FORCESTEREO: OBSERVATION ASSIMILATION THROUGH STATE EVOLUTION FOR STEREO MATCHING

Anonymous authors
Paper under double-blind review

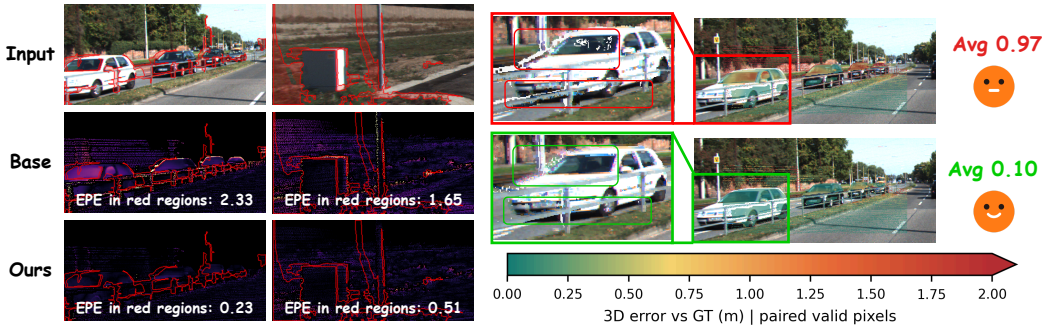


Figure 1: ***Why observation reliability matters for stereo inference?*** Baseline errors concentrate in regions with unreliable observations and further propagate into 3D geometric noise, while ForceStereo substantially reduces these localized failures. This highlights the importance of explicitly regulating the influence of unreliable observations during disparity inference.

ABSTRACT

Modern stereo matching methods achieve strong overall accuracy by combining stereo matching cues, contextual features, and monocular depth priors. However, they rarely assess how reliable different image regions are in open world scenes, so overall metrics and visual results can obscure large errors in regions with unreliable observations. These hidden errors may later surface as noise or artifacts in downstream 3D pipelines. We propose **ForceStereo**, a depth state inference framework based on observation calibration that reformulates disparity estimation as state evolution guided by reliability. ForceStereo first uses scene-level context to calibrate ambiguous local disparity hypotheses before state formation, so the model does not commit too early to unreliable matches. It then interprets residuals, hypothesis ambiguity, and structural information as reliability conditions, and uses a Neural ODE to regulate their influence throughout disparity evolution. This inference process controls how unreliable observations affect the evolving disparity state instead of directly turning them into disparity updates, improving robustness and generalization in open world scenes. Extensive experiments establish state-of-the-art performance, including up to **16.0%** error reduction on the ETH3D leaderboards and up to **25.3%** zero-shot error reduction over prior best results on challenging datasets. ForceStereo **ranks 1st on multiple key metrics in both the KITTI 2012 and KITTI 2015 stereo benchmarks**.

1 INTRODUCTION

Stereo matching estimates dense disparity from rectified image pairs. The resulting disparity provides essential geometric information for 3D perception, scene understanding, and robotic systems. Existing stereo methods commonly use binocular matching evidence and priors from monocular foundation models (Lipson et al., 2021; Xu et al., 2023; Guan et al., 2024; Jiang et al., 2025; Cheng et al., 2025; Guan et al., 2025; Wang et al., 2026). They are often further combined with iterative refinement to improve disparity estimation. However, challenging regions with weak texture, occlu-

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.

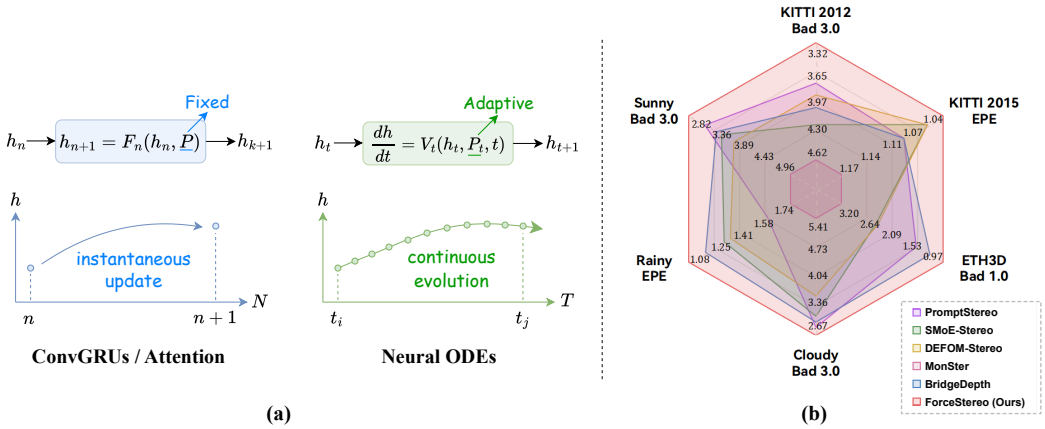


Figure 2: **Conventional disparity updates versus reliability guided state evolution.** (a) Conventional methods predict finite transitions, while ForceStereo integrates reliability conditioned velocities. (b) We achieve lower zero-shot errors on standard benchmarks and adverse driving conditions.

sion, or reflection can still reduce the reliability of these information sources. This motivates us to rethink stereo matching in terms of how observations influence disparity inference.

A key fact is that observation reliability varies across image regions. For example, a large cross-view residual may come from an incorrect disparity, but also from occlusion, reflection, or a true depth boundary. More importantly, this ambiguity changes during inference. Current disparity changes the binocular warping. This also changes the residuals and matching geometry computed afterward. Such observations should therefore be interpreted before they update the disparity state, so misleading evidence is not incorporated too early. To examine this issue, we rank pixels using a fixed, model-independent unreliability score based on cross-view photometric inconsistency and weak local texture. Errors of existing methods rise sharply in less reliable regions. These regions often coincide with weak texture, occlusion, and reflection. Because these regions usually cover only a small part of the image, large errors within them can be overlooked in overall metrics and visual results. Moreover, local disparity errors can propagate into 3D space and produce visible geometric noise and artifacts (see Fig. 1).

These concentrated errors suggest that stereo inference should assess observation reliability before using the observations to update the geometry state. For ambiguous local matches, broader scene context can first calibrate the candidate hypotheses and avoid committing too early to an incorrect disparity. Such calibration cannot be performed only once, because observation reliability changes with the current disparity state. As disparity evolves, cross-view warping changes the residuals and matching geometry, so these observations need to be re-evaluated under the updated state. This motivates treating disparity estimation as a geometry state that evolves continuously under state-dependent observations, rather than as a sequence of separate corrections, as illustrated in Fig. 2(a).

Based on this view, we propose ForceStereo, which models disparity estimation as depth-state evolution continuously guided by observation reliability. Before state formation, the **Overview-Conditioned Hypothesis Volume (OCHV)** uses scene-level context to recalibrate local candidates, avoiding early commitment when evidence is ambiguous. After state initialization, the **Reliability Prompt (RP)** uses residuals, hypothesis ambiguity, and structural cues to reinterpret observations from the current state, while matching-related cues change with the current disparity. **Residual Continuous Evolution (RCE)** then uses these conditions to determine state velocity at each step, rather than directly predicting a finite correction, and obtains the final disparity by integrating a shared Neural ODE vector field. ForceStereo therefore combines hypothesis calibration before state formation with reliability-guided state evolution afterward in a single inference process.

Extensive experiments show that ForceStereo's errors grow more slowly as observations become less reliable. As shown in Fig. 2(b), it also achieves leading performance across standard benchmarks and zero-shot settings. This trend suggests that ForceStereo benefits mainly from better handling unreliable observations during disparity formation, rather than from uniform gains on easy regions.

2 Related Work

Deep Stereo Matching. Deep stereo matching estimates dense disparity from rectified images by aggregating matching evidence or refining predictions. Early methods used 4D cost volumes and 3D regularization (Chang & Chen, 2018; Kendall et al., 2017), followed by multiscale context, groupwise correlation, attention volumes, and adaptive aggregation (Guo et al., 2019; Xu et al., 2022; Xu & Zhang, 2020). RAFT-Stereo (Lipson et al., 2021) corrects disparity through correlation lookup and recurrence, while CREStereo (Li et al., 2022) improves high-resolution matching with cascaded recurrence and adaptive correlation. IGEV (Xu et al., 2023) introduces nonlocal geometry through geometry encoding volumes, while NMRF (Guan et al., 2024) models disparity hypotheses as a neural Markov random field with message passing and proposal pruning. These methods use matching evidence more effectively, while confidence estimation treats reliability as an auxiliary output rather than guidance for state evolution.

Monocular Priors for Stereo Matching. Recent stereo methods use monocular depth priors from vision foundation models to complement local stereo evidence. Depth Anything models are common priors because they provide strong structural representations and generalize well (Yang et al., 2024a;b). Accordingly, DEFOM-Stereo (Jiang et al., 2025) uses depth foundation models to improve feature encoding and recurrent initialization, while MonSter (Cheng et al., 2025) handles scale shifts and ill-posed regions through mutual monocular and stereo refinement. BridgeDepth (Guan et al., 2025) aligns monocular context and stereo hypotheses in latent space with bidirectional cross-attention. PromptStereo (Wang et al., 2026) turns a monocular depth decoder into a prompt-based recurrent unit for iterative disparity refinement. However, existing studies focus on how to incorporate these priors into stereo networks. This leaves an important question: how should a model combine monocular priors and stereo evidence when their reliability differs?

Neural Ordinary Differential Equations. Neural Ordinary Differential Equations (NODEs) (Chen et al., 2018) model continuous hidden-state evolution with a learned vector field and obtain the terminal state through integration. Beyond physical-time modeling, NODEs describe continuous state changes along a general evolution coordinate and have been applied to irregular time series, neural operators, and graph states (Coelho et al., 2024; Chang et al., 2023; Cho et al., 2024; Chen et al., 2024). In stereo matching, the current disparity determines cross-view warping and consequently changes subsequent residuals and matching geometry. This state dependence motivates us to model disparity inference as evolution along a normalized state coordinate rather than as a sequence of separate updates. We therefore use NODEs to parameterize state velocity under reliability conditions, with finite disparity changes obtained through integration. This formulation couples changing observations with geometry evolution without a separate update module.

3 Method

Given a rectified stereo pair $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{3\times H\times W}$, we estimate the dense left-view disparity map $\mathbf{D}_{\text{final}} \in \mathbb{R}^{H\times W}$. Inference proceeds in two stages: hypothesis calibration and state initialization, followed by reliability-guided state evolution. ForceStereo calibrates local disparity hypotheses with scene overview into the overview-conditioned representation $\mathbf{H}^{\text{OCHV}}$. Calibrated stereo hypotheses are then aligned with monocular representations to form the initial geometry state $\mathbf{h}(0)$ and aligned context $\mathbf{C}_{\text{align}}$. After initialization, RP reinterprets observation reliability from the current disparity, while RCE evolves the geometry state under these conditions to obtain the final disparity (see Fig. 3). Further technical details and inherited baseline modules are deferred to the appendix.

3.1 Observation Encoding

ForceStereo encodes three complementary types of observations. They provide local stereo evidence, monocular structural context, and scene level overview, respectively. These representations serve different roles in the subsequent inference process.

Stereo Evidence. Given the left and right images $\mathbf{I}_{l(r)} \in \mathbb{R}^{3\times H\times W}$, a Siamese residual encoder with downsampling layers extracts multiscale stereo features $\mathbf{F}^s_{l(r)} \in \mathbb{R}^{C_s\times H/s\times W/s}$, where $s \in \{4, 8\}$. These features preserve local binocular correspondence evidence and generate the sparse disparity proposals used for subsequent hypothesis calibration.

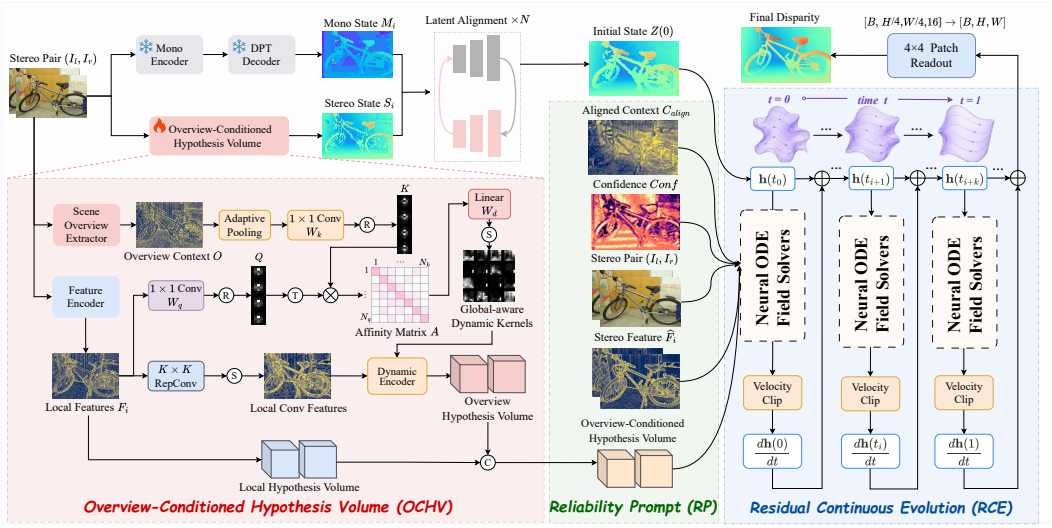


Figure 3: **Overall framework of ForceStereo.** ForceStereo calibrates local disparity hypotheses to initialize the geometry state, followed by reliability guided state evolution with RP and RCE.

Monocular Structural Context. We employ a frozen Depth Anything V2 (Yang et al., 2024b) to extract monocular structural representations from the reference image $\mathbf{I}_l$. The features produced by its DINOv2 encoder and DPT decoder are projected to $\mathbf{M} \in \mathbb{R}^{C_m \times H/8 \times W/8}$ and serve as structural context for the subsequent monocular stereo latent alignment.

Scene Overview. Local stereo evidence captures limited spatial context, so we extract a scene level overview representation from the reference image. A lightweight downsampling CNN and adaptive pooling encode it as $\mathbf{O} \in \mathbb{R}^{C_o \times K_o \times K_o}$, where K_o denotes the overview grid size. This overview representation provides scene context for local stereo features and supports disparity hypothesis calibration in the next section.

3.2 Hypothesis Calibration with Scene Overview

In ambiguous regions, local stereo evidence often provides similar matching support for multiple disparity candidates. We represent this hypothesis competition with a compact set of local proposals and use it for subsequent scene calibration. Specifically, we use DPN (Guan et al., 2024) to retain K local disparity proposals per pixel from dense disparity candidates. At scale s, we denote this support as $\mathbf{P} = \{d_{i,j}^k \mid (i,j) \in \Omega_s,\ k = 1, \ldots, K\}$, with $K = 2$ by default. Here, Ω_s denotes the stereo feature grid, and $d_{i,j}^k$ is the kth disparity proposal at pixel (i,j). These retained proposals form the hypothesis space for subsequent calibration, but are still derived mainly from local stereo evidence. Proposal selection alone therefore does not resolve ambiguity in local matching.

We construct local and overview conditioned hypothesis representations on the same proposal set $\mathbf{P}$. Local Hypothesis Volume (LHV) preserves local matching evidence, while Overview Hypothesis Volume (OHV) contextualizes the proposals with scene overview. As shown in Fig. 4, they are integrated into OCHV to form overview calibrated hypotheses. To introduce scene context, we first associate each local stereo position with the overview representation. For each view, the stereo feature and scene overview are projected into local queries and overview anchors, respectively. The resulting association generates dynamic kernels with two receptive fields:

$$\begin{gathered} \mathbf{Q}_v = \mathrm{Re}(W_q \mathbf{F}_v), \quad \mathbf{K}_o = \mathrm{Re}(W_k \mathbf{O}), \quad \mathbf{A}_v = \mathrm{Softmax}\left(\mathbf{Q}_v^\top \mathbf{K}_o\right), \\ (\mathbf{W}_v^{\mathrm{S}}, \mathbf{W}_v^{\mathrm{L}}) = \mathrm{Split}\left(\mathrm{FFN}(\mathbf{A}_v)\right), \quad v \in \{l, r\}. \end{gathered} \tag{1}$$

Here, W_q and W_k are 1×1 convolutions, and $\mathrm{Re}(\cdot)$ reshapes the projected features into token matrices for affinity computation. The affinity matrix $\mathbf{A}_v$ describes overview conditioned associations between local stereo representations and overview anchors. Split maps this association to a small kernel $\mathbf{W}_v^{\mathrm{S}}$ and a large kernel $\mathbf{W}_v^{\mathrm{L}}$. The two kernels capture fine endpoint information and

broader neighborhood context, respectively. Following previous dynamic encoding designs (Lou & Yu, 2025), we transform the local stereo features using these overview conditioned kernels:

$$\mathbf{F}_v^{\mathrm{ov}} = \mathrm{Encoder}\left(\mathbf{F}_v, \mathbf{W}_v^{\mathrm{S}}, \mathbf{W}_v^{\mathrm{L}}\right), \quad v \in \{l, r\}. \tag{2}$$

By integrating dynamic kernel responses with complementary receptive fields, the encoder produces $\mathbf{F}_v^{\mathrm{ov}}$ as stereo endpoint evidence conditioned on scene context. We next construct OHV and LHV at the same disparity proposal locations. OHV measures candidate support from the overview conditioned features, while LHV preserves the support provided by original local stereo features. Together with the disparity encoding, they form the final overview conditioned hypothesis representation:

$$\begin{aligned}
\mathbf{H}_{i,j,k}^{\mathrm{OHV}}(g) &= \frac{G}{C}\left\langle \mathbf{F}_l^{\mathrm{ov}}(i,j,g), \mathbf{F}_r^{\mathrm{ov}}(i,j-d_{i,j}^k,g)\right\rangle, \\
\mathbf{H}_{i,j,k}^{\mathrm{LHV}}(g) &= \frac{G}{C}\left\langle \mathbf{F}_l^{\mathrm{loc}}(i,j,g), \mathbf{F}_r^{\mathrm{loc}}(i,j-d_{i,j}^k,g)\right\rangle, \\
\mathbf{H}_{i,j,k}^{\mathrm{OCHV}} &= \mathrm{FFN}\left(\mathbf{H}_{i,j,k}^{\mathrm{OHV}} \parallel \mathbf{H}_{i,j,k}^{\mathrm{LHV}} \parallel \phi(d_{i,j}^k)\right),
\end{aligned} \tag{3}$$

where g indexes a channel group, G is the number of groups, and C is the channel dimension. $\mathbf{F}_v^{\mathrm{loc}}$ denotes the local stereo feature from Sec. 3.1, and $\phi(\cdot)$ denotes disparity positional encoding. Therefore, $\mathbf{H}^{\mathrm{OCHV}}$ preserves local matching support while adding scene conditioned evidence for each retained hypothesis. The calibrated stereo hypotheses are aligned with the monocular structural representation. Specifically, the stereo state $\mathbf{S}_0$ queried from OCHV and the monocular state $\mathbf{M}_0$ enter a cross attention latent alignment unit (Guan et al., 2025). The alignment forms the initial geometry state $\mathbf{h}(0)$ and aligned context $\mathbf{C}_{\mathrm{align}}$ for subsequent reliability modeling and state evolution, with aligned monocular states yielding $\mathbf{m}_i^{\mathrm{mono}}$ for affine-invariant supervision.

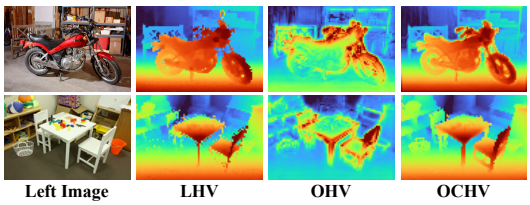


Figure 4: **Hypothesis calibration with OCHV.** Scene overview provides contextual evidence for locally supported disparity hypotheses.

3.3 State Conditioned Observation Interpretation

During state evolution, the same stereo observations can produce different cross-view residuals under different geometry states. RP reinterprets these observations at each evolution coordinate and produces reliability conditions $\mathbf{P}_t$ for subsequent evolution. State-induced residuals are therefore interpreted using hypothesis ambiguity and structural context before guiding evolution.

State Induced Residuals. Given the current geometry state $\mathbf{h}(t)$, we read out the current disparity $\mathbf{D}_t = \mathcal{R}(\mathbf{h}(t))$ and reconstruct the cross-view observations. Specifically, we warp the right image and feature to the left view and compute the corresponding image and feature residuals:

$$\begin{aligned}
\mathbf{I}_{r\to l}^t &= \mathrm{Warp}(\mathbf{I}_r, \mathbf{D}_t), \quad \mathbf{F}_{r\to l}^t = \mathrm{Warp}(\mathbf{F}_r, \mathbf{D}_t), \\
\mathbf{R}_t^{\mathrm{img}} &= \left|\mathbf{I}_l - \mathbf{I}_{r\to l}^t\right|, \quad \mathbf{R}_t^{\mathrm{feat}} = \left|\mathbf{F}_l - \mathbf{F}_{r\to l}^t\right|,
\end{aligned} \tag{4}$$

where Warp denotes bilinear warping guided by the current disparity. The image and feature residuals describe photometric and feature consistency under the current state, respectively. We concatenate them into $\mathbf{R}_t$ and further decompose it into low and high frequency components:

$$\mathbf{R}_t = \mathbf{R}_t^{\mathrm{img}} \parallel \mathbf{R}_t^{\mathrm{feat}}, \qquad \mathbf{R}_t^{\mathrm{low}} = \mathrm{AvgPool}_{5\times5}(\mathbf{R}_t), \qquad \mathbf{R}_t^{\mathrm{high}} = \mathbf{R}_t - \mathbf{R}_t^{\mathrm{low}}. \tag{5}$$

The low-frequency component captures broad state deviations, while the high-frequency component highlights boundaries, fine structures, and local changes.

Residual Interpretation Signals. Residuals alone cannot determine how current observations should affect the geometry state. We therefore interpret the state-induced residuals with hypothesis competition and scene structure to better identify their likely causes. After latent alignment, the aligned OCHV representation yields a pixelwise candidate distribution $\mathbf{p}$ over the K disparity hypotheses (details in appendix). We define the hypothesis ambiguity $\mathbf{U}$ using the normalized candidate entropy and the probability gap between the top two hypotheses:

$$\mathbf{U}(\mathbf{p}) = \frac{1}{2}\left[\mathcal{H}_K(\mathbf{p}) + 1 - \Delta_2(\mathbf{p})\right], \tag{6}$$

where $\mathcal{H}_K$ measures dispersion over the K candidates, while Δ_2 is the top-two probability gap. A larger $\mathbf{U}$ therefore indicates stronger hypothesis ambiguity and correspondingly greater adjustment freedom. Hypothesis ambiguity reflects adjustment demand rather than observation trust, while fixed image edges and state dependent matching geometry explain residual origins. The edge map $\mathbf{E} = \|\nabla \mathrm{Gray}(\mathbf{I}_l)\|$ captures local structural variations in the reference image. The dynamic geometry cue $\mathbf{G}_t = \delta_g(\mathbf{F}_l \parallel \mathbf{F}^t_{r\to l})$ captures the current correspondence between left and warped right features. Here, Gray denotes grayscale conversion and δ_g a lightweight convolutional geometry mapping.

Reliability Condition Construction. Before state evolution, we interpret the current residuals using hypothesis ambiguity, edge structure, and the aligned context $\mathbf{C}_{\mathrm{align}}$. The high- and low-frequency residuals are conditioned on $\mathbf{U}$ and $\mathbf{E}$ and adaptively fused with the aligned context and current matching geometry to form the reliability condition $\mathbf{P}_t$:

$$\begin{aligned} &\mathbf{B}^{\rho}_t = \delta_r\left(\mathbf{R}^{\rho}_t \parallel \mathbf{U} \parallel \mathbf{E}\right), \qquad \mathbf{\Omega}_t = \sigma\left(\delta_\omega\left(\mathbf{B}^{\mathrm{low}}_t \parallel \mathbf{B}^{\mathrm{high}}_t \parallel \mathbf{U} \parallel \mathbf{E}\right)\right), \\ &\mathbf{P}_t = \mathbf{C}_{\mathrm{align}} \parallel \mathbf{G}_t \parallel \left[\mathbf{\Omega}_t \odot \mathbf{B}^{\mathrm{high}}_t + (1-\mathbf{\Omega}_t)\odot \mathbf{B}^{\mathrm{low}}_t\right], \qquad \rho \in \{\mathrm{low}, \mathrm{high}\}. \end{aligned} \tag{7}$$

Here, δ_r and δ_ω perform residual interpretation and adaptive gating, while $\mathbf{\Omega}_t$ balances the two residual representations. Thus, $\mathbf{P}_t$ retains aligned structural context while combining interpreted residuals with current matching geometry. Since $\mathbf{R}_t$ and $\mathbf{G}_t$ are recomputed from $\mathbf{D}_t = \mathcal{R}(\mathbf{h}(t))$, the state-dependent components of $\mathbf{P}_t$ evolve with geometry and modulate RCE velocity.

3.4 Residual Continuous Evolution

RP is regenerated from current disparity, so the observation condition changes with the state along the trajectory. Inspired by continuous state modeling in Neural ODEs (Chen et al., 2018), we model refinement with a shared coordinate-conditioned velocity field under current reliability and obtain finite state changes through integration, rather than directly predicting a finite ConvGRU transition:

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{V}(\mathbf{h}(t), \mathbf{P}(t), t), \quad \mathbf{h}(1) = \Phi_{1\leftarrow 0}(\mathbf{h}(0)), \quad \Phi_{1\leftarrow 0} = \Phi_{1\leftarrow \tau}\circ\Phi_{\tau\leftarrow 0}, \quad \tau \in (0,1). \tag{8}$$

where $\Phi_{b\leftarrow a}$ denotes the mapping obtained by evolving from coordinate a to b along the shared vector field. Thus, the trajectory can be decomposed at any intermediate coordinate τ into two consecutive flows. This exposes intermediate geometry states, allowing RP to be regenerated before further integration under updated reliability conditions. In practice, we approximate the solution flow with three fixed-step Euler evaluations, regenerating RP at each intermediate state:

$$\mathbf{h}_{m+1} = \mathbf{h}_m + \int_{t_m}^{t_{m+1}} \mathbf{V}(\mathbf{h}(t), \mathbf{P}(t), t)\ \mathrm{d}t \approx \mathbf{h}_m + \Delta t\, \mathbf{V}(\mathbf{h}_m, \mathbf{P}_m, t_m). \tag{9}$$

where m indexes solver evaluations, with $t_m = m/M$ and $\Delta t = 1/M$. M controls solver evaluations rather than model depth, while the shared coordinate-conditioned vector field defines continuous state evolution. Each intermediate state yields $\mathbf{D}_m = \mathcal{R}(\mathbf{h}_m)$, which is used to regenerate $\mathbf{P}_m$ before the shared field re-evaluates state velocity. Thus, RCE parameterizes local state derivatives conditioned on reliability, while finite state changes are obtained only by integrating these velocities along the trajectory. To suppress unnecessary state drift in low-ambiguity regions, we use hypothesis ambiguity $\mathbf{U}$ to impose an explicit spatially adaptive bound on pixelwise state velocity:

$$(\mathbf{Z}^g_m, \mathbf{Z}^v_m) = \mathrm{Split}[\mathit{ResBlock}(\mathbf{h}_m \| \mathbf{P}_m, t_m)], \quad \mathbf{V}_m = \mathbf{U}\odot\sigma(\mathbf{Z}^g_m)\odot\tanh(\mathbf{Z}^v_m). \tag{10}$$

Here, $\mathbf{U}$ sets the local velocity bound, $\sigma(\mathbf{Z}^g_m)$ controls its activation strength, and $\tanh(\mathbf{Z}^v_m)$ independently determines the signed direction of evolution. Thus, $\mathbf{U}$ controls how much the state can change rather than observation reliability; low-ambiguity regions are therefore more constrained, while RP and the current state determine the local state change. $\mathit{ResBlock}$ uses three residual blocks with time embeddings and cross-fusion with RP to parameterize the local velocity field. Finally, we apply the same readout to the terminal state and use subpixel rearrangement to obtain $\mathbf{D}_{\mathrm{final}}$.

3.5 Loss Function

We use standard disparity sequence supervision for stereo matching and affine-invariant monocular supervision (Ranftl et al., 2022), together with ambiguity-aware supervision on the RCE trajectory:

$$\begin{aligned} &\mathcal{L}_{\mathrm{stereo}} = \textstyle\sum_i w_i \left\|\mathbf{d}^{\mathrm{st}}_i - \mathbf{d}_{\mathrm{gt}}\right\|_1, \qquad \mathcal{L}_{\mathrm{mono}} = \textstyle\sum_i \gamma_i \rho\left(\mathbf{m}^{\mathrm{mono}}_i, \mathbf{d}_{\mathrm{gt}}\right), \\ &\mathcal{L}_{\mathrm{evo}} = \textstyle\sum_{m=1}^{M} \beta_m \left\|(1+\mathrm{sg}(\mathbf{U}))\odot\left(\mathbf{d}^{\mathrm{RCE}}_m - \mathbf{d}_{\mathrm{gt}}\right)\right\|_1. \end{aligned} \tag{11}$$

Table 1: **Quantitative evaluation on the Scene Flow test set**, and the **best** result is bolded.

Method	RAFT-Stereo	NMRF	DEFOM-Stereo	MoCha-Stereo	MonSter	BridgeDepth	ForceStereo (Ours)
EPE	0.56	0.45	0.42	0.41	0.37	0.37	**0.34** (−8.11%)
Bad 1.0	6.63	4.50	5.13	5.07	4.16	3.67	**3.38** (−7.90%)

Table 2: **Quantitative results on four stereo matching benchmarks.** Results are from official leaderboards. All metrics are percentages except RMSE (px); "All" and "Noc" denote all valid and non-occluded pixels, respectively. The best, second-best, and third-best results are color-coded.

Method	KITTI 2012 Reflective				KITTI 2012				KITTI 2015				ETH3D			
	Out-2 Noc	Out-2 All	Out-3 Noc	Out-3 All	Out-2 Noc	Out-2 All	Out-3 Noc	Out-3 All	D1-bg Noc	D1-all Noc	D1-bg All	D1-all All	Bad 1.0 Noc	Bad 1.0 All	RMSE Noc	RMSE All
GANet (2019)	10.75	12.94	6.22	7.92	1.89	2.50	1.19	1.60	1.34	1.63	1.48	1.81	6.42	6.95	0.69	0.92
ACVNet (2022)	11.42	13.53	7.03	8.67	1.83	2.35	1.13	1.47	1.26	1.52	1.37	1.65	2.58	2.86	0.45	0.54
RAFT-Stereo (2021)	–	–	–	–	1.92	2.42	1.30	1.66	1.44	1.69	1.58	1.82	2.44	2.60	0.36	0.42
IGEV-Stereo (2023)	7.57	8.80	4.35	5.00	1.71	2.17	1.12	1.44	1.27	1.49	1.38	1.59	1.12	1.51	0.34	0.36
MoCha-Stereo (2024)	6.97	8.10	3.83	4.50	1.64	2.07	1.06	1.36	1.24	1.44	1.36	1.53	1.41	1.83	0.36	0.74
NMRF (2024)	10.12	12.34	6.35	8.11	1.59	2.07	1.01	1.35	1.18	1.46	1.28	1.57	–	–	–	–
MonSter (2025)	5.66	6.81	2.75	3.38	1.36	1.75	0.84	1.09	1.05	1.33	1.13	1.41	0.46	0.72	0.20	0.47
DEFOM-Stereo (2025)	5.76	6.72	3.04	3.56	1.43	1.79	0.94	1.18	1.15	1.33	1.25	1.41	0.70	0.78	0.22	0.26
BridgeDepth (2025)	5.80	6.85	2.91	3.48	1.32	1.65	0.83	1.03	1.05	1.31	1.13	1.40	0.38	0.50	**0.19**	0.26
PromptStereo (2026)	**5.11**	6.17	2.54	3.08	1.32	1.69	0.84	1.09	1.04	1.32	1.12	1.41	–	–	–	–
ForceStereo (Ours)	5.16	**5.90**	**2.49**	**2.92**	**1.24**	**1.56**	**0.77**	**0.99**	**1.03**	**1.28**	**1.11**	**1.35**	**0.32**	**0.42**	**0.19**	**0.24**

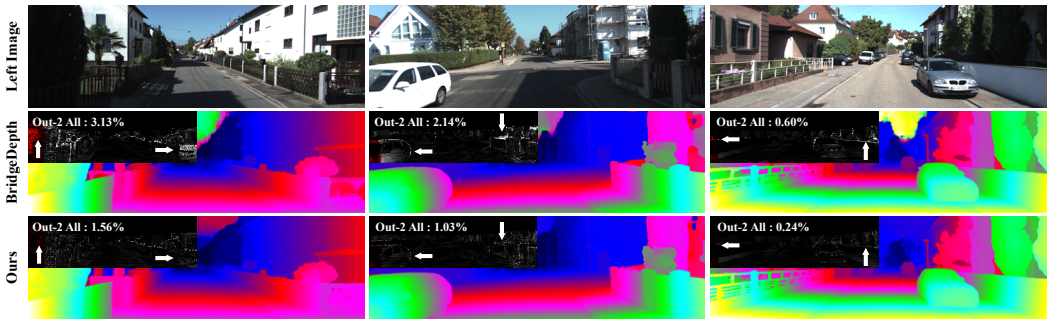


Figure 5: **Qualitative comparison on KITTI.** The inset of each result shows the error map and Out-2 All score. Although the disparity maps are visually similar, the error maps reveal that ForceStereo better suppresses residual errors in locally unreliable regions.

Here, i/m index stereo predictions/RCE states, ρ is affine-invariant loss, sg stop-gradient, and β_m trajectory weights. The final objective is $\mathcal{L} = \mathcal{L}_{\text{stereo}} + \lambda_{\text{mono}}\mathcal{L}_{\text{mono}} + \lambda_{\text{evo}}\mathcal{L}_{\text{evo}}$.

4 Experiments

We implement ForceStereo in PyTorch and train on NVIDIA L20 GPUs with AdamW, following RAFT-Stereo (Lipson et al., 2021) for augmentation. We train for 300K steps with batch size 8 using a one-cycle schedule peaking at 5×10^{-4}. Recent monocular-prior methods largely use DAv2; controlled comparisons use the same frozen DAv2 ViT-L (Yang et al., 2024b), while stereo-only methods remain unchanged and the stereo branch uses *BasicEncoder* from RAFT (Teed & Deng, 2020). Images are randomly cropped to 368×784. Most experiments use Scene Flow (Mayer et al., 2016) pretraining; for ETH3D and KITTI fine-tuning, we follow (Guan et al., 2025), pretraining on Scene Flow, VKITTI2 (Cabon et al., 2020), CREStereo (Li et al., 2022), TartanAir (Wang et al., 2020), Sintel Stereo (Butler et al., 2012), and FallingThings (Tremblay et al., 2018).

4.1 Benchmark Performance and Zero-Shot Generalization

Benchmark Performance. We compare ForceStereo with recent methods on Scene Flow (Mayer et al., 2016), KITTI 2012 (Geiger et al., 2012), KITTI 2015 (Menze & Geiger, 2015), and ETH3D (Schops et al., 2017). Tabs. 1 and 2 show leading results across multiple benchmarks.

Table 3: **Zero-shot generalization with Scene Flow training.** Adverse-condition results use DrivingStereo weather subsets. Colors mark the best, second-best, and third-best results.

Method	Standard Real-world Benchmarks								Adverse Driving Conditions Yang et al. (2019)							
	KITTI-12		KITTI-15		Middlebury		ETH3D		Cloudy		Foggy		Rainy		Sunny	
	EPE	BP-3	EPE	BP-3	EPE	BP-2	EPE	BP-1	EPE	BP-3	EPE	BP-3	EPE	BP-3	EPE	BP-3
RAFT-Stereo (2021)	0.90	4.34	1.13	5.68	1.40	11.07	0.27	2.61	0.97	3.58	0.95	2.63	1.80	11.01	1.01	3.97
IGEV-Stereo (2023)	1.03	5.13	1.21	6.03	1.48	9.95	0.33	4.05	1.11	4.84	1.12	4.58	2.34	14.87	1.18	5.05
DEFOM-Stereo (2025)	0.83	3.90	1.06	4.99	0.94	6.77	0.35	2.40	0.98	3.59	1.01	3.51	1.35	11.04	0.98	3.77
MonSter (2025)	0.93	4.62	1.17	5.52	1.04	8.97	0.35	3.20	1.10	5.41	1.38	9.44	1.74	17.59	1.12	4.96
BridgeDepth (2025)	0.83	4.04	1.09	4.69	0.92	7.84	0.23	1.26	0.94	2.98	1.07	4.34	1.19	6.40	0.99	3.40
MGStereo (2025)	0.86	4.14	1.12	5.64	1.12	8.37	0.25	1.88	0.92	3.36	0.94	2.85	1.29	9.39	0.94	3.70
SMoE-Stereo (2025)	0.83	4.23	1.06	4.94	1.54	7.38	0.27	2.44	0.93	3.12	1.11	4.77	1.31	6.08	1.00	3.51
PromptStereo (2026)	0.79	3.77	1.09	4.59	0.96	**6.03**	0.23	1.56	0.92	2.86	**0.89**	**2.43**	1.59	14.30	0.94	3.16
ForceStereo (Ours)	**0.74**	**3.32**	**1.04**	**4.43**	**0.87**	7.62	**0.21**	**0.97**	**0.90**	**2.67**	1.04	4.32	**1.08**	**4.54**	**0.93**	**2.82**

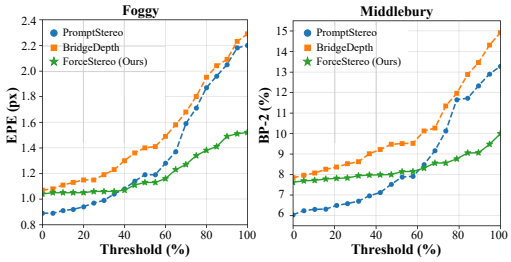


Figure 6: **Threshold-wise error comparison on Foggy and Middlebury** across model-independent unreliability levels.

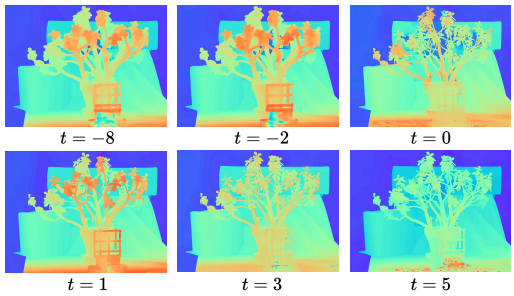


Figure 7: **RCE integration-range analysis.** Trained on $t \in [0, 1]$, states evolve coherently nearby but degrade under distant extrapolation.

Scene Flow. As shown in Tab. 1, compared with our baseline, ForceStereo achieves improvements of 8.11% and 7.90% on the two reported metrics, respectively.

KITTI. As shown in Tab. 2, ForceStereo consistently achieves the best results on almost all KITTI 2012/2015 metrics, setting new bests on all four KITTI 2012 standard metrics with 3.88%–7.23% lower errors than previous bests. On KITTI 2015, it reduces D1-all All to 1.35 as the only method below 1.40, while remaining particularly strong on the ill-posed reflective subset of KITTI 2012. Error maps in Fig. 5 further show fewer residual errors in locally unreliable regions.

ETH3D. As shown in Tab. 2, ForceStereo reduces ETH3D Bad 1.0 Noc/All errors by 15.79% and 16.00% over BridgeDepth. For RMSE, it retains the best Noc result and reduces All by 7.69%.

Zero-Shot Generalization Analysis. Tab. 3 compares zero-shot generalization after Scene Flow-only stereo training, with recent monocular-prior methods largely using DAv2. ForceStereo improves prior best results by up to 25.3%, with gains varying across benchmarks. Beyond overall metrics, Fig. 6 evaluates increasingly unreliable regions using a model-independent score based on GT-aligned photometric inconsistency and texture weakness, defined in the supplement. Errors increase for all methods with unreliability, but ForceStereo's errors grow more slowly. This suggests reliability-guided state evolution mainly improves zero-shot robustness in unreliable regions.

4.2 Ablation Study

With OCHV analyzed earlier in Fig. 4, we further ablate the remaining framework components. Unless otherwise specified, all ablation results are averaged over four independent random seeds.

Module Effectiveness. Tab. 4 shows that OCHV reduces hypothesis ambiguity, while RP and RCE enable reliability-guided observation integration. Together, they improve all benchmarks with only about 0.01 s overhead, and removal tests further confirm their complementary roles.

RP Component Analysis. Tab. 5 ablates RP cues and residual organization, with controls removing U only from RP, using unsplit residuals, or fixing $\Omega_t = 0.5$. Full RP performs best on all three

Table 4: **Module-effectiveness ablation.** In "+ OCHV + RP", RP reliability conditions are fused into the original BridgeDepth updater without RCE; inference time is measured on Scene Flow.

Model	OCHV	RP	RCE	ETH3D		Middlebury		DrivingStereo		Time (s)
				EPE	BP-1	EPE	BP-2	EPE	BP-3	
Baseline (BridgeDepth)				0.23	1.26	0.92	7.84	1.05	4.28	0.075
+ OCHV	✓			0.22	1.15	0.90	7.76	1.03	4.06	0.078
+ OCHV + RP	✓	✓		0.22	1.07	0.89	7.69	1.00	3.76	0.090
Full (ForceStereo)	✓	✓	✓	**0.21**	**0.97**	**0.87**	**7.62**	**0.98**	**3.59**	0.091
Full w/o OCHV		✓	✓	0.21	1.04	0.90	7.72	1.00	3.73	0.088
Full w/o RP	✓		✓	0.22	1.09	0.89	7.70	1.02	3.92	0.079

Table 5: **Ablation study of RP components.** Each variant changes one reliability cue or residual organization component while keeping the remaining design fixed.

Model	KITTI 2015 BP-3	Middlebury BP-2	ETH3D BP-1
w/o Ambiguity Cue	4.58	7.81	1.06
w/o Edge Cue	4.49	7.69	1.01
w/o Dynamic Geometry	4.61	7.84	1.08
w/o Frequency Separation	4.55	7.77	1.05
w/o Adaptive Gating	4.52	7.73	1.03
Full RP	**4.43**	**7.62**	**0.97**

Table 6: **Ablation study of RCE design.** Direct readout, matched GRU, and one-step RCE are compared with full RCE; the matched GRU uses the same RP regeneration, ambiguity gating, trajectory supervision, and update budget.

Model	KITTI 2015 BP-3	Middlebury BP-2	ETH3D BP-1
w/ Direct readout	4.65	7.80	1.16
w/ Matched GRU (3 steps)	4.56	7.73	1.10
w/ RCE (1 step)	4.50	7.65	1.01
ForceStereo	**4.43**	**7.62**	**0.97**

benchmarks, while dynamic geometry and ambiguity cause the largest drops, confirming their importance for residual interpretation and reliability conditioning. Frequency separation contributes more than adaptive gating, supporting low and high frequency separation before fusion. Edges provide smaller but consistent gains, showing that fixed structure still complements residual calibration.

RCE Design. Tab. 6 shows direct readout degrades across benchmarks, indicating one-shot prediction struggles with unstable observations. Matched GRU shares RP regeneration, ambiguity gating, trajectory supervision, and the three-step budget. One-step RCE still performs better, supporting velocity-field integration. Full RCE further improves all benchmarks via shared-field integration. Fig. 7 remains coherent near the trained interval but degrades under distant extrapolation, supporting coordinate-conditioned evolution.

Table 7: **Module universality ablation study** across other stereo backbones.

Model	Middlebury BP-2	ETH3D BP-1	Time (s)
IGEV-Stereo	9.95	4.05	0.375
Force-IGEV	**8.98**	**3.32**	0.389
NMRF	8.53	3.71	0.042
Force-NMRF	**8.01**	**2.56**	0.058

Module Universality. As shown in Tab. 7, we integrate full ForceStereo into IGEV-Stereo (Xu et al., 2023) and NMRF (Guan et al., 2024), using their native candidates for the dense cost representation of IGEV and the sparse hypotheses of NMRF at different candidate scales. Both variants consistently improve accuracy with minor overhead, showing that ForceStereo effectively adapts across disparity representations and stereo backbones.

5 Conclusion and Discussion

Stereo matching in open world scenes requires not only stronger matching evidence, but also reliability information to guide how observations contribute to disparity estimation. ForceStereo studies this view by reformulating disparity estimation as observation calibrated state integration, where reliability control spans hypothesis formation and subsequent geometric state evolution. Our results suggest that the gains arise not simply from additional priors or repeated updates, but from reliability-aware hypothesis calibration and observation-conditioned state evolution throughout inference. The current formulation still relies on explicitly constructed reliability cues and a fixed integration interval, motivating adaptive reliability discovery and more flexible coordinate-conditioned state evolution. More broadly, we hope this work encourages geometric vision systems to treat observation reliability as a central factor in inference rather than an auxiliary confidence output.

Reproducibility Statement

Most hyperparameters, experimental configurations, and implementation details required to reproduce our reported results are provided in the main text and appendix. To further facilitate reproducibility and independent verification, we provide anonymized code and execution logs in the supplementary material.

AI Use Statement

Generative AI tools were used only for limited language polishing during manuscript preparation, and all AI-assisted text was reviewed by the authors.

References

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (eds.), *Computer Vision – ECCV 2012*, pp. 611–625, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-33783-3. doi: 10.1007/978-3-642-33783-3_44.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2, January 2020.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid Stereo Matching Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

Zhuoqing Chang, Shubo Liu, Run Qiu, Song Song, Zhaohui Cai, and Guoqing Tu. Time-aware neural ordinary differential equations for incomplete time series modeling. *The Journal of Supercomputing*, 79(16):18699–18727, November 2023. ISSN 1573-0484. doi: 10.1007/s11227-023-05327-8.

Lanlan Chen, Kai Wu, Jian Lou, and Jing Liu. Signed Graph Neural Ordinary Differential Equation for Modeling Continuous-Time Dynamics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8292–8301, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i8.28670.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. MonSter: Marry Monodepth to Stereo Unleashes Power. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6273–6282, 2025.

Woojin Cho, Seunghyeon Cho, Hyundong Jin, Jinsung Jeon, Kookjin Lee, Sanghyun Hong, Dongeun Lee, Jonghyun Choi, and Noseong Park. Operator-Learning-Inspired Modeling of Neural Ordinary Differential Equations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11543–11551, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i10.29036.

C. Coelho, M. Fernanda P. Costa, and L. L. Ferrás. Enhancing continuous time series modelling with a latent ODE-LSTM approach. *Applied Mathematics and Computation*, 475:128727, August 2024. ISSN 0096-3003. doi: 10.1016/j.amc.2024.128727.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, June 2012. doi: 10.1109/CVPR.2012.6248074.

Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural Markov Random Field for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5459–5469, 2024.

Tongfan Guan, Jiaxin Guo, Chen Wang, and Yun-Hui Liu. BridgeDepth: Bridging Monocular and Stereo Reasoning with Latent Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27681–27691, 2025.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-Wise Correlation Stereo Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3273–3282, 2019.

Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. DEFOM-Stereo: Depth Foundation Model Based Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21857–21867, 2025.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-To-End Learning of Geometry and Context for Deep Stereo Regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 66–75, 2017.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16263–16272, 2022.

Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei. Contextual Transformer Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1489–1500, February 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3164083.

Lahav Lipson, Zachary Teed, and Jia Deng. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In *2021 International Conference on 3D Vision (3DV)*, pp. 218–227, December 2021. doi: 10.1109/3DV53792.2021.00032.

Meng Lou and Yizhou Yu. OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 128–138, June 2025. doi: 10.1109/CVPR52734.2025.00021.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, 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*, pp. 4040–4048, 2016.

Moritz Menze and Andreas Geiger. Object Scene Flow for Autonomous Vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3061–3070, 2015.

Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the Confidence of Stereo Matching in a Deep-Learning Era: A Quantitative Evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5293–5313, September 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3069706.

Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, March 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2020.3019967.

Thomas Schops, Johannes L. Schonberger, 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*, pp. 3260–3269, 2017.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 402–419, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

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.

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, October 2020. doi: 10.1109/IROS45743.2020.9341801.

Xianqi Wang, Hao Yang, Hangtian Wang, Junda Cheng, Gangwei Xu, Min Lin, and Xin Yang. PromptStereo: Zero-Shot Stereo Matching via Structure and Motion Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12565–12575, 2026.

Yun Wang, Longguang Wang, Chenghao Zhang, Yongjian Zhang, Zhanjie Zhang, Ao Ma, Chenyou Fan, Tin Lun Lam, and Junjie Hu. Learning Robust Stereo Matching in the Wild with Selective Mixture-of-Experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21276–21287, 2025.

Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention Concatenation Volume for Accurate and Efficient Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12981–12990, 2022.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21919–21928, 2023.

Haofei Xu and Juyong Zhang. AANet: Adaptive Aggregation Network for Efficient Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0688.

Chengtang Yao, Lidong Yu, Zhidan Liu, Jiaxi Zeng, Yuwei Wu, and Yunde Jia. Diving into the Fusion of Monocular Priors for Generalized Stereo Matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14887–14897, 2025.

Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H. S. Torr. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 185–194, 2019.

Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. MoCha-Stereo: Motif Channel Attention Network for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27768–27777, 2024.

APPENDIX

The supplementary material provides additional method details, evaluation protocols, and mechanism analyses to support reproducibility and the main design conclusions of ForceStereo. Specifically, we first detail latent alignment, hypothesis ambiguity, geometry-state readout, and RP/RCE implementation, followed by the observation-unreliability evaluation protocol and its independent validation. We then analyze hypothesis calibration and state-conditioned evolution, and conclude by discussing the relationship between ForceStereo and BridgeDepth and the broader perspective of observation-conditioned geometric inference.

A ADDITIONAL METHOD DETAILS

This section provides key method definitions and implementation details omitted from the main paper. We focus on latent alignment, the candidate distribution, state readout, and the implementation details of RP and RCE.

A.1 MONOCULAR–STEREO LATENT ALIGNMENT

After OCHV calibrates the disparity hypotheses, we adopt the bidirectional latent-alignment mechanism of BridgeDepth (Guan et al., 2025), allowing the calibrated stereo hypotheses to interact with structural representations from the frozen monocular branch.

We denote a windowed cross-attention operation from representation Y to X as $\Psi_{a\leftarrow b}(X, Y)$, where branch a reads information from branch b:

$$\begin{aligned} Q &= \mathrm{LN}(X)W_q^a, \qquad [K, V] = \mathrm{LN}(Y)[W_k^b, W_v^b], \\ X' &= X + \mathrm{WindowAttn}(Q, K, V), \qquad \Psi_{a\leftarrow b}(X, Y) = X' + \mathrm{FFN}(\mathrm{LN}(X')) \,. \end{aligned} \tag{12}$$

The two branches use separate query, key, and value projections, preserving their respective representation spaces during interaction. Let S_i and M_i denote the stereo hypotheses and monocular representation at alignment stage i, respectively; the stereo branch first reads monocular structural information and further aggregates candidate geometry through neural message passing:

$$\widetilde{S}_i = \Psi_{s\leftarrow m}(S_i, M_i), \qquad S_{i+1} = \mathrm{NMP}\Big(\widetilde{S}_i\Big) \,. \tag{13}$$

The interaction direction is then reversed, allowing the monocular branch to read geometric information from the updated stereo representation:

$$M_{i+1} = \Psi_{m\leftarrow s}\left(M_i, S_{i+1}\right) . \tag{14}$$

After the final bidirectional interaction, we denote the aligned stereo representation retaining the candidate dimension as $\mathbf{S}_{\mathrm{align}}$ and the aligned context for subsequent reliability conditioning as $\mathbf{C}_{\mathrm{align}}$. Here, $\mathbf{S}_{\mathrm{align}}$ is used to form the initial geometry state and estimate hypothesis ambiguity, while $\mathbf{C}_{\mathrm{align}}$ carries the aligned context into subsequent state evolution.

A.2 HYPOTHESIS DISTRIBUTION AND AMBIGUITY

To characterize the unresolved competition among aligned hypotheses, we map each candidate in $\mathbf{S}_{\mathrm{align}}$ to a pixelwise support score and normalize over the candidate dimension:

$$\ell_k = \mathcal{R}_{8\times 8}\Big(0.25\,\mathrm{Linear}\left(\mathbf{S}_{\mathrm{align}}^{(k)}\right)\Big) \,, \qquad p_k(x) = \frac{\exp(\ell_k(x))}{\sum_{j=1}^{K}\exp(\ell_j(x))} . \tag{15}$$

Here, $\mathcal{R}_{8\times 8}$ rearranges each local score block to its corresponding pixels, while p_k reflects only relative support within the retained hypothesis set rather than observation reliability. We define hypothesis ambiguity by combining normalized entropy with the probability gap between the top two candidates:

$$\begin{aligned} \mathcal{H}_K(p) &= -\frac{1}{\ln K}\sum_{k=1}^{K} p_k \ln\big(\max(p_k, 10^{-6})\big) \,, \\ \Delta_2(p) &= p_{(1)} - p_{(2)}, \qquad \mathbf{U}(p) = \frac{1}{2}\left[\mathcal{H}_K(p) + 1 - \Delta_2(p)\right] . \end{aligned} \tag{16}$$

A larger $\mathbf{U}$ indicates greater unresolved competition among candidates and thus greater adjustment freedom during subsequent evolution; when $K = 1$, we set $\mathbf{U} = 0$.

A.3 Geometry-State Readout

RCE directly evolves an explicit local subpixel geometry state rather than a generic hidden representation that requires an additional disparity prediction head. Specifically, each location at $1/4$ resolution maintains a local disparity state for a 4×4 full-resolution image block, giving $\mathbf{h}(t) \in \mathbb{R}^{B\times H/4\times W/4\times 16}$. The initial geometry state combines the coarse disparity with a local disparity correction predicted from the aligned stereo representation:

$$\mathbf{h}(0) = \mathrm{ReLU}\left(\mathbf{D}_0^{\mathrm{coarse}} + \Delta\mathbf{D}_{\mathrm{patch}}\right). \tag{17}$$

Thus, RCE evolves in local disparity coordinates with explicit geometric meaning from initialization onward. During evolution, RP reconstructs cross-view observations from the current geometry state, so each local 4×4 state is first reduced to a macro disparity:

$$\bar{\mathbf{D}}_t(i,j) = \frac{1}{16}\sum_{r=1}^{16} \mathbf{h}_{i,j}^{r}(t). \tag{18}$$

This readout remains at $1/4$ resolution and is used for feature warping, residual computation, and matching geometry reconstruction at the current state. At the terminal state, no additional learned prediction head is required; we directly recover the full-resolution disparity as $\mathbf{D}_{\mathrm{final}} = 4\,\Pi_4(\mathbf{h}(1))$. Here, Π_4 deterministically rearranges the 16 geometry components at each location into a 4×4 pixel block, while the factor of 4 restores disparity to the full-resolution pixel scale. Thus, the intermediate macro disparity and final dense disparity are deterministic readouts of the same evolving geometry state, so observation reinterpretation and final prediction are grounded in the same geometry state.

A.4 Reliability Prompt and RCE Implementation

From a dynamical-systems perspective, RP can be viewed as a state-dependent observation operator rather than an independent confidence predictor. Let $\mathcal{O}$ denote the fixed input evidence and initialization context; the current reliability condition can be written abstractly as

$$\mathbf{P}(t) = \mathcal{P}\left(\mathbf{h}(t); \mathcal{O}\right), \qquad \mathcal{O} = \{\mathbf{I}_l, \mathbf{I}_r, \mathbf{F}_l, \mathbf{F}_r, \mathbf{C}_{\mathrm{align}}, \mathbf{U}, \mathbf{E}\}. \tag{19}$$

Because the current geometry state changes cross-view warping and the resulting residuals and matching geometry, RP defines a feedback mapping from geometry state to observation interpretation. Substituting this state-dependent condition into RCE gives the following closed-loop neural dynamical system:

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{V}_\theta\left(\mathbf{h}(t), \mathcal{P}\left(\mathbf{h}(t); \mathcal{O}\right), t\right) \equiv \mathcal{F}_\theta\left(\mathbf{h}(t); \mathcal{O}, t\right). \tag{20}$$

A change in the geometry state affects subsequent dynamics through two pathways: one directly changes the local vector field, while the other feeds back through changes in the current observation interpretation. This coupling between direct state dependence and observation feedback can be characterized by the total derivative of the composed vector field with respect to the geometry state:

$$\begin{aligned} \mathrm{D}_{\mathbf{h}}\mathcal{F}_\theta &= \left.\frac{\partial \mathbf{V}_\theta}{\partial \mathbf{h}}\right|_{\mathbf{P}} + \left.\frac{\partial \mathbf{V}_\theta}{\partial \mathbf{P}}\right|_{\mathbf{h}} \mathrm{D}_{\mathbf{h}}\mathcal{P} \\ &= \left.\frac{\partial \mathbf{V}_\theta}{\partial \mathbf{h}}\right|_{\mathbf{P}} + \left.\frac{\partial \mathbf{V}_\theta}{\partial \mathbf{P}}\right|_{\mathbf{h}} \left[\frac{\partial \mathcal{P}}{\partial \mathbf{R}}\frac{\partial \mathbf{R}}{\partial \mathbf{D}} + \frac{\partial \mathcal{P}}{\partial \mathbf{G}}\frac{\partial \mathbf{G}}{\partial \mathbf{D}}\right]\frac{\partial \mathcal{R}}{\partial \mathbf{h}}, \qquad \mathbf{D} = \mathcal{R}(\mathbf{h}). \end{aligned} \tag{21}$$

The first term captures the direct effect of the geometry state on the local vector field with the reliability condition held fixed. The second term captures state-dependent observation feedback: the geometry state first changes the current disparity through the readout $\mathcal{R}$, which alters the cross-view residual $\mathbf{R}$ and matching geometry $\mathbf{G}$ and ultimately updates the reliability condition through $\mathcal{P}$. In contrast, hypothesis ambiguity $\mathbf{U}$, edge structure $\mathbf{E}$, and aligned context $\mathbf{C}_{\mathrm{align}}$ remain fixed over each evolution trajectory and thus introduce no additional feedback paths through $\mathbf{h}$. This decomposition shows that RP and RCE are coupled not only through their dependence on the current state, but also through the continual reconstruction of observation conditions as the state evolves.

Finally, we interpret hypothesis ambiguity as a local adjustment budget and impose a soft velocity bound with a nonzero floor:

$$b(\mathbf{U}) = \epsilon_v + (1 - \epsilon_v)\mathbf{U}, \qquad \mathbf{V} = b(\mathbf{U}) \odot \sigma(\mathbf{Z}^q) \odot \tanh(\mathbf{Z}^v), \qquad \epsilon_v = 0.25. \tag{22}$$

Since $|\mathbf{V}(x,t)| \leq b(\mathbf{U}(x))$, low-ambiguity regions are more strongly constrained, while high-ambiguity regions retain greater adjustment freedom, and the nonzero floor prevents them from being completely frozen.

B Reliability Evaluation and Independent Validation

This section details the observation-unreliability evaluation used in the main paper and validates the regional robustness findings using geometric regions defined independently of this metric.

B.1 Observation Unreliability Protocol

The unreliability ranking used in the main paper is solely for offline evaluation and is independent of any evaluated model's predictions, internal features, or confidence estimates. For each valid pixel $x = (u, v)$, we align the right image using ground-truth disparity and combine cross-view photometric inconsistency with local texture weakness to define the unreliability score:

$$\begin{aligned} P(x) &= \frac{1}{3}\sum_c \left| I_l^c(x) - I_r^c\left(u - d_{\mathrm{gt}}(x), v\right)\right|, \qquad G(x) = \left\|\nabla I_l^{\mathrm{gray}}(x)\right\|, \\ S_{\mathrm{unrel}}(x) &= \frac{1}{2}\mathcal{N}[P](x) + \frac{1}{2}\left(1 - \mathcal{N}[G](x)\right), \\ \Omega_q &= \left\{x \in \Omega_{\mathrm{valid}} \mid \mathrm{PctRank}\left(S_{\mathrm{unrel}}(x)\right) \geq q\right\}. \end{aligned} \tag{23}$$

Here, $\mathcal{N}[\cdot]$ denotes clipped normalization using the 5th and 95th percentiles over valid pixels, with scores smoothed by a 5×5 Gaussian kernel before percentile ranking. A higher q selects a smaller region with more unreliable observations; for example, $q = 90$ retains approximately the most unreliable 10% of valid pixels. All methods use identical region masks, so the protocol changes only the evaluation region and does not use any model-specific output to define difficult pixels.

B.2 Independent Reliability Validation

Table 8: **Independent validation on GT-defined geometric regions** using Middlebury zero-shot predictions. EPE (px) is reported and lower is better.

	GT Disparity Gradient			Boundary Proximity		
Method	Low	Medium	High	Interior	Transition	Boundary
BridgeDepth	0.57	0.93	1.78	0.70	1.00	1.35
PromptStereo	0.56	0.96	1.96	0.69	1.05	1.50
ForceStereo	**0.55**	**0.89**	**1.64**	**0.68**	**0.93**	**1.26**

To verify that the regional robustness findings do not depend on the specific definition of S_{unrel}, we further conduct an independent geometry-based analysis using only ground-truth disparity. This analysis directly reuses the Middlebury zero-shot predictions from the main paper, requiring no additional training or inference.

Pixels are stratified by ground-truth disparity gradients into low-, medium-, and high-gradient regions covering 50%, 30%, and 20% of valid pixels, respectively. In parallel, geometric discontinuities are extracted from ground-truth disparity, and pixels are divided by distance to the nearest boundary into interior, transition, and boundary regions with the same proportions. Geometric discontinuities are identified as high-gradient pixels in the ground-truth disparity, with spatial proximity computed using the Euclidean distance transform.

As shown in Tab. 8, errors increase with geometric difficulty for all methods, while ForceStereo's errors grow more slowly in high-gradient and boundary regions. Since these regions are defined solely

by ground-truth geometry and are independent of S_{unrel}, RP, and model predictions, the results provide independent support for the regional robustness findings under complementary geometry-based partitions.

Reliability versus Confidence. Observation reliability in our formulation plays a different role from prediction confidence in conventional stereo matching. Prediction confidence typically estimates the correctness of a model's own disparity prediction, whereas our formulation concerns how current observations should influence geometry-state evolution during inference. In particular, the offline unreliability metric above does not use the candidate distribution $\mathbf{p}$, hypothesis ambiguity $\mathbf{U}$, RP conditions, or ForceStereo disparity predictions. It therefore serves only as a common external analysis tool, whereas ForceStereo regulates state formation and evolution through hypothesis calibration and reliability conditioning during inference.

C Hypothesis Calibration Analysis

This section further analyzes proposal count and calibration quality to examine how scene-level context improves local hypotheses before geometry-state formation. We first study the trade-off among proposal coverage, proposal error, and computational cost, and then directly evaluate how OCHV redistributes support within a fixed candidate set.

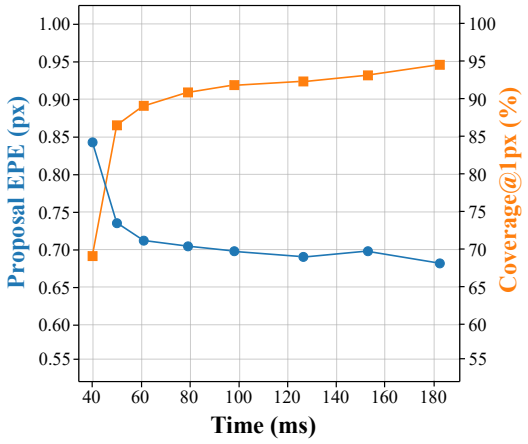


Figure 8: **Proposal-quality and efficiency trade-off for** $K = 1, \ldots, 8$. Coverage@1px is the percentage of valid pixels whose retained proposal set contains at least one hypothesis within 1 px of ground truth.

C.1 Proposal-Number Trade-off

DPN retains multiple local disparity proposals to avoid prematurely collapsing ambiguous matching evidence into a single hypothesis before state formation. Fig. 8 analyzes the trade-off among proposal accuracy, coverage, and computational cost for $K = 1, \ldots, 8$. Moving from a single proposal to a small competing set substantially reduces Proposal EPE while rapidly improving Coverage@1px, indicating that a single local hypothesis can prematurely discard plausible matching alternatives. As K increases further, proposal coverage continues to improve moderately, but accuracy gains gradually saturate while inference time continues to increase. We therefore use $K = 2$ by default, preserving the main competing-hypothesis information while avoiding the additional computational cost of larger candidate sets. These results suggest that the goal is not to enlarge the proposal space, but to retain sufficient alternative hypotheses for subsequent scene-conditioned calibration before state formation.

Table 9: **Analysis of hypothesis support assignment before state formation.** All representations are evaluated on the same retained disparity proposals. Higher selection accuracy and margin are better, while lower NLL, entropy, and ambiguity are better.

Representation	Sel. Acc. ↑	NLL ↓	$\mathcal{H}_K$ ↓	Δ_2 ↑	$\mathbf{U}$ ↓
LHV	78.6	0.517	0.641	0.336	0.653
OHV	77.4	0.534	0.632	0.349	0.671
OCHV	**83.3**	**0.448**	**0.568**	**0.431**	**0.557**

C.2 Hypothesis Calibration and OCHV Design

OCHV does not alter the disparity proposals produced by DPN, but combines local matching evidence with scene-level context over the same candidate set. We therefore analyze its calibration behavior in terms of candidate-selection quality and distributional ambiguity.

Candidate Selection Quality. For each pixel, we take the retained proposal closest to the ground-truth disparity as the reference candidate and evaluate whether each representation assigns it the highest support. Tab. 9 reports candidate-selection accuracy, negative log-likelihood of the reference candidate, normalized entropy, top-two probability margin, and the resulting hypothesis ambiguity. LHV relies on fine-grained local correspondence and therefore provides more reliable candidate support than the overview-conditioned OHV when used independently. In contrast, OHV is not intended to replace local matching, but to provide broader scene context for the retained candidates, resulting in slightly weaker standalone performance than LHV.

When the two evidence sources are combined in OCHV, candidate-selection accuracy increases substantially while NLL decreases, showing that scene-level context helps resolve candidate competition left ambiguous by local evidence. OCHV also yields lower candidate entropy and a larger top-two probability margin, further reducing hypothesis ambiguity before state formation. These results indicate that OCHV benefits from the complementarity between local matching and scene context, rather than from a stronger hypothesis representation produced by scene overview alone.

Table 10: **Candidate-selection accuracy (%) across empirical tertiles of the pre-calibration LHV ambiguity.** T_1, T_2, and T_3 denote equal-sized strata defined by the 33.3-th and 66.7-th percentiles. All representations share identical subsets, and Gain denotes the absolute improvement of OCHV over LHV.

Ambiguity Stratum	LHV	OCHV	Gain
T_1 (0–33.3%)	93.4	94.3	+0.9
T_2 (33.3–66.7%)	81.7	85.4	+3.7
T_3 (66.7–100%)	60.7	70.2	+9.5

Table 11: **Ablation study of the OCHV dynamic-kernel design.** Lower is better for all reported metrics.

Kernel Set	ETH3D BP-1	Middlebury BP-2	DrivingStereo BP-3
$\{3,5\}$	1.18	7.81	4.01
$\{3,7\}$	1.09	7.73	3.85
$\{5,7\}$	**0.97**	**7.62**	**3.59**
$\{7,11\}$	1.05	7.68	3.66
$\{3,5,7\}$	0.99	7.66	3.64
$\{5,7,11\}$	1.03	7.69	3.70

Calibration across Initial-Ambiguity Tertiles. To analyze how scene-level context affects different levels of initial hypothesis competition, we rank valid pixels using only the pre-calibration ambiguity $\mathbf{U}_{\mathrm{LHV}}$ produced by LHV. We partition the pixels into three equally populated strata using the 33.3rd and 66.7th percentiles of $\mathbf{U}_{\mathrm{LHV}}$ over the evaluation set:

$$T_1 : \mathbf{U}_{\mathrm{LHV}} < Q_{33.3}, \qquad T_2 : Q_{33.3} \le \mathbf{U}_{\mathrm{LHV}} < Q_{66.7}, \qquad T_3 : \mathbf{U}_{\mathrm{LHV}} \ge Q_{66.7}. \tag{24}$$

Here, $Q_{33.3}$ and $Q_{66.7}$ are empirical tertiles of the pre-calibration LHV ambiguity distribution, avoiding manually chosen thresholds. The three pixel subsets are fixed before OCHV calibration and used consistently across all representations. As shown in Tab. 10, the candidate-selection gain of OCHV over LHV increases consistently from T_1 to T_3. This indicates that scene-level context provides more useful complementary evidence when local matching cannot clearly distinguish competing hypotheses.

Dynamic Kernel Design. As shown in Tab. 11, $\{5,7\}$ denotes the 5×5 small kernel and 7×7 large kernel in OCHV and achieves the best results across all three datasets. This result shows that

Table 12: **Effect of state-dependent observation reinterpretation.** "Update" indicates that the corresponding condition is regenerated from the current geometry state. The $q = 90$ column evaluates approximately the most unreliable 10% of valid Middlebury pixels.

Variant	$\mathbf{R}_t$	$\mathbf{G}_t$	Middlebury BP-2 (%)		ETH3D BP-1 (%)
			All	$q = 90$	All
Fixed RP	Fixed	Fixed	7.91	11.24	1.13
Dynamic Residual	Update	Fixed	7.79	10.90	1.08
Dynamic Geometry	Fixed	Update	7.82	10.68	1.05
Full Dynamic	Update	Update	**7.62**	**10.03**	**0.97**

Table 13: **Continuous versus discrete refinement on Middlebury.** The all-region values follow the controlled comparison in the main paper, while $q = 90$ isolates highly unreliable observations. Lower is better.

Method	All BP-2 (%)	$q = 90$ BP-2 (%)
Direct Readout	7.80	11.51
Matched GRU	7.73	11.16
RCE (1 Step)	7.65	10.49
RCE (3 Steps)	**7.62**	**10.03**

small and large receptive fields provide complementary information for hypothesis calibration by preserving local candidate details and broader neighborhood context, respectively. Smaller, larger, or three-kernel configurations provide no further gains, suggesting that OCHV benefits from complementary scene context rather than simply larger receptive fields or more dynamic branches.

D State-Conditioned Evolution Analysis

This section analyzes reliability-guided state evolution from four perspectives: state-dependent observation conditioning, continuous versus discrete refinement, sensitivity to numerical settings, and the resulting evolution trajectories. Unlike the module-level ablations in the main paper, we focus on the mechanistic coupling among state changes, observation reinterpretation, and subsequent state evolution.

State-Dependent Observation Interpretation. Table 12 analyzes state-dependent observation interpretation by selectively freezing or recomputing the residual $\mathbf{R}_t$ and matching geometry $\mathbf{G}_t$, while $\mathbf{C}_{\text{align}}$, $\mathbf{U}$, and $\mathbf{E}$ remain fixed throughout the trajectory. Freezing the initial RP clearly degrades performance, whereas recomputing both state-dependent observations gives the best results, with a larger gap in highly unreliable regions. These results show that reinterpreting current observations as the geometry state changes is important for reliability-guided state evolution.

RCE versus Discrete Refinement. Table 13 compares direct readout, discrete GRU refinement, and RCE under matched observation conditions and update budgets, with additional evaluation on highly unreliable regions. Compared with the matched GRU, one-step RCE further reduces error, while three-step integration of the shared vector field provides additional gains, especially in highly unreliable regions. Thus, the benefit of RCE cannot be attributed solely to additional update rounds and is consistent with accumulating state changes through a shared local derivative field.

Sensitivity to Evolution Configuration. For the closed-loop evolution in Eq. 20, Table 14 varies the number of vector-field evaluations over the fixed interval $[0, 1]$ to examine the effect of integration budget on accuracy and inference efficiency. Increasing the number of evaluations from one to three consistently reduces error on both ETH3D and DrivingStereo, while increasing it to five yields only marginal additional gains. Meanwhile, five-step integration noticeably reduces inference speed, making three-step Euler a practical trade-off between near-saturated accuracy and computational efficiency and the default for all main experiments.

Evolution Dynamics. To examine hypothesis ambiguity as a local evolution budget in Eq. 22, we use the same tertile partition as Eq. 24, ranking valid pixels by hypothesis ambiguity $\mathbf{U}$ immediately

Table 14: **Sensitivity to the Euler integration budget over the fixed interval $[0, 1]$.** NFE denotes the number of vector-field evaluations, and FPS is measured for end-to-end inference.

NFE	ETH3D BP-1 (%)	DrivingStereo BP-3 (%)	FPS ↑
1	1.01	3.75	14.2
2	0.99	3.65	12.1
3 (Default)	0.97	3.59	11.0
5	**0.96**	**3.58**	7.4

Table 15: **Evolution dynamics across initial-ambiguity tertiles.** Valid pixels are ranked by the hypothesis ambiguity $\mathbf{U}$ immediately before RCE and divided into three equally populated groups, which are fixed throughout the evolution trajectory. EPE is measured at each intermediate geometry state, while $|\Delta\mathbf{h}|$ denotes the mean absolute state change over each integration interval.

	EPE along Evolution (px)				Mean State Change		
Initial $\mathbf{U}$ Tertile	$t=0$	$t=\frac{1}{3}$	$t=\frac{2}{3}$	$t=1$	$\lvert\Delta\mathbf{h}_0\rvert$	$\lvert\Delta\mathbf{h}_1\rvert$	$\lvert\Delta\mathbf{h}_2\rvert$
T_1^{evo} (0–33.3%)	0.41	0.39	0.38	0.36	0.028	0.017	0.015
T_2^{evo} (33.3–66.7%)	0.90	0.83	0.76	0.73	0.071	0.054	0.038
T_3^{evo} (66.7–100%)	2.01	1.84	1.67	1.52	0.151	0.113	0.091

before RCE. The three groups are determined once at $t = 0$ and remain fixed throughout the evolution trajectory. As shown in Tab. 15, higher initial-ambiguity tertiles exhibit larger state changes and greater cumulative error reduction. Meanwhile, state changes decrease along the trajectory for all groups, indicating that greater adjustment freedom is associated with unresolved hypothesis competition, while the actual updates diminish as the geometry state stabilizes.

Additional Zero-Shot Qualitative Results. We further provide zero-shot qualitative comparisons on ETH3D and DrivingStereo to complement the quantitative generalization results in the main paper. As shown in Fig. 9, both methods recover similar global geometry in stable regions of real ETH3D scenes, while their differences are more visible around object boundaries, fine structures, and locally difficult correspondences. ForceStereo preserves more coherent local geometry in these regions, consistent with the zero-shot quantitative results on ETH3D reported in the main paper.

Under more challenging DrivingStereo conditions, Fig. 10 highlights differences among the methods around distant vehicles, strongly illuminated regions, and low-contrast road structures. ForceStereo produces more stable disparity structures in these local regions while maintaining coherent geometry in the surrounding scene. These qualitative results complement the zero-shot generalization analysis in the main paper and show that the observed improvements are not restricted to a single dataset or imaging condition.

E Discussion

E.1 Relationship to BridgeDepth

ForceStereo inherits the latent alignment between monocular structural representations and stereo hypotheses from BridgeDepth Guan et al. (2025), providing the basis for interaction between the two representations. The main extension of ForceStereo is not a new monocular prior, but a different way for observations to influence the geometry state: local hypotheses are calibrated with scene-level context before state formation, and observations are reinterpreted according to the current geometry during state evolution. BridgeDepth primarily establishes latent interaction between representations, while ForceStereo further studies how reliability regulates the influence of observations on geometry-state formation and evolution.

E.2 Toward Observation-Conditioned Geometric Inference

The question studied by ForceStereo goes beyond confidence estimation in stereo matching and concerns how current observations should influence an evolving geometric representation according to

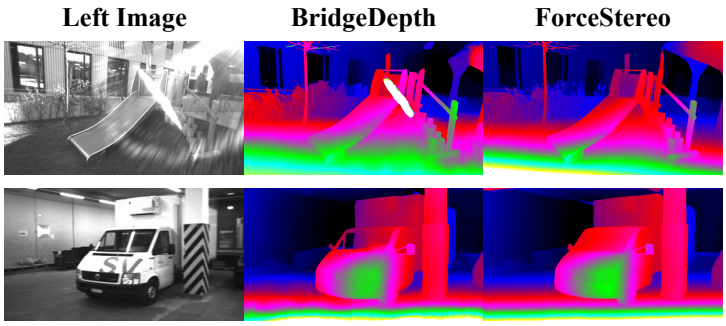


Figure 9: **Additional zero-shot qualitative results on ETH3D.** Compared with BridgeDepth, ForceStereo exhibits more coherent disparity structures around object boundaries and locally challenging regions.

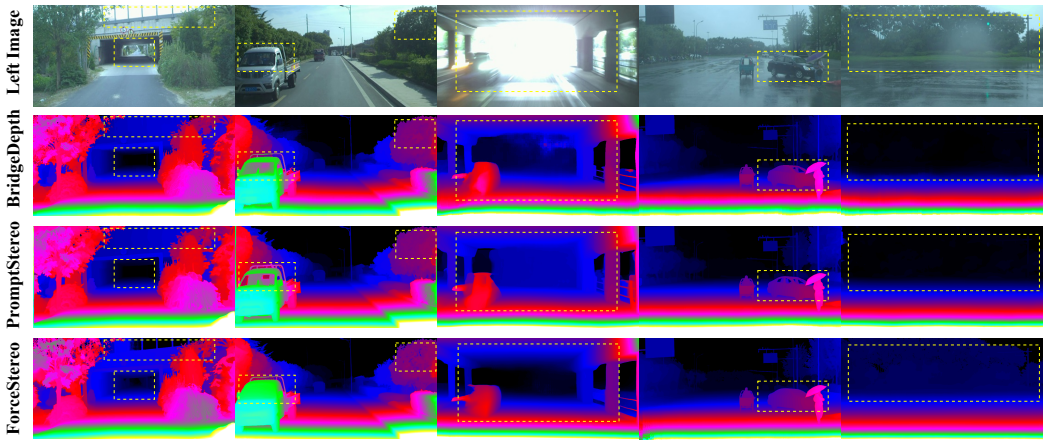


Figure 10: **Additional zero-shot qualitative results on DrivingStereo.** The examples include distant objects, strong illumination changes, and low-visibility regions, with predictions from BridgeDepth, PromptStereo, and ForceStereo.

the geometry state. From this perspective, reliability is not merely an auxiliary output after prediction, but can act as an inference variable controlling when and how strongly observations influence state changes. This shift from estimating reliability to using it to regulate state evolution may also offer a useful direction for multi-view reconstruction, depth fusion, and other geometric inference problems involving heterogeneous and potentially unreliable observations.