CONDITIONAL RESIDUAL DECOMPOSITION FOR DISTRIBUTED IMAGE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

In distributed image compression (DIC), the compressed target representation and decoder-only side information (SI) jointly determine the reconstruction. However, existing DIC methods do not explicitly characterize the residual that remains after the compressed target representation and SI have both been used. As a result, they do not distinguish the content that can be inferred from SI from the remaining residual that cannot be determined by the available conditions, motivating generative modeling. To achieve this separation, we apply a conditional-expectation projection to the residual left by the compressed target representation, isolating the component predictable from SI from an orthogonal remainder after joint decoding. This orthogonal decomposition yields a variance partition in which the explained fraction defines SI reliability, while the remaining variance determines the minimum MSE of deterministic reconstruction. Building on this decomposition, we propose a DIC method with Conditional Residual Decomposition (CRDIC), which estimates how much the utilized SI reduces the residual left by the compressed target representation and converts this reduction into a spatial estimate of the residual remaining after joint decoding. The deterministic branch uses the SI-predictable residual component to refine the conditional reconstruction, while a stochastic branch models the remaining residual with an amplitude determined by its estimated scale. An antithetic odd projection further suppresses noise-independent corrections in the stochastic branch. Experiments on KITTI Stereo, Cityscapes, and InStereo2K demonstrate superior reconstruction fidelity and perceptual quality across multiple bitrates.

1 INTRODUCTION

Multi-camera systems used for autonomous driving, three-dimensional perception, and distributed sensing continuously produce correlated image streams (Geiger et al., 2012; Matsuyama & Ukita, 2002). Transmitting every view independently repeats shared scene information and increases the communication load. Distributed image compression (DIC) addresses this setting by encoding a target image independently and reconstructing it with correlated side information (SI) already available at the decoder (Mital et al., 2022; 2023). It follows the separate-encoding and joint-decoding principles of distributed source coding (Slepian & Wolf, 1973; Wyner & Ziv, 1976).

The key objective of DIC is to use decoder-only SI to recover target information that is not retained in the transmitted representation. Accordingly, existing deterministic DIC methods focus on extracting and aligning correlated information from SI to complement the decoded target representation (Mital et al., 2022; 2023; Huang et al., 2023). For a given transmitted representation, the missing target information can be viewed as a reconstruction residual that cannot be inferred from the bitstream alone. As illustrated in Fig. 1(a) and (b), decoder-only SI can further explain part of this residual U, while a remaining component E persists after joint decoding. This residual partition determines both the contribution of SI to deterministic reconstruction and the uncertainty that remains after the available conditions have been used. As shown in Fig. 1(b), this remaining uncertainty reflects not only pixel-level reconstruction error, but also potential perceptual gain.

Generative image compression methods employ learned priors to reconstruct plausible details that are weakly constrained by the compressed representations, providing improved perceptual quality at extremely low bitrates (Agustsson et al., 2023; Mao et al., 2024; Gao et al., 2025). However, percep-

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.

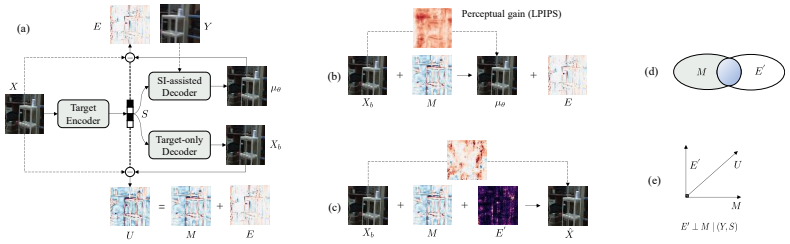

Figure 1: Conditional residual decomposition in DIC. (a) The target representation leaves the residual U, which is decomposed as $U = M + E$ after SI-assisted deterministic reconstruction. (b) The deterministic path incorporates M while E remains unresolved. (c) Stochastic reconstruction models the remaining variation through E'. (d) Unconstrained stochastic correction may overlap with deterministically recoverable content. (e) The proposed decomposition separates the deterministic component from the orthogonal residual modeled stochastically.

tually plausible reconstruction does not necessarily preserve fidelity to a particular source realization, as characterized by the perception-distortion (P-D) trade-off (Yan et al., 2022). Several studies have therefore introduced constrained generation (Chang et al., 2022; Mao et al., 2024). Recent generative DIC methods (Xu et al., 2026a;b) further incorporate SI to constrain the reconstruction. However, as illustrated in Fig. 1(c) and (d), without explicitly separating the residual that can be inferred from the available conditions from the component that requires stochastic reconstruction, the generative process may also modify content that could otherwise be reconstructed deterministically. Distinguishing the condition-supported residual from the remaining stochastic uncertainty is therefore necessary to preserve reconstruction fidelity while modeling unresolved image details.

To address these challenges, we formulate DIC through a conditional residual decomposition. Given the transmitted representation, the target-only MMSE residual is further decomposed by conditional expectation into an SI-predictable component and an orthogonal remainder, as illustrated in Fig. 1(e). The resulting variance decomposition defines SI reliability by the fraction of residual variance it explains, while the remaining variance determines stochastic residual scale. Controlling the fraction of this remaining variance restored then provides a controllable mechanism for the P-D tradeoff.

Based on this decomposition, we develop Conditional Residual Decomposition for DIC (CRDIC), which assigns the SI-predictable component to deterministic reconstruction and the remaining residual to stochastic modeling. Multi-scale correspondence and gated fusion incorporate SI into the conditional mean, while the estimated reliability determines the remaining residual scale. The stochastic branch is constrained by this scale, and an antithetic odd projection suppresses noise-independent shifts of the conditional mean. The amount of stochastic residual restoration can be continuously adjusted at inference without retraining.

The main contributions are summarized as follows:

- We formulate a conditional residual decomposition for DIC that separates the target-only residual into an SI-predictable component and an orthogonal remainder, linking SI reliability to deterministic reconstruction error and remaining uncertainty.
- Based on this decomposition, we develop a Conditional Residual Decomposition strategy for DIC (CRDIC), which assigns the SI-supported component to deterministic reconstruction and models the remaining residual with reliability-controlled stochastic reconstruction.
- Experiments on KITTI Stereo, Cityscapes, and InStereo2K validate the proposed decomposition and demonstrate superior reconstruction performance across multiple bitrates.

2 Related Work

2.1 Distributed Image Compression

Distributed source coding enables separate encoding and joint decoding of correlated sources. Slepian–Wolf established the lossless setting (Slepian & Wolf, 1973), while Wyner–Ziv extended

it to lossy coding with decoder-only SI (Wyner & Ziv, 1976). Learning-based DIC follows this principle by encoding the target independently and exploiting decoder-only SI.

Existing methods primarily improve how SI contributes to target reconstruction. DSIN (Ayzik & Avidan, 2020) extracts correlated features from the reference image, while NDIC (Mital et al., 2022) separates common and source-specific information to facilitate joint decoding. ATN (Mital et al., 2023) introduces cross-attention to align the decoded target representation with SI, and MSFDPM (Huang et al., 2023) further addresses spatial displacement through multi-scale patch matching. NDSC (Whang et al., 2024) formulates neural distributed source coding with discrete latent representations. LDMIC (Zhang et al., 2023) employs a symmetric multi-view DIC paradigm with a joint context module. Related stereo codecs, including SASIC (Wödlinger et al., 2022), ECSIC (Wödlinger et al., 2024), BiSIC (Liu et al., 2024), and CAMSIC (Zhang et al., 2025), also exploit inter-image correlation, although their encoder-access assumptions differ from those of the DIC method. These approaches improve the use of correlated information, but do not explicitly separate the residual predictable from SI from the uncertainty remaining after joint reconstruction.

2.2 Generative Image Compression

Generative image compression replaces deterministic reconstruction with a conditional distribution. Invertible generative models (Gao et al., 2023), conditional generators (Agustsson et al., 2023), and diffusion-based approaches (Li et al., 2025; Bordin & Maugey, 2025; Vonderfecht & Liu, 2025) use stochastic reconstruction to represent image details that are not uniquely determined by the compressed representation. By avoiding a single conditional-mean solution and restoring local image statistics, these methods can reduce the over-smoothing commonly observed at extremely low bitrates. Recent generative DIC methods (Xu et al., 2026a;b) further incorporate decoder-side correlated information to constrain the generated reconstruction.

Despite this progress, stochastic reconstruction is generally optimized through perceptual and distributional objectives without explicitly separating condition-supported content from the residual uncertainty after joint decoding. In DIC, this residual depends directly on the correlation between the target and SI, making its explicit characterization important for stochastic reconstruction.

2.3 Conditional Perception-Distortion Reconstruction

Blau and Michaeli formalized the perception-distortion trade-off and connected it to lossy compression through the rate-distortion-perception (R-D-P) formulation. (Blau & Michaeli, 2019). Chen *et al.* (Chen et al., 2022) further investigated the role of randomness in R-D-P coding and characterized conditions under which deterministic codes are sufficient. For Gaussian vector sources, Qian *et al.* (Qian et al., 2025) derived the R-D-P trade-off under distributional perception constraints.

SI further extends R-D-P analysis to correlated reconstruction settings. Niu *et al.* (Niu et al., 2023) considered shared SI and common randomness, while Hamdi and Gündüz (Hamdi & Gündüz, 2023) studied decoder-only SI under a perfect-realism constraint. In contrast to these information-theoretic proposals, our work characterizes how SI partitions the target residual into deterministic and stochastic components.

3 Methodology

3.1 Problem Formulation

Let $X \in [0,1]^{H\times W\times 3}$ denote the target image and Y a correlated side image. The target encoder observes only X and produces a transmitted representation T with expected bitrate R. At the decoder, an SI representation $S = h_y(Y)$ is extracted from Y, and the reconstruction is

$$\hat{X} \sim p_\theta(\hat{X} \mid T, S). \tag{1}$$

We characterize reconstruction uncertainty through the conditional residual. With only T, this residual is the MMSE error unexplained by the transmitted representation. After incorporating S, part of it becomes predictable from (T, S), while the remainder persists after joint conditioning.

3.2 Conditional Residual Decomposition

3.2.1 Target-only Reconstruction Residual

Given T, the MMSE reconstruction and its residual are

$$m_T = \mathbb{E}[X \mid T], \qquad U = X - m_T, \tag{2}$$

with

$$V_T = \mathbb{E}[U^2 \mid T] = \mathrm{Var}(X \mid T). \tag{3}$$

where V_T denotes the residual variance left by the transmitted representation. Because the information retained in T depends on the bitrate R, V_T can also vary with R.

3.2.2 Residual Reduction by Side Information

Conditioning further on S decomposes U as

$$\begin{aligned} M &= \mathbb{E}[U \mid T, S] = \mathbb{E}[X \mid T, S] - \mathbb{E}[X \mid T], \\ E &= U - M = X - \mathbb{E}[X \mid T, S], \end{aligned} \tag{4}$$

where M is the residual component predictable from SI, while E remains after joint conditioning. Since $\mathbb{E}[E \mid T, S] = 0$ and M is determined by (T, S),

$$\mathbb{E}[ME \mid T] = \mathbb{E}[M\,\mathbb{E}[E \mid T, S] \mid T] = 0. \tag{5}$$

The resulting orthogonality gives

$$V_S = \mathbb{E}[M^2 \mid T],\ V_{T,S} = \mathbb{E}[E^2 \mid T] = \mathbb{E}\left[\mathrm{Var}(X \mid T, S) \mid T\right],\ V_T = V_S + V_{T,S}. \tag{6}$$

Therefore, V_S is the residual variance explained by SI and $V_{T,S}$ is the variance remaining after joint conditioning. We define SI reliability as

$$\rho^2 = \frac{V_S}{V_T} = 1 - \frac{V_{T,S}}{V_T}, \qquad 0 \le \rho \le 1. \tag{7}$$

where ρ^2 measures the fraction of the target-only residual variance reduced by SI.

3.2.3 Remaining Residual and Deterministic Reconstruction

After incorporating SI, the conditional mean is

$$\mu = m_T + M = \mathbb{E}[X \mid T, S]. \tag{8}$$

Its expected squared error at fixed T is

$$D^*_{\mathrm{det}} = \mathbb{E}[(X - \mu)^2 \mid T] = V_{T,S} = (1 - \rho^2)V_T. \tag{9}$$

Thus, the remaining variance determines both the deterministic reconstruction error and the residual scale to be modeled stochastically.

3.2.4 Controlled Stochastic Reconstruction

To model the remaining residual, let E' be a conditionally independent sample from the distribution of E given (T, S). We define

$$\hat{X} = \mu + qE' = m_T + M + qE', \qquad q \in [0, 1], \tag{10}$$

where q^2 controls the fraction of remaining residual variance restored by the stochastic sample. At $q = 0$, $\hat{X} = \mu$, while $q = 1$ recovers the conditional residual distribution under the ideal model.

Let $\hat{U} = \hat{X} - m_T$. Using the orthogonality above and the conditional statistics of E', the reconstructed residual satisfies

$$\beta^2 := \frac{\mathbb{E}[\hat{U}^2 \mid T]}{V_T} = \rho^2 + q^2(1 - \rho^2), \tag{11}$$

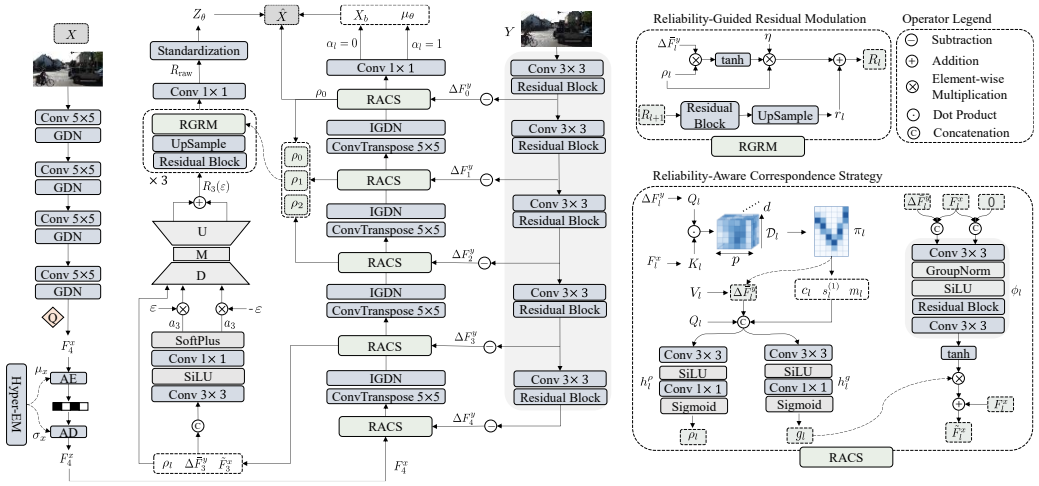

Figure 2: Overall architecture of CRDIC. The target branch produces X_b and multi-scale features from the transmitted representation. Decoder-only SI is aligned and fused through the Reliability-Aware Correspondence Strategy (RACS), which also estimates reliability ρ_l for the deterministic reconstruction. The resulting reliability determines the remaining residual scale σ_c. Reliability-Guided Residual Modulation (RGRM) then models the stochastic residual, whose amplitude is controlled by $q(\rho_0, \tau)\sigma_c$ to produce $\hat{X}$.

where $\beta \in [\rho, 1]$ is the standard-deviation ratio between $\hat{U}$ and the target-only residual U. Since $X - \hat{X} = E - qE'$,

$$D(\rho, q) = (1 + q^2)V_{T,S} = D^*_{\text{det}} + q^2 V_{T,S}. \tag{12}$$

We control this restoration by

$$\beta(\rho, \tau) = \rho + \tau(1 - \rho), \qquad \tau \in [0, 1], \tag{13}$$

which, together with Eq. (11), gives

$$q^2(\rho, \tau) = \frac{\beta^2 - \rho^2}{1 - \rho^2} = \frac{\tau[2\rho + \tau(1 - \rho)]}{1 + \rho}. \tag{14}$$

Thus, $\tau = 0$ gives deterministic reconstruction and $\tau = 1$ restores the full remaining residual variance under the model calibration.

3.3 The proposed CRDIC framework

3.3.1 Overview

As shown in Fig. 2, CRDIC implements the conditional residual decomposition through deterministic and stochastic reconstruction. The target codec first produces the base reconstruction X_b and multi-scale target features from the transmitted representation. Decoder-only SI is then aligned with the target features to estimate the SI-supported correction and its reliability. The resulting deterministic reconstruction μ_θ approximates the conditional mean, while the estimated reliability determines the scale of the residual that remains for stochastic reconstruction. The stochastic branch models this residual and produces the final reconstruction $\hat{X}$. Network details are provided in the Appendix.

3.3.2 SI-assisted Deterministic Reconstruction

The target codec follows a mean-scale hyperprior and produces a base reconstruction X_b together with multi-scale features $\{F_l^x\}_{l=0}^{L-1}$. Since X_b depends only on the transmitted representation, it corresponds to the target-only estimate m_T in Sec. 3.2.

Given decoder-only SI Y, the side encoder extracts multi-scale features, which are aligned with the target features by the Reliability-Aware Correspondence Strategy (RACS). RACS estimates the correspondence and predicts a spatial fusion gate g_l and reliability map ρ_l. The aligned SI features are incorporated into the target decoder, yielding

$$\mu_\theta = X_b + g_0 \odot \Delta_\mu, \tag{15}$$

where Δ_μ denotes the SI-induced correction. Therefore, $g_0 \odot \Delta_\mu$ corresponds to the predictable residual component M, and μ_θ approximates $\mathbb{E}[X \mid T, S]$.

The reliability target follows the residual reduction in Eq. (7). Let $D^b = \mathcal{A}(\|X - X_b\|_2^2/3)$ and $D^c = \mathcal{A}(\|X - \mu_\theta\|_2^2/3)$ denote locally averaged reconstruction errors. We define:

$$\rho^*(p) = \sqrt{\left[\frac{D^b(p) - D^c(p)}{D^b(p) + \epsilon_0}\right]_+}, \tag{16}$$

$$\mathcal{L}_\rho = \frac{1}{L}\sum_{l=0}^{L-1} \|\rho_l - \mathrm{sg}[\mathcal{R}_l(\rho^*)]\|_2^2, \tag{17}$$

where $\mathcal{R}_l$ resizes ρ^* to the l-th feature scale and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Ground truth is used only to construct ρ^* during training, while ρ_l is predicted from decoder-available features at inference. Details of RACS and multi-scale fusion are provided in the Appendix.

3.3.3 Remaining Residual Reconstruction

After obtaining μ_θ, we estimate the scale of the remaining residual. A target-only reference scale $\sigma_0 = h_\sigma(X_b)$ is first predicted from the base reconstruction. Following Eq. (9), we define

$$\sigma_c = \sigma_0\sqrt{1 - \rho_0^2}, \tag{18}$$

so that stronger SI reliability reduces the residual scale available for stochastic reconstruction. During training, σ_c is calibrated to the residual $r_c = X - \mu_\theta$ using a Gaussian negative log-likelihood:

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2}\mathbb{E}\left[\log(\sigma_c^2) + \frac{r_c^2}{\sigma_c^2}\right], \tag{19}$$

The stochastic branch models the residual structure around the fixed deterministic reconstruction. To avoid a noise-independent shift of μ_θ, we retain the odd component of the stochastic mapping using antithetic noise:

$$R(\varepsilon) = \frac{1}{2}\left[h_s(\varepsilon) - h_s(-\varepsilon)\right], \tag{20}$$

which gives $\mathbb{E}_\varepsilon[R(\varepsilon) \mid T, S] = 0$ for symmetric Gaussian noise. Reliability-Guided Residual Modulation (RGRM) then incorporates the aligned SI features and reliability into the residual representation. The stochastic structure is denoted by Z_θ, and the final reconstruction is

$$\hat{X} = \mu_\theta + q(\rho_0, \tau) \odot \sigma_c \odot Z_\theta. \tag{21}$$

Thus, σ_c specifies the estimated remaining scale, while $q(\rho_0, \tau)$ controls the residual restoration at inference. The detailed noise modulation and RGRM operations are given in the Appendix.

3.4 Optimization

Training follows the decomposition above in three stages:

$$\mathcal{L}_{\mathrm{target}} = \lambda_R R + \mathcal{L}_{\mathrm{D}}(X, X_b), \tag{22}$$

$$\mathcal{L}_{\mathrm{det}} = \lambda_D \mathcal{L}_{\mathrm{D}}(X, \mu_\theta) + \lambda_\rho \mathcal{L}_\rho + \lambda_N \mathcal{L}_{\mathrm{NLL}}, \tag{23}$$

$$\mathcal{L}_{\mathrm{stoch}} = \lambda_D^s \mathcal{L}_{\mathrm{D}}(X, \hat{X}) + \lambda_P^s \mathcal{L}_{\mathrm{P}}(X, \hat{X}). \tag{24}$$

We first optimize the target codec to obtain X_b. The target codec is then fixed while SI-assisted deterministic reconstruction, reliability prediction, and residual-scale estimation are optimized. Finally, the deterministic model is fixed and only the stochastic reconstruction is trained, with $\tau \sim \mathcal{U}(0, \tau_{\mathrm{train}})$. We use MSE for $\mathcal{L}_{\mathrm{D}}$ and pretrained VGG-LPIPS (Zhang et al., 2018) for $\mathcal{L}_{\mathrm{P}}$.

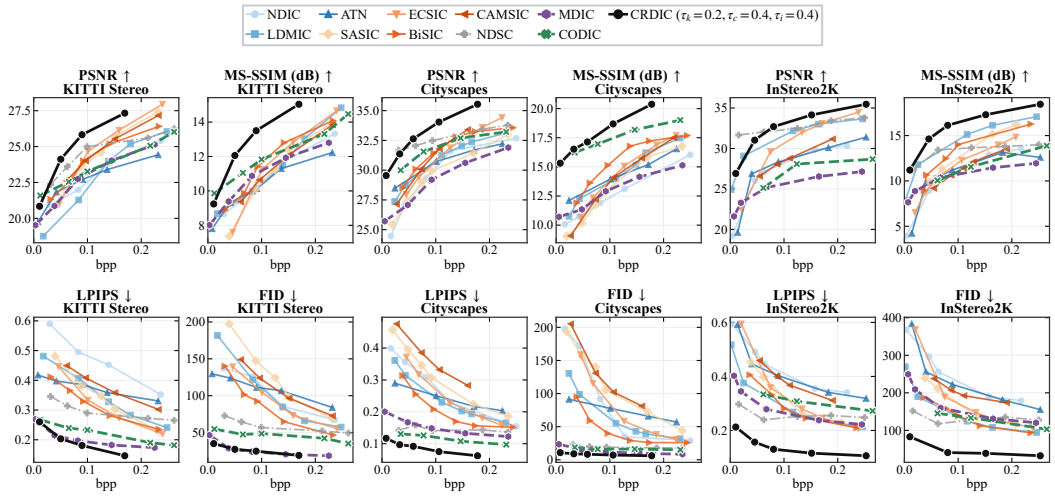

Figure 3: Rate-distortion-perception performance on KITTI Stereo, Cityscapes, and InStereo2K. Symbol conventions: ↑ = higher is better, ↓ = lower is better.

4 Experiments

4.1 Experimental Settings

Training Setup and Datasets. We train and evaluate CRDIC on KITTI Stereo (Geiger et al., 2012; Menze et al., 2015), Cityscapes (Cordts et al., 2016), and InStereo2K (Bao et al., 2020). The train/validation/test splits are 1576/790/790 for KITTI Stereo, 2975/500/1525 for Cityscapes, and 1960/50/50 for InStereo2K, where the InStereo2K validation set is randomly sampled from its original training split. KITTI images are center-cropped to 370×740 and resized to 128×256, Cityscapes images are resized to 128×256, and InStereo2K images are resized to 256×256.

CRDIC runs use NVIDIA L40 GPUs and Adam with a training batch size of 8. The learning rates for Stages 1–3 are 10^{-4}, 2×10^{-5}, and 2×10^{-5}, respectively. The entropy bottleneck uses an auxiliary learning rate of 10^{-3} in Stage 1. We use $L = 5$ feature levels and stochastic injection at $l_s = 3$. Stage 1 starts from random initialization at each target bitrate. Stage 2 uses $\lambda_\rho = 0.01$ and $\lambda_N = 5 \times 10^{-4}$ with MSE weight 1. Stage 3 uses $\lambda_D^s = 1$, $\lambda_P^s = 0.1$, and $\tau \sim \mathcal{U}(0, 1)$.

Evaluation Metrics. We report bits per pixel (BPP) for compression rate, and use PSNR (Korhonen & You, 2012) and MS-SSIM (Wang et al., 2003) to evaluate reconstruction fidelity. For perceptual quality, LPIPS (Zhang et al., 2018) measures full-reference perceptual similarity, while FID (Heusel et al., 2017) evaluates the distributional discrepancy between reconstructed and real images. Additional results on DISTS (Ding et al., 2022), KID (Bińkowski et al., 2018), and NIQE (Mittal et al., 2013) are provided in the Appendix.

Benchmarks. We compare CRDIC with representative compression methods under correlated-image settings. The decoder-only DIC baselines include NDIC (Mital et al., 2022), ATN (Mital et al., 2023), and NDSC (Whang et al., 2024). We further include LDMIC (Zhang et al., 2023), SASIC (Wödlinger et al., 2022), ECSIC (Wödlinger et al., 2024), BiSIC (Liu et al., 2024), and CAMSIC (Zhang et al., 2025) as representative multi-view image compression methods. For perception-oriented DIC, we compare with MDIC (Xu et al., 2026b) and CODIC (Xu et al., 2026a).

4.2 Main Results

Quantitative Comparison. We select τ according to the lowest average LPIPS on the validation set, yielding $\tau_k = 0.2$, $\tau_c = 0.4$, and $\tau_i = 0.4$ for three datasets. As illustrated in Fig. 3, CRDIC achieves the best PSNR and MS-SSIM performance across the three datasets, showing that SI-assisted deterministic reconstruction effectively reduces the residual left by the transmitted representation. The remaining uncertainty reflects not only pixel-level reconstruction error, but also



Figure 4: Qualitative comparison on Cityscapes and InStereo2K.

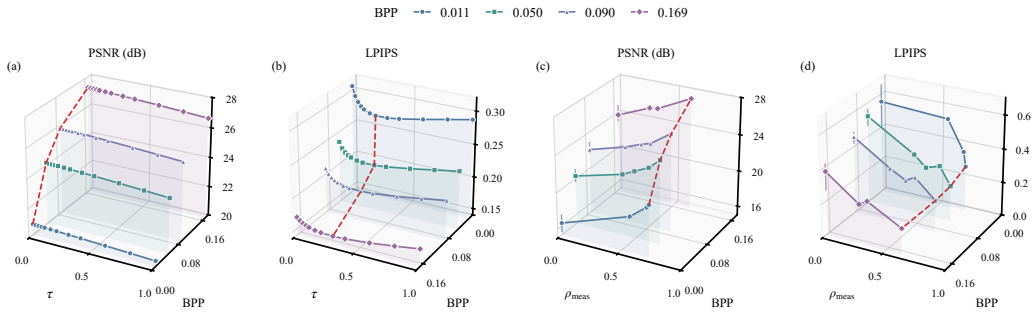

Figure 5: Controlled stochastic reconstruction on KITTI Stereo. (a) PSNR and (b) LPIPS under different restoration levels τ with the original SI. (c) PSNR and (d) LPIPS under different decoder-side SI conditions at $\tau = 0.2$. Results are shown across four fixed bitrates.

potential perceptual gain. By modeling this residual stochastically, CRDIC further achieves consistently lower LPIPS and FID while preserving the content supported by the target representation and SI, indicating favorable distributional quality compared with the diffusion-based methods (Xu et al., 2026a;b). Together, these results show that the proposed residual decomposition supports both accurate deterministic reconstruction and perceptually improved stochastic restoration.

Visual Comparison. Fig. 4 compares representative reconstructions. Distortion-oriented methods (Mital et al., 2023; Liu et al., 2024) tend to lose fine structures and produce smoother local details, while generative reconstruction can recover richer textures but may deviate from the paired target (Xu et al., 2026a;b). CRDIC better preserves object boundaries and structural detail while maintaining consistency with the target, which is supported by its higher PSNR and lower LPIPS.

Controlled Residual Restoration. At a fixed bitrate and SI condition, changing τ only adjusts the stochastic residual restoration in Eq. (21). Figs. 5(a) and 5(b) show that $\tau = 0$ gives the highest PSNR at all four bitrates. Increasing τ to 0.2 reduces LPIPS by 0.0146–0.0382, with a PSNR decrease of only 0.065–0.077 dB. Further increasing τ continues to reduce PSNR but does not yield additional LPIPS improvement.

This behavior follows the different quantities represented by the two metrics. PSNR directly penalizes deviation from the paired target, whereas LPIPS measures perceptual similarity in a learned feature space (Zhang et al., 2018). A moderate stochastic residual can restore local variations that are weakly represented by the deterministic estimate and thereby reduce the perceptual feature distance at a small pixel-fidelity cost. As the residual amplitude increases further, the sampled details deviate more from the paired target and LPIPS increases again. Therefore, the variance controlled by τ does not imply a monotonic change in LPIPS.

Figs. 5(c) and 5(d) fix $\tau = 0.2$ and evaluate CRDIC under different SI conditions. Changing SI affects both the deterministic correction and the residual passed to the stochastic stage, leading to different PSNR and LPIPS responses. Here, ρ_{meas} summarizes the residual reduction achieved by

Table 1: Stage-wise ablation on KITTI Stereo. All BD-Metric values are computed relative to the Stage-1 target-only model over the common log-BPP interval.

Stage	Variant	BD-LPIPS ↓	BD-DISTS ↓	BD-PSNR ↑	BD-MS-SSIM (dB) ↑
Stage 1	Target-only	0.0000	0.0000	0.0000	0.0000
Stage 2	w/o RACS	−0.1915	−0.1154	1.3177	1.9674
	w RACS	−0.2485	−0.1449	2.8353	3.7061
Stage 3	w/o RGRM	−0.2736	−0.1574	2.8294	3.6975
	w/o odd projection	−0.2277	−0.1444	3.0361	4.0247
	CRDIC (Ours)	−0.2891	−0.1710	2.8183	3.6807

deterministic reconstruction rather than serving as an independent control variable. The resulting trajectories therefore reflect the joint effect of SI quality, correspondence, and residual structure. Similar ρ_{meas} values can still yield different perceptual quality, showing that this scalar does not fully capture the spatial information provided by SI.

Taken together, the learned reconstruction approximately preserves the residual variance partition derived from conditional expectation. The predicted reliability and residual scale track the reduction and magnitude of the remaining residual, while τ provides controlled restoration of this unresolved component. The observed calibration gaps indicate the approximation between learned decoder statistics and their theoretical conditional quantities, without changing the deterministic and stochastic roles defined by the decomposition.

4.3 Ablation Study

Table 1 evaluates the main components introduced across the three training stages. Stage 1 provides the target-only baseline. In Stage 2, directly incorporating SI already improves both distortion and perceptual quality, while RACS further reduces BD-LPIPS and BD-DISTS by 0.0570 and 0.0295, and improves BD-PSNR and BD-MS-SSIM by 1.5176 dB and 1.7387 dB, respectively. This confirms that correspondence-aware alignment and reliability-guided fusion are important for recovering the SI-predictable residual component M.

Stage 3 mainly improves perceptual similarity while causing only minor changes in distortion. Compared with the Stage 2 model, the stochastic branch further improves LPIPS and DISTS, indicating that modeling the remaining residual can recover perceptually relevant variation that is not captured by the conditional mean. Removing RGRM reduces these perceptual gains, supporting its role in shaping the stochastic reconstruction according to the remaining SI-conditioned residual.

The odd projection further preserves the intended deterministic–stochastic separation. Without it, distortion improves but perceptual quality degrades, suggesting that the stochastic branch can deviate from its intended role of modeling variation around μ_θ. Constraining this residual to remain centered is therefore consistent with the proposed decomposition, where SI-predictable content is recovered deterministically and the remaining uncertainty is modeled stochastically.

5 Conclusion

This work studies distributed image compression (DIC) through the residual uncertainty remaining after the transmitted representation and decoder-only side information are jointly used. We formulate a conditional residual decomposition that separates the target-only residual into an SI-supported component and an orthogonal remainder, linking SI reliability to deterministic reconstruction error and remaining uncertainty. Based on this formulation, we propose Conditional Residual Decomposition for DIC (CRDIC), which assigns the SI-supported component to deterministic reconstruction and the remaining uncertainty to stochastic modeling. Experiments verify the learned reliability and residual scale while showing improved fidelity and perceptual quality. Overall, CRDIC makes the roles of SI and stochastic reconstruction explicit: SI removes the predictable part of the target-only residual, while stochastic modeling is applied to the uncertainty that remains after joint decoding.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, figure layout, code and LaTeX debugging, and feedback on manuscript organization and the presentation and interpretation of experimental results. All mathematical derivations, methodological claims, implementations, figures, and reported experimental results were reviewed and verified by the authors. The authors take full responsibility for the content of this submission.

ETHICS STATEMENT

This work studies image compression using established public research datasets. It does not involve human-subject experiments or the collection of new personally identifiable data.

REPRODUCIBILITY STATEMENT

The paper specifies the model components, optimization objectives, training stages, dataset splits, preprocessing procedures, evaluation metrics, and operating-point selection protocol. Additional implementation details and an anonymous code archive are provided with the supplementary material. The archive includes training and evaluation commands, but does not redistribute datasets or trained model weights.

REFERENCES

Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22324–22333, 2023.

Sharon Ayzik and Shai Avidan. Deep image compression using decoder side information. In *European Conference on Computer Vision*, pp. 699–714. Springer, 2020.

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

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Tom Bordin and Thomas Maugey. Linearly transformed color guide for low-bitrate diffusion-based image compression. *IEEE Transactions on Image Processing*, 34:468–482, 2025. doi: 10.1109/TIP.2024.3521301.

Jianhui Chang, Zhenghui Zhao, Chuanmin Jia, Shiqi Wang, Lingbo Yang, Qi Mao, Jian Zhang, and Siwei Ma. Conceptual compression via deep structure and texture synthesis. *IEEE Transactions on Image Processing*, 31:2809–2823, 2022. doi: 10.1109/TIP.2022.3159477.

Jun Chen, Lei Yu, Jia Wang, Wuxian Shi, Yiqun Ge, and Wen Tong. On the rate-distortion-perception function. *IEEE Journal on Selected Areas in Information Theory*, 3(4):664–673, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Fangyuan Gao, Xin Deng, Junpeng Jing, Xin Zou, and Mai Xu. Extremely low bit-rate image compression via invertible image generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6993–7004, 2023.

Junlong Gao, Zhimeng Huang, Qi Mao, Siwei Ma, and Chuanmin Jia. Exploring multimodal knowledge for image compression via large foundation models. *IEEE Transactions on Image Processing*, 34:5904–5919, 2025. doi: 10.1109/TIP.2025.3607616.

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. IEEE, 2012.

Yassine Hamdi and Deniz Gündüz. The rate-distortion-perception trade-off with side information. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1056–1061. IEEE, 2023.

Christian Herglotz, Hannah Och, Anna Meyer, Geetha Ramasubbu, Lena Eichermüller, Matthias Kränzler, Fabian Brand, Kristian Fischer, Dat Thanh Nguyen, Andy Regensky, et al. The bjøntegaard bible why your way of comparing video codecs may be wrong. *IEEE Transactions on Image Processing*, 33:987–1001, 2024.

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

Yujun Huang, Bin Chen, Shiyu Qin, Jiawei Li, Yaowei Wang, Tao Dai, and Shu-Tao Xia. Learned distributed image compression with multi-scale patch matching in feature domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4322–4329, 2023.

Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 37–38, 2012. doi: 10.1109/QoMEX.2012.6263880.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Toward extreme image compression with latent feature guidance and diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):888–899, 2025. doi: 10.1109/TCSVT.2024.3455576.

Zhening Liu, Xinjie Zhang, Jiawei Shao, Zehong Lin, and Jun Zhang. Bidirectional stereo image compression with cross-dimensional entropy model. In *European Conference on Computer Vision*, pp. 480–496. Springer, 2024.

Qi Mao, Chongyu Wang, Meng Wang, Shiqi Wang, Ruijie Chen, Libiao Jin, and Siwei Ma. Scalable face image coding via stylegan prior: Toward compression for human-machine collaborative vision. *IEEE Transactions on Image Processing*, 33:408–422, 2024. doi: 10.1109/TIP.2023.3343912.

Takashi Matsuyama and Norimichi Ukita. Real-time multitarget tracking by a cooperative distributed vision system. *Proceedings of the IEEE*, 90(7):1136–1150, 2002.

Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 2: 427–434, 2015.

Nitish Mital, Ezgi Özyılkan, Ali Garjani, and Deniz Gündüz. Neural distributed image compression using common information. In *2022 Data Compression Conference (DCC)*, pp. 182–191. IEEE, 2022.

Nitish Mital, Ezgi Özyilkan, Ali Garjani, and Deniz Gündüz. Neural distributed image compression with cross-attention feature alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2498–2507, 2023.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Xueyan Niu, Deniz Gündüz, Bo Bai, and Wei Han. Conditional rate-distortion-perception trade-off. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1068–1073. IEEE, 2023.

Jingjing Qian, Sadaf Salehkalaibar, Jun Chen, Ashish Khisti, Wei Yu, Wuxian Shi, Yiqun Ge, and Wen Tong. Rate-distortion-perception tradeoff for gaussian vector sources. *IEEE Journal on Selected Areas in Information Theory*, 6:1–17, 2025. doi: 10.1109/JSAIT.2024.3509420.

David Slepian and Jack Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, 1973.

Jeremy Vonderfecht and Feng Liu. Lossy compression with pretrained diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 687–702, 2025.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Jay Whang, Alliot Nagle, Anish Acharya, Hyeji Kim, and Alexandros G. Dimakis. Neural distributed source coding. *IEEE Journal on Selected Areas in Information Theory*, 5:493–508, 2024. doi: 10.1109/JSAIT.2024.3412976.

Matthias Wödlinger, Jan Kotera, Jan Xu, and Robert Sablatnig. Sasic: Stereo image compression with latent shifts and stereo attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 661–670, 2022.

Matthias Wödlinger, Jan Kotera, Manuel Keglevic, Jan Xu, and Robert Sablatnig. Ecsic: Epipolar cross attention for stereo image compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3436–3445, 2024.

Aaron Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

Guojun Xu, Cheng Tan, Jianwen Xiang, and Junwei Zhou. Codic: Distributed image compression with distortion-perception co-optimization. *IEEE Transactions on Multimedia*, pp. 1–12, 2026a. doi: 10.1109/TMM.2026.3718254.

Guojun Xu, Mingyang Zhang, Jianwen Xiang, Cheng Tan, Yanchao Yang, and Junwei Zhou. Distributed image compression with multimodal side information at extremely low bitrates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33662–33671, June 2026b.

Zeyu Yan, Fei Wen, and Peilin Liu. Optimally controllable perceptual lossy compression. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24911–24928. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/yan22b.html`.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Xinjie Zhang, Jiawei Shao, and Jun Zhang. Ldmic: Learning-based distributed multi-view image coding. In *International Conference on Learning Representations*, 2023.

Xinjie Zhang, Shenyuan Gao, Zhening Liu, Jiawei Shao, Xingtong Ge, Dailan He, Tongda Xu, Yan Wang, and Jun Zhang. Camsic: Content-aware masked image modeling transformer for stereo image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10239–10247, 2025.

A Appendix

A.1 Additional Implementation Details of CRDIC

This section provides the implementation details omitted from the main paper, including target coding, SI correspondence, multi-scale fusion, residual-scale calibration, and stochastic residual propagation.

A.1.1 Target Representation and Base Reconstruction

The target encoder and entropy model follow a mean-scale hyperprior. The analysis transform produces a quantized latent $\hat{y}_x$, while the hyper-latent $\hat{z}_x$ predicts the mean and scale used for entropy coding. The expected bitrate is

$$R = \frac{1}{HW}\mathbb{E}\left[-\log_2 p(\hat{y}_x \mid \hat{z}_x) - \log_2 p(\hat{z}_x)\right]. \tag{25}$$

The target decoder produces the base reconstruction X_b together with multi-scale target features $\{F_l^x\}_{l=0}^{L-1}$. As described in the main paper, X_b corresponds to the target-only estimate m_T, while the feature pyramid is used for subsequent SI correspondence and reconstruction.

A.1.2 Reliability-Aware Correspondence Strategy

Let $\{F_l^y(Y)\}_{l=0}^{L-1}$ denote the multi-scale features extracted from the decoder-side image Y. To remove the feature response independent of the available SI, we define

$$\Delta F_l^y = F_l^y(Y) - F_l^y(0), \qquad l = 0, \ldots, L-1. \tag{26}$$

Hence, $\Delta F_l^y = 0$ when SI is unavailable.

At each scale, the target and SI features are projected as

$$Q_l = W_l^Q F_l^x, \qquad K_l = W_l^K \Delta F_l^y, \qquad V_l = W_l^V \Delta F_l^y. \tag{27}$$

After channel-wise normalization, correspondence is searched along the horizontal disparity range $\mathcal{D}_l$. For target position p and disparity d, the normalized matching distribution is

$$\pi_l(d,p) = \frac{\exp\big(\gamma\langle \bar{Q}_l(p), \bar{K}_l(p-d)\rangle\big)}{\sum_{d'\in\mathcal{D}_l}\exp\big(\gamma\langle \bar{Q}_l(p), \bar{K}_l(p-d')\rangle\big)}, \tag{28}$$

where γ is the cost-volume temperature. The aligned SI feature is

$$\Delta\bar{F}_l^y(p) = \sum_{d\in\mathcal{D}_l}\pi_l(d,p)V_l(p-d). \tag{29}$$

The disparity distribution also provides local correspondence cues. We use its normalized concentration, peak probability, and peak margin:

$$\begin{aligned} c_l(p) &= 1 - \frac{H(\pi_l(\cdot,p))}{\log|\mathcal{D}_l|}, \\ s_l^{(1)}(p) &= \max_{d\in\mathcal{D}_l}\pi_l(d,p), \\ m_l(p) &= s_l^{(1)}(p) - s_l^{(2)}(p), \end{aligned} \tag{30}$$

where $H(\pi_l) = -\sum_{d\in\mathcal{D}_l}\pi_l(d,p)\log\pi_l(d,p)$ and $s_l^{(2)}(p)$ is the second-largest correspondence probability.

These quantities are combined with the target query and aligned SI feature:

$$A_l = \mathrm{Concat}\left[Q_l, \Delta\bar{F}_l^y, |Q_l - \Delta\bar{F}_l^y|, c_l, s_l^{(1)}, m_l\right]. \tag{31}$$

An SI availability indicator $a_l \in \{0,1\}$ ensures zero SI contribution when the side image is unavailable. The fusion gate and reliability map are predicted as

$$g_l = a_l h_l^g(A_l), \qquad \rho_l = a_l h_l^\rho(A_l). \tag{32}$$

The gate g_l controls SI contribution to deterministic reconstruction, while ρ_l estimates the corresponding local residual reduction.

A.1.3 MULTI-SCALE SI FUSION

The aligned SI feature is incorporated into the target decoder through a gated correction. At each scale,

$$\Delta_l = \phi_l(F_l^x, \Delta\bar{F}_l^y) - \phi_l(F_l^x, 0), \tag{33}$$

and

$$\tilde{F}_l^x = F_l^x + \alpha_l g_l \odot \tanh(\Delta_l), \tag{34}$$

where ϕ_l denotes the fusion mapping and α_l controls the correction magnitude. The subtraction removes the response of the fusion mapping without SI, so that Δ_l contains only the SI-dependent correction.

At the finest scale, the fused representation is mapped to

$$\Delta_\mu = s_{\max} \tanh\Big(h_\mu(\tilde{F}_0^x)\Big), \tag{35}$$

which is used in the deterministic reconstruction defined in Eq. (15) of the main paper.

A.1.4 RELIABILITY SUPERVISION

The reliability target in Eq. (16) is constructed from the reduction in local reconstruction error. We first compute

$$\begin{aligned} e^b(p) &= \frac{1}{3}\|X(p) - X_b(p)\|_2^2, \\ e^c(p) &= \frac{1}{3}\|X(p) - \mu_\theta(p)\|_2^2, \end{aligned} \tag{36}$$

and obtain the locally averaged errors

$$D^b = \mathcal{A}(e^b), \qquad D^c = \mathcal{A}(e^c), \tag{37}$$

where $\mathcal{A}(\cdot)$ denotes local average filtering.

The full-resolution reliability target is then obtained from their relative reduction as defined in the main paper. For feature level l, it is resized and detached from the reconstruction graph:

$$\rho_l^* = \mathrm{sg}\left[\mathcal{R}_l(\rho^*)\right]. \tag{38}$$

The resulting targets supervise the multi-scale reliability predictions through $\mathcal{L}_\rho$ in Eq. (16).

A.2 RESIDUAL-SCALE CALIBRATION

The reference scale $\sigma_0 = h_\sigma(X_b)$ is predicted solely from the target-only reconstruction. The conditional residual scale σ_c is then obtained from Eq. (18) in the main paper.

Let

$$r_c = X - \mu_\theta \tag{39}$$

denote the residual remaining after deterministic reconstruction. Under the Gaussian scale objective, the optimal prediction satisfies

$$\sigma_c^2(T, S) = \mathbb{E}\big[(X - \mu_\theta)^2 \mid T, S\big]. \tag{40}$$

When μ_θ approaches $\mathbb{E}[X \mid T, S]$, this quantity estimates $\mathrm{Var}(X \mid T, S)$. Averaging it over S conditioned on T recovers the theoretical remaining variance $V_{T,S}$ in Eq. (6).

A.2.1 STOCHASTIC RESIDUAL PROPAGATION AND RGRM

At decoder level l_s, we draw $\varepsilon \sim \mathcal{N}(0, I)$ and modulate it using the deterministic conditioning features

$$\mathcal{C}_{l_s} = \left\{\tilde{F}_{l_s}^x, \Delta\bar{F}_{l_s}^y, \rho_{l_s}\right\}. \tag{41}$$

The modulation scale and stochastic input are

$$a_{l_s} = g_\varepsilon(\mathcal{C}_{l_s}), \qquad n_\varepsilon = a_{l_s} \odot \varepsilon. \tag{42}$$

The antithetic projection in Eq. (20) ensures

$$R_{l_s}(-\varepsilon) = -R_{l_s}(\varepsilon), \tag{43}$$

and therefore

$$\mathbb{E}_\varepsilon \left[R_{l_s}(\varepsilon) \mid T, S \right] = 0 \tag{44}$$

for symmetric Gaussian noise.

The stochastic feature is then propagated to finer decoder scales. Let r_l denote the upsampled residual from the preceding scale. RGRM predicts

$$\gamma_l = c_l(\Delta \bar{F}_l^y, \rho_l) \tag{45}$$

and updates the residual as

$$R_l = r_l \odot \left[1 + \eta \rho_l \odot \tanh(\gamma_l) \right], \tag{46}$$

where η controls the modulation strength. When $\rho_l = 0$, the update reduces to $R_l = r_l$.

At the finest level, R_0 is projected to an image-domain residual R_{raw}. Its low-frequency component and sample-wise mean are removed, after which the residual is normalized to obtain Z_θ. This removes sample-dependent global scale while retaining the learned spatial and cross-channel structure. The final amplitude is subsequently determined by $q(\rho_0, \tau)\sigma_c$ as defined in Eq. (21).

A.2.2 Detailed Process of Optimization

We first optimize the target codec to obtain the target-only reconstruction, then introduce SI to learn the deterministic residual reduction and the remaining residual scale, and finally optimize stochastic residual reconstruction with the deterministic model fixed. We use MSE $\mathcal{L}_{\text{D}}$ for reconstruction fidelity and pretrained LPIPS $\mathcal{L}_{\text{P}}$ for perceptual quality in the stochastic stage. The reliability loss $\mathcal{L}_\rho$ and residual-scale loss $\mathcal{L}_{\text{NLL}}$ are defined in Eq. (17) and Eq. (19), respectively.

Target-only Codec Optimization. We first train the target codec from random initialization without using SI:

$$\mathcal{L}_{\text{target}} = \lambda_R R + \mathcal{L}_{\text{D}}(X, X_b), \tag{47}$$

where the rate weight λ_R is set separately for each operating point. This stage determines the transmitted target representation and the base reconstruction X_b, corresponding to the target-only estimate m_T in the theoretical formulation.

SI-assisted Deterministic Optimization. We then fix the target codec and train the SI-assisted deterministic model to refine X_b while estimating the associated reliability and remaining residual scale. The objective is

$$\mathcal{L}_{\text{det}} = \lambda_D \mathcal{L}_{\text{D}}(X, \mu_\theta) + \lambda_\rho \mathcal{L}_\rho + \lambda_N \mathcal{L}_{\text{NLL}}. \tag{48}$$

The SI-dependent correction is initialized to zero, such that $\mu_\theta = X_b$ at the beginning of this stage. Since the target codec is fixed, the bitrate R and base reconstruction X_b remain unchanged while the SI-dependent modules learn the residual component supported by Y. The loss $\mathcal{L}_\rho$ constrains the predicted reliability, while $\mathcal{L}_{\text{NLL}}$ calibrates σ_c to the residual remaining around μ_θ.

Stochastic Residual Optimization. Finally, we freeze the target codec and all SI-assisted deterministic modules, so that μ_θ, ρ_l, and σ_c remain fixed, and optimize only the stochastic residual modules. During training, $\tau \sim \mathcal{U}(0, \tau_{\text{train}})$, and the objective is

$$\mathcal{L}_{\text{stoch}} = \lambda_D^s \mathcal{L}_{\text{D}}(X, \hat{X}) + \lambda_P^s \mathcal{L}_{\text{P}}(X, \hat{X}). \tag{49}$$

For each sampled τ, the factor $q(\rho_0, \tau)\sigma_c$ determines the residual amplitude, while the stochastic modules learn the residual structure represented by Z_θ.

A.3 Theory Discussion

We examine whether the learned CRDIC model follows the residual relations derived in Sec. 3. For controlled comparisons, the transmitted target representation and X_b are fixed, and only decoder-side SI or the stochastic restoration level is varied.

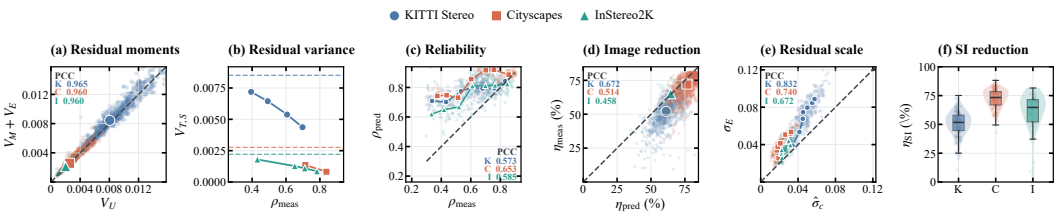

Figure 6: Validation of the conditional residual decomposition on KITTI Stereo, Cityscapes, and InStereo2K. (a) Residual second-moment relation. (b) Remaining residual variance under different SI reliability levels. (c) Predicted versus measured reliability. (d) Predicted versus measured residual reduction. (e) Predicted residual scale versus measured residual RMS. (f) Distribution of the image-level SI residual reduction η_{SI}. Dashed diagonals denote identity, and dashed horizontal lines in (b) indicate the target-only residual variance.

Conditional Residual Decomposition. For the learned decoder, we define

$$U = X - X_b, \qquad M = \mu_\theta - X_b, \qquad E = X - \mu_\theta. \tag{50}$$

The conditional residual decomposition gives $\mathbb{E}[ME \mid T] = 0$ and hence $V_U = V_M + V_E$ for the ideal conditional estimates. To examine this relation in the learned model, we measure

$$\kappa_{ME} = \frac{\langle M, E \rangle}{\sqrt{V_M V_E}}, \tag{51}$$

where $V_M = \langle M, M \rangle$ and $V_E = \langle E, E \rangle$.

Fig. 6(a) compares V_U with $V_M + V_E$ for individual images. The pooled κ_{ME} values are -0.038, 0.0002, and -0.015 on KITTI Stereo, Cityscapes, and InStereo2K, respectively. The corresponding correlations are 0.965, 0.960, and 0.960, with median relative differences below 5.5%. The small cross moments and the concentration around the identity line show that the learned residual components approximately retain the orthogonality underlying the theoretical variance decomposition.

For each image, we further measure the fraction of the target-only residual removed after SI-assisted reconstruction as

$$\eta_{\mathrm{SI},i} = 1 - \frac{V_{E,i}}{V_{U,i}}. \tag{52}$$

Fig. 6(f) summarizes η_{SI} over the complete test sets. SI removes a substantial fraction of the target-only residual across all three datasets, while the within-dataset spread and the differences across datasets show that its contribution varies considerably with the target–SI pair. This variation supports estimating reliability from the available target and SI features instead of assuming a fixed contribution from SI.

Fig. 7 further provides spatial examples of the learned decomposition. The SI-supported component M preserves most of the structured content in the target-only residual U, while the remaining residual E is substantially reduced and becomes concentrated around object boundaries, occlusions, thin structures, and local appearance details. This behavior indicates that SI does not merely attenuate the residual magnitude, but selectively explains the component that can be inferred from cross-view correspondence, leaving a more localized residual after joint conditioning.

The spatial overlap that remains between M and E does not contradict the orthogonality in Eq. (5), since the theoretical constraint concerns their conditional cross moment rather than disjoint spatial support. In practice, the same boundary or texture region may contain both an SI-predictable component and an unresolved component, while their aggregate interaction remains close to orthogonal as verified in Fig. 6(a).

The last row further relates this remaining residual to Stage 3 stochastic reconstruction. The Stage 3 correction is concentrated on fine structures and local details that remain prominent in E, and these regions frequently coincide with positive fine-scale LPIPS gain. This suggests that the residual remaining after deterministic reconstruction contains perceptually relevant variation, providing the

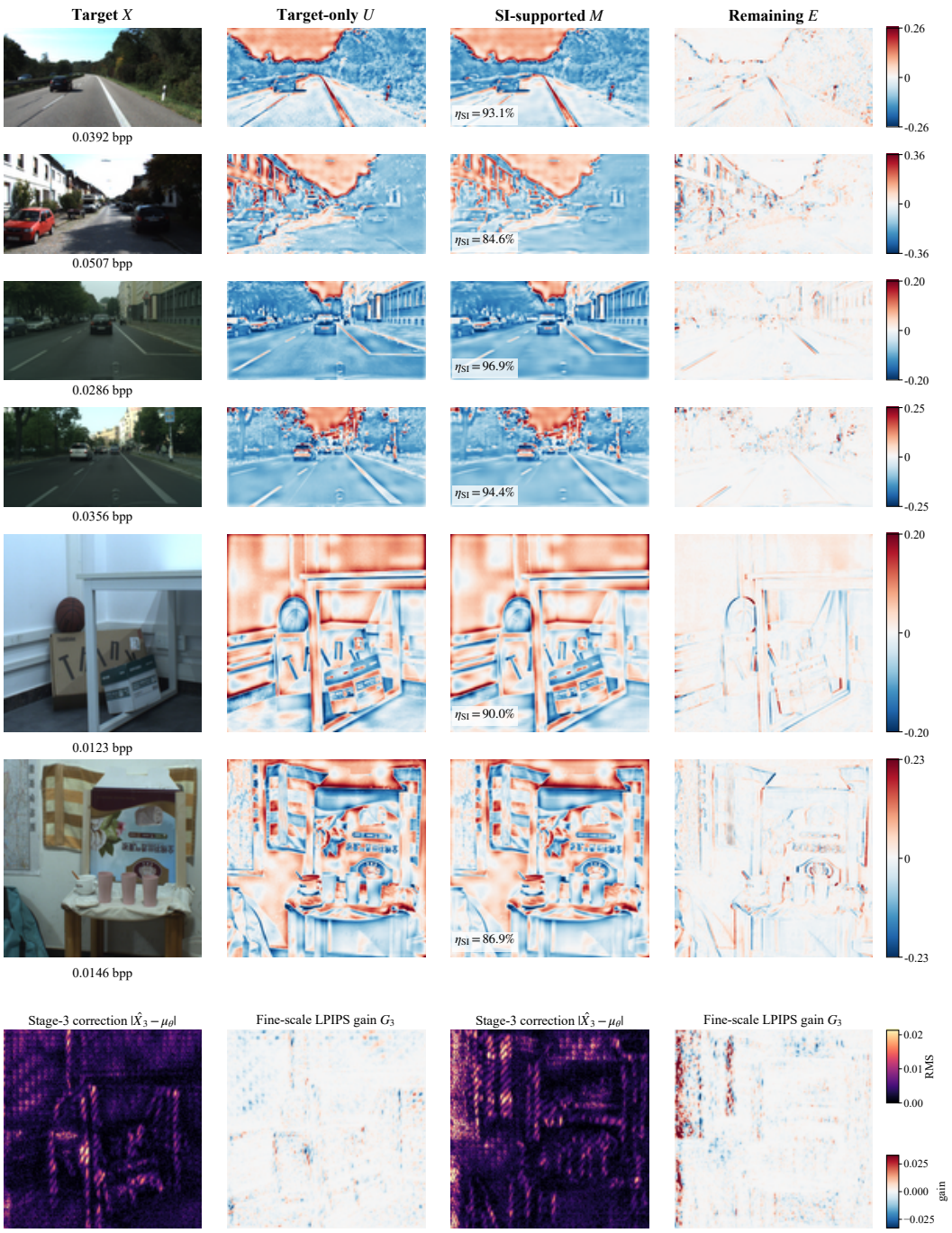

Figure 7: Additional conditional-residual decompositions and Stage-3 corrections. The first six rows show X, the target-only residual $U = X - X_b$, the SI-supported component $M = \mu_\theta - X_b$, and the remaining residual $E = X - \mu_\theta$ on KITTI Stereo, Cityscapes, and InStereo2K. Each M panel reports $\eta_{\mathrm{SI}} = 1 - V_E/V_U$. The last row shows the Stage-3 correction magnitude $|\hat{X}_3 - \mu_\theta|$ and the corresponding fine-scale LPIPS gain G_3 for two InStereo2K examples.

spatial motivation for modeling E stochastically rather than treating it solely as a pixel-level reconstruction error.

SI Reliability and Remaining Residual Scale. According to Eq. (7), reliability measures the fraction of target-only residual variance explained by SI. The theoretical $\rho(T)$ is a conditional population quantity and cannot be observed directly for an individual reconstruction. We therefore use the measured residual reduction under a fixed target bitstream to define

$$\rho_{\mathrm{meas}} = \sqrt{\left[1 - \frac{D_c}{D_b}\right]_+}, \tag{53}$$

where D_b and D_c denote the target-only and SI-assisted residual energies. The decoder prediction is denoted by ρ_{pred}.

Fig. 6(b) shows that the remaining residual variance decreases with increasing ρ_{meas}. The three datasets have different absolute levels because their target-only residual variances differ, while each follows the relation $V_{T,S} = V_T(1 - \rho^2)$ derived in Sec. 3. This result connects the measured SI contribution to the reduction in the residual left after joint conditioning.

In Fig. 6(c), ρ_{pred} increases consistently with ρ_{meas} under different SI conditions, showing that the decoder can distinguish changes in the amount of residual reduction provided by SI. The binned predictions remain above the identity line over part of the range, indicating a tendency to overestimate reliability when SI becomes less effective. This discrepancy is expected because ρ_{pred} is inferred from local target and SI features, while ρ_{meas} is obtained from the realized reduction of the reconstruction residual. Changes in SI may preserve local feature correspondence without producing an equivalent reduction in reconstruction error, and the clipping in Eq. (53) further compresses different low-reliability cases into the same measured range.

Fig. 6(d) further examines whether the spatial reliability prediction translates into the expected reduction at the image level. The predicted and measured reductions maintain a clear positive relation across the three datasets, while their spread and dataset-dependent offsets show that local reliability cannot determine the final reduction exactly. The final reconstruction also depends on how accurately the SI-supported correction is recovered at each location. These results support the use of ρ_l as a relative measure of SI contribution for spatial control.

Fig. 6(e) compares the predicted residual scale with the RMS of the residual remaining after deterministic reconstruction. The predicted scale follows the change in residual magnitude across all three datasets, but is generally lower than the measured value, with a larger gap for samples with larger residuals. This behavior is related to the coupling $\sigma_c = \sigma_0\sqrt{1 - \rho_0^2}$: an overestimated reliability directly reduces the predicted remaining scale, so the calibration bias observed in Figs. 6(c) and 6(d) is propagated to σ_c. In addition, σ_0 itself is predicted from the target-only reconstruction, which limits the absolute accuracy of the resulting scale. Despite this bias, σ_c preserves the relative variation of the remaining residual and therefore provides a meaningful amplitude constraint for stochastic reconstruction.

A.4 Limitations and Future Directions

The theoretical formulation uses $E' \sim p(E \mid T, S)$ to characterize the residual variation that remains unresolved after conditioning on both the transmitted representation and SI. In CRDIC, this ideal sample is approximated by a learned stochastic residual centered around the deterministic reconstruction μ_θ, with its structure and amplitude constrained by the decoder. The current model does not explicitly reproduce the complete conditional distribution of E. To examine the behavior of this approximation, we evaluate the full model and the w/o odd projection variant at two InStereo2K operating points. For each fixed (T, S) and μ_θ, $K = 32$ stochastic reconstructions are sampled at $\tau = 1$ and compared with the observed remaining residual $E = X - \mu_\theta$.

Fig. 8(a) shows that the mean-offset ratio B_μ remains small for the full model across both bitrates and different reliability levels, while removing the odd projection introduces a much larger offset. This agrees with the intended role of the odd construction in keeping stochastic reconstruction centered around μ_θ under symmetric noise. The rank distributions in Fig. 8(b) provide a complementary view. For the full model, the approximately symmetric accumulation at the two boundary ranks is consistent with the small mean offset, but also indicates that the generated residual covers a narrower range of variation than the ideal conditional model. This behavior is consistent with the conservative residual-scale estimation observed earlier in Fig. 6(e), and becomes more pronounced when the

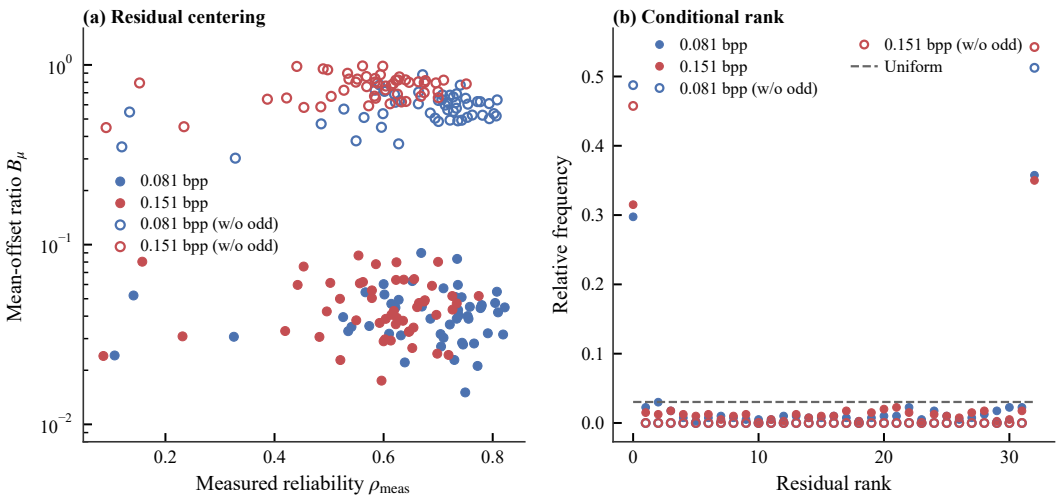

Figure 8: Stochastic residual behavior on InStereo2K at the two trained bitrates, comparing the full model with the "w/o odd projection" ablation (filled vs. open markers). (a) Per-image mean-offset ratio B_μ against the measured reliability ρ_{meas} on a logarithmic axis. (b) Rank distribution of the remaining residual E among $K = 32$ stochastic samples, with the dashed line denoting the uniform reference.

odd projection is removed. The results therefore suggest that the current approximation preserves the intended residual centering while only partially representing the variation left after deterministic reconstruction.

Overall, the current formulation provides a stable centered residual model and controlled stochastic restoration, with the experiments showing favorable fidelity and perceptual performance. Its main limitation is that the learned stochastic residual captures only part of the conditional variation remaining after deterministic reconstruction. Future work may improve this approximation through a more distribution-aware characterization of the remaining residual while preserving the deterministic–stochastic decomposition of CRDIC.

A.5 Complexity Analysis

Table 2 compares computational complexity and BD-Metrics on KITTI Stereo. The Bjøntegaard Delta metrics (Bjontegaard, 2001; Herglotz et al., 2024) summarize the average difference over the common bitrate range. DIC favors lightweight independent encoding and shifts correlation exploitation to the decoder. CRDIC follows this principle, using only 2.75M encoder parameters and 0.0033 s of encoding time. Although NDSC is lighter, its distortion and perceptual performance are both weaker.

The additional cost of CRDIC lies mainly in decoder-side SI modeling. With 12.20M decoder parameters and 0.0246 s decoding time, CRDIC remains much faster than diffusion-based DIC, decoding about 18.7× faster than MDIC and 23.2× faster than CODIC. Meanwhile, it achieves the highest BD-PSNR and the lowest BD-LPIPS, showing a favorable trade-off between decoding cost and reconstruction quality.

All runtimes in Table 2 are measured with batch size 1 and 128×256 input resolution. Each method is run for 50 warm-up iterations followed by 200 timed forward passes. We report the mean and standard deviation over the timed runs. Arithmetic coding is excluded so that the comparison reflects the neural encoder and decoder computation.

A.6 Additional Perceptual Metrics

Fig. 9 reports the remaining perceptual metrics used in our evaluation. CRDIC achieves consistently low DISTS and KID across the three datasets, showing that the perceptual gains observed with

Table 2: Complexity and BD-Metric comparison on KITTI Stereo. BD-Metrics are computed relative to LDMIC; ↓/↑ indicate lower/higher is better.

Method	Params. (M)		FLOPs (G)	Time (s)		BD-Metric	
	Enc.	Dec.		Enc.	Dec.	BD-LPIPS ↓	BD-PSNR ↑
NDIC (Mital et al., 2022)	17.40	7.63	19.88	0.0115±0.0010	0.0042±0.0001	0.1063	0.5940
ATN (Mital et al., 2023)	8.70	21.01	322.53	0.0115±0.0010	0.0246±0.0014	-0.0212	1.0896
NDSC (Whang et al., 2024)	**0.93**	**1.66**	16.40	**0.0016±0.0001**	**0.0032±0.0002**	-0.0778	2.1922
LDMIC (Zhang et al., 2023)	7.73	5.96	21.22	0.0062±0.0004	0.0073±0.0008	0.0000	0.0000
SASIC (Wödlinger et al., 2022)	3.58	3.01	108.72	0.0342±0.0032	0.0139±0.0005	-0.0058	1.6510
ECSIC (Wödlinger et al., 2024)	15.08	10.67	**7.26**	0.0102±0.0025	0.0052±0.0003	-0.0330	1.9554
BiSIC (Liu et al., 2024)	51.20	27.02	130.26	0.1055±0.0011	0.0100±0.0017	-0.0495	1.6185
CAMSIC (Zhang et al., 2025)	48.53	7.34	235.99	0.2039±0.0071	0.2240±0.0112	0.0334	1.6451
MDIC (5 steps) (Xu et al., 2026b)	5.87	1281.61	827.04	0.0120±0.0010	0.4610±0.0020	-0.2032	0.9528
CODIC (50 steps) (Xu et al., 2026a)	11.38	59.63	2467.35	0.0074±0.0005	0.5700±0.0048	-0.0368	0.9888
CRDIC	2.75	12.20	71.14	0.0033±0.0001	0.0246±0.0004	**-0.2109**	**3.9478**

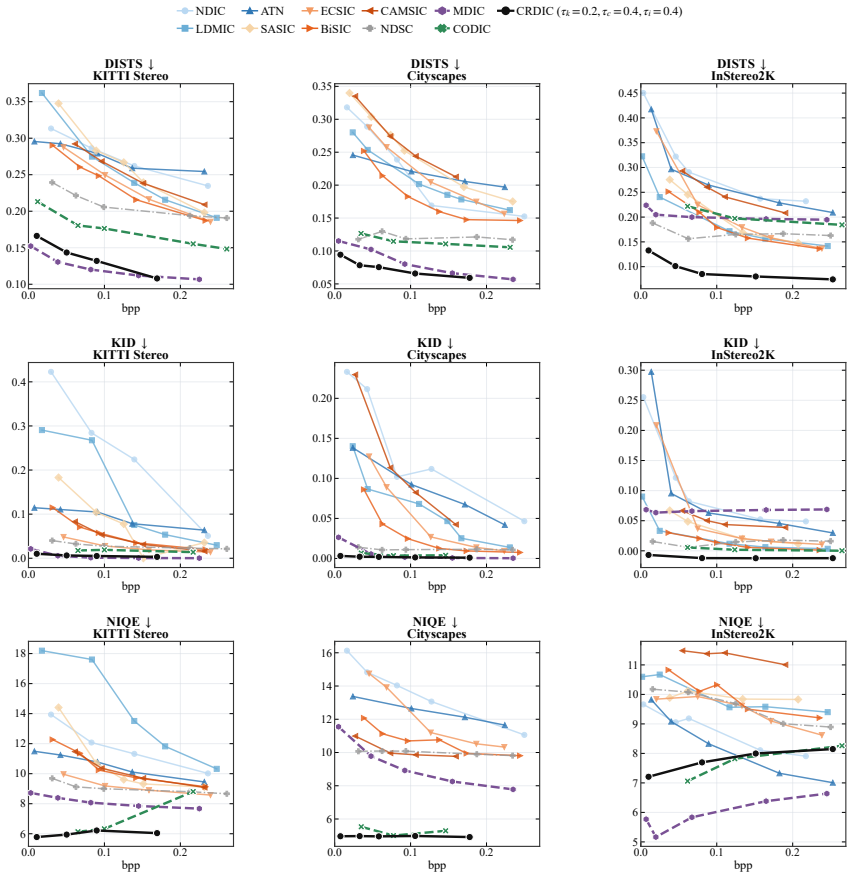

Figure 9: Additional rate–perception results on KITTI Stereo, Cityscapes, and InStereo2K. Lower DISTS, KID, and NIQE are better.

LPIPS and FID also extend to complementary full-reference and distribution-level measures. The NIQE results are less uniform: CRDIC remains competitive on KITTI Stereo and Cityscapes, while several baselines obtain lower values on InStereo2K. Since NIQE evaluates no-reference natural-image statistics without access to the paired target, this difference is compatible with methods whose generative priors favor such statistics, even when their paired perceptual similarity is weaker.

A.7 Additional Qualitative Comparisons

Fig. 10 provides additional qualitative comparisons across the three datasets. Distortion-oriented methods tend to smooth fine structures and object boundaries at low bitrates, while the generative baselines recover richer local appearance but can introduce details that deviate from the paired target. CRDIC preserves the structures supported by the transmitted representation and SI while recovering finer local details, which is also reflected by its PSNR and LPIPS values. These examples are consistent with the residual decomposition: deterministic reconstruction recovers the SI-supported content, while stochastic restoration is confined to the residual left after joint decoding.



Figure 10: Additional qualitative comparisons on KITTI Stereo, Cityscapes, and InStereo2K. Enlarged regions highlight local structures and reconstruction details. BPP, PSNR, and LPIPS are reported below each reconstruction.