SW-MAMBA: SELECTIVE WEATHER FORGETTING FOR SELF-SUPERVISED MONOCULAR DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised monocular depth estimation relies on cross-frame photometric consistency, an assumption severely challenged by adverse weather where appearance changes deviate from scene geometry. To address these challenges, we propose Selective Weather-forgetting Mamba (SW-Mamba), which formulates robust adverse-weather depth estimation as selective latent-state correction. Rather than treating inter-frame variation as a weather mask, SW-Mamba interprets temporal cues jointly with multidirectionally propagated spatial context to estimate a spatiotemporal correction gate. The gate modulates reference-guided state correction, selectively updating unreliable responses while preserving geometric consistency. During paired training, aligned clean states provide supervision for gate calibration and serve as structural references. At inference, an internal structural proxy enables deployment without aligned clean observations. Extensive experiments on WeatherKITTI, nuScenes, and DrivingStereo show consistent gains under adverse weather and strong cross-domain generalization.

1 INTRODUCTION

Self-supervised monocular depth estimation (MDE) learns scene geometry directly from monocular video by exploiting appearance consistency across adjacent views (Godard et al., 2019; Choi et al., 2025). Its effectiveness relies on a key assumption: cross-frame appearance variation should primarily reflect geometric transformations. Under adverse conditions, however, this assumption is easily violated. Rain, snow, fog, and illumination changes introduce non-geometric appearance variation, entangling valid geometric cues with transient appearance corruption.

Existing adverse-condition methods mitigate this problem through cleaner supervision (Saunders et al., 2023), multi-condition training (Gasperini et al., 2023), weather-aware objectives (Wang et al., 2024a), generative regularization (Wang et al., 2024b), and domain adaptation (Yan et al., 2025; Jiang et al., 2025). These strategies improve robustness through data, supervision, or objective design, but do not explicitly model how latent representations should be updated when geometric and non-geometric variations become entangled. The key question is therefore which latent states should be retained and which should be corrected when appearance consistency becomes unreliable.

Figure 1 highlights an asymmetry between weather-induced appearance variation and geometric structure. Across adjacent frames, weather perturbations are often spatially irregular and temporally unstable, whereas scene geometry remains comparatively coherent. This motivates selective latent-state correction. Temporal variation provides a useful cue for unstable responses, but is not weather-specific, since viewpoint changes, object motion, and occlusion can induce similar variations. Effective correction therefore requires interpreting temporal variation together with the surrounding latent structure.

Based on this observation, we propose Selective Weather-forgetting Mamba (SW-Mamba) for robust self-supervised MDE. SW-Mamba decomposes latent-state correction into three functions: establishing spatial context, estimating correction strength, and determining correction direction. First, **State Propagation** aggregates multidirectional spatial context around each latent response. Second, **State Gating** combines this context with temporal variation to estimate how strongly each state should be corrected. Finally, since correction strength does not determine the update target, **State Correction** moves selected states toward a structural reference, using aligned clean observations during training and an internal proxy at inference. Together, these mechanisms correct transient weather corruption

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.

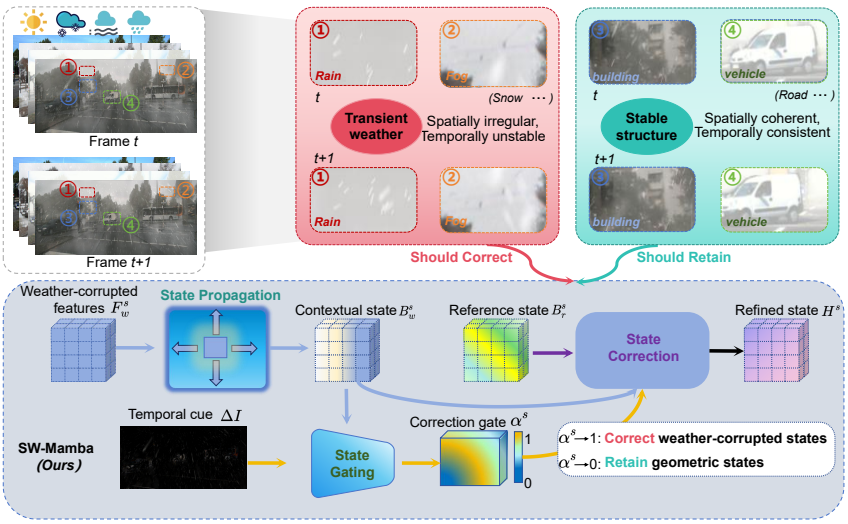

Figure 1: **Motivation.** Adverse-weather transients are spatially irregular and temporally unstable, whereas geometric structure remains comparatively coherent across adjacent frames. SW-Mamba uses *State Propagation* to establish spatial context, *State Gating* to determine correction strength from temporal and contextual cues, and *State Correction* to update selected states toward a structural reference while leaving reliable states largely unchanged.

while preserving geometrically meaningful representations. Extensive experiments on WeatherKITTI, nuScenes, and DrivingStereo demonstrate consistent improvements under adverse conditions.

The main contributions of this work are:

- We formulate adverse-weather MDE as a selective latent-state correction problem, distinguishing transient weather corruption from geometrically meaningful representations.
- We propose SW-Mamba, which integrates multidirectional state propagation, spatiotemporal gating, and reference-guided correction to determine spatial context, correction strength, and correction direction, while an internal proxy enables inference without clean observations.
- Extensive experiments on WeatherKITTI, nuScenes, and DrivingStereo demonstrate consistent improvements across adverse-weather depth estimation, real-world conditions, and zero-shot transfer, with favorable accuracy–efficiency trade-offs.

2 Related Work

Self-supervised Monocular Depth Estimation. Self-supervised MDE commonly learns geometry from view synthesis across adjacent frames (Godard et al., 2019). Subsequent work has improved temporal correspondence, feature representation, efficiency, and geometric regularization through multi-frame matching (Watson et al., 2021), stronger visual backbones (Zhao et al., 2022; Zhang et al., 2023), and explicit structural priors (Wang et al., 2023). Large pretrained models have further strengthened transferable monocular representations (Yang et al., 2024; Piccinelli et al., 2024). Despite these advances, self-supervised learning remains sensitive to inter-frame appearance changes that cannot be explained by scene geometry.

Robust Depth under Adverse Conditions. Robust MDE under degraded visibility has been approached through cleaner supervision and cross-observation consistency (Saunders et al., 2023), multi-condition training (Gasperini et al., 2023), weather-aware self-supervision (Wang et al., 2024a), and generative regularization (Wang et al., 2024b). Recent studies further improve robustness through synthetic-to-real transfer and adaptation to unseen conditions (Yan et al., 2025; Jiang et al., 2025). These methods mainly address adverse conditions through data, supervision, or objective design. Instead, SW-Mamba improves weather robustness through selective correction of latent states.

Visual State-space Models. Recent visual state-space models adapt selective sequence modeling to two-dimensional representations through bidirectional and multidirectional state propagation (Zhu et al., 2024; Liu et al., 2024). Recent designs further improve spatial modeling through local selective scanning and structure-aware state fusion (Huang et al., 2024; Xiao et al., 2025). These developments provide increasingly effective spatial context within recurrent state representations. SW-Mamba builds on this capability by using the propagated state to contextualize temporal variation, so that the correction gate is determined jointly by temporal change and latent spatial structure.

3 METHOD

3.1 PROBLEM FORMULATION AND METHOD OVERVIEW

Let $I_t^w \in \mathbb{R}^{3\times H\times W}$ denote a target frame captured under adverse weather and I_j^w an adjacent source frame. Self-supervised MDE predicts target-view depth and relative camera motion,

$$D_t = f_\theta(I_t^w), \qquad P_{t\to j} = g_\xi(I_t^w, I_j^w), \qquad \widehat{I}_{t\leftarrow j}^w = \mathcal{W}\left(I_j^w, D_t, P_{t\to j}, K\right), \tag{1}$$

where K denotes the camera intrinsics and $\mathcal{W}$ is differentiable view synthesis. Training constrains D_t through the photometric agreement between the target I_t^w and its synthesized counterpart $\widehat{I}_{t\leftarrow j}^w$.

Under adverse weather, cross-frame appearance variation no longer provides an unambiguous geometric cue, as weather-induced changes violate the warp in Eq. (1). Temporal variation, however, cannot simply be suppressed: valid geometric parallax also produces strong cross-frame changes, while object motion and occlusion introduce additional non-rigid variations. This motivates context-dependent use of temporal variation, as indiscriminate suppression removes useful geometric evidence, whereas indiscriminate propagation admits non-geometric disturbances. The key problem is thus to determine *which latent states should be retained and which should be corrected*.

SW-Mamba therefore addresses adverse-weather depth estimation through selective state correction rather than explicit weather detection. Let $F_w^s = E_s(I_t^w)$ denote the encoder feature at stage s. For the selected stages $s \in \mathcal{S}_r = \{2, 3, 4\}$, multidirectional state propagation maps F_w^s to a spatially contextualized state B_w^s. A temporal cue Q^s is then combined with B_w^s to estimate a correction gate $\alpha^s \in (0,1)^{C_s\times H_s\times W_s}$, which controls how strongly each latent state is shifted toward a reference B_r^s. The corrected state and its encoder reintegration are

$$H^s = (1-\alpha^s)\odot B_w^s + \alpha^s \odot B_r^s, \qquad \widehat{F}_w^s = F_w^s + W_C^s(H^s). \tag{2}$$

Thus, α^s controls the retain-or-correct decision in SW-Mamba: small values preserve the propagated state, whereas large values progressively shift it toward the reference. The projection W_C^s is initialized to zero, so $\widehat{F}_w^s = F_w^s$ at initialization.

As shown in Fig. 2, the method separates three operations: constructing the contextual state B_w^s (Sec. 3.2), estimating the correction gate α^s to determine where and how strongly the state should be corrected (Sec. 3.3), and constructing the reference B_r^s that defines the correction target (Sec. 3.4). Sec. 3.5 then describes the training objectives used to calibrate the gate and reference proxy.

3.2 MULTIDIRECTIONAL STATE PROPAGATION

Selective correction requires a contextual state that relates each local response to its surrounding structure. For each selected stage $s \in \mathcal{S}_r$, the encoder feature $F_w^s \in \mathbb{R}^{C_s\times H_s\times W_s}$ is projected into a latent space,

$$X_w^s = \mathrm{GN}\left(W_B^s F_w^s\right), \tag{3}$$

where W_B^s is a pointwise projection and $\mathrm{GN}(\cdot)$ denotes group normalization.

We use a multidirectional selective SSM, denoted by $\mathcal{M}_s$, to aggregate spatial context. The projected state is first divided into a scan branch U^s and a companion branch Z^s,

$$(U^s, Z^s) = \mathrm{Split}\left(W_{\mathrm{in}}^s X_w^s\right), \qquad \widetilde{U}^s = \mathrm{SiLU}\left(\mathrm{DWConv}_{3\times 3}(U^s)\right). \tag{4}$$

The scan branch is serialized along $\mathcal{D} = \{\rightarrow, \leftarrow, \downarrow, \uparrow\}$. Let Π_d and Π_d^{-1} denote serialization and restoration along direction d, with $\mathcal{S}_{s,d}$ as the selective SSM. The multidirectional response is

$$\mathcal{M}_s(X_w^s) = W_O^s\left[\mathrm{GN}\left(\frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\Pi_d^{-1}\mathcal{S}_{s,d}\left(\Pi_d(\widetilde{U}^s)\right)\right)\odot \mathrm{SiLU}(Z^s)\right], \tag{5}$$

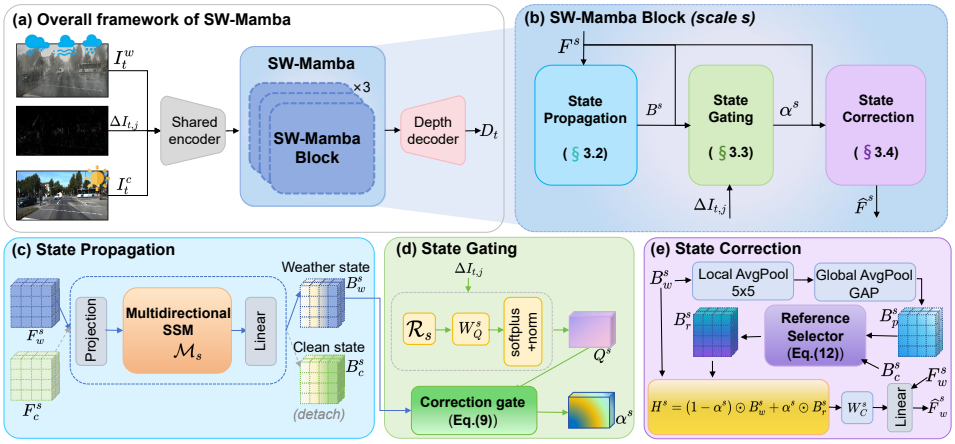

Figure 2: **Overview of SW-Mamba.** SW-Mamba refines encoder features via shared multidirectional state propagation, spatiotemporal state gating, and reference-guided state correction. The propagated context and temporal cue jointly determine correction strength through α^s, while B_r^s defines the correction target, using clean states during training and an internal proxy at inference.

where W_O^s projects the fused response back to C_s channels. Each $\mathcal{S}_{s,d}$ follows the standard selective SSM formulation with input-dependent discretization, state injection, and readout.

The contextual state is then defined as

$$B_w^s = X_w^s + \mathcal{M}_s(X_w^s). \tag{6}$$

The resulting B_w^s serves as the contextual state for the subsequent retain-or-correct decision. By aggregating directional spatial context, it allows local responses to be assessed in relation to their surrounding representation.

3.3 Spatiotemporal State Gating

The correction gate evaluates inter-frame variation against the propagated state B_w^s, yielding a context-dependent correction strength, as shown in Figure 3. For an adjacent frame I_j^w, we compute the absolute inter-frame difference

$$\Delta I_{t,j} = \left| I_t^w - I_j^w \right|. \tag{7}$$

For each selected stage $s \in \mathcal{S}_r$, $\Delta I_{t,j}$ is resized to the spatial resolution of B_w^s and projected into the corresponding channel space:

$$R^s = \mathrm{softplus}\left[W_Q^s \, \mathcal{R}_s(\Delta I_{t,j}) \right], \qquad Q^s = \frac{R^s - \mu_{HW}(R^s)}{\sqrt{\mathrm{Var}_{HW}(R^s) + \epsilon}}, \tag{8}$$

where $\mathcal{R}_s(\cdot)$ denotes spatial resampling, W_Q^s is a learnable 1×1 projection, and μ_{HW} and Var_{HW} are computed over the spatial dimensions of each channel.

The normalized cue is combined with the contextual state through a channel-wise correction gate,

$$\alpha^s = \sigma\left[\ell_s(B_w^s) + \beta_s Q^s \right], \qquad \beta_s = \mathrm{softplus}(\widetilde{\beta}_s) > 0, \tag{9}$$

where $\widetilde{\beta}_s$ is an unconstrained learnable scalar and $\ell_s(\cdot)$ is a channel-preserving 1×1 projection. Here, $\ell_s(B_w^s)$ provides the state-dependent gate logit, while Q^s modulates it according to relative inter-frame variation. The resulting α^s controls correction strength without treating frame difference as a hard weather mask or direct evidence of non-geometric corruption. Instead, temporal variation is interpreted jointly with the propagated state to produce a context-dependent correction gate.

3.4 Reference-Guided State Correction

Selective correction requires a reference state B_r^s that defines the direction of the latent update. Aligned clean observations provide this reference during training but are unavailable at inference.

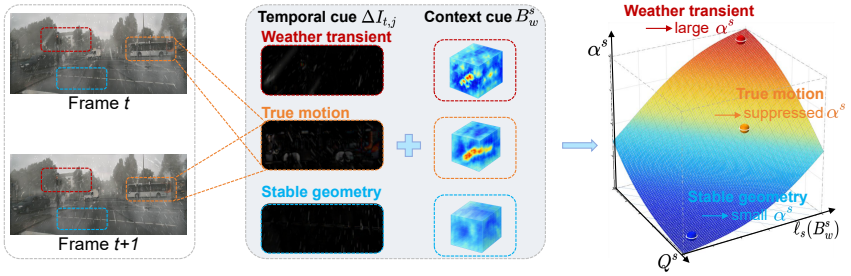

Figure 3: **Spatiotemporal state gating in SW-Mamba.** It maps the frame difference $\Delta I_{t,j}$ to the temporal cue Q^s and the propagated state B_w^s to contextual evidence $\ell_s(B_w^s)$, which jointly determine the correction strength α^s.

We therefore use the clean state as a training reference and construct an internal proxy from B_w^s for deployment.

When an aligned clean observation is available, its encoder feature F_c^s is mapped through the same state transformation as the adverse-weather feature:

$$X_c^s = \mathrm{GN}\left(W_B^s F_c^s\right), \qquad B_c^s = \mathrm{sg}\left[X_c^s + \mathcal{M}_s(X_c^s)\right], \tag{10}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The shared transformation places B_c^s and B_w^s in the same latent space, while detachment keeps the clean reference fixed.

Because B_c^s is unavailable at inference, we construct a deployment proxy directly from the propagated weather state:

$$B_p^s = \eta\ \mathrm{AvgPool}_{5\times5}(B_w^s) + (1-\eta)\ \mathrm{Broadcast}\left[\mathrm{GAP}(B_w^s)\right], \tag{11}$$

where $\eta = 0.75$. The local average preserves coarse spatial layout, whereas the global term supplies scene-level context. Their combination provides a structural reference without reconstructing the unavailable clean state.

Using only B_c^s during paired training would leave the correction process unexposed to the proxy used at inference. We therefore randomly replace the clean reference with the proxy during training. Let $\chi_c \in \{0,1\}$ indicate clean observation availability and $z_b^s \sim \mathrm{Bernoulli}(p_r)$, with $p_r = 0.25$, indicate proxy selection for sample b at stage s. The reference is defined as

$$B_{r,b}^s = \begin{cases} (1-z_b^s)B_{c,b}^s + z_b^s\ \mathrm{sg}(B_{p,b}^s), & \chi_c = 1, \\ \mathrm{sg}(B_{p,b}^s), & \chi_c = 0. \end{cases} \tag{12}$$

When $\chi_c = 0$, including at inference, Eq. (12) reduces to $B_r^s = \mathrm{sg}(B_p^s)$. Reference dropout therefore exposes the correction process to the deployment proxy while retaining clean-state guidance when available.

Given B_r^s, the corrected state and its reintegration into the encoder feature are

$$H^s = (1-\alpha^s) \odot B_w^s + \alpha^s \odot B_r^s, \qquad \widehat{F}_w^s = F_w^s + W_C^s(H^s), \tag{13}$$

where $\odot$ denotes elementwise multiplication. Thus, B_r^s defines the correction target, while α^s controls how strongly the propagated state is shifted toward that reference.

3.5 Learning Objective

SW-Mamba uses aligned clean states to calibrate two components: the correction gate and the deployment proxy. The former is supervised by latent disagreement between adverse-weather and clean representations, while the latter is aligned with a spatially smoothed version of the clean state.

Gate Calibration. Rather than deriving gate supervision from image difference, we measure the discrepancy between the contextual weather state and its aligned clean counterpart. For each selected stage $s \in \mathcal{S}_r$, the spatial discrepancy is

$$M_i^s = \frac{1}{C_s}\sum_{c=1}^{C_s}\left|\mathrm{sg}(B_{w,c,i}^s) - \mathrm{sg}(B_{c,c,i}^s)\right|, \tag{14}$$

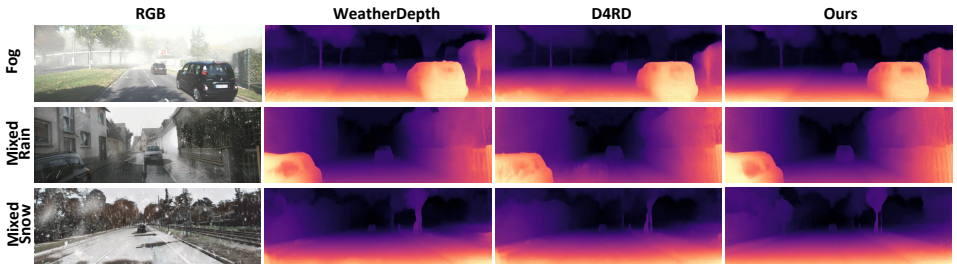

Figure 4: **Qualitative comparison on WeatherKITTI** under fog, mixed rain and snow.

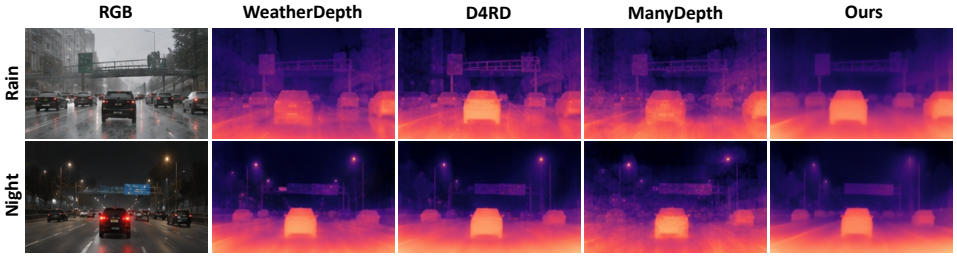

Figure 5: **Qualitative comparison on nuScenes** under rain and night.

where i indexes spatial locations. We use the spatial mean of M^s to define relative high- and low-discrepancy regions:

$$\tau^s = \frac{1}{H_s W_s} \sum_i M_i^s, \qquad \Omega_h^s = \{i \mid M_i^s > \tau^s\}, \qquad \Omega_l^s = \{i \mid M_i^s \le \tau^s\}. \tag{15}$$

Let $\bar{\alpha}_i^s = C_s^{-1} \sum_{c=1}^{C_s} \alpha_{c,i}^s$ denote the channel-averaged correction weight at location i. We calibrate the gate with a margin-based hinge loss,

$$\mathcal{L}_{\text{gate}} = \frac{1}{|\mathcal{S}_r|} \sum_{s \in \mathcal{S}_r} \left[\frac{\sum_{i \in \Omega_h^s} [\gamma_h - \bar{\alpha}_i^s]_+}{|\Omega_h^s| + 1} + \frac{\sum_{i \in \Omega_l^s} [\bar{\alpha}_i^s - \gamma_l]_+}{|\Omega_l^s| + 1} \right], \tag{16}$$

where $[x]_+ = \max(x, 0)$, $\gamma_h = 0.65$, and $\gamma_l = 0.25$. This objective favors stronger correction where the adverse-weather representation departs substantially from the clean state, while suppressing correction where the two remain close.

Proxy Alignment. The deployment proxy summarizes spatial structure through local and global averaging and is aligned with a spatially smoothed clean-state target:

$$\mathcal{L}_{\text{proxy}} = \frac{1}{|\mathcal{S}_r|} \sum_{s \in \mathcal{S}_r} \text{SmoothL1}\left(B_p^s,\, \text{AvgPool}_{5\times5}\left[\text{sg}(B_c^s)\right]\right). \tag{17}$$

This constraint matches the proxy to structural content it can represent without imposing high-frequency clean details that are unavailable from the proxy construction. Although $\text{sg}(B_p^s)$ is used as the correction reference in Sec. 3.4, $\mathcal{L}_{\text{proxy}}$ is evaluated on the undetached B_p^s, allowing the proxy representation to be optimized during training.

Overall Objective. Let $\mathcal{L}_{\text{base}}$ denote the base self-supervised objective, including photometric, smoothness, and contrastive terms when enabled. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \chi_c \left(\lambda_g \mathcal{L}_{\text{gate}} + \lambda_p \mathcal{L}_{\text{proxy}}\right), \tag{18}$$

where $\chi_c \in \{0, 1\}$ indicates aligned clean observation availability, with $\lambda_g = 0.05$ and $\lambda_p = 0.01$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SW-Mamba on three complementary benchmarks. **WeatherKITTI** is the primary benchmark and provides aligned clean–weather pairs for training-time calibration.

Table 1: **Quantitative comparison on WeatherKITTI.** Pretrained depth priors are reported separately because they do not share the self-supervised training setting. Best results are shown in **bold**.

Category	Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
General	Monodepth2 (Godard et al., 2019)	0.124	0.941	6.372	0.217	0.851	0.949	0.981
	PackNet-SfM (Guizilini et al., 2020)	0.151	1.468	5.908	0.228	0.752	0.921	0.969
	HR-Depth (Lyu et al., 2021)	0.143	1.313	5.548	0.215	0.775	0.932	0.973
	ManyDepth (Watson et al., 2021)	0.139	1.238	5.423	0.209	0.784	0.938	0.976
	MonoViT (Zhao et al., 2022)	0.120	0.899	5.111	0.200	0.857	0.953	0.980
	Lite-Mono (Zhang et al., 2023)	0.133	1.102	5.183	0.202	0.802	0.942	0.977
	PlaneDepth (Wang et al., 2023)	0.150	1.360	6.513	0.277	0.757	0.891	0.945
	SC-DepthV3 (Sun et al., 2024)	0.131	1.049	5.093	0.199	0.809	0.944	0.978
Adverse-weather	DeFeat-Net (Spencer et al., 2020)	0.171	1.821	6.508	0.251	0.704	0.895	0.957
	ADDS-DepthNet (Liu et al., 2021)	0.149	1.432	5.827	0.224	0.761	0.925	0.969
	RNW (Wang et al., 2021)	0.144	1.353	5.619	0.216	0.778	0.932	0.973
	WSGD (Vankadari et al., 2023)	0.136	1.168	5.282	0.205	0.801	0.941	0.977
	STEPS (Zheng et al., 2023)	0.130	1.062	5.089	0.198	0.817	0.947	0.979
	Robust-Depth (Saunders et al., 2023)	0.107	0.791	4.604	0.183	0.883	0.963	0.983
	md4all-AD (Gasperini et al., 2023)	0.121	0.922	4.819	0.189	0.842	0.953	0.981
	md4all-DD (Gasperini et al., 2023)	0.111	0.819	4.612	0.182	0.861	0.958	0.983
	EC-Depth (Song et al., 2023)	0.110	0.790	4.745	0.186	0.874	0.960	0.983
	WeatherDepth (Wang et al., 2024a)	0.103	0.738	4.414	0.178	0.892	0.965	0.984
	D4RD (Wang et al., 2024b)	0.099	0.688	4.377	0.174	0.897	0.966	0.984
	DMMDE-v3 (Tosi et al., 2024)	0.097	0.652	4.258	0.171	0.901	0.968	0.985
	ACDepth (Jiang et al., 2025)	0.095	0.628	4.192	0.168	0.904	0.969	0.986
	WeatherMono (Qiu, 2026)	0.094	0.621	4.148	0.167	0.906	0.971	0.986
Pretrained prior	Depth Anything V2 (Yang et al., 2024)	0.103	0.742	4.418	0.176	0.891	0.965	0.984
	UniDepth (Piccinelli et al., 2024)	0.098	0.678	4.313	0.172	0.898	0.967	**0.987**
	SW-Mamba	**0.091**	**0.589**	**4.078**	**0.163**	**0.912**	**0.972**	0.986

Table 2: **Condition-wise comparison on nuScenes.** Results are reported under the all-condition protocol for day-clear, night, and rain scenes. Best results are shown in **bold**.

Method	Day-clear			Night			Rain		
	Abs Rel↓	RMSE↓	δ_1 ↑	Abs Rel↓	RMSE↓	δ_1 ↑	Abs Rel↓	RMSE↓	δ_1 ↑
Monodepth2 (Godard et al., 2019)	0.137	6.692	0.850	0.283	9.729	0.518	0.173	7.743	0.776
PackNet-SfM (Guizilini et al., 2020)	0.157	7.230	0.826	0.262	11.063	0.566	0.165	8.288	0.771
RNW (Wang et al., 2021)	0.287	9.185	0.562	0.333	10.098	0.437	0.295	9.341	0.572
md4all-AD (Gasperini et al., 2023)	0.152	6.853	0.831	0.219	9.003	0.688	0.160	7.832	0.790
md4all-DD (Gasperini et al., 2023)	0.137	6.452	0.846	0.192	8.507	0.711	0.141	7.228	0.810
DMMDE-v3 (Tosi et al., 2024)	0.128	6.449	0.840	0.191	8.433	0.711	0.139	7.129	0.814
ACDepth (Jiang et al., 2025)	0.134	6.284	0.851	0.187	8.125	0.711	0.138	6.970	0.813
WeatherDepth (Wang et al., 2024a)	0.130	6.310	0.852	0.181	8.010	0.725	0.133	6.850	0.824
D4RD (Wang et al., 2024b)	0.128	6.270	0.855	0.178	7.910	0.733	0.129	6.740	0.832
SW-Mamba	**0.124**	**6.110**	**0.861**	**0.168**	**7.620**	**0.751**	**0.121**	**6.480**	**0.846**

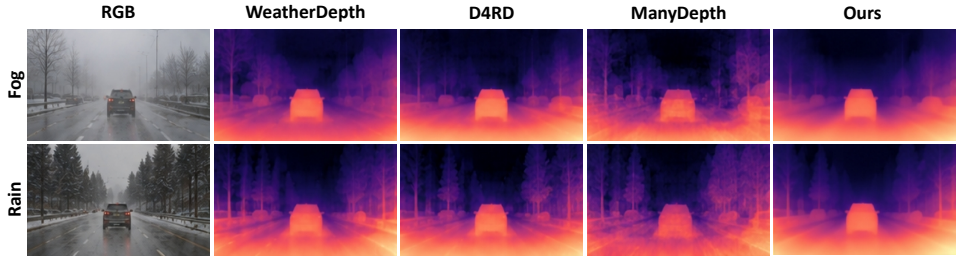

Figure 6: **Zero-shot qualitative comparison on DrivingStereo** under unseen fog and rain.

nuScenes evaluates real-world performance across day-clear, night, and rain conditions under the established all-condition protocol. **DrivingStereo-Rain/Fog** is reserved for zero-shot evaluation of the WeatherKITTI-trained model without target-domain adaptation.

Evaluation Metrics. We report standard metrics: Abs Rel, Sq Rel, RMSE, RMSE log, and threshold accuracies δ_1, δ_2, and δ_3, with lower error and higher accuracy indicating better performance.

Table 3: **Zero-shot transfer from WeatherKITTI to DrivingStereo.** Best results are shown in **bold**.

Condition	Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Rain	MonoViT (Zhao et al., 2022)	0.175	2.136	9.618	0.232	0.730	0.931	0.979
	PlaneDepth (Wang et al., 2023)	0.220	3.302	11.671	0.278	0.654	0.883	0.965
	md4all (Gasperini et al., 2023)	0.176	2.134	9.216	0.225	0.752	0.941	0.981
	WeatherDepth (Wang et al., 2024a)	0.158	1.833	8.837	0.211	0.764	0.945	0.985
	WeatherMono (Qiu, 2026)	0.149	1.824	8.634	0.202	0.775	0.951	0.988
	D4RD (Wang et al., 2024b)	0.145	1.760	8.480	0.198	0.787	0.954	0.988
	SW-Mamba	**0.137**	**1.620**	**8.210**	**0.190**	**0.802**	**0.960**	**0.990**
Fog	MonoViT (Zhao et al., 2022)	0.109	1.204	7.760	0.167	0.870	0.967	0.990
	PlaneDepth (Wang et al., 2023)	0.151	1.836	7.758	0.167	0.803	0.945	0.983
	md4all (Gasperini et al., 2023)	0.131	1.241	7.415	0.165	0.842	0.981	0.991
	WeatherDepth (Wang et al., 2024a)	0.110	1.195	7.323	0.160	0.878	0.973	0.992
	WeatherMono (Qiu, 2026)	0.101	1.112	7.265	0.144	0.881	0.988	0.992
	D4RD (Wang et al., 2024b)	0.099	1.080	7.180	0.145	0.886	0.979	0.993
	SW-Mamba	**0.094**	**1.020**	**7.020**	**0.139**	**0.895**	**0.982**	**0.994**

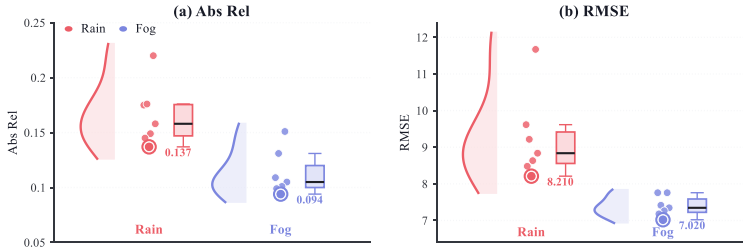

Figure 7: **Zero-shot error distributions on DrivingStereo.** Raincloud plots, combining density estimates, individual method results, and boxplots, summarize Abs Rel and RMSE under rain and fog, with SW-Mamba highlighted by the labeled markers.

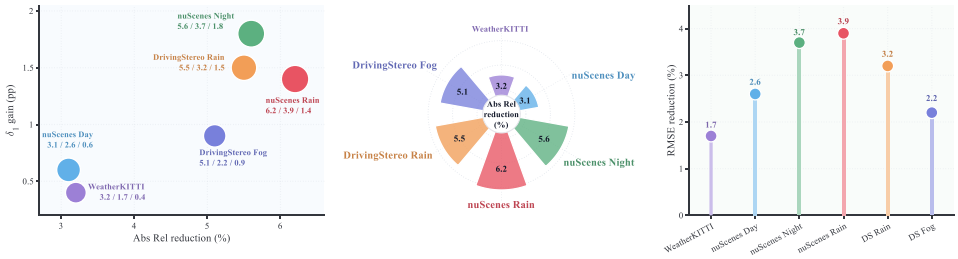

Figure 8: **Performance gains across evaluation settings.** Gains of SW-Mamba over WeatherMono on WeatherKITTI and D4RD on nuScenes and DrivingStereo in terms of Abs Rel, RMSE, and δ_1.

Table 4: **Component ablation on WeatherKITTI.** Best results are shown in **bold**.

Variant	State ctx.	Temporal gate	Ref. corr.	Deploy. proxy	Abs Rel↓
Base model	✗	✗	✗	✗	0.128
+ State propagation	✓	✗	✗	✗	0.117
+ Temporal gating	✓	✓	✗	✗	0.108
+ Reference correction	✓	✓	✓	✗	0.097
Full SW-Mamba	✓	✓	✓	✓	**0.091**

Implementation Details. SW-Mamba is built on the WeatherDepth baseline (Wang et al., 2024a), with a ResNet-18 pose network. On WeatherKITTI, models are trained at 640×192 for 30 epochs with batch size 8 using AdamW. The encoder and SW-Mamba blocks use a learning rate of 5×10^{-5}, while the depth decoder and pose network use 10^{-4}, with both learning rates decayed by a factor of 0.9 per epoch. For nuScenes, we train at 320×576 with a physical batch size of 4 and two-step gradient accumulation, yielding an effective batch size of 8, while retaining the same optimization schedule. Since aligned clean–weather pairs are unavailable on nuScenes, $\mathcal{L}_{\text{gate}}$ and $\mathcal{L}_{\text{proxy}}$ are disabled, while the photometric, smoothness, and contrastive objectives remain unchanged.

Table 5: **Ablation of reference design on WeatherKITTI.** Best results are shown in **bold**.

Inference reference	Ref. dropout	Proxy align.	Abs Rel↓	δ_1 ↑
Current state B_w^s	–	–	0.105	0.889
Global mean	✓	✓	0.100	0.896
Local average	✓	✓	0.096	0.903
Local + global proxy	✗	✓	0.094	0.906
Local + global proxy	✓	✗	0.093	0.908
Local + global proxy	✓	✓	**0.091**	**0.912**

Table 6: **Efficiency comparison on WeatherKITTI.** Best results are shown in **bold**.

Method	Params (M)	FLOPs (G)	Latency (ms)	FPS	Abs Rel↓
WeatherDepth baseline	31.60	51.8	23.8	42.0	0.128
+ CNN adapter	31.78	53.6	25.1	39.8	0.116
+ Transformer adapter	32.41	55.9	31.4	31.8	0.109
+ Mamba adapter	31.94	54.1	26.7	37.5	0.103
SW-Mamba	31.82	54.0	25.9	38.6	**0.091**

4.2 Performance across Conditions and Domains

Adverse-Weather Depth Estimation. On WeatherKITTI, SW-Mamba achieves the best performance on most metrics (Table 1), reducing Abs Rel, Sq Rel, RMSE, and RMSE log over WeatherMono by 3.2%, 5.2%, 1.7%, and 2.4%, respectively. On nuScenes, it performs best under day-clear, night, and rain conditions (Table 2), with Abs Rel reductions of 3.1%, 5.6%, and 6.2% over D4RD. Figures 4 and 5 further show improved boundary and structural preservation.

Zero-Shot Generalization. Without target-domain adaptation, SW-Mamba performs best on DrivingStereo under both rain and fog (Table 3), reducing the four error metrics over D4RD by 3.2–8.0% and 2.2–5.6%, respectively. Figure 6 shows consistent improvements in boundary and depth structure under these unseen conditions.

Cross-Condition Analysis. Figure 7 places DrivingStereo results within the distribution of all compared methods. SW-Mamba lies at the low-error end for both Abs Rel and RMSE under rain and fog, indicating consistent zero-shot gains across the two unseen conditions. Figure 8 summarizes the relative improvements across all evaluated settings, with the gains increasing under more challenging conditions. This shows SW-Mamba maintains its advantage as visual degradation worsens.

4.3 Ablation and Efficiency Analysis

Ablation Studies. Table 4 evaluates the main components of SW-Mamba. State propagation improves the base model, while temporal gating and reference correction provide further gains. The deployment proxy completes the full model and yields the best performance. Table 5 further examines the reference design. Local averaging outperforms global averaging, while combining both provides a stronger deployment reference. Removing either reference dropout or proxy alignment degrades performance, confirming that both contribute to effective proxy learning.

Efficiency Analysis. As shown in Table 6, SW-Mamba adds only 0.7% parameters and 4.2% FLOPs over WeatherDepth while retaining 38.6 FPS. Compared with a Mamba adapter of similar complexity, it achieves substantially lower Abs Rel, showing gains are not from added size or computation.

5 Conclusion

This paper proposed SW-Mamba, which formulates adverse-weather self-supervised monocular depth estimation as selective latent-state correction. Multidirectional state propagation establishes spatial context, while spatiotemporal state gating combines this context with temporal variation to estimate correction strength. Reference-guided state correction then determines the update target through a structural reference, using aligned clean states during training and an internal structural proxy at inference. Experiments across synthetic adverse weather, real-world conditions, and zero-shot transfer demonstrate consistent improvements in depth estimation and cross-domain robustness.

AI USE STATEMENT

Generative AI tools were used for language editing and figure layout prototyping during manuscript preparation. They were not used as a source of experimental measurements. All technical content, citations, experimental results, and final figures were verified by the authors.

ETHICS STATEMENT

This work studies monocular depth estimation under adverse driving conditions using existing public driving benchmarks. It does not involve new human-subject experiments, collection of personal data, or the release of a new dataset. The proposed method is intended for research on robust visual perception and does not by itself constitute a safety-critical driving system. Although improved depth estimation under degraded visibility may benefit autonomous perception, performance can still be affected by unseen environments, severe image degradation, and distribution shifts. The reported results should therefore not be interpreted as a guarantee of safe deployment, and any use in safety-critical applications requires appropriate system-level validation and safeguards. We are not aware of additional privacy, fairness, or security concerns specific to the proposed methodology.

REPRODUCIBILITY STATEMENT

The main paper specifies the model formulation, learning objective, and evaluation protocol, and distinguishes paired clean-reference training from reference-free deployment. Appendix B provides additional analysis of state propagation, spatiotemporal gating, reference-guided correction, and clean-to-proxy transfer. Appendix C details the training and inference procedure, while Appendix D reports the architectural configuration and optimization settings. Appendices E–F document the dataset roles, evaluation protocols, and additional experiments. Together, these materials provide the methodological and experimental details required for reproduction.

REFERENCES

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF conference on computer vision and pattern recognition*, pp. 11618–11628. IEEE, 2020.

Wonhyeok Choi, Kyumin Hwang, Wei Peng, Minwoo Choi, and Sunghoon Im. Self-supervised monocular depth estimation robust to reflective surface leveraged by triplet mining. In *International Conference on Learning Representations*, volume 2025, pp. 9587–9605, 2025.

Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust monocular depth estimation under challenging conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8177–8186, 2023.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3828–3838, 2019.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020. doi: 10.1109/CVPR42600.2020.00256.

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. In *European conference on computer vision*, pp. 12–22. Springer, 2024.

Kui Jiang, Jing Cao, Zhaocheng Yu, Junjun Jiang, and Jingchun Zhou. Always clear depth: Robust monocular depth estimation under adverse weather. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1251–1259, 2025. doi: 10.24963/ijcai.2025/140.

Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Self-supervised monocular depth estimation for all day images using domain separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12717–12726, 2021. doi: 10.1109/ICCV48922.2021.01250.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems*, 2024.

Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2294–2301, 2021. doi: 10.1609/aaai.v35i3.16329.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Yongsheng Qiu. Weathermono: A cnn-transformer architecture for self-supervised monocular depth estimation in rainy and foggy conditions. *Sensors*, 26(5):1705, 2026. doi: 10.3390/s26051705.

Kieran Saunders, George Vogiatzis, and Luis J. Manso. Self-supervised monocular depth estimation: Let's talk about the weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8907–8917, 2023.

Ziyang Song, Ruijie Zhu, Chuxin Wang, Jiacheng Deng, Jianfeng He, and Tianzhu Zhang. Ec-depth: Exploring the consistency of self-supervised monocular depth estimation in challenging scenes. *arXiv preprint arXiv:2310.08044*, 2023.

Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14402–14413, 2020. doi: 10.1109/CVPR42600.2020.01441.

Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):497–508, 2024. doi: 10.1109/TPAMI.2023.3322549.

Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *European Conference on Computer Vision*, pp. 236–257, 2024.

Madhu Vankadari, Stuart Golodetz, Sourav Garg, Sangyun Shin, Andrew Markham, and Niki Trigoni. When the sun goes down: Repairing photometric losses for all-day depth estimation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 1992–2003, 2023.

Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions. In *2024 IEEE International Conference on Robotics and Automation*, pp. 4976–4982, 2024a. doi: 10.1109/ICRA57147.2024.10611100.

Jiyuan Wang, Chunyu Lin, Lang Nie, Kang Liao, Shuwei Shao, and Yao Zhao. Digging into contrastive learning for robust depth estimation with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024b. doi: 10.1145/3664647.3681168.

Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16055–16064, 2021. doi: 10.1109/ICCV48922.2021.01575.

Ruoyu Wang, Zehao Yu, and Shenghua Gao. Planedepth: Self-supervised depth estimation via orthogonal planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21425–21434, 2023.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1164–1174, 2021.

Chaodong Xiao, Minghan Li, Deyu Meng, Lei Zhang, et al. Spatial-mamba: Effective visual state space models via structure-aware state fusion. In *International Conference on Learning Representations*, volume 2025, pp. 73777–73795, 2025.

Weilong Yan, Ming Li, Haipeng Li, Shuwei Shao, and Robby T. Tan. Synthetic-to-real self-supervised robust depth estimation via learning with motion and structure priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21880–21890, 2025.

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*, 2019.

Lihe Yang, Zhen Zhao, Bingyi Kang, Xiaogang Xu, Zilong Huang, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024.

Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18537–18546, 2023.

Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision*, 2022.

Yupeng Zheng, Chengliang Zhong, Pengfei Li, Huan-ang Gao, Yuhang Zheng, Bu Jin, Ling Wang, Hao Zhao, Guyue Zhou, Qichao Zhang, and Dongbin Zhao. Steps: Joint self-supervised nighttime image enhancement and depth estimation. In *2023 IEEE International Conference on Robotics and Automation*, pp. 4916–4923, 2023. doi: 10.1109/ICRA48891.2023.10160708.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A Appendix Overview

This appendix complements the main paper with additional theoretical analysis, procedural details, reproducibility information, and supporting experiments for SW-Mamba. Section B analyzes the stability of directional state propagation, contextual modulation of the temporal cue, reference-guided correction, and clean-to-proxy transfer. Section C gives the complete training and inference procedure, including the training-only clean reference and deployment-time proxy. Section D specifies the architectural configuration and optimization settings, while Section E details the datasets and evaluation protocols used throughout the experiments. Section F provides additional evidence through benchmark and condition-wise analysis, clear-condition preservation, compound adverse-condition evaluation, and design-choice ablations.

B Additional Theoretical Analysis

B.1 Notation and Scope

We follow the notation of Secs. 3.2–3.4. At a encoder stage $s \in \mathcal{S}_r$, B_w^s denotes the propagated weather state, Q^s the normalized temporal cue, and $\alpha^s \in (0,1)^{C_s \times H_s \times W_s}$ the correction gate. The

reference B_r^s is obtained from the detached clean state B_c^s when available or from the internal proxy B_p^s. State correction is

$$H^s = (1 - \alpha^s) \odot B_w^s + \alpha^s \odot B_r^s, \qquad \widehat{F}_w^s = F_w^s + W_C^s(H^s), \tag{19}$$

which is identical to Eq. equation 13.

The analysis below separates three questions. First, we characterize how an isolated perturbation propagates through one selective scan direction. Second, we analyze how the normalized temporal cue affects the correction gate when contextual evidence is fixed. Third, we establish conditions under which reference-guided correction reduces latent error and quantify the discrepancy introduced when the clean training reference is replaced by the deployment proxy.

B.2 Stability of Directional State Propagation

Consider one scan direction $d \in \mathcal{D}$ at stage s. For one serialized position k, the implemented selective SSM can be written in the form

$$h_{d,k}^s = \mathcal{T}_{d,k}^s h_{d,k-1}^s + \iota_{d,k}^s, \qquad \mathcal{T}_{d,k}^s = \exp\left(\Delta_{d,k}^s A_d^s\right), \tag{20}$$

where A_d^s is diagonal with strictly negative entries and $\Delta_{d,k}^s > 0$ is the input-dependent discretization step. The term $\iota_{d,k}^s$ collects the corresponding input-dependent state injection.

Proposition 1 (Bounded propagation of an isolated state perturbation)**.** *Suppose two directional recurrences receive identical inputs and therefore identical input-dependent scan parameters after position q, but their hidden states differ by an additive perturbation ε_q at q. Then, for any $k > q$,*

$$\|\varepsilon_{q\to k}\|_2 \le \left(\prod_{j=q+1}^{k} \rho_{d,j}^s\right) \|\varepsilon_q\|_2, \qquad \rho_{d,j}^s = \|\mathcal{T}_{d,j}^s\|_2 < 1. \tag{21}$$

Proof. Let $\delta h_{d,j}^s$ denote the difference between the two hidden states. Since the subsequent inputs and scan parameters are identical, the injection terms cancel and

$$\delta h_{d,j}^s = \mathcal{T}_{d,j}^s \delta h_{d,j-1}^s, \qquad \delta h_{d,q}^s = \varepsilon_q. \tag{22}$$

Iteration yields

$$\varepsilon_{q\to k} = \left(\prod_{j=q+1}^{k} \mathcal{T}_{d,j}^s\right) \varepsilon_q. \tag{23}$$

For every diagonal component $a_{d,\ell}^s < 0$,

$$0 < \exp\left(\Delta_{d,j}^s a_{d,\ell}^s\right) < 1 \tag{24}$$

because $\Delta_{d,j}^s > 0$. Hence $\rho_{d,j}^s = \|\mathcal{T}_{d,j}^s\|_2 < 1$. Applying submultiplicativity of the induced ℓ_2 norm gives Eq. equation 21. □

B.3 Contextual Modulation of the Temporal Cue

The main paper defines

$$\alpha^s = \sigma\left[\ell_s(B_w^s) + \beta_s Q^s\right], \qquad \beta_s > 0, \tag{25}$$

where Q^s is obtained by spatially standardizing the positive response R^s in Eq. equation 8.

Fix one channel-spatial coordinate and write

$$\alpha(q) = \sigma(z + \beta_s q), \qquad z = [\ell_s(B_w^s)]_i, \qquad q = [Q^s]_i. \tag{26}$$

Proposition 2 (Monotone and bounded temporal modulation)**.** *For fixed contextual state B_w^s,*

$$\frac{\partial \alpha}{\partial q} = \beta_s \alpha(1 - \alpha), \qquad 0 < \frac{\partial \alpha}{\partial q} \le \frac{\beta_s}{4}. \tag{27}$$

Thus, if $Q_1^s \preceq Q_2^s$ elementwise while B_w^s is fixed,

$$\alpha^s(Q_1^s) \preceq \alpha^s(Q_2^s). \tag{28}$$

Proof. The logistic sigmoid is strictly increasing. Differentiating Eq. equation 26 gives

$$\frac{\partial \alpha}{\partial q} = \beta_s \sigma(z + \beta_s q)\left[1 - \sigma(z + \beta_s q)\right]. \tag{29}$$

Since $\beta_s > 0$ and $\alpha(1-\alpha) \le 1/4$ for $\alpha \in (0,1)$, Eq. equation 27 follows. □

Lemma 1 (Contextual suppression condition)**.** *For any normalized cue $q_i^s \in \mathbb{R}$ and any desired upper gate level $\tau \in (0,1)$,*

$$\alpha_i^s \le \tau \quad \Longleftrightarrow \quad [\ell_s(B_w^s)]_i \le \mathrm{logit}(\tau) - \beta_s q_i^s, \tag{30}$$

where

$$\mathrm{logit}(\tau) = \log \frac{\tau}{1-\tau}. \tag{31}$$

Proof. Since $\sigma(\cdot)$ is strictly increasing, $\alpha_i^s \le \tau$ if and only if

$$[\ell_s(B_w^s)]_i + \beta_s q_i^s \le \mathrm{logit}(\tau), \tag{32}$$

which directly yields Eq. equation 30. □

For a positive normalized temporal response, increasing q_i^s therefore requires stronger opposing contextual evidence to maintain the same upper correction level. This makes explicit why the temporal cue is not a weather detector. Motion, parallax, occlusion, or illumination variation can increase temporal evidence, while the propagated state determines whether that evidence results in strong correction.

B.4 Error Reduction under Reference-Guided Correction

To analyze the correction step, let B_*^s denote an ideal latent target used only for analysis. We use the Euclidean norm after vectorization, equivalently the Frobenius norm on the corresponding latent tensor. Define

$$e_w^s = \mathrm{vec}(B_w^s - B_*^s), \qquad \delta_r^s = \mathrm{vec}(B_r^s - B_w^s), \tag{33}$$

and

$$D_{\alpha^s} = \mathrm{Diag}\left(\mathrm{vec}(\alpha^s)\right). \tag{34}$$

Equation equation 19 gives

$$\mathrm{vec}(H^s - B_*^s) = e_w^s + D_{\alpha^s}\delta_r^s. \tag{35}$$

Therefore,

$$\|H^s - B_*^s\|_2^2 - \|B_w^s - B_*^s\|_2^2 = 2\left\langle e_w^s, D_{\alpha^s}\delta_r^s\right\rangle + \|D_{\alpha^s}\delta_r^s\|_2^2. \tag{36}$$

Proposition 3 (Exact condition for latent error reduction)**.** *Reference-guided correction strictly reduces the squared latent error if and only if*

$$2\left\langle e_w^s, D_{\alpha^s}\delta_r^s\right\rangle + \|D_{\alpha^s}\delta_r^s\|_2^2 < 0. \tag{37}$$

For a nonzero correction $D_{\alpha^s}\delta_r^s$, a sufficient condition is

$$\left\langle e_w^s, D_{\alpha^s}\delta_r^s\right\rangle \le -\kappa\|D_{\alpha^s}\delta_r^s\|_2^2, \qquad \kappa > \frac{1}{2}. \tag{38}$$

Proof. Equation equation 36 follows by expanding $\|e_w^s + D_{\alpha^s}\delta_r^s\|_2^2$. Hence Eq. equation 37 is necessary and sufficient for strict error reduction. Under Eq. equation 38,

$$2\left\langle e_w^s, D_{\alpha^s}\delta_r^s\right\rangle + \|D_{\alpha^s}\delta_r^s\|_2^2 \le (1-2\kappa)\|D_{\alpha^s}\delta_r^s\|_2^2 < 0, \tag{39}$$

because $\kappa > 1/2$ and $D_{\alpha^s}\delta_r^s \ne 0$. □

Proposition 4 (Contraction under a locally more accurate reference)**.** *Let Ω^s denote a set of latent coordinates satisfying*

$$\left|B_{r,i}^s - B_{*,i}^s\right| \le \rho_s \left|B_{w,i}^s - B_{*,i}^s\right|, \qquad 0 \le \rho_s < 1. \tag{40}$$

If

$$\alpha_i^s \ge a_s > 0, \qquad i \in \Omega^s, \tag{41}$$

then

$$\|(H^s - B_*^s)_{\Omega^s}\|_2 \le \left[1 - a_s(1-\rho_s)\right]\|(B_w^s - B_*^s)_{\Omega^s}\|_2. \tag{42}$$

Proof. For each $i \in \Omega^s$,

$$\begin{aligned}\left|H_i^s - B_{*,i}^s\right| &= \left|(1-\alpha_i^s)(B_{w,i}^s - B_{*,i}^s) + \alpha_i^s(B_{r,i}^s - B_{*,i}^s)\right| \\ &\le (1-\alpha_i^s)\left|B_{w,i}^s - B_{*,i}^s\right| + \alpha_i^s\left|B_{r,i}^s - B_{*,i}^s\right| \\ &\le [1-\alpha_i^s(1-\rho_s)]\left|B_{w,i}^s - B_{*,i}^s\right| \\ &\le [1-a_s(1-\rho_s)]\left|B_{w,i}^s - B_{*,i}^s\right|. \end{aligned} \tag{43}$$

Squaring and summing over Ω^s gives Eq. equation 42. □

Variational interpretation. For fixed α^s, Eq. equation 19 is the unique minimizer of

$$\mathcal{E}(H) = \left\|\sqrt{1-\alpha^s} \odot (H - B_w^s)\right\|_2^2 + \left\|\sqrt{\alpha^s} \odot (H - B_r^s)\right\|_2^2. \tag{44}$$

Indeed, for one coordinate with $a \in [0,1]$,

$$E(h) = (1-a)(h-b_w)^2 + a(h-b_r)^2, \tag{45}$$

and

$$\frac{\partial E}{\partial h} = 2h - 2\left[(1-a)b_w + ab_r\right], \qquad \frac{\partial^2 E}{\partial h^2} = 2. \tag{46}$$

Thus

$$h^* = (1-a)b_w + ab_r. \tag{47}$$

This gives α^s a coordinate-wise interpolation interpretation.

B.5 Clean-to-Proxy Transfer

The reference used for paired training differs from the reference available at deployment. To isolate this change, consider two counterfactual corrected states obtained from the same B_w^s and α^s:

$$\begin{aligned} H_c^s &= (1-\alpha^s) \odot B_w^s + \alpha^s \odot B_c^s, \\ H_p^s &= (1-\alpha^s) \odot B_w^s + \alpha^s \odot B_p^s. \end{aligned} \tag{48}$$

Proposition 5 (Bounded clean-to-proxy transfer)**.** *The latent-state discrepancy induced by replacing the clean reference with the deployment proxy satisfies*

$$\|H_p^s - H_c^s\|_2 \le \|\alpha^s\|_\infty \|B_p^s - B_c^s\|_2. \tag{49}$$

The corresponding encoder-feature discrepancy satisfies

$$\|\widehat{F}_p^s - \widehat{F}_c^s\|_2 \le \|W_C^s\|_{2\to 2}\|\alpha^s\|_\infty \|B_p^s - B_c^s\|_2. \tag{50}$$

Proof. Subtracting the two states in Eq. equation 48 gives

$$H_p^s - H_c^s = \alpha^s \odot (B_p^s - B_c^s). \tag{51}$$

Hence

$$\begin{aligned}\|H_p^s - H_c^s\|_2 &= \left\|D_{\alpha^s}\,\mathrm{vec}(B_p^s - B_c^s)\right\|_2 \\ &\le \|D_{\alpha^s}\|_2 \|B_p^s - B_c^s\|_2 \\ &= \|\alpha^s\|_\infty \|B_p^s - B_c^s\|_2. \end{aligned} \tag{52}$$

Since the reintegration map W_C^s is linear,

$$\widehat{F}_p^s - \widehat{F}_c^s = W_C^s(H_p^s - H_c^s), \tag{53}$$

and Eq. equation 50 follows from its induced operator norm. □

The proxy-alignment loss in Eq. equation 17 does not ask B_p^s to reproduce the complete clean state. Define the low-frequency clean target

$$\overline{B}_c^s = \mathrm{AvgPool}_{5\times 5}(B_c^s). \tag{54}$$

The triangle inequality gives

$$\|B_p^s - B_c^s\|_2 \le \|B_p^s - \overline{B}_c^s\|_2 + \|\overline{B}_c^s - B_c^s\|_2. \tag{55}$$

Combining Eqs. equation 50 and equation 55,

$$\|\widehat{F}_p^s - \widehat{F}_c^s\|_2 \le \|W_C^s\|_{2\to 2}\|\alpha^s\|_\infty \times \left(\|B_p^s - \overline{B}_c^s\|_2 + \|\overline{B}_c^s - B_c^s\|_2\right). \tag{56}$$

The first term measures the discrepancy between the internal proxy and the low-frequency clean target and is directly addressed by the SmoothL1 objective in Eq. equation 17. The second contains clean-state variation removed by the low-pass operator and is not directly optimized by that objective. This decomposition makes explicit that proxy alignment reduces the representable part of the clean-to-proxy discrepancy without requiring reconstruction of unavailable high-frequency clean information.

Corollary 1 (Conditional decoder-level transfer bound)**.** *Let $\widehat{\mathbf{F}}_c$ and $\widehat{\mathbf{F}}_p$ denote the complete encoder feature tuples obtained with clean and proxy references, respectively, while all unrefined features remain identical. If the depth decoder $\mathcal{D}$ is locally L_D-Lipschitz under the concatenated feature norm, then*

$$\left\|\mathcal{D}(\widehat{\mathbf{F}}_p) - \mathcal{D}(\widehat{\mathbf{F}}_c)\right\|_2 \le L_D \left[\sum_{s\in\mathcal{S}_r} \|W_C^s\|_{2\to 2}^2 \|\alpha^s\|_\infty^2 \|B_p^s - B_c^s\|_2^2\right]^{1/2}. \tag{57}$$

B.6 Reference Dropout and Conservative Reintegration

Reference dropout in Eq. equation 12 exposes the correction branch to its deployment reference while clean supervision remains available. For sample b and stage s, let

$$z_b^s \sim \mathrm{Bernoulli}(p_r), \qquad p_r = 0.25, \tag{58}$$

where $z_b^s = 1$ selects the proxy. Conditioned on $B_{c,b}^s$ and $B_{p,b}^s$,

$$\mathbb{E}\left[B_{r,b}^s \mid B_{c,b}^s, B_{p,b}^s\right] = (1-p_r)B_{c,b}^s + p_r B_{p,b}^s. \tag{59}$$

Reference dropout therefore does not remove the clean-to-proxy shift in expectation. Instead, it exposes the correction branch to the proxy during optimization. Proxy alignment plays a different role by reducing the discrepancy between B_p^s and the low-frequency clean target. When $\chi_c = 0$, the proxy is used directly and the clean-dependent auxiliary losses are inactive.

The output projection W_C^s is initialized to zero. Consequently,

$$W_C^s = 0 \quad\Longrightarrow\quad \widehat{F}_w^s = F_w^s \qquad \text{at initialization.} \tag{60}$$

Thus the inserted correction branch is an exact forward identity at the start of optimization.

Overall, the analysis establishes properties that match the scope of the implemented SW-Mamba mechanism. The selective recurrent transition attenuates an isolated hidden-state perturbation under fixed subsequent inputs. The normalized temporal cue modulates the correction gate monotonically while contextual evidence can suppress its effect. Reference-guided interpolation reduces latent error under explicit local conditions rather than by assumption. Finally, the clean-to-proxy discrepancy is filtered by both the correction gate and the reintegration map, while proxy alignment, reference dropout, and zero-initialized reintegration respectively address low-frequency reference matching, deployment exposure, and conservative insertion.

C Training and Inference Procedure

SW-Mamba follows the same weather-feature pathway during training and inference, while the use of the reference state depends on whether an aligned clean observation is available. During paired training, the clean observation provides a stop-gradient state for gate calibration and reference construction. The internal proxy remains active during training through reference dropout and becomes the sole reference at inference. Algorithm 1 summarizes the complete procedure.

Here, $\mathrm{sg}(\cdot)$ denotes stop-gradient, and $\chi = 1$ only when an aligned clean observation is available. The base objective $\mathcal{L}_{\mathrm{base}}$ contains the original self-supervised photometric and smoothness terms, together with the contrastive term when enabled. Reference dropout is applied independently to each sample with $p_r = 0.25$, exposing the correction branch to the proxy reference during paired training.

Algorithm 1 Training and inference procedure of SW-Mamba.

Require: Weather target I_t^w, adjacent frame I_j^w, optional aligned clean target I_t^c
Ensure: Predicted depth D_t

$\{F_w^s\}_{s=1}^S \leftarrow E(I_t^w)$
$\Delta I_{t,j} \leftarrow \mathrm{sg}\big(|I_t^w - I_j^w|\big)$
if I_t^c is available **then**
  $\{F_c^s\}_{s=1}^S \leftarrow \mathrm{sg}(E(I_t^c))$
end if
for $s \in \mathcal{S}_r$ **do**
  $X_w^s \leftarrow \mathcal{P}_s(F_w^s)$
  $B_w^s \leftarrow X_w^s + \mathcal{M}_s(X_w^s)$
  $R^s \leftarrow \mathrm{softplus}\big(W_Q^s \mathcal{R}_s(\Delta I_{t,j})\big)$
  $Q^s \leftarrow \dfrac{R^s - \mu_{HW}(R^s)}{\sqrt{\mathrm{Var}_{HW}(R^s) + \epsilon}}$
  $\alpha^s \leftarrow \sigma[\ell_s(B_w^s) + \beta_s Q^s]$
  $B_p^s \leftarrow 0.75\ \mathrm{AvgPool}_{5\times5}(B_w^s) + 0.25\ \mathrm{Broadcast}(\mathrm{GAP}(B_w^s))$
  if I_t^c is available **then**
    $X_c^s \leftarrow \mathcal{P}_s(F_c^s)$
    $B_c^s \leftarrow \mathrm{sg}[X_c^s + \mathcal{M}_s(X_c^s)]$
    $M^s \leftarrow \mathrm{Mean}_C\left(|\mathrm{sg}(B_w^s) - B_c^s|\right)$
    $\tau^s \leftarrow \mu_{HW}(M^s)$
    accumulate $\mathcal{L}_{\mathrm{gate}}$ from (α^s, M^s, τ^s)
    accumulate $\mathcal{L}_{\mathrm{proxy}}$ from $\big(B_p^s, \mathrm{AvgPool}_{5\times5}(B_c^s)\big)$
    $m^s \sim \mathrm{Bernoulli}(1 - p_r)$
    $B_r^s \leftarrow m^s B_c^s + (1 - m^s)\,\mathrm{sg}(B_p^s)$
  else
    $B_r^s \leftarrow \mathrm{sg}(B_p^s)$
  end if
  $H^s \leftarrow (1 - \alpha^s) \odot B_w^s + \alpha^s \odot B_r^s$
  $\widehat{F}_w^s \leftarrow F_w^s + W_C^s(H^s)$
end for
$D_t \leftarrow f_{\mathrm{dec}}\left(\{\widehat{F}_w^s\}_{s\in\mathcal{S}_r}, \{F_w^s\}_{s\notin\mathcal{S}_r}\right)$
if training **then**
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{base}} + \chi\left(\lambda_g \mathcal{L}_{\mathrm{gate}} + \lambda_p \mathcal{L}_{\mathrm{proxy}}\right)$
  return $D_t, \mathcal{L}$
else
  return D_t
end if

D Implementation Details

D.1 Architecture Configuration

SW-Mamba augments the WeatherDepth encoder by introducing selective state refinement at encoder stages $\mathcal{S}_r = \{2, 3, 4\}$, while the two higher-resolution stages are passed directly to the decoder. The depth decoder follows HR-Depth, and relative camera motion is estimated by a ResNet-18 pose network. This placement assigns the state-space refinement to deeper feature scales, where broader spatial context can be modeled at moderate spatial resolution while preserving shallow spatial detail. The main architectural settings are summarized in Table 7.

Each refined stage uses an independent SW-Mamba block with state dimension $d_{\mathrm{state}} = 16$, discretization rank $d_{\mathrm{rank}} = 8$, expansion ratio 1, and a 3×3 depthwise convolution. The input feature is first mapped by a channel-preserving 1×1 projection followed by group normalization. The number of groups is chosen as the largest divisor of the channel dimension not exceeding eight. Spatial state propagation is performed along four traversal directions, left-to-right, right-to-left, top-to-bottom, and bottom-to-top. The four directional outputs are restored to the original spatial layout and averaged

Table 7: **SW-Mamba architecture configuration.**

Configuration	Setting
Refined encoder stages	$\{2, 3, 4\}$
Scan directions	LR, RL, TB, BT
State dimension d_{state}	16
Discretization rank d_{rank}	8
Expansion ratio	1.0
Depthwise convolution	3×3
State projection W_B^s	1×1
Temporal projection W_Q^s	1×1
Gate-logit projection ℓ_s	1×1
Correction projection W_C^s	1×1
Initial temporal coefficient β_s	2.0
Proxy kernel	5×5
Proxy local/global weights	$0.75/0.25$
Reference dropout p_r	0.25

before output normalization, multiplicative branch fusion, and projection back to the encoder channel space.

The spatiotemporal gate operates at the same channel resolution as the propagated state. The inter-frame RGB difference is bilinearly resized to each selected stage and projected from three channels to the corresponding feature dimension by a 1×1 convolution. After Softplus activation, the resulting cue is standardized independently over the spatial dimensions of each channel. The state-dependent gate logit is produced by a separate channel-preserving 1×1 projection. The positive temporal coefficient is parameterized through Softplus and initialized to $\beta_s = 2.0$.

The deployment proxy uses a 5×5 local average and the spatial global mean of the propagated state with weights 0.75 and 0.25, respectively. Reference dropout is set to $p_r = 0.25$ during paired training.

D.2 Optimization Details

On WeatherKITTI, we train at a resolution of 640×192 for 30 epochs with batch size 8. Optimization uses AdamW with two learning-rate groups. The encoder wrapper, including the inserted SW-Mamba blocks, uses a learning rate of 5×10^{-5}, while the depth decoder and pose network use 1×10^{-4}. Both learning rates are multiplied by 0.9 after each epoch.

The base self-supervised objective follows the WeatherDepth training configuration. Photometric reconstruction combines SSIM and ℓ_1 errors with weights 0.85 and 0.15, respectively. Temporal source frames at offsets $\{-1, +1\}$ are used for self-supervised view synthesis, while the SW-Mamba temporal cue is computed from the augmentation-matched target frame and the preceding frame. Edge-aware smoothness is weighted by $10^{-3}2^{-s}$ at scale s. For the paired WeatherKITTI setting, the contrastive term is weighted by $\lambda_{\text{contrast}} = 0.02$. The SW-Mamba auxiliary weights are $\lambda_g = 0.05$ and $\lambda_p = 0.01$, with gate margins $\gamma_h = 0.65$ and $\gamma_l = 0.25$.

For nuScenes, training is performed at 320×576. We use a physical batch size of 4 and accumulate gradients over two iterations, giving an effective batch size of 8, while retaining the same optimization schedule. Because nuScenes does not provide aligned clean–weather observations, $\mathcal{L}_{\text{gate}}$ and $\mathcal{L}_{\text{proxy}}$ are disabled. The model is therefore optimized using the base self-supervised objective, while state correction uses the internal proxy reference. DrivingStereo is used only for zero-shot evaluation and introduces no additional optimization or target-domain adaptation.

E Datasets and Evaluation Protocols

We evaluate SW-Mamba on three complementary driving benchmarks with distinct roles in training and evaluation. WeatherKITTI provides aligned clean–weather observations for paired calibration,

Table 8: **Datasets and experimental roles.** A clean anchor denotes an aligned clean observation of the same scene that can be used for the auxiliary gate and proxy calibration.

Dataset	Conditions used	Clean anchor	Training use	Evaluation role
WeatherKITTI	Rain, snow, fog, mixed weather	Yes	Paired self-supervision and calibration	In-domain adverse weather
nuScenes	Day-clear, night, rain	No	All-condition self-supervision	Real-world robustness
DrivingStereo	Rain, fog	No	None	Zero-shot transfer

nuScenes evaluates robustness under real-world adverse conditions without clean anchors, and DrivingStereo measures zero-shot transfer to unseen weather data. Table 8 summarizes these settings.

E.1 WeatherKITTI

WeatherKITTI (Wang et al., 2024a) is our primary benchmark for adverse-weather training and evaluation. It extends KITTI driving sequences with weather-corrupted observations while retaining the corresponding clean RGB frames, providing the alignment required by the clean-state calibration in Sec. 3.5. We follow the standard monocular Eigen protocol used by WeatherDepth, with weather variants constructed from the same underlying KITTI frames. The benchmark includes rain, snow, mixed rain and snow, and fog with different visibility levels, together with the corresponding clean observations. Its test partition follows the standard 697-image Eigen test set.

During paired training, the weather observation is used by the main self-supervised depth pathway, while its aligned clean counterpart is used only to construct the detached clean state B_c^s and the auxiliary calibration targets. The clean image is therefore not required by the depth decoder and is not available to the model at inference. All main component ablations and reference-design analyses are conducted on this benchmark because it provides the clean–weather correspondence needed to isolate the proposed gate and reference mechanisms.

E.2 nuScenes

We use nuScenes (Caesar et al., 2020) to evaluate SW-Mamba under naturally occurring visibility and illumination changes. Following the all-condition protocol of md4all (Gasperini et al., 2023), the data are grouped into day-clear, night, and rain conditions. The adopted split contains 15,129 training images and 6,019 validation images, with the validation set comprising 4,449 day-clear, 602 night, and 1,088 rain samples. The night subset includes nighttime scenes with and without rain, following the established protocol.

Unlike WeatherKITTI, nuScenes does not provide temporally and geometrically aligned clean counterparts for adverse-condition frames. We therefore do not construct pseudo clean–weather pairs. Accordingly, $\mathcal{L}_{\text{gate}}$ and $\mathcal{L}_{\text{proxy}}$ are disabled on nuScenes and optimization reduces to the base self-supervised objective. State correction remains active and uses the internal proxy B_p^s, matching the reference path available at deployment. This setting evaluates whether the correction mechanism learned without paired clean anchors remains effective on real-world adverse conditions.

E.3 DrivingStereo

DrivingStereo (Yang et al., 2019) provides high-resolution stereo imagery and corresponding depth maps under multiple driving conditions. The released weather subset contains sunny, cloudy, foggy, and rainy scenes, with 500 frames for each condition. Following prior adverse-weather depth evaluation protocols, we use only the rainy and foggy subsets.

DrivingStereo is used exclusively for zero-shot evaluation. The model trained on WeatherKITTI is transferred directly to DrivingStereo without fine-tuning, adaptation, clean-reference calibration, or any use of target-domain images during optimization. The rainy and foggy subsets therefore test whether the latent-state correction learned from WeatherKITTI transfers across changes in scene distribution, image formation, and weather appearance. At inference, SW-Mamba uses only the observed image sequence and its internal low-frequency reference proxy.

E.4 Evaluation Protocol

Following standard MDE evaluation protocols, we report Abs Rel, Sq Rel, RMSE, RMSE log, and the threshold accuracies δ_1, δ_2, and δ_3:

$$\mathrm{AbsRel} = \frac{1}{N}\sum_i \frac{|D_i - \widehat{D}_i|}{D_i}, \tag{61}$$

$$\mathrm{SqRel} = \frac{1}{N}\sum_i \frac{(D_i - \widehat{D}_i)^2}{D_i}, \tag{62}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i (D_i - \widehat{D}_i)^2}, \tag{63}$$

$$\mathrm{RMSE}_{\log} = \sqrt{\frac{1}{N}\sum_i (\log D_i - \log \widehat{D}_i)^2}, \tag{64}$$

$$\delta_k = \frac{1}{N}\sum_i \mathbf{1}\left[\max\left(\frac{D_i}{\widehat{D}_i}, \frac{\widehat{D}_i}{D_i}\right) < 1.25^k\right], \qquad k \in \{1,2,3\}, \tag{65}$$

where D and $\widehat{D}$ denote ground-truth and predicted depth, respectively.

For re-evaluated baselines, we use the same image preprocessing, valid-depth mask, depth range, and scale-alignment rule as SW-Mamba within each dataset. Published numbers are used only when the corresponding work follows the same split and evaluation convention. Depth Anything V2 and UniDepth are reported separately from supervision-matched self-supervised methods because their pretraining and supervision settings differ.

F Additional Experiments

This section complements the main experimental results with four analyses. We first reorganize the reported benchmark measurements to examine method ranking on WeatherKITTI and condition-wise behavior on nuScenes, without introducing additional measurements. We then evaluate clear-condition preservation, robustness under compound adverse conditions, and selected low-level design choices of SW-Mamba.

F.1 Benchmark and Condition-wise Analysis

We further reorganize the main quantitative results to examine the relative position of SW-Mamba across competing methods and environmental conditions. Figure 9 provides a compact view of the WeatherKITTI comparison across conventional, adverse-weather, and pretrained depth models. SW-Mamba attains an Abs Rel of 0.091 and an RMSE of 4.078, compared with 0.094 and 4.148 for WeatherMono. This corresponds to relative reductions of 3.2% in Abs Rel and 1.7% in RMSE. The same ordering across both metrics supports the consistency of the improvement. Figure 10 further shows how the performance gap changes with environmental condition. Relative to D4RD, the Abs Rel reduction of SW-Mamba increases from 3.1% under day-clear conditions to 5.6% at night and 6.2% in rain. The corresponding δ_1 gains are 0.006, 0.018, and 0.014. The advantage is therefore retained in clear scenes and becomes more pronounced under degraded visibility, consistent with the intended role of selective state correction when visual evidence becomes less reliable.

F.2 Clear-Condition Preservation

A weather-robust depth estimator should preserve geometric accuracy when the input is already reliable. Table 9 compares SW-Mamba with the same depth backbone without selective state correction on the clear WeatherKITTI subset.

SW-Mamba preserves the clear-condition performance of the underlying depth model. Abs Rel changes from 0.102 to 0.101 and RMSE from 4.68 to 4.64, while δ_1 and δ_2 increase by 0.007 and 0.003, respectively. These changes are substantially smaller than those observed under adverse

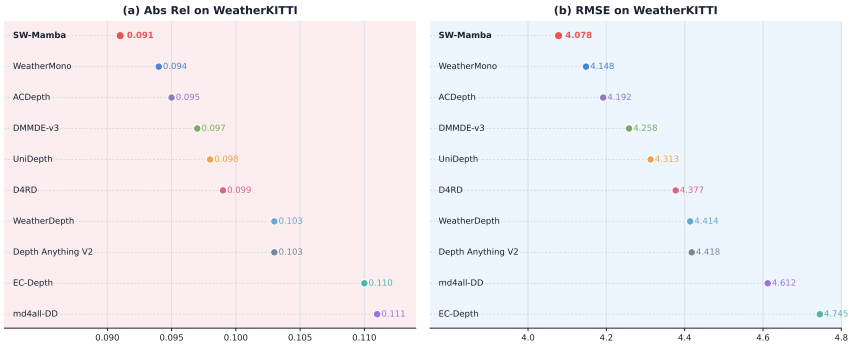

Figure 9: **Ranked comparison on WeatherKITTI.** Abs Rel and RMSE of representative competing methods using the quantitative results reported in the main paper. Lower values are better for both metrics.

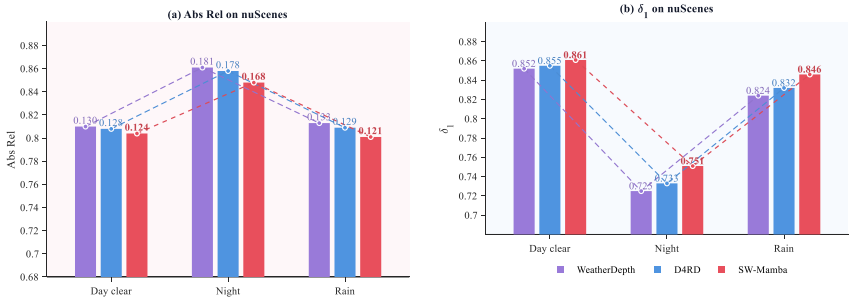

Figure 10: **Condition-wise comparison on nuScenes.** Abs Rel and δ_1 under day-clear, night, and rain conditions using the quantitative results reported in the main paper. Lower Abs Rel and higher δ_1 indicate better performance.

Table 9: **Quantitative comparison under clear conditions on WeatherKITTI.** Base denotes the same depth backbone without SW-Mamba. Best results are shown in **bold**.

Method	Abs Rel↓	RMSE↓	δ_1 ↑	δ_2 ↑
Base	0.102	4.68	0.884	0.961
SW-Mamba	**0.101**	**4.64**	**0.891**	**0.964**

conditions, indicating that selective state correction does not degrade performance when the visual evidence is reliable.

F.3 Compound Adverse Conditions

The main experiments evaluate individual adverse conditions, while practical driving scenes can exhibit multiple degradations simultaneously. We therefore evaluate four compound settings on WeatherKITTI. Table 10 and Figure 11 report the measured Abs Rel, with absolute reduction defined as $E_{\text{base}} - E_{\text{SW}}$ and relative reduction as $(E_{\text{base}} - E_{\text{SW}})/E_{\text{base}}$.

SW-Mamba improves over the paired base model in all four compound settings, with relative Abs Rel reductions ranging from 10.2% to 12.3% and an average reduction of 11.1%. Rain + Fog yields the largest relative reduction at 12.3%, whereas Snow + Blur gives the largest absolute reduction of 0.028. The improvement therefore remains consistent across the tested combinations of rain, fog, snow, low illumination, and blur. These results show that the measured benefit is not confined to a single degradation combination.

Table 10: **Compound adverse-condition evaluation on WeatherKITTI.** Best results are shown in **bold**.

Condition	Base	SW-Mamba	Abs. red.	Rel. red.
Rain + Fog	0.212	**0.186**	0.026	12.3%
Night + Rain	0.231	**0.207**	0.024	10.4%
Snow + Blur	0.244	**0.216**	0.028	11.5%
Fog + Low Light	0.226	**0.203**	0.023	10.2%

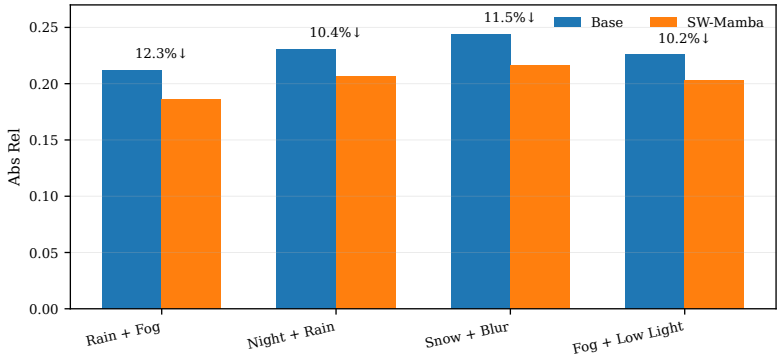

Figure 11: **Performance under compound adverse conditions.** Paired Abs Rel results for the four compound settings in Table 10.

Table 11: **Design choice ablations on WeatherKITTI.** Best results are shown in **bold**.

	Abs Rel↓			RMSE↓			$\Delta\delta_1$ ↑
Design choice	w/o	with	Rel. red.	w/o	with	Rel. red.	
Gate alignment	0.121	**0.116**	4.1%	5.067	**4.882**	3.7%	+0.008
Zero-initialized reintegration	0.120	**0.116**	3.3%	4.973	**4.882**	1.8%	+0.005
Multiscale refinement	0.123	**0.116**	5.7%	5.141	**4.882**	5.0%	+0.014

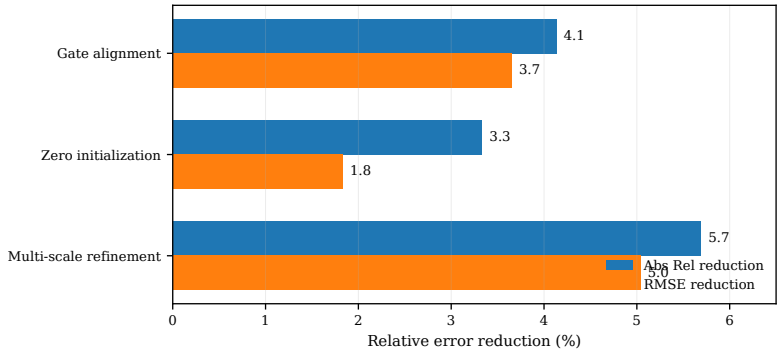

Figure 12: **Effects of design choices on WeatherKITTI.** Relative reductions in Abs Rel and RMSE are computed from the paired comparisons in Table 11.

F.4 Design Choice Ablations

The main paper examines the core components of SW-Mamba. We further isolate three design choices that affect the calibration and integration of selective correction within the depth encoder: gate alignment, zero-initialized reintegration, and multiscale refinement.

As shown in Table 11 and Figure 12, all three design choices yield favorable changes in the ablation setting. Multiscale refinement produces the largest measured effect, reducing Abs Rel by 5.7% and RMSE by 5.0%, together with a 0.014 increase in δ_1. Gate alignment reduces Abs Rel and RMSE by 4.1% and 3.7%, respectively. Zero-initialized reintegration has a smaller effect on the final error, with reductions of 3.3% in Abs Rel and 1.8% in RMSE. This is consistent with its intended role as a conservative initialization strategy rather than a primary source of representational capacity.