StereoWorld: Geometry-Aware Monocular-to-Stereo Video Generation

Supplementary Material

6. More Implementation Details.

We build our method upon Wan2.1-T2V-1.3B [48], which generates 5-second video clips at 16 FPS with a spatial resolution of 832×480. To obtain depth supervision, we employ Video Depth Anything [9] to estimate per-frame depth maps for all training videos. We further employ Stereo Any Video [24] to generate ground-truth disparity maps used for supervising the disparity during training. Each input video is represented as $V \in \mathbb{R}^{c\times f\times h\times w}$, where $c = 3, f = 81, h = 480$, and $w = 832$. The corresponding latent representation encoded by the 3D VAE is $z \in \mathbb{R}^{c'\times f'\times h'\times w'}$, with $c' = 16, f' = 21, h' = 60$, and $w' = 104$. During fine-tuning, we adopt the LoRA [20] framework with a rank of 128, using "dit" as the base model type. The LoRA adaptation targets the modules "q, k, v, o, ffn.0, ffn.2, head.head". The number of shared DiT blocks N is 20 and the number of RGB&depth DiT blocks L is 10. We train the model with a batch size of 2 and gradient accumulation steps of 2 on 8 gpus, resulting in an effective batch size of 32. The learning rate is set to 1×10^{-4}, setting $\lambda_{\text{l}} = \lambda_{\text{ll}} = 0.1$ and $\lambda_{\text{dis}} = 0.5$, and the model is trained for one epoch, totaling approximately 9k optimization steps. During temporal tiling training, the frame replacement probability p is set to 0.3 and the amount of overlapping frames is 9. Training is performed on 8 NVIDIA A800 GPUs using the AdamW optimizer [28] under bfloat16 precision. The entire training process takes approximately 11 days to complete.

7. More Comparisons and Cases.

We have prepared many additional comparison examples, along with more high-resolution and extended-duration video results, in our supplementary video materials. We sincerely invite the reviewers to watch them—these videos offer the clearest and most vivid demonstration of our method's capabilities.

8. Source code of the stereo projector

The source code of the stereo projector is provided in the accompanying code files.