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

Ke Xing[1,2], Longfei Li[1], Yuyang Yin[1], Hanwen Liang[3], Guixun Luo[1†], Chen Fang[2]
Jue Wang[2], Konstantinos N. Plataniotis[3], Xiaojie Jin[1†‡], Yao Zhao[1], Yunchao Wei[1†]
[1]Beijing Jiaotong University, [2]Dzine AI, [3]University of Toronto
† Corresponding Author, ‡ Project Lead
https://ke-xing.github.io/StereoWorld/

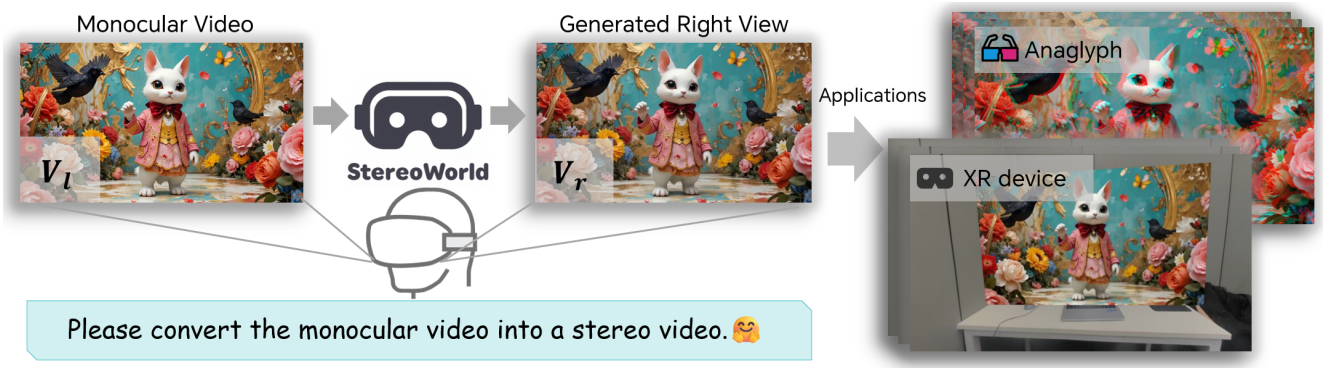

Figure 1. **Examples generated by StereoWorld.** Our method directly generates stereo videos from arbitrary monocular videos without requiring additional information. The results can be displayed on 3D glasses, XR headsets, and other stereoscopic devices.

Abstract

The growing adoption of XR devices has fueled strong demand for high-quality stereo video, yet its production remains costly and artifact-prone. To address this challenge, we present ***StereoWorld, an end-to-end framework*** *that repurposes a pretrained video generator for high-fidelity monocular-to-stereo video generation. Our framework jointly conditions the model on the monocular video input while explicitly supervising the generation with a* ***geometry-aware regularization*** *to ensure 3D structural fidelity. A spatio-temporal tiling scheme is further integrated to enable efficient, high-resolution synthesis. To enable large-scale training and evaluation, we curate a high-definition stereo video dataset containing over 11M frames aligned to natural human interpupillary distance (IPD). Extensive experiments demonstrate that StereoWorld substantially outperforms prior methods, generating stereo videos with superior visual fidelity and geometric consistency.*

1. Introduction

The rapid adoption of Extended Reality (XR) devices, such as the Apple Vision Pro and Meta Quest, has fueled growing demand for immersive stereo video. However, producing high-quality stereo footage relies on dedicated dual-camera rigs with tight calibration and synchronization, making it inaccessible to most creators. Given the vast abundance of monocular videos online, a scalable algorithm that can transform ordinary monocular footage into realistic, high-fidelity stereo video would be highly beneficial.

Existing approaches for monocular-to-stereo conversion can be broadly grouped into two paradigms. The first treats the task as novel-view synthesis (NVS) via 3D scene reconstruction. Traditional SfM pipelines [41] and modern neural rendering methods such as NeRF [32] and 3D Gaussian Splatting (3DGS) [26] attempt to recover scene geometry and camera parameters before rendering a new perspective for the right eye. However, this pipeline is vulnerable to pose inaccuracies and unconstrained scene dynamics. Alternative pose-free models [49, 53, 60] also struggle with the geometric ambiguities and non-rigid motions in real-world videos. Therefore, these approaches often generate stereos with unstable geometry and temporal inconsistency.

An alternative and more recent paradigm is the depth-warping-inpainting pipeline, powered by advances in diffusion models [11, 43, 50, 61, 63]. Depth is first estimated from the input video and used to warp frames into the target viewpoint; occluded regions are then hallucinated by

an inpainting model to plausibly fill in the missing areas. While conceptually simple, this pipeline suffers from distinct drawbacks. Specifically, the inpainting phase is decoupled from stereo geometry estimation, breaking pixel-level correspondence and resulting in texture distortions, color shifts, and stereo artifacts that degrade viewing comfort.

To address these limitations, we introduce **StereoWorld**, a novel end-to-end diffusion-based framework that converts a general monocular video generative model into a high-fidelity stereo generator. Instead of relying on fragile pose estimation or multi-stage warping pipelines, we leverages the rich spatio-temporal priors of the foundational video model to explicitly learn stereo geometry and generate coherent right-eye views directly from monocular input.

Our framework achieves this with the following proposed techniques. First, we extend the base video diffusion model with **monocular-conditioning** that allows the model to incorporate a monocular video as strong guidance. Second, to overcome the geometric inaccuracies of prior methods, we introduce a novel **geometry-aware regularization** strategy composed of a disparity and depth supervision. Disparity supervision enforces accurate stereo correspondence, mitigating cross-view misalignment and temporal disparity drift to improve stability and visual comfort. To further supplement geometric information, the model jointly diffuses RGB videos and their associated depth maps, explicitly learning 3D structure and providing stronger geometric guidance than RGB reconstruction alone. Furthermore, a **spatio-temporal tiling strategy** enables the efficient generation of high-resolution, long-duration videos. This optimization allows our framework to overcome the typical constraints of diffusion models, making it scalable for producing practical, high-fidelity content suitable for modern XR displays.

Another major challenge for stereo generation is the lack of suitable training data. Existing datasets feature *baselines* [1] that far exceed the human interpupillary distance (IPD). This wide *baseline* is unsuitable for XR devices as it leads to exaggerated parallax, which can easily cause visual discomfort for the viewer. To overcome this limitation, we curate a large-scale, high-resolution stereo video benchmark dataset aligned to human IPD, as summarized in Tab. 1, enabling both reliable training and fair evaluation.

Our main contributions are summarized as follows:

- We propose **StereoWorld**, the first fully end-to-end diffusion framework that adapts a pretrained monocular video generative model into a stereo generator with high visual fidelity and geometric accuracy.
- We build a large-scale, high-definition **stereo video dataset** aligned with human-IPD, featuring over 11M curated Blu-ray SBS video frames across diverse genres with comprehensive evaluation metrics.
- Extensive experiments demonstrate that StereoWorld substantially outperforms prior works in visual quality, geometric consistency, and temporal stability, with clear advantages in objective metrics and subjective perception.

[1] In stereo vision, the *baseline* is the precise physical distance between the optical centers of the two camera lenses.

Dataset	Domain	IPD-aligned	Available	Frames
Spring [30]	Optical Flow	✗	✓	5K
Sintel [5]	Optical Flow	✗	✓	1K
VKITTI2 [6]	Driving	✗	✓	21K
PLT-D3 [47]	Driving	✗	✓	3K
IRS [51]	Robotics	✗	✓	103K
TartanAir [54]	Robotics	✗	✓	306K
3D Movies [38]	Moives	✓	✗	75K
StereoWorld-11M	Moives	✓	✓	**11M**

Table 1. **Comparison of the stereo datasets.** Existing datasets are generally not IPD-aligned (e.g., Spring, VKITTI2), while datasets that are IPD-aligned are not publicly available (e.g., 3D Movies). Our StereoWorld is the first large-scale, IPD-aligned dataset.

2. Related Work

2.1. Novel View Synthesis

With the advent of deep learning, novel view synthesis (NVS) has progressed rapidly, evolving from traditional Structure-from-Motion (SfM) [41] methods to neural approaches such as NeRF [32] and 3D Gaussian Splatting (3DGS) [26]. While NeRF implicitly models scene geometry, 3DGS introduces explicit representations that improve reconstruction quality and efficiency. Recent works extend NVS to dynamic scenes and feed-forward reconstruction. For example, 4DGS [57] and Shape of Motion [52] reconstruct dynamic geometry from monocular videos, while methods like MVSplat [10], PixelSplat [8], and VGGT [49] leverage large pretrained models for fast, pose-free reconstruction. However, reconstructed geometry remains sparse, and the visual fidelity of novel views is still limited, restricting their applicability to stereo video generation.

2.2. Diffusion-based Stereo Generation

Existing diffusion-based stereo generation methods—whether training-free approaches such as SVG [11], StereoCrafter-Zero [43], T-SVG [23], and StereoDiffusion [50], or pretrained models including StereoCrafter [63], SpatialMe [61], StereoConversion [31], SpatialDreamer [29], ImmersePro [42], ReStereo [21], and GenStereo [37]—generally follow a similar pipeline: monocular depth estimation, view warping, and diffusion-based inpainting of occluded regions. But this paradigm disrupts the natural video distribution, causing spatial–temporal inconsistencies and degraded fidelity. In contrast, our method departs fundamentally from this paradigm by directly generating

stereo videos in an end-to-end manner, thereby ensuring cross-view consistency and preserving visual fidelity.

2.3. Video Diffusion Models

Diffusion models [16, 44–46] have achieved remarkable success in image generation [7, 33, 35, 39, 40, 56], inspiring their extension to 3D domain [13, 36, 58, 64] and the video domain [3, 4, 17, 18, 48, 59]. Early works [17] trained diffusion models directly on video datasets, while subsequent approaches augmented pretrained image diffusion models with temporal modules to capture motion dynamics [3, 15]. More recently, native video diffusion architectures built upon 3D-VAEs, such as Sora [4] and CogVideo [18], have demonstrated impressive spatio-temporal coherence and high-quality synthesis. The strong generative capacity and temporal consistency of pretrained video diffusion models make them a promising foundation for tasks like monocular-to-stereo video generation.

3. Methodology

In this section, we first review the fundamentals of video diffusion models (Sec. 3.1). We then detail our benchmark dataset construction (Sec. 3.2), present the core training strategy (Sec. 3.3), and describe practical optimizations for scalable generation (Sec. 3.4).

3.1. Preliminary: Video Diffusion Model

We build our framework on a pretrained text-to-video diffusion model based on the Diffusion Transformer(DiT) architecture [34]. This model first uses a 3D Variational Auto-Encoder (3D VAE) to encode videos into a latent representation. Then DiT blocks integrate self-attention and cross-attention modules to jointly capture spatio–temporal dependencies and text–video interactions. The model is trained under the Rectified Flow framework [27], where the forward process defines a linear trajectory between the data distribution and a standard normal distribution:

$$z_t = (1-t)z_0 + t\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and t denotes the diffusion timestep. The goal is to learn a velocity field v_Θ parameterized by the network weights Θ, which transforms random noise $z_1 \sim p_1$ into data samples $z_0 \sim p_0$ through an ordinary differential equation (ODE): $\frac{dz_t}{dt} = v_\Theta(z_t, t)$. During training, we minimize the Conditional Flow Matching (CFM) loss [12] by regressing the target vector field u_t that generates a valid probability path between p_0 and p_1:

$$\mathbb{E}_{t, p_t(z,\epsilon), p(\epsilon)} ||v_\Theta(z_t, t) - u_t(z_0|\epsilon)||_2^2, \tag{2}$$

where $u_t(z, \epsilon) := \psi_t'(\psi_t^{-1}(z|\epsilon)|\epsilon)$, and $\psi(\cdot|\epsilon)$ represents the mapping defined in Eq. 1.

3.2. StereoWorld-11M Dataset Construction

High-fidelity stereo video generation models are critically dependent on large-scale, high-quality training data. However, existing datasets are ill-suited for generating stereo video optimized for the human eye. These datasets [5, 6, 14, 25, 30, 47, 51, 54] are primarily developed for applications like depth estimation, autonomous driving, or robotics. Therefore, their *baseline* often exceeds 10 cm—far beyond the typical human IPD (55–75 mm), which can cause visual discomfort or dizziness when viewed stereoscopically.

To address this gap, we curated a new dataset tailored for stereo video generation with *baseline* aligned to natural human perception. We collected and cleaned over a hundred high-definition Blu-ray side-by-side (SBS) stereo movies from the Internet, spanning diverse genres such as animation, realism, war, sci-fi, historical, and drama, ensuring both visual diversity and richness for training. All videos are unified into the SBS format by stretching and horizontally cropping to obtain left–right views, each with a resolution of 1080p, 16:9 aspect ratio, and 24 fps frame rate. To match the training requirements of our base model (480p resolution, 81-frame inputs), we uniformly downscale each video to 480p. To enhance motion diversity and increase temporal information density, we uniformly sample 81 frames per clip at fixed intervals.

3.3. Framework

Our primary objective is to generate a corresponding right-view video $V_r \in \mathbb{R}^{c\times f\times h\times w}$ from the left-view video $V_l \in \mathbb{R}^{c\times f\times h\times w}$. Since existing video generative models are inherently monocular and lack this capability. To address this limitation, we adapt a pretrained monocular video generator to our stereo synthesis task through a simple yet effective conditioning way.

Monocular-conditioning. The first challenge in adapting a foundational monocular video generator for stereo synthesis is devising an effective conditioning way. The model must generate a geometrically consistent right-view V_r conditioned on the provided left-view V_l. The predominant existing paradigm is a multi-stage depth-warping-inpainting pipeline [11, 21, 63]. This approach lacks effective reference and fusion of information from the original left view during the inpainting process, leading to degraded visual quality in the generated results. A more integrated alternative, such as injecting left-view features via cross-attention, avoids this issue but requires significant architectural modifications and adds substantial computational overhead.

Inspired by ReCamMaster [1], we employ a simple yet effective conditioning strategy that aggregates left and right-view latents along the frame dimension. This allows the diffusion model to leverage its existing attention mechanisms to fuse information across space, time, and viewpoint simultaneously. Specifically, we encode the left

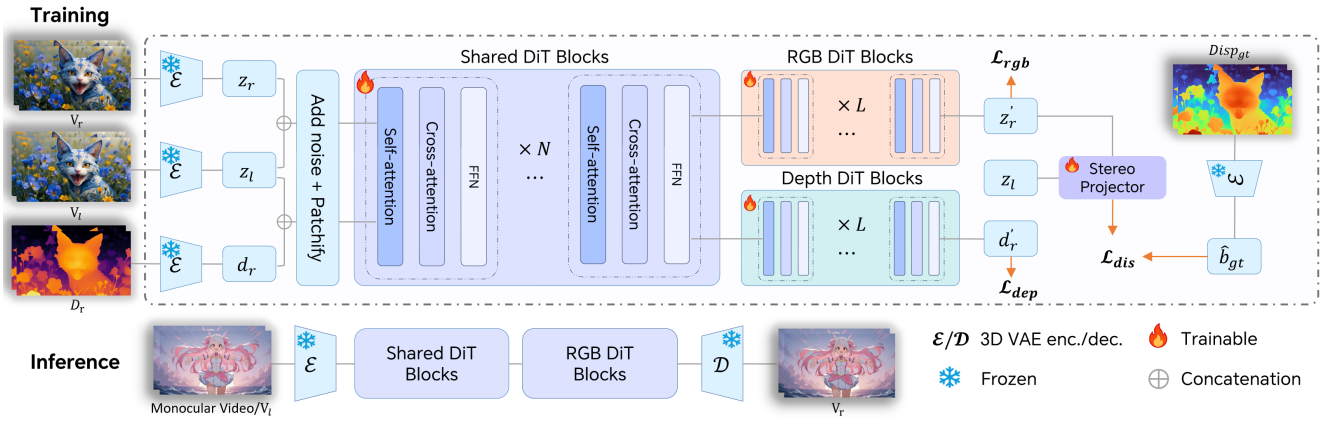

Figure 2. **Overall framework of StereoWorld.** Before training, we use Video Depth Anything [9] and Stereo Any Video [24] to obtain the depth maps D_r and disparity maps $Disp_{\text{gt}}$, and the left-view videos are then concatenated with the right-view videos and corresponding depth maps along the frame dimension in the latent space as conditioning inputs. During training, a lightweight differentiable stereo projector estimates the disparity between the input left-view and the generated right-view, which is supervised by disparity maps $Disp_{\text{gt}}$ via disparity loss to enforce accurate geometric correspondence. Additionally, the last few DiT blocks are duplicated to form dual branches, allowing the model to learn RGB and depth distributions separately to further supplement geometric information. During inference, only the shared and RGB DiT blocks are used, taking the monocular video as the sole input.

and right videos into latent space via the VAE encoder $\mathcal{E}$: $z_l = \mathcal{E}(V_l)$ and $z_r = \mathcal{E}(V_r)$, where $z_l, z_r \in \mathbb{R}^{c' \times f' \times h' \times w'}$. We then concatenate these latents along the frame dimension, $z_i = [z_l, z_r]_{\text{frame-dim}}$, where $z_i \in \mathbb{R}^{b \times c' \times 2f' \times h' \times w'}$ to serve as the direct input to the diffusion model , as illustrated in Fig. 2. This strategy is highly efficient, requiring no architectural changes , as the model's existing 3D spatio-temporal self-attention layers naturally fuse information by operating across all tokens from both views.

The essence of stereoscopy lies in the depth variations among scene objects, which define their hierarchical spatial relationships relative to the observer. However, relying solely on monocular conditioning and a standard $\mathcal{L}_{\text{rgb}}$ reconstruction loss is insufficient to capture such geometric structure. As shown in Fig. 8, we observe that the model struggles to implicitly learn complex geometric structures from only RGB data, resulting in outputs with weak stereo perception, such as flattened object boundaries or unstable disparities. This suggests that the model requires a more explicit signal to guide its understanding of 3D geometry. Therefore, we propose a geometry-aware regularization strategy to enhance the model's 3D perception capacity.

Geometry-aware Regularization. Our geometry-aware regularization consists of two complementary components—disparity and depth supervision—jointly designed to enhance stereo correspondence and geometric fidelity during right-view generation.

- Disparity supervision. To enforce accurate stereo correspondence and mitigate cross-view misalignment and temporal disparity drift, we introduce a disparity-based loss. First, we pre-compute the ground-truth disparity map $\hat{b}_{\text{gt}}$ by applying a pre-trained stereo matching network [24] to

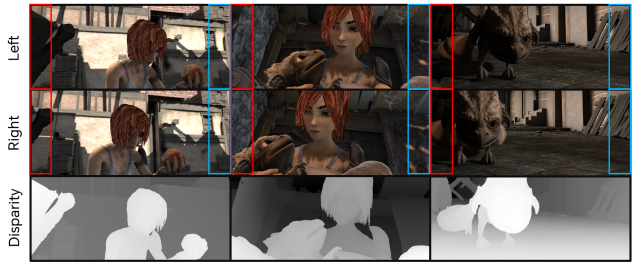

Figure 3. **Non-overlapping regions between stereo views.** Horizontal camera translation introduces non-overlapping content, which disparity supervision alone cannot constrain, motivating the use of depth-based supervision.

the ground-truth left (V_l) and right (V_r) video frames. This provides a geometrically accurate target. During training, after the model predicts the denoised right-view latent z'_r, we employ a lightweight, differentiable stereo projector κ to estimate the predicted disparity $\hat{b}_{\text{pred}}$ (Fig. 2). This projector takes the original left-view latent z_l and the generated right-view latent z'_r as input, $\hat{b}_{\text{pred}} = \kappa(z_l, z'_r)$. The overall disparity loss is defined as

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\log} + \lambda_{\text{l1}} \mathcal{L}_{\text{l1}}, \tag{3}$$

where λ_{l1} is a weighting hyperparameter. The two loss terms are defined as: $\mathcal{L}_{\log} = \mathbb{E}[d^2] - \lambda_1 (\mathbb{E}[d])^2$ that enforces global geometric consistency across disparity, $\mathcal{L}_{\text{l1}} = \mathbb{E}[|\hat{b}_{\text{pred}} - \hat{b}_{\text{gt}}|]$ that penalizes pixel-wise disparity errors, where $d = \log \hat{b}_{\text{pred}} - \log \hat{b}_{\text{gt}}$. Together, they explicitly guide the model to learn the stereo correspondence between left and right views, producing geometrically con-

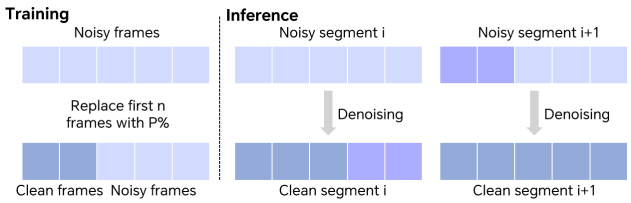

Figure 4. **Temporal tiling strategy.** During training, the first few frames of noisy latents are replaced with ground-truth frames with a probability p. During inference, long videos are split into overlapping segments, with the last frames of the previous segment used to guide the next, ensuring temporal consistency.

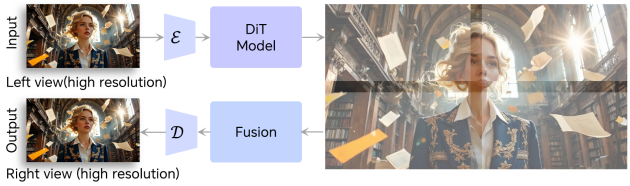

Figure 5. **Spatial tiling strategy.** During inference, high-resolution videos are encoded into latents, which are split into overlapping tiles. Each tile is denoised independently, and then the tiles are stitched back to the original size with overlapping regions fused before decoding.

sistent stereo videos.

Although disparity maps encode geometric cues, they only capture correspondences within the overlapping regions between left and right views. As illustrated in Fig. 3, when the camera translates horizontally to capture the right view, new regions appear on one side while others disappear on the opposite side. Since stereo matching operates only on overlapping content, the resulting disparity supervision provides incomplete geometric guidance for right-view generation. To address this limitation, we introduce an additional depth-based constraint.

- Depth supervision. To compensate for the missing geometric cues in non-overlapping regions, we introduce a depth-based supervision. Unlike disparity, depth provides a complete per-pixel geometric description of the target regions, including areas invisible to stereo matching. We predict the right-view depth maps and constrain the model to generate them consistently with the RGB frames. This reformulates the generation as a multi-objective joint prediction problem, where the model is simultaneously guided to learn the velocity field for both the RGB video and its corresponding depth map. Specifically, we denote the latent representation of the right-view depth map as d_r. Following the same Conditional Flow Matching logic as Sec. 3.1, we define a new objective, $\mathcal{L}_{\text{dep}}$, to train the model to predict the velocity field for this depth distribution. We denote the updated model parameters as Θ'. This results in a revised training objective with two main components:

$$\mathcal{L}_{\text{rgb}} = \mathbb{E}_{t,p_t(z,\epsilon),p(\epsilon)}||v_{\Theta'}(z_t,t) - u_t(z_0|\epsilon)||_2^2, \quad (4)$$

$$\mathcal{L}_{\text{dep}} = \mathbb{E}_{t,p_t(d,\epsilon),p(\epsilon)}||v_{\Theta'}(d_t,t) - u_t(d_0|\epsilon)||_2^2. \quad (5)$$

To provide the target d_r for this objective, we first precompute a per-frame depth map $D_r \in \mathbb{R}^{c\times f\times h\times w}$ for each right-view video V_r using a state-of-the-art depth estimation model [9]. This depth map is then encoded into its latent representation $d_r = \mathcal{E}(D_r)$ using the same VAE encoder.

A naive architectural approach for this multi-objective prediction would be to use the exact same set of DiT parameters (i.e., full parameter sharing) to learn the velocity fields for both RGB (z_r) and depth (d_r). However, this approach can hinder convergence and reduce learning efficiency, as the model is forced to reconcile potentially conflicting optimization gradients from two different data distributions within a single set of weights. To address this, we implement a specialized network architecture, as shown in Fig. 2, that balances shared representation learning with task-specific refinement. We keep the initial transformer blocks shared, allowing the model to capture joint texture and geometric representations from both tasks. We then duplicate the weights of the final few DiT blocks to create two specialized branches: one dedicated to predicting the RGB velocity field and the other for the depth velocity field. This design enables the model to build a robust, structured understanding of both scene layout and depth hierarchy, leading to more geometrically accurate synthesis.

Training Objectives. The overall training objective jointly supervises RGB reconstruction, depth consistency, and stereo disparity learning:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{dep}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}}, \quad (6)$$

This integrated objective promotes both visual fidelity and geometric correctness, leading to more perceptually coherent and stereoscopically realistic video generation.

3.4. Practical and Scalable Optimization.

Temporal Tiling Strategy. Our base model generates only short clips (81 frames, about 3s at 24 FPS). To handle longer videos, we split them into overlapping segments, using the last frames of each segment to guide the next for smooth transitions [63]. To further reduce flickering, during training we probabilistically replace first few frames of noisy latents with clean frames (probability p), enabling the model to learn robust long-range temporal consistency (Fig. 4).

Spatial Tiling Strategy. Our model is trained at 480p. To handle high-resolution videos beyond the 480p training resolution, we adopt block-wise latent diffusion [63]. High-resolution latents are split into overlapping tiles, each denoised independently, then stitched and fused before decoding (Fig. 5). This enables efficient high-resolution synthesis

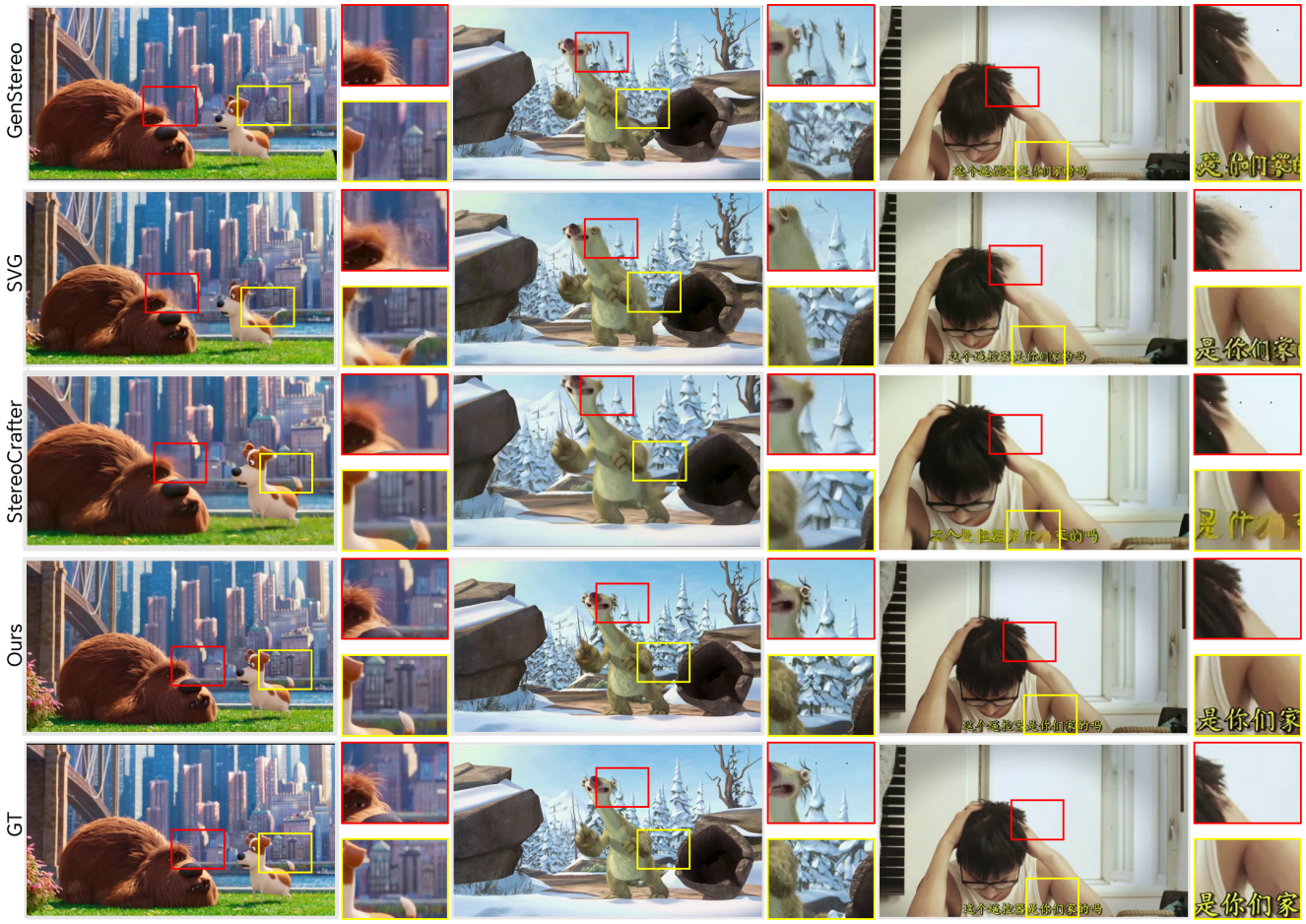

Figure 6. **Qualitative comparisons with state-of-the-art methods.** It shows that our method achieves the best generation quality, preserving fine details while maintaining strong visual consistency with the left view. Crucially, our method achieves far better text rendering quality than all baselines.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	IQ-Score ↑	TF-Score ↑	EPE ↓	D1-all ↓
GenStereo [37]	19.4486	0.6803	0.3008	0.4047	0.9642	35.0022	0.8954
SVG [11]	18.0256	0.5881	0.3467	0.4714	**0.9706**	33.2508	0.9630
StereoCrafter [63]	23.0372	0.6561	0.1869	0.4370	0.9685	24.7784	0.5271
Ours	**25.9794**	**0.7964**	**0.0952**	**0.5019**	0.9704	**17.4527**	**0.4213**

Table 2. Quantitative comparisons with state-of-the-art methods on visual quality and geometry accuracy. IQ-Score and TF-Score refer to image quality and temporal flickering scores from VBench [22].

while preserving spatial details and visual coherence.

4. Experiments

4.1. Experimental Setup

Implementation Details. We build our method upon Wan2.1-T2V-1.3B [48], which generates 5-second video clips at 16 FPS with a spatial resolution of 832×480. To obtain depth supervision, we employ Video Depth Anything [9] to estimate per-frame depth maps for all training videos. We further employ Stereo Any Video [24] to generate ground-truth disparity maps used for supervising the disparity during training. During fine-tuning, we adopt the LoRA [20] framework with a rank of 128, setting $\lambda_1 = \lambda_{11} = 0.1$ and $\lambda_{\text{dis}} = 0.5$. The learning rate is set to 1×10^{-4}, and the model is trained for one epoch, totaling approximately 9k optimization steps. Training is performed on 8 NVIDIA A800 GPUs using the AdamW optimizer [28]

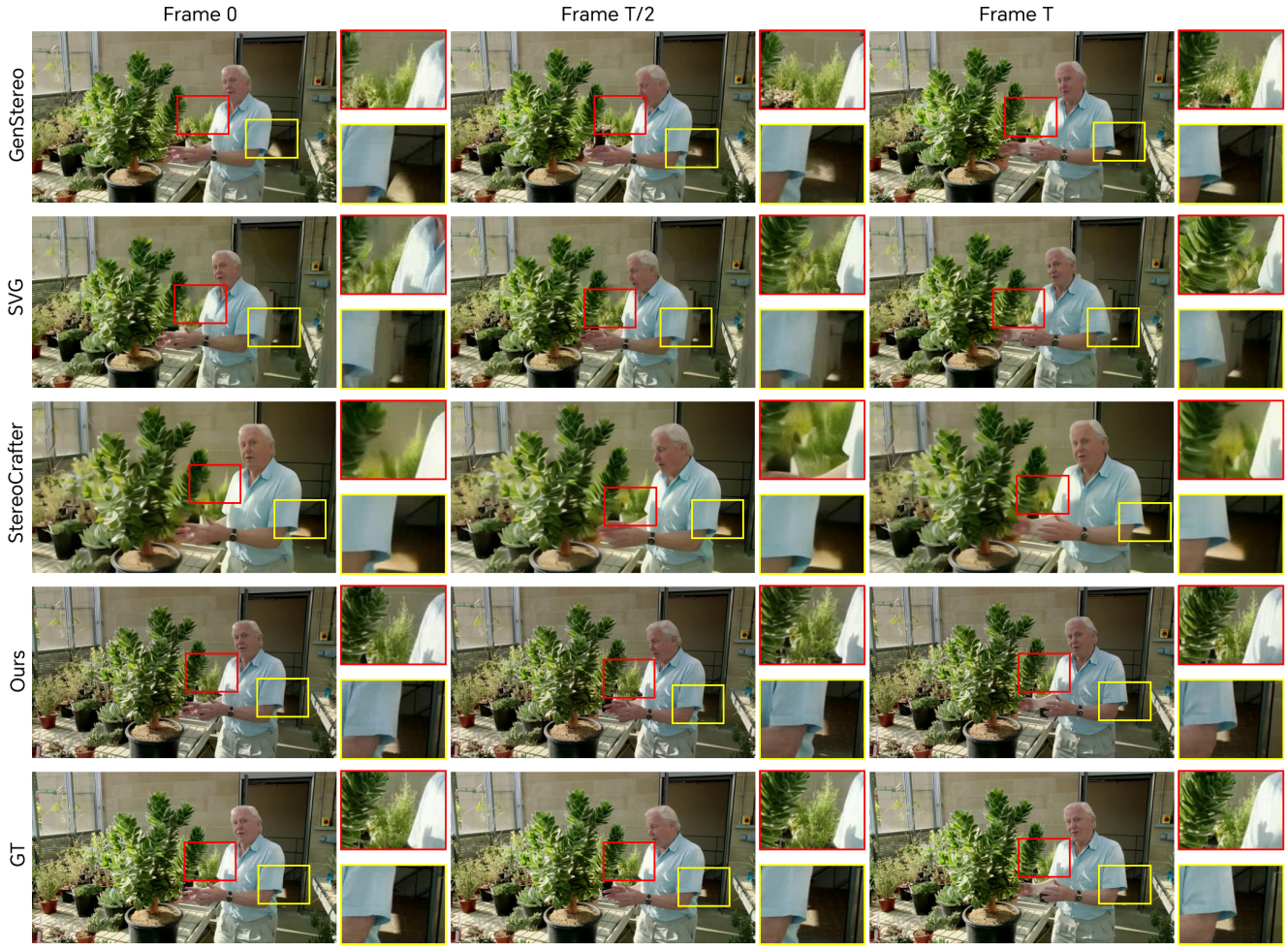

Figure 7. **Qualitative comparisons with state-of-the-art methods in the temporal dimension.** Our method maintains superior temporal consistency while preserving high visual quality and fine-grained detail fidelity compared to other methods.

under bfloat16 precision. The entire training process takes approximately 11 days to complete.

Dataset. We train our model on the web-collected video dataset aforementioned. After preprocessing, the dataset contains 142,520 video clips, each with a spatial resolution of 480×832 and 81 frames, corresponding to approximately 7 seconds at 12 FPS. We randomly select 1,000 clips as the test set, with the remaining clips used for training.

Evaluation Metrics. To quantitatively assess the generated right-view videos, we adopt PSNR [19], SSIM [55], and LPIPS [62] to measure the generation fidelity with respect to the ground-truth right views. In addition, we evaluate image quality (IQ-Score) and temporal flickering (TF-Score) from VBench [22] to measure visual quality and temporal consistency. For disparity-level evaluation, we employ Stereo Any Video [24] to estimate disparity maps from both the ground-truth stereo pairs and the generated pairs. We then compute EPE (End-Point-Error) [2]—the average pixel-wise disparity error—and D1-all [14], which denotes the percentage of pixels whose disparity error exceeds a given threshold (typically 3 pixels or 5% of the true disparity). These metrics together provide a comprehensive assessment of both visual fidelity and geometric accuracy.

Baselines. Our approach targets video-to-video stereo generation, and several related methods have not released official implementations, we select three representative baselines for comparison. Specifically, we use GenStereo [37] as a training-based image-to-image baseline, SVG [11] as a training-free video-to-video baseline, and StereoCrafter [63] as a training-based video-to-video baseline.

4.2. Comparisons

Qualitative Results. As shown in Fig. 6 and Fig. 7, the image-based method GenStereo fails to generalize to video data, exhibiting severe temporal instability and frame-wise distortions. The training-free method SVG struggles to in-

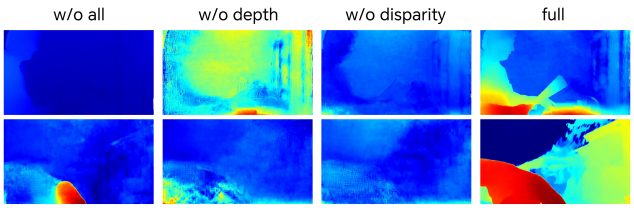

Figure 8. **Qualitative comparison results of ablation study.** Our full model exhibits better disparity shifts and structural perception.

paint occluded regions introduced by warping, producing visible artifacts and incomplete structures in the synthesized right views. While StereoCrafter generates visually coherent results, it tends to oversmooth fine textures, resulting in noticeable loss of high-frequency details. In contrast, our method achieves the most visually faithful and temporally stable results, accurately preserving scene geometry and fine-grained textures while maintaining strong semantic alignment with the left view. Most notably, StereoWorld excels in text rendering (a particularly challenging case for stereo generation) maintaining sharpness, legibility, and consistent spatial placement across both views, where all other baselines exhibit blurring or ghosting artifacts.

Quantitative Results. As presented in Tab. 2, the image-based baseline GenStereo and the training-free method SVG obtain the lowest overall scores, consistent with the qualitative observations. Although StereoCrafter achieves competitive results on perceptual quality metrics, it exhibits substantially higher errors on geometry-related measures such as EPE and D1-all, indicating inaccurate disparity estimation and weaker stereo correspondence. In contrast, our method consistently outperforms all baselines across both visual and geometric metrics, achieving the best balance between visual quality and geometry accuracy. These results highlight that StereoWorld not only enhances the realism and temporal coherence of generated videos but also produces geometrically consistent stereo pairs that align more closely with human interpupillary distance.

4.3. Ablation Studies

We analyze the contributions of geometry-aware regularization during training. As shown in Fig. 8 and Tab. 3, removing either component leads to noticeable degradation in both visual fidelity and geometric accuracy. The the disparity supervision enhances disparity magnitude, while depth supervision improves depth boundary perception and spatial structure. The full model achieves the best overall performance, demonstrating that these two complementary supervision signals are essential for producing geometrically accurate and visually faithful videos.

Depth Sup.	Disp. Loss	PSNR ↑	SSIM ↑	LPIPS ↓	EPE ↓	D1-all ↓
✗	✗	23.413	0.742	0.152	42.318	0.613
✓	✗	24.104	0.758	0.132	37.593	0.574
✗	✓	24.509	0.781	0.113	29.998	0.522
✓	✓	**25.979**	**0.796**	**0.095**	**17.453**	**0.421**

Table 3. Ablation on geometry-aware regularization. The full model achieves the best overall performance.

Method	SE ↑	VQ ↑	BC ↑	TC ↑
GenStereo [37]	3.8	3.6	4.3	3.7
SVG [11]	4.0	3.9	3.9	4.1
StereoCrafter [63]	4.2	4.0	4.1	4.2
Ours	**4.8**	**4.7**	**4.9**	**4.8**

Table 4. Results of Human evaluation with metrics: Stereo Effect (SE), Visual Quality (VQ), Binocular Consistency (BC), and Temporal Consistency (TC).

4.4. Human Evaluation

To further assess perceptual quality, we conducted a human evaluation with 20 participants who rated 15 generated scenes. Following the protocol of SVG [11], participants scored each scene on four aspects using a 1–5 scale: Stereo Effect (SE), Visual Quality (VQ), Binocular Consistency (BC), and Temporal Consistency (TC). Specifically, Stereo Effect measures the perceived 3D depth and immersion in XR displays, Visual Quality assesses image clarity and realism, Binocular Consistency evaluates alignment between left and generated right views, and Temporal Consistency reflects frame-to-frame stability over time. As summarized in Tab. 4, our method achieves the highest scores across all subjective dimensions. Participants consistently reported that StereoWorld delivers more natural depth perception, fewer cross-view mismatches, and smoother motion continuity compared to other approaches, validating its superior perceptual and stereoscopic experience.

5. Conclusion and Limitations

We present StereoWorld, an end-to-end diffusion-based framework for monocular-to-stereo video generation that produces high-quality results with strong visual consistency and geometric accuracy between left and right views. The model is further optimized for long-duration and high-resolution videos, demonstrating significant practical potential. Nevertheless, our approach has limitations. The disparity is learned in an end-to-end manner, limiting explicit control over the stereo *baseline*, and the current generation speed is relatively slow, requiring around six minutes per clip. Future work will explore model distillation and other acceleration strategies to improve efficiency and expand real-world applicability.

Acknowledgement

Xiaojie Jin's work was supported by the Talent Fund of Beijing Jiaotong University under Grant No. 2025XKRC015.

References

[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 3

[2] John L Barron, David J Fleet, and Steven S Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994. 7

[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. *arXiv preprint arXiv:2311.15127*, 2023. 3

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 3

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012. 2, 3

[6] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 2, 3

[7] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 3

[8] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024. 2

[9] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 4, 5, 6

[10] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 2

[11] 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. *arXiv preprint arXiv:2407.00367*, 2024. 1, 2, 3, 6, 7, 8

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[13] Jiashi Feng, Xiu Li, Jing Lin, Jiahang Liu, Gaohong Liu, Weiqiang Lou, Su Ma, Guang Shi, Qinlong Wang, Jun Wang, et al. Seed3d 1.0: From images to high-fidelity simulation-ready 3d assets. *arXiv preprint arXiv:2510.19944*, 2025. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 3, 7

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 3

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[19] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 7

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[21] 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. *arXiv preprint arXiv:2506.06023*, 2025. 2, 3

[22] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6, 7

[23] Qiao Jin, Xiaodong Chen, Wu Liu, Tao Mei, and Yongdong Zhang. T-svg: Text-driven stereoscopic video generation. *arXiv preprint arXiv:2412.09323*, 2024. 2

[24] Junpeng Jing, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. Stereo any video: Temporally consistent stereo matching. *arXiv preprint arXiv:2503.05549*, 2025. 4, 6, 7

[25] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13229–13239, 2023. 3

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2

[27] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[29] Zhen Lv, Yangqi Long, Congzhentao Huang, Cao Li, Chengfei Lv, Hao Ren, and Dian Zheng. Spatialdreamer: Self-supervised stereo video synthesis from monocular input. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 811–821, 2025. 2

[30] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2023. 2, 3

[31] Lukas Mehl, Andrés Bruhn, Markus Gross, and Christopher Schroers. Stereo conversion with disparity-aware warping, compositing and inpainting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4260–4269, 2024. 2

[32] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[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. *arXiv preprint arXiv:2307.01952*, 2023. 3

[36] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[37] Feng Qiao, Zhexiao Xiong, Eric Xing, and Nathan Jacobs. Towards open-world generation of stereo images and unsupervised matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26579–26589, 2025. 2, 6, 7, 8

[38] 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*, 44(3):1623–1637, 2020. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[40] 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*, 35:36479–36494, 2022. 3

[41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2

[42] Jian Shi, Zhenyu Li, and Peter Wonka. Immersepro: End-to-end stereo video synthesis via implicit disparity learning. *arXiv preprint arXiv:2410.00262*, 2024. 2

[43] Jian Shi, Qian Wang, Zhenyu Li, Ramzi Idoughi, and Peter Wonka. Stereocrafter-zero: Zero-shot stereo video generation with noisy restart. *arXiv preprint arXiv:2411.14295*, 2024. 1, 2

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[47] Joshua Tokarsky, Ibrahim Abdulhafiz, Satya Ayyalasomayajula, Mostafa Mohsen, Navya G Rao, and Adam Forbes. Plt-d3: A high-fidelity dynamic driving simulation dataset for stereo depth and scene flow. *arXiv preprint arXiv:2406.07667*, 2024. 2, 3

[48] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 6

[49] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 2

[50] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereodiffusion: Training-free stereo image generation using latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7416–7425, 2024. 1, 2

[51] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019. 2, 3

[52] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 2

[53] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1

[54] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 2, 3

[55] 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*, 13(4):600–612, 2004. 7

[56] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3

[57] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 2

[58] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 3

[59] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[60] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 1

[61] Jiale Zhang, Qianxi Jia, Yang Liu, Wei Zhang, Wei Wei, and Xin Tian. Spatialme: Stereo video conversion using depth-warping and blend-inpainting. *arXiv preprint arXiv:2412.11512*, 2024. 1, 2

[62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[63] 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. *arXiv preprint arXiv:2409.07447*, 2024. 1, 2, 3, 5, 6, 7, 8

[64] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 3