Generative Rendering: Controllable 4D-Guided Video Generation with 2D Diffusion Models

Supplementary Material

Method	Pairwise Frame Interval				
	1	5	10	15	20
Per-Frame	0.9547	0.9173	0.9109	0.9145	0.9062
Pix2Video*	0.9630	0.9503	0.9494	0.9415	0.9471
TokenFlow*	0.9822	0.9754	0.9728	0.9706	0.9712
Gen1	**0.9907**	0.9715	0.9582	0.9624	0.9601
Ours	0.9845	**0.9815**	**0.9738**	**0.9749**	**0.9727**

Table 2. **Frame consistency with different intervals.** Our approach offers better or competitive consistencies at different intervals. *: baseline adapted to our setting.

A. Additional qualitative results

We refer to the supplemental webpage for various qualitative results obtained by using different 3D scenes, motions, and target prompts.

B. Additional quantitative results

We report CLIP [29] embedding similarities for a pair of frames separated by different frame intervals in Tab. 2. As shown, semantic consistency achieved by our method is stable across different intervals. This is potentially because of our texture-based feature aggregation. This provides a canonical representation for the objects that links frames that are temporally further apart even in long sequences. Even though video-based methods such as Gen1 [7] achieve temporally smoother results, we speculate that it is not easy for them to capture such long range interactions.

C. Robustness to video length

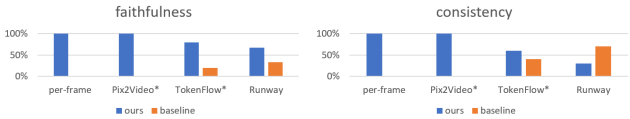
Our method can be scaled up to handle long input sequences. To generate a long output video, our method requires the sampling of a sparse set of keyframes, and the unified features in the texture space can be propagated. The only GPU memory bottleneck in our method is the parallel processing of the keyframes using inflated attention. Our algorithm is robust enough to generate long videos with $>$200 frames with $<$12GB GPU memory.

D. Additional ablations

Effectiveness of UV noise initialization. We find that the performance of many components in our pipeline, e.g. inflated attention, are tightly related to the initial noise. Video editing works [6, 10, 39] typically require accurate DDIM inversion [17] as noise initialization. Using DDIM inverted noise encourages temporal coherence and similarities to the input video. However, since our input is textureless UV maps and depths, DDIM inversion constantly fails and will not provide useful information. We instead propose to utilize the canonical UV space to initialize the noise in each frame of the sequence. In our supplemental page, we show the effectiveness of this initialization strategy.

Effectiveness of latent normalization. Our latent normalization can address a large portion of the color flickering issue, where the overall color distribution shifts randomly for different frames. We notice that small color shifts in the latent space will be inflated by the VAE decoder used in Stable Diffusion [32], causing the generation process very sensitive to small differences in the latent space. The video comparison is included in our supplemental page.

E. Perceptual user experiment



We conduct a perceptual study following the setups in Pix2Video. Given 5 sets of 3D scenes and prompts, we generate animations from our *Generative Rendering* and baselines (please refer to Sec. 4 in our paper for the baseline setups). We then ask 20 users to compare our animations against the baselines. We ask two questions: **1)** Which animation looks more pleasing and represents the prompt better? and **2)** Which animation looks more consistent? We show the preference below, where preference for the first question is labeled as "faithfulness", and the second as "consistency". We plot the % of answers that think our method outperforms a baseline.

F. Limitation and failure cases

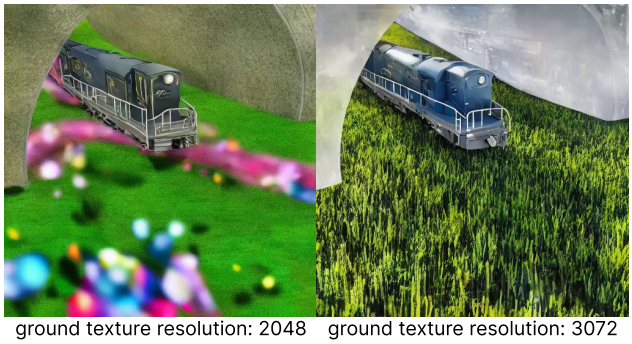


Figure 7. **Effect of different texture resolution.** If the ground's texture resolution is too low (2048×2048), artifacts appear at ground regions in the final rendered image compared with using a higher texture resolution (3072×3072). Prompt: a train running in a field of green grass.

Generative rendering still suffers from several aspects. Firstly, it cannot achieve real-time animations due to the multi-step inference of current diffusion models. However, accelerating this inference stage has been an active research area and advances in this area, such as consistency models [34], can be directly applied to speed up our method. Furthermore, generative rendering is not yet able to guarantee perfect consistency and preservation of details. This is mainly because our method works purely in the low-dimensional latent space, which is only 64×64 pixels in our model resulting in imprecise UV correspondences. For the same reason, generative rendering may suffer from unwanted misalignment and artifacts. We believe that applying some of findings to pre-trained video diffusion models to augment them with 3D controllability is an exciting future direction. A large portion of our inconsistency comes from the VAE decoder. As mentioned in D, small inconsistencies in the latent space will be inflated by the VAE decoder while transiting into RGB frames. This phenomenon scales to image details, where small differences in the latent space will be inflated after decoding with the VAE. Additionally, since our noise initialization and feature fusion both work in the UV space, it could be tricky to set the texture resolutions. Setting the texture resolution too low will create corrupted regions while too high will cause the UV coordinates to be too far away and no blending effect will take place. An example is shown in Fig. 7. Finally, our work does not yet generalize to large environmental changes and dramatic perspective changes. This is because of the highly overlapping pixel-wise matches and bad feature projections. We believe finetuning or adding a video module to be an exciting future direction to improve the performances on these more challenging scenes.