


SampleMono: Multi-Frame Spatiotemporal Extrapolation of 1-spp Path-Traced Sequences via Transfer Learning

Mehmet Oguz Derin 

Izmir, Türkiye

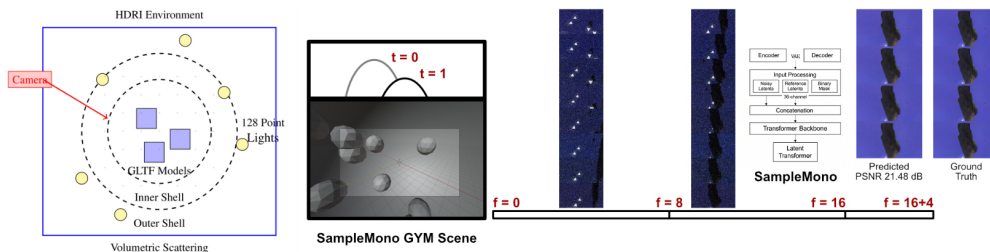


Figure 1: We present *SampleMono* and *SampleMono GYM*: Given 16 noisy 1-spp frames from the render timeline, our pipeline extrapolates and denoises them into four high-resolution outputs on the presentation timeline. By leveraging transfer learning from a frozen VAE encoder-decoder and a pruned transformer fine-tuned on synthesized data, it achieves spatial upsampling and denoising within a 5 GB VRAM budget.

Abstract

Path-traced sequences at one sample per pixel (1-spp) are attractive for interactive previews but remain severely noisy, particularly under caustics, indirect lighting, and volumetric media. We present *SampleMono*, a novel approach that performs multi-frame spatiotemporal extrapolation of low-resolution and low-sample Monte Carlo sequences without requiring auxiliary buffers or scene-specific information. We transfer and prune a pre-trained video generation backbone and fine-tune it on *SampleMono GYM*, a synthetic Monte Carlo dataset, to generate four clean high-resolution frames from a longer window of noisy inputs, thereby decoupling render and presentation timelines. Our experiments demonstrate that by combining a frozen VAE encoder-decoder and training of a video generation model pruned to two transformer layers, our pipeline can both provide spatial upsampling and temporal extrapolation to a long sequence of 16 RGB frames of 50 milliseconds time delta between frames at 256×144 resolution with severe Monte Carlo noise, generating subsequent four RGB frames of 12.5 milliseconds time delta between frames at 1280×720 resolution with substantially reduced noise at varying quality while fitting VRAM budget of 5GB. We plan to publish the code for data GYM, model pruning, pipeline training, and rendering.

CCS Concepts

• **Computing methodologies** → **Transfer learning; Ray tracing; Tracking;**

1 Introduction

Interactive, physically based path tracing at 1-spp yields severe Monte Carlo noise, especially with caustics, indirect illumination, and volumetrics. Conventional denoisers and frame generation often assume auxiliary buffers (normals, albedo, motion) that are costly on memory- and bandwidth-constrained platforms, can bias reconstruction, and limit sample streaming over the network [CKS*17, BVM*17]. Recent temporal generation methods improve presentation but typically depend on future frames or renderer-specific features [WKZ*23, WVS*24].

SampleMono reframes reconstruction as an RGB-only, causal

video-to-video translation problem: from a long, noisy, low-resolution input window, synthesize a short burst of higher-resolution, denoised future frames, decoupling render and presentation timelines. The method processes 16 past 1-spp frames at 256×144 (50 ms spacing) and predicts four future frames at 1280×720 (12.5 ms spacing), achieving $5 \times$ spatial upsampling and temporal extrapolation without G-buffers.

2 Method

Backbone transfer and pruning We adapt a pre-trained video generation model to deterministic regression. The system comprises:

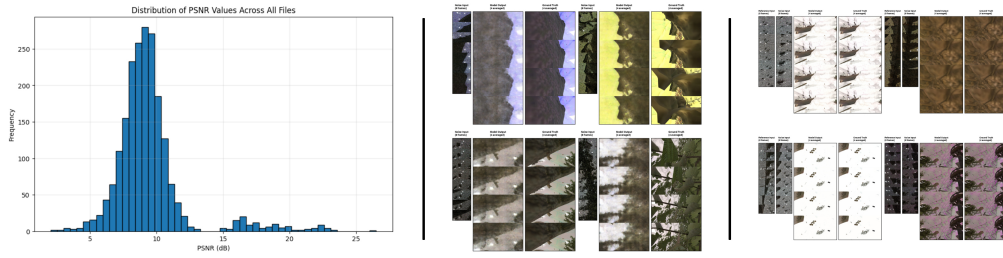


Figure 2: Qualitative and quantitative results. Left is the histogram of PSNR values, middle is the worst-performing sequences (PSNRs are 4.67 dB, 2.94 dB, 4.81 dB, and 3.67 dB), right is the best-performing sequences (PSNRs are 22.73 dB, 22.15 dB, 26.51 dB, and 22 dB).

(1) a frozen VAE encoder–decoder that maps RGB frames to a compact latent space, instantiated from the Wan-2.1 video model suite [WAN25]; (2) a pruned Wan-2.1 14B I2V spatiotemporal transformer reduced to two blocks (from 14), yielding ~ 1.18 B trainable parameters that fit within a ~ 5 GB weight budget; and (3) an input packer that concatenates noisy input latents, optional reference latents, and binary masks into a model-ready tensor. We convert the diffusion backbone into single-step, text-free, deterministic prediction by fixing the timestep to zero noise and removing auxiliary conditioning.

Learning objective Let $\mathcal{E}(\cdot)$ and $\mathcal{D}(\cdot)$ denote the frozen VAE encoder/decoder. Given an input window I_{in} (16 past RGB frames) and targets O_{out} (four future high-res RGB frames), the transformer g_{θ} operates in latent space with an L2 reconstruction loss:

$$\mathcal{L}(\theta) = \|\mathcal{E}(O_{out}) - g_{\theta}(\mathcal{E}(I_{in}))\|_2^2. \quad (1)$$

No renderer buffers (G-buffers) are used; the approach is strictly causal and RGB-only.

Data synthesis (SampleMono GYM) We generate paired sequences in Blender Cycles [Ble25]. Scenes combine randomly sampled GLTF assets (including Poly Haven models/HDRIs [Pol25]), nested glass shells, 128 point lights, environment lighting, and volumetric scattering. Camera/object trajectories use C^1 -continuous interpolation. Inputs are rendered at 1-spp with independent per-frame seeds; references are high-spp (128-spp) and denoised with OIDN [Áfr25]. The dataset comprises 2,048 training and 2,048 evaluation sequences. Optimization uses AdamW with cosine warmup; only the two transformer blocks are trained while the VAE remains frozen.

3 Experimental Results

We evaluate on 2,048 held-out sequences. The model performs joint denoising, $5\times$ spatial upsampling, and causal temporal extrapolation without auxiliary buffers.

Quality Mean PSNR is 9.42 ± 4.87 dB (range 2.04–26.51 dB). Success cases feature smooth motion and stable illumination; failure modes include flicker under caustics/volumetrics, global color drift during rapid lighting changes, and hallucinations on thin structures.

Compute and memory End-to-end inference is 2,967 ms per 4-frame group on an A100 (40 GB), i.e., ~ 742 ms/frame amortized. The pruned transformer totals ~ 1.18 B parameters with ~ 4.49 GB of weights; the frozen VAE decoder dominates runtime.

4 Discussion and Conclusion

SampleMono shows that video priors can be pruned and transferred to enable buffer-free, causal spatiotemporal reconstruction of 1-

spp path-traced sequences. Operating in a frozen VAE latent space with a two-block transformer meets tight VRAM constraints while delivering $5\times$ upsampling and temporal extrapolation.

Limitations include fixed 16:4 context/output sizing, latency (~ 742 ms/frame) that precludes real-time, and quality variance on fast dynamics and complex light transport. Training on synthetic data may limit generalization. Future work includes decoder-path pruning/quantization, confidence estimation, extended temporal context, broader datasets, alternative architectures, frame packing, and remote sample streaming. We plan to release a version of SampleMono, including GYM, pruning utilities, training code, and render-engine patch, under a permissive open license.

Acknowledgments This research relies on open-source tools and open-access data with permissive licenses. We gratefully acknowledge the contributors and community for providing these invaluable resources.

References

- [Áfr25] ÁFRA A. T.: Intel® Open Image Denoise, 2025. <https://www.openimagedenoise.org.2>
- [Ble25] BLENDER ONLINE COMMUNITY: Blender 4.5. <https://www.blender.org/,2025.2>
- [BVM*17] BAKO S., VOGELS T., MCWILLIAMS B., MEYER M., NOVÁK J., HARVILL A., SEN P., DEROSE T., ROUSSELLE F.: Kernel-predicting convolutional networks for denoising monte carlo renderings. *ACM Trans. Graph.* 36, 4 (July 2017). doi:10.1145/3072959.3073708. 1
- [CKS*17] CHAITANYA C. R. A., KAPLANYAN A. S., SCHIED C., SALVI M., LEFOHN A., NOWROUZSAHRAI D., AILA T.: Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. *ACM Trans. Graph.* 36, 4 (July 2017). doi:10.1145/3072959.3073601. 1
- [Pol25] POLY HAVEN: Poly haven asset library. <https://polyhaven.com/,2025.2>
- [WAN25] WAN TEAM: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025). 2
- [WKZ*23] WU S., KIM S., ZENG Z., VEMBAR D., JHA S., KAPLANYAN A., YAN L.-Q.: Extrass: A framework for joint spatial super sampling and frame extrapolation. In *SIGGRAPH Asia 2023 Conference Papers* (New York, NY, USA, 2023), SA '23, Association for Computing Machinery. doi:10.1145/3610548.3618224. 1
- [WVS*24] WU S., VEMBAR D., SOCHENOV A., PANNEER S., KIM S., KAPLANYAN A., YAN L.-Q.: Gffe: G-buffer free frame extrapolation for low-latency real-time rendering. *ACM Trans. Graph.* 43, 6 (Nov. 2024). doi:10.1145/3687923. 1