DreamStereo: Towards Real-Time Stereo Inpainting for HD Videos

Supplementary Material

Appendix

This supplementary material provides additional analyses and visual results that complement the experiments presented in the main paper. Appendix A details the runtime evaluation and component ablations, while Appendix B presents extended qualitative comparisons across datasets and baselines, together with representative failure cases.

A. Extended Experiments

A.1. Runtime Analysis

Figure 8 illustrates the inference pipeline of 2D-to-3D conversion, which integrates our proposed Gradient-Aware Parallax Warping (GAPW) for precise occlusion handling and view synthesis. Within this pipeline, DreamStereo serves as the stereo inpainting module responsible for reconstructing occluded regions in the right view. We further analyze the runtime characteristics of this module under identical inference configurations.

We compare the end-to-end latency of SASI with our internal baseline that adopts dense tokens (no sparsity) and the original WanVAE. All measurements were conducted on an NVIDIA A100 using HD-100 videos at 768×1280, batch size 1, and FP16 precision. As shown in Fig. 9, our method achieves an overall latency reduction of **84.0%** per frame. Specifically, the sparsity-aware DiT reduces computation time by **84.2%**, while the distilled 3D-aware VAE lowers encoding and decoding cost by **83.9%**. Together, these two optimizations shorten total inference from 250 to 40 ms per frame, enabling real-time HD stereo generation at 25 fps. We also report throughput on commodity GPUs, achieving **54.2** ms/frame on an RTX 3090 and **33.0** ms/frame on an RTX 4090, compared with **40.1** ms/frame on an A100.

A.2. Distilled VAE Ablation

We adopt a method similar to CV-VAE [46] to distill the WanVideo [28] VAE, aiming to reduce the time consumption of video encoding and decoding. Tab. 6 reports stereo inpainting results on the HD-100 test set using different VAE while keeping the same final setting as in the main paper, showing that the distilled VAE achieves nearly identical quality to the original WanVAE. For standalone VAE profiling at 1024×1024 resolution (Tab. 7), our model reduces parameter count by **36%** and achieves over 4× faster encoding and decoding, providing substantial acceleration with negligible quality impact.

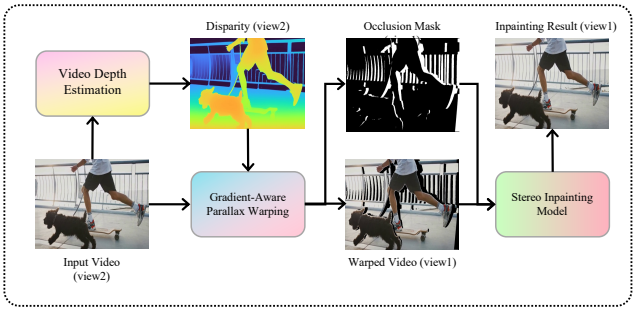

Figure 8. Inference pipeline of 2D-to-3D conversion.

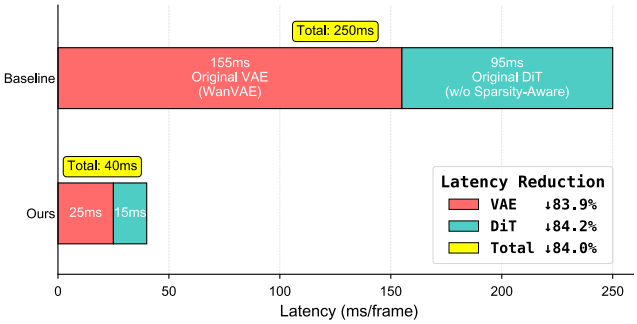

Figure 9. Module-wise latency breakdown and reduction on 768×1280 HD videos.

VAE	PSNR ↑	SSIM ↑	LPIPS ↓
WanVAE	30.59	0.906	0.040
Distilled (Ours)	30.48	0.900	0.053

Table 6. Ablation on stereo inpainting quality on HD-100 (768×1280). Results are evaluated under the same final setting as in the main paper (see Tab. 1), using different VAEs. The distilled VAE achieves nearly identical quality to the original WanVAE.

VAE	Params	Encoder	Decoder	Total
WanVAE	126.8 M	47.7 ms	75.0 ms	122.7 ms
Distilled (Ours)	80.2 M	9.3 ms	18.5 ms	27.8 ms

Table 7. Ablation on VAE parameter efficiency and latency at 1024×1024 resolution.

A.3. Training Sparsity Ablation

We further conduct an ablation on applying sparsity during training. Tab. 8 compares the default setting, which uses sparsity only at inference, with a variant that applies sparsity in both training and inference. The results show that

introducing sparsity during training brings no clear performance gain across metrics. Therefore, we use sparsity only at inference in the final model, which reduces redundant computation without changing the training pipeline.

Train-time token filtering	PSNR↑	SSIM↑	LPIPS↓
No (dense training)	32.48	0.933	0.049
Yes (sparse training)	32.31	0.932	0.050

Table 8. Ablation on applying sparsity during training.

B. Additional Visual Results

B.1. Comparison with ZeroStereo

Figure 10 complements Fig. 5 in the main paper by adding the stereo inpainting baseline ZeroStereo [30]. ZeroStereo is a training-free, single-image stereo inpainting method that fails to maintain temporal coherence and often yields distorted textures and geometry. In contrast, our approach produces temporally consistent, geometrically accurate, and perceptually clean stereo reconstructions.

B.2. Comparison with SpatialDreamer and M2SViD

We further compare our results with recent non-released stereo synthesis models using publicly available demo videos. The left view is used as input, and the generated right view is compared. As illustrated in Figs. 11 and 12, SpatialDreamer [15] exhibits truncation artifacts near image boundaries, while M2SViD [26] shows color shifts and degraded structure fidelity. Our approach maintains complete, sharp, and geometrically aligned stereo structures.

B.3. Qualitative Results on Dynamic Replica and SVD

While the main paper reports quantitative results on SVD AVP and Dynamic Replica test sets, we provide qualitative comparisons in Figs. 13 and 14. Our approach produces faithful stereo completions with sharper details and fewer artifacts, consistent with the quantitative improvements observed in the main paper.

B.4. Geometry-Consistency Comparison

We therefore add a geometry-consistency metric on **Dynamic Replica** (Table 2), where disparity is estimated from the generated left-right pair and depth accuracy is evaluated after alignment to metric depth using scale and shift (Table 9). While ZeroStereo achieves the best depth score, it is an image-only, training-free baseline and still exhibits noticeable edge artifacts in the stereo views (Fig. 14). In contrast, our method achieves competitive depth accuracy while producing cleaner stereo geometry.

Metrics	Deep3D	Depthify.ai	Owl3D	StereoCrafter	ZeroStereo	Ours
AbsRel. ↓	0.029	0.035	0.042	0.039	**0.018**	<u>0.019</u>
$\delta < 1.05$ ↑	0.847	0.770	0.716	0.751	**0.929**	<u>0.914</u>

Table 9. Geometry-consistency comparison on Dynamic Replica.

B.5. Failure Cases

Due to the lack of stronger priors for challenging real-world scenes, directly performing pixel-domain warping based on estimated depth still has inherent limitations. As shown in Figure 15, our method exhibits noticeable artifacts in several representative cases.

- **Transparent objects.** Due to the geometric ambiguity introduced by transparency, the utility pole seen through the bus window becomes noticeably distorted.
- **Reflective surfaces.** The shadows reflected on the corridor glass are inconsistent with the true scene geometry, since the model cannot properly account for such view-dependent reflection effects.
- **High-frequency details.** The synthesized results appear insufficiently sharp, indicating limited fidelity in preserving fine-grained high-frequency textures.

These examples suggest that transparency, reflection, and dense high-frequency textures remain challenging for depth-based warping methods. We consider such cases as long-tail scenarios in view synthesis, which may be further alleviated in future work by improving geometric reasoning and incorporating stronger visual priors.

Figure 10. **Qualitative comparison on HD-100** (768×1280)**.** This figure complements Fig. 5 by including ZeroStereo [30], a training-free stereo inpainting method that often produces random, spatially inconsistent fillings and distorted textures.

Figure 11. **Comparison with SpatialDreamer [15].** Highlighted regions show truncation artifacts in SpatialDreamer, whereas our results preserve complete and consistent stereo geometry.

Figure 12. **Comparison with M2SViD [26].** Our results show sharper details, larger disparities, and better color consistency, while M2SViD suffers from color drift and structural degradation.

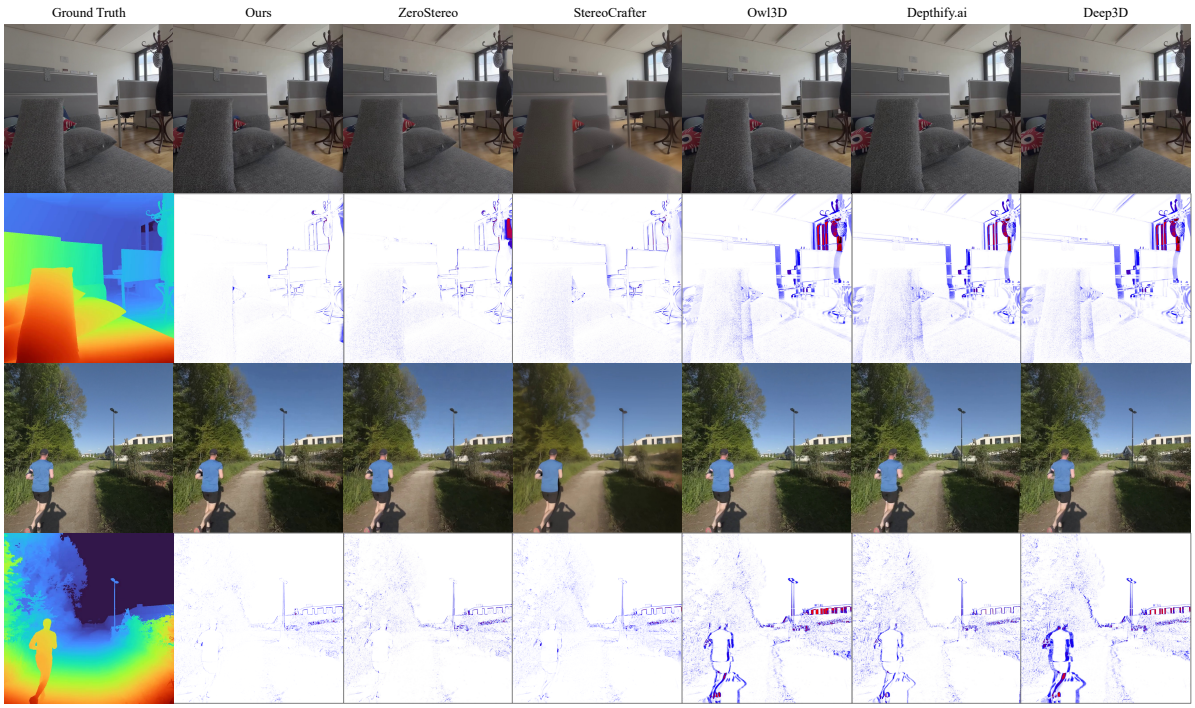

Figure 13. **Qualitative comparison on the SVD AVP test set** (768×768)**.** The second row shows per-pixel MSE maps, where blue denotes small errors and red large deviations from the ground truth. Our results exhibit the smallest errors and best overall consistency.

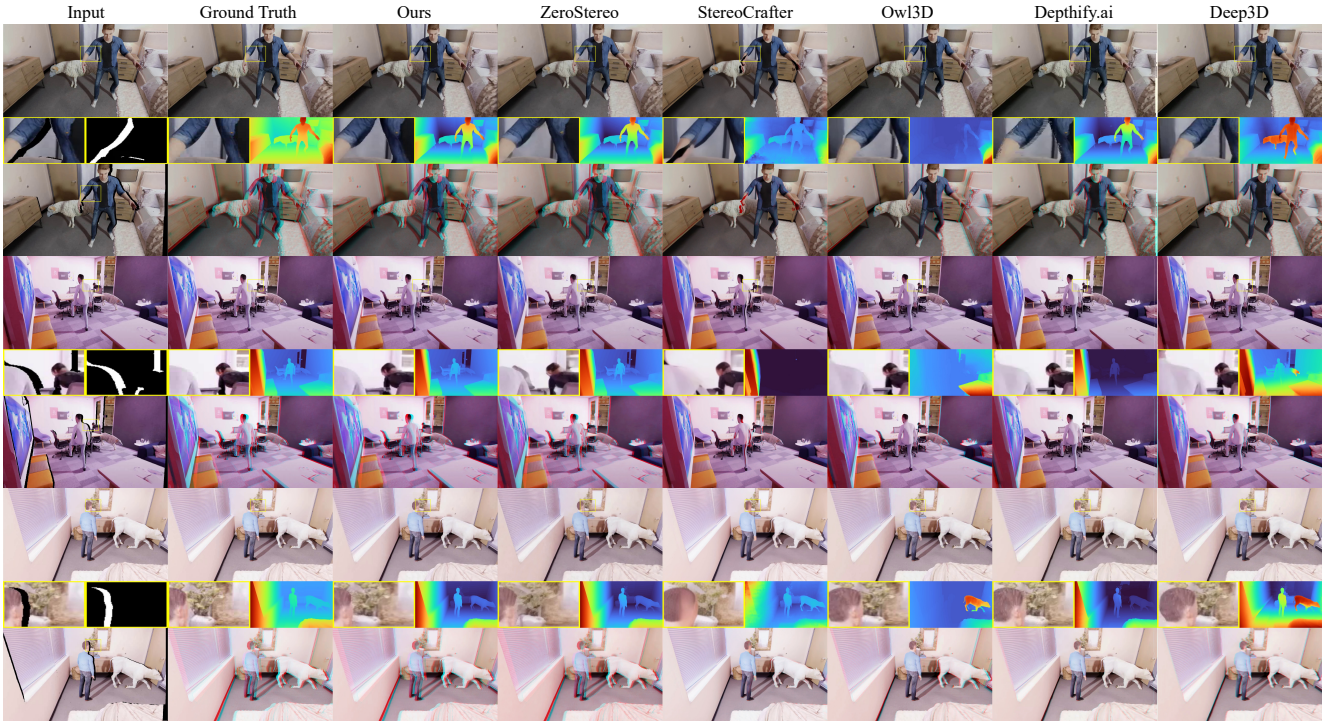

Figure 14. **Qualitative comparison on the Dynamic Replica valid set**(720×1280)**.** Our method yields the most accurate and artifact-free stereo completions, consistent with quantitative improvements in Tab. 2.

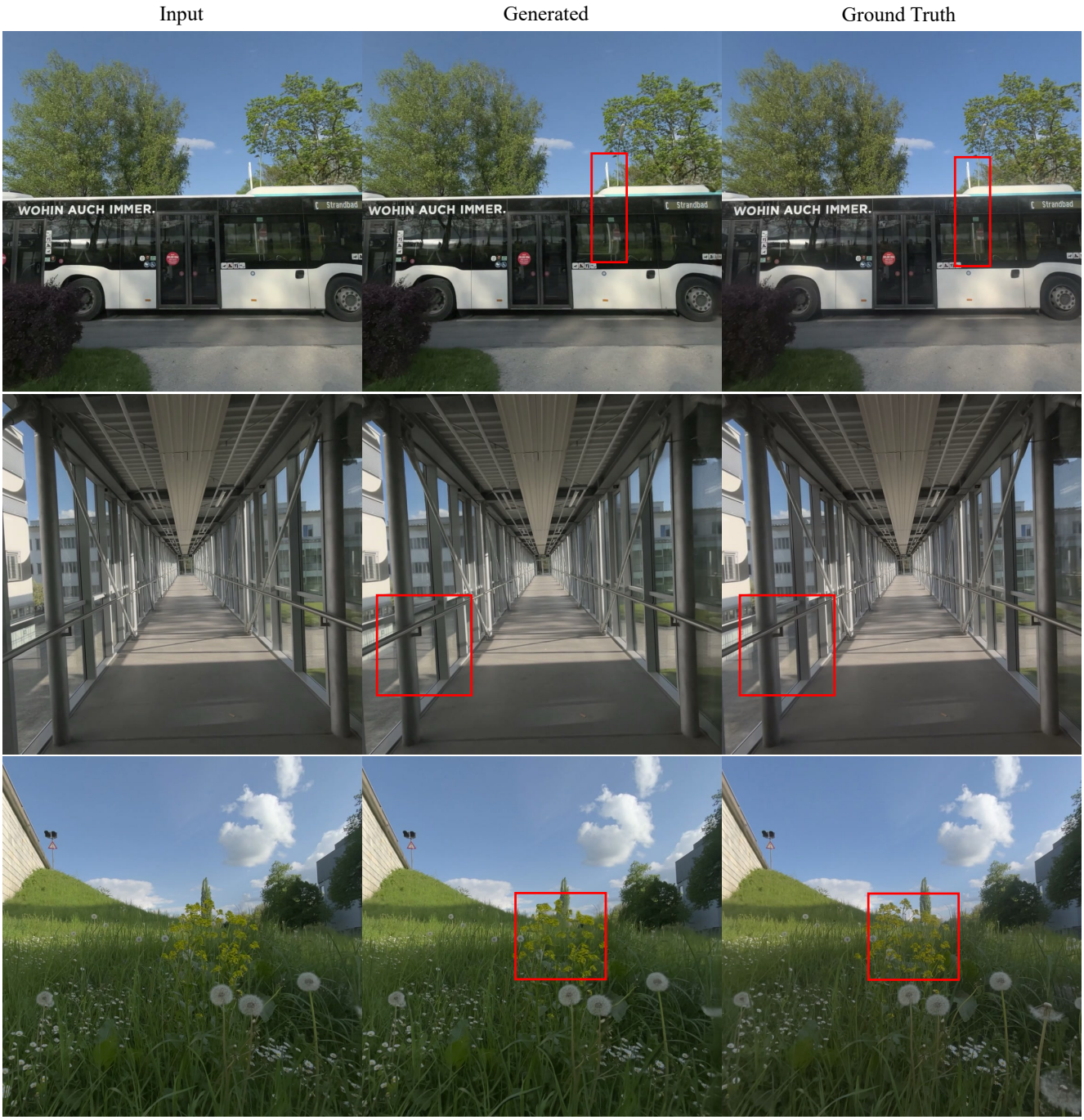

Figure 15. Representative failure cases on transparent, reflective, and highly textured scenes from the SVD AVP test set.