Elastic3D: Controllable Stereo Video Conversion with Guided Latent Decoding

Nando Metzger[1,2,*] Prune Truong[2] Goutam Bhat[2] Konrad Schindler[1] Federico Tombari[2,3]

[1]ETH Zurich [2]Google [3]TU Munich

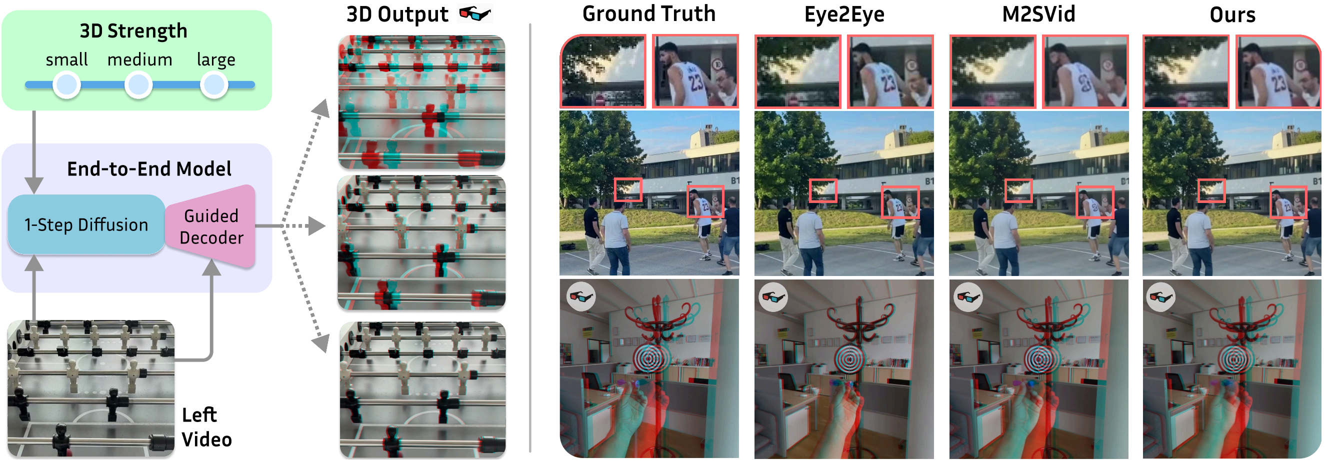

Figure 1. We introduce a direct, warping-free and feed-forward approach for mono-to-stereo video conversion. Our latent diffusion model (left) bypasses the need for depth estimation, generating the right views directly from the input left views in a single step. We achieve control over the 3D strength by conditioning on a scalar disparity factor, and employ a left-view guided VAE decoder to preserve high-frequency details and eliminate binocular rivalry artifacts. As opposed to M2SVid [50], our approach seamlessly transfers high-frequency details from the left to the right videos (top-right). Unlike Eye2Eye [13] which lacks the ability to control the 3D strength in generated content and limits its practical use, we can generate videos for any disparity distribution, as shown in the anaglyph (bottom-right).

Abstract

The growing demand for immersive 3D content calls for automated monocular-to-stereo video conversion. We present Elastic3D, a controllable, direct end-to-end method for upgrading a conventional video to a binocular one. Our approach, based on (conditional) latent diffusion, avoids artifacts due to explicit depth estimation and warping. The key to its high-quality stereo video output is a novel, guided VAE decoder that ensures sharp and epipolar-consistent stereo video output. Moreover, our method gives the user control over the strength of the stereo effect (more precisely, the disparity range) at inference time, via an intuitive, scalar tuning knob. Experiments on three different datasets of real-world stereo videos show that our method outperforms both traditional warping-based and recent warping-free baselines and sets a new standard for reliable, controllable stereo video conversion. Please check the project page for the video samples `elastic3d.github.io`.

1. Introduction

The proliferation of virtual and augmented reality hardware has created a surge in demand for immersive 3D content. With the vast majority of video content being monocular, computational mono-to-stereo conversion has become an important capability. The dominant paradigm follows a two-step warp-and-refine process: first estimate scene depth from the monocular input; second, reproject the input to a new viewpoint via the depth map [7, 19, 50, 71]. This warping process inevitably creates holes and artifacts, particularly in disoccluded regions (areas newly visible in the target view), which must then be inpainted or refined to produce a complete image.

This approach, however, faces challenges. The reliance on a separate depth estimator makes the pipeline more complex and brittle. Its quality often critically depends on the accuracy of the intermediate (monocular) depth estimator, which can struggle with certain scenarios, e.g. thin structures and non-lambertian surfaces. Moreover, most of these approaches [50, 71] rely on Latent Diffusion Models (LDMs), operating in a compressed latent space for effi-

[*] Work conducted during an internship at Google.

ciency. However, their general-purpose decoders create an information bottleneck, failing to reconstruct details from the source view (see the missing and distorted details of M2SVid in Fig. 1 right). This loss of detail, or the hallucination of new, inconsistent details, leads to a disorienting visual artifact known as *binocular rivalry* [5], where the viewer's eyes are presented with conflicting information.

To overcome these issues, the recent Eye2Eye [13] method directly generates the second viewpoint conditioned on the monocular input, sidestepping explicit depth estimation and warping. While promising, it has left a major challenge unsolved, namely, the lack of geometric control. A key advantage of warp-based methods is that users can easily adjust the 3D effect by simply scaling the intermediate depth map. In contrast, Eye2Eye is trained for a fixed, implicit baseline and offers no mechanism to control the stereo disparity, limiting its artistic and practical flexibility. Moreover, while its reliance on a two-stage refinement approach with a pixel-space diffusion model largely improves the stereoscopic fidelity of the generated video, it is prohibitively slow.

Our primary contribution is a surprisingly straightforward, warping-free model (Elastic3D– Epipolar-guided LAtent STereo vIdeo Conversion) for stereo video conversion that elegantly solves both of these challenges. The model, based on Stable Video Diffusion [3], directly synthesizes the second view in a latent space and features two innovations: (1) a conditioning mechanism that provides intuitive user control over the strength of the stereo effect by simply setting a scalar (continuous) "median disparity" at inference time. And, (2) a new guided VAE decoder, which receives high-resolution information from the source view, bypassing the latent bottleneck, to support the reconstruction of fine details and minimize binocular rivalry.

The field of video conversion lacks a comprehensive evaluation protocol for the entire mono-to-stereo conversion pipeline. Previous approaches focus only on some aspects, such as inpainting quality. Hence, as a necessary side effect of our approach, we contribute a comprehensive evaluation protocol for the entire stereo conversion pipeline. Our framework is designed to achieve a holistic quality assessment, by quantifying and comparing stereo videos along four essential dimensions: overall quality taking into account 3D-strength controllability, stereoscopic fidelity, geometric correctness, and temporal consistency.

To summarize, the contributions of this paper are: (1) a direct, *warping-free and feed-forward model for stereo video conversion* that does not explicitly depend on any depth estimation; (2) a novel *conditioning method that gives the user control over the stereoscopic disparity*, trained only on the wild rectified video data without known calibration; (3) a *guided VAE architecture that minimizes binocular rivalry* in the latent diffusion framework; (4) an improved, *comprehensive evaluation protocol* that covers the geometric, perceptual and temporal dimensions of stereo video quality. Our approach to stereo video conversion sets a new state of the art across three in-the-wild datasets.

2. Related Work

We situate our work at the intersection of stereo image synthesis, controllable latent diffusion, and guided decoding.

2.1. Stereo Conversion

Classic conversion approaches typically follow a "depth-then-warp" paradigm via depth-image-based rendering (DIBR) [10, 27, 67]. While the depth estimation stage has been significantly improved by modern estimators [4, 18, 24, 25, 33, 34, 37, 38, 63], subsequent warping steps still introduce artifacts such as disocclusions and artifacts on non-Lambertian surfaces. Even recent diffusion-based methods often retain explicit warping, inheriting these limitations [36, 57]. Similar trends are observed in video generation [7, 19, 50, 67, 70, 71]. Building on pioneering warping-free works like Deep3D [62] and Eye2Eye [13], our method avoids warping artifacts by directly synthesizing the target view in latent space, advancing the state of the art in robustness, controllability, and detail preservation.

2.2. Diffusion Models

Denoising diffusion probabilistic models (DDPMs) [16, 51] have become the state-of-the-art in generative modeling. Latent Diffusion Models (LDMs) [40] scale the capabilities of DDPMs [16, 51] by operating in a compressed latent space, enabling efficient high-resolution image [35, 43] and video synthesis [3, 17]. LDMs consist of two main components: an autoencoder (typically a VAE [26]) and a diffusion model trained in the autoencoder's latent space. The VAE's encoder compresses a high-resolution image x into a lower-dimensional, usually quantized latent representation z, and its decoder reconstructs the image $\hat{x}$ from z. The diffusion model learns to denoise this latent z. While this is highly efficient, the compression to z creates an information bottleneck, discarding information details that the decoder must regenerate. Our method leverages the efficiency, temporal pretraining and generative priors of video LDMs while introducing a guided decoding mechanism to explicitly bypass this bottleneck and recover lost detail.

2.3. Guided Decoding

Re-injecting high-fidelity source information is standard practice in guided super-resolution [8, 21, 30] and pansharpening [6, 56]. In diffusion models, this is achieved via convolutional skip connections for image translation [32, 64] or "Texture Bridges" for view synthesis [70]. Geometric constraints also serve as effective guidance, notably in Epipolar Transformers [15, 49]. Most relevant to our approach, [61] utilizes epipolar attention for simultaneous multi-view generation. We distinguish our work by integrating a lightweight epipolar attention module directly into the VAE decoder; this creates structured skip connections that bypass the bottleneck by sampling details along geometrically plausible lines.

2.4. Controllable Generation

Significant progress has been made in controlling generative models using spatial conditions, such as depth or edge maps in ControlNet [68] and T2I-Adapters [31]. Alternatively, cross-attention mechanisms allow for conditioning on learned embeddings to control style and object identity [11, 42, 65] or geometric parameters like camera trajectories [66]. However, these methods do not address stereoscopic layouts. Our work is the first to propose a mechanism for controlling the stereo effect via a simple and intuitive scalar disparity proxy.

3. What Makes a Good Stereo Model?

Given an input "left-eye" video $V_L \in \mathbb{R}^{N\times H\times W\times 3}$, consisting of N frames of resolution $H \times W$, the goal of monocular-to-stereo video conversion is to synthesize a "right-eye" video $\hat{V}_R \in \mathbb{R}^{N\times H\times W\times 3}$ that shows the same scene as if viewed from a horizontally shifted camera position. In this section, we first describe the desired properties of such a stereo conversion model. For a comfortable and immersive 3D experience, the generated video pair $(V_L, \hat{V}_R)$ must satisfy the following key properties:

1. **Geometric Correctness.** The stereo pair must adhere to the epipolar constraint and provide plausible depth perception corresponding to a rectified stereo setup [28, 72], with accurate relative depth ordering.
2. **3D-Effect Control.** A stereo conversion model must provide continuous (ideally intuitive) control over the 3D strength. This is crucial for both content creators for creative control, and end-users to adjust the immersion level for personal comfort [54].
3. **Stereoscopic Fidelity and Detail Preservation.** To avoid distracting binocular rivalry [5], each synthesized view must losslessly transfer textures from the source left view [2, 45, 47]. Any inconsistency in shared regions degrades the user experience [72].
4. **Plausible Disocclusion Handling.** The model must realistically inpaint regions that were occluded in the input left view by utilizing information from the neighboring frames whenever possible [29, 46, 48, 59].
5. **Temporal Stability.** The generated video must be temporally consistent for a comfortable viewing experience [28, 54].

Achieving all five properties simultaneously is the central challenge. Most existing stereo conversion methods utilize pretrained video diffusion models [7, 13, 50, 71] which inherently provide strong inpainting capability and temporal stability. The warp-based methods [7, 50, 71] rely on explicit depth estimation to reproject the input left views, providing effective '3D-Effect Control'. However, this requires using a separate monocular depth model, increasing system complexity and latency. Moreover, the geometric correctness of the method is bounded to that of the depth model. These methods also suffer from binocular rivalry due to the lossy compression introduced by latent diffusion models. The recent Eye2Eye [13] approach mitigates the dependency on a separate monodepth model by directly synthesizing the right view. However, it does not provide any mechanism to control the 3D effect. Furthermore, it is prohibitively slow due to a two-stage refinement process and the usage of a multi-step pixel-space diffusion model [1]. In this work we aim to address the aforementioned issues.

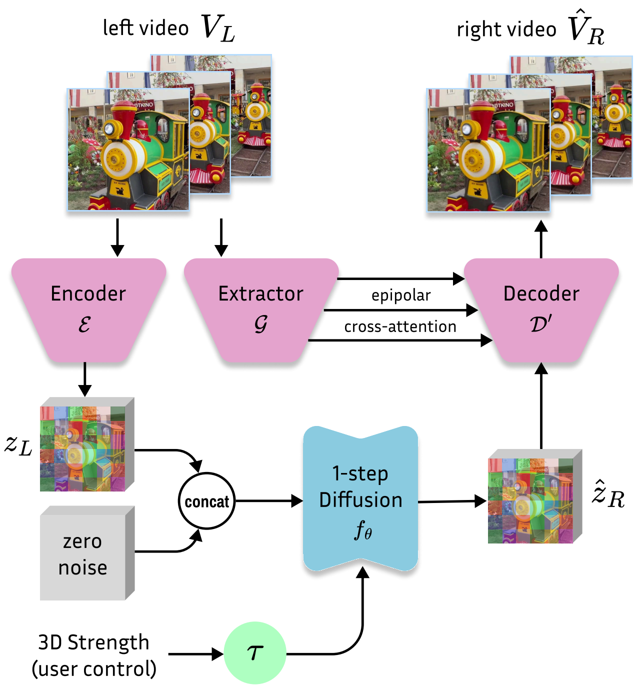

Figure 2. **Inference Pipeline.** A frozen VAE Encoder $\mathcal{E}$ computes the latent code z_L of the input video V_L. The synthesis network f_θ then generates the right-view latent $\hat{z}_R$, conditioned on z_L and on a 3D strength control token $\tau(\delta)$ (Sec. 4.2). Finally, our Guided Decoder $\mathcal{D}'$ (Sec. 4.3) renders the high-fidelity output $\hat{I}_R$, using both the generated latent $\hat{z}_R$ and the original video V_L as guidance.

4. Methodology

We present a novel, warping-free latent diffusion pipeline that aims to satisfy the key stereo model requirements listed in Sec. 3. An overview of our inference pipeline is shown in Fig. 2. Given an input video, our 1-step latent diffusion model directly generates the right camera views, without relying on any explicit depth input (Sec. 4.1). Crucially, our diffusion model is conditioned on a scalar disparity factor, enabling seamless control over the amount of '3D Effect' in the generated stereo video (Sec. 4.2). Furthermore, we introduce a left-view guided VAE decoder to effectively transfer the high-frequency details from the input left view to the generated right view, mitigating binocular rivalry (Sec. 4.3). We describe these key contributions in the next sections.

4.1. Feed-forward Warping-Free Synthesis

We build our stereo conversion model upon a latent diffusion model, Stable Video Diffusion [3]. We utilize the 1-step denoising process [12], effectively turning the latent denoiser into a feed-forward model. Unlike warp-based pipelines [50, 71] which often use the diffusion model to *refine* a warped video, our model f_θ *directly* synthesizes the entire right-video latent $\hat{z}_R$ conditioned on the left-video latent z_L and a disparity-conditioning signal $\tau(\delta)$ (Sec. 4.2). This yields an output:

$$\hat{V}_R = \mathcal{D}'(f_\theta(\mathbf{0}, z_L, \tau(\delta)), V_L), \quad \text{with} \quad z_L = \mathcal{E}(V_L) \quad (1)$$

where $\mathcal{E}$ is the VAE encoder and $\mathbf{0}$ is a zero vector. $\tau(\delta)$ refers to our *Disparity Conditioning Mechanism* (Sec. 4.2) while $\mathcal{D}'$ is the *Guided Decoder*, described in Sec. 4.3. Our feed-forward approach can efficiently generate the right views without performing numerous denoising iterations. It also enables training with image-space losses, improving the quality of the generated views, as shown in [50].

4.2. 3D-Controllable Synthesis

As discussed in Sec. 3, it is important that a stereo conversion model provides an intuitive control over the strength of the 3D effect in the generated videos. We introduce a novel disparity conditioning mechanism to enable this in our warp-free method. Our synthesis model f_θ is conditioned on a scalar disparity proxy, $\delta \in \mathbb{R}$, which controls the amount of pixel disparity between the input view and the generated right view. This provides a simple and intuitive '3D strength' knob at inference and, crucially, removes any dependency on camera parameters, enabling training on uncalibrated 'in-the-wild' rectified video pairs (V_L, V_R).

We project δ into a token embedding $\tau(\delta)$ and inject it into the model's spatial attention layers:

$$\hat{z}_R = f_\theta(\mathbf{0}, z_L, \tau(\delta)) \quad (2)$$

Conditioning δ during training: We compute δ using the ground-truth disparity maps $D^0_{L\to R} \in \mathbb{R}^{H\times W}$ from the first frame of the left to the right video.

$$\delta = \mathrm{P}_{50}(D^0_{L\to R}) \quad (3)$$

The median disparity is chosen as it is insensitive to outliers and provides an interpretable measure of the scene's overall stereo effect. However, we found that both the average and the max disparity are also adequate choices (see Appendix).

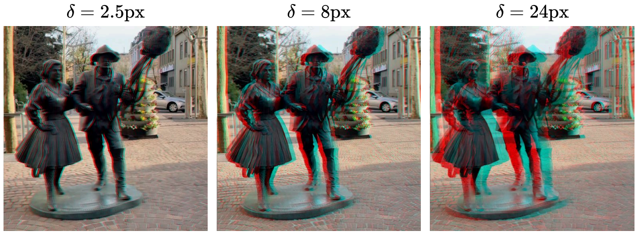

Figure 3. The strength of the stereo effect can be controlled by varying the parameter δ that acts as a conditioning for the median disparity of the generated video (see Sec. 4.2).

Conditioning δ at inference: A user can control the 3D effect by adjusting the δ in pixels, as shown in Fig. 3.

Training strategy: To train the model, we use pairs of left V_L and ground-truth right V_R frames. The model is trained to minimize a composite loss function for 1-step diffusion models [12, 50] consisting of equally weighted L2-latent loss, and pixel space L1, SSIM, and LPIPS objectives (backpropagated through the frozen VAE decoder).

To teach the model to generalize beyond the narrow range of baselines found in typical datasets, we employ a data augmentation strategy. For each sample, we randomly scale its ground-truth disparity map $D^0_{L\to R}$ by a factor s, generate a new warped view as the right-view GT using a simple forward warp, and use the corresponding scaled disparity map to get the new conditioning token $\tau(s \cdot \delta)$. We then train on a mixture of original (real) and augmented (synthetic) samples. For the synthetic pairs, we apply a simple L1 pixel loss, masked to exclude invalid pixels from the forward warp, while real pairs use our full composite loss.

4.3. Guided Latent Decoding

We now solve the primary limitation of latent synthesis models: Detail Preservation (Property 2, see Sec. 3). The VAE's information bottleneck is the main cause of binocular rivalry. As shown in Fig. 4, the Stable Video Diffusion Decoder struggles to even decode the GT latents, omitting or even hallucinating micro-details. This is unsurprising given the SVD bottleneck's high compression ratio of 1:48. We introduce a novel guided decoder, $\mathcal{D}'$, which bypasses this bottleneck by drawing details directly from the input left video, as illustrated in the top part of Fig. 2.

Epipolar-guided decoder: We reformulate decoding as a guided latent upsampling task. The decoder $\mathcal{D}'$ is conditioned on both the synthesized latent code $\hat{z}_R$ and the original left video V_L. An extractor network $\mathcal{G}$, initialized from the VAE encoder, processes V_L to produce frame-wise pyramids of multi-scale guidance feature maps, $\{g_1, ..., g_N\}$, which act as an information reservoir.

Inspired by novel view synthesis literature [15, 20], we introduce a lightweight epipolar attention mechanism. At

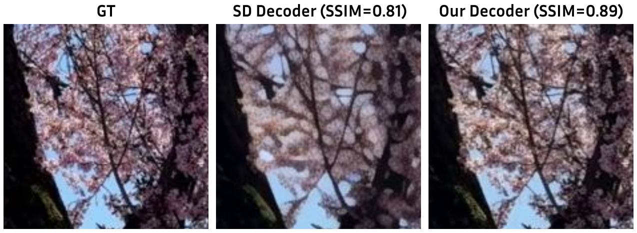

Figure 4. Decoding the ground truth latent. The compression with the standard VAE of an LDM is unsuitable for discriminative tasks.

each upsampling block i, for each feature vector $h_i(p)$ in the decoder (query), its cross-attention is computed only over the keys and values from the corresponding epipolar line in the guidance map g_i. In our rectified case, this simplifies to a one-dimensional key-query correspondence along the horizontal row. This geometrically constrained attention $\mathcal{A}_{\text{epipolar}}$ allows the decoder to "look up" and re-inject the information needed to reconstruct the image. More explicitly, the refined feature is computed residually:

$$h'_i(p) = h_i(p) + \mathcal{A}_{\text{epipolar}}(h_i(p), g_i)$$

Light-weight epipolar-attention: The epipolar constraint is not only geometrically motivated but also crucial for efficiency, reducing the complexity of full-cross-attention from a prohibitive $\mathcal{O}(H^2W^2)$ to a more manageable $\mathcal{O}(HW^2)$. For our experiments where the decoder's highest feature map is 512×512, this reduces the memory required for the 16-bit attention matrix from an infeasible 128 GB to just 256 MB.

Implementation details: We initialize our decoder $\mathcal{D}'$ with the weights of the standard decoder $\mathcal{D}$. The guidance network $\mathcal{G}$ is initialized with the weights of the encoder $\mathcal{E}$. The output projection of the attention layers $\mathcal{A}$ is initialized with zero weights, so to preserve the mapping of the original decoder at the beginning of the training.

Training strategy: We train the decoder as a separate module to disentangle the task of geometric synthesis from low-level texture reconstruction. This keeps the method modular and plug-and-play with respect to 3rd party synthesis cores f_θ, as shown in Tab. 4. The objective is to reconstruct the ground-truth right view V_R from its latent $z_R = \mathcal{E}(V_R)$, using V_L as guidance. We use a standard reconstruction loss combining L1 and LPIPS as in [70].

5. Evaluation of Stereo Conversion

A model is only as good as the benchmark used to measure it. However, there is a lack of a common evaluation protocol for the full stereo conversion task. Recent warp-based methods [36, 50] compute standard image quality metrics between the ground truth and generated right views. However, they utilize stereo depth from the ground-truth image pair [36, 50] to warp the source images. This leaks target-view information, fundamentally changing the task from robust geometric inference to simple guided inpainting. Other methods [7, 13, 50] utilize human user studies, which are very hard to scale. These methods also do not evaluate the 3D controllability of the stereo generation, which is a key requirement as described in Sec. 3.

To address this, we propose a standardized "black-box" evaluation protocol for the full stereo conversion pipeline. We require all methods—whether warp-based or end-to-end—to operate solely on the monocular input I_L at inference time. For warp-based methods, this means using their intended monocular depth estimator. In order to account for the differences in '3D strength' in the test videos (due to *e.g.* differences in stereo cameras), the methods are also provided access to global 3D information. For the warp-based methods, this entails providing the scale and shift factors to align the relative monocular depths to the pseudo ground truth metric depth. For our approach, we provide the median disparity between left and right views. We evaluate the methods on diverse videos captured from different stereo cameras in order to validate that the approaches can provide effective 3D control.

5.1. Evaluation Metrics

We assess this full pipeline using four complementary metric categories.

Overall Quality: To evaluate the overall quality of the left-to-right stereo conversion, we use standard full-reference metrics: Peak Signal-to-Noise Ratio (*PSNR*), Structural Similarity (*SSIM*) [58], and Learned Perceptual Image Patch Similarity (*LPIPS*) [69]. Note that these metrics are highly sensitive to pixel-wise alignment between the generated and GT views, and hence evaluate the combination of all properties listed in Sec. 3, i.e. geometric correctness, 3D-strength controllability, perceptual fidelity in shared and inpainted regions and temporal stability.

In order to better understand the trade-off of the different methods, we also propose "component-wise" metrics, which evaluate only one of the properties listed in Sec. 3, while being less sensitive to the others.

Stereoscopic Fidelity: Standard image quality metrics such as PSNR cannot capture specific stereo artifacts such as binocular rivalry, which is caused by mismatches in the left and right images. We thus introduce specialized metrics to evaluate this.

- *Matchability Error* ($\mathcal{E}_{\text{Match}}$) As a proxy for binocular rivalry, we measure the inconsistency of feature keypoints between views. Using a robust matcher (DeDoDe v2 [9]), we identify keypoints in I_L that have epipolar-consistent matches in the ground truth (M_{gt}) and in the prediction (M_{pred}). We define the error as the complement of their

Jaccard index,

$$\mathcal{E}_{\text{Match}} = 1 - \frac{|M_{gt} \cap M_{pred}|}{|M_{gt} \cup M_{pred}|} \tag{4}$$

A lower error indicates that the synthesized view maintains consistent, matchable texture details along the correct epipolar geometry, minimizing rivalry.

- *Patch-wise PSNR (P-PSNR)* To assess local photometric consistency between input left and generated right, while being robust to small local and global disparity shifts, we measure PSNR between 16×16 patches in I_L and their best-matching counterparts found along the horizontal epipolar line in the generated $\hat{I}_R$.

Note that as opposed to the earlier perceptual metrics, these metrics are less sensitive to precise pixel-wise alignment between the generated and GT right view and are thus partly decorrelated from the geometric correctness and 3D controllability. Nevertheless, both proposed metrics perform searches strictly along horizontal epipolar lines, inherently penalizing any vertical misalignment or rectification errors.

Geometric Correctness: We also validate that the model generates stereo videos with correct geometry and depth ordering. We estimate the dense disparity maps between the generated (D_{pred}) stereo pairs using FoundationStereo [60], and compare it with the (pseudo) ground truth disparity maps. In order to disentangle the depth ordering errors from incorrect '3D strength' and account for global scale differences, we perform a least-squares alignment between predicted and ground truth disparities and report the Mean Absolute Error between aligned disparities, as *Disp. err.*

Temporal Stability: For video, consistency over time is paramount. We measure this by computing optical flow fields for both the ground-truth and generated right-view videos using RAFT [53]. We report the End-Point Error, measuring the deviation of the generated motion from the true scene motion, which we denote as *Temp. err.*

6. Experiments

In this section, we evaluate the quality of the stereoscopic videos generated by our method, both qualitatively as well as quantitatively. Further results, analysis, visualizations and implementation details are provided in the Appendix.

6.1. Implementation Details

Training data: Following [50], we train on the **Stereo4d** and **Ego4d** datasets containing stereoscopic videos. Stereo4d comprises diverse real-world scenes and dynamic objects captured by internet videos with a fixed baseline of 63mm, while Ego4d features egocentric footage characterized by large disparities. For our work, we utilize the splits and stereo-rectified version of the data provided by [50], at a resolution of 512×512, with $N = 16$. We use the FoundationStereo [60] model to estimate the disparity maps used to compute the 3D strength δ.

Evaluation set-up: For in-domain evaluation, we rely on the test-split of Stereo4D. Results on Ego4D are provided in the Supplementary. For out-of-distribution evaluation, we use the Spatial Video Dataset [22] containing stereoscopic videos captured from Apple Vision Pro (AVP) and iPhone. The AVP portion features data taken with a similar baseline (63.8mm, i.e. close to inter-eye distance) as the training data, while the iPhone data is captured with a significantly lower baseline of 19.2mm. For the warping-based approaches [7, 50, 71], we extract relative depths with DepthCrafter [18] and align it to the GT disparity, as described in 5. The aligned mono-depth is then used to generate the warped right view. For our method, the GT disparity is used to compute the conditioning factor δ with eq. (3).

6.2. Model Ablations and Analysis

We first ablate the key components of our approach. More analysis is provided in Appendix.

Warping-free conditioning: In Tab. 1, we evaluate the impact of the proposed disparity conditioning (Sec. 4.2), which enables controlling the 3D strength in the generated stereoscopic videos. We train a baseline model without the disparity conditioning on the Stereo4D and Ego4D datasets, both of which contain videos predominantly captured with stereo cameras with baseline close to 63mm. On the test split of the Stereo4D dataset, which includes videos with similar 3D strength as those in the train split, the baseline without conditioning obtains a PSNR of 25.1 (Tab. 1, top). Adding the conditioning nevertheless allows to get more fine-grained spatial alignment of the generated and GT right videos, leading to better overall metrics (+1 dB in PSNR). On the Spatial Video iPhone set which uses a substantially different stereo baseline of 19.3mm, the baseline model without conditioning obtains poor perceptual metrics since its predictions correspond to $\approx$ 63 mm baseline and hence are spatially misaligned. Our proposed conditioning elegantly solves this issue, allowing to predict stereoscopic content of any 3D strength (see Fig. 3). Consequently, our approach obtains +3.8 dB PSNR improvement over the non-conditioned baseline (Tab. 1, bottom). Note that our conditioning does not impact the geometric accuracy (Disp. err), which only evaluates the relative depth ordering.

Warping-based versus direct: In Tab. 2, we compare

Table 1. Impact of our conditioning approach (Sec. 4.2) on Stereo4D [23] (**top**) and iPhone Spatial Video [22] (**bottom**).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{E}_{\text{Match}}$ ↓	P-PSNR ↑	Disp. err ↓	Temp. err ↓
Elastic3D w/o Cond	25.1	0.880	0.192	28.6	27.2	1.27	**1.28**
Elastic3D (ours)	**26.1**	**0.913**	**0.176**	**27.8**	**27.4**	**1.24**	1.30
Elastic3D w/o Cond	18.7	0.703	0.289	30.3	25.2	**0.64**	3.23
Elastic3D (ours)	**22.5**	**0.890**	**0.193**	**26.5**	**26.2**	0.77	**3.10**

Table 2. Impact of warping free paradigm (Sec. 4.1) on the Apple Vision Pro Spatial Video dataset [22].

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{E}_{\text{Match}}$ ↓	P-PSNR ↑	Disp. err ↓	Temp. err ↓
Warp-based	24.5	0.827	0.198	**26.8**	28.1	2.33	1.32
Elastic3D (ours)	**25.9**	**0.894**	**0.196**	30.9	**28.4**	**1.74**	**1.31**

Table 3. Reconstruction results on Stereo4D [23], where the VAE is applied on the ground truth right views.

Model	PSNR ↑	SSIM ↑	LPIPS ↓
Stable Diffusion	30.2	0.963	0.106
Guided Decoder (Ours)	**34.3**	**0.981**	**0.068**

Table 4. Impact of our guided-VAE decoder $\mathcal{D}'$ (Sec. 4.3) for mono-to-stereo conversion on the Stereo4d dataset [23].

Model	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{E}_{\text{Match}}$ ↓	P-PSNR ↑	Disp. err ↓	Temp. err ↓
Elastic3D w/o $\mathcal{D}'$	25.2	0.895	0.212	41.9	26.1	**1.23**	1.37
Elastic3D (ours)	**26.1**	**0.913**	**0.176**	**27.8**	**27.4**	1.24	**1.30**
M2SVid	24.6	0.819	0.206	39.6	26.3	**1.56**	1.35
M2SVid + $\mathcal{D}'$	**25.2**	**0.832**	**0.175**	**24.8**	**27.5**	1.61	**1.27**

Table 5. State-of-the-art comparison on the Apple Vision Pro Spatial Video dataset [22]. The baseline of the dataset is similar to during training while the content is out-of-distribution.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{E}_{\text{Match}}$ ↓	P-PSNR ↑	Disp. err ↓	Temp. err ↓
SVG	19.3	0.690	0.410	56.3	20.2	3.71	8.49
StereoCrafter	22.5	0.826	0.323	51.8	22.6	2.30	1.71
M2SVid	24.4	0.821	0.221	41.5	27.3	2.30	1.35
Eye2Eye	20.6	0.733	0.392	39.2	23.9	3.82	2.18
Elastic3D	**25.9**	**0.894**	**0.196**	**30.9**	**28.4**	**1.74**	**1.31**

our warp-free approach to a warp-based baseline that uses DepthCrafter [18] to estimate monocular depth. The geometric accuracy of the warping-based method directly depends on and is bounded by the accuracy of the pre-trained monocular depth method. In contrast, our approach implicitly learns a depth estimator for the direct generation task, achieving the lowest disparity errors. In Fig. 5, we also see that our method produces better depth ordering than warping-based alternatives. This also results in $+1.4$ dB PSNR improvement over the warp-based baseline.

Guided Latent Decoding: We analyse the impact of our guided decoder which addresses the issue of lossy compression in VAE, as shown in Fig. 4. In Tab. 3, we evaluate our guided decoder in isolation by decoding the latents of the ground truth right image. Compared to the vanilla Stable Diffusion decoder, the guided decoder improves the PSNR by 4.1 dB, and LPIPS score by 35%, by utilizing the information from the input left video. This enables better preservation of the high-frequency details, as shown in Fig. 4.

We also evaluate the impact of the guided decoder on the end stereo conversion task in Tab. 4. When employed with our proposed warping-free approach, the guided decoder generates sharper images with high-frequency details (16% lower LPIPS, $+0.9$ dB in PSNR). It also improves stereoscopic fidelity, with an increase in $+1.3$ dB in P-PSNR and a drastic reduction of 44% in Matchability error, a proxy for binocular rivalry. Moreover, the proposed guided decoder can be used as a plug and play component, replacing the standard decoder in other frameworks. For example, replacing the decoder in M2SVid [50] with ours at inference (without any specific retraining) leads to improvements in all metrics, with a particularly impressive relative improvement of 15% and 34% in LPIPS and Matchability error, respectively, showing the generality of the contribution. Finally, note that the proposed decoder does not impact the depth accuracy (Disp. err), and leads to a small improvement in temporal stability, which can be attributed to the recovery of non-flickering high-frequency details.

6.3. State-of-the-Art Comparison

We benchmark our method against recent leading approaches, i.e. SVG [7], StereoCrafter [71], M2SVid [50], ReStereo[1] [19], and Eye2Eye [13].

Qualitative results: As illustrated in Fig. 5, SVG and StereoCrafter yield blurry results. M2SVid fails to reconstruct high-frequency details (e.g., leaves and facial textures) and struggles with text (top row and Fig. 1). Conversely, our approach utilizes the guided-decoder to recover fine texture and texts accurately. Furthermore, Eye2Eye generates pixelated results and exhibits disparity mismatch with the ground truth due to its inherently fixed stereoscopic baseline. In contrast, our method produces aligned disparity with better depth ordering than warp-based approaches.

Results on AVP data: We evaluate the recent methods on the AVP dataset in Tab. 5. SVG, based on a pre-trained diffusion model without any task specific finetuning, obtains the worst metrics in $3/4$ categories. The latent diffusion model based StereoCrafter struggles with detail preservation due to the lossy VAE encoding, obtaining poor PSNR score as well as Matchability error. While M2SVid partially mitigates this through image-based losses, it cannot fully address the VAE compression losses. In contrast, despite sharing the same latent diffusion backbone, Elastic3D generates sharp images with high-frequency details, thanks to the proposed guided decoder, as evidenced by the Matchability error and P-PSNR results. Furthermore, we observe that our direct generation approach obtains lower disparity errors compared to the warping based methods, which are inherently bounding by the limitations of the off-the-shelf monocular depth estimator. The warp-free method Eye2Eye lacks any mechanism to control the 3D strength. Consequently, it produces spatially mis-aligned right views, leading to low overall quality metrics. We also noticed that Eye2Eye struggles to get the correct relative ordering of the objects in a scene (See Fig. 5), leading to higher disparity errors. On the other hand, thanks to the proposed disparity conditioning, our approach produces pixel-aligned outputs, leading to the best overall quality metrics.

[1] The authors of [19] performed the inference for us.

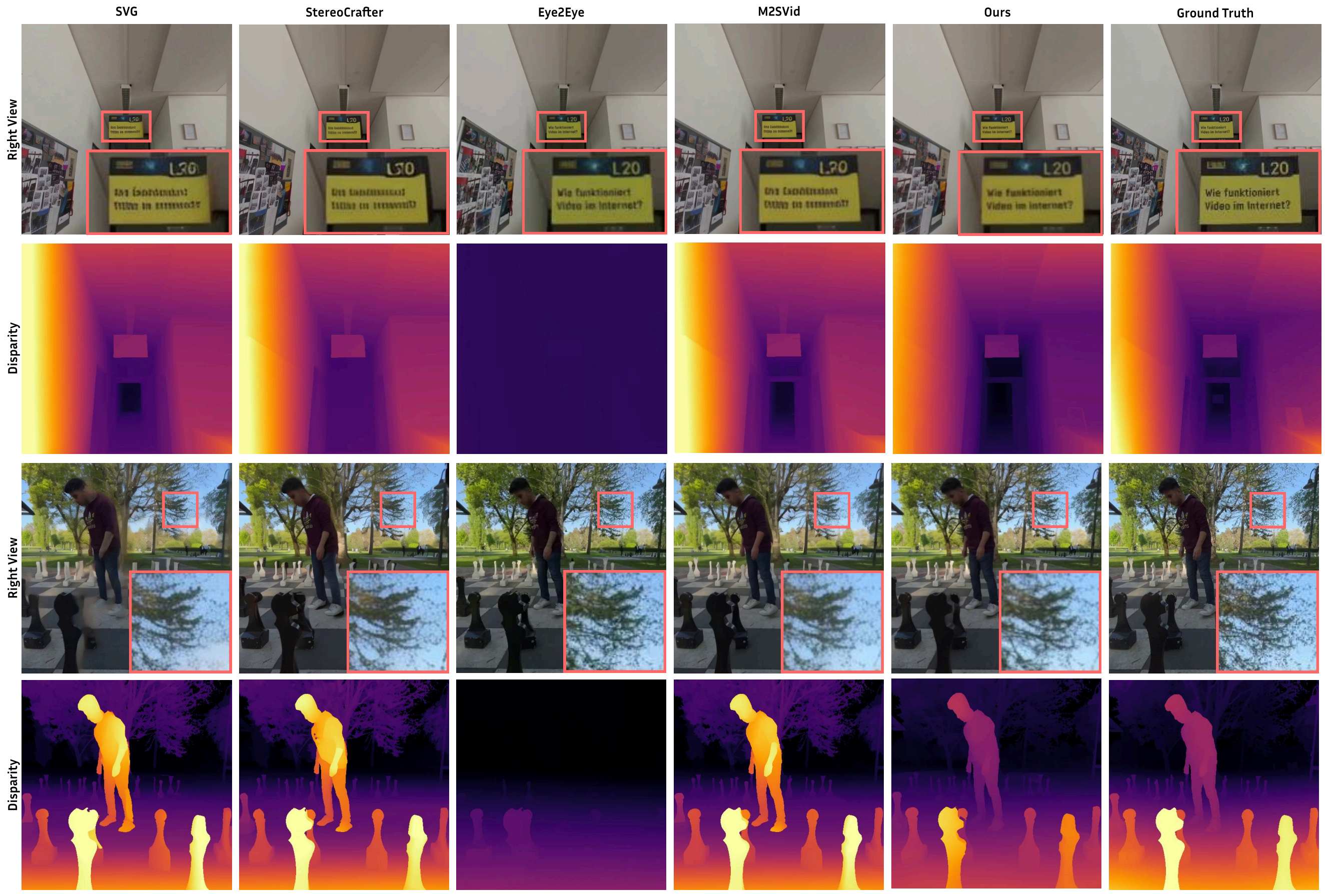

Figure 5. **Qualitative comparison.** Our method generates better textures with high-frequency details. The geometry of the generated stereo sample is also more correct – in line with the ground truth disparity.

Results on Stereo4D: ReStereo, which combines video generation with restoration, often hallucinates textures, resulting in poor stereoscopic fidelity on the Stereo4D dataset (Tab. 6). Other methods largely follow a similar trend as on the AVP set. Our method obtains the best metrics on all categories, notably achieving a +1.5 dB PSNR improvement over the second-best method M2SVid.

Results on iPhone data: In Tab. 7, we report results on the iPhone dataset, which contains videos captured with substantially different 3D strength compared to our training datasets due to differences in the employed stereo cameras. Note that compared to the warp-based methods that solely perform inpainting/refinement, it is more challenging for the direct methods to synthesize right views for different stereo setups. Nevertheless, our approach obtains the best SSIM and LPIPS scores and second best PSNR score, behind only M2SVid. Conversely, Eye2Eye obtains low overall quality metrics due to lack of 3D strength control.

Table 6. State-of-the-art comparison on the Stereo4D [23] test set.

Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	$\mathcal{E}_{\textbf{Match}}$ ↓	**P-PSNR** ↑	**Disp. err** ↓	**Temp. err** ↓
SVG	22.9	0.874	0.258	40.0	23.2	1.69	1.45
StereoCrafter	23.6	0.874	0.258	43.9	22.8	1.75	1.42
M2SVid	24.6	0.819	0.206	39.6	26.3	1.56	1.35
Eye2Eye	21.1	0.780	0.313	35.8	23.0	2.45	2.11
ReStereo	21.2	0.788	0.307	41.3	23.0	1.95	1.52
Elastic3D	**26.1**	**0.913**	**0.176**	**27.8**	**27.4**	**1.24**	**1.30**

7. Conclusion

We present a 3D-controllable, warping-free framework for converting monocular videos into high-fidelity binocular stereo videos. By integrating a novel guided VAE decoder into a disparity-conditioned latent diffusion architecture, our method ensures sharp, epipolar-consistent output while enabling intuitive scalar control over disparity, setting a new state-of-the-art across diverse datasets.

Table 7. State-of-the-art comparison on the iPhone portion of Spatial Video dataset [22]. Both calibration and video content are out-of-distribution.

Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	$\mathcal{E}_{\textbf{Match}}$ ↓	**P-PSNR** ↑	**Disp. err** ↓	**Temp. err** ↓
SVG	16.3	0.600	0.387	49.0	17.8	1.45	13.2
StereoCrafter	21.9	0.877	0.257	45.1	21.4	0.63	3.43
M2SVid	**22.9**	0.865	0.205	38.4	25.1	**0.60**	3.06
Eye2Eye	20.2	0.818	0.281	32.0	22.8	1.11	**3.02**
Elastic3D	22.5	**0.890**	**0.193**	**26.5**	**26.2**	0.77	3.10

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *ACM SIGGRAPH Asia*, 2024. 3

[2] Randolph Blake and Nikos K Logothetis. Visual competition. *Nature Reviews Neuroscience*, 3(1):13–21, 2002. 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *preprint arXiv:2311.15127*, 2023. 2, 4

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. *preprint arXiv:2410.02073*, 2024. 2

[5] B. B Breese. Binocular rivalry. *Psychological Review*, 16(6): 410–415, 1909. 2, 3

[6] Pats Chavez, Stuart C Sides, Jeffrey A Anderson, et al. Comparison of three different methods to merge multiresolution and multispectral data- landsat TM and SPOT panchromatic. *Photogrammetric Engineering and Remote Sensing*, 57(3): 295–303, 1991. 2

[7] Peng Dai, Feitong Tan, Qiangeng Xu, David Futschik, Ruofei Du, Sean Fanello, Xiaojuan Qi, and Yinda Zhang. SVG: 3d stereoscopic video generation via denoising frame matrix. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 3, 5, 6, 7

[8] Riccardo de Lutio, Stefano D'Aronco, Jan Dirk Wegner, and Konrad Schindler. Guided super-resolution as pixel-to-pixel transformation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[9] Johan Edstedt, Georg Bökman, and Zhenjun Zhao. DeDoDe v2: Analyzing and improving the DeDoDe keypoint detector. In *IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 5, 3

[10] Christoph Fehn. Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In *Stereoscopic Displays and Virtual Reality Systems XI*, pages 93–104. SPIE, 2004. 2

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *preprint arXiv:2208.01618*, 2022. 3

[12] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 4, 3

[13] Michal Geyer, Omer Tov, Linyi Jin, Richard Tucker, Inbar Mosseri, Tali Dekel, and Noah Snavely. Eye2Eye: A simple approach for monocular-to-stereo video synthesis. *preprint arXiv:2505.00135*, 2025. 1, 2, 3, 5, 7, 8

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 7

[15] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[18] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2005–2015, 2025. 2, 6, 7

[19] Xingchang Huang, Ashish Kumar Singh, Florian Dubost, Cristina Nader Vasconcelos, Sakar Khattar, Liang Shi, Christian Theobalt, Cengiz Oztireli, and Gurprit Singh. Restereo: Diffusion stereo video generation and restoration. *preprint arXiv:2506.06023*, 2025. 1, 2, 7

[20] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. EpiDiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

[21] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[22] M.H. Izadimehr, Milad Ghanbari, Guodong Chen, Wei Zhou, Xiaoshuai Hao, Mallesham Dasari, Christian Timmerer, and Hadi Amirpour. SVD: Spatial video dataset. In *ACM International Conference on Multimedia (ACM MM)*, 2025. Submitted. 6, 7, 8

[23] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4D: Learning how things move in 3d from internet stereo videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6, 7, 8

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[25] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. 2

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2013. 2

[27] Janusz Konrad, Meng Wang, and Prakash Ishwar. 2d-to-3d image conversion by learning depth from examples. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012. 2

[28] Marc Lambooij, Wijnand A IJsselsteijn, and Ingrid Heynderickx. Visual discomfort of 3D TV: Assessment methods and modeling. *Displays*, 32(4):209–218, 2011. 3

[29] Lukas Mehl, Andrés Bruhn, Markus Gross, and Christopher Schroers. Stereo conversion with disparity-aware warping, compositing and inpainting. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[30] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, 2024. 3

[32] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *preprint arXiv:2403.12036*, 2024. 2

[33] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[34] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. 2

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. 2

[36] Feng Qiao, Zhexiao Xiong, Eric Xing, and Nathan Jacobs. Towards open-world generation of stereo images and unsupervised matching. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 5

[37] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(3):1623–1637, 2020. 2

[38] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 2

[39] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020. 1

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[44] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[45] Feng Shao, Weisi Lin, Shanbo Gu, Gangyi Jiang, and Thambipillai Srikanthan. Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics. *IEEE Transactions on Image Processing (TIP)*, 22(5):1940–1953, 2013. 3

[46] Jian Shi, Zhenyu Li, and Peter Wonka. ImmersePro: End-to-end stereo video synthesis via implicit disparity learning. *preprint arXiv:2410.00262*, 2024. 3

[47] Takashi Shibata, Joohwan Kim, David M Hoffman, and Martin S Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):11–11, 2011. 3

[48] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[49] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[50] Nina Shvetsova, Goutam Bhat, Prune Truong, Hilde Kuehne, and Federico Tombari. M2SVid: End-to-end inpainting and refinement for monocular-to-stereo video conversion. In *International Conference on 3D Vision (3DV)*, 2026. 1, 2, 3, 4, 5, 6, 7, 10

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[52] Netanel Tamir, Shir Amir, Ranel Itzhaky, Noam Atia, Shobhita Sundaram, Stephanie Fu, Ron Sokolovsky, Phillip Isola,

Tali Dekel, Richard Zhang, et al. What makes for a good stereoscopic image? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 7

[53] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, 2020. 6

[54] Kasim Terzić and Miles Hansard. Methods for reducing visual discomfort in stereoscopic 3d: A review. *Signal Processing: Image Communication*, 47:402–416, 2016. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[56] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio Licciardi, Rocco Restaino, and Lucien Wald. A critical comparison of pansharpening algorithms. In *IEEE Geoscience and Remote Sensing Symposium (IGARSS)*, 2014. 2

[57] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. StereoDiffusion: Training-free stereo image generation using latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. 5

[59] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[60] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 6

[61] Haoyu Wu, Meher Gitika Karumuri, Chuhang Zou, Seungbae Bang, Yuelong Li, Dimitris Samaras, and Sunil Hadap. Direct and explicit 3d generation from a single image. In *International Conference on 3D Vision (3DV)*, 2025. 3

[62] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[64] Yang Yang, Siming Zheng, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Any-to-Bokeh: One-step video bokeh via multi-plane image guided diffusion. *preprint arXiv:2505.21593*, 2025. 2

[65] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *preprint arXiv:2308.06721*, 2023. 3

[66] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *IEEE/CVF International Conference on Cputer Vision (ICCV)*, 2025. 3

[67] Liang Zhang, Carlos Vazquez, and Sebastian Knorr. 3D-TV content creation: automatic 2D-to-3D video conversion. *IEEE Transactions on Broadcasting*, 57(2):372–383, 2011. 2

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 2

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[70] Xiang Zhang, Yang Zhang, Lukas Mehl, Markus Gross, and Christopher Schroers. High-fidelity novel view synthesis via splatting-guided diffusion. In *ACM SIGGRAPH*, 2025. 2, 5

[71] Sijie Zhao, Wenbo Hu, Xiaodong Cun, Yong Zhang, Xiaoyu Li, Zhe Kong, Xiangjun Gao, Muyao Niu, and Ying Shan. StereoCrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos. *preprint arXiv:2409.07447*, 2024. 1, 2, 3, 4, 6, 7, 10

[72] Wei Zhou, Hadi Amirpour, Christian Timmerer, Guangtao Zhai, Patrick Le Callet, and Alan C Bovik. Perceptual visual quality assessment: Principles, methods, and future directions. *preprint arXiv:2503.00625*, 2025. 3

Elastic3D: Controllable Stereo Video Conversion with Guided Latent Decoding

Supplementary Material

In Sec. A, we provide implementation details for our approach. Then, in Sec B, we explain in detail our epipolar-guided decoder cross-attention (see Sec. 4.3 of the main paper). We follow with a short background on diffusion models and a derivation of our 1-step feed-forward model in Sec. C. Next, we provide more details on the evaluation protocol in Sec. D. In particular, we give intuitive explanations and insights on the proposed metrics (see Sec. 5 of the main paper). In Sec. E, we provide additional results for our method. In particular, we conduct a user study on headset to validate that our superiority according to metrics translates to a better user experience on device. Notably, we also provide latency comparisons between our method and state-of-the-art methods. In Sec. F, we then present more ablation experiments, on more diverse datasets than in the main paper. Finally, we provide ample visual comparisons in Sec. G and highlight limitations and future work in Sec. H.

Video Results. We strongly encourage the reader to also view the accompanying `index.html` file included in the supplementary material. This interactive viewer contains more extensive video comparisons.

A. Implementation Details

Our training pipeline consists of two distinct stages: first, the training of the warping-free synthesis core (U-Net), and second, the training of the Guided Decoder. For both stages, we utilize the Stereo4D and Ego4D datasets.

A.1. Training Data Handling

Following [50], we use a batch size of 1 and apply random temporal sub-sampling to each sample to achieve a target frame rate between 5 and 30 FPS. Additionally, we apply random spatial scaling with factors in $[0.3, 1.0]$ and employ a resolution and frame length bucketing strategy for data augmentation, as detailed in Table 8.

Synthetic Baseline Augmentation. To improve the model's robustness to varying disparity ranges, we generate a synthetic subset of data. We pre-compute 2,500 random samples from Stereo4D and create synthetic stereo pairs by forward-warping the left view pixels using the ground truth depth, scaled by a random factor s. The corresponding conditioning scalar is adjusted to $s \cdot \delta$. We employ the following discrete set of scaling factors: $s \in \{0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.25, 1.5, 2.0, 3.0\}$.

Zero-Disparity Augmentation. For 1% of the training batches, we employ a zero-disparity augmentation where the target frame is identical to the input frame, and the disparity conditioning is explicitly set to $\delta = 0$.

Table 8. **Training Bucketing Strategy.** Combinations of resolution and frame counts used for augmentation.

Resolution $[H, W]$	**Frames** N	**Aspect Ratio**
$[512, 512]$	4	$1:1$
$[768, 320]$, $[320, 768]$	4	$2.4:1$
$[1024, 256]$, $[256, 1024]$	4	$4:1$
$[1280, 256]$, $[256, 1280]$	3	$5:1$
$[256, 256]$	16	$1:1$
$[192, 192]$	25	$1:1$
$[128, 256]$, $[256, 128]$	25	$2:1$

A.2. U-Net Training

We utilize Stable Video Diffusion (SVD) as our backbone, freezing the VAE and training the U-Net for the 1-step latent synthesis task.

Optimization Framework. To manage the memory footprint during training, we utilize DeepSpeed Stage 2 [39] with FP16 mixed precision. Furthermore, we optimize memory usage by offloading optimizer states to the CPU.

Loss and Hyperparameters. We employ a composite loss function consisting of an L_2 loss in the latent space, combined with pixel-space losses computed via the frozen Stable Diffusion VAE Decoder. The pixel-space objectives include L_1, SSIM, and LPIPS. We train with a learning rate of 2×10^{-6}.

Training Phases and Data Sampling. The training is performed in two distinct phases. First, we train a base unconditional model for 200k iterations. Subsequently, we fine-tune the model for an additional 170k iterations with the disparity conditioning mechanism enabled. During this second phase each batch is sampled with equal probability $(1/3)$ from three sources: the standard Stereo4D dataset, the Ego4D dataset, and the Synthetic Stereo4D subset described above.

Conditioning Mechanism. The scalar median disparity δ is projected into a high-dimensional token embedding $\tau(\delta)$. This projection is achieved by broadcasting the scalar value across the vector until it matches the required token dimensionality. This token is then concatenated with the standard CLIP embeddings of the input frame and injected into the U-Net. This strategy requires no additional learnable parameters. We found that scaling the raw disparities by a factor of $\approx 10^{-2}$ is crucial for numerical stability.

A.3. Guided Decoder Training

The Guided Decoder is trained separately to reconstruct the ground-truth right view V_R from its latent z_R, using the left video V_L as guidance.

Precision and Optimization. To avoid overflow in the introduced attention, we found that training in `bfloat16` precision is crucial. Following [70], we optimize the decoder using an equally weighted reconstruction loss combining L_1 and LPIPS objectives, with a learning rate of 1×10^{-4}. We train for 50k iterations.

B. Epipolar Guided Decoder Cross-Attention

In this section, we provide the formal definition of the epipolar cross-attention mechanism [15] within the guided VAE decoder. In short, the employed attention mechanism is a standard cross attention implementation [55] that chooses the query, key, and value locations according to epipolar geometry.

Let $h_i \in \mathbb{R}^{H_i \times W_i \times C}$ be the intermediate feature map in the decoder at layer i, and $g_i \in \mathbb{R}^{H_i \times W_i \times C}$ be the corresponding guidance feature map extracted from the input left view V_L.

For a specific pixel location $p = (u, v)$, we first compute the query vector $Q_p \in \mathbb{R}^d$ by projecting the decoder feature $h_i(p)$ using a learnable weight matrix $\mathcal{W}_Q \in \mathbb{R}^{C \times d}$:

$$Q_p = h_i(p)\mathcal{W}_Q \tag{5}$$

Similarly, for the guidance features, we restrict our scope to the corresponding horizontal scanline (epipolar line) v. We compute the keys $K_{row} \in \mathbb{R}^{W_i \times d}$ and values $V_{row} \in \mathbb{R}^{W_i \times d}$ by projecting the entire row of guidance features $g_i(v) \in \mathbb{R}^{W_i \times C}$ via learnable matrices $\mathcal{W}_K, \mathcal{W}_V \in \mathbb{R}^{C \times d}$:

$$K_{row} = g_i(v)\mathcal{W}_K, \quad V_{row} = g_i(v)\mathcal{W}_V \tag{6}$$

We then calculate the intermediate attention result $\alpha_p \in \mathbb{R}^d$ via scaled dot-product attention over the row:

$$\alpha_p = \text{Softmax}\left(\frac{Q_p K_{row}^\top}{\sqrt{d}}\right) V_{row} \tag{7}$$

The full epipolar attention module $\mathcal{A}_{\text{epipolar}}$ is obtained by projecting this result back to the original channel dimension C using a linear output projection $\mathcal{W}_{\text{out}} \in \mathbb{R}^{d \times C}$:

$$\mathcal{A}_{\text{epipolar}}(h_i(p), g_i) = \alpha_p \mathcal{W}_{\text{out}} \tag{8}$$

Finally, the decoder features are updated via a residual connection:

$$h_i'(p) = h_i(p) + \mathcal{A}_{\text{epipolar}}(h_i(p), g_i) \tag{9}$$

Zero-Initialization. To ensure the Guided Decoder preserves the identity mapping of the pre-trained decoder at the beginning of training, we initialize the weights of the linear output projection layer $\mathcal{W}_{\text{out}}$ (inside $\mathcal{A}_{\text{epipolar}}$) to zero.

C. Background on 1-Step Diffusion Formulation

Our synthesis core utilizes the diffusion backbone f_θ for 1-step inference, transforming it into an efficient feed-forward generator. We here provide background on diffusion and derive the formulation of our 1-step inference approach.

DDPMs: The core mechanism of DDPMs [16] involves learning to reverse a stochastic forward chain that gradually destroys data structure. This forward process maps a data distribution p_0 to a Gaussian noise distribution p_T across discrete timesteps $t = 1, \dots, T$. By introducing Gaussian noise with variance β_t at each step, the transition can be expressed in closed form as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$. Here, $\mathbf{x}_0$ represents the clean data sample, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes the noise term, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau$. A parameterized denoising network $\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t)$ is trained to approximate the reverse transition, effectively predicting $\mathbf{x}_{t-1}$ given the noisy state $\mathbf{x}_t$ to iteratively recover the original signal.

DDIMs: To mitigate the computational cost associated with the Markovian sampling of DDPMs, Song *et al.* [51] proposed Denoising Diffusion Implicit Models (DDIMs). By reformulating the diffusion as a non-Markovian process, DDIMs maintain the original training objective but facilitate deterministic sampling. This permits the generation of high-fidelity samples in significantly fewer steps (typically 25 to 50) compared to the full trajectory required by DDPMs.

LDMs: Directly applying diffusion models to high-resolution pixel space is computationally expensive. Latent Diffusion Models (LDMs) [40] address this by shifting the generative process into the compressed latent space of a Variational Autoencoder (VAE) [26]. The VAE comprises an encoder E and a decoder D such that $D(E(\mathbf{x})) \approx \mathbf{x}$. During the training of the diffusion backbone $\hat{\mathbf{v}}_\theta$, the VAE weights are frozen, allowing the model to learn the data distribution within a lower-dimensional space. This approach preserves the semantic structure of the data while substantially reducing computational complexity.

Conditional diffusion models: Recent advancements [43, 68] have extended the capabilities of diffusion models by incorporating conditional inputs. The denoising network is modified to the form $\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, c)$, where c represents auxiliary information. This conditioning signal enables control over the generation process using diverse modalities, including text prompts [43], reference images, or spatial guidance such as depth maps and pose estimations [68].

Training and inference with conditional LDMs: In the training phase, a data sample $\mathbf{x}$ (e.g., an image) is processed by encoding $\mathbf{x}$ into a latent representation $\mathbf{z} = E(\mathbf{x})$. This latent vector is then perturbed via the forward diffusion process:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z} + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t$ is determined by the noise schedule. Modern implementations often utilize *v-parameterization* [44] for improved stability. Instead of directly predicting the noise, the model $\hat{\mathbf{v}}_\theta(\mathbf{z}_t, t, c)$ learns to regress a target velocity $\mathbf{v}$, defined as:

$$\mathbf{v} = \alpha_t\boldsymbol{\epsilon} - \sqrt{1 - \bar{\alpha}_t}\mathbf{z}.$$

Consequently, the network—typically a U-Net [41]—is optimized by minimizing the following reconstruction objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z},c,\boldsymbol{\epsilon},t}\left[\|\mathbf{v} - \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t, c)\|^2\right]. \tag{10}$$

During inference, the system reconstructs the sample by iteratively denoising a pure noise tensor $\mathbf{z}_T$ over T steps, guided by the learned denoiser $\hat{\mathbf{v}}_\theta$ and the condition c.

1-step diffusion as feed-forward models: Garcia *et al.* [12] propose a method to adapt pre-trained diffusion U-Nets as deterministic feed-forward models for pixel-to-pixel tasks. This is achieved by fixing the sampling timestep to $t = T$ and replacing the stochastic noise component $\boldsymbol{\epsilon}$ with its expectation (zero). Under the assumption that $t = T$, we have $\bar{\alpha}_T \approx 0$, which simplifies the forward process terms to:

$$\mathbf{z}_T = \sqrt{\bar{\alpha}_T}\mathbf{z} + \sqrt{1 - \bar{\alpha}_T}\boldsymbol{\epsilon} \approx \mathbf{0}$$

and

$$\mathbf{v} = \alpha_t\boldsymbol{\epsilon} - \sqrt{1 - \bar{\alpha}_t}\mathbf{z} \approx -\mathbf{z}.$$

In this regime, the model learns to predict the clean latent structure directly from a zero-initialized input $\mathbf{z}_T = \mathbf{0}$. The standard diffusion loss in Equation 10 effectively converges to an L_2 regression between the ground truth $-\mathbf{z}$ and the prediction $\hat{\mathbf{v}}_\theta(\mathbf{0}, T, c)$:

$$\mathcal{L}_{latent} = \mathbb{E}_{\mathbf{z},c}\left[\|(-\mathbf{z}) - \hat{\mathbf{v}}_\theta(\mathbf{0}, T, c)\|^2\right]. \tag{11}$$

Rather than optimizing in latent space, Garcia *et al.* decode the predicted latents $\hat{\mathbf{z}} = -\hat{\mathbf{v}}_\theta(\mathbf{0}, T, c)$ via the decoder D to obtain a reconstruction $\hat{\mathbf{x}} = D(\hat{\mathbf{z}})$. This enables the optimization of a task-specific loss $\mathcal{L} = L_{task}(\mathbf{x}, \hat{\mathbf{x}})$ directly in the image domain.

Our own formulation: Our synthesis core utilizes the diffusion backbone f_θ for **1-step inference**, transforming it into an efficient feed-forward generator. While the original diffusion model predicts the noise ϵ iteratively, we train f_θ to directly predict the clean latent data $\hat{z}_R$ from a zero-noise input $\mathbf{0}$ and a fixed maximum timestep (or pseudo-timestep T), conditioned on the input z_L and the disparity token $\tau(\delta)$:

$$\hat{z}_R = f_\theta(\mathbf{0}, T, z_L, \tau(\delta)) = -\hat{\mathbf{v}}_\theta(\mathbf{0}, T, c) \tag{12}$$

where we have simplidied $f_\theta = -\hat{\mathbf{v}}_\theta$.

This approach utilizes the rich video and image priors learned during SVD pre-training but bypasses the computationally expensive iterative scheduler. By employing a composite loss (L2 latent + pixel-space L1/SSIM/LPIPS) to minimize the distance between $\hat{z}_R$ and the ground truth z_R, we effectively train the U-Net as a powerful image-to-image translation network for direct stereo synthesis.

D. More Details on Evaluation Protocol

First, in Sec. D.1, we analyse the robustness of our proposed metrics. Next, we provide an intuitive explanation of the matchability error and what it measures in Sec. D.2.

D.1. Metric Sensitivity Analysis

In the main paper, we introduced two component-wise metrics: Matchability Error ($\mathcal{E}_{\text{Match}}$) and Patch-wise PSNR (P-PSNR). These were designed to complement standard full-reference metrics (SSIM, PSNR) by decoupling the evaluation of *stereoscopic fidelity* (texture preservation) from *geometric correctness* and *3D-strength controllability* (3D alignment).

To validate these design choices and demonstrate the robustness of our proposed metrics, we conducted a sensitivity analysis using the Stereo4D dataset. We took the predictions generated by our model and applied two distinct post-processing degradations before computing the metrics:

1. **Horizontal Shifting:** We artificially shift the right view horizontally by N pixels. This simulates geometric rectification errors or incorrect 3D strength (disparity) predictions.
2. **Gaussian Blur:** We apply a Gaussian blur with varying σ. This simulates texture loss and over-smoothing, which are common artifacts in latent diffusion models.

D.1.1. Sensitivity to Geometric Misalignment

Standard pixel-wise metrics are heavily penalized by slight spatial misalignments, even if the textural content is perfectly preserved. In Fig. 6, we observe that standard PSNR drops precipitously with even minor pixel shifts. In contrast, our P-PSNR exhibits high robustness, maintaining a relatively good score for shifts up to $\sim$ 13 pixels. This shows that P-PSNR disentangles photometric consistency from geometric alignment, effectively finding the best matching patches along the epipolar line.

Similarly, in Fig. 7, SSIM degrades rapidly as the shift increases. Conversely, the Matchability Error remains largely stable regardless of the horizontal shift. This is expected, as the feature matcher (DeDoDe [9]) is designed to

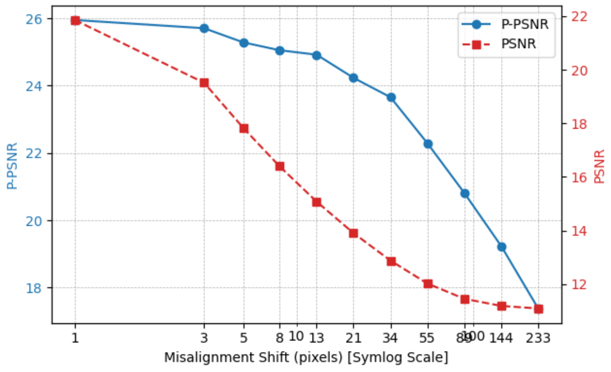

Figure 6. **Robustness to Misalignment.** Sensitivity of Standard PSNR vs. our PatchPSNR (P-PSNR) to horizontal pixel shifts. P-PSNR remains robust to small geometric errors, focusing on texture quality.

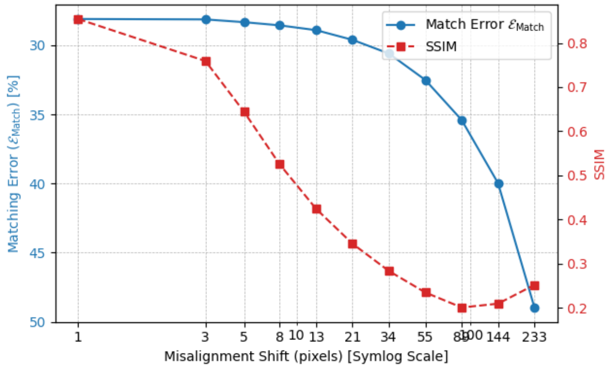

Figure 7. **Robustness to Misalignment.** Sensitivity of SSIM vs. our Matchability Error to horizontal pixel shifts. Matchability remains stable, effectively decoupling texture evaluation from geometry.

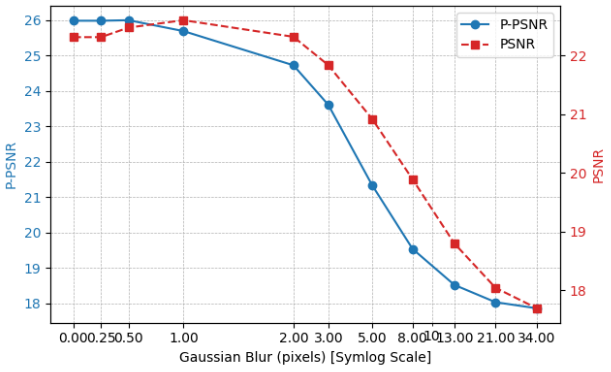

Figure 8. **Sensitivity to Blur.** Sensitivity of Standard PSNR vs. our PatchPSNR (P-PSNR) to Gaussian blur. Both metrics follow a similar trend, validating P-PSNR as a reliable quality estimator.

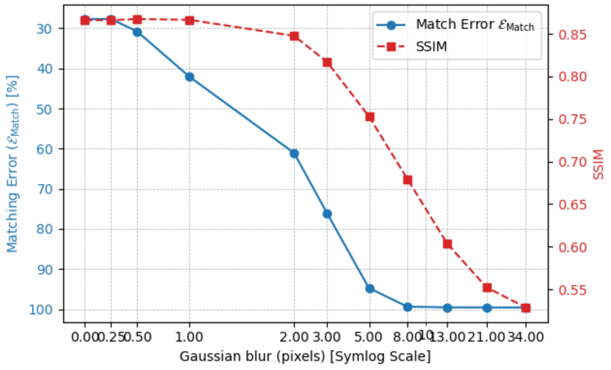

Figure 9. **Sensitivity to Blur.** Sensitivity of SSIM vs. our Matchability Error to Gaussian blur. Matchability is more sensitive to blur than SSIM, effectively penalizing over-smoothed generations.

be invariant to translation. This stability ensures that we are measuring the *presence* of matchable texture details, rather than their precise pixel location, which is evaluated separately by our geometric correctness metric (Disp. err).

D.1.2. Sensitivity to Texture Degradation

While robustness to shift is desirable, the metrics must still remain sensitive to image degradation (blur). In Fig. 8, we compare the sensitivity to Gaussian blur. We observe that P-PSNR follows the same trend as standard PSNR, confirming that it remains a valid proxy for image quality despite its spatial relaxation. Interestingly, we observe the common phenomenon where a slight blur ($\sigma = 1.0$) results in a marginal increase in PSNR ($+0.4$ dB).

Crucially, Fig. 9 demonstrates the superiority of Matchability Error in detecting over-smoothing. While SSIM decays relatively slowly with increased blur, the Matchability Error rises sharply. This indicates that our metric is significantly more sensitive to the loss of high-frequency details than SSIM. A blurry image might still look “structurally” similar (high SSIM), but it will fail to produce reliable keypoints, leading to a high Matchability Error. This makes it an effective metric for detecting the ”waxy” or ”washed-out” artifacts often produced by video diffusion models.

D.2. Matchability Interpretation and Visualization

In this section, we provide more intuitive understanding of the Matchability error, and what it measures.

Recall that the Matchability Error ($\mathcal{E}_{\text{Match}}$) is defined as the complement of the Jaccard index between the set of epipolarly-consistent matches in the ground truth (M_{gt}) and the prediction (M_{pred}):

$$\mathcal{E}_{\text{Match}} = 1 - \frac{|M_{gt} \cap M_{pred}|}{|M_{gt} \cup M_{pred}|}. \tag{13}$$

To analyze the specific nature of stereoscopic artifacts, we decompose this metric using True Positives ($N_{TP} = |M_{gt} \cap$

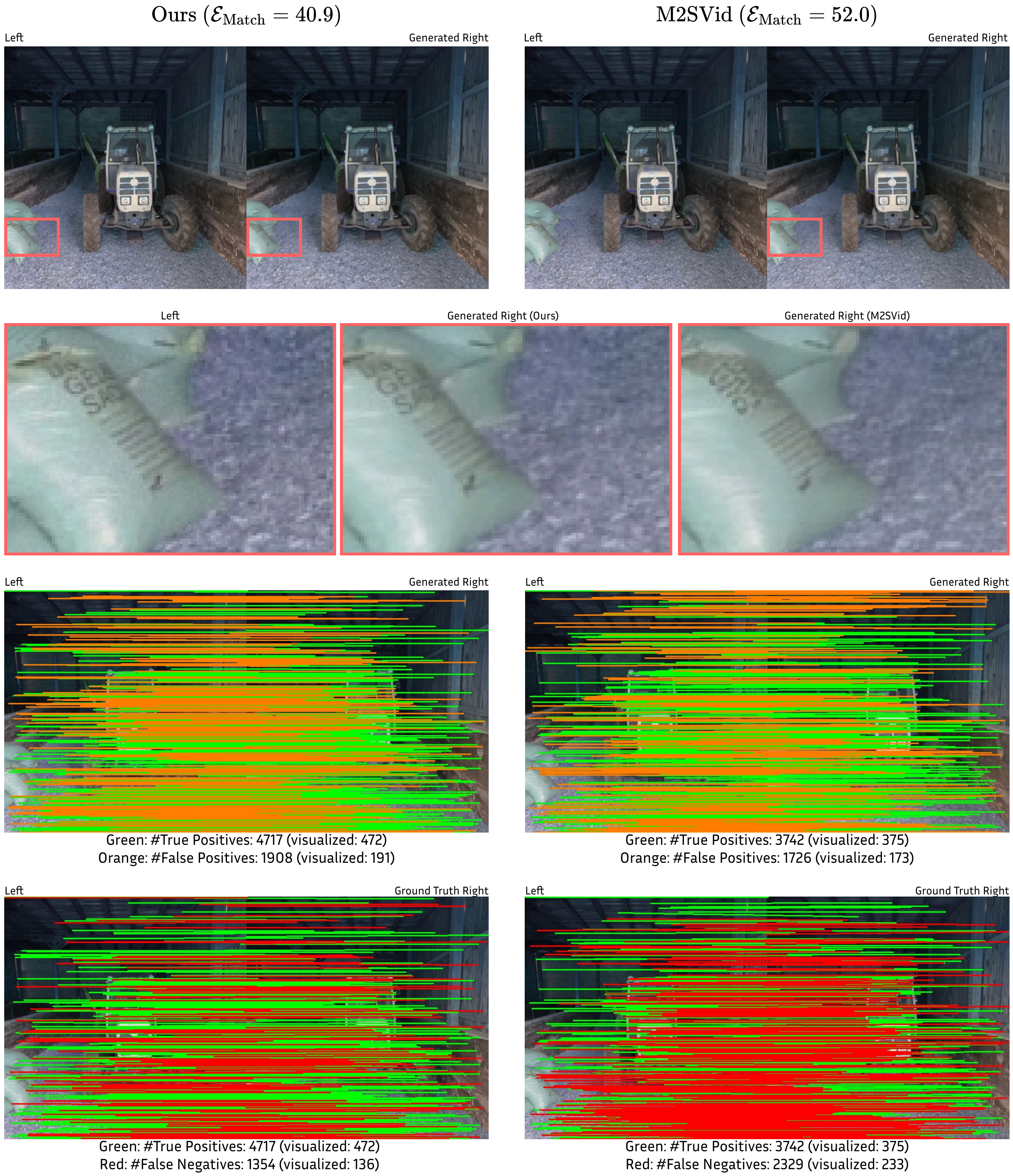
Ours ($\mathcal{E}_{\text{Match}} = 40.9$)
M2SVid ($\mathcal{E}_{\text{Match}} = 52.0$)
Left
Generated Right
Left
Generated Right
Left
Generated Right (Ours)
Generated Right (M2SVid)
Left
Generated Right
Left
Generated Right
Green: #True Positives: 4717 (visualized: 472)
Orange: #False Positives: 1908 (visualized: 191)
Green: #True Positives: 3742 (visualized: 375)
Orange: #False Positives: 1726 (visualized: 173)
Left
Ground Truth Right
Left
Ground Truth Right
Green: #True Positives: 4717 (visualized: 472)
Red: #False Negatives: 1354 (visualized: 136)
Green: #True Positives: 3742 (visualized: 375)
Red: #False Negatives: 2329 (visualized: 233)

Figure 10. **Visualization of Matchability Error Components. Green**: TP (Intersection), **Orange**: FP (Hallucination), **Red**: FN (Omission). Visualizing the error components reveals that M2SVid yields a higher error driven by significant *omission* of details (Red). We note that our model produces slightly more *hallucinations* (Orange) than M2SVid. We are visualizing only every 10th match for readability purposes. The figure is best viewed zoomed in.

Table 9. Results of Pairwise Human Perception Study. Participants compared Elastic3D against two other methods: M2SVid and Eye2Eye. The table shows the number of times (% of total for each row) each method was preferred, or if they were rated as equal.

	Total	**Preference in Pairwise Comparisons**		
Comparison (Elastic3D vs.)	*(Count)*	**Competitor Preferred** *(Count / %)*	**Equal / No Preference** *(Count / %)*	**Elastic3D Ours Preferred** *(Count / %)*
M2SVid	120	2 (1.7%)	20 (16.7%)	**98 (81.7%)**
Eye2Eye	120	15 (12.5%)	45 (37.5%)	**60 (50.0%)**

$M_{pred}|$), False Positives ($N_{FP} = |M_{pred} \setminus M_{gt}|$), and False Negatives ($N_{FN} = |M_{gt} \setminus M_{pred}|$). The error can be expressed as:

$$\mathcal{E}_{\text{Match}} = \frac{N_{FP} + N_{FN}}{N_{TP} + N_{FP} + N_{FN}}. \tag{14}$$

This decomposition allows us to categorize the fidelity of the synthesis into three distinct components:

1. **Omission Errors (False Negatives):** A high N_{FN} indicates that the model fails to generate sufficient texture detail to support feature matching, often resulting from over-smoothing or blurring in complex regions.
2. **Hallucination Errors (False Positives):** A high N_{FP} indicates that the model generates high-frequency details that satisfy epipolar constraints but does not correspond to the ground truth scene structure (e.g., texture artifacts).
3. **Feature Density (True Positives):** Crucially, the error rate must be contextualized by the absolute number of correct matches (N_{TP}). A method could artificially lower its error score by generating featureless (blurry) content, thereby minimizing both TP and FP.

A higher feature density (N_{TP}) combined with a low overall error indicates the successful preservation of sharp, matchable, and geometrically correct details.

Qualitative Analysis. We illustrate the functioning of this metric with a qualitative example in Figure 10. We visualize the ground truth left/right, as well as the right views generated by M2SVid or our approach for a particular example. From the zoomed-in example, it is evident that M2SVid loses a lot of the high-frequency details, making the text unreadable. For the generated views, we also show the true positive, false positive and false negative matches. We observe that M2SVid yields a higher error ($\mathcal{E}_{\text{Match}} = 52.0$) driven primarily by a high rate of False Negatives (Red). This confirms a loss of high-frequency detail required for robust feature matching. In contrast, our method achieves a lower error ($\mathcal{E}_{\text{Match}} = 40.9$) with a substantially higher density of True Positives (Green). While our method exhibits a slightly higher False Positive rate—indicating the generation of plausible but non-ground-truth textures—it successfully recovers significantly more valid structural details, resulting in superior overall stereoscopic fidelity.

E. More Evaluation Results

To verify that the quantitative superiority of our approach translates to better user experience on headsets, we performed a user study in Sec. E.1. We also provide results according to the iSQoE metric in Sec. E.3. In Sec. E.2, we present state-of-the-art results on the Ego4D dataset [14]. Finally, in Sec. E.4, we compare the latency of our approach to alternative state-of-the-art methods.

E.1. User Study on Headset

To evaluate the perceptual quality and geometric consistency of our generated videos, we conducted an anonymized blind pairwise comparison study on a headset. We recruited 9 participants to evaluate a diverse set of 26 scenes, selected from both the iPhone and Apple Vision Pro (AVP) datasets to ensure coverage of varying scene complexities.

Methodology. In each trial, participants were first presented with the Ground Truth (GT) video to establish a reference for correct geometry and visual details. Subsequently, they were shown two generated videos (3D vizualization of the side-by-side): one produced by our method and one by a baseline method (either M2SVid or Eye2Eye). The position of the generated videos (Sample A vs. Sample B) was randomized to prevent bias. Participants were asked to select the video that exhibited better visual quality and 3D consistency compared to the GT reference, or to select "Equal / No Preference" if the results were indistinguishable. In total, we collected 120 pairwise judgments for each baseline comparison.

Results. The results, summarized in Table 9, demonstrate a clear preference for our approach over both baselines.

When compared against M2SVid, our method achieved a dominant preference rate of 81.7% (98 votes), while the baseline was preferred in only 1.7% of cases.

Against Eye2Eye, the comparison was more competitive, reflecting Eye2Eye's high per-frame visual quality. However, our method was still preferred 50.0% of the time – four times more frequently than Eye2Eye (12.5%). A significant portion of trials (37.5%) resulted in a tie, suggesting that while Eye2Eye produces high-quality pixel-space outputs, our method achieves comparable or superior quality.

Table 10. Quantitative results on the Ego4D dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{E}_{\text{Match}}$ (%) ↓	P-PSNR ↑	MAE ↓	Flow EPE ↓
SVG	12.7	0.294	0.595	71.7	14.6	11.9	49.8
StereoCrafter	16.1	0.539	0.450	50.5	17.4	8.10	8.10
M2SVid	18.0	0.681	0.334	43.6	24.6	5.28	6.59
ReStereo	15.2	0.477	0.479	44.9	17.3	11.2	8.53
Elastic3D	**19.8**	**0.760**	**0.276**	**27.0**	**25.9**	**3.30**	**5.67**

E.2. Results on Ego4D

We present results on Ego4D [14] in Tab. 10. Our approach obtains the best performance on all metrics for all categories. SVG, which relies on a pre-trained diffusion model without task-specific fine-tuning, demonstrates the poorest performance, ranking lowest in all categories. ReStereo, which combines video generation with restoration, often hallucinates textures, resulting in poor stereoscopic fidelity. Similarly, the latent diffusion-based StereoCrafter struggles with detail preservation due to lossy VAE encoding and often produces blurry outputs, resulting in poor P-PSNR scores and high Matchability errors.

Although M2SVid attempts to mitigate this via image-based losses, obtaining largely better P-PSNR and Matchability error, it fails to fully overcome the VAE compression artifacts. In contrast, despite utilizing the same latent diffusion backbone, our approach generates sharp images with high-frequency details—evidenced by superior Matchability error and P-PSNR results—thanks to our proposed guided decoder.

Furthermore, our direct generation approach achieves lower disparity errors compared to warping-based methods, which are inherently constrained by the limitations of off-the-shelf monocular depth estimators.

Finally, by leveraging the proposed disparity conditioning, our approach ensures pixel-aligned outputs, leading to the best overall quality metrics.

E.3. Results of the iSQoE Metric

To further validate the perceptual quality of our generated stereo pairs, we utilize the iSQoE (Immersive Stereoscopic Quality of Experience) metric [52]. Unlike traditional metrics that rely on pixel-wise differences, iSQoE is a no-reference, learning-based model explicitly trained to predict human preferences in Virtual Reality environments. It outputs a single scalar value where a lower score indicates higher quality. In Tab. 11, we report the iSQoE scores across the AVP, Stereo4D, and iPhone test sets.

Metric Saturation Analysis. The quantitative results reveal a critical insight regarding the metric's sensitivity in high-fidelity regimes. As shown in the table, the scores for all state-of-the-art methods are tightly clustered between 0.505 and 0.521. Most notably, the Ground Truth (GT) scores are numerically nearly indistinguishable from the generated outputs. For the iPhone dataset, M2SVid (0.505) and our model (0.506) both achieve lower (better) scores than the Ground Truth (0.510), which should pose a theoretical lower bound. On Stereo4D, our model and M2SVid tie for the best score (0.515).

Table 11. **iSQoE Evaluation.** Comparison of stereoscopic quality on AVP, Stereo4D [23], and iPhone datasets. **Lower is better** (↓). The metric exhibits significant saturation near the optimal score.

	iSQoE ↓		
Method	**AVP**	**Stereo4D**	**iPhone**
SVG	0.519	0.521	0.508
StereoCrafter	0.509	0.519	0.508
M2SVid	**0.507**	**0.515**	**0.505**
Eye2Eye	**0.507**	0.517	0.507
ReStereo	–	0.517	–
Elastic3D	0.509	**0.515**	0.506
Ground Truth	*0.505*	*0.513*	*0.510*

This suggests that iSQoE heavily saturates when evaluating high-quality imagery. The metric was trained on the SCOPE dataset, which includes both traditional signal corruptions (noise, blur) and generative artifacts from methods like SDEdit and MotionCtrl. However, the metric appears to be highly tuned to the specific artifacts present in its training set and are presumably insensitive to inconsistencies of other latent video diffusion models. Effectively, once the image quality passes a certain threshold of "cleanliness" – a threshold met by all modern methods—the metric bottoms out around 0.50, treating high-quality synthesis and real images as perceptually equivalent.

E.4. Latency Comparison

In this section, we analyze the inference latency of our proposed method compared to state-of-the-art baselines. All evaluations were performed on a single NVIDIA H100 GPU, generating video sequences of 16 frames at a resolution of 512 × 512. Table 12 summarizes the computational cost breakdown.

Depth-Dependent Baselines (SVG, StereoCrafter, M2SVid). Existing multi-stage pipelines rely on explicit depth estimation and geometric warping as necessary preprocessing steps. Using DepthCrafter [18] with 8 denoising steps, this introduces a fixed computational overhead of approximately 3.3s for depth estimation and 0.9s for warping. SVG incurs a significantly higher warping cost (7.2s) as it necessitates warping inputs 8 times. While M2SVid boasts a very fast base inference time (0.8s), its total runtime is dominated by these preprocessing bottlenecks, resulting in a total latency of 5.0s.

Pixel-Space Diffusion (Eye2Eye). Comparison with Eye2Eye is challenging due to the non-public availability of

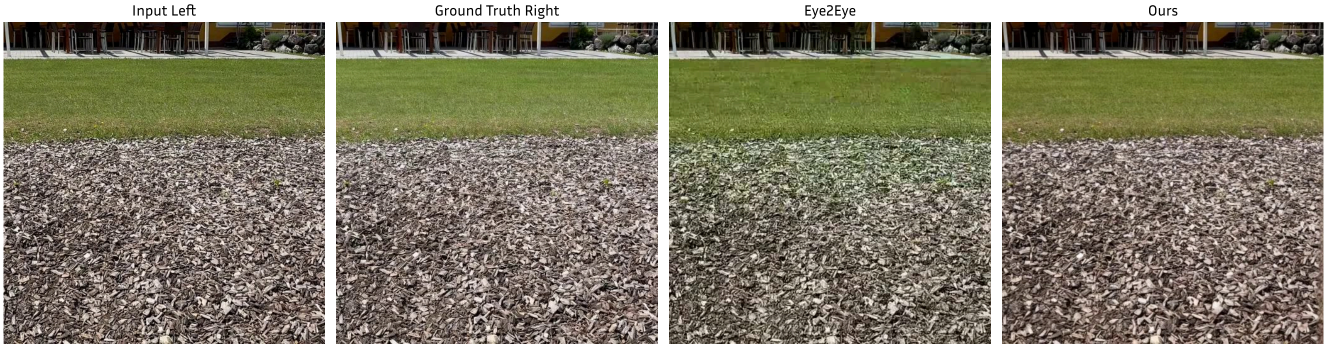

Figure 11. Eye2Eye exhibits color shifts.

Table 12. Model Latency Comparison. Runtimes are measured for generating a video of 16 frames with a resolution of 512×512. All evaluations were performed on an NVIDIA H100 GPU. “Monodepth” and “Warp” denote the time taken for depth estimation and geometric warping, respectively.

Method	Inference	Monodepth	Warp	Total
SVG	45s	3.3s	7.2s	55.5s
StereoCrafter	2.6s	3.3s	0.9s	6.8s
M2SVid	0.8s	3.3s	0.9s	5.0s
Elastic3D	1.7s	–	–	**1.7s**

the code; however, a theoretical analysis of computational complexity highlights significant efficiency gaps. First, the compute-intense Stage 2 of Eye2Eye operates in pixel space (512×512) rather than the compressed latent space (64×64) used by our method, resulting in a $64\times$ increase in spatial dimensionality. Second, it requires 50 denoising steps compared to our single-step generation. A back-of-the-envelope calculation suggests a massive difference in computational load assuming similar model sizes:

$$\underbrace{64\times}_{\text{Spatial Res.}} \times \underbrace{50\times}_{\text{Temporal Steps}} \approx 3200\times \text{ theoretical FLOPs increase.} \tag{15}$$

Ours. It is observed that our model’s pure inference time (1.7s) is higher than the M2SVid baseline (0.8s). The overhead comes from the added epipolar cross-attention and its associated feature extractor module. In other words, we trade internal model complexity for pipeline simplicity, resulting in a net reduction in total latency.

E.5. Analysis of Eye2Eye

In this section, we provide additional analysis on competitor method Eye2Eye [13].

Inference scheme: Comparison with Eye2Eye required adaptations to match their fixed inference constraints. First, the pre-trained Eye2Eye model accepts a fixed input length of 80 frames, whereas our evaluation benchmark consists of 16-frame clips. To accommodate this, we temporally padded the input by concatenating the clip and its reverse in a loop pattern $[V, V_{rev}, V, V_{rev}, V]$ to reach the required length. After inference, we extract the first 16 frames for evaluation. Second, the trained Eye2Eye model is trained to generate a *Left* view conditioned on a *Right* view condition, which opposes our benchmark task (Left $\rightarrow$ Right). To bridge this gap, we horizontally flipped our input Left views before inference—effectively treating them as Right views—and subsequently flipped the generated outputs back to obtain the final Right view. We qualitatively verified that this mirroring process did not introduce specific artifacts, such as inverted text or texture distortions.

Visual quality: We noticed that Eye2Eye generates views with widely different visual quality, depending on the inputs. On many samples, it generates high quality outputs, with preserved texture and details. However, we also noticed that on a non-neglectable amount of samples, it exhibits color shifts, as shown in Fig. 11, presumably resulting from bleeding artifacts from neighboring frames. Moreover, on other examples, it produced blurry outputs, such as on the face of the man in the second row of Fig. 5 of the main paper, the swan of Fig. 17, the leaves of Fig. 24 or the text on the bus in Fig. 23. These type of artifacts explain the relatively poor stereoscopic metrics that Eye2Eye obtains on average, in particular in terms of P-PSNR.

Geometric correctness and 3D effect control: Eye2Eye does not have any mechanism to control the amount of 3D effect that it can generate. As a result, as evidenced by the different samples in Fig. 13, 14, 16 and 22, it generates disparity maps in a different distribution than that of the ground-truth. Since the generated right views are not aligned to the ground truth, Eye2Eye obtains poor overall quality metrics (PSNR, SSIM, LPIPS), since those require pixel-wise alignment.

Table 13. Impact of our conditioning approach (Sec. 4.2 of the main paper) on AVP Spatial Video (**top**), Stereo4D (**middle**) and iPhone Spatial Video (**bottom**).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	$\mathcal{E}_{\text{Match}}$ ↓	P-PSNR ↑	Disp. err ↓	Temp. err ↓
Elastic3D w/o Cond	21.1	0.710	0.301	**23.4**	27.6	3.11	1.70
Elastic3D (ours)	**25.9**	**0.894**	**0.196**	30.9	**28.4**	**1.74**	**1.31**
Elastic3D w/o Cond	25.1	0.880	0.192	28.6	27.2	1.27	**1.28**
Elastic3D (ours)	**26.1**	**0.913**	**0.176**	**27.8**	**27.4**	**1.24**	1.30
Elastic3D w/o Cond	18.7	0.703	0.289	30.3	25.2	**0.64**	3.23
Elastic3D (ours)	**22.5**	**0.890**	**0.193**	**26.5**	**26.2**	0.77	**3.10**

F. More Ablation Studies

In this section, we provide ablation studies similar to those in the main paper on different datasets. We also ablate the signal used for conditioning.

F.1. Warping-free conditioning

In Tab. 13, we evaluate the impact of the proposed disparity conditioning (Sec. 4.2 of the main paper), which enables controlling the 3D strength in the generated stereoscopic videos, on the AVP, Stereo4D and iPhone datasets. Note that the results on Stereo4D and the iphone datasets are already provided in Tab. 1 of the main paper and reported here for completeness. We train a baseline model without the disparity conditioning on the Stereo4D and Ego4D datasets, both of which contain videos predominantly captured with stereo cameras with baseline close to 63mm. The results confirm that adding the conditioning results in better overall metrics for all datasets – the gain is especially pronounced for the out-of-domain cameras setups.

In Fig. 12, we qualitatively show the impact of the proposed disparity conditioning. We show the anaglyphs generated from the ground truth left and right view, our approach without or with conditioning. The baseline approach without conditioning has overfitted to the training baseline, close to 63mm. It is thus incapable of generating content of the correct 3D strength on the iPhone dataset, which features a smaller baseline. Conversely, the proposed disparity conditioning allows the model to generate pixel-aligned right views for any disparity distributions.

F.2. Conditioning types eq. (3) of the main paper

We compute the conditioning δ as the median of the disparity map relating the first frame of the left to the right video. In Tab. 14, we compare using the Median, Maximum or Average disparity as the conditioning signal, on the AVP, Stereo4D, iphone and Ego4D datasets. All three types of conditioning obtain similar results on the four datasets. As a result, the median disparity was chosen as it is insensitive to outliers and provides an interpretable measure of the scene's overall stereo effect. However, both the average and the max disparity are also adequate choices.

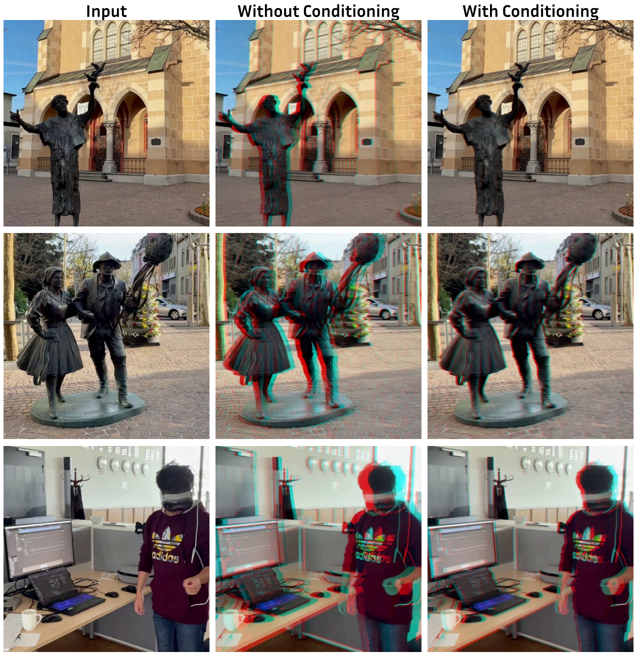

Figure 12. Input frames, and anaglyphs generated using our model trained without or with our proposed conditioning on the iPhone dataset. The iPhone's baseline of 19.2mm is significantly smaller than the one from the training datasets which is presumably 63mm. It is evident that our conditioning approach is crucial for rescaling to desired baselines.

Table 14. Impact of the disparity conditioning types (Sec. 4.2 of the main paper) on the AVP Spatial Video, Stereo4D, iPhone Spatial Video and Ego4D datasets.

Dataset	Model	PSNR ↑	SSIM ↑	LPIPS ↓
AVP	Average	**26.2**	**0.903**	**0.194**
	Max	25.2	0.860	0.217
	Median	25.9	0.894	0.196
Stereo4D	Average	**26.4**	**0.913**	**0.176**
	Max	25.6	0.897	0.185
	Median	26.1	**0.913**	0.**176**
iPhone	Average	**23.4**	**0.916**	**0.180**
	Max	22.9	0.902	0.184
	Median	22.5	0.890	0.193
Ego4d	Average	19.6	0.751	0.290
	Max	19.6	0.754	0.293
	Median	**19.8**	**0.760**	**0.276**

F.3. Guided Latent Decoding

In this section, we evaluate the impact of the proposed guided decoder (Sec. 4.3 of the main paper) on the end stereo conversion task in Tab. 15. We present results on the AVP, Stereo4D and iPhone datasets. Note that results on Stereo4D were already provided in Tab. 4 of the main

Table 15. Impact of our guided-VAE decoder $\mathcal{D}'$ (Sec. 4.3 of the main paper) for mono-to-stereo conversion on the AVP Spatial Video (**top**), Stereo4d (**middle**) and iPhone Spatial Video (**bottom**) datasets.

Model	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	$\mathcal{E}_{\textbf{Match}}$ ↓	**P-PSNR** ↑	**Disp. err** ↓	**Temp. err** ↓
Elastic3D w/o $\mathcal{D}'$	25.5	0.884	0.229	44.9	27.3	**1.71**	1.38
Elastic3D (ours)	**25.9**	**0.894**	**0.196**	**30.9**	**28.4**	1.74	**1.31**
M2SVid	24.4	0.821	0.221	41.5	27.3	**2.30**	1.35
M2SVid + $\mathcal{D}'$	**24.5**	**0.827**	**0.198**	**26.8**	**28.1**	2.33	**1.32**
Elastic3D w/o $\mathcal{D}'$	25.2	0.895	0.212	41.9	26.1	**1.23**	1.37
Elastic3D (ours)	**26.1**	**0.913**	**0.176**	**27.8**	**27.4**	1.24	**1.30**
M2SVid	24.6	0.819	0.206	39.6	26.3	**1.56**	1.35
M2SVid + $\mathcal{D}'$	**25.2**	**0.832**	**0.175**	**24.8**	**27.5**	1.61	**1.27**
Elastic3D w/o $\mathcal{D}'$	21.9	0.868	0.239	62.5	24.7	**0.74**	3.34
Elastic3D (ours)	22.5	**0.890**	**0.193**	**26.5**	**26.2**	0.77	**3.10**
M2SVid	22.9	0.865	0.205	38.4	25.1	**0.60**	3.06
M2SVid + $\mathcal{D}'$	**23.9**	**0.892**	**0.158**	**21.0**	**26.8**	**0.60**	**2.58**

paper and are added here for completeness.

On all datasets, when integrated into our warping-free framework, the guided decoder significantly enhances image sharpness and high-frequency detail, resulting in a drastic reductions in LPIPS and increases of at least $+0.9dB$ PSNR. It also creates substantial gains in stereoscopic fidelity: we observe improvements of at least $+1.2dB$ in P-PSNR and massive reductions in Matchability error, with a particularly impressive relative improvement of 58% on the iPhone dataset. This shows that our proposed guided decoder effectively minimizes binocular rivalry.

Furthermore, the proposed decoder serves as a generalized plug-and-play component. For instance, replacing the standard decoder in M2SVid [50] with ours at inference time—without any retraining—improves performance across all metrics and datasets, yielding notably strong relative improvements in LPIPS and in Matchability error.

Finally, we note that the decoder maintains depth accuracy (Disp. err) while slightly enhancing temporal stability, a benefit attributed to the consistent recovery of high-frequency details.

G. Qualitative results

Video Results (HTML). As stereoscopic video generation involves temporal dynamics that are difficult to convey through static frames, the attached HTML video viewer provides a more comprehensive comparison.

Static Results. In Fig. 13, 14, 15, 16, 17 and 18, we compare the generated right view by our approach to state-of-the-art SVG, StereoCrafter, Eye2Eye and M2SVid on multiple samples from the AVP datasets. We also show the ground truth right view for reference. Fig. 19, 20, 21, 22, 23 and 24 show similar comparisons on the iPhone dataset instead.

By skimming through the examples, it is evident that Eye2Eye struggles to generate a view that is pixel-aligned to the ground truth. The disparity in all cases is very different from the ground truth, and the distribution varies a lot depending on the input. In contrast, thanks to our proposed disparity conditioning, our approach can generate right view for any desired disparity distribution. The geometric accuracy of warping-based approaches SVG, StereoCrafter and MS2Vid entirely depends on the quality of the underlying monocular depth model that they use to warp the left video. As a result, the disparity maps, while mostly accurate, can show variable quality depending on the input samples.

In terms of visual quality, our approach recovers better texture, and high-frequency details than counterparts. This is particularly evident on texts such as in Fig. 13, or on vegetation/leaves as seen in Fig. 21, 22 and 24.

In Fig. 25, 26 and 27, we show the anaglyphs generated by our approach on in-the-wild images with varying disparity conditioning. As the pixel-disparity conditioning increases, the strength of the 3D effect also increases, as evidenced by the anaglyphs.

H. Limitations and Future Work

While our method establishes a new baseline for warping-free stereoscopic video generation, several limitations remain which outline promising directions for future research.

Inference on long and high-resolution videos: Currently, our generation is constrained by GPU memory. On a single H100 (80GB) GPU, we can generate a maximum of approximately 45 frames at 512×512 resolution. However, our methodology is theoretically extendable to arbitrary video lengths. Future work could adopt temporal autoregressive strategies similar to those employed by StereoCrafter [71] and M2SVid [50], where the video is generated in chunks, conditioning the synthesis of the current chunk on the last k frames of the previous one. Similarly, higher-resolution inference could be achieved through spatial tiling and stitching strategies [50, 71], though we leave the implementation of these scaling techniques to future work.

Extreme Disparities: While our direct synthesis approach excels at standard stereoscopic baselines, it struggles to generalize to extreme disparities (baselines $\gg$ 63mm). In these regimes, warp-based methods still hold an advantage, as they rely on explicit geometric reprojection rather than learned synthesis. Improving the data diversity to include wider baselines could mitigate this.

Ambiguity of Median Conditioning: Our use of a single scalar δ (median disparity) to control 3D strength is efficient but can be ambiguous. For example, a scene dominated by a flat background may have a median disparity near zero, even if foreground objects are present. As seen in Fig. 27, this ambiguity is pronounced in zoom-in shots: the sequence begins with a background-dominated view ($\delta \approx 0$), providing a weak conditioning signal. As the camera zooms into the subject and non-zero disparities appear, the model occasionally struggles to recover the correct stereo effect because the global conditioning token fails to capture the dynamic change in scene geometry.

Geometric Inconsistencies: While our warping-free approach eliminates the artifacts associated with occlusion filling and warping, it occasionally sacrifices strict geometric rigidity. Warp-based methods rely on explicit, pretrained depth estimators which enforce strong geometric priors — potentially ensuring, for instance, that flat walls remain perfectly planar. In contrast, our model learns geometry implicitly from the training data. As observed in our qualitative samples, this can sometimes lead to minor geometric hallucinations, such as "wobbly" depth on flat surfaces or inconsistent relative depths in complex scenes. Since the mapping from a single image to a stereo pair is ill-posed, our model effectively predicts the most probable depth configuration, which may not always align perfectly with the physical ground truth.

End-to-End Joint Training: Currently, we train the U-Net and the Guided Decoder separately. While it is technically possible to train them jointly—allowing the U-Net to be optimized directly via the decoder's pixel-space reconstruction loss—we found this to be computationally prohibitive. Joint training significantly increases the memory footprint, making the training-time versus memory trade-off impractical on standard hardware. Furthermore, our factorized approach maintains modularity, allowing the Guided Decoder to be used as a plug-and-play enhancement for other latent diffusion pipelines.

Computational Optimization: As noted in our latency analysis, the current implementation of the epipolar cross-attention mechanism relies on a row-wise Python loop. This introduces interpreter overhead and prevents full GPU saturation ($O(H)$ sequential iterations). A significant speedup could be achieved by implementing this operation as a custom CUDA kernel, enabling fully parallelized execution.

I. Acknowledgment

Nina Shvetsova for the help throughout the project and for providing the initial codebase. Special thanks goes to Diego Martín Arroyo for helping with the the infrastructure, experiments and technical discussion. Thanks also to Haofei Xu, Felix Wimbauer, Luca Zanella, Nikolai Kalischek, Mattia Segu, Michael Niemeyer for technical inputs.

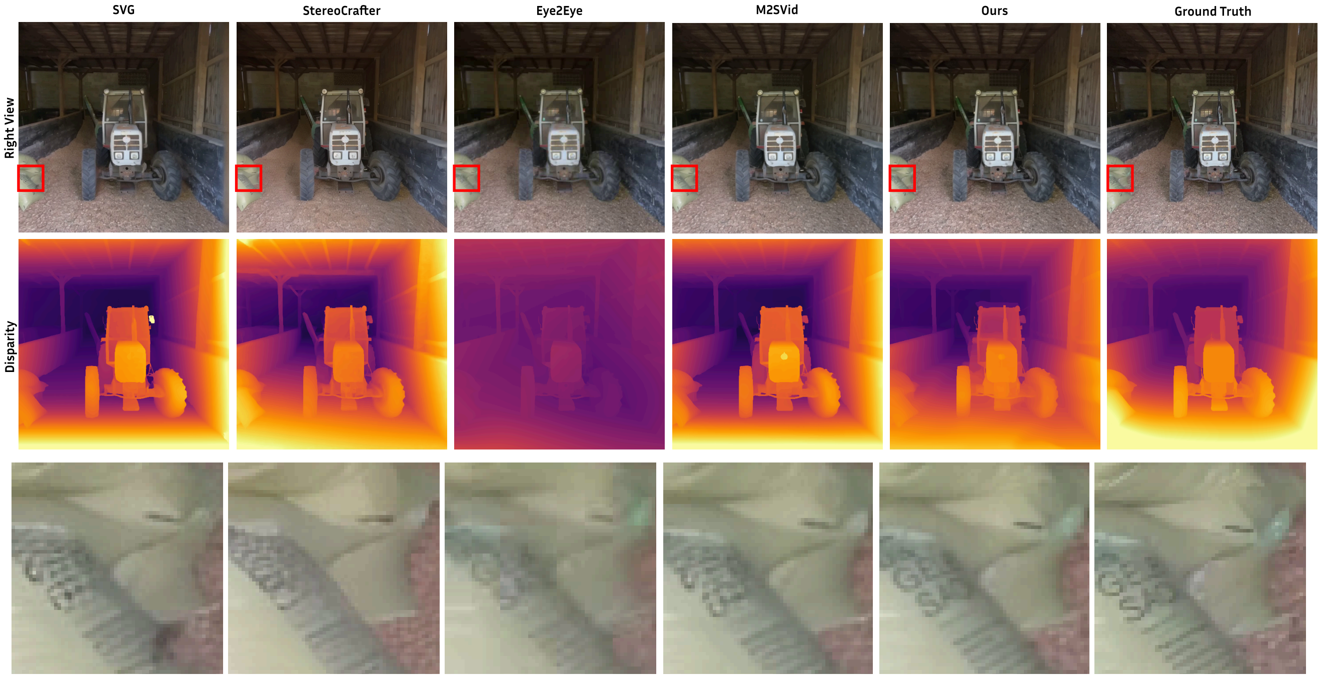

Figure 13. **Comparison on Scene 158 of the AVP dataset.** Note the more readable text on the bag and the texture of the ground.

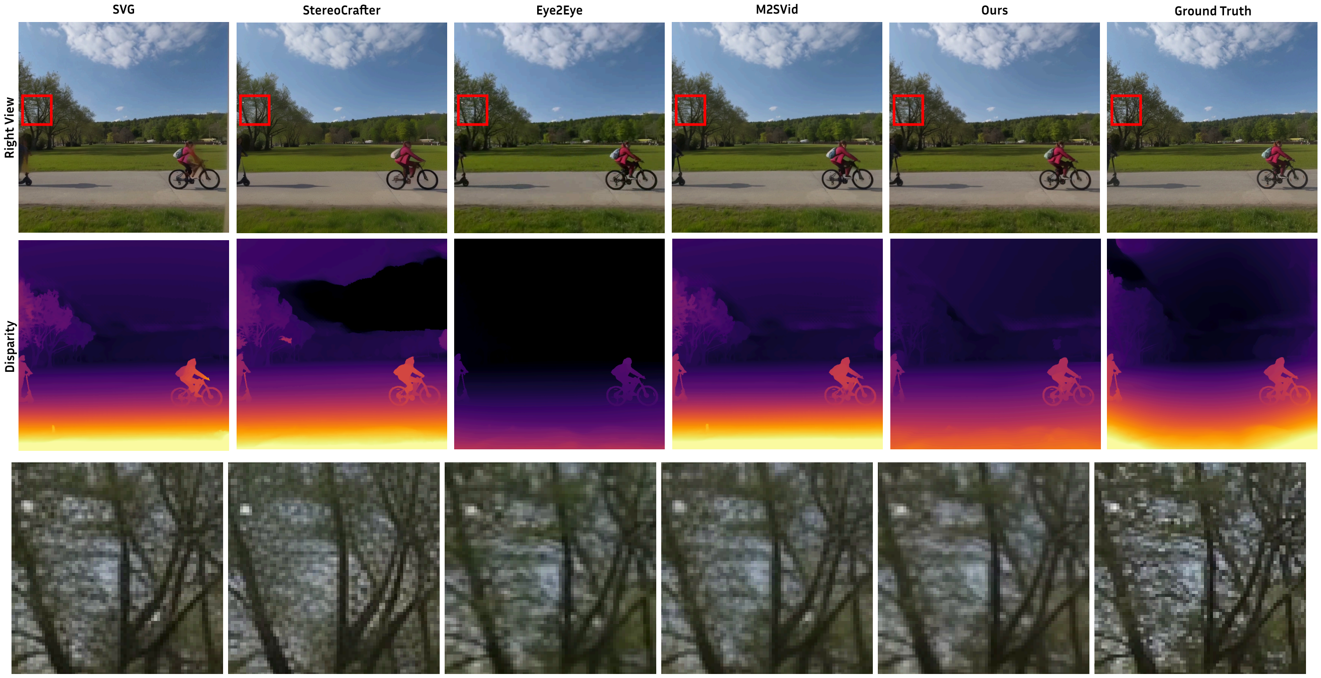

Figure 14. **Comparison on Scene 29 of the AVP dataset.** Our method preserves the branch structures in the trees.

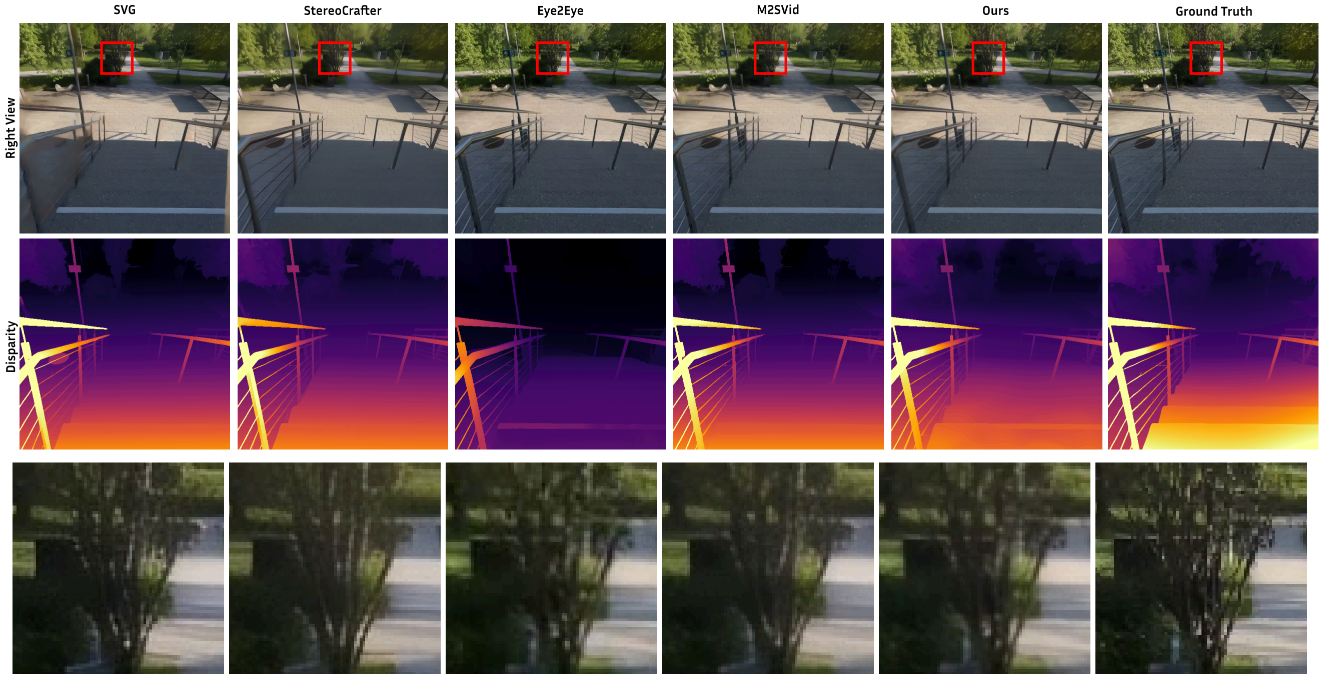

Figure 15. **Comparison on Scene 42 of the AVP dataset.** Our method better preserves the tree and background.

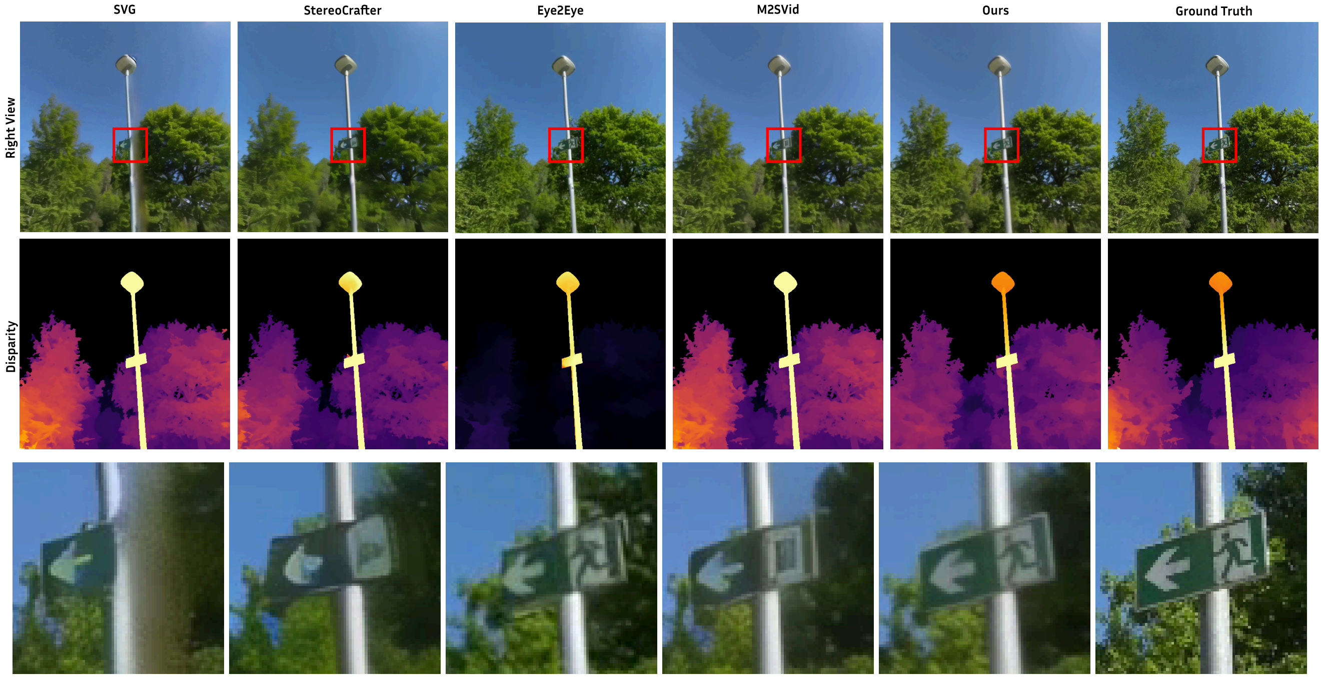

Figure 16. **Comparison on Scene 47 of the AVP dataset.** Our method recovers the details on the sign.

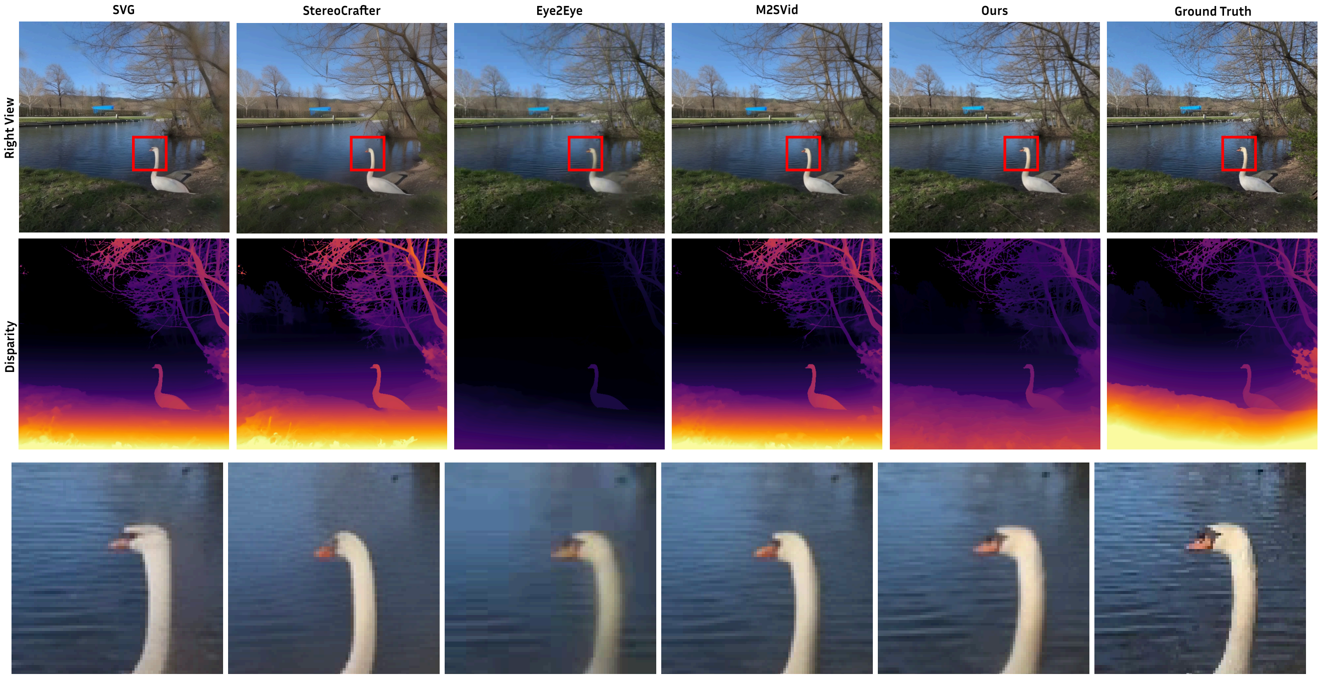

Figure 17. **Comparison on Scene 121 of the AVP dataset.** Our method generates a sharp output, as evidenced by the swan head.

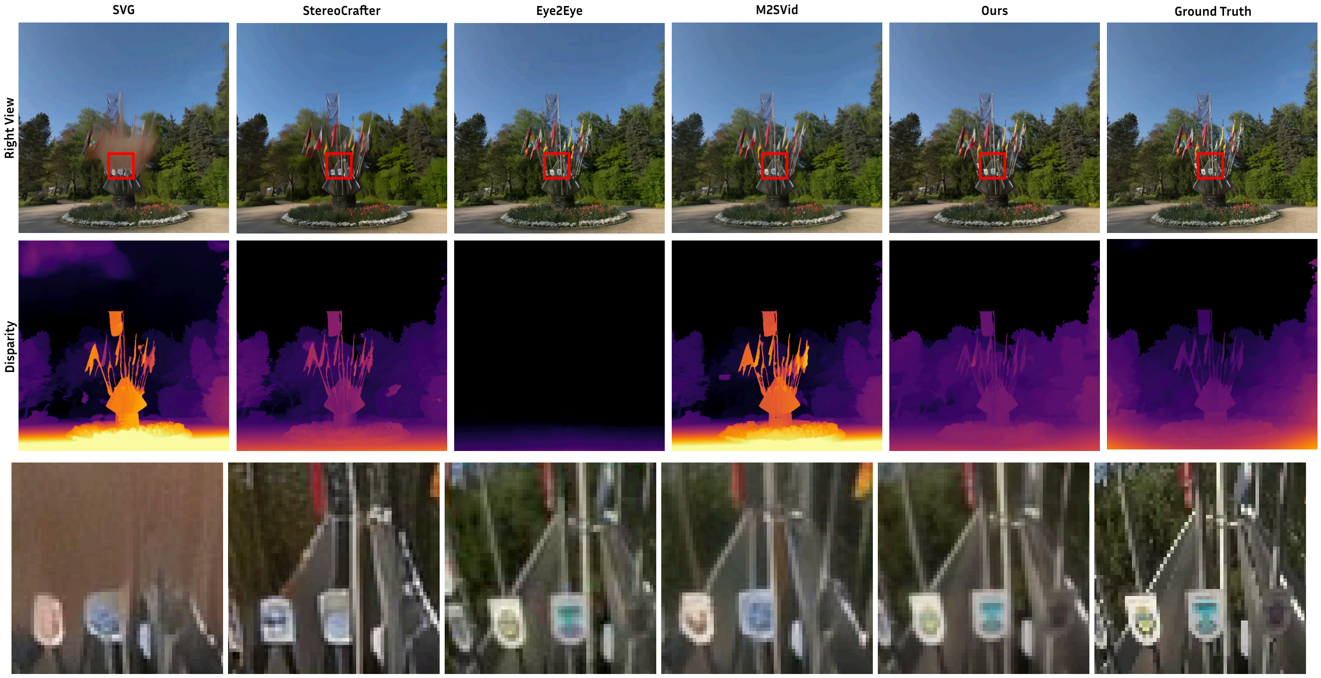

Figure 18. **Comparison on Scene 22 of the AVP dataset.** Our method recovers the high-frequency details on the signs and better geometry.

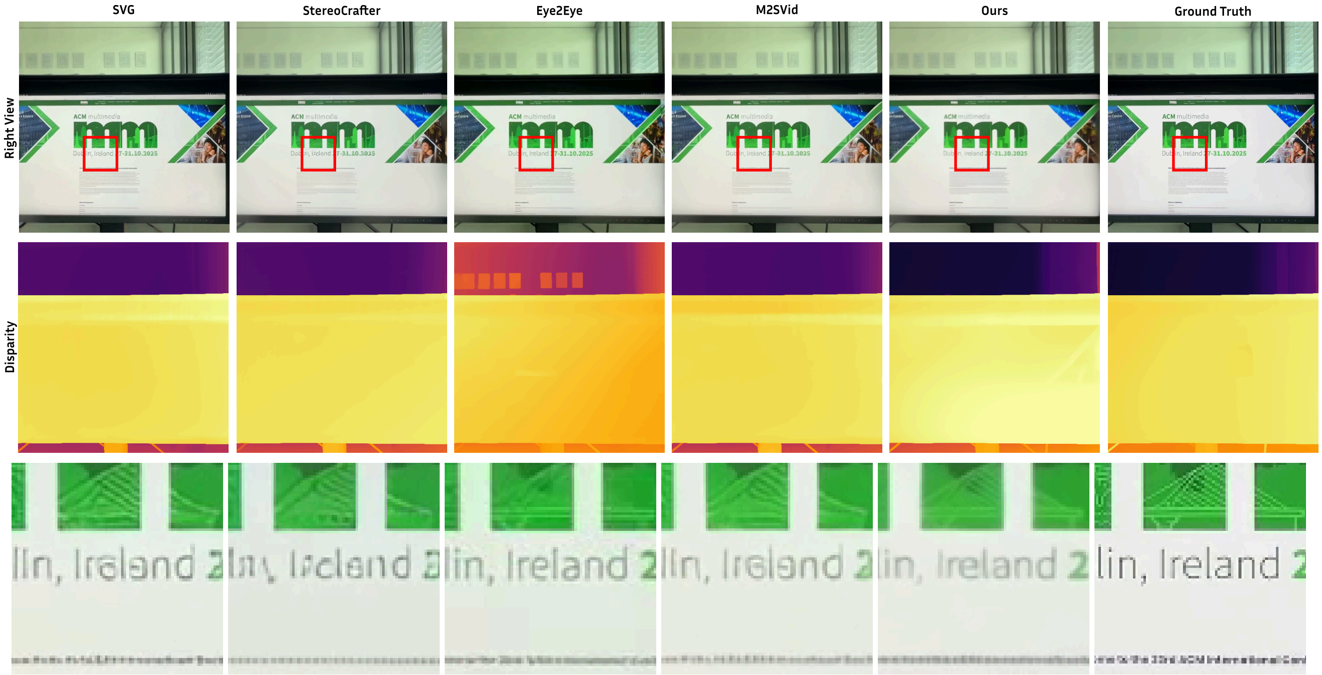

Figure 19. **Comparison on Scene 14 of the iPhone dataset.** Our method produces readable text.

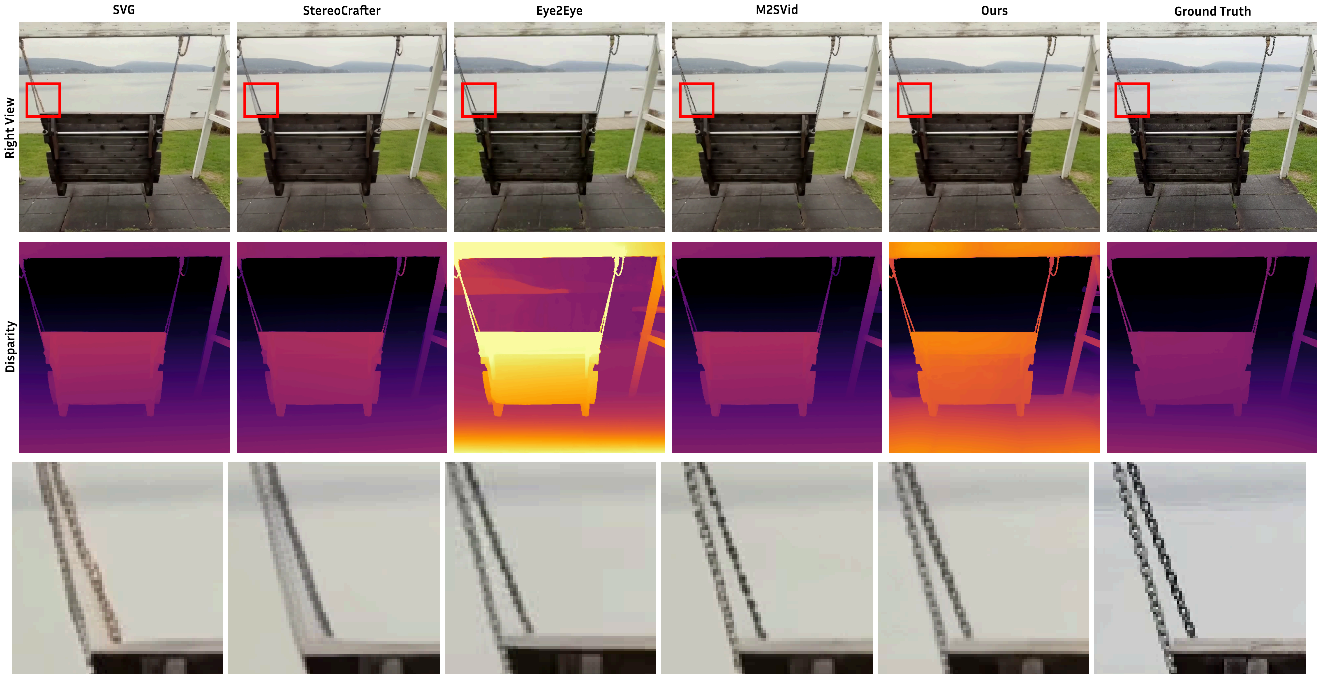

Figure 20. **Comparison on Scene 30 of the iPhone dataset.** Our method recovers high-frequency details on the chain.

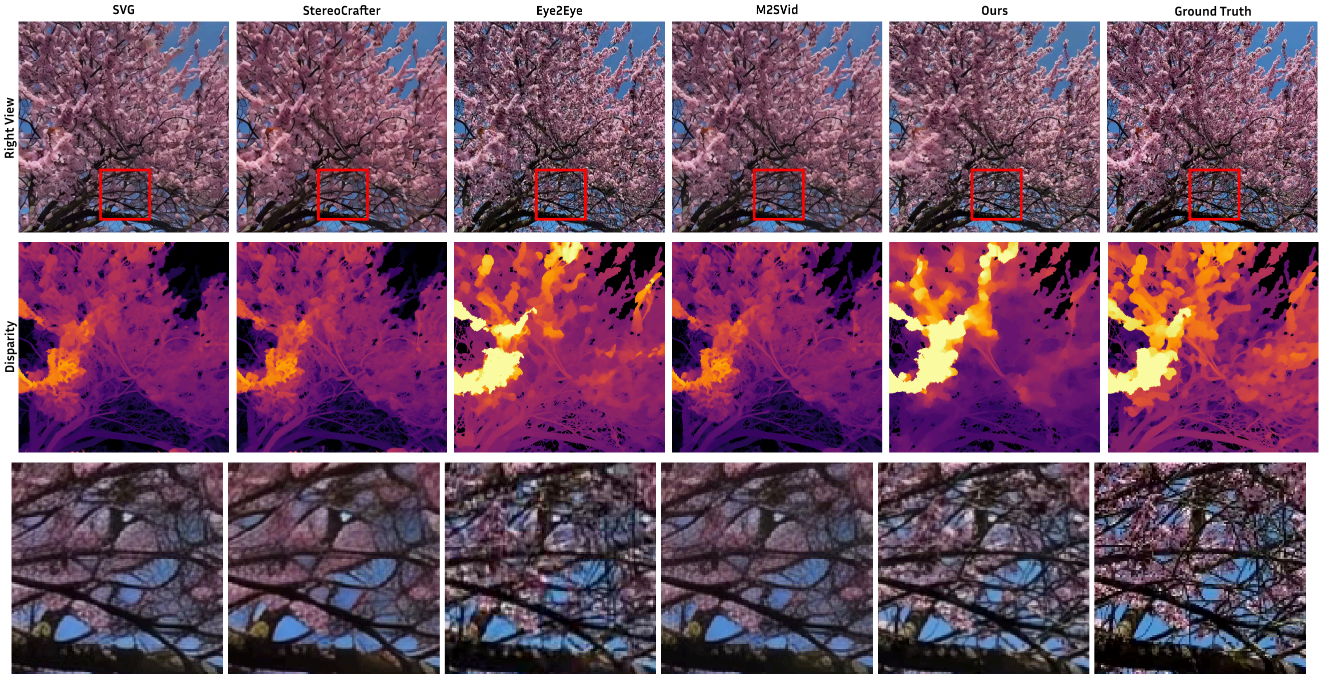

Figure 21. **Comparison on Scene 49 of the iPhone dataset.** Our method recovers the texture on the leaves.

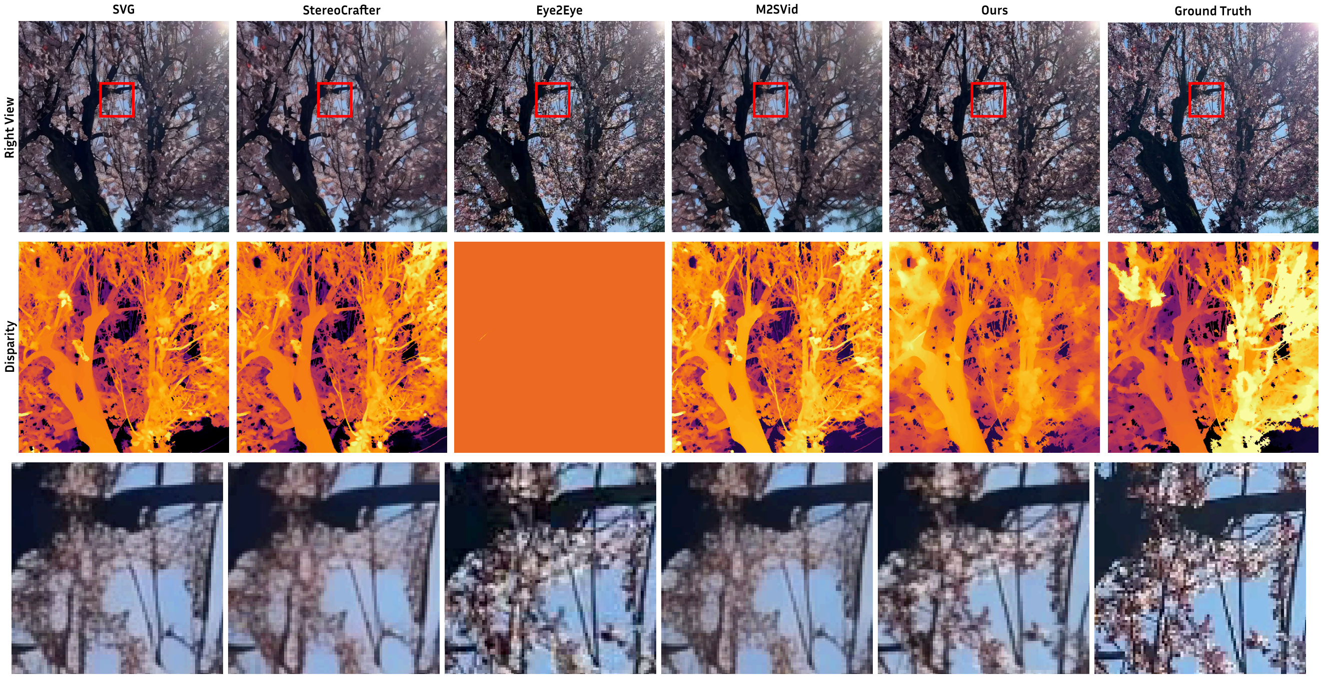

Figure 22. **Comparison on Scene 52 of the iPhone dataset.** Our method recovers the texture on the leaves.

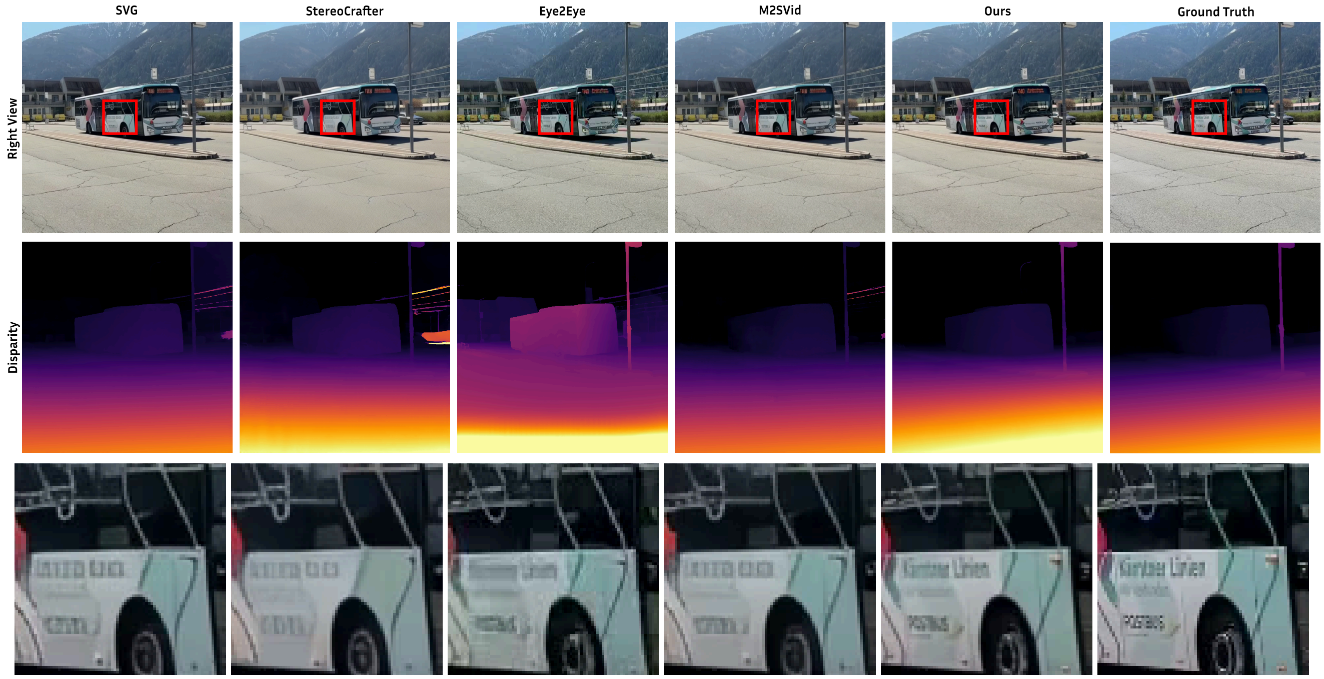

Figure 23. **Comparison on Scene 55 of the iPhone dataset.** Our method preserves better the high-frequency details on text.

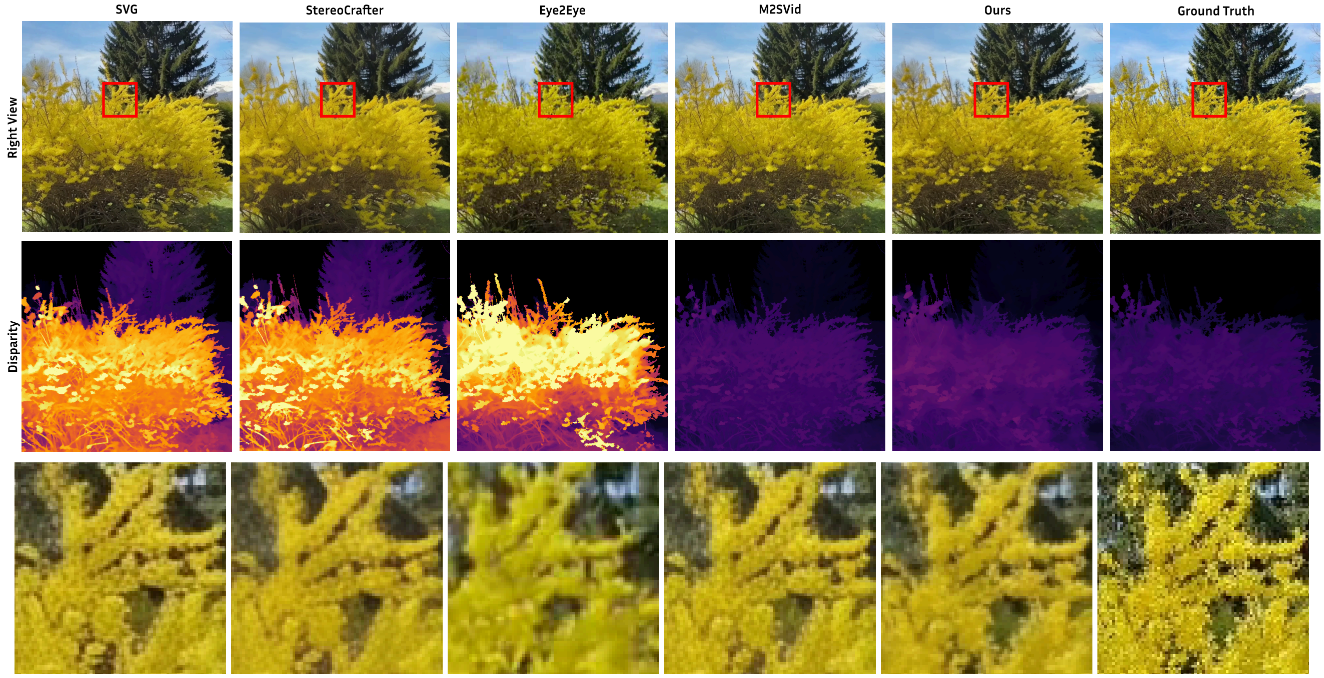

Figure 24. **Comparison on Scene 56 of the iPhone dataset.** Our method produces sharper details and texture.

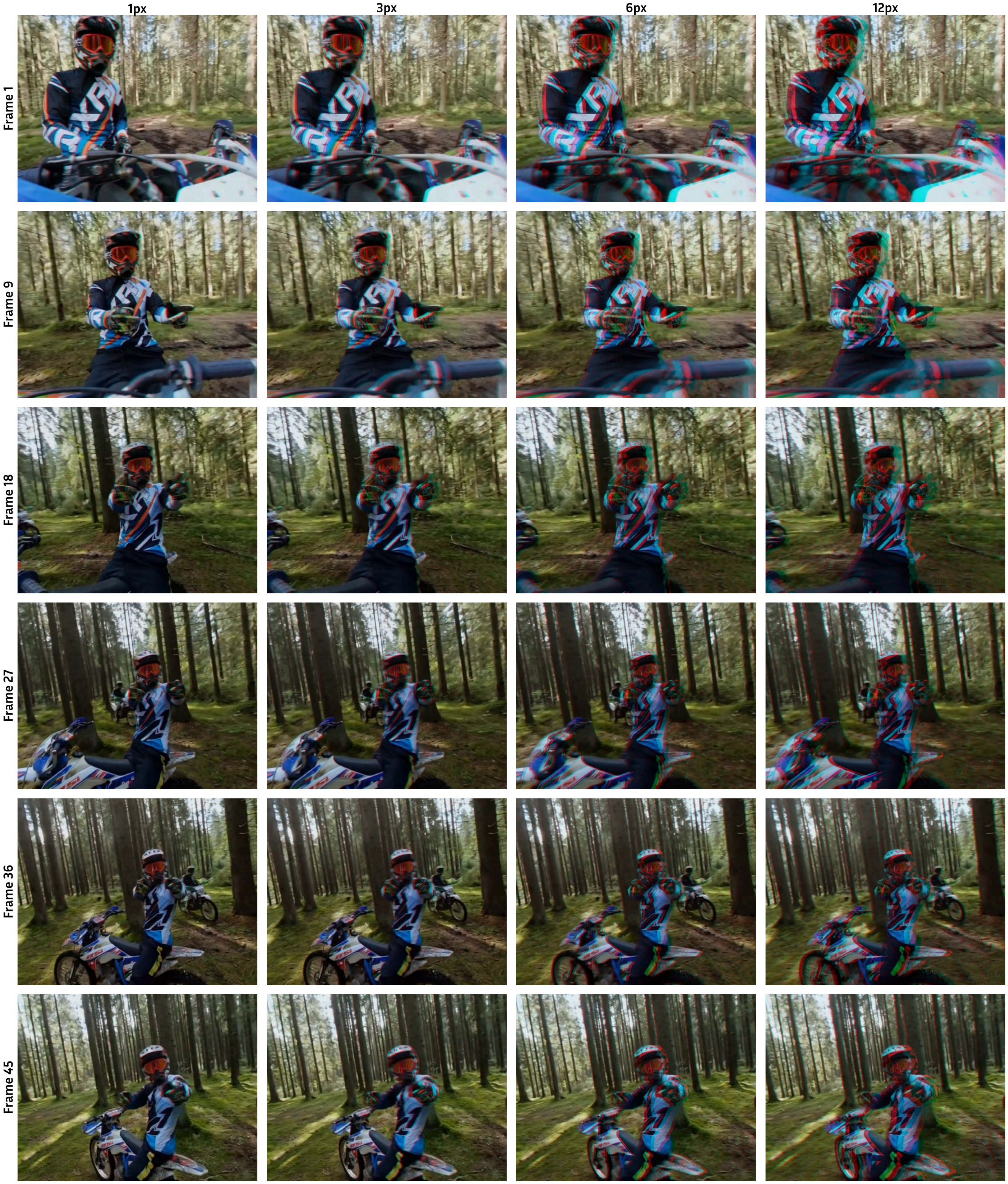

Figure 25. Anaglyphs generated by our approach on an in-the-wild image with varying disparity conditioning (in pixels). The results demonstrate the model's ability to control stereoscopic depth on real-world data. The frame number is indicated on the left while the disparity conditioning in pixel is written on the top row.

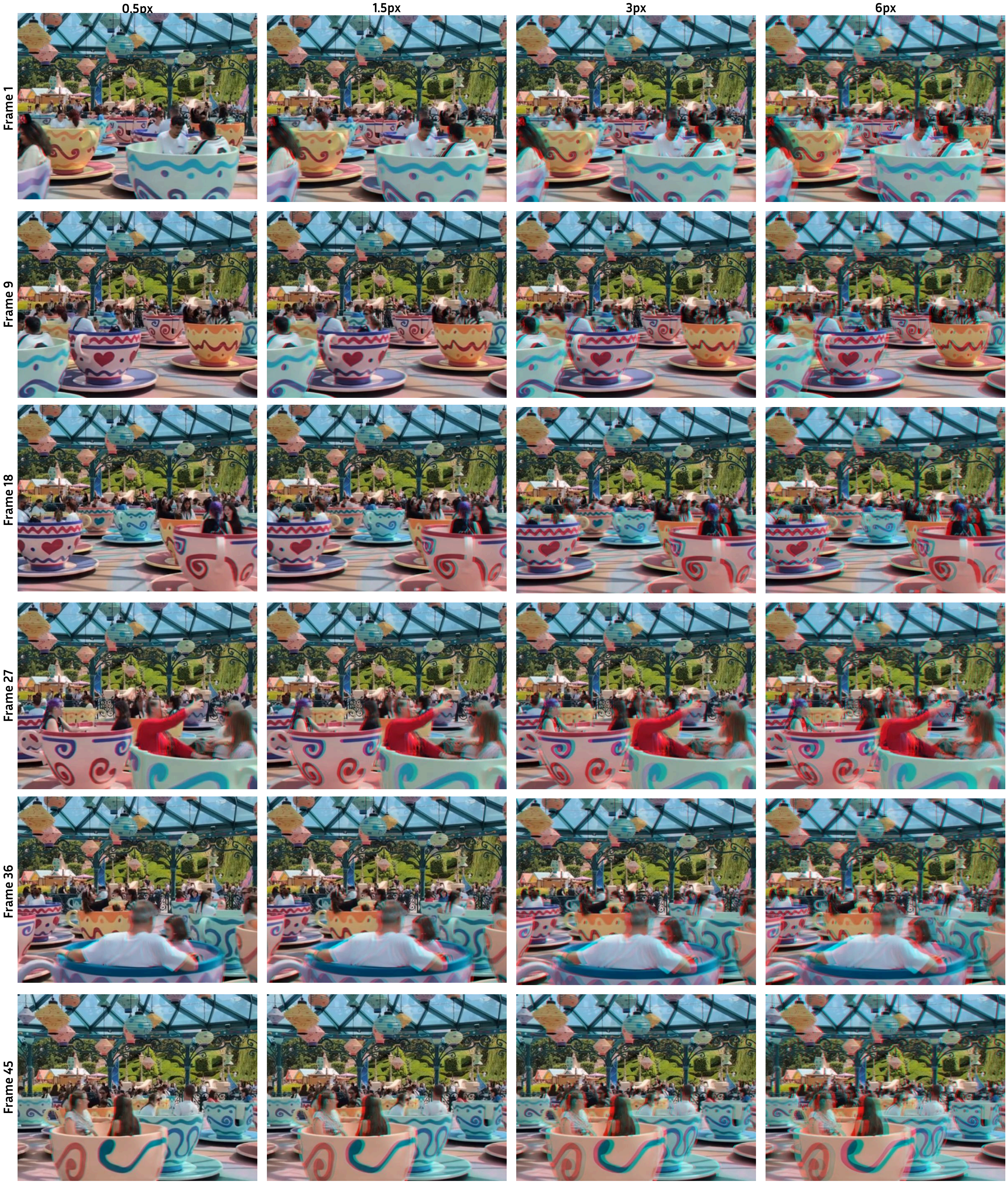

Figure 26. Anaglyphs generated by our approach on an in-the-wild image with varying disparity conditioning (in pixels). The results demonstrate the model's ability to control stereoscopic depth on real-world data. The frame number is indicated on the left while the disparity conditioning in pixel is written on the top row.

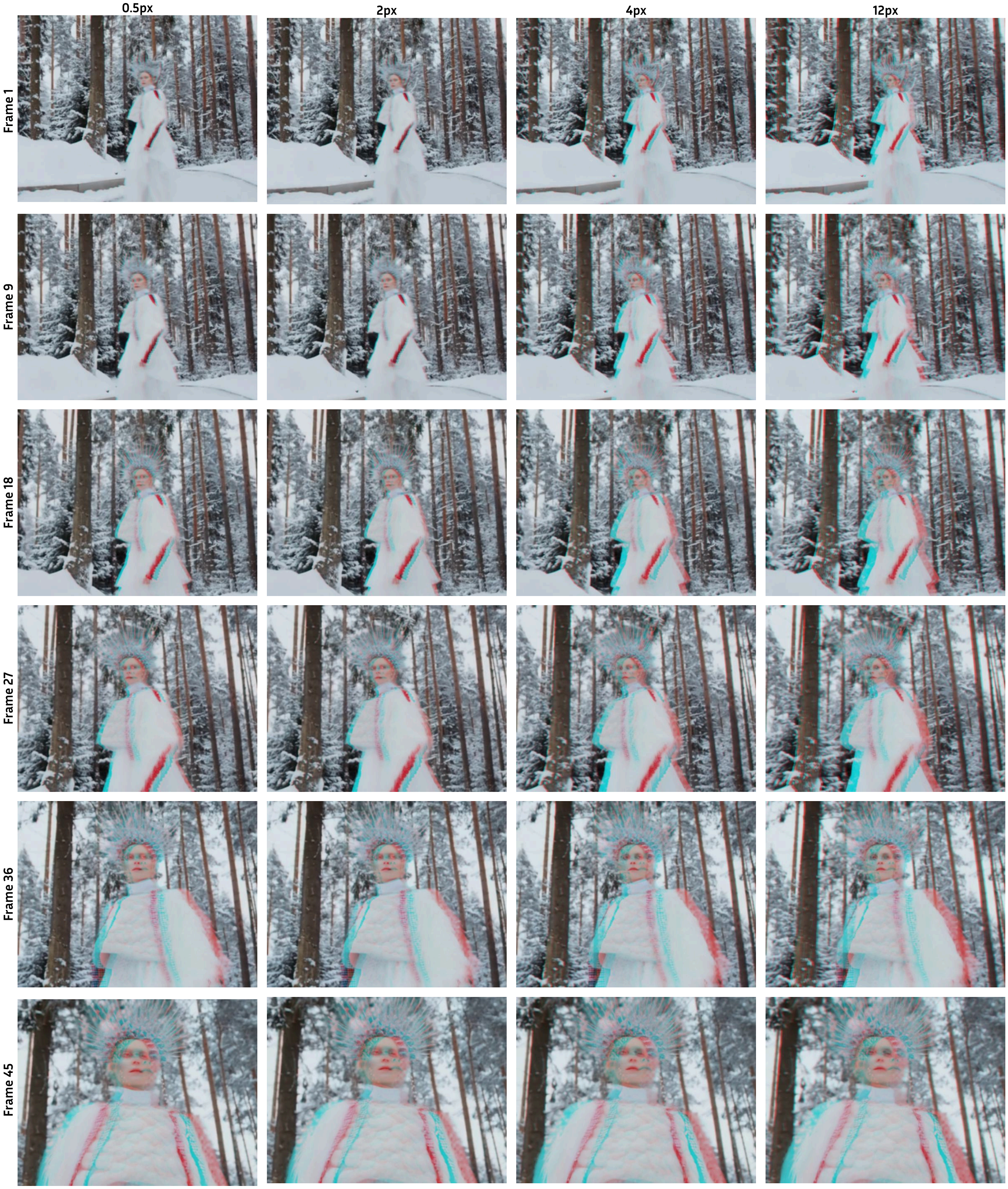

Figure 27. Anaglyphs generated by our approach on an in-the-wild image with varying disparity conditioning (in pixels). The results demonstrate the model's ability to control stereoscopic depth on real-world data. The frame number is indicated on the left while the disparity conditioning in pixel is written on the top row.