Stereo World Model: Camera-Guided Stereo Video Generation

Supplementary Material

S1. Experiment

S1.1. Dataset Construction

The datasets used for training are summarized in Tab.1 of the main paper. For Stereo4D [29], we filtered out videos in which the camera remained static, exhibited minimal motion, or suffered from excessive jitter, as such samples are unsuitable for camera-conditioned training. Each video was divided into 49-frame clips, which were then cropped and resized to a uniform resolution of 480×640. For each clip, we used the left-eye video to generate caption annotations. All training data were accompanied by metric-scale camera parameters.

For the test set, we selected approximately 280 video clips from the processed TartanAirGround [43] video clips, sampled at intervals of 200. In addition, we used the UnrealStereo4K [54] and Middlebury [45] stereo image datasets, for which we generated a set of random camera trajectories to conduct out-of-domain evaluations (approximately 160 clips). Each camera trajectory was composed of both translation and rotation components. The translation sampling range along the z-axis was $[-20m, -4m] \cup [4m, 20m]$, and the rotation sampling range around the y-axis was $[-150°, -50°] \cup [50°, 150°]$.

S1.2. Stereo Attention FLOPs

For each attention head, let L be the sequence length or number of query tokens, d be the head dimension, a vanilla full attention head costs:

$$\text{FLOPs}_{\text{full}} = 4L^2 d. \tag{10}$$

In our experiment, the input feature has the shape $f \in \mathbb{R}^{b \times 2f \times h \times w \times c}$. As for 4D Attention, $L = 2f \times h \times w$, we have

$$\text{FLOPs}_{\text{Attn}_{\text{4D}}} = 16bf^2h^2w^2d. \tag{11}$$

While for the stereo attention, we have

$$\text{FLOPs}_{\text{Attn}_{\text{3D}}} = 8bf^2h^2w^2d, \tag{12}$$

$$\text{FLOPs}_{\text{Attn}_{\text{row}}} = 4bfhw^2d. \tag{13}$$

Supposing we use $b = 1$, $f = 13$, $h = 15$, $w = 20$, $d = 128$, we can calculate that $\text{FLOPs}_{\text{Attn}_{\text{4D}}} \approx 3.115 \times 10^{10}$, while in comparison, the stereo attention costs $\text{FLOPs}_{\text{Attn}_{\text{3D}}} + \text{FLOPs}_{\text{Attn}_{\text{row}}} = 1.561 \times 10^{10}$. Hence the stereo attention block reduces multiply-adds by a factor about $\mathbf{2}\times$.

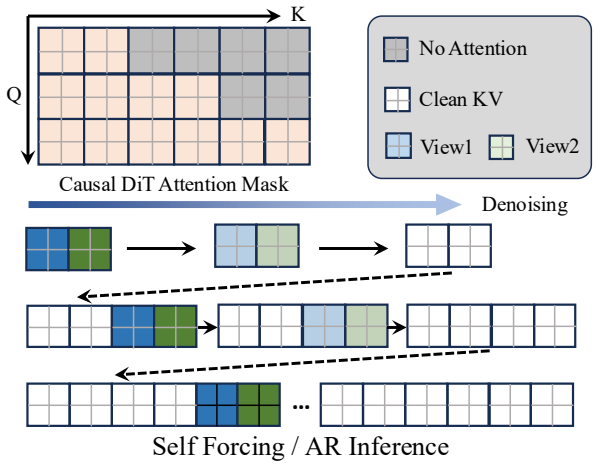

Figure 8. Attention mask configuration in distillation process.

S2. Application

S2.1. VR/AR Display

The binocular videos generated by StereoWorld can be directly utilized in VR/AR applications to deliver immersive experiences. In Fig. 9, we provide additional generated scene examples, together with anaglyph image to demonstrate the diversity and practicality of our approach. We also report the user study results in Fig. 10, in which we compare our results with baselines in terms of "Camera Conformity", "Temporal Consistency", "Image Quality" and "Overall Experience".

S2.2. Embodied Scenarios

By fine-tuning our model on binocular robotic arm datasets [31], our approach can also be applied to embodied scenarios for stereo video generation, supporting downstream tasks such as action planning. As shown in Fig. 11, given an action command and the initial stereo frame, our model is able to generate the corresponding subsequent motion sequence. The results demonstrate that the generated videos not only follow the specified action instructions but also maintain high stereo consistency between the left and right views. We further performed disparity estimation on the generated outputs to verify their geometric plausibility and assess their feasibility for action planning.

S2.3. Long Video Distillation

Our trained model employs a bidirectional attention mechanism, which limits it to relatively short video sequences (49 frames in our setting). In contrast, autoregressive video gen-

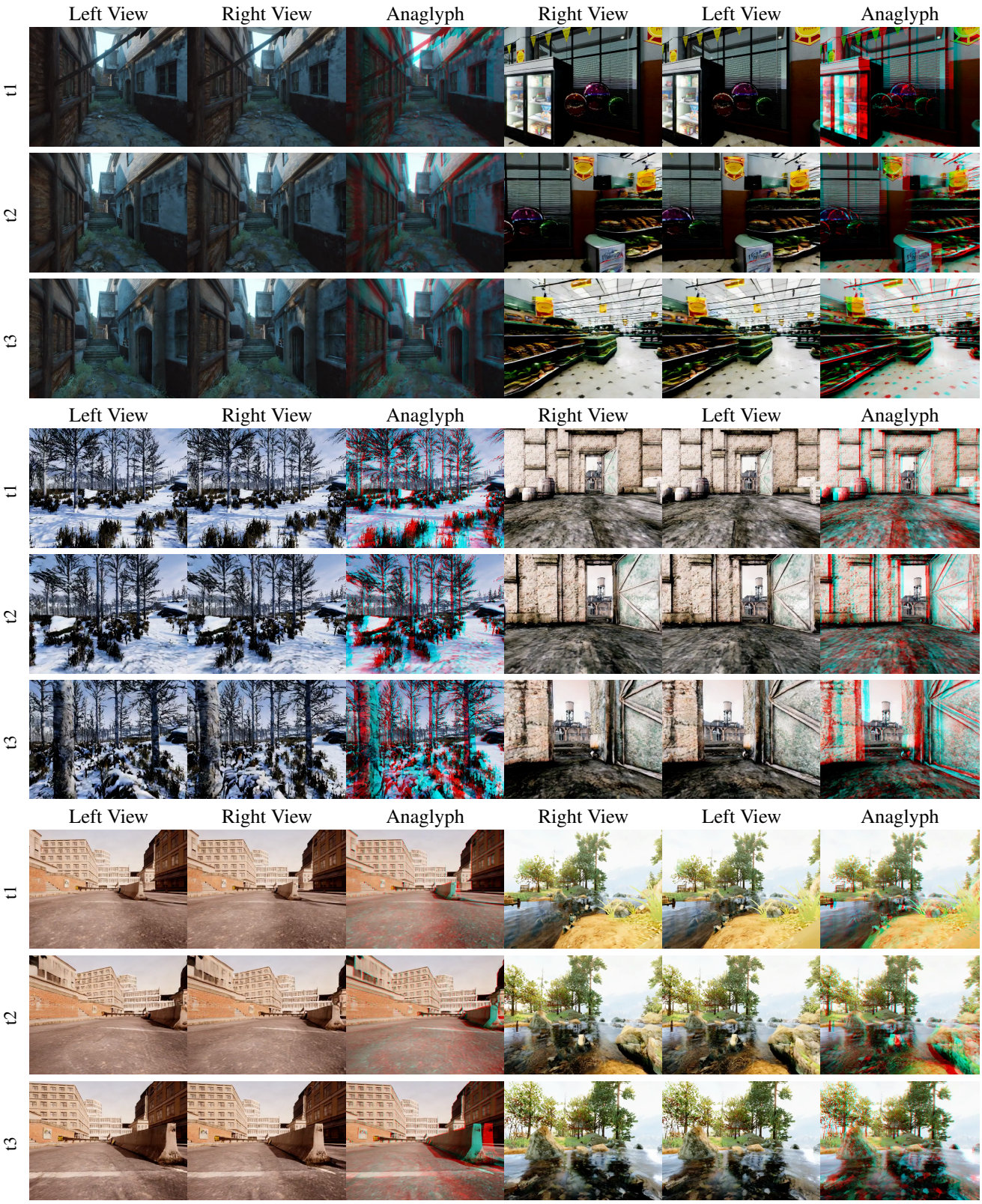

Figure 9. More StereoWorld Results with Anaglyph.

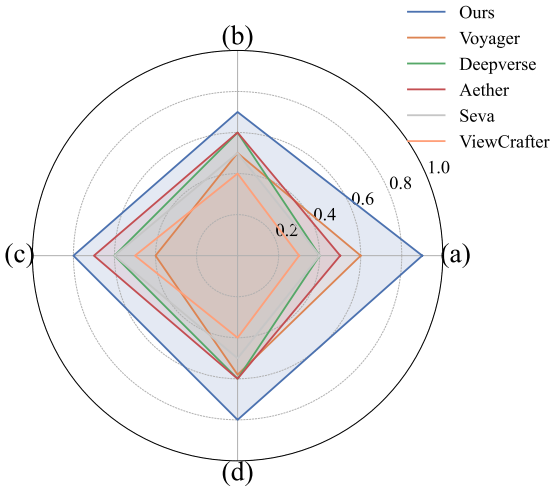

Figure 10. The summary of quantitative feedback in the user study. (a) Camera Conformity (b) Temporal Consistency (c) Image Quality (d) Overall.

eration models [27, 74] can effectively overcome this limitation and improve efficiency through a rolling KV-cache mechanism. Inspired by these advancements, we further distill StereoWorld into an autoregressive binocular video generation model, enabling long-horizon video synthesis and improving generation speed.

Following Self-Forcing [27], we adopt a two-stage paradigm. In the first stage (ODE distillation), we replace the bidirectional attention with a causal attention mechanism and distill the denoising process into four steps. The attention mask is illustrated in Fig 8, which generates two views at one step. In the second stage [27], we condition each pair of stereo frame's (or chunks in practice) generation on previously self-generated outputs by performing autoregressive rollout with KV-cache. In this stage, a distribution matching distillation [73] (DMD loss) is applied to address exposure bias via distribution matching. Unlike monocular autoregressive video generation, our method simultaneously synthesizes binocular views and incorporates camera pose–aware positional encoding. As a result, the KV-cache must be updated with two separate sets of keys and values at each step – one for the left-view tokens and one for the right-view tokens, each containing our *Unified Camera-Frame RoPE*.

The distilled model achieves a significant improvement in binocular video generation speed, increasing from 0.49 FPS to 5 FPS, and is no longer limited to generating video clips of 49 frames. We present the results of long-video distillation in Fig 12, and in the supplementary video materials.

However, we observe that as the video length increases, the generated results still exhibit noticeable degradation. This issue is also present in prior works such as Self-Forcing. Improving the stability of long-horizon video generation therefore remains an open challenge shared by both monocular and stereo video synthesis.

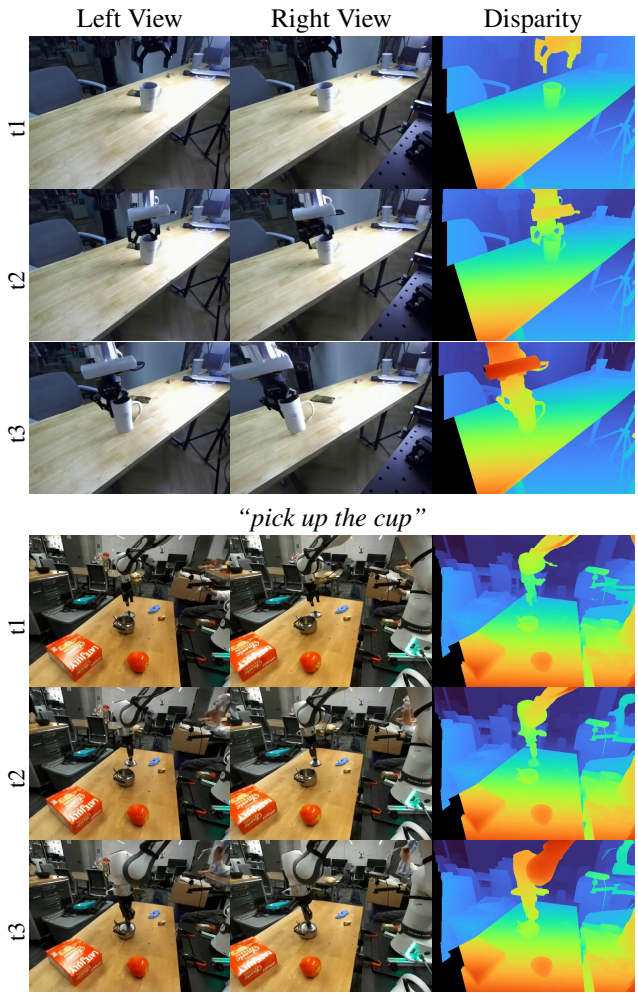

Figure 11. Stereo Video Generation on Embodied Scenarios.

S3. Monocular & Stereo Generation Comparison.

"Ours Monocular" and "Ours Stereo" in Tab 2 employ the exact same parameter count and compute budget. The superior FID for "Ours Stereo" is because binocular views provide a physical "anchor". As demonstrated below (Fig 13) monocular pipelines relies on a single condition frame and often hallucinate unrealistic structures due to occlusion, whereas stereo settings incorporates additional view and better maintains alignment with real scene by stereo-aware attention.

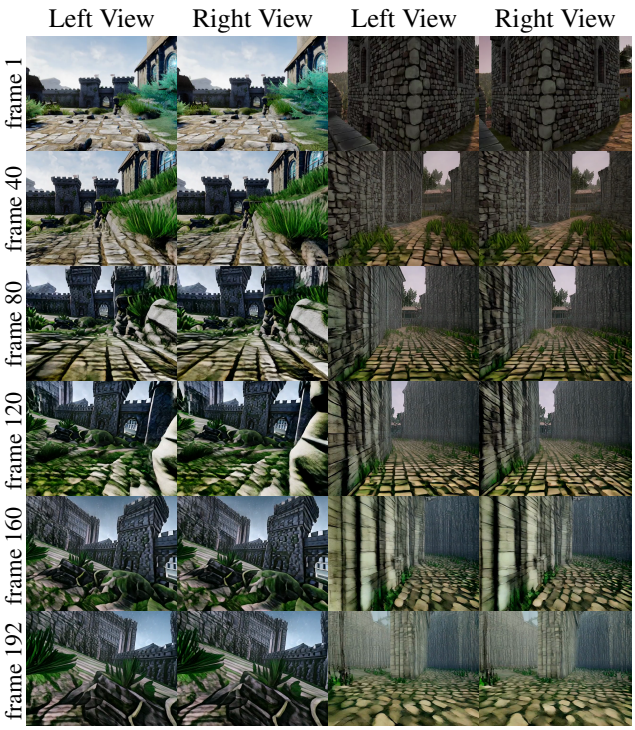

Figure 12. Long Video Distillation Results.

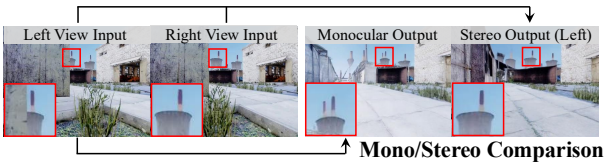

Figure 13. Monocular and stereo generation comparison.

S4. Large & Varying Baselines.

To evaluate the model's performance under varying baselines, we construct a camera trajectory by expanding the right camera baseline from 0.25m to 0.75m– well beyond the training distribution (0.063m-0.25m). As illustrated below (Fig 14), StereoWorld maintains geometric plausibility and achieves accurate metric-scale recovery up to 0.42m, outperforming SOTA like DepthAnything V2. This confirms our Unified Camera-Frame RoPE performs genuine geometric reasoning rather than simple image stretching, also demonstrating robust generalization to unseen camera trajectories and baseline configurations.

S5. Discussion

Our method currently does not incorporate any explicit constraints on scene-level consistency. Although it handles most cases well, certain examples may still exhibit spatial inconsistencies across video frames, as illustrated in Fig 15. This issue may be alleviated by introducing a spatial memory mechanism [34, 66]. Since stereo video generation inherently provides geometric information about the scene, our approach can be naturally integrated with methods such as VMem [34] or SPMem [66], replacing their additional reconstruction modules and maintaining consistency through a dedicated spatial memory.

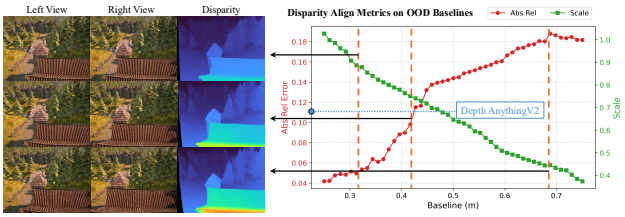

Figure 14. Effect of different baselines on StereoWorld.

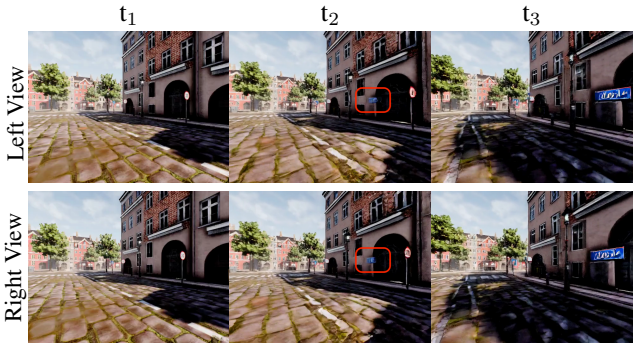

Figure 15. Failure Case. Note that the blue road sign does not exist at the beginning of the sequence; however, as the viewpoint advances, it gradually emerges and increases in size.

We also note that our method predominantly generates static scenes. This is primarily due to the limited availability of binocular video data for training stereo models. Most of our training corpus consists of static, rendered scenes, which restricts the model's ability to synthesize dynamic environments. Exploring strategies for collecting more dynamic stereo video data, or leveraging richer monocular dynamic video datasets, represents a highly promising direction for future work. Scaling the training to substantially larger datasets may also help mitigate the aforementioned consistency issues.

Moreover, since the stereo world model generates binocular videos simultaneously, it inherently models fewer frames compared to monocular methods. Although distillation into autoregressive frameworks enables the generation of longer videos, we still observe noticeable degradation in the later stages of video generation, similar as reported in self-forcing [27]. Developing approaches to robustly distill stereo video models into long-term video generators will therefore be a key focus for our future work.