DISPARITY HAS A SIGN: STEREO MATCHING BEYOND THE ZERO-DISPARITY PLANE

Anonymous authors
Paper under double-blind review

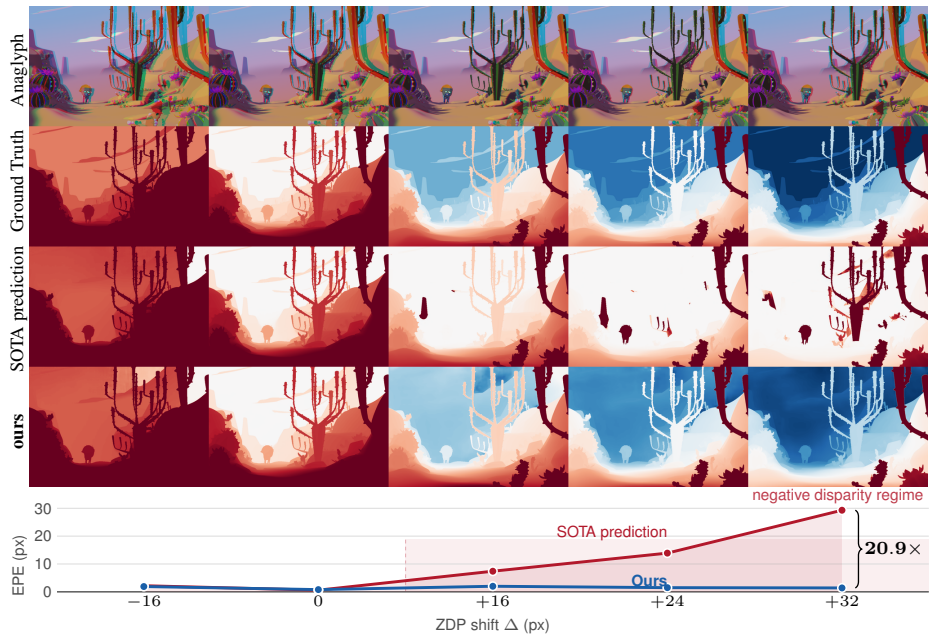

Figure 1: **Moving the zero-disparity plane (ZDP) breaks existing stereo matching models.** One scene rendered at five ZDP shifts Δ. The disparity shifts from positive only (red, in front of the ZDP) to predominantly negative (blue, behind it). We show that SOTA FoundationStereo (the 3rd row) cannot express the negative regime. Our method (the 4th row) restores this capability.

ABSTRACT

Modern stereo matching models fail when disparity crosses zero, with end-point error (EPE) rising by 4.6–37×. Yet stereoscopic content, from cinema 3D to VR, routinely contains objects behind the zero-disparity plane (ZDP), corresponding to negative disparities. The blind spot cascades through datasets, architectures, and evaluation protocols, all of which inherit the non-negative geometry. Rectified parallel cameras conventionally center both principal points, placing ZDP at infinity, so every finite depth yields $d = fB/z \geq 0$ by construction, and nothing within the standard pipeline can violate, or even measure, a negative disparity. To measure it, we propose *ZDPShift*, a benchmark of 21,375 stereo pairs from seven cinematographer-authored open movies, each frame rendered at five zero-disparity-plane positions with dense signed ground truth. Six state-of-the-art image and video stereo matching models collapse once the plane moves. On identical scene content, FoundationStereo goes from 2.15 px EPE to 75.53 px, with every backbone leaving roughly half of all pixels exceeding a three-pixel disparity error. What is missing, however, is not the underlying matching capability. Training on supervision synthesized from SceneFlow, which adds no new data or parameters, keeps the error flat across the signed range. Training only the decoder, with the pretrained matching features frozen, performs comparably across all six backbones, with EPE jittering within 0.2 px. Thus, the pretrained features already extend to the negative regime they were never trained on, and only the

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.

output convention discarded it. Meanwhile, positive-regime accuracy on KITTI, Middlebury, ETH3D, and Sintel is largely preserved.

1 INTRODUCTION

Stereoscopic content is organized around one axis: where each object sits relative to the screen. Cinema 3D, virtual reality, and head-mounted displays all manipulate the zero-disparity plane (ZDP) to place content either in front of the screen ("pop-out") or behind it ("recede" into the scene). Pop-out draws the eye and is used for emphasis and dramatic impact, bringing a subject out into the viewer's space, while a predominantly behind-the-screen placement opens up a deep, immersive volume that the viewer looks into, strengthening the sense of depth and presence (Mendiburu, 2009). Pushed too far in either direction, both cause visual discomfort, so artists must keep depth within a limited viewer comfort budget (Shibata et al., 2011; Lang et al., 2010). Signed disparity is not a corner case of stereoscopic content. It is the design language.

Yet modern stereo matching cannot operate in this language. As shown in Figure 2, on frames from real-world stereo movies, some pixels (can be above 90%) fall behind the screen (the ZDP), and the state-of-the-art models (FoundationStereo and Stereo-Any-Video) cannot represent this content, whereas our signed version recovers it. Notably, ZDP is an authoring choice baked into the in-



Figure 2: **Negative disparity in released 3D films.** Four stereo films (shown as anaglyph), followed by disparity predictions (red: in front of the ZDP; blue: behind). Zero-shot FoundationStereo and Stereo-Any-Video cannot represent behind-the-screen disparities. Our signed model recovers it.

trinsics as a relative principal-point offset that is never recorded in 3D movie datasets. Prior work on web stereo avoids this entirely, computing disparities with optical flow instead (Wang et al., 2019; Xian et al., 2018). Yet, no existing benchmark contains a negative-disparity pixel, so no stereo matching model has ever been measured. We propose the ZDPShift benchmark, a controlled multi-ZDP benchmark drawn from open-movie content. As shown in Figure 3, we benchmark three image (RAFT-Stereo (Lipson et al., 2021), IGEV-Stereo (Xu et al., 2023), FoundationStereo (Wen et al., 2025)) and three video stereo matching backbones (DynamicStereo (Karaev et al., 2023), BiDAStereo (Jing et al., 2024), StereoAnyVideo (Jing et al., 2025)), and find that every backbone fails the moment disparity crosses into the negative regime[1], with EPE increasing by 4.6–$37\times$. FoundationStereo, the SOTA zero-shot model at $\Delta = 0$ (positive disparity only, EPE 2.15 px), produces an EPE of 75.53 px at $\Delta = +32$, which is a $35\times$ degradation on the same scene content. The failure is not a tail-case quirk. It occurs the instant the disparity crosses zero, and the field has no released model for negative disparities, though it has been considered essential for decades in 3D movie production.

[1]We use the stereo-matching sign convention throughout: disparity $d = x_L - x_R$ is positive for scene content nearer than the zero-disparity plane and negative for content beyond it. This is opposite to the display-parallax convention common in stereoscopic production, where behind-the-screen content has positive (uncrossed) parallax and pop-out has negative (crossed) parallax (Mendiburu, 2009). Our negative-disparity regime is the behind-the-screen regime in production terms.

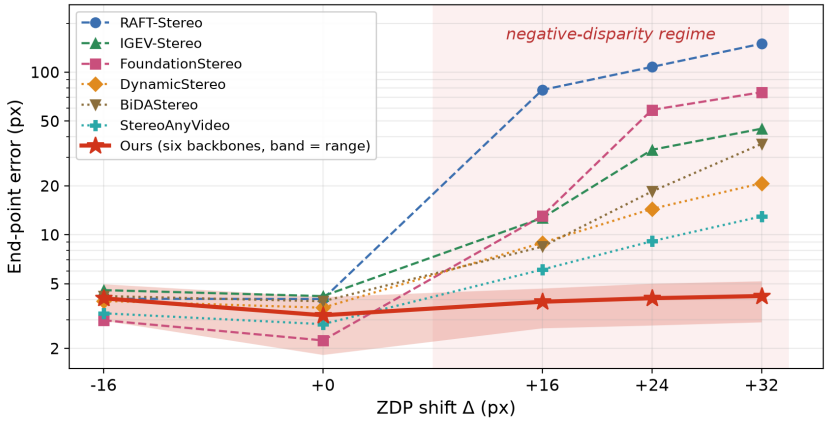

Figure 3: **End-point error (EPE) on varying the ZDP shift Δ.** We evaluate on six image and video stereo matching models and our signed RAFT-Stereo (solid line). Every backbone degrades sharply once Δ grows, reaching 13–150 px EPE at $\Delta = +32$, while all six of the signed versions stay within 2.9–5.2 px across the entire signed range.

Negative disparity is not permitted in the standard stereo matching pipeline. It is enforced through three interlocking mechanisms: *(i) datasets* — KITTI (Menze & Geiger, 2015), Middlebury (Scharstein et al., 2014), ETH3D (Schöps et al., 2017), and SceneFlow (Mayer et al., 2016) are all captured or rendered with rectified parallel rigs whose centered imaging geometry guarantees non-negative disparities, *(ii) architectures* — prevailing matching models encode a positive-disparity prior. This is most explicit in cost-volume methods (Chang & Chen, 2018; Guo et al., 2019; Xu & Zhang, 2020; Wen et al., 2025), whose correlation volume spans only a one-sided hypothesis range $[0, d_{\max}]$, making negative disparity not just unlikely but inexpressible, and *(iii) evaluation* — every standard benchmark measures only the positive regime, so the failure cannot even be observed.

We address these issues in this paper. Our contributions are:

- **Restoring negative disparity, cheaply.** We remove the degradation on all six backbones. Specifically, we synthesize signed supervision by adapting SceneFlow to the full signed disparity range, and extend the one-sided hypothesis range without adding parameters. Freezing the matching features and training only the decoder can also perform reasonably well.
- **Measuring the negative regime.** ZDPShift is the first public stereo benchmark offering multi-ZDP renders, including 21,375 pairs (4,275 open-movie frames $\times$ 5 ZDP shifts) with signed ground truth disparity.
- **Signed models.** We show that many released SOTA stereo matching models can fail with negative disparities, with EPE rising by 4.6–37$\times$. Our resulting signed models work across the full signed range, while preserving positive-regime accuracy on standard benchmarks.

Overview. Section 3 introduces the ZDPShift benchmark and its collection protocol. Section 4 presents the architectural and data techniques that enable negative-disparity estimation. Section 5 gives our results across six image and video backbones. We present discussions and conclusions in Sections 6 and 7, respectively.

2 Related Work

2.1 Stereoscopic perception and ZDP placement

Stereoscopic comfort is bounded on both sides of the zero-disparity plane (Shibata et al., 2011; Lambooij et al., 2009): content too far in front of it and too far behind it is both fatiguing. A scene's depth range is fixed by its content, but where that range falls relative to the screen is not. The ZDP is the offset that maps scene depth onto the display's comfort budget, setting how much of the frame sits in front of the screen and how much behind. Stereoscopic cinematography, therefore, treats its placement as a primary control (Mendiburu, 2009), adjusted per shot alongside disparity remapping (Lang et al., 2010) or applied directly (Shao et al., 2015). Automating that choice requires measuring disparity on both sides of the plane, which no released stereo matching model can do.

2.2 STEREO MATCHING ARCHITECTURES

Modern deep stereo matching models split into two dominant families. Cost-volume aggregation techniques (PSMNet (Chang & Chen, 2018), GwcNet (Guo et al., 2019), AANet (Xu & Zhang, 2020), FoundationStereo (Wen et al., 2025)) mostly build correlation volumes over a one-sided disparity hypothesis range $[0, d_{\max}]$ that halves memory and compute but makes negative disparity inexpressible. Iterative recurrent methods (RAFT-Stereo (Lipson et al., 2021), IGEV-Stereo (Xu et al., 2023)) relax the volume but initialize the disparity field at zero and train exclusively on non-negative corpora. Prior stereo work broadens competence within the non-negative regime. However, ours is orthogonal. We target the regime boundary itself, an axis no prior model has addressed.

A parallel line of work reframes two-view geometry in scene space, such as pointmap regression (DUSt3R (Wang et al., 2024)), its matching-augmented successor MASt3R (Leroy et al., 2024), and feed-forward multi-view transformers (VGGT (Wang et al., 2025)). Those methods are indifferent to the ZDP by construction. However, they operate far from the sub-pixel precision that stereoscopic use requires (EPE$>$ 10px), which we quantify in the supplementary.

2.3 DATASETS

Common datasets (KITTI (Geiger et al., 2012; Menze & Geiger, 2015), Middlebury (Scharstein et al., 2014), ETH3D (Schöps et al., 2017), SceneFlow (Mayer et al., 2016)) use centered rectified parallel rigs, which place the ZDP at infinity, so that negative-disparity pixels cannot occur. In-the-wild datasets such as Holopix50k (Hua et al., 2020) converge the ZDP on mean scene disparity, so they contain behind-screen content but lack dense ground-truth disparity. As Table 1 shows, no prior dataset renders the same scene at multiple ZDP positions with dense ground-truth.

Table 1: ZDPShift versus existing stereo datasets along the axes that matter for the negative-disparity regime: capture/render convention, ZDP coverage, disparity sign, and scale.

Dataset	Indoor	Outdoor	Dense	Annotation	Disp. sign	# images
Sintel (Butler et al., 2012)	✓	✓	✓	Synthetic	pos	1,064
KITTI 2012 (Geiger et al., 2012)	✗	✓	✗	LiDAR	pos	194
Middlebury (Scharstein et al., 2014)	✓	✗	✗	LiDAR	pos	15
KITTI 2015 (Menze & Geiger, 2015)	✗	✓	✗	LiDAR	pos	200
SceneFlow (Mayer et al., 2016)	✓	✓	✓	Synthetic	pos	35,454
ETH3D (Schöps et al., 2017)	✓	✓	✗	LiDAR	pos	27
FallingThings (Tremblay et al., 2018)	✓	✓	✓	Synthetic	pos	61,500
DrivingStereo (Yang et al., 2019)	✗	✓	✗	LiDAR	pos	7,751
Argoverse (Wilson et al., 2023)	✗	✓	✗	LiDAR	pos	5,530
VirtualKITTI2 (Cabon et al., 2020)	✗	✓	✓	Synthetic	pos	21,260
InStereo2K (Bao et al., 2020)	✓	✗	✗	Structured Light	pos	2,010
UnrealStereo4K (Tosi et al., 2021)	✓	✓	✓	Synthetic	pos	8,200
Spring (Mehl et al., 2023)	✗	✓	✓	Synthetic	pos	5,000
Holopix50k (Hua et al., 2020)	✓	✓	✗	None	pos & neg[†]	49,368
ZDPShift (Ours)	✓	✓	✓	Synthetic	pos & neg	21,375

[†] Unlabelled. We measured 30 validation pairs. All of them contains negative disparities.

3 MEASURING THE SIGNED AXIS

We propose ZDPShift, a stereo matching benchmark covering the positive and negative regimes, rendered from open movies. Negative disparity arises chiefly in 3D films, so a benchmark built from film content evaluates the main application. We render 32 scenes drawn from seven Blender Studio open movies (Settlers, Sprite Fright, Spring, Agent 327, Charge, Project Gold, Caminandes Llamigos). Each scene is rendered at five ZDP shifts $\Delta \in \{-16, 0, +16, +24, +32\}$ pixels with a fixed baseline $B = 0.10$ m, at 1920×1280 resolution. The principal points of the left and right cameras are offset by $\mp\Delta/2$ in pixels. The physical camera positions are untouched, and the cameras remain parallel (no toe-in), preserving the rectified-pair assumption every modern model relies on. Per-pixel ground-truth disparity follows from the standard rectified-pair geometry:

$$d(Z, \Delta) \;=\; \frac{fB}{Z} - \Delta, \tag{1}$$

computed analytically from the rendered depth. Each render carries the left/right RGB image and the disparity on different Δ.



Figure 4: **One scene, five zero-disparity planes**. Top $\rightarrow$ bottom: reference view, red–cyan anaglyph, prediction error. Left $\rightarrow$ right: the same scene sweeping from positive (warm color on the 3rd row) at $\Delta = -16$ to predominantly negative (cool color on the 3rd row) at $\Delta = +32$.

3.1 Collection Protocol

A rendered stereo pair is usable as ground truth only if every pixel has a single, sharp, photometrically consistent correspondence in the other view. Cinematic rendering is optimized for a pleasing image, not a measurable one, and breaks this requirement. An artist's depth of field blurs out-of-focus regions, spreading a pixel's true correspondence across a circle of confusion many pixels wide. We use a narrow aperture ($f/64$) for minimal defocus effects. Meanwhile, we bypassed the artist's compositor to disable view-dependent effects, such as lens flares. Thus, rendered images remain sharp across the entire depth range, and the left and right views can stay photometrically consistent. Every other factor that defines the content — lighting, materials, scene composition, camera path, and per-shot focal length — is inherited from the original `.blend` files unchanged. Full settings are provided in the supplementary material.



Figure 5: **ZDPShift scene gallery**, left views. Representative frames span the seven sources, including Caminandes, Spring, Sprite Fright, Settlers, Agent 327, Charge, and Project Gold.

3.2 Dataset Statistics

The corpus comprises 32 scenes and 21,375 stereo pairs (4,275 frames $\times$ 5 ZDPs). All were captured at a resolution of 1920×1280. Table 2 reports how the multi-ZDP rendering populates the signed-disparity axis per source. At $\Delta \leq 0$ the corpus is entirely non-negative — the regime every prior benchmark occupies — while increasing the shift progressively converts the scene: overall 32% of pixels are negative at $\Delta = +16$, rising to 41% at $\Delta = +32$ and present in 78% of frames, with disparities reaching -32 px, all while retaining a wide positive range. The per-source breakdown shows this is a corpus-wide property rather than an artifact of a few scenes. Coverage grows monotonically with the shift for every scene, spanning content from near-screen interiors (Agent 327, essentially all-positive even at $\Delta = +32$) to deep exteriors (Project Gold, 93% negative).

Table 2: Composition of ZDPShift (each frame rendered at five ZDP shifts) and the share of negative-disparity pixels at $\Delta > 0$. Coverage varies with the content, from Agent 327's near-screen interiors to Project Gold's deep exteriors.

			Neg. pixels (%) at Δ		
Source	Scenes	Frames	$+16$	$+24$	$+32$
The Settlers	4	1,283	54	59	63
Sprite Fright	7	821	26	33	38
Spring	5	649	33	38	43
Agent 327	6	590	0	0	1
Charge	5	435	40	49	57
Project Gold	1	226	66	80	93
Caminandes Llamigos	4	271	36	39	44
Total	32	4,275	30	35	40

4 Enabling Negative Disparity Estimation

As discussed, for a complete system, three aspects need to be addressed: one-sided training data, one-sided hypothesis ranges, and evaluation that never crosses zero. ZDPShift addresses the third. This section addresses the other two with a pair of lightweight modifications.

Signed supervision from existing data. Shifting the right image of a rectified pair to the right by Δ pixels moves the ZDP, so relabelling the ground truth as $d' = d - \Delta$ yields a physically consistent pair whose disparities are negative wherever $d < \Delta$. We apply this to SceneFlow (Mayer et al., 2016) with $\Delta \in \{-16, 0, +16, +24, +32\}$ px, drawn uniformly per sample, and keep half of every batch unshifted to preserve positive-disparity competence. For the video models, a single Δ is used per clip, keeping the synthesized plane constant. Two additional datasets, Dynamic Replica (Karaev et al., 2023) and CREStereo (Li et al., 2022), are included for training the video models.

A signed hypothesis range. Unlike IGEV-Stereo and FoundationStereo, enabling negative disparities on RAFT-Stereo and the three video models needs no architectural change. IGEV-Stereo and FoundationStereo build a group-wise correlation (GWC) volume indexed over the non-negative range $[0, d_{\max}/4]$, which is out of bounds once the estimate goes negative. We simply extend the GWC index range to $[-d_{\text{neg}}, d_{\text{pos}}]$ and offset the soft-argmax over the signed bins. The reformulation adds no additional training parameters.

An interesting question is therefore raised: *Does the signed regime demand new correspondences, or only an output convention that admits them?* To answer it, we train a variant with the matching-feature pathway *frozen*, optimizing only the decoder. This will shed light on how much of the negative regime the pretrained correspondences already cover.

Further evaluation. The supplementary reports evaluations of MASt3R and VGGT, manual ZDP re-positioning, training data, detailed frozen ablations, qualitative results, etc. We encourage readers to watch the accompanying videos, where the gain on real production stereo is clearest.

5 Results

5.1 Experiment Setting

All experiments run on a single NVIDIA H200 (160 GB) in PyTorch with FP16 mixed precision. All training uses AdamW with a OneCycle schedule (peak learning rate 2×10^{-4}). We use training resolutions of 320×480 and 320×448 for image and video models, respectively. StereoAnyVideo's DepthAnything branch and FoundationStereo's DINOv2 backbone are kept frozen in all experiments. Image and video models were trained for 30k and 40k steps, respectively. Each run takes approximately 3 h (image) or 10 h (video) on a single H200 GPU. The signed cost volume (for FoundationStereo and IGEV-Stereo) used is $d_{\text{neg}} = 64$ and $d_{\text{pos}} = 192$. Widening the signed volume costs $+6\%$ peak memory and $+1\%$ runtime, with no added parameters.

Table 3: **Quantitative comparisons on ZDPShift**. In addition to *Ours (full training)*, *Ours (frozen)* keeps the pretrained matching features frozen. Zero-shot accuracy collapses as the ZDP shifts into the negative regime, while both of our signed variants stay nearly flat with unchanged temporal consistency. Best and second-best performances are annotated in **red** and **cyan** color, respectively. A light shade is applied to the cells where the negative-disparity regime collapsed significantly.

	EPE ↓ @ Δ					TEPE ↓ @ Δ					Bad3↓ @ Δ	
Model	−16	0	+16	+24	+32	−16	0	+16	+24	+32	0	+32
Original (zero-shot)												
Video Models												
DynamicStereo	3.84	3.46	9.02	14.61	20.90	1.29	1.28	1.59	1.93	2.12	15.9%	45.1%
BiDAStereo	4.12	3.80	8.51	18.69	36.62	1.22	1.21	1.62	2.12	2.79	15.0%	45.8%
StereoAnyVideo	3.15	2.73	6.15	9.17	12.91	1.14	1.10	1.18	1.30	1.44	11.7%	43.9%
Image Models												
RAFT-Stereo	3.98	3.98	78.31	108.41	150.46	1.75	1.48	5.98	6.68	11.08	18.5%	49.4%
IGEV-Stereo	3.85	3.63	12.38	33.11	44.79	1.40	1.37	2.84	5.21	6.40	17.4%	49.0%
FoundationStereo	2.91	2.15	13.03	58.80	75.53	1.10	1.12	2.42	5.61	10.69	9.2%	46.4%
Ours (frozen matching features)												
Video Models												
DynamicStereo	3.98	3.27	3.95	4.15	4.35	1.12	1.10	1.10	1.10	1.13	12.9%	17.4%
BiDAStereo	4.68	4.02	4.30	4.67	4.75	1.10	1.10	1.09	1.10	1.09	12.1%	16.6%
StereoAnyVideo	3.67	2.74	3.56	3.77	4.06	**1.03**	**1.03**	**1.01**	**1.02**	**1.02**	11.3%	16.4%
Image Models												
RAFT-Stereo	5.01	3.96	4.87	4.81	4.78	1.87	1.76	1.71	1.76	1.67	15.6%	20.4%
IGEV-Stereo	4.36	3.53	3.98	3.85	3.85	1.80	1.72	1.55	1.63	1.47	13.1%	18.0%
FoundationStereo	**2.82**	**1.82**	**2.71**	**2.78**	**2.94**	1.13	1.05	1.08	1.09	1.08	**9.0%**	**14.7%**
Ours (full training)												
Video Models												
DynamicStereo	4.16	3.39	3.95	4.30	4.75	1.12	1.11	1.10	1.10	1.12	13.2%	17.7%
BiDAStereo	4.54	3.85	4.16	4.41	4.56	1.11	1.09	1.10	1.11	1.11	12.1%	16.9%
StereoAnyVideo	3.69	2.75	3.39	3.66	3.90	**1.01**	**1.00**	**0.99**	**0.99**	**1.00**	10.7%	15.5%
Image Models												
RAFT-Stereo	4.90	4.06	4.60	4.55	4.57	1.91	1.85	1.83	1.82	1.77	15.1%	19.1%
IGEV-Stereo	4.12	3.07	3.69	3.69	3.73	1.91	1.71	1.56	1.60	1.48	12.1%	17.5%
FoundationStereo	**2.91**	**1.77**	**2.61**	**2.72**	**2.84**	1.15	1.05	1.06	1.09	1.03	**8.0%**	**14.4%**

All models infer at a fixed resolution of 960×640. Predictions are upsampled to 1920×1280 for evaluation. Temporal EPE uses one 16-frame clip per scene.

5.2 Quantitative Results

Every backbone degrades steeply beyond the zero-disparity plane. Wherever the plane leaves every disparity positive ($\Delta \leq 0$), every backbone performs well. At $\Delta = +32$, the same models on the same scenes degrade by around 4–37×, with roughly half of all pixels exceeding a three-pixel disparity error. Notably, moving towards the positive side ($\Delta = -16$) does not break the performance. However, the error increases monotonically as Δ increases for all six backbones. The negative regime is outside what they can express.

The features already support negative disparity. Although our method flattens the signed axis for every backbone, we observe that the frozen variant matches full training on all six backbones, within 0.2 px throughout, and temporal consistency is equally unaffected. This indicates that the matching features can already deal with negative disparity. We use a fixed 800-pair sample of ZDPShift as the evaluation protocol. As shown in Figure 6, models with a frozen matching feature extractor can achieve comparable performance within 500 steps, indicating that the features computed by the pretrained models can

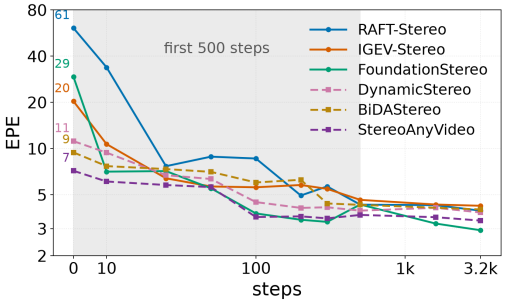

Figure 6: **The negative-disparity capability surfaces almost immediately.** Both axes are logarithmic, and step 0 is the pretrained model.

match content behind the ZDP. Thus, the representation from pretrained models is already sign-agnostic, which transfers to disparities they never saw.

The positive regime is preserved. We evaluate on six standard benchmarks containing no negative disparity: KITTI-2015 (Menze & Geiger, 2015), Middlebury (Scharstein et al., 2014), ETH3D (Schöps et al., 2017) and Sintel (Butler et al., 2012) for the image models, and Sintel and Virtual KITTI 2 (Cabon et al., 2020) for the video models. Image models run at native resolution. Video models are evaluated on 16-frame clips at 512 px. In general, the cost of gaining the signed regime is small and mostly comparable. Set against the gain in the negative regime, the trade is favourable for every backbone.

Table 4: **Positive-regime generalization across standard benchmarks.** We compare the original model (Src) against ours. Ours retains competitive positive-regime accuracy and temporal consistency, with negative capability enabled.

	Sintel				Virtual KITTI 2			
Video Models	EPE↓		TEPE↓		EPE↓		TEPE↓	
	Src	Ours	Src	Ours	Src	Ours	Src	Ours
DynamicStereo	1.95	**1.69**	1.02	**0.98**	1.71	**1.61**	1.54	**1.52**
BiDAStereo	**1.52**	1.72	**0.92**	0.98	1.68	**1.64**	**1.49**	1.52
StereoAnyVideo	**1.47**	1.53	**0.86**	**0.86**	**1.12**	1.29	**1.11**	1.22

	KITTI 2015				Middlebury				ETH3D				Sintel	
Image Models	EPE↓		D1↓		EPE↓		Bad2.0↓		EPE↓		Bad1.0↓		EPE↓	
	Src	Ours	Src	Ours	Src	Ours	Src	Ours	Src	Ours	Src	Ours	Src	Ours
RAFT-Stereo	**1.13**	1.92	**5.26**	10.01	**1.09**	2.49	**8.7**	13.8	**0.24**	0.44	**2.3**	5.6	**1.42**	1.65
IGEV-Stereo	**1.21**	1.79	**5.65**	6.11	2.27	**3.99**	**9.4**	**9.4**	0.29	**0.28**	3.6	**3.3**	15.20	**2.24**
FoundationStereo	**0.86**	1.06	**2.88**	4.61	**0.38**	0.58	**1.1**	2.9	**0.14**	**0.14**	**0.5**	0.7	**0.66**	1.11

Are the negative disparity offsets memorized? Instead of training and inference on a fixed set of Δ, we repeat the experiment with $\Delta{=}{+}32$ withheld, training on the remaining four for 20k steps and keeping $+32$ unseen. As shown in Table 5, models achieve comparable performance on unseen shifts, not simply memorizing training offsets.

Table 5: **Generalization to an unseen shift**, with $\Delta{=}{+}32$ withheld from training. $\text{EPE}_{-16\ldots+24}$ denotes the mean EPE of the predictions for $\Delta \in \{-16, 0, 8, 16, 24\}$.

	Trained on all Δ			$\Delta = +32$ withheld		
Model	$\text{EPE}_{-16\ldots+24}$	EPE_{+32}	TEPE_{+32}	$\text{EPE}_{-16\ldots+24}$	EPE_{+32}	TEPE_{+32}
RAFT-Stereo	4.53	4.57	1.93	4.21	4.31	1.60
IGEV-Stereo	3.64	3.73	1.48	3.89	3.96	1.41
FoundationStereo	2.50	2.84	1.03	2.64	3.05	1.04
DynamicStereo	4.28	5.02	1.14	4.02	4.95	1.11
BiDAStereo	3.92	4.38	1.15	3.90	4.36	1.14
StereoAnyVideo	3.28	3.89	1.03	3.22	3.81	0.99

6 Discussion

Measuring 81 clips (16 frames each) from ten released stereoscopic films with our signed models, 89% of frames contain negative disparity, and 67% contain more than half of the pixels in the negative regime (see supplementary for per-film distribution). The stereo matching field, however, grew up on a more controlled geometry. Conventionally rectified rigs center the principal points, placing the ZDP at infinity and producing only positive disparity. Thus, the convention hardened into datasets, hypothesis spaces, and evaluation protocols, where it became invisible as the standard stereo matching pipeline.

Freezing the matching features comes within a 0.2 px difference of full tuning everywhere. Note that those matching features are never updated, yet they support correspondences of a sign they never trained on. Thus, what these matching features encode is the correspondence, and the sign convention lives in the output layer, not in the representation.

ZDP for visual comfort. The degree of visual discomfort can be predicted from a few factors, such as spatial frequency, disparity response, and visual attention (Shao et al., 2015). During production, artists tend to place ZDPs to keep the whole video stereoscopic and within the viewer's comfort budget (Shibata et al., 2011; Lambooij et al., 2009); because scene depth evolves within and across shots, that placement cannot stay fixed. Comfort budget bounds the parallax angle to about 1° either side of the screen (Lambooij et al., 2009).[2] The budget is itself defined on signed disparity: how far content sits in front of the screen and how far behind. Thus, with models that never output negative disparity, stereo comfort on real films cannot be measured. A signed stereo matching model makes it computable, turning ZDP placement from a judgment made by hand into a quantity that can be measured, scored, and optimized over a sequence.

Stereo generation shares the blind spot. We evaluate the ZDP tolerance on stereo generation models, where the same source frames are synthesized entirely with positive disparities ($t = 0$) or entirely negative ($t = 1$) with StereoCrafter (Zhao et al., 2024). As shown in Figure 7, over ten clips synthesized at five plane positions (50 generations), MEt3R (Asim et al., 2025) degrades from 0.081 at t=0 to 0.091 at t=1 (+12.5%). 9/10 clips are individually worse at t=1. Details can be found in the supplementary material. As behind-screen content is essential to viewing comfort, improving synthesis quality in the negative regime is a necessary next step for stereo video generation.

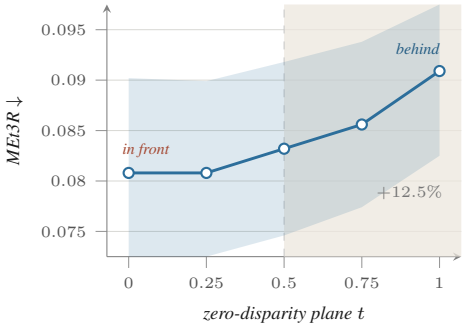

Figure 7: **Generation degrades in negative regimes.** The band is one standard error.

Limitations. First, a signed output space admits a new failure mode. On ambiguous, low-texture content (*e.g.* fog or haze), a stereo matching model can occasionally flip the sign of a region, making the disparity of the low-texture area jitter. This accounts for most of the small residual cost we observe on real stereo films. Second, signed ground truth exists only for rendered content. Yet, the performance on real delivered stereo (Figure 2) can only be validated qualitatively, since released films ship no disparity ground truth. Future work on real-world captured stereo content with negative disparities would accelerate both the benchmarking and its adoption in stereoscopic production tooling.

7 Conclusion

Negative disparity is not an edge case, but the regular operating regime of stereoscopic displays and movies. Surprisingly, no prior released stereo matching model handles it. Not because signed matching is hard, but because two one-sided conventions make it unobservable: 1) training data where $d > 0$ holds by construction, and hypothesis ranges where $d < 0$ are inexpressible. Both lift cheaply. Interestingly, we show that freezing the matching features still recovers most of the gain, showing that the representation could already match across the full signed-disparity range and only lacked a way to express the sign. ZDPShift makes the regime measurable, and we hope it makes signed disparity a standard axis of stereo evaluation.

AI use statement

We used generative AI tools to assist with boilerplate code, namely the standard rendering, training, and evaluation scripts, and with writing scripts to automate experiments. The critical implementation

[2] At the standard viewing distance of three screen heights on a 16:9 display, 1° subtends $3 \cdot \tan(1^\circ)\, W/1.78 \approx 0.029\, W$, i.e. 2.9% of screen width.

originates from the authors, and we used AI tools to tidy the codebase. Additionally, we used generative AI tools to create artifacts (the project page and the demo videos), to style figures, and to polish sentences. We have reviewed all AI-assisted work. We verified the correctness of LLM-generated relevant literature regarding stereoscopic comfort and the relation to ZDP. We reviewed the LLM-generated code and ensured its reproducibility. All drafted text was revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11): 212101, 2020.

Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pp. 611–625. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

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

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5418, 2018.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3273–3282, 2019.

Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, 2020.*, June 2020.

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.

Junpeng Jing, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. Stereo any video: Temporally consistent stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20836–20846, 2025.

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.

Marc Lambooij, Marten Fortuin, Ingrid Heynderickx, and Wijnand IJsselsteijn. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Technology*, 53(3):30201–1–30201–14, May 2009. ISSN 1943-3522. doi: 10.2352/j.imagingsci.technol.2009.53.3.030201. URL http://dx.doi.org/10.2352/J.ImagingSci.Technol.2009.53.3.030201.

Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross. Nonlinear disparity mapping for stereoscopic 3d. *ACM Transactions on Graphics (TOG)*, 29(4):1–10, 2010.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16263–16272, 2022.

Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 218–227, 2021.

Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4040–4048, 2016.

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*, pp. 4981–4991, 2023.

Bernard Mendiburu. *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*. Focal Press, 2009.

Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3061–3070, 2015.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Proceedings of the German Conference on Pattern Recognition (GCPR)*, pp. 31–42, 2014.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3260–3269, 2017.

Feng Shao, Zhutuan Li, Qiuping Jiang, Gangyi Jiang, Mei Yu, and Zongju Peng. Visual discomfort relaxation for stereoscopic 3d images by adjusting zero-disparity plane for projection. *Displays*, 39:125–132, 2015. ISSN 0141-9382. doi: https://doi.org/10.1016/j.displa.2015.10.001. URL https://www.sciencedirect.com/science/article/pii/S0141938215300159.

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

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 2038–2041, 2018.

Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *2019 International Conference on 3D Vision (3DV)*, pp. 348–357. IEEE, 2019.

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*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 311–320. IEEE, 2018.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21919–21928, 2023.

Haofei Xu and Juyong Zhang. AANet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1959–1968, 2020.

Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 899–908, 2019.

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.