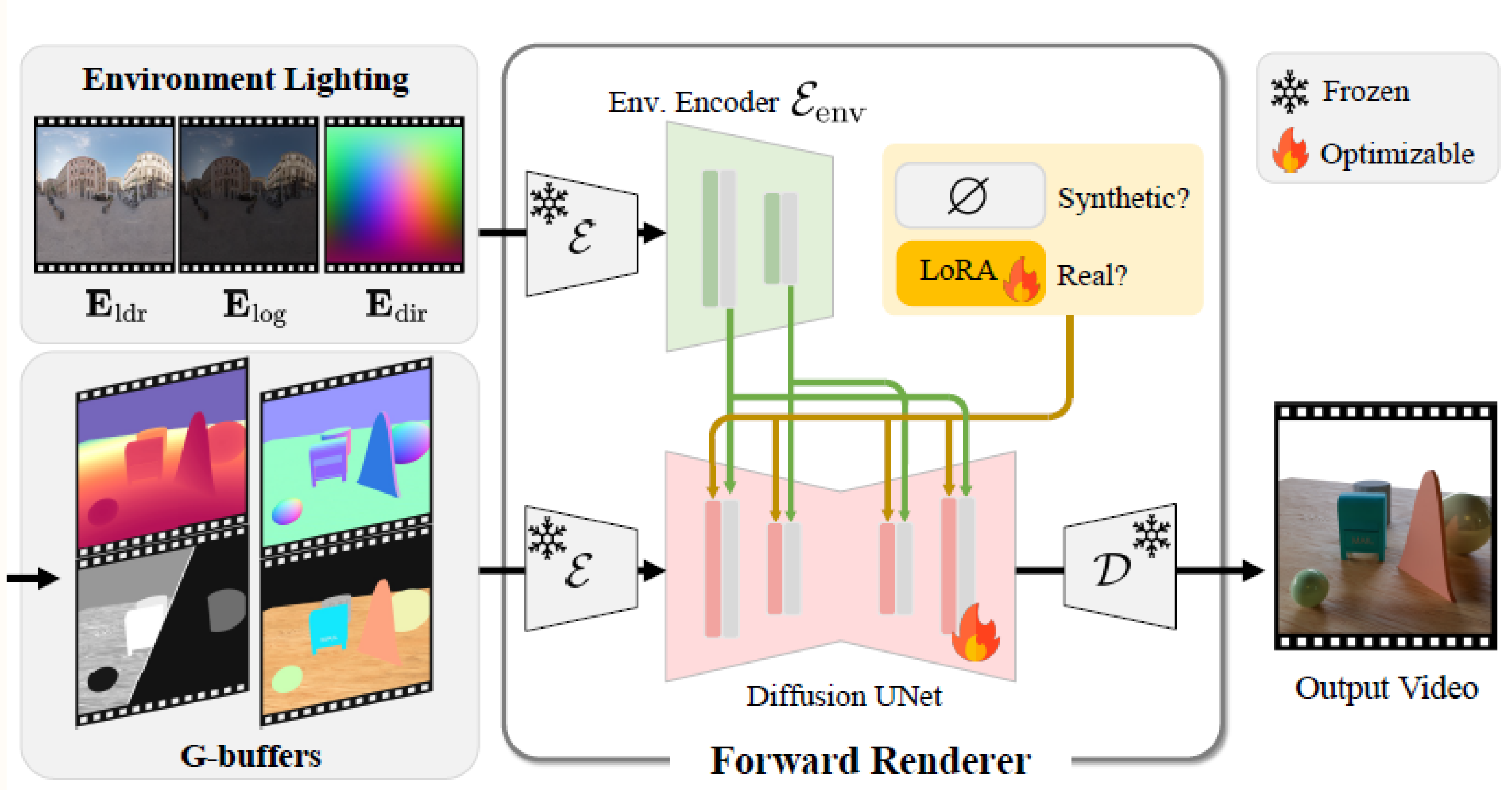
    ② Existing methods are specialized to a domain

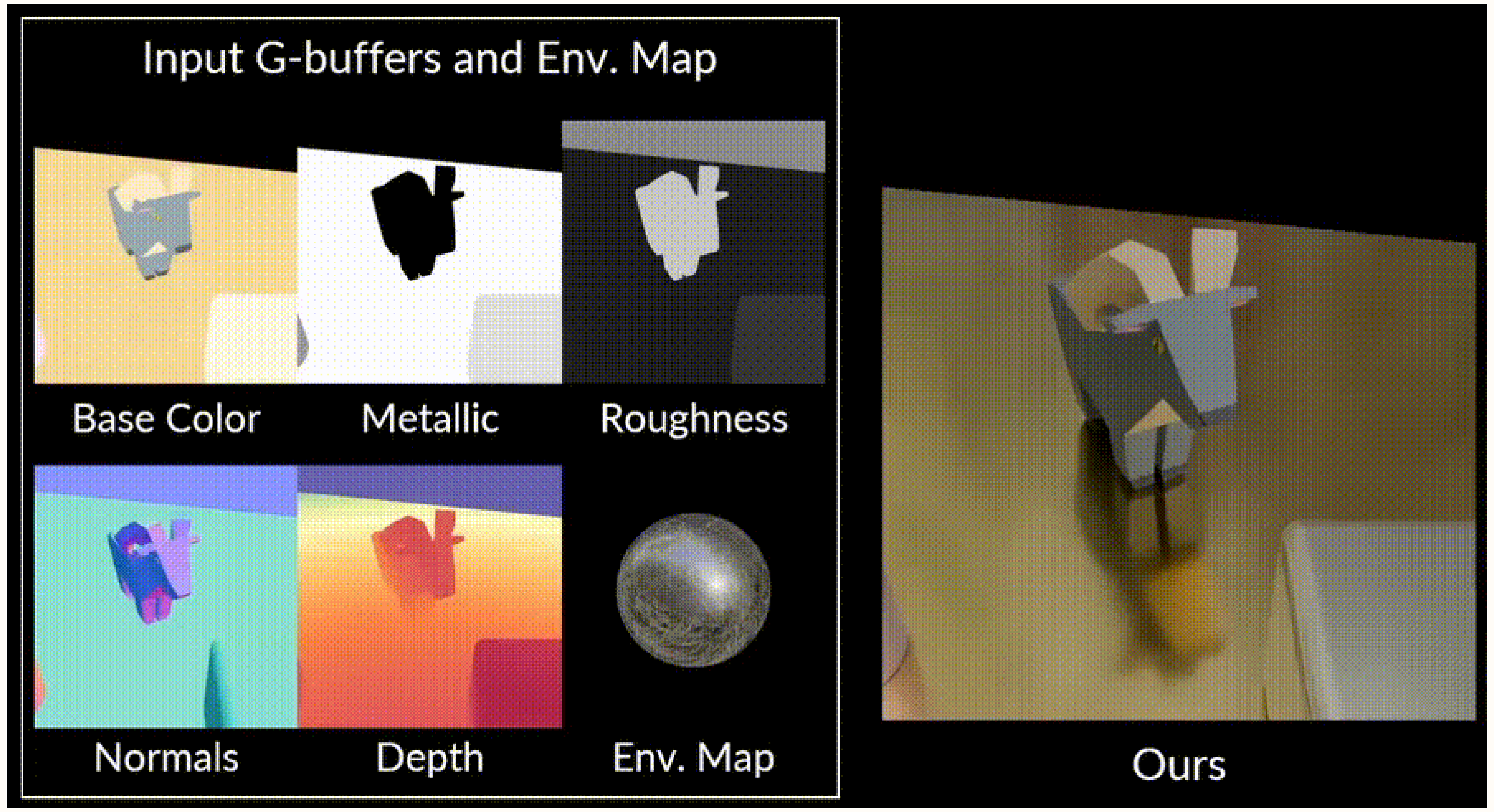
3. Method - Overview

- **Diffusion Renderer**
  ① Neural forward renderer
  ② Neural inverse renderer

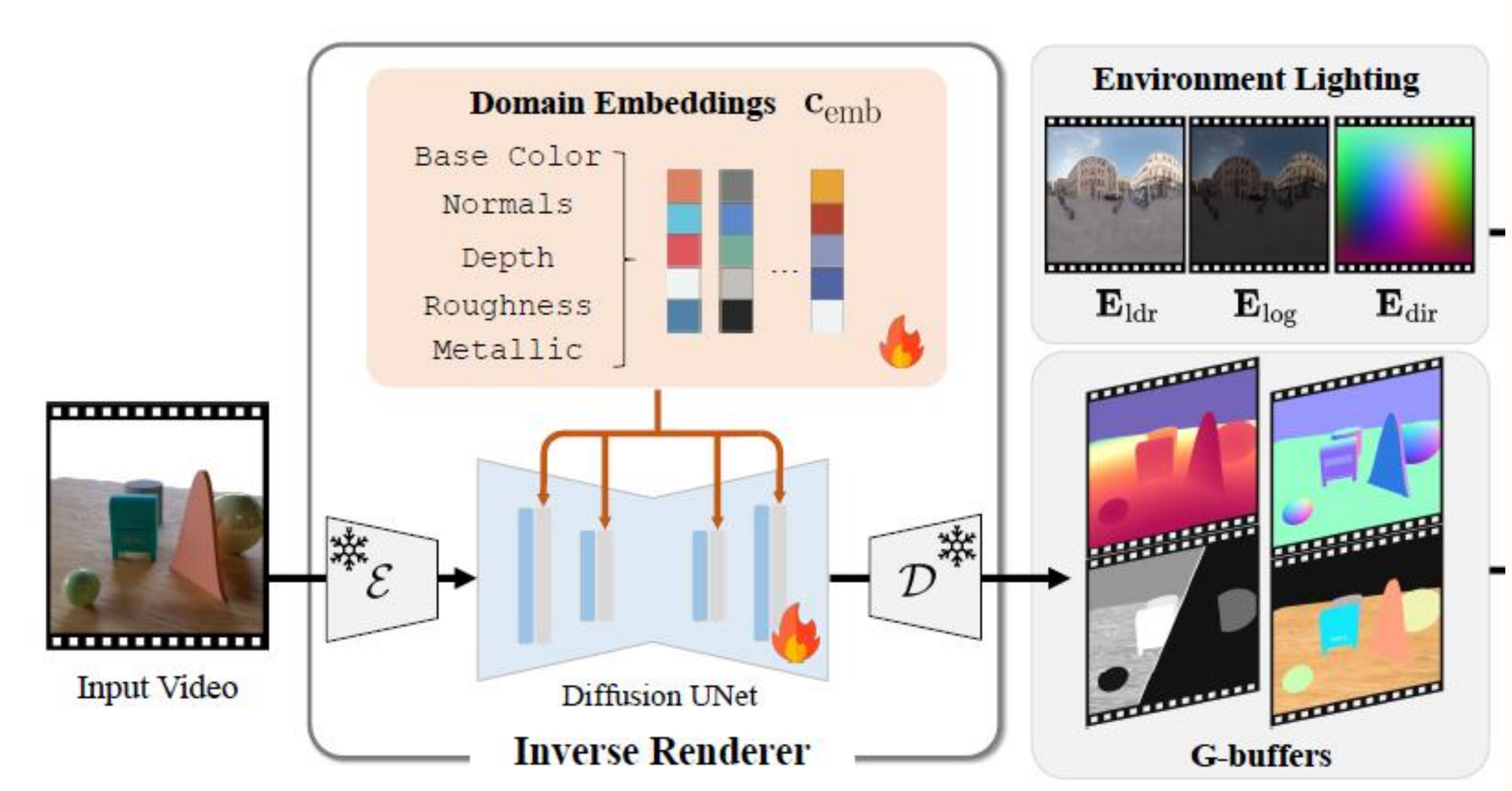


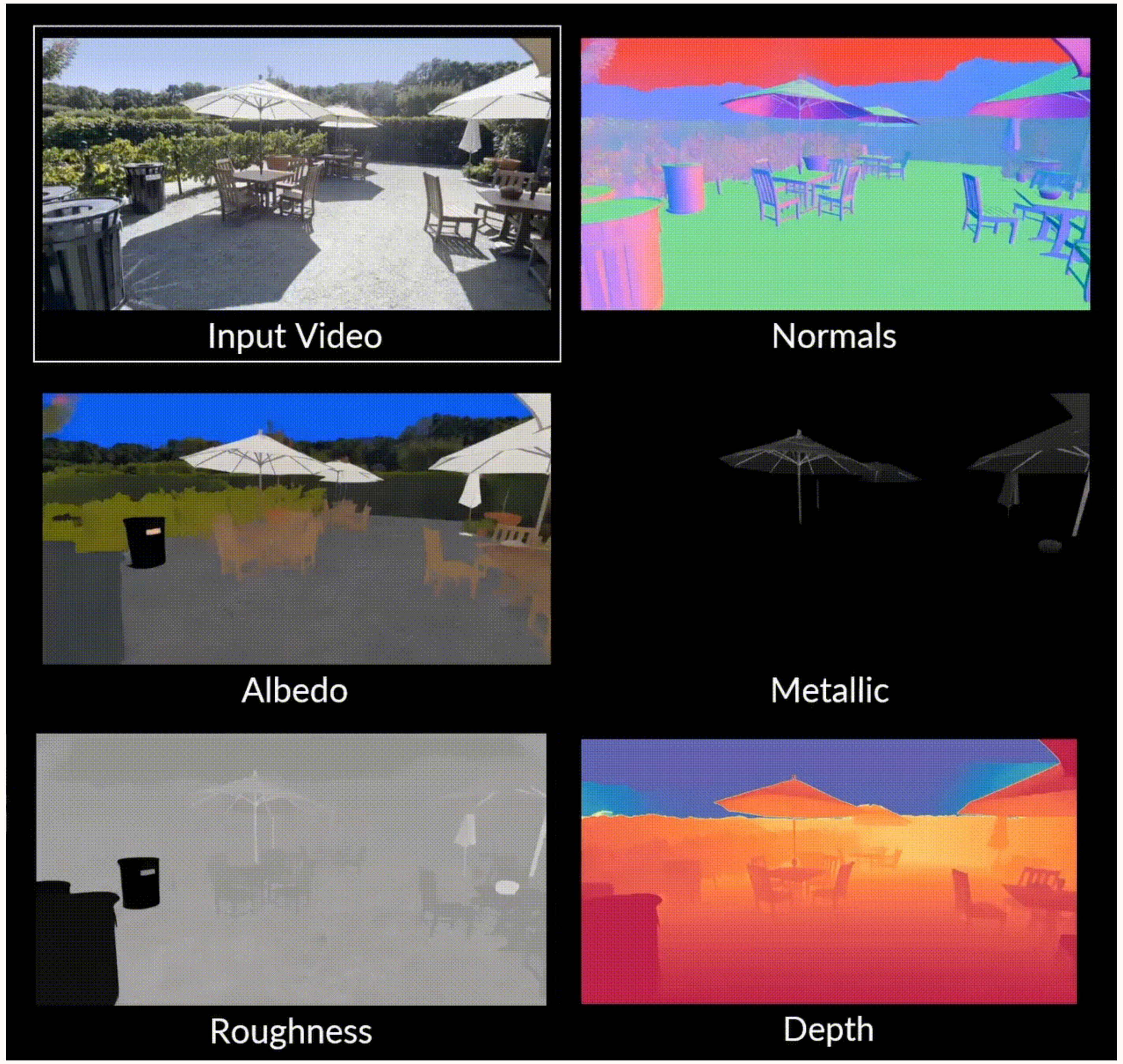
3. Neural Forward Rendering





3. Neural Inverse Rendering





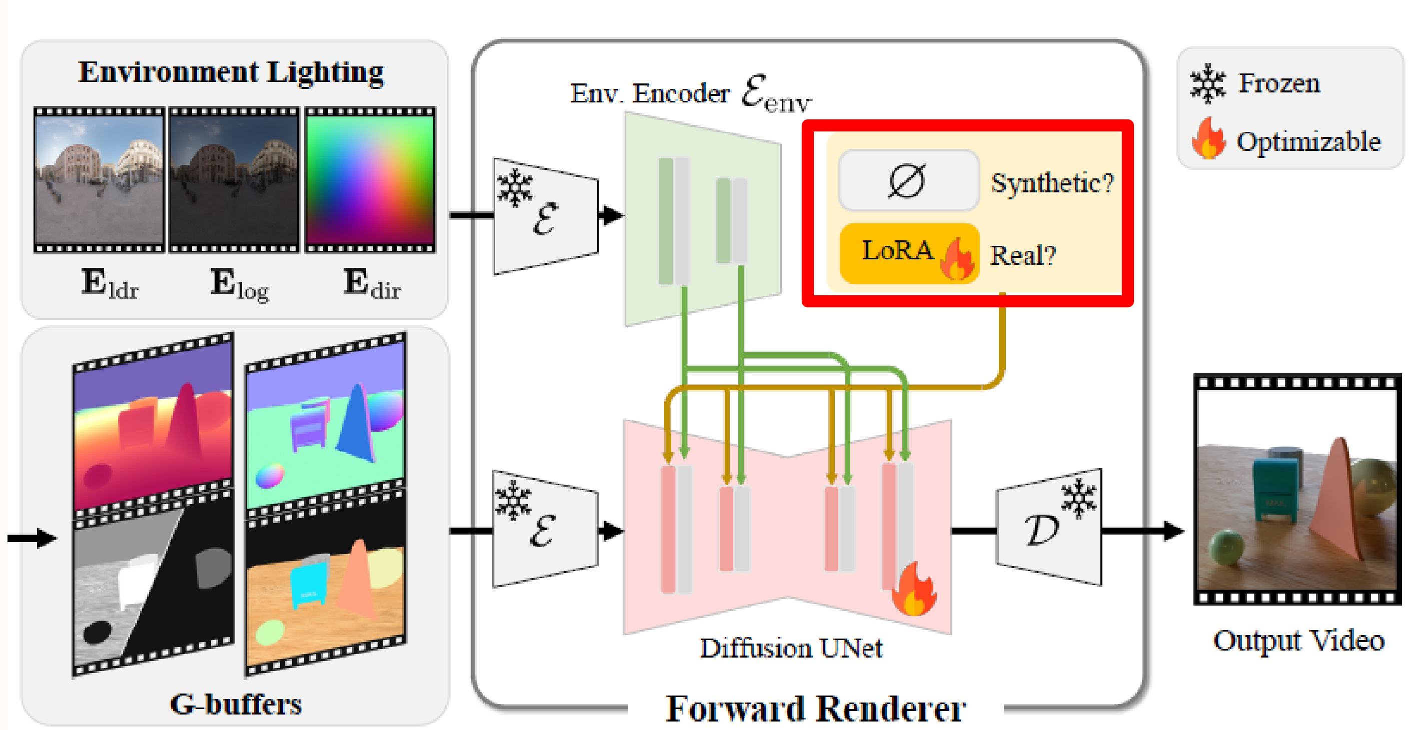
3. Training Dataset

- **Synthetic dataset**
  - Generated 150,000 videos using traditional 3D rendering engines
  - High-quality video data with paired ground-truth for material, geometry, and lighting information

- **Auto-labeling real-world dataset**
  - Use trained inverse rendering model to generate G-buffer labels
  - Use DiffusionLight to estimate environment map

3. Training Pipeline

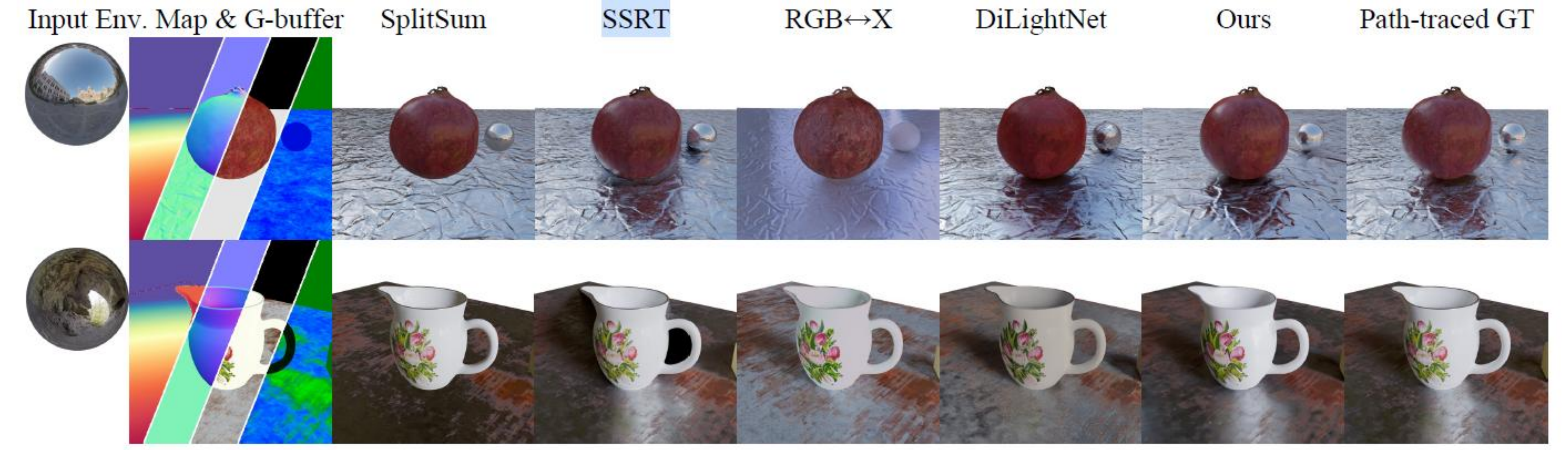
- **Neural inverse renderer training**
  - Using synthetic video dataset and public image intrinsic datasets
  - After training, the inverse renderer can be used for auto-labeling real-world videos
- **Neural forward renderer training**
  - Using synthetic video dataset and auto-labeling real-world videos
  - Integrating LoRA into the training pipeline



4. Experiments

- **Evaluation of neural rendering**

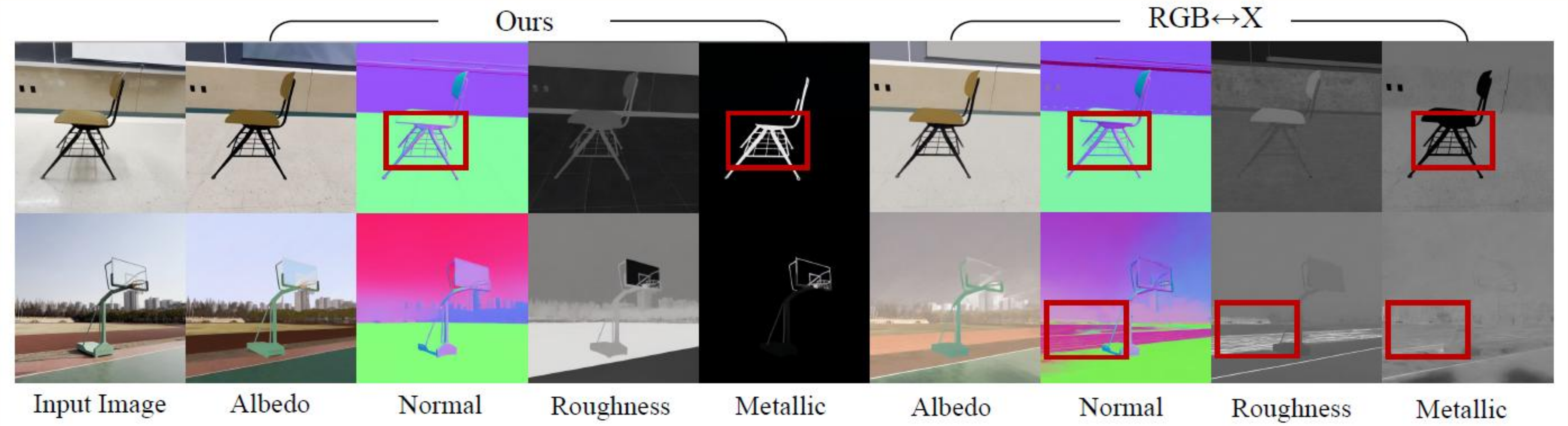
	SyntheticObjects			*SyntheticScenes*		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SSRT	**29.4**	**0.951**	**0.037**	**24.8**	**0.899**	**0.113**
SplitSum [32]	28.7	**0.951**	0.038	23.1	0.883	0.116
RGB↔X [83]	25.2	0.896	0.077	18.5	0.645	0.302
DiLightNet [82]	26.6	0.914	0.067	20.7	0.630	0.300
Ours	28.3	**0.935**	**0.048**	**26.0**	**0.780**	**0.201**



4. Experiments

- **Evaluation of inverse rendering**

	Albedo				Metallic	Roughness	Normals
	PSNR ↑	LPIPS ↓	si-PSNR ↑	si-LPIPS ↓	RMSE ↓	RMSE ↓	Angular Error ↓
RGB↔X [83]	14.3	0.323	19.6	0.286	0.441	0.321	23.80°
Ours	25.0	**0.205**	26.7	**0.204**	0.039	0.078	5.97°



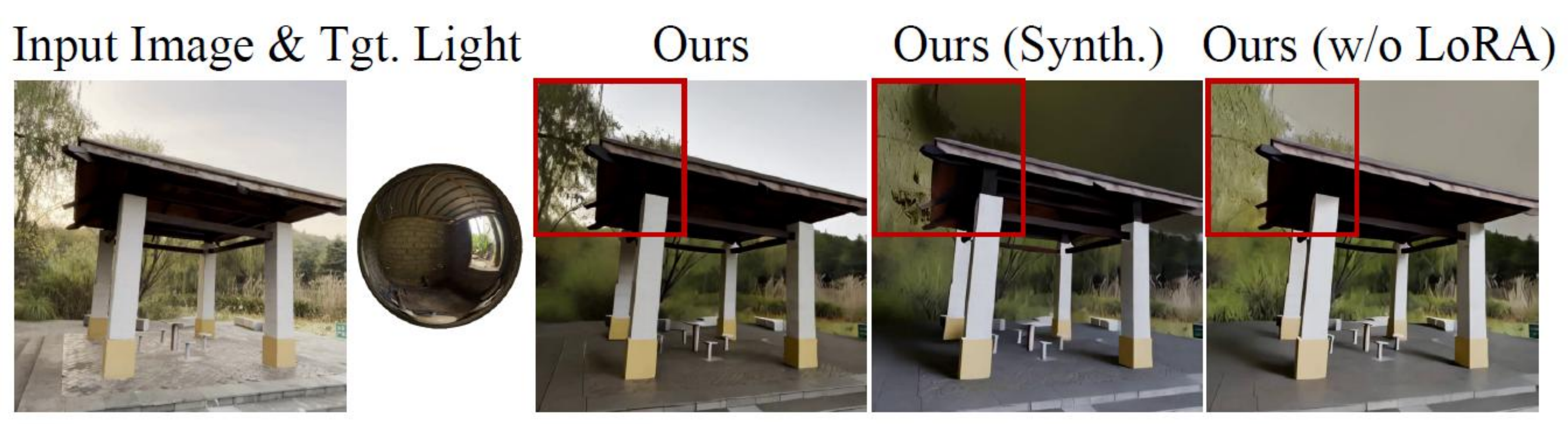
4. Experiments

- **Evaluation of relighting**

	SyntheticObjects PSNR ↑	SSIM ↑	LPIPS ↓	*SyntheticScenes* PSNR ↑	SSIM ↑	LPIPS ↓
DiLightNet [82]	23.79	0.872	0.087	18.88	0.576	0.344
Neural Gaffer [30]	26.39	0.903	0.086	20.75	0.633	0.343
Ours	**27.50**	**0.918**	**0.067**	**24.63**	**0.756**	**0.257**

4. Experiments

- **Ablation of relighting**



4. Experiments

- **Applications**

  ① **Material editing**

  ② **Object insertion**

5. Conclusion

① DIFFUSIONRENDERER provides a scalable, data-driven framework that successfully addresses the dual tasks of high-quality G-buffer estimation (inverse rendering) and photorealistic image generation (forward rendering).

② The system achieves these results without the need for traditional constraints like explicit path tracing or precise 3D geometry, relying instead on the power of video diffusion models.

5. Limitation

① **Inference Speed**: The system is currently slow (offline) due to relying on SVD, requiring distillation techniques to improve speed.

② **Content Consistency:** Editing may cause slight color/texture variations, suggesting a need for neural intrinsic features to enhance visual consistency.

③ **Auto-Labeling Accuracy:** The reliance on an off-the-shelf lighting model for real-world auto-labeling needs improvement in accuracy and robustness.

Thanks for Listening!