

# Eye2Eye: A Simple Approach for Monocular-to-Stereo Video Synthesis

Michal Geyer<sup>1,2</sup> Omer Tov<sup>1</sup> Linyi Jin<sup>1,3</sup> Richard Tucker<sup>1</sup>

Inbar Mosseri<sup>1</sup> Tali Dekel<sup>1,2</sup> Noah Snavely<sup>1</sup>

<sup>1</sup>Google DeepMind

<sup>2</sup>Weizmann Institute of Science

<sup>3</sup>University of Michigan



Figure 1. **3D anaglyph visualization of stereo videos produced by our method.** Our framework, *Eye2Eye*, takes as input a monocular video representing a right-eye view (top), and produces a left-eye video (visualized in the anaglyph on the bottom), enabling stereoscopic viewing using 3D glasses or a VR headset. Our method directly produces the new viewpoint, avoiding steps like explicit depth estimation and warping, and thus can plausibly handle challenging scenes with specular or transparent surfaces, such as the wine glass in the left example or the shiny floor in the right example, where assumptions of a single well-defined depth per pixel do not hold.

## Abstract

The rising popularity of immersive visual experiences has increased interest in stereoscopic 3D video generation. Despite significant advances in video synthesis, creating 3D videos remains challenging due to the relative scarcity of 3D video data. We propose a simple approach for transforming a text-to-video generator into a video-to-stereo generator. Given an input video, our framework automatically produces the video frames from a shifted viewpoint, enabling a compelling 3D effect. Prior and concurrent approaches for this task typically operate in multiple phases, first estimating video disparity or depth, then warping the video accordingly to produce a second view, and finally inpainting the disoccluded regions. This approach inherently fails when the scene involves specular surfaces or transparent objects. In such cases, single-layer disparity estimation is insufficient, resulting in artifacts and incorrect pixel shifts during warping. Our work bypasses these restrictions by directly synthesizing the new viewpoint, avoiding any intermediate steps. This is achieved by leveraging a pre-trained video model's priors on geometry, object ma-

terials, optics, and semantics, without relying on external geometry models or manually disentangling geometry from the synthesis process. We demonstrate the advantages of our approach in complex, real-world scenarios featuring diverse object materials and compositions. See videos on <https://video-eye2eye.github.io/>.

## 1. Introduction

Immersive viewing hardware—such as VR headsets and 3D displays—is rapidly improving, offering users increasingly high-quality 3D experiences. However, capturing high-quality 3D content remains challenging and often requires specialized equipment, limiting the availability of immersive media. This generates a growing demand for methods that can generate high-quality 3D content, such as stereoscopic video. Ultimately, we can envision a future where any video content can be experienced in 3D regardless of its original capture. Towards this goal, we address the problem of up-converting monocular 2D video to more immersive stereoscopic 3D video, leveraging recent advances in

generative video models.

The prevalent approach to mono-to-stereo video conversion adopts a two-step process: they first estimate geometry for an input video via monocular depth models, then use this geometry to re-project pixels to a second view, inpainting dis-occluded regions to generate a video for the second eye. However, such *warp-and-inpaint* approaches have an inherent restriction—they are inapplicable to scenes with reflections and complex light transport. Fundamentally, using a disparity map to warp an image to a new viewpoint assumes that there is a single, distinct depth at every input pixel. For scenes that exhibit simple Lambertian reflectance, this assumption largely holds true. However, for scenes with more complex light transport—specular reflection, partial transparency, etc.—we often cannot characterize each pixel with a single depth. For instance, in Fig. 6, a person is viewed through a glass window, thus the window’s pixels are a mixture of the person and the objects reflected on the glass—each at completely different (virtual) depths. To correctly handle such cases, methods based on explicit pixel warping would need to decompose the scene into multiple layers—e.g., reflected and transmitted light—warp each separately, then composite the results [35]. Without such handling, these methods can produce physically implausible views, for instance, where reflections appear pasted on a reflective surface, rather than at their correct virtual depth. The effects of such artifacts have been widely studied in cognitive science, where they have been shown to affect the way shape, material, and geometry are perceived [4, 30, 32, 52, 53].

We propose to address these limitations by *directly* producing the output RGB view, sidestepping the need for explicit disparity estimation or pixel warping. We leverage recent video diffusion models for this goal, as well as the observation that while full multi-camera 3D video datasets are scarce, stereo videos captured from two-view setups are relatively abundant online. Such videos represent ideal training data for mono-to-stereo methods, and allow us to learn to directly produce the desired output in a way that optimizes for the actual ground truth second-eye view, no matter how complex the underlying light transport is. We call this method **Eye2Eye**.

Our direct approach yields superior performance over warp-and-inpaint baselines in challenging real-world scenes featuring specular or transparent surfaces and dynamic lighting conditions. We validate these findings through a user study as well as a stereo perception metric introduced in [39].

In summary, our contributions are: (1) demonstrating, for the first time, mono-to-stereo video generation of specular dynamic scenes; (2) showing how to effectively leverage a pre-trained generative video model for this task, using curated online stereo videos; and (3) providing quantitative and qualitative evaluations via a user study and a perceptual

stereo metric that highlight the advantages of our approach over existing warp-and-inpaint methods.

## 2. Related work

**Multi-view video synthesis.** Progress in Generative-AI has been expanded to 3D generation, with trained image and video models being repurposed for static and dynamic multi-view generation. CAT3D [9] inflates an image diffusion model to take an arbitrary number of frames of a static scene as input, and to generate as output a 360° set of views, from which a 3D reconstruction can be estimated using off-the-shelf methods [15, 29]. A follow up work, CAT4D [54] expanded the method to dynamic scenes, but as the base CAT3D model is an image model, it still lacks motion prior and fails to handle complex motion. Other work tackles scaling video-diffusion architectures to the dynamic 4D setting [19, 49, 54, 58]. As fully multi-view video datasets are scarce, these methods often build largely on synthetic data, static scenes and monocular videos, which limits their performance on real-world videos. Our two-view stereo setting allows us to use a dataset of real-world online videos from [13], enabling stereo generation of arbitrarily complex videos in terms of scene-dynamics, camera motion, and light conditions.

**Mono-to-Stereo Conversion.** Early mono-to-stereo conversion methods primarily relied on motion parallax [59], perceptual heuristics [14, 21, 46, 57], or user interaction [24]. Deep3D [55] uses a CNN to predict each right video frame from the left by first estimating a soft disparity map and then compositing the output frame. These early approaches share a common limitation: the absence of a generative prior.

More recent mono-to-stereo synthesis methods employ a multi-stage pipeline, involving: (1) estimating video disparity (and temporally smoothing it), (2) using it to warp frames to the output view, and (3) inpainting disoccluded regions. Early works using this approach include [18, 50]. Recent methods following this pipeline build on top of generative diffusion models: StereoDiffusion [45] for images, and SVG [7], StereoCrafter [60], and SpatialDreamer [25] for videos. SVG leverages a pretrained text-to-video model without any further training, by devising a specific inpainting scheme, while StereoCrafter and SpatialDreamer fine-tune an image-to-video model, modifying it (1) to be video-rather than image-conditioned, and (2) to inpaint left-right dis-occlusion regions.

Our approach offers a key advantage over those pipelines, as in many real-world scenarios a single-layered disparity estimate is insufficient to represent scene geometry. While some progress on multi-layer flow prediction has been recently made [51], correctly estimating layered video

disparity remains an overlooked challenge. Instead, we directly leverage a pre-trained video model’s implicit, joint priors on geometry, object materials, and light, helping to alleviate this issue.

**Novel view synthesis with reflections and specularities.** Another possible approach for stereo synthesis is to apply a 3D video reconstruction pipeline and render stereo views from it. A line of work focuses on such 3D video reconstruction pipelines [20, 41, 47, 48]. Other work has focused on improving the ability of 3D reconstruction methods to render and reconstruct scenes with specular reflections, including: (1) re-parameterizing outgoing radiance as a function of the reflected view direction [23, 26, 42, 44], (2) combining 3D reconstruction with inverse graphics (simultaneously estimating material properties) [3, 12, 27, 36], (3) directly tracing reflection rays [43]. In the context of 3D video reconstruction, a recent work incorporates physically-based rendering into a Gaussian-splatting 3D video reconstruction pipeline to handle specular reflections [8]. In contrast, our approach leverages the implicit modeling capabilities of a large pretrained video model, eliminating the need for explicit physics-based representations. Furthermore, existing 4D reconstruction pipelines rely heavily on the input video to constrain the learned geometry and appearance, and often fail when the input lacks sufficient information (for example, when the camera motion is minimal, as demonstrated in Fig. 6). These limitations make 3D video reconstruction pipelines less robust for stereo generation.



Figure 2. **Limitations of warp-and-inpaint approach for mono-to-stereo video synthesis.** Given an input video frame (left), we use a state-of-the-art disparity estimation model [11] to compute its disparity (middle), and use it to warp the original frame to a new view (right). Since the predicted disparity map captures only the surface of the table, without considering the reflection of other objects off of it, the warped frame depicts incorrect reflection (skewed diagonally instead of reflecting vertically). When viewed in VR, the reflection on the table appears “flat”, as if it is a part of the table. This demonstrates the fundamental limitation of the common warp-and-inpaint approach for stereo view synthesis.

### 3. Preliminaries

#### 3.1. Stereo geometry

The geometric relationship between corresponding points in a stereo pair is governed by epipolar geometry. For a rectified stereo setup with parallel camera projection planes, a 3D point  $(x, y, z)$  projects to image coordinates  $(u_L, v)$  in the left view and  $(u_R, v)$  in the right view, where the horizontal disparity  $d = u_L - u_R$  is inversely proportional to depth:  $d = \frac{fb}{z}$ , where  $f$  is the focal length and  $b$  is the baseline distance between cameras. In the case of specular surfaces, a single depth value  $z$  cannot be assigned to each pixel, since the depth of the surface itself  $z_{\text{surface}}$  and that of the reflected object  $z_{\text{reflected-object}}$  may differ. Thus, to correctly re-render the video from another view point, the surface and the reflected content should shift according to their depth as in the above equation:  $d_{\text{surface}} = \frac{fb}{z_{\text{surface}}}$ ,  $d_{\text{reflection}} = \frac{fb}{z_{\text{reflected-object}}}$ . Fig 2 demonstrates how warping pixels that have reflections only with  $d_{\text{surface}}$  distorts the rendered image.

#### 3.2. Diffusion models

A diffusion model learns to reverse a noising process. Given a clean image  $x_0$ , the forward noising process adds Gaussian noise according to a variance schedule  $\beta_t$ , producing noisy samples  $x_t = \sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon$ , where  $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$  and  $\epsilon \sim \mathcal{N}(0, I)$ . The simplified diffusion objective minimizes:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2] \quad (1)$$

When conditioned on additional inputs  $c$ , the model learns the conditional distribution  $\epsilon_\theta(x_t, t, c)$ . During inference, the model iteratively denoises random noise  $x_T$  back to a clean sample. We build our framework on top of Lumiere [2], which is a cascaded video diffusion model.

Cascaded diffusion models consist of two components: a base model that generates videos at low resolution, and a spatial-super-resolution (SSR) model that upsamples low-resolution outputs to a higher resolution. The SSR model is a conditional diffusion model that is trained to denoise high resolution videos conditioned on downsampled videos. At inference time, the SSR model iteratively denoises Gaussian noise into a high resolution video, conditioned on the low-resolution video produced by the base model.

### 4. Method

Given a monocular input video  $V^{\text{right}}$ , our goal is to synthesize its corresponding stereo pair by generating a left view  $V^{\text{left}}$ , as if captured by a camera horizontally shifted from the original camera position by approximately human interpupillary distance (roughly 6.5cm), as per the rectified stereo geometry described in Section 3.

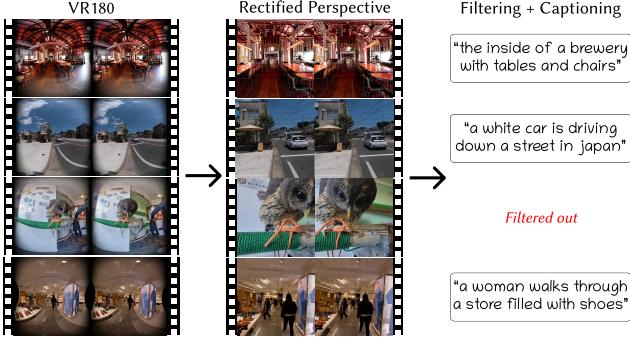


Figure 3. **Data processing pipeline.** We curate stereo VR180 footage captured with high-resolution cameras and stored in a equirectangular format. Following Stereo4D [13], we rectify the stereo videos and map the equirectangular format to perspective videos. We filter out videos with large disparity using RAFT [40] and caption the remaining videos with BLIP2 [22].

Our task presents two key challenges: (1) understanding the video’s geometry and light transport sufficiently well to determine how to transform the pixel content into the new view, and (2) synthesizing realistic content for regions that are occluded in the original view, but become visible in the new viewpoint. Given that generative video diffusion models have been shown to capture priors on both scene geometry and occluded content [6, 10, 17, 34], we propose to leverage such models to jointly address both challenges, as well as a stereo video dataset, containing left and right eye viewpoints of dynamic, in-the-wild videos. Specifically, we extend Lumiere [2], a cascaded text-to-video diffusion model, and construct training data from [13], to address this task.

While we maintain Lumiere’s two-stage process of low-resolution generation followed by super-resolution, we make two principal changes to adapt it to our task. First, our model takes a video as input, in addition to text (rather than text alone). Second, we find that Lumiere’s super-resolution design is not suitable for stereo synthesis, leading us to develop a different approach. We detail these modifications in the following sections, as well as our stereo dataset collection and processing. We call our overall method **Eye2Eye**.

#### 4.1. Low-resolution stereo generation

Our first step focuses on fine-tuning the base Lumiere model  $\phi(x_t, t, c)$  (where  $t$  is the diffusion timestep and  $c$  is the text conditioning) to produce left-from-right views. We do so by modifying its architecture to accept additional conditioning channels in its first input convolution layer. The model is trained to denoise the left view while being conditioned on the clean right view, following the standard conditional diffusion training formulation (as described in Section 3). This results in a model that produces novel left views at 128-pixel resolution (Fig. 4 top left). We call this model

the **base Eye2Eye generator** and denote it by  $\tilde{\phi}_{\text{base}}$ . After training, given an input down-sampled video,  $V_{\downarrow}^{\text{right}}$ , and a caption  $c$ ,  $\tilde{\phi}_{\text{base}}$  produces a low resolution left-view video:

$$\tilde{\phi}_{\text{base}}(x_T, T, V_{\downarrow}^{\text{right}}, c) = V_{\text{base}}^{\text{left}} \quad (2)$$

#### 4.2. High-resolution stereo refinement

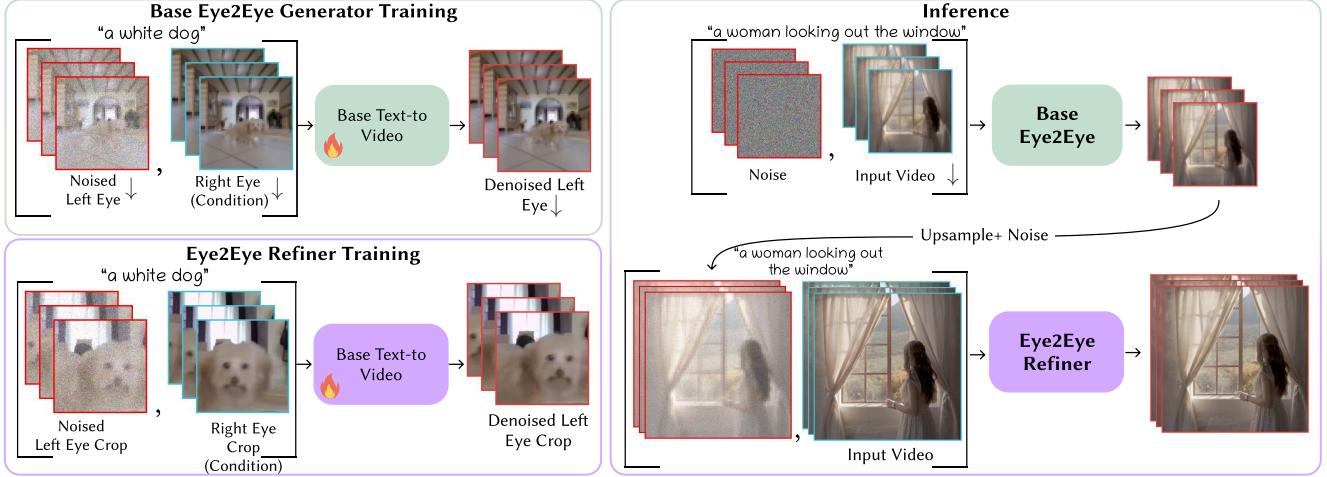
While the base stereo generator successfully creates left-from-right views, achieving high-resolution stereo synthesis presents additional challenges. Directly applying the pre-trained Lumiere super-resolution (SSR) model to  $V_{\text{base}}^{\text{left}}$  would produce details inconsistent with the original input video  $V^{\text{right}}$ , as the SSR denoises Gaussian noise based solely on  $V_{\text{base}}^{\text{left}}$ . We experimented with modifying the SSR model to be conditioned on the right view, but this resulted in degraded quality, which we attribute to its fully convolutional, simpler architecture compared to the base model. Therefore, we take a different route and adapt the base model for high-resolution left-from-right video synthesis.

Consider a pixel in a video with a resolution of  $128 \times 128$  that has a disparity of  $d$  pixels between the original and generated view. When generating a video at  $512 \times 512$  resolution ( $4 \times$  higher), the disparity should scale proportionally ( $4d$ ) pixels to maintain the same real-world depth effect. However, we observe that when sampling from  $\tilde{\phi}_{\text{base}}$  at different input resolutions, the pixel disparity remains at  $d$  pixels rather than scaling with the resolution. This leads to an undesirable effect: sampling at higher resolutions effectively reduces the perceived 3D depth in the stereo pairs, as shown in Fig. 5, columns 2 and 3. This behavior is analogous to changing the scale of the disparity itself.

To address this issue, we instead fine-tune  $\phi$  on *high-resolution crops* of size  $128 \times 128$  to learn correct-scale disparity and inpainting (Fig. 4 top right). We observe that although training on high-resolution crops indeed allows high-resolution sampling with larger pixel shifts, this approach introduces its own challenge: small crops often contain limited disparity variation and distant content. This causes the model to develop a bias toward uniformly shifting the input view (Fig. 5, column 1).

While simply bilinearly up-sampling  $V_{\text{base}}^{\text{left}}$  yields the correct disparity scale and stereo geometry, the model trained on high resolution crops yields better quality in inpainted areas or in areas where the disparity is large. To combine the strengths of both models, we use them in a two-stage inference pipeline, which exploits a fundamental property of diffusion models—early denoising steps establish global layout and structure, while later steps refine details [28]. Specifically, we use  $\tilde{\phi}_{\text{base}}$  to produce a low-resolution layout with correctly scaled disparity, and use the model trained on crops as an *Eye2Eye refiner model*, denoted by  $\tilde{\phi}_{\text{refiner}}$ . That is, our method:

1. generates an initial low-resolution left view video using the base Eye2Eye generator,  $V_{\text{base}}^{\text{left}}$  as in eq 2,



**Figure 4. Eye2Eye mono-to-stereo pipeline.** We leverage the pre-trained Lumiere cascaded text-to-video model, as well as a curated dataset of rectified stereo pairs, to perform mono-to-stereo synthesis. We finetune two different copies of a base (low-resolution) pre-trained Lumiere model, in two different contexts. For the first base model, we add additional input channels to condition the model on an input right eye, and train the base Eye2Eye generator on downsampled, low-resolution  $128 \times 128$  stereo pairs (top left). We call the resulting trained model the **base Eye2Eye generator** model. We train the second model to be a refinement model with the same conditioning mechanism, but instead trained on  $128 \times 128$  *crops* from full, high-resolution images (bottom left). We call the resulting model the **Eye2Eye refiner** model. The base Eye2Eye model models correct pixels disparity at a low resolution, and the Eye2Eye refiner has better quality in inpainted areas or areas with large disparities. At inference time, our sampling process (right) combines both models’ strengths by first generating a low-resolution output from the base Eye2Eye model to establish appropriate stereo disparity for a compelling 3D effect, then noising and denoising it with the Eye2Eye refiner to achieve high visual quality.

- upsamples this output to the target resolution and noises the upsampled output,

$$x_t^{\text{left}} = \sqrt{\alpha_t} \cdot V_{\text{base}}^{\text{left}} + \sqrt{1 - \alpha_t} \cdot \epsilon$$

where  $\alpha_t$  is the diffusion noise schedule parameter as described in Sec. 3.2,

- denoises the noised upsampled-resolution video using the *stereo refiner* model:

$$V^{\text{left}} = \tilde{\phi}_{\text{refiner}}(x_t^{\text{left}}, t, V^{\text{right}}, c)$$

In other words, we perform SDEdit [28] on  $V_{\text{base}}^{\text{left}}$   $\uparrow$  with  $\tilde{\phi}_{\text{refiner}}$ . This combined approach preserves correct scale disparity from the low-resolution generation while enabling high-resolution refinement of fine details and textures (Fig. 5, rightmost column). The result is a pipeline that consistently balances stereo disparity with high-resolution detail, effectively bridging the gap between training and inference resolution.

#### 4.3. Training dataset

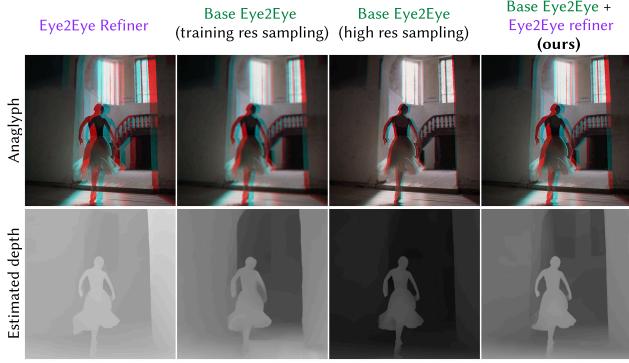
We construct our training data from the Stereo4D [13] dataset, which contains over 100k high-resolution, rectified stereo videos capturing diverse scenes and moving objects. As shown in Fig. 3, this dataset provides real-world video data that naturally includes challenging cases such as reflective surfaces, which are difficult to simulate in synthetic

datasets. Following Stereo4D, we project VR180 videos to rectified perspective videos of resolution  $512 \times 512$ . We filter out examples with excessively large disparities caused by objects being too close to the camera, as these often lead to stereo window violations [61] and are challenging for the model to learn. Specifically, we compute optical flow between the left and right frames with RAFT [38, 40] to estimate pixel disparities and discard videos where the disparity exceeds a specified threshold (60 pixels). Additionally, we use BLIP2 [22] to generate captions from the middle frame of each video. During training, we sample 80 frames per clip to align with Lumiere’s input video length.

## 5. Results

### 5.1. Baselines

The most prominent baseline for our approach is StereoCrafter [60], which adopts a warp-and-inpaint approach and trains an inpainting model specifically for handling left-right disocclusions. Since StereoCrafter builds upon Video-Stable-Diffusion, a different pretrained video diffusion model than the one we use, we re-implement StereoCrafter using Lumiere—the pretrained model utilized in our method, in order to ensure a fair comparison between approaches. We fine-tune the low-resolution Lumiere model specifically on warped views and subsequently employ the Lumiere super-resolution stage combined with a blended



**Figure 5. Resolving training and inference gap.** We ablate the use of the two models in our pipeline, illustrating the training-inference gap of each of them, and visualize the resulting anaglyph and depth estimation (estimated using [40]) of their outputs. When sampling from the Eye2Eye-refiner model (trained on crops without any downsampling), far away content is still shifted by a large amount (column 1). When sampling from the base Eye2Eye generator at a higher resolution than its training resolution, the scale of the disparity and novel content in the frame reduces, weakening the 3D effect compared to sampling at the training resolution (columns 2 and 3, in column 2 the outputs were upsampled). By upsampling the outputs of the base stereo model and noising and denoising it with the Eye2Eye-refiner model, we maintain both a good depth perception from the base model and the stereo refiner’s ability to generate high quality frames (column 4).

diffusion approach [1] to preserve details from the original warped videos. We refer to this baseline as *warp and inpaint*. See the appendix for more details.

We additionally include qualitative comparisons with Deep3D [55], a deep CNN trained for mono-to-stereo video prediction; and Dynamic Gaussian Marbles (DGM) [37], a method for novel view synthesis of monocular videos.

## 5.2. Evaluation data

We assess our method on a held-out test set of 30 publicly sourced videos, encompassing diverse scenes, camera motions, and dynamic content. These videos are chosen to feature complex lighting conditions and varied materials, including specular surfaces that introduce challenges such as reflections. Some of the videos are taken from the data provided in [56], which proposed a method to decompose the different layers of reflected and refracted light. See a sample of the evaluation videos in Figure 7.

## 5.3. Qualitative comparisons

Figure 6 shows qualitative comparisons to the baselines. Both the warp-and-inpaint baseline and StereoCrafter incorrectly shift scene content in areas with reflections or transparencies. These methods struggle to handle layered structures, failing to accurately separate and shift objects with reflections or transparencies.

In the top example, depth-warping methods (column b, c) shift the upper portion of the building more than the lower part. This occurs because the top is occluded by a transparent umbrella, and the single-layer disparity model assigns it a larger disparity. This causes distortion and incorrect 3D effect: the top part of the building appears as close to the viewer as the umbrella when viewed with red-cyan anaglyph glasses or in VR. Similarly, in the bottom example, the reflected distant pole is shifted too much, along with the woman’s head.

In contrast, our results (column a) preserve the correct depth layering by shifting each visible layers according to its own disparity. The transparent umbrella is shifted more than the building behind it, and the reflection of the pole is shifted only slightly, while the woman’s head is shifted more substantially, consistent with their relative depths.

DGM (column c) cannot inpaint missing content at occlusion boundaries since it has no generative prior, leading to holes in the video frames such as white borders near the people (top example) and along the left edge of the frame (bottom example). Additionally, as it uses single-layer depth estimation for geometry regularization, it also suffers from the distortions and fails to correctly model the geometry. Deep3D (column d) fails to produce a sufficient 3D effect—the output videos are almost identical to the input ones in most cases.

## 5.4. Quantitative comparisons

**User study.** To evaluate the benefit of our direct synthesis approach over the warp-and-inpaint approach, we conducted a user study using a Two-alternative Forced Choice (2AFC) protocol [16, 33]. Participants viewed two videos side-by-side with VR headsets: our model’s output and the *warp and inpaint* baseline’s output. Specifically, our model’s and the *warp and inpaint* left view predictions were presented to the participants’ left eyes, while the input right video was presented to their right eyes.

Prior to the main comparison, participants were shown a ground truth stereo pair featuring a large reflection alongside a warp-and-inpaint result that does not account for reflection. This preliminary step ensured that participants understood the task and excluded those with binocular vision dysfunctions (see the test examples in the supplemental material). During the main task, participants were asked to determine which video exhibited a more realistic 3D effect, including in areas with reflections or transparent surfaces. Overall, participants favored our videos 66% of the time based on 239 judgments. To further statistically assess the study’s results, we classified a video as favoring our method if more than half of its votes were positive (each video received between 5 and 15 votes). Out of 31 videos, 23 (about 74 %) met this criterion. Under the assumption of a 50% chance for a positive majority, a binomial test produced a



**Figure 6. Qualitative Comparison.** Our method successfully generates left from right views in complex scenarios where light is both reflected on a transparent material and refracted through it. The warp-and-inpaint baseline, relying on a single-layer disparity prediction, fails in such cases. For instance, in the top example, the top of the building appears as near as the transparent umbrella overlaying it (see anaglyph), and the building is distorted (see output left eye). Our method, in contrast, successfully shifts the umbrella without shifting the building behind it. In the bottom example, the pole reflected on the glass appears as near as the woman behind the window (see anaglyph); in our result, the pole is almost not shifted, as it is far away. Dynamic Gaussian Marbles (DGM, c), a 4D reconstruction method, lacks generative capabilities. Thus, their output left eye has white regions of missing content (see top example along the borders of the people, and in the bottom example along the left edge of the frame). In addition, since DGM relies on metric depth estimation as a regularization, it often fails to correctly model the geometry in complex scenarios—producing distortions similar to those of the warp-and-inpaint baseline in the top example, and a “flat” output in the bottom example. Finally, Deep3D (d) generally fails to generate a sufficient 3D effect, as seen in the anaglyph visualizations.

one-sided p-value of 0.0053 (and a two-sided p-value of 0.0107), indicating that the result is statistically significant and unlikely to be due to chance.

**iSQoE stereo perception metric** We also evaluate our results using the recently proposed stereo perception metric, iSQoE [39], which trained a model to assess stereoscopic quality of experience (SQoE) of a stereo pair by aligning it closely with human perceptual preferences. The authors showed that iSQoE effectively evaluates different mono-to-stereo conversion techniques. iSQoE is an image (not a video) metric, and it is meaningful only when comparing the same stereo pair generated through different pro-

cessing or conversion methods. Thus, to obtain per-video preferences, we average the iSQoE scores across frames and compare the mean scores between methods. Our approach achieves higher average scores on 84% of the videos from our test set when compared to StereoCrafter, and 74% when compared to our implemented warp-and-inpaint baseline, supporting our approach’s superior performance. Interestingly, StereoCrafter performed worse than the warp-and-inpaint baseline; we attribute this to the stronger pre-trained model used in our implementation (i.e., Lumiere in our warp-and-inpaint baseline vs. Stable-Video-Diffusion [5] in StereoCrafter).

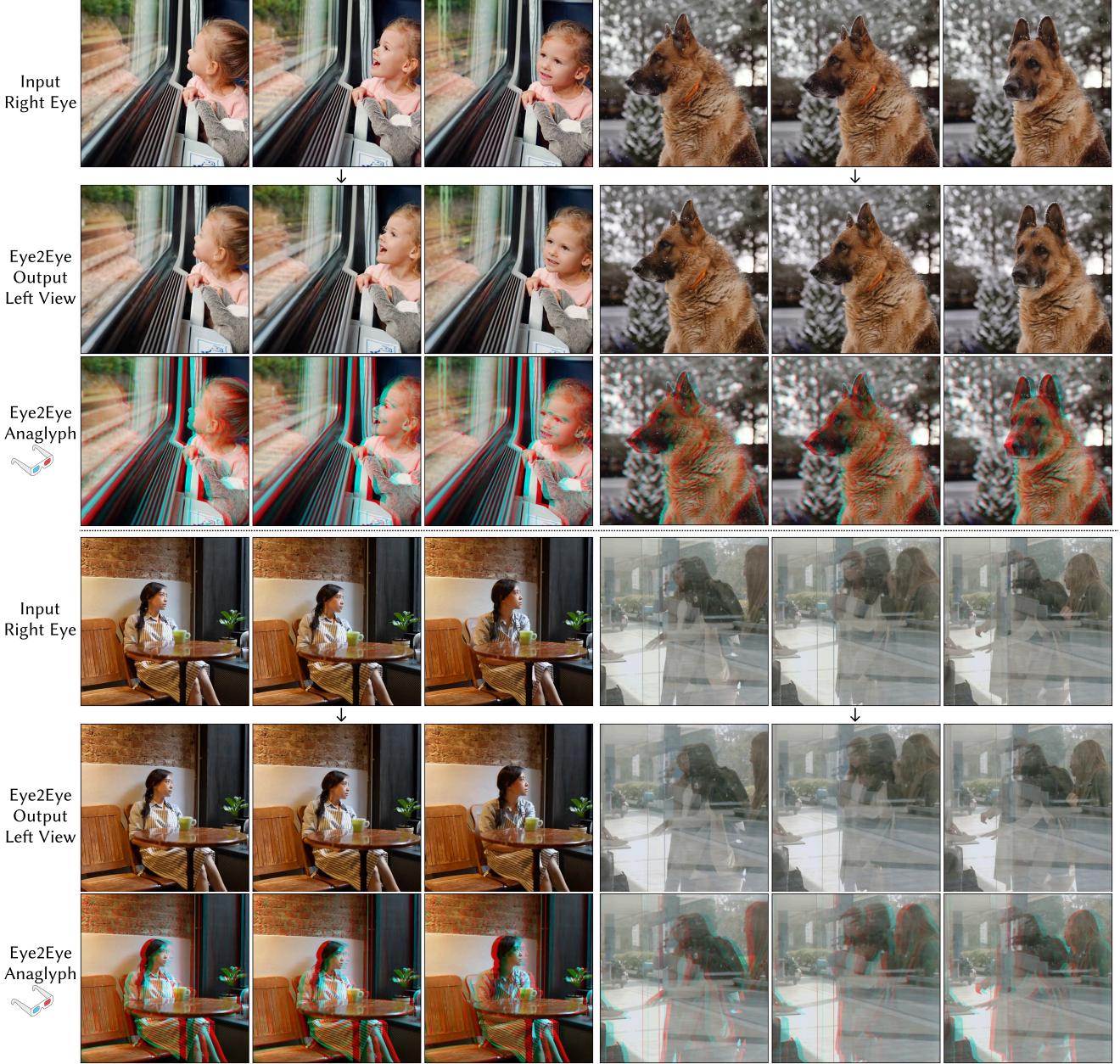


Figure 7. **Our generated stereo views.** Our approach is particularly successful in complex scenarios involving reflective objects such as glass doors or specular tables, where traditional methods often produce distortions. See videos in our website.

We note that we do not report reconstruction errors with respect to the ground truth left view, since our method does not directly control the disparity scale, making pixel-wise comparisons unreliable.

## 6. Discussion and Conclusions

We presented a simple approach for video mono-to-stereo conversion, highlighting complexities that were often overlooked in prior and concurrent work. As video models con-

tinue to grow in size and training data, they not only produce higher-quality outputs but also appear to implicitly approximate certain aspects of our physical world; our results highlight these emerging capabilities. Our user study suggests handling reflections in modern VR headsets would help increase the realism of immersive experiences, encouraging VR development and research to consider these nuances. Nonetheless, an inherent limitation of our current approach is that we do not control the baseline between the cam-

eras, constraining the extent of the 3D effect. Future work could explore methods for dynamically adjusting this baseline, thereby offering more flexibility in creating immersive stereo content.

## 7. Acknowledgments

We thank Shir Amir for her valuable assistance in running the iSQoE model for our evaluation.

## References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18187–18197. IEEE, 2022. [6](#)
- [2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Hermann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. [3, 4](#)
- [3] Sai Bi, Zexiang Xu, Pratul P. Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Milos Hasan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv*, 2020. [3](#)
- [4] Andrew Blake and Heinrich Bülthoff. Does the brain know the physics of specular reflection? *Nature*, 1990. [2](#)
- [5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling latent video diffusion models to large datasets, 2023. [7](#)
- [6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *arXiv*, 2024. [4](#)
- [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, 2024. [2](#)
- [8] Cheng-De Fan, Chen-Wei Chang, Yi-Ruei Liu, Jie-Ying Lee, Jiun-Long Huang, Yu-Chee Tseng, and Yu-Lun Liu. Spec-tromotion: Dynamic 3D reconstruction of specular scenes. *arXiv*, 2024. [3](#)
- [9] Ruiqi Gao\*, Aleksander Holynski\*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole\*. CAT3D: Create anything in 3D with multi-view diffusion models. *arXiv*, 2024. [2](#)
- [10] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models, 2023. [4](#)
- [11] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xi-aodong Cun, Yong Zhang, Long Quan, and Ying Shan. DepthCrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. [3, 11](#)
- [12] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. TensoIR: Tensorial Inverse Rendering. *CVPR*, 2023. [3](#)
- [13] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4D: Learning how things move in 3D from internet stereo videos. *arXiv preprint*, 2024. [2, 4, 5](#)
- [14] Petr Kellnhofer, Thomas Leimkühler, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. What makes 2D-to-3D stereo conversion perceptually plausible? In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pages 59–66, New York, NY, USA, 2015. ACM. [2](#)
- [15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. [2](#)
- [16] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. [6](#)
- [17] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation, 2024. [4](#)
- [18] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, Peizhao Zhang, Zijian He, Peter Vajda, Ayush Saraf, and Michael Cohen. One shot 3D photography. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 39(4), 2020. [2](#)
- [19] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *arXiv*, 2024. [2](#)
- [20] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. Fast view synthesis of casual videos with soup-of-planes, 2024. [3](#)
- [21] Thomas Leimkühler, Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. Perceptual real-time 2D-to-3D conversion using cue fusion. *IEEE Transactions on Visualization and Computer Graphics*, 24(6):2037–2050, 2018. [2](#)
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 2023. [4, 5](#)

- [23] Ruofan Liang, Huiting Chen, Chunlin Li, Fan Chen, Selvakumar Panneer, and Nandita Vijaykumar. ENVIDR: Implicit Differentiable Renderer with Neural Environment Lighting. *ICCV*, 2023. 3
- [24] Miao Liao, Jizhou Gao, Ruigang Yang, and Minglun Gong. Video stereolization: Combining motion analysis with user interaction. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1079–1088, 2012. 2
- [25] Zhen Lv, Yangqi Long, Congzhentao Huang, Cao Li, Chengfei Lv, Hao Ren, and Dian Zheng. SpatialDreamer: Self-supervised stereo video synthesis from monocular input, 2024. 2
- [26] Li Ma, Vasu Agrawal, Haithem Turki, Changil Kim, Chen Gao, Pedro Sander, Michael Zollhöfer, and Christian Richardt. SpecNeRF: Gaussian directional encoding for specular reflections. *arXiv 2312.13102*, 2023. 3
- [27] Alexander Mai, Dor Verbin, Falko Kuester, and Sara Fridovich-Keil. Neural microfacet fields for inverse rendering. *ICCV*, 2023. 3
- [28] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Image synthesis and editing with stochastic differential equations. *CoRR*, abs/2108.01073, 2021. 4, 5
- [29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2
- [30] Martin Mišiak, Arnulph Fuhrmann, and Marc Erich Latoschik. The impact of reflection approximations on visual quality in virtual reality. In *ACM Symposium on Applied Perception 2023*, New York, NY, USA, 2023. Association for Computing Machinery. 2
- [31] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 12
- [32] Masakazu Ohara, Juno Kim, and Kowa Koida. The role of specular reflections and illumination in the perception of thickness in solid transparent objects. *Frontiers in Psychology*, 13, 2022. 2
- [33] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems*, 2020. 6
- [34] Abhishek Sharma, Adams Yu, Ali Razavi, Andeep Toor, Andrew Pierson, Ankush Gupta, Austin Waters, Aäron van den Oord, Daniel Tanis, Dumitru Erhan, Eric Lau, Eleni Shaw, Gabe Barth-Maron, Greg Shaw, Han Zhang, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jakob Bauer, Jeff Donahue, Junyoung Chung, Kory Mathewson, Kurtis David, Lasse Espeholt, Marc van Zee, Matt McGill, Medhini Narasimhan, Miaosen Wang, Mikołaj Bińkowski, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Nando de Freitas, Nick Pezzotti, Pieter-Jan Kindermans, Poorva Rane, Rachel Hornung, Robert Riachi, Ruben Villegas, Rui Qian, Sander Dieleman, Serena Zhang, Serkan Cabi, Shixin Luo, Shlomi Fruchter, Signe Nørly, Srivatsan Srinivasan, Tobias Pfaff, Tom Hume, Vikas Verma, Weizhe Hua, William Zhu, Xinchen Yan, Xinyu Wang, Yelin Kim, Yuqing Du, and Yutian Chen. Veo. *arXiv*, 2024. 4
- [35] Sudipta N. Sinha, Johannes Kopf, Michael Goesele, Daniel Scharstein, and Richard Szeliski. Image-based rendering for scenes with reflections. *ACM Transactions on Graphics (TOG)*, 31:1 – 10, 2012. 2
- [36] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. *CVPR*, 2021. 3
- [37] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, page 1–11. ACM, 2024. 6
- [38] Deqing Sun, Charles Herrmann, Fitzsum Reda, Michael Rubinstein, David J Fleet, and William T Freeman. Disentangling architecture and training for optical flow. In *ECCV*, 2022. 5
- [39] Netanel Y. Tamir, Shir Amir, Ranel Itzhaky, Noam Atia, Shobhit Sundaram, Stephanie Fu, Ron Sokolovsky, Phillip Isola, Tali Dekel, Richard Zhang, and Miriam Farber. What makes for a good stereoscopic image?, 2024. 2, 7
- [40] Zachary Teed and Jia Deng. RaFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 4, 5, 6, 11
- [41] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 3
- [42] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 3
- [43] Dor Verbin, Pratul P. Srinivasan, Peter Hedman, Ben Mildenhall, Benjamin Attal, Richard Szeliski, and Jonathan T. Barron. NeRF-Casting: Improved view-dependent appearance with consistent reflections, 2024. 3
- [44] Fangjinhua Wang, Marie-Julie Rakotosaona, Michael Niemeyer, Richard Szeliski, Marc Pollefeys, and Federico Tombari. UniSDF: Unifying Neural Representations for High-Fidelity 3D Reconstruction of Complex Scenes with Reflections. *arXiv:2312.13285*, 2023. 3
- [45] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. StereoDiffusion: Training-free stereo image generation using latent diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7416–7425, 2024. 2
- [46] Oliver Wang, Manuel Lang, M. Frei, Alexander Sorkine-Hornung, Aljosa Smolic, and Markus Gross. StereoBrush: Interactive 2D to 3D conversion using discontinuous warps. pages 47–54, 2011. 2
- [47] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4D reconstruction from a single video. 2024. 3

- [48] Shizun Wang, Xingyi Yang, QiuHong Shen, Zhenxiang Jiang, and XinChao Wang. GFlow: Recovering 4D world from monocular video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 3
- [49] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J. Fleet. Controlling space and time with diffusion models, 2024. 2
- [50] 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. 2
- [51] Hongyu Wen, Erich Liang, and Jia Deng. Layeredflow: A real-world benchmark for non-lambertian multi-layer optical flow. *arXiv preprint arXiv:2409.05688*, 2024. 2
- [52] Gunnar Wendt, Franz Faul, and Rainer Mausfeld. Highlight disparity contributes to the authenticity and strength of perceived glossiness. *Journal of Vision*, 8(1):14–14, 2008. 2
- [53] Gunnar Wendt, Franz Faul, Vebjørn Ekroll, and Rainer Mausfeld. Disparity, motion, and color information improve gloss constancy performance. *Journal of Vision*, 10(9):7–7, 2010. 2
- [54] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create anything in 4D with multi-view video diffusion models. 2024. 2
- [55] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, 2016. 2, 6
- [56] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 34(4), 2015. 6
- [57] Qiong Zeng, Wenzheng Chen, Huan Wang, Changhe Tu, Daniel Cohen-Or, Dani Lischinski, and Baoquan Chen. Hallucinating stereoscopy from a single image. *Computer Graphics Forum*, 34, 2015. 2
- [58] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Kannan, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024. 2
- [59] Guofeng Zhang, Wei Hua, Xueying Qin, Tien-Tsin Wong, and Hujun Bao. Stereoscopic video synthesis from a monocular video. *IEEE Transactions on Visualization and Computer Graphics*, 13(4):686–696, 2007. 2
- [60] 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, 2024. 2, 5
- [61] Frederik Zilly, Marcus Müller, Peter Eisert, and Peter Kauff. The stereoscopic analyzer—an image-based assistance tool for stereo shooting and 3D production. In *2010 IEEE International Conference on Image Processing*, pages 4029–4032. IEEE, 2010. 5

## 8. Additional Details

### 8.1. Training details

We fine-tune Lumiere on a dataset of 100K clips from Stereo4D as mentioned in Section 4.3 of the main paper. We temporally subsample the videos into 80 frames at 16 fps to match Lumiere’s pre-training temporal resolution. We train the model for 120K steps with batch size 32 and learning rate  $2 \cdot 10^{-5}$ . The original clips resolution is  $512 \times 512$  pixels. To train the Eye2Eye base model, we additionally downsample the frames spatially to  $128 \times 128$  pixels. For the Eye2Eye refiner, we randomly sample crops of 128 pixels.

### 8.2. Sampling hyper-parameters for our method

#### 8.2.1. Base Eye2Eye sampling

We sample with 50 diffusion timesteps and without classifier-free guidance. We sample from this model at a resolution of 256 pixels, as we found that this resolution best mitigates visual quality and 3D effect.

#### 8.2.2. Eye2Eye refiner

We upsample the output of the base Eye2Eye model to  $512 \times 512$  pixels resolution and noise it to diffusion timestep  $t = 0.9$ . We then denoise it with 48 diffusion timesteps and without classifier-free guidance

## 9. Baselines

### 9.1. Warp-and-inpaint implementation

For a fair comparison with the warp-and-inpaint approach, we implement and train this baseline using the same pre-trained model as in our method. We use the same dataset described in 4.3 to fine tune the base Lumiere inpainting model to inpaint left-right disocclusion masks. We use [40] to estimate disparity of each pair of stereo frames,  $V^{\text{left}}$ ,  $V^{\text{right}}$  and obtain the disocclusion mask by computing left-right consistency of the disparity prediction. At training, the model is conditioned on the right video warped according to the estimated disparity,  $V_{\text{warped}}^{\text{right}}$ , and the corresponding disocclusion mask  $M$ , to denoise the left frame, with the standard diffusion objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, x_0, \epsilon} \left[ \|\epsilon - \epsilon_\theta(x_t, t, V_{\text{warped}}^{\text{right}}, M, c)\|_2^2 \right] \quad (3)$$

Here  $c$  is the text caption,  $x_t = \sqrt{\alpha_t} V^{\text{left}} + \sqrt{1 - \alpha_t} \epsilon$ , and  $\epsilon \sim \mathcal{N}(0, I)$ . Denote by  $\theta(x_t, t, V_{\text{warped}}^{\text{right}}, M, c)$  this model after training. At inference time, given a video  $V^{\text{right}}$ , we use SOTA monocular disparity estimation [11] to estimate video disparity  $D^V$ . As this estimation is scale and shift invariant, we fit a scale and shift parameter to the disparity map to align it with the disparity of our outputs (we first estimate the disparity of our outputs using [40]). We

then forward-warp the frames using depth ordered softmax splatting [31] and downsample the warped frames to obtain  $V_{\text{warped}\downarrow}^{\text{right}}$ . The inpainting mask here are the pixel locations that were not mapped onto by  $D^V$ . We open and dilate the mask to reduce temporal inconsistencies before feeding it along with the downsampled right eye video to  $\theta$  model, to obtain a low resolution inpainted video:

$$\theta(x_T, T, V_{\text{warped}\downarrow}^{\text{right}}, M, c) = V_{\text{base}}^{\text{inpainted}}$$

For spatial super resolution, we use the pretrained Lumiere SSR model and take a blended diffusion approach for maintaining faithfulness to the original video. Specifically, the input to the SSR model is the low resolution base inpainting model output  $V_{\text{base}}^{\text{inpainted}}$ , and at each timestep  $t$ , we blend the predicted clean super-resolved output

$$\hat{x}_0^t(x_t, t, V_{\text{base}}^{\text{inpainted}})$$

with the high resolution warped right video

$$V_{\text{warped}}^{\text{right}} = \text{softmax\_z\_splatting}(V, D^v)$$

according to the disocclusion mask  $M$ :

$$\hat{x}_0^t \leftarrow M \cdot \hat{x}_0^t(x_t, t, V_{\text{base}}^{\text{inpainted}}) + (1 - M) \cdot V_{\text{warped}}^{\text{right}}$$

This blending ensures that details in areas that appear in the input right video are preserved in the super-resolved left view. We use a the standard lumiere sampling of 256 and 32 diffusion timesteps for the base model and the SSR model, respectively, and a classifier free guidance of 8.

## 9.2. Stereo-Crafter

We use the official Stereo-Crafter repository <https://github.com/TencentARC/StereoCrafter>. For the depth splatting stage, we scale and shift the predicted disparity in the same way described in 9.1.

## 9.3. Deep3D

As the original paper implementation uses a deprecated codebase, we turn to a more recent implementation found in the link: <https://github.com/HypoX64/Deep3D>. Their training data consists of 3D movies, which are typically processed in a different manner then our data—the zero disparity plane is usually shifted to increase human comfort, making the RGB comparison difficult. We thus encourage the viewer to use anaglyph glasses for these results.

## 9.4. Dynamic Gaussian marbles

We optimize the Dynamic Gaussian Marbles using the official paper implementation <https://github.com/coltonstearns/dynamic-gaussian-marbles>, using their default real-world videos configuration. We observed the optimizing the representation for the full number of steps (100K) in this configuration diverges, and thus synthesize stereo views from it after 40K steps.