UNIGENSTEREO: UNIFIED STEREO VIDEO GENERATION WITH FLEXIBLE CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-fidelity stereo video generation is crucial for immersive extended reality (XR), robotics, and embodied AI. In practice, the geometric information available at inference time varies widely: dense per-frame depth or disparity is sometimes available, sometimes only the camera baseline is known, and sometimes a monocular clip arrives with neither. Existing methods each commit to one of these cases, and those taking no geometric input cannot control the stereo geometry they produce. We propose UniGenStereo, which synthesizes the right view from a monocular video with one set of weights for all three. Disparity is the camera baseline times the ratio between focal length and metric depth. The ratio does not depend on the baseline, so one prediction serves them all; we predict it from the input video with a lightweight head on the generation backbone. (i) Supplied disparity is used as given and the head is never invoked; (ii) a calibrated baseline scales the predicted ratio; and (iii) with no baseline, a standard interpupillary value stands in. Whatever its source, the disparity warps the input view into the same conditioning channels, so one backbone covers all three and varying the baseline alone controls the synthesized disparity. Experiments show state-of-the-art results in all three settings, including on real-world capture.

1 INTRODUCTION

The rapid growth of extended reality (XR), robotics, and embodied AI has created a strong demand for high-fidelity stereoscopic video. Modern video diffusion models synthesize monocular content with striking realism, yet extending them to stereo is non-trivial: the generated view pair must satisfy binocular geometry at every frame while remaining temporally coherent across the sequence.

A closer look at how stereo video is produced in practice reveals that the *geometric information available at inference time* varies fundamentally across applications. A film post-production or simulation pipeline typically has high-quality per-frame disparity; a calibrated robotic or autonomous-driving rig knows its camera baseline but has no target-view geometry; and an in-the-wild XR application converting user videos for VR playback receives nothing beyond a monocular clip. Existing methods, however, each commit to a single point on this spectrum. *Depth-Warp-Refine* pipelines (Zhang et al., 2025; Dai et al., 2025; Zhao et al., 2024; Huang et al., 2025) and their end-to-end descendants (Shvetsova et al., 2026; Metzger et al., 2026) require dense target-view geometry (G0), so their quality is bounded by the geometry supplied to them. *Geometry-free* approaches (Geyer et al., 2025; Xing et al., 2026; Shen et al., 2025) synthesize the right view directly from monocular input (G2), but having internalized a fixed baseline distribution from training data, they offer no control over the stereo geometry. In between, *camera-conditioned* methods such as ReCamMaster (Bai et al., 2025) and the concurrent StereoSpace (Behrens et al., 2026) accept compact target-camera metadata (G1), yet remain restricted to that single conditioning form. To our knowledge, no existing method covers all three with one set of weights.

Our key observation is that these seemingly disparate settings all answer the same underlying question: *where does the target-view geometry come from?* With dense disparity, the answer is supplied externally. With only a baseline, the scene structure must still be inferred, and the baseline merely fixes its physical scale. With monocular input alone, the structure must again be inferred and the scale has to be assumed. This suggests a single unifying mechanism: an internal geometry predictor that estimates, directly from the input video, the ratio $g = f/Z$ between focal length and metric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

depth. This ratio, rather than depth itself, is the right prediction target on two counts. It is independent of the target baseline, so one prediction serves every baseline an application may request; and it is the complete geometric input the warp needs, so nothing has to be estimated separately and then divided. Disparity then follows from the classical relation $D = B \cdot g$. When external disparity is available it is used directly and the predictor is bypassed (G0); when a calibrated baseline B is known, it scales the predicted g into a metric disparity field (G1); and when nothing is provided, a standard interpupillary baseline stands in for B, so the model runs with no geometric input at all (G2). In all three cases the resulting disparity drives the same warp-based conditioning (warped latents, warped coordinates, and disocclusion masks), so one backbone with one set of weights serves the entire spectrum, adapting to whatever geometric information is available at inference time.

We propose UniGenStereo, built on a video diffusion backbone with the geometry predictor operating in latent space and trained jointly with generation. Separating structure (g) from scale (B) has a further consequence: varying B alone produces continuously controllable disparity. Experiments support the overall design. On a multi-camera stereo generation benchmark, UniGenStereo attains state-of-the-art results under all three settings; it transfers to real-world head-mounted capture; and same-scene sweeps show the generated disparity tracking the commanded baseline with a median per-scene R^2 of 0.96, while removing the geometry-driven conditioning degrades quality by a large margin.

Our main contributions are summarized as follows:

- We present UniGenStereo, a unified stereo video generation framework in which one model with one set of weights treats dense target-view geometry, a calibrated baseline, and purely monocular input as interchangeable conditioning settings.
- We propose an internal geometry predictor that estimates the ratio $g = f/Z$ from the input video, unifying all three settings through the relation $D = B \cdot g$ and enabling continuous control without external geometric preprocessing.
- We introduce a two-stage training scheme that first supervises the predictor against dense ground-truth geometry and then refines it jointly with generation through a differentiable warp, using mixed stereo datasets for geometry alone so that their uneven visual quality never reaches the generator.

2 Related Work

2.1 Video Generation Models and Geometric Priors

Diffusion models and transformer architectures have rapidly advanced video generation. Open foundation models such as Stable Video Diffusion (Blattmann et al., 2023), CogVideo (Yang et al., 2025), and Wan (Team Wan, 2025) have made strong spatiotemporal priors broadly accessible. We build on Wan for its video prior and for its 3D self-attention backbone, which lets a source and a target view be denoised as one sequence.

These priors also carry geometry. Marigold (Ke et al., 2025) showed that a pre-trained image diffusion model can be fine-tuned into a strong monocular depth estimator, a recipe extended by GeoWizard (Fu et al., 2024) and Lotus (He et al., 2025) and carried to video by DepthCrafter (Hu et al., 2025) and ChronoDepth (Shao et al., 2025), alongside large-scale feed-forward models such as the Depth Anything series (Lin et al., 2025). Taken together, these results establish that a generative video backbone already encodes the scene structure a stereo warp needs; our framework reads that structure out of the backbone itself rather than obtaining it from a separate network.

2.2 Stereo Video Generation

Existing methods can be organized by the geometric information they require at inference time. A first family relies on dense target-view geometry: the classical "Depth-Warp-Refine" paradigm forward-warps the source view with estimated depth or disparity and inpaints disocclusions, spanning image-level methods (Watson et al., 2020; Wang et al., 2024; 2025) and video-level systems (Zhao et al., 2024; Zhang et al., 2025; Dai et al., 2025; Huang et al., 2025). GenStereo (Qiao et al., 2025) belongs to this family and conditions synthesis on an explicit warp together with a disocclusion mask rather than on an implicit code, a design we carry into latent space. Recent end-to-end

methods stay in the same family: M2SVid (Shvetsova et al., 2026) folds the warp and the inpainting into a single decoder conditioned on a ground-truth disparity warp, and Elastic3D (Metzger et al., 2026) dispenses with the warp altogether, steering generation with a scalar disparity token calibrated from ground-truth disparity statistics. These methods achieve strong pixel alignment when accurate geometry is supplied, but their quality is bounded by the geometry given to them.

At the opposite end of the spectrum, a second family generates the right view directly from a monocular video without any geometric conditioning, including ImmersePro (Shi et al., 2024a), StereoCrafter-Zero (Shi et al., 2024b), Eye2Eye (Geyer et al., 2025), StereoPilot (Shen et al., 2025), Mono2Stereo (Yu et al., 2025b), and StereoWorld (Xing et al., 2026). Free of external geometry, these models embed a fixed baseline distribution in their weights and provide no mechanism to adapt the stereo geometry to a target device or rig.

Between these extremes, a third family conditions generation on compact target-camera metadata. Generic camera-controlled models steer synthesis with explicit trajectories (Xiao et al., 2025; Yu et al., 2025a; Ren et al., 2025), most prominently ReCamMaster (Bai et al., 2025), which re-renders videos under novel 6-DoF camera paths using a large multi-camera dataset; stereo generation is a constrained special case of this task, and we fine-tune ReCamMaster on stereo data as a strong baseline in our experiments. Closest to our intermediate setting, the concurrent StereoSpace (Behrens et al., 2026) conditions image-level stereo synthesis on a physical baseline within a canonical rectified space, demonstrating that metric viewpoint conditioning alone can drive stereo synthesis. UniGenStereo differs in scope and mechanism: it operates on video, where temporal coherence must be maintained; it derives its guidance from an internal geometry predictor rather than viewpoint encoding alone; and, most importantly, it unifies all three conditioning settings in a single model rather than committing to one.

2.3 Stereo Generation Benchmarks

Evaluating controllable stereo generation requires ground truth across multiple baselines of the same scene, which real capture rarely provides. StereoGenBench (Cui et al., 2026) addresses this with a synthetic multi-camera benchmark that renders each scene under multiple baselines and organizes evaluation by conditioning tier: dense target-view geometry (G0), target-camera metadata (G1), and monocular input (G2). These tiers correspond exactly to the three settings our framework covers, so we adopt the benchmark as our primary evaluation protocol, complemented by a real-world stereo benchmark at an interpupillary baseline.

3 Method

3.1 Overview

Problem formulation. Given a left-view video $v_l \in \mathbb{R}^{F\times 3\times H\times W}$ and whatever geometric information is available, we synthesize the right-view sequence v_r, working in the latent space of a pre-trained VAE encoder $\mathcal{E}$ with x_l and x_r the two latents. For a rectified pair with baseline B and focal length f, the views differ by a horizontal disparity

$$D(\mathbf{p}) \;=\; \frac{B\,f}{Z(\mathbf{p})} \;=\; B\cdot g(\mathbf{p}), \qquad g(\mathbf{p}) \;\triangleq\; \frac{f}{Z(\mathbf{p})}, \tag{1}$$

where $\mathbf{p}$ indexes pixels and Z is metric depth. Eq. (1) splits the target-view geometry into a field g fixed by the scene and the capturing camera, and an extrinsic scalar B that sets its physical scale. This split is at the core of our design.

Pipeline. The three settings differ only in how much of Eq. (1) is given at test time: disparity D itself (G0), the baseline B alone (G1), or neither (G2). We build on a Diffusion Transformer (DiT) with an attached geometry head $\mathcal{G}_\phi$ that predicts g from the left video; the baseline only scales it:

$$D \;=\; \begin{cases} D_{\text{ext}}, & \text{(G0) external disparity,} \\ B\cdot\hat{g}, & \text{(G1) calibrated baseline,} \\ B_{\text{IPD}}\cdot\hat{g}, & \text{(G2) monocular input, standard interpupillary baseline,} \end{cases} \tag{2}$$

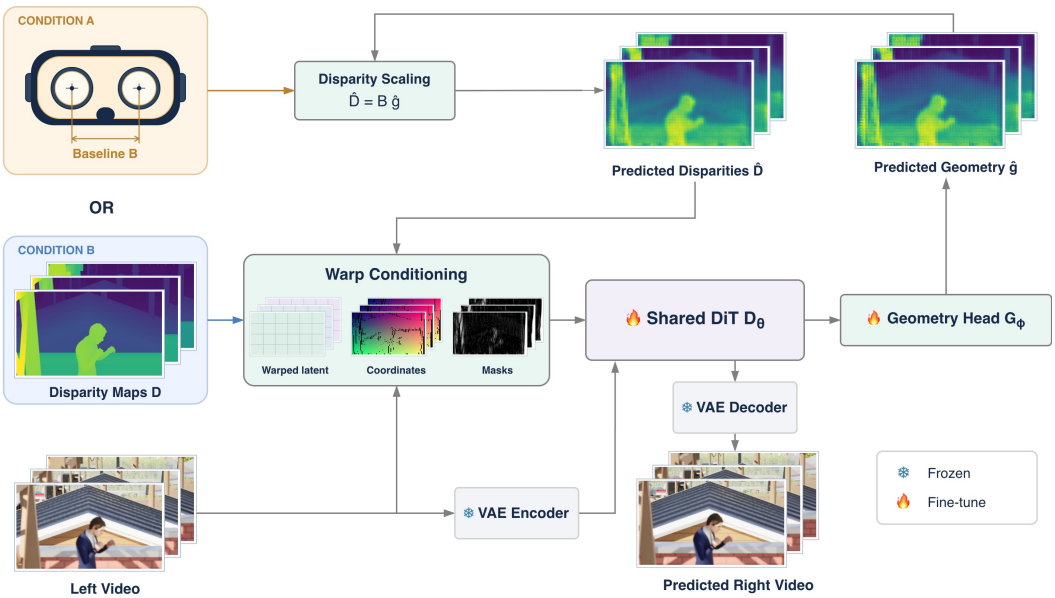


Figure 1: Overview of UniGenStereo. The geometry head $\mathcal{G}_\phi$, attached to the shared DiT $\mathcal{D}_\theta$, predicts $\hat{g}$, its estimate of $g = f/Z$, from the left video; a target baseline scales it into a disparity field (Condition A), which external disparity replaces when available, bypassing the head (Condition B). Either field warps the left video into the conditioning channels of Eq. (3). Only $\mathcal{D}_\theta$ and $\mathcal{G}_\phi$ are fine-tuned. Stage 1 keeps the two forward paths separate; Stage 2 routes $\hat{D}$ into the warp so the generation loss also reaches $\mathcal{G}_\phi$.

where $\hat{g} = \mathcal{G}_\phi(v_l)$ is the head's estimate of g; in the last two branches Eq. (2) is Eq. (1) with g replaced by that estimate, while the first bypasses it entirely. We write $\hat{D} = B \cdot \hat{g}$ for the resulting field. Following ReCamMaster (Bai et al., 2025), the target and source latents are concatenated along the *frame* axis before patchification, so 3D self-attention operates within and across views; the source half stays clean and the loss falls on the target half only. Whatever its source, D warps the left view into the same conditioning channels (Sec. 3.2), so one set of weights covers all three settings with no setting-specific module (Fig. 1).

3.2 Geometry-Driven Warp Conditioning

Given a disparity field from any source in Eq. (2), we follow GenStereo (Qiao et al., 2025) and turn it into an explicit, dense condition rather than an implicit code. Let $\mathcal{W}(\cdot\,; D)$ be the forward warp $(u, v) \mapsto (u - D, v)$, with collisions resolved by a z-buffer so the nearer surface wins, and let M mark the pixels it leaves uncovered:

$$c \;=\; \big[\; \underbrace{\mathcal{E}\big(\mathcal{W}(v_l; D)\big)}_{16} \;\big\|\; \underbrace{\mathcal{W}\big(\gamma(\mathbf{p}); D\big)}_{10} \;\big\|\; \underbrace{M}_{1} \;\big], \tag{3}$$

where γ stacks the normalized pixel coordinates with their sine and cosine at two frequencies ($2 + 8 = 10$ channels) and $\|$ is channel concatenation. The warped latent gives a photometric starting point, the coordinates state where each target pixel came from, and the mask which pixels must be inpainted. We concatenate c to the noisy latent on the target half and zeros on the source half, widening the patch-embedding convolution from 16 to 43 channels with the new channels zero-initialized, so the widened model matches the pre-trained one at initialization. Setting $M \equiv 1$ with the rest at zero encodes the absence of geometry: what the model sees when the warp is dropped in training, and our *No-warp* ablation. Since c depends on D alone, external and predicted disparity yield conditions of identical form.

3.3 INTERNAL GEOMETRY HEAD

The head predicts $g = f/Z$ rather than depth for two reasons. It is exactly the quantity that converts a baseline into disparity, so nothing intermediate has to be estimated and then divided; and it is independent of the target baseline, so a single prediction serves every baseline an application may request. Neither g nor depth is uniquely identifiable from monocular pixels, since an overall scene scale leaves the image unchanged: both must be learned from data carrying metric geometry rather than recovered from the pixels alone. What differs is how the error behaves. A depth pipeline must estimate metric depth and a focal length separately and then divide, so its accuracy depends on how the two errors combine; Appendix A.2 shows this on StereoGenBench, where supplying a state-of-the-art zero-shot depth model with the ground-truth focal length makes its native-scale disparity worse rather than better. Internally the head regresses the resolution-normalized quantity $\tilde{g} = f/(100\,Z\,W)$ (normalized disparity per centimeter of baseline) so that its output does not depend on the working resolution; writing $\hat{\tilde{g}}$ for the head's output, $\hat{g} = 100\,W\,\hat{\tilde{g}}$ is its estimate of g in the units of Eq. (1), with W the frame width.

$\mathcal{G}_\phi$ is a lightweight head with 9.5M parameters, not a separate network: it extracts source-half tokens from DiT blocks 8, 14, 20 and 26, projects each to 256 channels, fuses them in a DPT-style manner (Ranftl et al., 2021), applies two $3 \times 3 \times 3$ convolutions across frames for temporal stability, and upsamples to the latent grid. It outputs $\tilde{g}$ through a softplus, hence strictly positive, together with a heteroscedastic log-scale $s = \log \sigma$. Because the head must run before any right-view content exists, we read it from a dedicated pass reproducing the test-time state: target half pure noise, source half the clean left latent, timestep at its maximum, condition $M \equiv 1$. This matters: when a warp *is* supplied, the warped coordinate channels of Eq. (3) contain the disparity field itself, and a head reading them could copy it instead of inferring it. As a second safeguard against residual leakage through 3D self-attention, we apply the geometry losses only on warp-dropped samples.

3.4 TRAINING

Two stages. We train in two stages because the two data sources differ in what they can supervise. Dense, metrically correct geometry exists only in simulation, so the in-domain split is the only source that can supervise g directly and teach synthesis under exact geometry; the external datasets provide a calibrated baseline and a true right view but no dense depth, and cover a far wider range of rigs and scenes. Stage 1 therefore trains the two tasks separately on the in-domain split, and Stage 2 couples them while opening the geometry head to the external pool. Both stages generate with the flow-matching objective of the backbone, using a Huber loss $\mathcal{H}_\delta$ ($\delta = 1$) in place of the squared error for robustness to structural artifacts in training pairs, on the target half only: $\mathcal{L}_{\text{gen}} = \mathbb{E}[\, w(t)\, \mathcal{H}_\delta(v_\theta(x_t, t, c),\, u_t)]$, with u_t the target velocity and $w(t)$ the scheduler's timestep weight.

Stage 1. We train on the in-domain multi-baseline split alone, the only source with dense ground-truth geometry at many baselines per scene. It is rendered under two complementary baseline distributions and we use both: a normal distribution centered on the human interpupillary distance ($\approx 0.063\,\text{m}$) (Dodgson, 2004), and a uniform distribution over a wide range that forces metric accuracy across arbitrary rigs. Because g^* is available per pixel, the head is supervised directly and in the log domain, so the residual measures a ratio, and near and far surfaces contribute comparably. Unlike the scale-invariant losses common in monocular depth, no per-image mean is subtracted, so a global error in the predicted scale is penalized rather than absorbed. The residual is weighted by the head's own confidence:

$$\mathcal{L}_g \;=\; \frac{1}{|\Omega|} \sum_{\mathbf{p} \in \Omega} \Big(\big|\log \hat{g}(\mathbf{p}) - \log g^*(\mathbf{p})\big|\, e^{-s(\mathbf{p})} \;+\; s(\mathbf{p}) \Big), \tag{4}$$

over the pixels Ω with valid ground truth; the e^{-s} factor lets the head discount what it cannot resolve, such as sky and textureless regions, while $+s$ keeps it from declaring everything uncertain. The generator meanwhile receives a warp built from ground-truth geometry, dropped with probability 0.2, and $\mathcal{G}_\phi$ is supervised on exactly those dropped samples. The two forward paths stay separate: the head's output never enters the generation path, so the generator never sees its early, unreliable estimates. Gradients from $\mathcal{L}_g$ still reach the tapped blocks, so the shared trunk learns features that serve both tasks. The stage objective is $\mathcal{L}^{(1)} = \mathcal{L}_{\text{gen}} + 0.3\,\mathcal{L}_g$.

Table 1: Training data. Stage 1 uses the in-domain multi-baseline split alone, the only source with dense ground-truth geometry at many baselines per scene; Stage 2 adds the external datasets as geometry-only samples, each resampled to a common per-epoch quota. External datasets are sorted by size. **Type** (R = Real, S = Synthetic), **Scenario** (I = Indoors, O = Outdoors).

Dataset	Type	Scenario	# Videos	Baseline (m)
Stage 1 & 2 · in-domain, dense ground-truth geometry				
StereoGenBench (Cui et al., 2026)	S	I + O	84,213	0.01 – 5.64
Stage 2 · external, geometry-only supervision				
StereoCarla (Guo et al., 2025)	S	O	8,610	0.01 – 0.3
TartanAir (Wang et al., 2020)	S	I + O	3,944	0.25
SID (El-Shair et al., 2024)	R	O	2,218	0.12
Dynamic Replica (Karaev et al., 2023)	S	I	1,932	0.04 – 0.3
Cityscapes (Cordts et al., 2015)	R	O	1,885	0.22
IRS (Wang et al., 2021)	S	I	1,315	0.10
SouthKensington (Jing et al., 2024)	R	I + O	1,293	0.12
WMGStereo-Nature (Yan et al., 2025)	S	O	427	0.075 – 0.5
InfinigenSV (Jing et al., 2024)	S	O	308	0.075 – 1.205
VKITTI 2 (Cabon et al., 2020)	S	O	272	0.53
PLT-D3 (Tokarsky et al., 2024)	R	O	182	0.12
SimStereo (Jospin et al., 2022)	S	I	173	0.16
Spring (Mehl et al., 2023)	S	O	73	0.065
MPI-Sintel (Butler et al., 2012)	S	I + O	20	0.10
Total (external)			**22,652**	**0.01 – 1.205**

Stage 2. We add the external stereo datasets of Table 1, excluding single-frame datasets, sequences with dropped frames, and abstract synthetic data lacking realistic textures, and keeping only datasets with measured rather than estimated calibration, since an untrustworthy baseline corrupts the disparity $\hat{D} = B \cdot \hat{g}$ that Eq. (5) is built on. They span driving, robotics and human-centric scenes with baselines from 0.01 to 1.205 m, each resampled to a common per-epoch quota so no large one dominates; the SVD evaluation benchmark is held out. Carrying no g^*, they are supervised indirectly, by whether the predicted disparity transports the left view onto the right. Because this gradient must reach $\mathcal{G}_\phi$, the warp here is a differentiable forward splat $\mathcal{W}_{\text{soft}}$ weighted by a softmax-splatting importance (Niklaus & Liu, 2020), which acts as a soft z-buffer, applied on the latent grid with $\hat{D}_{\text{lat}} = \hat{D}\, W_{\text{lat}}/W$ the predicted field rescaled to the latent width W_{lat}:

$$\mathcal{L}_{\text{wc}} \;=\; \big\| \mathcal{W}_{\text{soft}}(x_l; \hat{D}_{\text{lat}}) - x_r \big\|_{1,\mathcal{V}} \,/\, \big\| x_r \big\|_{1,\mathcal{V}}, \tag{5}$$

normalized by the target magnitude over the visible set $\mathcal{V}$ so the term is scale-free across datasets. This stage also closes the loop between the two components: on 75% of the warp-dropped in-domain samples the conditioning tensor is rebuilt online from the model's own $\hat{g}$, splatting the clean left latent and the coordinate grid with $\mathcal{W}_{\text{soft}}$ rather than warping in pixel space and re-encoding. The condition is then a differentiable function of $\hat{g}$, so the generation loss backpropagates *through the warp* into $\mathcal{G}_\phi$ and the head is trained by what synthesis needs, not by a reference field alone; the hard z-buffer warp of Sec. 3.2 is kept for externally supplied disparity and at test time. The geometry losses run the other way: $\mathcal{G}_\phi$ reads the deployment-matched pass of Sec. 3.3 with its taps detached, so $\mathcal{L}_g$ and $\mathcal{L}_{\text{wc}}$ update the head alone and cannot reshape the backbone. External samples receive no generation loss, so datasets of uneven visual quality sharpen the head across diverse rigs while the generator stays shaped by in-domain data alone. The stage objective is $\mathcal{L}^{(2)} = \mathcal{L}_{\text{gen}} + 0.1\,\mathcal{L}_g + 0.5\,\mathcal{L}_{\text{wc}}$.

4 Experiments

4.1 Benchmarks and Metrics

Our primary evaluation is StereoGenBench (Cui et al., 2026), a synthetic multi-camera benchmark that renders each scene under multiple baselines and organizes evaluation by conditioning

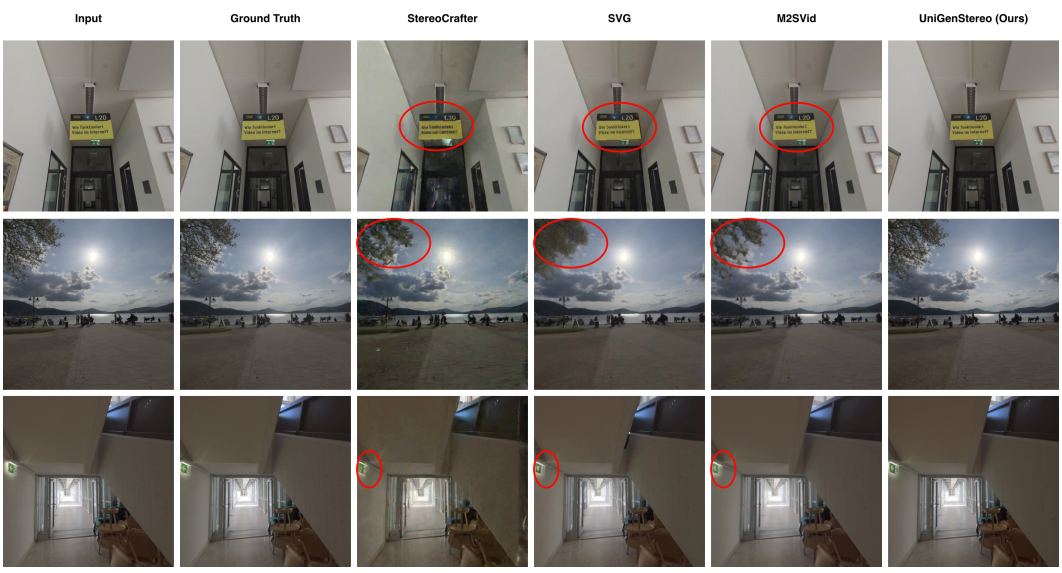


Figure 2: Qualitative comparison on SVD under supplied geometry (G0). Disocclusions along depth discontinuities are where the methods separate: warp-and-inpaint pipelines leave stretched or duplicated texture at object boundaries, while our warp-based conditioning supplies the transported source appearance directly and the generator only has to fill what the warp leaves uncovered.

tier (G0/G1/G2). We evaluate on the Seen split (546 scenes: 281 Uniform, 265 IPD-Gaussian), with the unseen split (193 scenes) in the appendix; all clips contain 81 frames at 832×480. For real-world generalization we use the Spatial Video Dataset (SVD) (Izadimehr et al., 2025), 159 head-mounted captures at $B = 0.0638\,\text{m}$. We follow the protocol of Metzger et al. (2026) throughout: all methods are evaluated at 512×512 resolution, and, since SVD carries no ground-truth depth, the G0 tier supplies the same FoundationStereo (Wen et al., 2025) pseudo disparity to every method.

Following recent stereo conversion protocols (Shvetsova et al., 2026; Cui et al., 2026), we report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) for overall fidelity, patch-wise PSNR (P-PSNR) for local photometric consistency along epipolar lines, and FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) for distributional realism. Geometry prediction and baseline controllability are evaluated separately; details are given in Appendices A.2 and A.3. StereoGenBench's calibration controls (Cui et al., 2026) verify that this suite cannot be gamed by degenerate strategies such as copying the input view.

4.2 Comparisons with Other Methods

On StereoGenBench (Table 2) UniGenStereo leads on PSNR, SSIM, LPIPS, FID and FVD in all three settings and ties M2SVid on P-PSNR under supplied geometry. The margin widens as geometry is withheld: 0.24–$1.34\,\text{dB}$ ahead under supplied disparity, 2.46–$4.52\,\text{dB}$ under a calibrated baseline, and $7.24\,\text{dB}$ under monocular input alone, where competing methods have no mechanism to recover the geometry they are not given. StereoPilot, the fine-tuned ReCamMaster, and StereoWorld build on the same Wan2.1-1.3B backbone as our model, so the G1 and G2 margins reflect the conditioning design, not backbone capacity. (Fig. 2)

The ordering transfers to real-world dataset (Table 3). With supplied geometry we obtain the best PSNR and LPIPS; with a calibrated baseline we lead the fine-tuned ReCamMaster by $1.3\,\text{dB}$; without any target metadata we lead Eye2Eye by $1.9\,\text{dB}$. SSIM is the exception, where several baselines score higher, though our LPIPS is roughly half theirs. On SVD the calibrated rig baseline $0.0638\,\text{m}$ and the standard interpupillary baseline $B_{\text{IPD}} = 0.0638\,\text{m}$ that G2 assumes coincide, so the two settings reduce to the same run and we report it in both rows.

Table 2: Quantitative comparison on the StereoGenBench (Cui et al., 2026) Seen split, restricted to video-level methods. **G0** (dense target-view geometry), **G1** (target-camera metadata only), and **G2** (monocular input). Orange rows denote Uniform; blue rows denote IPD-Gaussian. **Bold**: best within each tier–branch group; underline: second best. ‡ReCamMaster is fine-tuned on the StereoGenBench training set. G2 results are available only on the IPD-Gaussian branch.

Method	Base model	PSNR↑	SSIM↑	LPIPS↓	P-PSNR↑	FID↓	FVD↓
G0 · Dense target-view geometry							
StereoCrafter (Zhao et al., 2024)	SVD-XT	21.93	0.658	0.320	22.19	44.66	462.82
SVG (Dai et al., 2025)	Zeroscope v2	23.40	0.732	0.230	23.37	30.08	105.06
M2SVid (Shvetsova et al., 2026)	SVD-XT	27.06	0.812	0.173	**23.68**	15.77	57.72
UniGenStereo (Ours)	Wan2.1-1.3B	**28.40**	**0.853**	**0.081**	**23.68**	**4.14**	**9.03**
StereoCrafter (Zhao et al., 2024)	SVD-XT	22.96	0.717	0.260	24.32	25.13	106.85
SVG (Dai et al., 2025)	Zeroscope v2	25.07	0.775	0.191	26.41	19.36	71.93
M2SVid (Shvetsova et al., 2026)	SVD-XT	29.63	0.876	0.122	**28.07**	6.80	22.91
UniGenStereo (Ours)	Wan2.1-1.3B	**29.87**	**0.885**	**0.078**	**28.07**	**2.59**	**5.31**
G1 · Target-camera metadata							
ReCamMaster‡ (Bai et al., 2025)	Wan2.1-1.3B	19.97	0.622	0.225	23.57	8.62	30.52
UniGenStereo (Ours)	Wan2.1-1.3B	**22.43**	**0.681**	**0.174**	**23.62**	**6.79**	**21.57**
ReCamMaster‡ (Bai et al., 2025)	Wan2.1-1.3B	20.85	0.721	0.183	25.84	6.35	14.24
UniGenStereo (Ours)	Wan2.1-1.3B	**25.37**	**0.816**	**0.111**	**28.44**	**3.17**	**8.66**
G2 · Monocular input, no per-scene target metadata							
StereoCrafter-Zero (Shi et al., 2024b)	DynamiCrafter	11.76	0.398	0.700	13.14	108.01	991.66
StereoPilot (Shen et al., 2025)	Wan2.1-1.3B	13.89	0.468	0.543	15.67	37.82	332.50
StereoWorld (Xing et al., 2026)	Wan2.1-1.3B	16.92	0.619	0.312	25.36	5.31	26.24
ImmersePro (Shi et al., 2024a)	–	17.66	0.603	0.282	23.59	4.58	12.59
UniGenStereo (Ours)	Wan2.1-1.3B	**24.90**	**0.806**	**0.117**	**28.43**	**3.27**	**10.61**

Table 3: Quantitative comparison on SVD, under the evaluation protocol of Metzger et al. (2026). Methods are grouped by conditioning tier as in Table 2. ‡Fine-tuned as in Table 2.

Method	PSNR↑	SSIM↑	LPIPS↓
G0 · Dense target-view geometry			
StereoCrafter (Zhao et al., 2024)	22.5	0.826	0.323
SVG (Dai et al., 2025)	19.3	0.690	0.410
M2SVid (Shvetsova et al., 2026)	24.4	0.821	0.221
Elastic3D (Metzger et al., 2026)	25.9	**0.894**	0.196
UniGenStereo (Ours)	**26.2**	0.782	**0.111**
G1 · Target-camera metadata			
ReCamMaster‡ (Bai et al., 2025)	21.2	0.608	0.201
UniGenStereo (Ours)	**22.5**	**0.654**	**0.178**
G2 · Monocular input, no per-scene target metadata			
Eye2Eye (Geyer et al., 2025)	20.6	**0.733**	0.392
StereoWorld (Xing et al., 2026)	19.9	0.542	0.247
UniGenStereo (Ours)	**22.5**	0.654	**0.178**

4.3 Ablation Studies and Analysis

Conditioning channels. The warped latent carries essentially the entire conditioning signal (Table 4). Zeroing it costs 8.04 dB, while withholding geometry altogether costs 8.08 dB, so the remaining eleven channels are worth 0.04 dB once the warped latent is gone. Removing the coordinate embedding costs 0.19 dB and removing the mask is within noise. The construction anticipates this. Among the three groups only the warped latent depends on the source video; the coordinate embedding is an identity grid warped by the same disparity field and the mask is a binary map of

Table 4: Conditioning-channel ablation under the G0 setting (supplied ground-truth disparity) on the full StereoGenBench evaluation set. One slice of c is zeroed at inference; the last row replaces c entirely with the no-geometry encoding. **Ch.** is the number of conditioning channels that remain non-zero; in that row the single remaining channel is the constant mask, which carries no geometry.

Condition	Ch.	PSNR↑	SSIM↑	LPIPS↓
Full condition	27	**28.93**	**0.8413**	0.0939
w/o disocclusion mask	26	28.92	**0.8413**	**0.0938**
w/o warped coordinates	17	28.74	0.8392	0.0962
w/o warped latent	11	20.89	0.6549	0.2279
No warp (no geometry supplied)	1	20.85	0.6514	0.2226

Table 5: Two-stage training under the model's own predicted geometry (G1). Generation metrics are computed on the full StereoGenBench evaluation set (739 videos) and the full SVD evaluation set (159 videos).

	StereoGenBench				SVD			
Stage	PSNR↑	SSIM↑	LPIPS↓	P-PSNR↑	PSNR↑	SSIM↑	LPIPS↓	P-PSNR↑
Stage 1	22.55	0.703	0.172	**26.94**	21.95	0.637	0.191	30.50
Stage 2	**23.89**	**0.718**	**0.161**	25.97	**22.46**	**0.654**	**0.178**	**30.62**

the pixels that the warp leaves uncovered, so both are determined by D alone. The warped latent in turn makes D observable and marks the disocclusions itself, since uncovered pixels are filled with a constant before encoding. What the generator cannot synthesize on its own is source appearance transported to the target view, and only the warped latent supplies it.

Two-stage training. Stage 2 improves generation on both benchmarks (Table 5): under predicted geometry it gains 1.34 dB in domain and 0.51 dB on out-of-domain real capture, with LPIPS down by more than 6% on each. The mechanism is the differentiable self-warp. Stage 1 trains synthesis only on exact warps, so the backbone comes to depend on their exactness and a predicted warp costs it more; Stage 2 rebuilds the condition from the model's own $\hat{g}$, which exposes the backbone to the imperfect warps it will meet at test time and sends the generation loss back through the splat into the head. The geometry sharpens with it, from 1.582 to 1.000 g MAE in domain and 2.933 to 2.224 on SVD. P-PSNR is the one metric that does not follow, falling in domain while rising on SVD; it searches a bounded horizontal range for the best-matching patch, so it rewards locally plausible structure without requiring the disparity to be correct. Stage 2 adds the external datasets and the self-warp together, so the table measures their combined effect.

5 Conclusion

We presented UniGenStereo, a stereo video generation framework in which a single model with a single set of weights covers three conditioning settings that prior work treats as separate problems. The unification rests on one observation: the settings differ only in where the target-view geometry comes from. An internal head predicts the ratio $g = f/Z$, which is baseline-independent and is the one quantity that turns a requested baseline into disparity, so that supplied disparity, a calibrated baseline and a standard interpupillary baseline all produce the same warp-based conditioning. Training that head jointly with synthesis improves both: the stage that makes the predicted-geometry path deployable also makes the predicted geometry more accurate. Across a synthetic multi-camera benchmark and real-world stereo capture, our unified model attains the best PSNR and LPIPS in every setting, and varying the baseline alone gives continuous control over the synthesized disparity.

Limitations. Stage 2 broadens the scenes and rigs the head sees. When a video still offers few or no geometric cues, however, $\hat{g}$ degrades, a limit of monocular geometry estimation in general, which supplying dense geometry avoids.

AI USE STATEMENT

We used generative AI tools for writing only: to draft sections of this paper and to aid and polish the writing throughout. Every AI-drafted passage was read and edited by the authors, and every passage making a technical claim was verified against our source code and evaluation outputs before inclusion.

We did not use generative AI tools to generate or retrieve any reference.

We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The architecture, the conditioning tensor and the two-stage objective are specified in Sec. 3.2–3.4; optimizer, precision, hardware, training length and inference settings are in Appendix A.1. Evaluation protocols, splits and reference geometry are stated with each table. Code and model weights will be released upon acceptance.

REFERENCES

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025.

Tjark Behrens, Anton Obukhov, Bingxin Ke, Fabio Tosi, Matteo Poggi, and Konrad Schindler. Stereospace: Depth-free synthesis of stereo geometry via end-to-end diffusion in a canonical space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3975–3987, 2026.

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. *arXiv*, 2023. doi: 10.48550/arXiv.2311.15127. URL https://arxiv.org/abs/2311.15127.

D Butler, Jonas Wulff, G Stanley, and M Black. Mpi-sintel optical flow benchmark: Supplemental material. In *MPI-IS-TR-006, MPI for Intelligent Systems (2012)*. Citeseer, 2012.

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

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

Yangzhi Cui, Feng Qiao, and Nathan Jacobs. Stereogenbench: A synthetic multi-camera benchmark for stereo generation under controlled baseline regimes. *arXiv preprint arXiv:2605.23237*, 2026.

Peng Dai, Feitong Tan, Qiangeng Xu, David Futschik, Ruofei Du, Sean Fanello, XIAOJUAN QI, and Yinda Zhang. SVG: 3d stereoscopic video generation via denoising frame matrix. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sx2jXZuhIx.

Neil A. Dodgson. Variation and extrema of human interpupillary distance. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pp. 36–46. SPIE, 2004. doi: 10.1117/12.529999. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5291/1/Variation-and-extrema-of-human-interpupillary-distance/10.1117/12.529999.full.

Zaid A El-Shair, Abdalmalek Abu-raddaha, Aaron Cofield, Hisham Alawneh, Mohamed Aladem, Yazan Hamzeh, and Samir A Rawashdeh. Sid: Stereo image dataset for autonomous driving in adverse conditions. In *NAECON 2024-IEEE National Aerospace and Electronics Conference*, pp. 403–408. IEEE, 2024.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Michal Geyer, Omer Tov, Linyi Jin, Richard Tucker, Inbar Mosseri, Tali Dekel, and Noah Snavely. Eye2eye: A simple approach for monocular-to-stereo video synthesis. *arXiv preprint arXiv:2505.00135*, 2025.

Xianda Guo, Chenming Zhang, Ruilin Wang, Youmin Zhang, Wenzhao Zheng, Matteo Poggi, Hao Zhao, Qin Zou, and Long Chen. Stereocarla: A high-fidelity driving dataset for generalizable stereo. *arXiv preprint arXiv:2509.12683*, 2025.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, volume 2025, pp. 89454–89467, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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*, 2025. doi: 10.48550/arXiv.2506.06023. URL https://arxiv.org/abs/2506.06023.

MohammadHossein Izadimehr, Milad Ghanbari, Guodong Chen, Wei Zhou, Xiaoshuai Hao, Mallesham Dasari, Christian Timmerer, and Hadi Amirpour. Svd: Spatial video dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12988–12994, 2025.

Junpeng Jing, Ye Mao, and Krystian Mikolajczyk. Match-stereo-videos: Bidirectional alignment for consistent dynamic stereo matching. In *European Conference on Computer Vision*, pp. 415–432. Springer, 2024.

Laurent Jospin, Allen Antony, Lian Xu, Hamid Laga, Farid Boussaid, and Mohammed Bennamoun. Active-passive simstereo-benchmarking the cross-generalization capabilities of deep learning-based stereo methods. *Advances in Neural Information Processing Systems*, 35:29235–29247, 2022.

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*, pp. 13229–13239, 2023.

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

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Nando Metzger, Prune Truong, Goutam Bhat, Konrad Schindler, and Federico Tombari. Elastic3d: Controllable stereo video conversion with guided latent decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32693–32703, 2026.

Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5436–5445. IEEE, 2020.

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*, pp. 26579–26589, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.48550/arXiv.2503.03751. URL https://arxiv.org/abs/2503.03751.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22841–22852. IEEE, 2025.

Guibao Shen, Yihua Du, Wenhang Ge, Jing He, Chirui Chang, Donghao Zhou, Zhen Yang, Luozhou Wang, Xin Tao, and Ying-Cong Chen. Stereopilot: Learning unified and efficient stereo conversion via generative priors. *arXiv*, 2025. doi: 10.48550/arXiv.2512.16915. URL https://arxiv.org/abs/2512.16915.

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

Jian Shi, Qian Wang, Zhenyu Li, and Peter Wonka. Stereocrafter-zero: Zero-shot stereo video generation with noisy restart. 2024b.

Nina Shvetsova, Goutam Bhat, Prune Truong, Hilde Kuehne, and Federico Tombari. M2svid: End-to-end inpainting and refinement for monocular-to-stereo video conversion. In *2026 International Conference on 3D Vision (3DV)*, pp. 145–155. IEEE, 2026.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv*, 2025. doi: 10.48550/arXiv.2503.20314. URL https://arxiv.org/abs/2503.20314.

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.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereodiffusion: Training-free stereo image generation using latent diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7416–7425. IEEE, 2024.

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. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.

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)*, pp. 4909–4916. IEEE, 2020.

Xianqi Wang, Hao Yang, Gangwei Xu, Junda Cheng, Min Lin, Yong Deng, Jinliang Zang, Yurui Chen, and Xin Yang. Zerostereo: Zero-shot stereo matching from single images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28177–28187. IEEE, 2025.

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. doi: 10.1109/TIP.2003.819861.

Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *European Conference on Computer Vision*, pp. 722–740. Springer, 2020.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Wen_FoundationStereo_Zero-Shot_Stereo_Matching_CVPR_2025_paper.pdf.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=2z1HT5lw5M.

Ke Xing, Longfei Li, Yuyang Yin, Hanwen Liang, Guixun Luo, Chen Fang, Jue Wang, Konstantinos N Plataniotis, Xiaojie Jin, Yao Zhao, et al. Stereoworld: Geometry-aware monocular-to-stereo video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40245–40255, 2026.

David Yan, Alexander Raistrick, and Jia Deng. What makes good synthetic training data for zero-shot stereo matching?, 2025. URL https://arxiv.org/abs/2504.16930.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2408.06072. URL https://arxiv.org/abs/2408.06072.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv*, 2025a. doi: 10.48550/arXiv.2503.05638. URL https://arxiv.org/abs/2503.05638.

Songsong Yu, Yuxin Chen, Zhongang Qi, Zeke Xie, Yifan Wang, Lijun Wang, Ying Shan, and Huchuan Lu. Mono2stereo: A benchmark and empirical study for stereo conversion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21847–21856. IEEE, 2025b.

Jiale Zhang, Qianxi Jia, Yang Liu, Wei Zhang, Wei Wei, and Xin Tian. Spatialme: Stereo video conversion using depth-warping and blend-inpainting. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209850.

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 (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

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.

A Appendix

A.1 Implementation Details

UniGenStereo is built on the pre-trained Wan2.1-1.3B text-to-video foundation model, with the source and noisy target latents concatenated along the frame axis as network input; the geometry head operates on the fixed $21 \times 60 \times 104$ latent grid. Both stages fine-tune the self-attention layers of all blocks, the input conditioning projections and the geometry head, while cross-attention and feed-forward layers stay frozen. We optimize with AdamW at a constant learning rate of 1×10^{-5}, BF16 mixed precision, gradient clipping at max norm 1.0, and DeepSpeed ZeRO Stage-1 on 4×NVIDIA H100 (80GB) GPUs. Each stage trains for 3 epochs over its data. At test time we run 50 denoising steps with classifier-free guidance scale 5.0, and one checkpoint serves every setting: G0 warps the supplied disparity and never invokes the head, while G1 and G2 run the test-time pass to obtain $\hat{g}$, scale it by the requested baseline, upsample to pixel resolution, and build c through the same warp and re-encoding used for ground-truth geometry.

A.2 Evaluation of Geometry Prediction

We evaluate $\hat{g}$ on its own against Depth Anything V3, converted into g in the two ways a deployed system could obtain it: with the ground-truth focal length, and with the focal length the depth model predicts itself. We report the mean absolute error, the root-mean-square error and the absolute relative error of g; the δ_1 accuracy, that is the fraction of pixels with $\max(\hat{g}/g^*,\, g^*/\hat{g}) < 1.25$; and the disparity error in pixels that $\hat{g}$ produces at the evaluation baseline. All are computed at native scale and without any per-clip scale or shift alignment, so that an error in scale is penalized rather than absorbed by the metric (Table 6). All numbers are from the Stage-2 checkpoint used throughout the paper.

f/Z is directly learnable. In domain the head leads on every metric, roughly halving the g MAE of the better Depth Anything variant and nearly doubling its δ_1. It does so with a 9.5M-parameter head attached to the generator, and against both variants of a far larger dedicated depth model: the one that estimates its own focal length and the one handed the ground truth.

Estimating f and Z separately. On StereoGenBench, Depth Anything V3 is more accurate with its own *predicted* focal length than with the ground-truth one, by a factor of three in g MAE. This is what one expects when the estimated depth is off by a scale factor. Disparity depends only on the ratio f/Z, so errors in f and in Z that point the same way partially divide out; supplying the correct focal length removes that compensation and leaves the error in the depth scale exposed. A model that regresses f/Z in one step has no second quantity whose error could either compensate or compound.

Transfer to unseen real capture. On SVD the ordering changes. Depth Anything V3 leads on g MAE, RMSE and native-scale disparity error, which is what one expects of a model trained on orders of magnitude more real imagery than our external stereo data, and on a rig whose single nominal focal length is a far more stable divisor than StereoGenBench's per-scene focals. Our head is nonetheless second on AbsRel and δ_1, and competitive with the variant a deployed system without camera metadata could actually run: against the predicted-focal Depth Anything V3 it stays within 25% relative on the remaining metrics while nearly doubling δ_1. What separates it from the best variant is around four pixels of disparity at 832×480.

A.3 Controllability of Baseline

The baseline is the only geometric quantity our model accepts in G1, so we show directly what varying it does. Fixing the scene and the source camera, we sweep only the calibrated target camera across the five non-zero cameras of each capture, over 20 scenes from each of the two baseline families; Fig. 3 shows the result on Uniform.

The content is unchanged from column to column — the same objects, the same motion, the same illumination — while every surface shifts horizontally by an amount that grows with the commanded baseline, and the disoccluded strips along depth discontinuities widen with it. Near surfaces move

Table 6: Geometry prediction on 739 StereoGenBench scenes and 159 SVD clips. The predicted $\hat{g}$ is compared against reference $g = f/Z$ at native scale, with no per-clip scale or shift alignment. Errors in g are reported in pixels of disparity per centimeter of baseline at 832×480, i.e. $g/100$ in the meter-based units of Eq. (1); Disp. MAE is that quantity times the baseline in centimeters. StereoGenBench references are dataset disparity divided by the per-sample baseline; SVD references are FoundationStereo (Wen et al., 2025) disparity divided by the 6.38 cm rig baseline, so the reference there is an estimate rather than a measurement. Depth Anything V3 (Lin et al., 2025) is converted to g both with the ground-truth focal length — the renderer's per-scene value on StereoGenBench (276 distinct values over 739 scenes) and the rig focal on SVD (577.1 px at 832×480, identical for all clips) — and with the focal length it predicts itself.

Method	MAE↓ (px/cm)	RMSE↓ (px/cm)	AbsRel↓	δ_1 ↑	Disp. MAE↓
StereoGenBench · in-domain					
DA V3, GT focal	5.793	7.777	4.923	0.096	90.68
DA V3, predicted focal	<u>1.861</u>	<u>2.913</u>	<u>2.862</u>	<u>0.304</u>	<u>42.77</u>
UniGenStereo (Ours)	**1.000**	**1.761**	**0.809**	**0.582**	**18.91**
SVD · unseen, real capture					
DA V3, GT focal	**1.567**	**2.238**	7.570	**0.318**	**10.00**
DA V3, predicted focal	<u>1.795</u>	<u>2.403</u>	**5.468**	0.118	<u>11.45</u>
UniGenStereo (Ours)	2.224	2.682	<u>6.305</u>	<u>0.222</u>	14.19

Table 7: Same-scene baseline controllability on StereoGenBench under predicted geometry (G1). The scene and source camera are fixed and only the calibrated target camera varies; each row aggregates 20 independent five-point per-scene fits over 100 non-zero-baseline pairs, so scenes of differing depth are never pooled into a single trajectory. The effective baseline is recovered per clip by fitting $D_{\mathrm{GT}} \approx a\, D_{\mathrm{pred}} + b$ between the generated and reference disparity fields and taking $B_{\mathrm{eff}} = B/a$; it is a response diagnostic rather than a physical camera calibration. Per scene we then fit $B_{\mathrm{eff}} = s\, B + c$ over its five target cameras and summarize the 20 fits by the mean slope s (ideal 1) and the median coefficient of determination R^2. We use the median because R^2 is bounded above by one, so a single degenerate fit moves the mean far more than it moves the typical behavior.

Family	**Mean slope** ($\to 1$)	**Median** R^2↑
Uniform	0.723	0.959
IPD-Gaussian	1.225	0.965

most and distant surfaces barely at all, which is the behavior Eq. (1) prescribes. This follows from the construction rather than from anything learned per baseline: B enters the model only through $D = B \cdot \hat{g}$, and the predicted field $\hat{g}$ does not depend on it, so a single model produces a continuum of stereo strengths without retraining or re-conditioning.

Within each scene the response to the commanded baseline is linear. Aggregating 20 independent five-point per-scene fits over 100 non-zero-baseline pairs per family, the median scene-level R^2 is 0.965 on IPD-Gaussian and 0.959 on Uniform (Table 7). The gain is less well determined than the linearity: the mean fitted slope is 0.723 on Uniform and 1.225 on IPD-Gaussian. The difference between the two families follows from how they are sampled. Uniform sweeps span baselines from a few centimeters to several meters, so each five-point fit is anchored over a wide range; IPD-Gaussian baselines are clustered around the interpupillary distance, which leaves the slope poorly constrained even when the points themselves lie close to a line. The clustered IPD-Gaussian fits are therefore the less reliable of the two, but the wide-range Uniform sweeps, the better-anchored estimate, still return a slope of 0.723, so a systematic under-gain of roughly 28% cannot be ruled out. Together with a non-zero intercept in the per-scene fit, this means B_{eff} should be read as a response diagnostic rather than a physical camera calibration. Both families nonetheless show the same behavior: the generated disparity is monotonically ordered by, and close to affine in, the commanded baseline, with the gain and offset of that response remaining uncalibrated.

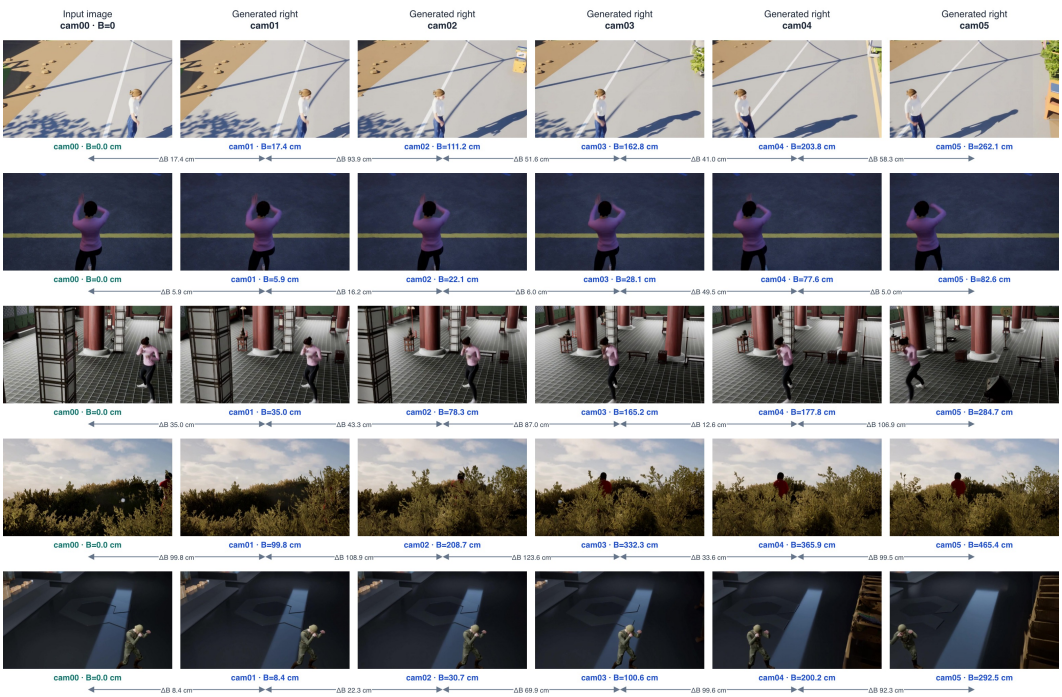


Figure 3: Baseline control on the Uniform split. Each row is one scene: the left-view input followed by the right views generated for the five target cameras of the same capture, with the commanded baseline annotated under each column and increasing from left to right. Everything except B is held fixed, and the same checkpoint produces all five.



Figure 4: Qualitative comparison on StereoGenBench under supplied geometry (G0).

A.4 Qualitative Comparison

Under supplied geometry the methods differ mainly at disocclusions: warp-and-inpaint pipelines leave stretched or duplicated texture along depth boundaries, while our conditioning transports the source appearance directly and leaves the generator only the uncovered pixels to fill (Fig. 4).

A.5 Results on the Unseen Split

Table 8 repeats the comparison on the unseen split, whose scenes share no map with the training set. The ranking is preserved in every tier and branch: PSNR, SSIM, and LPIPS remain best throughout, the sole exception being SSIM on Unseen-Uniform G0, where SVG is marginally higher. Our own degradation relative to the seen split is at most 1.2 dB, and G1-Uniform in fact improves, while

Table 8: Quantitative comparison on the StereoGenBench (Cui et al., 2026) unseen split (102 Uniform / 91 IPD-Gaussian scenes, 81 frames, 832×480), restricted to video-level methods and grouped by the geometric information available at inference time: **G0** (dense target-view geometry), **G1** (target-camera metadata only), and **G2** (monocular input). Orange rows: Unseen-Uniform branch; blue rows: Unseen-IPD-Gaussian branch. **Bold**: best within each tier–branch group; underline: second best. ‡ReCamMaster fine-tuned on the StereoGenBench training set.

Method	Base model	PSNR↑	SSIM↑	LPIPS↓	P-PSNR↑	FID↓	FVD↓
G0 · Dense target-view geometry							
StereoCrafter (Zhao et al., 2024)	SVD-XT	22.06	0.525	0.505	22.93	74.59	879.08
SVG (Dai et al., 2025)	Zeroscope v2	24.48	**0.746**	0.271	24.90	50.25	288.64
M2SVid (Shvetsova et al., 2026)	SVD-XT	27.47	0.733	0.308	**25.31**	47.00	341.21
UniGenStereo (Ours)	Wan2.1-1.3B	**28.02**	0.735	**0.157**	25.18	**11.12**	**111.05**
StereoCrafter (Zhao et al., 2024)	SVD-XT	22.76	0.551	0.408	23.52	45.69	383.28
SVG (Dai et al., 2025)	Zeroscope v2	24.97	0.673	0.296	25.62	44.65	123.63
M2SVid (Shvetsova et al., 2026)	SVD-XT	28.49	0.790	0.189	26.43	18.14	205.35
UniGenStereo (Ours)	Wan2.1-1.3B	**28.83**	**0.796**	**0.109**	**26.48**	**6.66**	**43.67**
G1 · Target-camera metadata							
ReCamMaster‡ (Bai et al., 2025)	Wan2.1-1.3B	22.32	0.579	0.277	**25.09**	17.07	288.13
UniGenStereo (Ours)	Wan2.1-1.3B	**23.83**	**0.603**	**0.251**	24.74	**14.54**	**127.01**
ReCamMaster‡ (Bai et al., 2025)	Wan2.1-1.3B	21.39	0.604	0.240	25.51	16.22	134.22
UniGenStereo (Ours)	Wan2.1-1.3B	**24.15**	**0.676**	**0.168**	**27.38**	**8.66**	**49.59**
G2 · Monocular input, no per-scene target metadata							
StereoCrafter-Zero (Shi et al., 2024b)	DynamiCrafter	12.62	0.344	0.737	13.84	194.39	1250.15
StereoPilot (Shen et al., 2025)	Wan2.1-1.3B	17.00	0.444	0.586	19.02	64.65	443.78
StereoWorld (Xing et al., 2026)	Wan2.1-1.3B	18.32	0.527	0.359	**27.99**	10.26	**32.31**
ImmersePro (Shi et al., 2024a)	–	19.08	0.511	0.370	25.03	9.63	34.27
UniGenStereo (Ours)	Wan2.1-1.3B	**23.88**	**0.667**	**0.175**	27.12	**9.24**	33.86

the distributional metrics keep the widest margins with FID 11.12 against 47.00 and FVD 111.05 against 288.64 on Unseen-Uniform G0. P-PSNR is where the advantage does not hold: M2SVid, ReCamMaster and StereoWorld each lead it by a small amount in one group.