

# Ouroboros: Single-step Diffusion Models for Cycle-consistent Forward and Inverse Rendering

## Supplementary Material

### A. Video Inference

#### A.1. Pipeline

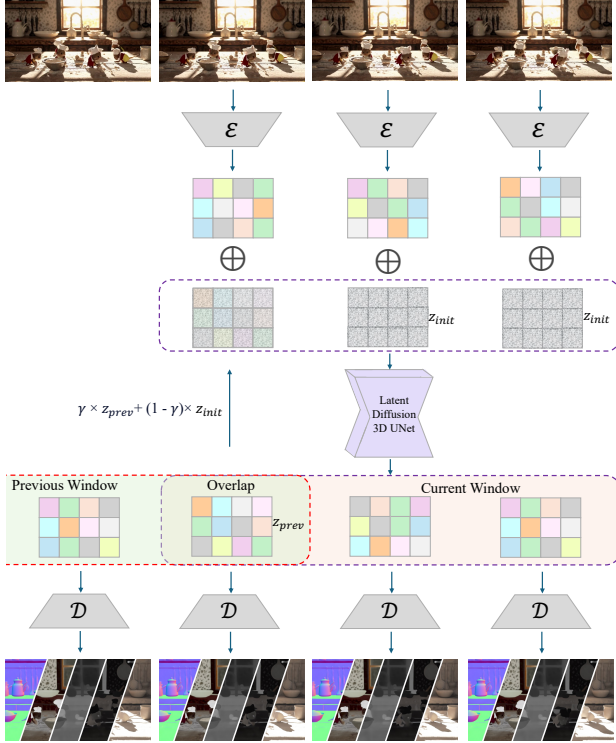


Figure 1. **Iterative video generation pipeline.** Overlapping windows are processed sequentially, with latent representations from previous windows guiding the initialization of overlapping regions. In practice, the window size and overlap are larger than the figure shown.

##### A.1.1. Limitations of Naive Per-frame Inference

Fig. 2 compares naive per-frame inference with our approach. In per-frame inference, each frame is initialized with the same noise. Inspired by Lotus [3], we incorporate the previous frame’s latent, scaled by a predefined factor, into the initial noise for the next frame. Although per-frame inference captures scene details well, it introduces temporal inconsistencies, such as the shining effect on the car in consecutive frames. Our approach mitigates this issue by jointly processing multiple frames, ensuring stable object appearances throughout the video.

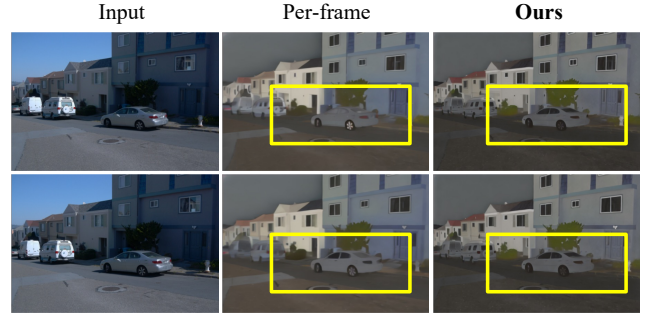


Figure 2. Qualitative comparison with naive per-frame inference.

Method	Roughness		Metallicity	
	PSNR	LPIPS	PSNR	LPIPS
RGB↔X [5]	12.07	0.35	8.04	0.45
Kocsis et al. [4]	11.29	0.32	8.93	0.71
Zhu et al. [6]	7.51	0.51	6.45	0.78
Ours	19.56	0.19	21.96	0.18

Table 1. Comparison of different methods on roughness and metallicity on interiorverse test set.

Method	Roughness		Metallicity	
	PSNR	LPIPS	PSNR	LPIPS
RGB↔X [5]	23.82	0.3655	6.83	0.57
Zhu et al. [6]	7.0028	0.6024	4.87	0.68
Kocsis et al. [4]	8.4766	0.4419	10.67	0.38
Ours	24.04	0.2301	26.32	0.14

Table 2. Comparison of different methods on roughness and metallicity on matrixcity test set

Method	PSNR↑	LPIPS↓
RGB↔X [5]	11.64	0.23
Du et al. [2]	9.51	0.56
<b>Ours</b>	12.07	0.29

Table 3. Comparison of different methods in Irradiance prediction in Hypersim datasets.

### B. Forward Rendering Results

#### B.1. Cycle Qualitative Results

The results can be seen in Fig. 5. Compared to RGB↔X [5], we demonstrate advantages in both inverse

rendering and forward rendering. In inverse rendering, our method shows better understanding of object materials such as building textures, while in forward rendering, there is significant improvement in the recovery of lighting conditions.

## B.2. Ablation study

### B.2.1. Effectiveness of e2e Loss in Cycle training

As shown in Fig. 3, we evaluated whether to use e2e loss in cycle training. We can observe that with e2e loss, there is better continuity in metallicity and irradiance predictions, and the results are more faithful to the actual physical properties and materials.

### B.3. Effectiveness of Wild Data in Cycle Training

Fig. 4 showcases the effectiveness of training with wild data. For tall buildings, we can clearly see that training with wild data produces more realistic irradiance, while metallicity is more continuous, demonstrating better understanding of surface properties.

## C. Forward Rendering Results

### C.1. Cycle Qualitative Results

The results can be seen in Fig. 5. Compared to RGB $\leftrightarrow$ X [5], we demonstrate advantages in both inverse rendering and forward rendering. In inverse rendering, our method shows better understanding of object materials such as building textures, while in forward rendering, there is significant improvement in the recovery of lighting conditions.

## C.2. Ablation study

### C.2.1. Effectiveness of e2e Loss in Cycle training

As shown in Fig. 3, we evaluated whether to use e2e loss in cycle training. We can observe that with e2e loss, there is better continuity in metallicity and irradiance predictions, and the results are more faithful to the actual physical properties and materials.

### C.3. Effectiveness of Wild Data in Cycle Training

Fig. 4 showcases the effectiveness of training with wild data. For tall buildings, we can clearly see that training with wild data produces more realistic irradiance, while metallicity is more continuous, demonstrating better understanding of surface properties.

## References

- [1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *OpenAI Blog*, 2024.
- [2] Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let’s find out! *arXiv preprint arXiv:2311.17137*, 2023.
- [3] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.
- [4] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for single-view material estimation. *arXiv preprint arXiv:2312.12274*, 2023.
- [5] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. Rgbx: Image decomposition and synthesis using material- and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery.
- [6] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022.



Figure 3. Ablation study on cycle training with or w/o e2e loss. Methods incorporating e2e loss can better understand lighting conditions and provide more continuous estimation. We can observe that the colors in the restored images are also more accurate/faithful.

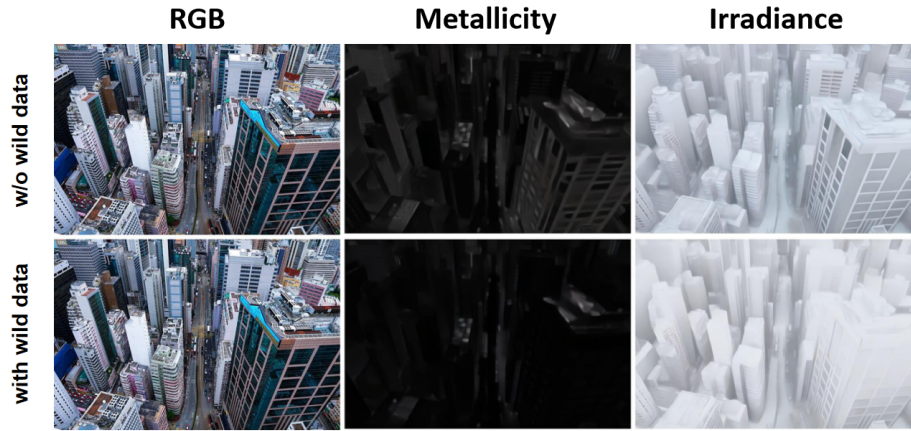


Figure 4. Ablation study on cycle training with wild data or w/o wild data. Training on wild data helps improve the understanding of overhead lighting and surface materials.



Figure 5. Comparison for X2RGB and Cycle on wild data. Our method demonstrates superior performance in terms of material understanding, lighting comprehension, and generation continuity.



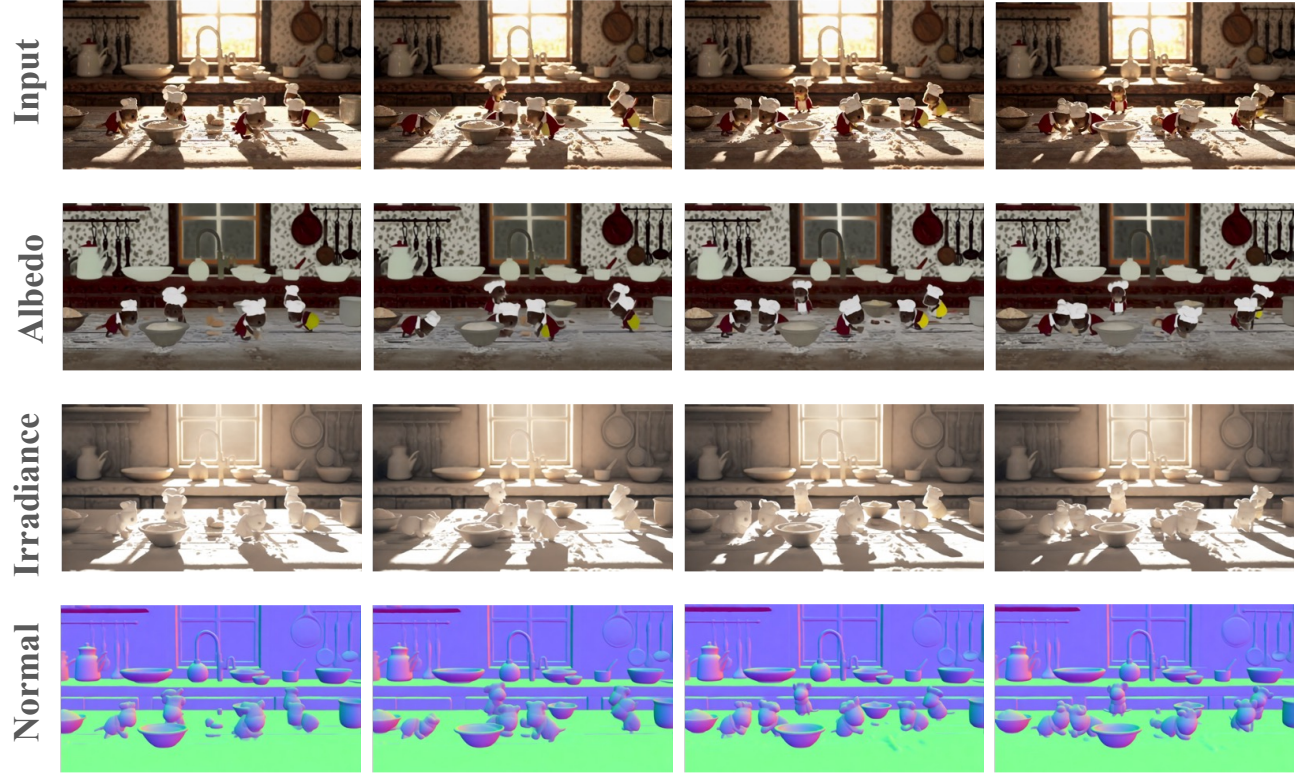


Figure 6. Example of video inference. Our model demonstrates the ability to synthesized videos generated by Sora[1].



Figure 7. Qualitative comparison with RGB↔X. In the first example, the RGB↔X model fails to maintain a stable albedo prediction, particularly for the building, where a large portion appears incorrectly as white. In contrast, our model produces more consistent and realistic results. In the second example, the Roman columns and the main wall exhibit distinct material properties. While the RGB↔X model struggles to differentiate them, our approach successfully preserves their unique characteristics in the output.





Figure 8. Qualitative comparison with RGB↔X. In the first example, the RGB↔X model struggles with the red cars at the end of the road, producing unrealistic predictions. In contrast, our model accurately preserves the structure and albedo of the vehicles, ensuring a more coherent and visually consistent output. In the second example, the RGB↔X model fails to maintain the consistency of the car’s albedo. In the 6th and 7th frames, the car’s color shifts from white to gray. In contrast, our model preserves the car’s appearance consistently across frames.