Controllable Generative Reference for Stereo Image Compression via Reliability-Aware Gating

Zhineng Zhao[1,3], Mingyao Hong[3], Fengfan Shi[2,3], and Zhihai He[1,3⋆]

[1] Southern University of Science and Technology, Shenzhen, China
[2] Harbin Institute of Technology, Shenzhen, China
[3] Peng Cheng Laboratory, Shenzhen, China
zhaozhn@pcl.ac.cn, hezh@sustech.edu.cn

Abstract. Stereo image compression fundamentally relies on exploiting cross-view redundancies. We observe that the performance of conventional correspondence-driven methodologies degrades severely in out-of-view (OOV) regions or when reference features are heavily quantized, resulting in blurred textures and prediction failure. In this paper, we propose a novel generative stereo image compression framework driven by a controllable generative reference. Instead of relying on finding explicit correspondences from a degraded reference, we leverage the powerful priors of a pre-trained diffusion model to synthesize a high-quality target-view reference. To bridge the gap between stochastic generative processes and strict stereo consistency, we introduce Joint Semantic-Spatial Control for reference generation. This mechanism synergistically steers the synthesis using a robust geometric anchor from the compressed left view and an ultra-compact semantic condition extracted from the target view. Furthermore, to minimize the impact of structural inconsistency during reference generation on the quality of the reconstructed image, we design a Reliability-Aware Gating module. By dynamically evaluating the synthesized prior in the feature space, this module adaptively regulates generative information flow within the entropy model to improve the reconstructed image quality. Extensive experiments on standard benchmarks demonstrate that our framework achieves state-of-the-art rate-distortion performance, delivering remarkable bitrate savings and high visual quality at low bitrates.

Keywords: Stereo Image Compression · Diffusion Models · Generative Prior · Entropy Coding

1 Introduction

Stereo image compression shows great potential for powering critical applications in binocular vision systems, ranging from immersive 3D telepresence to safety-critical autonomous driving [13, 14, 16, 50]. Unlike traditional single-image codecs [8, 39], stereo compression architectures strive to maximize coding efficiency by exploiting the significant inter-view redundancy inherent in binocular

⋆ Corresponding author.

pairs. The core objective is to reconstruct a high-quality target view by leveraging information from a compressed reference view. Over the past decade, the field has transitioned from traditional standards like MV-HEVC [37] to learned paradigms that utilize explicit geometric constraints, such as disparity-based warping or optical flow [3,12,22], to align and predict target-view features [17,42].

Despite these advancements, we observe that the performance of these conventional correspondence-driven methodologies degrades severely under stringent bandwidth constraints, regardless of whether they utilize explicit spatial alignment or implicit attention-based feature interaction [12, 25, 41]. Specifically, we identify two critical limitations. First, under heavy compression, the compressed reference view itself suffers from severe degradation, resulting in heavily quantized latent representations. Attempting to extract and transfer context from these degraded reference features inevitably propagates quantization artifacts to the target view, resulting in blurred textures and a loss of high-frequency details. Second, due to the stereo camera baseline, certain regions of the target view are physically occluded in the reference view—a phenomenon known as out-of-view (OOV) regions [35]. Since both feature warping and cross-attention mechanisms rely strictly on finding explicit cross-view correspondences [41, 42, 48], these occluded areas suffer from severe information depletion. Consequently, current interaction modules fail to establish reliable cross-view contexts, inevitably leading to prediction failure and severe structural inconsistency.

To overcome these limitations, we propose a novel generative stereo image compression framework driven by a controllable generative reference. Instead of relying on finding explicit correspondences from a degraded reference to predict the target view, we leverage the powerful priors of pre-trained diffusion models [33, 34, 46] to synthesize a high-quality target-view reference. To ensure practical runtime efficiency for the overall compression system, this generative process is executed in the latent space via single-step inference. To bridge the gap between stochastic generative processes and strict stereo consistency, we introduce Joint Semantic-Spatial Control for reference generation [31, 46]. This mechanism synergistically steers the synthesis using a robust geometric anchor from the compressed left view and an ultra-compact semantic condition (requiring only $\sim$0.00065 bpp) extracted from the target view. This joint guidance allows the model to infer missing out-of-view (OOV) structures and restore degraded textures with high perceptual fidelity. Consequently, this generative stage produces a visually compelling reconstruction that can stand alone for applications prioritizing high visual quality [18, 27, 28] at negligible bitrates.

Nevertheless, for objective-oriented compression, a critical challenge remains: the generative process inherently introduces structural inconsistency [1,7]. While generative priors can synthesize visually plausible details, these generated structures are not always mathematically aligned with the ground truth at the pixel level. This inconsistency inevitably leads to severe penalties in objective metrics such as PSNR and MS-SSIM. To minimize the impact of structural inconsistency during reference generation on the quality of the reconstructed image, we

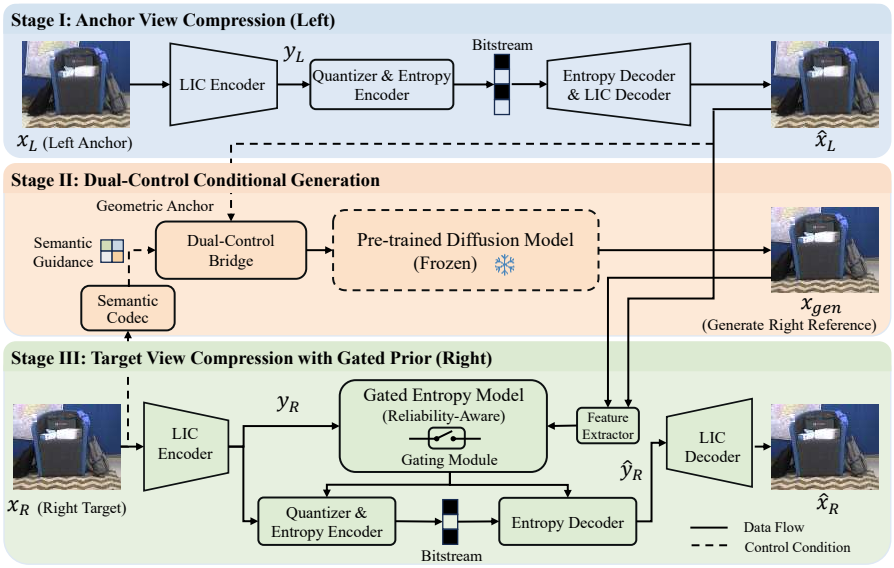

Fig. 1: Overview of the proposed generative stereo image compression framework. The system integrates Joint Semantic-Spatial Control for high-quality reference synthesis and a Reliability-Aware Gating module for adaptive feature integration across three optimized stages.

design a Reliability-Aware Gating module. By dynamically evaluating the synthesized prior in the feature space, this module adaptively regulates generative information flow within the entropy model. Specifically, if the generative features associated with a synthesized region demonstrate high uncertainty or misalignment, the gate suppresses this unreliable guidance, falling back to the robust geometric context. This sophisticated gating mechanism ensures that our framework maintains strict adherence to the objective rate-distortion curve without sacrificing generative richness. The overall pipeline of our proposed framework is illustrated in Fig. 1.

By seamlessly integrating controllable reference synthesis with a conditional entropy model robust to structural inconsistency, we establish a cohesive framework for generative stereo image compression. Crucially, rather than a naive application of generative models, this architecture operates on two synergistic pillars. First, it deterministically steers the stochastic synthesis using Joint Semantic-Spatial Control to overcome extreme reference degradation and synthesize a high-quality target-view reference. Second, it resolves the inherent conflict between stochastic generative processes and strict pixel-level alignment through the proposed Reliability-Aware Gating module. By strategically evaluating and regulating the generative prior, our approach redefines inter-view prediction as a controllable generative completion task while ensuring robust objective performance.

In summary, the **main contributions** of this paper are as follows:

(1) We propose a generative stereo image compression framework driven by a controllable generative reference. Instead of extracting explicit correspondences from a degraded reference, we leverage pre-trained diffusion priors to synthesize a high-quality target-view reference, effectively overcoming the fundamental limits of out-of-view (OOV) regions and severe quantization degradation.

(2) We introduce Joint Semantic-Spatial Control to synergistically steer the generative synthesis. This control seamlessly combines a robust geometric anchor from the compressed left view with an ultra-compact target-view semantic condition, restoring degraded textures with high perceptual fidelity.

(3) We design a Reliability-Aware Gating module to mitigate the impact of structural inconsistency inherent in generative priors. By evaluating the synthesized features, it adaptively regulates the information flow within the entropy model, ensuring strict adherence to the objective rate-distortion curve.

(4) Extensive experiments on standard benchmarks demonstrate that our framework achieves superior rate-distortion performance, delivering remarkable bitrate savings and high visual quality, especially in the low-bitrate regime.

2 Related Work

2.1 Correspondence-Driven Stereo Image Compression

Stereo image compression (SIC) [22, 29, 37, 49] fundamentally relies on exploiting cross-view redundancies. Existing methodologies are predominantly driven by establishing spatial correspondences. Early learned frameworks adopted explicit geometric alignment, utilizing disparity-based warping [3, 22, 37, 38, 45] or latent shift operations [42] to predict the target view. While effective for regions with high spatial correlation, the performance of these explicit methods degrades severely in out-of-view (OOV) regions, where physical correspondences are fundamentally absent.

To capture complex dependencies, recent state-of-the-art methods shift toward implicit feature interaction. These architectures employ bidirectional attention [20, 30, 41], 3D convolutions [25], or masked image modeling [11, 48] to transfer contextual information. Despite improved robustness, they still fundamentally rely on extracting valid features from the compressed reference. Under heavy compression, the reference features are heavily quantized. Forcing the network to extract contexts from these degraded correspondences inevitably propagates quantization artifacts to the target view, resulting in blurred textures and prediction failure. To overcome this bottleneck, our approach abandons the reliance on finding explicit correspondences from a degraded reference. Instead, we leverage the powerful priors of a pre-trained diffusion model to synthesize a high-quality target-view reference.

2.2 Generative Image Compression and Conditional Synthesis

Traditional learned image compression typically optimizes for pixel-wise distortion, which inevitably leads to over-smoothed textures at low bitrates [21, 23,

24,26]. To prioritize perceptual quality, generative codecs incorporate generative adversarial networks (GANs) [2, 19, 28] or diffusion models [32, 43, 47] to reconstruct high-frequency details. While effective for single images, extending these unconstrained priors to stereo compression poses a theoretical challenge. Their inherent stochasticity risks introducing severe structural inconsistency, violating necessary cross-view consistency and penalizing objective metrics.

To regulate this stochasticity, conditional frameworks like ControlNet [31, 44, 46] incorporate spatial anchors to deterministically guide diffusion [15, 33]. In stereo vision, DiffStereo [9] explores diffusion for restoration. However, such approaches target unconstrained image enhancement rather than bandwidth-constrained compression. Bridging this gap, our framework integrates conditional diffusion priors into the stereo compression pipeline. By utilizing the compressed left view as a robust geometric anchor, we formulate inter-view prediction as a controllable generative completion task. This approach effectively bridges the gap between stochastic generative processes and strict stereo consistency, ensuring robust objective performance without sacrificing generative richness.

3 Proposed Method

3.1 Overview

Let a stereo image pair be denoted as (x_L, x_R), where x_L is the left view and x_R is the right target view. Our goal is to encode this pair and reconstruct $(\hat{x}_L, \hat{x}_R)$ by optimizing the rate-distortion tradeoff. Departing from conventional correspondence-driven paradigms, we propose a novel framework that leverages the powerful priors of a pre-trained diffusion model to synthesize a high-quality target-view reference. To bridge the gap between stochastic generative processes and strict stereo consistency, we systematically integrate these generative priors into a unified compression pipeline. As illustrated in Fig. 1, our proposed framework operates in three distinct stages, adopting an asymmetric coding approach:

(1) Anchor View Compression. We first compress the left view x_L using a standard learned image compression (LIC) network. The reconstructed left view, $\hat{x}_L$, serves a dual purpose: it provides the final output for the left view and acts as a robust geometric anchor for compressing the right view.

(2) Generative Reference via Joint Semantic-Spatial Control. This stage is central to our approach. Instead of performing conventional feature alignment, we synthesize a high-quality target-view reference, x_{gen}. To ensure x_{gen} maintains strict stereo consistency with x_R, we employ a frozen Stable Diffusion model steered by the proposed Joint Semantic-Spatial Control. This mechanism synergistically combines two complementary conditions: a spatial condition derived from the robust geometric anchor ($\hat{x}_L$), and a content condition derived from an ultra-compact semantic condition extracted from x_R, denoted as $\hat{f}_{sem}$. Through this joint guidance mechanism, we successfully regulate the stochastic generative process.

(3) Target View Compression with Reliability-Aware Gating. Finally, the right view x_R is conditionally compressed, employing an entropy model

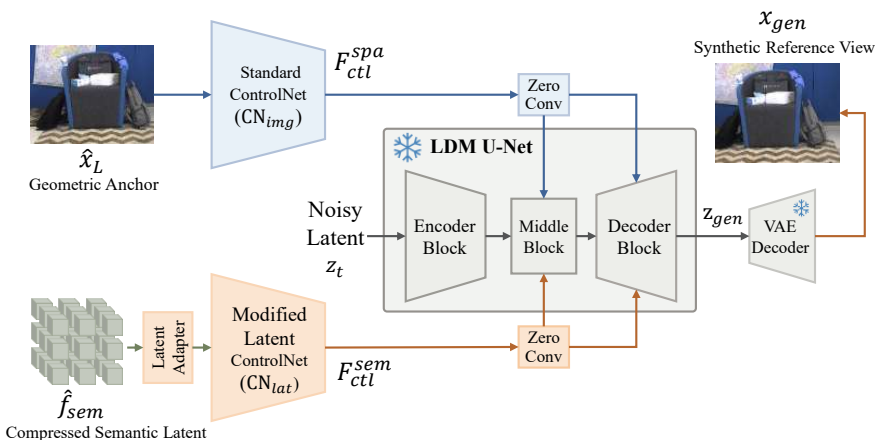

Fig. 2: Detailed architecture of the Joint Semantic-Spatial Control (Stage II). The module synthesizes the high-quality target-view reference x_{gen} by steering a frozen LDM backbone with dual control networks, integrating spatial geometry from the anchor $\hat{x}_L$ and semantic priors from the ultra-compact condition $\hat{f}_{sem}$.

that utilizes a hybrid context from both the geometric anchor ($\hat{x}_L$) and the synthesized reference (x_{gen}). Recognizing that the generative process may introduce structural inconsistency during reference generation, we design the Reliability-Aware Gating module within the entropy model. By dynamically evaluating the synthesized prior in the feature space, this module adaptively regulates generative information flow within the entropy model. This ensures that the system minimizes the impact of structural inconsistency on the final reconstructed image quality, effectively falling back to the robust geometric context when the generative prior demonstrates misalignment.

3.2 Generative Reference via Joint Semantic-Spatial Control

This stage aims to synthesize a high-quality target-view reference, x_{gen}. It serves as a powerful generative prior for the subsequent target view compression. To harness generative capabilities without prohibitive training costs, we employ a pre-trained, frozen Latent Diffusion Model (LDM) [33] as our backbone.

A core challenge lies in bridging the gap between stochastic generative processes and strict stereo consistency, avoiding structural inconsistency during reference generation. Relying solely on the reconstructed left view is fundamentally insufficient; it not only fails to infer occluded out-of-view (OOV) regions but also struggles to provide high-frequency textures because the reference features are heavily quantized under severe compression. Conversely, relying solely on an ultra-compact representation extracted from the target view lacks explicit spatial definition. To address this, we design the Joint Semantic-Spatial Control to synergistically combine these complementary conditions. In our implementation,

as illustrated in Fig. 2, we instantiate this mechanism by deploying two independent control networks [46] operating in parallel to elegantly process the spatial and semantic conditions.

(1) Spatial Control Using the Geometric Anchor. The reconstructed left view $\hat{x}_L$ serves as a robust geometric anchor, providing explicit spatial constraints for co-visible regions. Processed via a standard control network (CN_{img}), it yields a set of multi-scale spatial control features:

$$\{F_{ctl}^{spa}\} = \mathrm{CN}_{img}(\hat{x}_L, z_t, t), \tag{1}$$

where z_t is the noisy latent at timestep t.

(2) Semantic Control Using an Ultra-Compact Condition. To guide generation beyond mere local alignment, we utilize an ultra-compact semantic condition $\hat{f}_{sem}$. Specifically, the target view x_R is mapped into the latent space via a frozen VAE encoder and compressed by a lightweight semantic codec to yield $\hat{f}_{sem}$. This condition is crucial for inferring out-of-view (OOV) regions and restoring degraded textures caused by heavily quantized reference features. To maintain the overall rate-distortion optimization, this auxiliary compression operates in an extremely low-bitrate regime, incurring a negligible average overhead of approximately 0.00065 bpp. At this minimal cost, $\hat{f}_{sem}$ acts as a semantic seed, injecting macro-layout priors into the diffusion process while reserving the primary bitrate budget for the final target view compression. Because $\hat{f}_{sem}$ is derived directly from the VAE latent space, we design a modified latent control network (CN_{lat}). By replacing the initial image-domain feature extractors with a lightweight Latent Adapter, we align the dimensions of $\hat{f}_{sem}$ and reconstruct the necessary local spatial context:

$$\{F_{ctl}^{sem}\} = \mathrm{CN}_{lat}(\mathrm{Adapter}(\hat{f}_{sem}), z_t, t). \tag{2}$$

(3) Joint Semantic-Spatial Control for Reference Generation. The multi-scale features $\{F_{ctl}^{spa}\}$ and $\{F_{ctl}^{sem}\}$ are fused via element-wise summation and injected into the frozen U-Net decoder blocks through zero-convolutions. This strategy synergistically steers the synthesis: the spatial branch deterministically anchors the base geometry, while the semantic branch dictates the target disparity, supplies texture priors, and infers missing OOV structures. Finally, the fully denoised latent z_{gen} is decoded to produce the high-quality target-view reference: $x_{gen} = \mathrm{VAE}_{dec}(z_{gen})$.

3.3 Target View Compression with Reliability-Aware Gating

The final stage conditionally compresses the target view x_R by adopting a generalized Gaussian entropy model for its quantized latent $\hat{y}_R$. Our goal is to refine the initial mean and scale parameters $(\mu_{Base}, \sigma_{Base})$ estimated by a hyperprior network, leveraging complementary contexts from both reference views (Fig. 3).

(1) Hierarchical Gated Refinement. We first extract features from the robust geometric anchor and the synthesized high-quality target-view reference

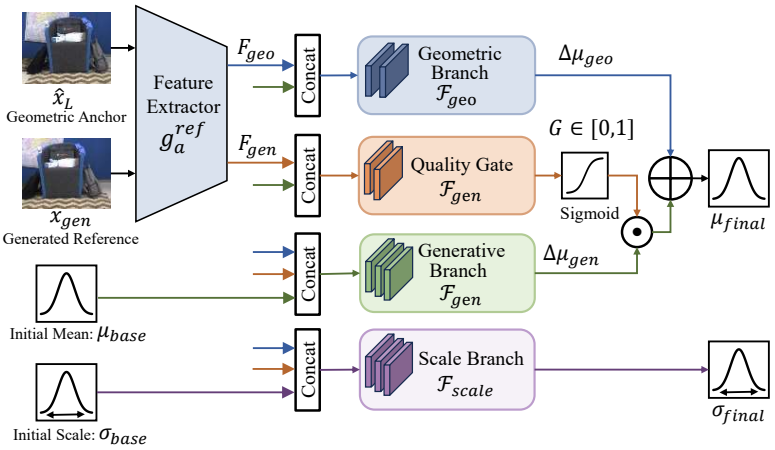

Fig. 3: Architecture of the Reliability-Aware Gating module (Stage III). It refines entropy parameters by using a spatial confidence map G to adaptively fuse geometric (F_{geo}) and generative (F_{gen}) priors, regulating information flow to mitigate structural inconsistency.

using a shared extractor g_a^{ref}: $F_{geo} = g_a^{ref}(\hat{x}_L)$ and $F_{gen} = g_a^{ref}(x_{gen})$. To refine μ_{Base}, a Geometric Refinement Branch initially computes a base offset $\Delta\mu_{geo}$ using the reliable spatial context:

$$\Delta\mu_{geo} = \mathcal{F}_{geo}([\mu_{Base}, F_{geo}]), \tag{3}$$

where $\mathcal{F}_{geo}$ is a residual network and $[\cdot]$ denotes concatenation. Subsequently, a Generative Refinement Branch computes a potentially more informative but riskier offset $\Delta\mu_{gen}$, conditioned on both references:

$$\Delta\mu_{gen} = \mathcal{F}_{gen}([\mu_{Base}, F_{geo}, F_{gen}]). \tag{4}$$

Crucially, to assess the trustworthiness of the generated features, we introduce the Reliability-Aware Gating module. By dynamically evaluating the synthesized prior in the feature space, this lightweight CNN predicts a spatial confidence map $G \in [0, 1]$ indicating the reliability of F_{gen}:

$$G = \sigma(\mathcal{F}_{Gate}([\mu_{Base}, F_{gen}])), \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function. A value of G approaching 0 indicates unreliable synthesis (e.g., severe structural inconsistency). To adaptively regulate the generative information flow, the final refined mean μ_{Final} is calculated by weighting the generative offset with the gate G and adding it to the previous estimates:

$$\mu_{Final} = \mu_{Base} + \Delta\mu_{geo} + (G \odot \Delta\mu_{gen}), \tag{6}$$

where $\odot$ denotes element-wise multiplication. This sophisticated gating mechanism ensures the model selectively incorporates generative priors only when they

are deemed structurally consistent, effectively falling back to the robust geometric context otherwise. Finally, the scale parameter σ_{Final} is refined by jointly evaluating all available contexts:

$$\sigma_{Final} = \mathcal{F}_{Scale}([\sigma_{Base}, F_{geo}, F_{gen}]). \tag{7}$$

(2) Reliability-Aware Training Scheme. To ensure the gate G effectively identifies unreliable generated content, we employ a targeted corruption approach during training. We randomly degrade F_{gen} by either dropping it (setting it to zero with probability p_{Drop}) or injecting Gaussian noise (with probability p_{Noise}). This forces $\mathcal{F}_{Gate}$ to recognize non-informative generative features and adaptively suppress this unreliable guidance, thereby minimizing the impact of structural inconsistency and ensuring robust objective performance during inference.

3.4 Loss Functions

To ensure training stability and mitigate interference between anchor compression and the generative process, we adopt a stage-wise training protocol. During each subsequent stage, the modules optimized in all preceding stages are strictly frozen. The specific optimization objective for each stage is detailed below.

(1) Anchor View Compression. This stage establishes the robust geometric anchor. We optimize the anchor compression branch using the standard rate-distortion (R-D) loss:

$$\mathcal{L}_{\text{Stage I}} = \lambda \cdot D(x_L, \hat{x}_L) + R(\hat{y}_L) + R(\hat{z}_L), \tag{8}$$

where $D(\cdot)$ denotes the Mean Squared Error (MSE), while $R(\hat{y}_L)$ and $R(\hat{z}_L)$ represent the bitrates of the anchor view's main latent and its hyperprior, respectively. The Lagrange multiplier λ governs the rate-distortion trade-off.

(2) Joint Generative and Semantic Optimization. In this stage, the Joint Semantic-Spatial Control and the lightweight semantic codec are jointly optimized, while the LDM U-Net and VAE remain strictly frozen. The objective is twofold: 1) ensure the synthesized high-quality target-view reference x_{gen} maintains strict stereo consistency, and 2) minimize the bitrate overhead of the ultra-compact semantic condition $\hat{f}_{sem}$. The total objective function combines the diffusion loss and the semantic compression loss:

$$\mathcal{L}_{\text{Stage II}} = \lambda_{gen} \cdot \mathcal{L}_{Diff} + \lambda_{sem} \cdot \mathcal{L}_{Codec}. \tag{9}$$

The diffusion loss $\mathcal{L}_{Diff}$ evaluates the generation quality via the MSE between the added and predicted noise:

$$\mathcal{L}_{Diff} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \hat{x}_L, \hat{f}_{sem})\|_2^2\right]. \tag{10}$$

Crucially, gradients from $\mathcal{L}_{Diff}$ are back-propagated through the modified latent control network directly to the semantic codec, compelling $\hat{f}_{sem}$ to capture optimal macro-layout priors and bridge the gap between stochastic generation and

strict stereo consistency. The semantic codec loss $\mathcal{L}_{Codec}$ constrains the bitrate and ensures representational stability:

$$\mathcal{L}_{Codec} = R(\hat{f}_{sem}) + \beta \cdot \|f_{sem} - \hat{f}_{sem}\|_2^2, \tag{11}$$

where f_{sem} is the uncompressed latent extracted from x_R (utilizing the mode of the VAE distribution for stability), and β balances feature reconstruction fidelity against the bitrate constraint.

(3) Target View Compression with Reliability-Aware Gating. Finally, we train the target view compression branch equipped with the Reliability-Aware Gating module. The goal is to minimize the rate-distortion (R-D) cost of the target view x_R by leveraging the hybrid contexts derived from the robust geometric anchor ($\hat{x}_L$) and the high-quality target-view reference (x_{gen}):

$$\mathcal{L}_{\text{Stage III}} = \lambda \cdot D(x_R, \hat{x}_R) + R(\hat{y}_R) + R(\hat{z}_R). \tag{12}$$

4 Experimental Results

4.1 Experimental Setup

The experimental setup is summarized as follows.

(1) Datasets. We evaluate our framework on two standard stereo benchmarks: InStereo2K [4] (1080×860 resolution, with 2,010 training and 50 testing pairs) and Cityscapes [10] (2048×1024 resolution, with 2,975 training and 1,525 testing pairs). We follow established preprocessing protocols (e.g., boundary cropping) from recent works [25, 41], deferring detailed dataset descriptions and exact cropping parameters to the supplementary material.

(2) Baselines. We compare against a comprehensive set of state-of-the-art baselines, including traditional codecs (BPG [5], HEVC [36], VVC [8], MV-HEVC [37]) and recent learned stereo compression methods (BCSIC [20], LDMIC [49], ECSIC [41], BiSIC [25], and CAMSIC [48]).

(3) Evaluation Metrics. We evaluate objective quality using Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity (MS-SSIM) [40]. Coding efficiency is quantified via the Bjøntegaard Delta Bitrate (BDBR) [6], which calculates the average percentage of bitrate savings relative to an anchor method. Lower BDBR values indicate superior compression performance.

(4) Implementation Details. To synthesize the high-quality target-view reference, we utilize a frozen Stable Diffusion v1.5 backbone steered by the proposed Joint Semantic-Spatial Control. Following the stage-wise training protocol (Sec. 3.4), we optimize the framework using the AdamW optimizer, employing randomly cropped 256×256 patches for the compression networks and 512×512 patches for the generative module to accommodate the diffusion prior. To evaluate varying rate-distortion tradeoffs, we train distinct models utilizing standard λ sets for both MSE and MS-SSIM objectives. Comprehensive training details—including exact λ values, learning rate schedules, iteration steps, and detailed network architectures—are provided in the supplementary material.

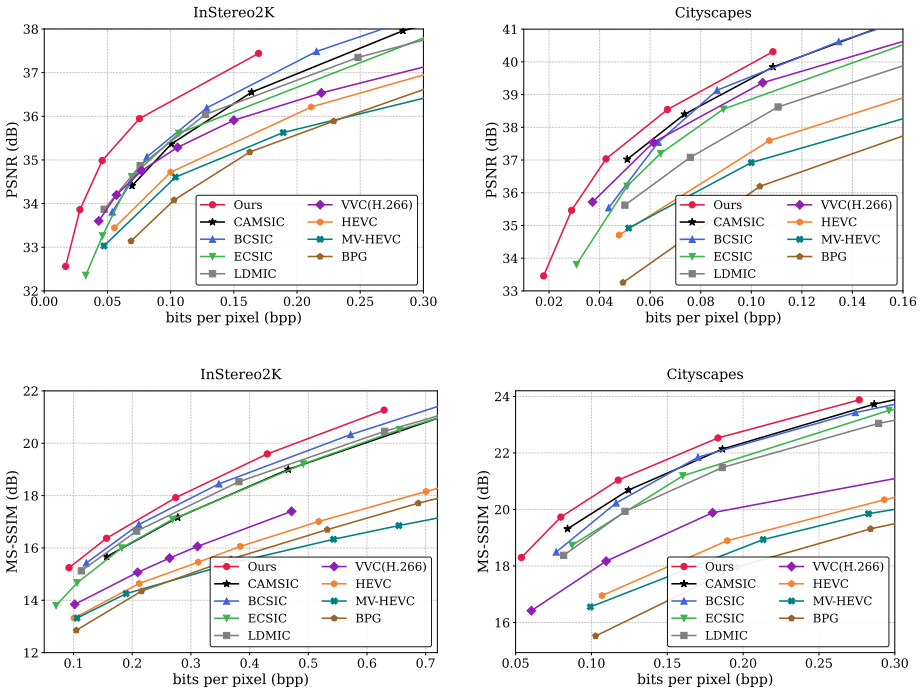

Fig. 4: Rate-distortion curves in terms of PSNR and MS-SSIM on the InStereo2K and Cityscapes datasets.

4.2 Results

(1) Rate-Distortion Performance. We evaluate Rate-Distortion (R-D) performance using Bjøntegaard Delta Bitrate (BDBR) (Table 1) and R-D curves (Fig. 4). As demonstrated in Table 1, our framework establishes a new state-of-the-art across both datasets. Anchored on BPG, it achieves remarkable BDBR savings on InStereo2K of 67.64% in PSNR and 65.22% in MS-SSIM. This significantly outperforms the second-best BiSIC [25] in PSNR (48.08% savings) while maintaining a clear lead in MS-SSIM (61.13%). Similar dominance is observed on Cityscapes with a 65.90% PSNR BDBR reduction, surpassing CAMSIC [48] (59.74%) and VVC (57.52%). These substantial improvements validate the superior coding efficiency of our synthesized high-quality target-view reference.

Fig. 4 further illustrates that our gains are particularly pronounced in the low-bitrate regime. On InStereo2K, at a low bitrate of roughly 0.07 bpp, our model achieves a PSNR of 35.75 dB, outperforming CAMSIC and BiSIC by over 1 dB. It also maintains top-tier MS-SSIM, yielding 21.1 dB at 0.6 bpp versus BiSIC's 20.6 dB. On Cityscapes, where conventional correspondence-driven methodologies (e.g., ECSIC [41]) often struggle at low bitrates, our generative approach remains robust. At approximately 0.03 bpp, it surpasses ECSIC by 1.5 dB in PSNR, while at around 0.3 bpp, it outperforms VVC in MS-SSIM

Table 1: Comparison of BDBR cost relative to BPG on different datasets, with the best results in **Bold** and second-best ones in underlined.

Method	InStereo2K		Cityscapes	
	PSNR	MS-SSIM	PSNR	MS-SSIM
HEVC [36]	-21.50%	-12.55%	-28.33%	-24.50%
VVC [8]	-35.74%	-30.31%	-57.52%	-46.27%
MV-HEVC [37]	-8.52%	5.58%	-19.02%	-16.89%
LDMIC [49]	-45.58%	-58.72%	-42.29%	-63.23%
ECSIC [41]	-44.66%	-55.41%	-51.13%	-65.20%
BiSIC [25]	-48.08%	-61.13%	-57.08%	-67.74%
CAMSIC [48]	-43.88%	-53.89%	-59.74%	-69.67%
Ours	**-67.64%**	**-65.22%**	**-65.90%**	**-74.53%**

by 2.0 dB. Notably, this state-of-the-art performance requires negligible bitrate overhead. The ultra-compact semantic condition $\hat{f}_{sem}$ consumes an average of merely 0.00065 bpp. This proves that ultra-compact semantic guidance, when synergized with a frozen diffusion prior, is sufficient to synthesize high-quality target-view references that effectively resolve the inherent conflict between generative stochasticity and strict stereo consistency.

Table 2: Comparison of average runtime per image pair (in seconds) across diverse deep learning and standard codec architectures.

Codec	BPG	HEVC	VVC	ECSIC	BiSIC	CAMSIC	Ours
Codec Time (s)	16.24	31.15	208.42	1.47	1.29	1.34	1.60

(2) Runtime Comparison. Table 2 compares average runtimes on InStereo2K. Learned models are executed on an NVIDIA A100 GPU, whereas traditional codecs like VVC are executed on an AMD EPYC 7742 CPU. Our core compression network (0.78s) is highly efficient compared to ECSIC (1.47s) and BiSIC (1.29s). Even including reference synthesis (0.82s), our total inference time (1.60s) remains competitive with state-of-the-art learned models and orders of magnitude faster than exhaustive CPU-bound codecs like VVC (208.42s). Given the bitrate savings exceeding 65% relative to BPG, this generative overhead is a highly worthwhile trade-off for superior reconstruction quality.

4.3 Ablation Studies

In this section, we conduct ablation studies on the InStereo2K dataset to analyze the contribution of each proposed component, evaluating the Bjøntegaard Delta PSNR (BD-PSNR) relative to our full model (Table 3).

Table 3: Ablation study of key components on InStereo2K and Cityscapes. BD-PSNR is calculated relative to our full model. (Negative values indicate performance drop).

Model Variant	InStereo2K	Cityscapes
	BD-PSNR (↑)	BD-PSNR (↑)
Full Model (Ours)	0 dB	0 dB
w/o Generative Reference	-1.31 dB	-1.23 dB
w/o Semantic Control	-0.76 dB	-0.83 dB
w/o Reliability-Aware Gate	-1.07 dB	-0.91 dB

(1) Component Ablation. Specifically, removing the generative prior entirely yields the most severe performance degradation (-1.31 dB and -1.23 dB), confirming that replacing conventional correspondence-driven methodologies with controllable generative synthesis is the fundamental driver of our coding efficiency. Furthermore, without the ultra-compact semantic condition, the generative process suffers from spatial ambiguity, resulting in a performance reduction of -0.76 dB and -0.83 dB. This underscores that our minimal semantic guidance is crucial for deterministically steering the synthesis toward high stereo consistency. Finally, disabling the Reliability-Aware Gating module forces the entropy model to blindly incorporate the synthesized reference, which inherently introduces structural inconsistency during reference generation. This leads to significant penalties (-1.07 dB and -0.91 dB), highlighting the absolute necessity of dynamically regulating the generative information flow before feature integration to maintain robust objective performance.

(2) Visualization of Generative Priors. Fig. 5 presents a qualitative comparison in the low-bitrate regime. We highlight two critical challenges: severe texture degradation (highlighted boxes) and out-of-view (OOV) regions at the right boundary.

The severely compressed left view (Col. 1) is fundamentally degraded, resulting in heavily quantized features that lack sufficient information for the OOV area. In contrast, guided by an ultra-compact semantic condition (0.0008 bpp), our high-quality target-view reference (Col. 2) successfully restores degraded textures (e.g., basketball lines) and synthesizes missing OOV structures with high visual quality. Col. 4 visualizes the inherent perception-distortion tradeoff. While our synthesized reference provides excellent perceptual quality, its generated details may not always be strictly aligned with the ground truth (Col. 3), potentially introducing structural inconsistency. Consequently, under a strict bit budget (0.0079 bpp), the subsequent MSE-optimized reconstruction suppresses these unaligned textures to maintain robust objective performance.

This demonstrates the controllability of our framework: the generative prior establishes a rich perceptual foundation, while the Reliability-Aware Gating module adaptively regulates the information flow to minimize the impact of structural inconsistency on the final reconstructed image quality. Furthermore, our framework demonstrates remarkable zero-shot Sim-to-Real generalization

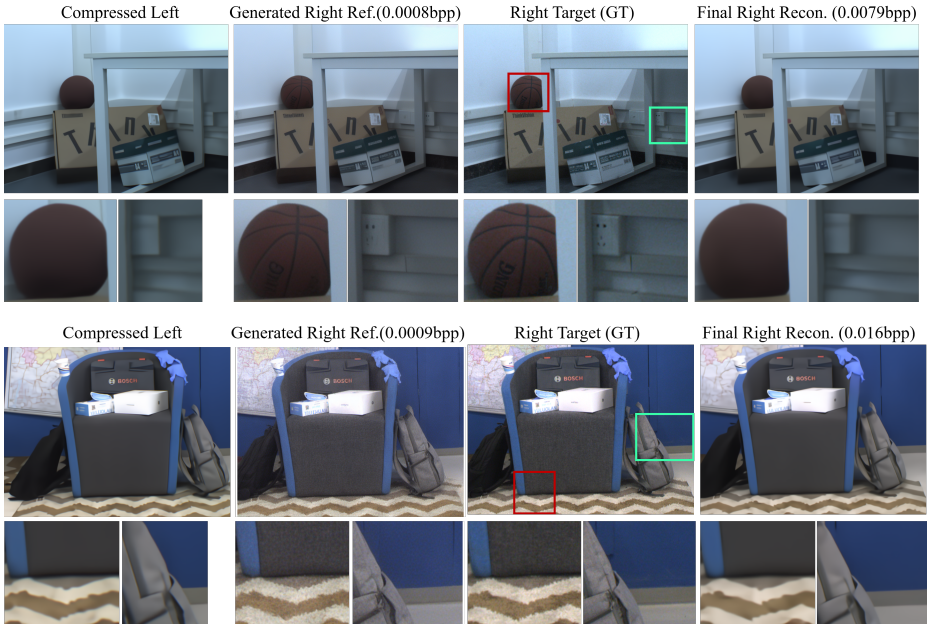

Fig. 5: Qualitative comparison at low bitrates across two distinct examples, demonstrating our capability in restoring degraded textures (highlighted boxes) and synthesizing missing out-of-view regions.

capabilities (trained entirely on the synthetic IRS_small dataset and evaluated zero-shot on the real-world InStereo2K). Detailed generalization analyses and additional qualitative results are deferred to the supplementary material.

5 Conclusion

In this paper, we presented a novel generative framework for stereo image compression. By steering a pre-trained diffusion prior via Joint Semantic-Spatial Control, our method synthesizes a high-quality target-view reference guided by an ultra-compact semantic condition. This paradigm fundamentally overcomes the limitations of conventional correspondence-driven methodologies, effectively synthesizing missing out-of-view (OOV) structures and restoring degraded textures. To ensure robust objective performance, we introduced the Reliability-Aware Gating module to adaptively regulate the generative information flow within the entropy model, thereby mitigating the impact of structural inconsistency. Extensive experiments validate the superior rate-distortion performance and high perceptual fidelity of our approach. Ultimately, this architecture successfully resolves the inherent conflict between stochastic generative processes and strict pixel-level alignment, demonstrating great potential for advanced binocular vision applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62331014, 12426312). We acknowledge the computational support of the Center for Computational Science and Engineering at Southern University of Science and Technology.

References

1. Agustsson, E., Minnen, D., Toderici, G., Mentzer, F.: 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)
2. Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., Gool, L.V.: Generative adversarial networks for extreme learned image compression. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 221–231 (2019)
3. Ayzik, S., Avidan, S.: Deep image compression using decoder side information. In: European Conference on Computer Vision. pp. 699–714. Springer (2020)
4. Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S., Zhang, X.: Instereo2k: a large real dataset for stereo matching in indoor scenes. Science China Information Sciences **63**(11), 212101 (2020)
5. Bellard, F.: BPG image format. `https://bellard.org/bpg/` (2014), accessed 26 June 2026
6. Bjontegaard, G.: Calculation of average psnr differences between rd-curves. ITU-T SG16, Doc. VCEG-M33 (2001)
7. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
8. Bross, B., Wang, Y.K., Ye, Y., Liu, S., Chen, J., Sullivan, G.J., Ohm, J.R.: Overview of the versatile video coding (vvc) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology **31**(10), 3736–3764 (2021)
9. Cao, H., Shi, Y., Xia, B., Jin, X., Yang, W.: DiffStereo: High-frequency aware diffusion model for stereo image restoration. arXiv preprint arXiv:2501.10325 (2025)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
11. Deng, X., Deng, Y., Yang, R., Yang, W., Timofte, R., Xu, M.: MASIC: Deep mask stereo image compression. IEEE Transactions on Circuits and Systems for Video Technology **33**(10), 6026–6040 (2023)
12. Deng, X., Yang, W., Yang, R., Xu, M., Liu, E., Feng, Q., Timofte, R.: Deep homography for efficient stereo image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1492–1501 (2021)
13. Geiger, A., Lenz, P., Urtasun, R.: 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)
14. Hegde, D., Yasarla, R., Cai, H., Han, S., Bhattacharyya, A., Mahajan, S., Liu, L., Garrepalli, R., Patel, V.M., Porikli, F.: Distilling multi-modal large language models for autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27575–27585 (2025)

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
16. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
17. Huang, Y., Chen, B., Qin, S., Li, J., Wang, Y., Dai, T., Xia, S.T.: Learned distributed image compression with multi-scale patch matching in feature domain. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 4322–4329 (2023)
18. Jin, Y., Liu, X., Liu, J.: Generative ai for multimedia communication: Recent advances, an information-theoretic framework, and future opportunities. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 12237–12246 (2025)
19. Körber, N., Kromer, E., Siebert, A., Hauke, S., Mueller-Gritschneder, D., Schuller, B.: EGIC: enhanced low-bit-rate generative image compression guided by semantic segmentation. In: European Conference on Computer Vision. pp. 202–220. Springer (2024)
20. Lei, J., Liu, X., Peng, B., Jin, D., Li, W., Gu, J.: Deep stereo image compression via bi-directional coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19669–19678 (2022)
21. Li, H., Li, S., Dai, W., Li, C., Zou, J., Xiong, H.: Frequency-aware transformer for learned image compression. In: The Twelfth International Conference on Learning Representations (2024)
22. Liu, J., Wang, S., Urtasun, R.: DSIC: Deep stereo image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3136–3145 (2019)
23. Liu, J., Sun, H., Katto, J.: Learned image compression with mixed transformer-cnn architectures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14388–14397 (2023)
24. Liu, Y., Yang, W., Bai, H., Wei, Y., Zhao, Y.: Region-adaptive transform with segmentation prior for image compression. In: European Conference on Computer Vision. pp. 181–197. Springer (2024)
25. Liu, Z., Zhang, X., Shao, J., Lin, Z., Zhang, J.: Bidirectional stereo image compression with cross-dimensional entropy model. In: European Conference on Computer Vision. pp. 480–496. Springer (2024)
26. Lu, J., Zhang, L., Zhou, X., Li, M., Li, W., Gu, S.: Learned image compression with dictionary-based entropy model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12850–12859 (2025)
27. Ma, Y., Yang, W., Liu, J.: Correcting diffusion-based perceptual image compression with privileged end-to-end decoder. In: International Conference on Machine Learning. pp. 34075–34093. PMLR (2024)
28. Mentzer, F., Toderici, G.D., Tschannen, M., Agustsson, E.: High-fidelity generative image compression. In: Advances in Neural Information Processing Systems. vol. 33, pp. 11913–11924 (2020)
29. Mital, N., Özyılkan, E., Garjani, A., Gündüz, D.: Neural distributed image compression using common information. In: 2022 Data Compression Conference (DCC). pp. 182–191. IEEE (2022)
30. Mital, N., Özyilkan, E., Garjani, A., Gündüz, D.: 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)

31. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
32. Relic, L., Azevedo, R., Gross, M., Schroers, C.: Lossy image compression with foundation diffusion models. In: European Conference on Computer Vision. pp. 303–319. Springer (2024)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
34. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
35. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision **47**(1), 7–42 (2002)
36. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on circuits and systems for video technology **22**(12), 1649–1668 (2012)
37. Tech, G., Chen, Y., Müller, K., Ohm, J.R., Vetro, A., Wang, Y.K.: Overview of the multiview and 3d extensions of high efficiency video coding. IEEE Transactions on Circuits and Systems for Video Technology **26**(1), 35–49 (2015)
38. Vetro, A., Wiegand, T., Sullivan, G.J.: Overview of the stereo and multiview video coding extensions of the h. 264/mpeg-4 avc standard. Proceedings of the IEEE **99**(4), 626–642 (2011)
39. Wallace, G.K.: The JPEG still picture compression standard. IEEE transactions on consumer electronics **38**(1), xviii–xxxiv (1992)
40. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers. vol. 2, pp. 1398–1402. Ieee (2003)
41. Wödlinger, M., Kotera, J., Keglevic, M., Xu, J., Sablatnig, R.: 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)
42. Wödlinger, M., Kotera, J., Xu, J., Sablatnig, R.: 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)
43. Yang, R., Mandt, S.: Lossy image compression with conditional diffusion models. In: Advances in Neural Information Processing Systems. vol. 36, pp. 64971–64995 (2023)
44. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
45. Zhai, Y., Tang, L., Ma, Y., Peng, R., Wang, R.: Disparity-based stereo image compression with aligned cross-view priors. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2351–2360 (2022)
46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
47. Zhang, T., Luo, X., Li, L., Liu, D.: StableCodec: Taming one-step diffusion for extreme image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17379–17389 (2025)

48. Zhang, X., Gao, S., Liu, Z., Shao, J., Ge, X., He, D., Xu, T., Wang, Y., Zhang, J.: CAMSIC: Content-aware masked image modeling transformer for stereo image compression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10239–10247 (2025)
49. Zhang, X., Shao, J., Zhang, J.: LDMIC: Learning-based distributed multi-view image coding. In: The Eleventh International Conference on Learning Representations (2023)
50. Zheng, W., Chen, W., Huang, Y., Zhang, B., Duan, Y., Lu, J.: Occworld: Learning a 3d occupancy world model for autonomous driving. In: European conference on computer vision. pp. 55–72. Springer (2024)