RECASTING DIT FOR EXTREME IMAGE COMPRESSION WITH HIERARCHICAL VISUAL PRIORS

Anonymous authors
Paper under double-blind review

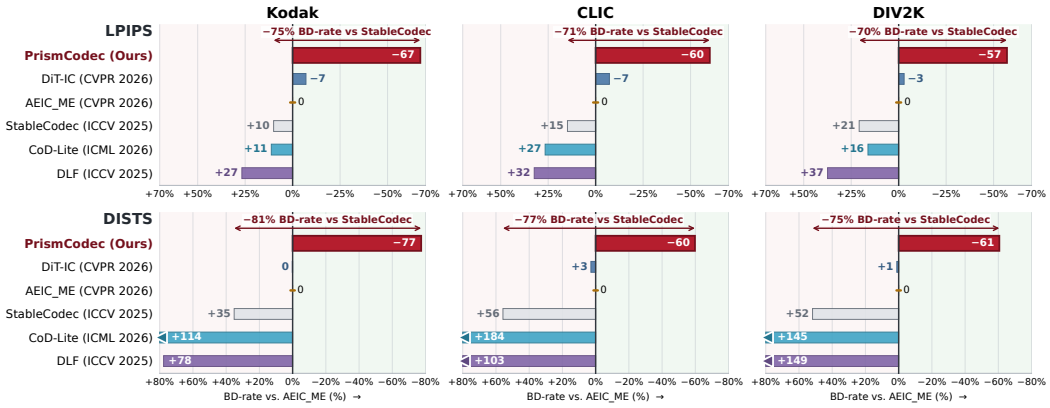

Figure 1: **Perceptual bitrate savings.** LPIPS- and DISTS-based BD-rate comparisons on Kodak, CLIC, and DIV2K, with AEIC-ME as the bar reference (0%). Negative values indicate bitrate savings and extend to the right. Red arrows report PrismCodec's direct BD-rate relative to StableCodec. Triangles mark bars clipped at the axis limit; labels retain the unclipped values, rounded to integers.

ABSTRACT

Pretrained diffusion models provide powerful generative priors for extreme image compression, but their stochastic, iterative generation processes are not designed to decode compressed representations directly. We present PrismCodec, which recasts a pretrained Diffusion Transformer (DiT) as a bitstream-grounded one-step generative decoder. PrismCodec uses the decoded representation in two complementary roles: a projected codec state defines the starting state for generation, while hierarchical visual priors derived from the same decoded latent provide semantic and structural guidance. To bridge the gap between stochastic, iterative generation and deterministic one-step decoding, we develop a progressive adaptation strategy. We first learn codec-conditioned transport from a noisy neighborhood centered on the codec state, distill it into a one-step mapping, and finally anneal the source perturbation to zero during end-to-end optimization. The resulting decoder reconstructs deterministically from the received representation without additional conditioning bits. Across Kodak, CLIC, and DIV2K, PrismCodec consistently improves rate–perception performance over prior generative codecs, achieving BD-rate reductions of 70–75% in LPIPS and 75–81% in DISTS over StableCodec. Code will be released upon acceptance.

1 INTRODUCTION

At ultra-low bitrates, learned codecs built on optimized transforms and entropy models (Ballé et al., 2017; Ballé et al., 2018; Minnen et al., 2018) can preserve only limited visual evidence of the source image. Generative compression therefore relies on image generation priors to synthesize details that are not explicitly transmitted. Yet compression differs fundamentally from unconstrained generation: plausible content is insufficient if it is inconsistent with the evidence retained in the bitstream. A

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.

central challenge is therefore to exploit powerful generative priors while grounding reconstruction in this limited transmitted evidence.

Pretrained diffusion and flow models offer powerful visual priors (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Esser et al., 2024), but their generative processes are not naturally aligned with compression decoding. Existing methods expose image-specific compressed evidence to the generative decoder through different interfaces. OneDC (Xue et al., 2025b) and DiT-IC (Shi et al., 2026a) derive semantic guidance directly from compressed representations, while AEIC (Zhang et al., 2026a) removes the text-conditioning pathway and reconstructs from codec latents. NeFIC (Chen et al., 2026c) and StableCodec (Zhang et al., 2025) retain fixed text conditions but also inject image-specific information through a decoded anchor and a codec latent, respectively. Together, these approaches show that effective generative decoding depends on how transmitted visual evidence is exposed to the generator. Two questions nevertheless remain: *how should that evidence determine both the generative starting state and the visual guidance, and how can a pretrained model built for stochastic, iterative generation be adapted to deterministic one-step decoding?*

We introduce **PrismCodec**, which recasts a pretrained DiT as a compression decoder whose starting state and visual guidance are both derived from the bitstream. These two roles are complementary: the codec-centered state anchors generation in the transmitted representation, while the visual priors specify semantic content and spatial structure that should be preserved during reconstruction. Inspired by recent image-generation studies (Shi et al., 2026b; Gong et al., 2026), we combine semantic and structural visual priors through a shared visual-evidence pathway. A lightweight decoder renders the decoded codec features as an internal evidence image, providing a common image-domain interface through which pretrained visual encoders can interpret the same bitstream-derived evidence. DINOv3 (Siméoni et al., 2025) supplies multi-depth dense semantic features, which are delivered to successive DiT block groups through depth-matched cross-attention rather than collapsed into a single condition. This preserves complementary levels of visual abstraction across the pretrained representation hierarchy. In parallel, the diffusion-native VAE supplies a spatially aligned structural prior through the latent-input pathway. Both priors are computed entirely at the decoder and therefore require no additional conditioning bits.

Visual conditioning alone, however, does not make a pretrained generator a practical compression decoder. Diffusion and flow models are typically trained to generate from stochastic sources through iterative transport, whereas practical compression typically favors deterministic reconstruction from a transmitted representation under a tight decoding budget. We therefore progressively reshape the generative process around the codec state. First, we learn conditional transport from a stochastic neighborhood centered on the codec representation to the clean generative latent, so that the learned trajectory is already grounded in transmitted evidence rather than generic noise. We then distill this codec-conditioned transport into a direct one-step mapping. Finally, during end-to-end compression training, we gradually contract the stochastic source distribution onto the deterministic codec state. This annealing avoids an abrupt shift from the stochastic source used for flow learning and distillation to the deterministic input used at deployment, while allowing the codec and generative decoder to adapt jointly. Fig. 1 summarizes the resulting rate–perception performance. Across Kodak, CLIC, and DIV2K, PrismCodec achieves LPIPS and DISTS BD-rate reductions of approximately 70–75% and 75–81%, respectively, over StableCodec. Our main contributions are:

- We recast a pretrained DiT as a bitstream-grounded generative decoder, using the compressed representation to determine both its starting state and its visual guidance.
- We introduce hierarchical visual priors that combine multi-depth dense semantic features with a diffusion-native structural prior through complementary DiT conditioning interfaces.
- We develop a progressive adaptation strategy that moves from codec-centered stochastic flow learning through one-step distillation to deterministic end-to-end decoding, bridging pretrained stochastic generation and practical extreme image compression.

2 Related Work

Generative Image Compression. Generative codecs use learned priors to recover plausible content at very low bitrates. Adversarial methods improve perceptual quality through distribution matching (Mentzer et al., 2020; Muckley et al., 2023), while conditional and residual diffusion reconstruct

images from compressed representations (Yang & Mandt, 2023; Ghouse et al., 2023). Recent work increasingly adapts pretrained generative priors, making bitstream grounding and efficient decoding central design concerns.

Bitstream Grounding and One-Step Adaptation. Pretrained-prior codecs expose transmitted evidence through different conditioning interfaces to guide reconstruction (Zhang et al., 2025; Chen et al., 2026c;d). StableCodec combines a codec latent with structural correction under fixed text conditioning. OneDC (Xue et al., 2025b) derives tokenizer-supervised semantic tokens from the hyperprior, DiT-IC (Shi et al., 2026a) uses text-aligned compressed features, and AEIC (Zhang et al., 2026a) removes cross-attention. PrismCodec instead separates bitstream-derived evidence into two roles: a codec state determines the generative source, while an evidence image provides hierarchical semantic and diffusion-native structural guidance through complementary pretrained encoders and separate DiT conditioning interfaces.

Fast decoding is addressed through diffusion pretraining and distillation. CoD (Jia et al., 2026b) pretrains diffusion for compression, CoD-Lite (Jia et al., 2026a) introduces lightweight one-step decoding, OSCAR (Guo et al., 2025) supports one-step reconstruction across multiple bitrates, and FlowCodec (Huang et al., 2026) adapts a pretrained flow prior for near-terminal transport. More detailed comparisons with prior methods are provided in Appendix A.

3 METHOD

3.1 OVERVIEW

PrismCodec recasts a pretrained DiT as a one-step generative decoder whose reconstruction is grounded in the received bitstream. Given a source image $\mathbf{x}$ encoded into a bitstream $\mathbf{b}$, entropy decoding produces compressed features from which we derive two complementary forms of information: a codec state $\mathbf{z}_c$ that determines the starting state for generation, and visual priors $\mathcal{P}$ that specify the semantic content and spatial structure to preserve during reconstruction. The decoder computes

$$\widehat{\mathbf{z}}_0 = \mathcal{G}_\theta(\mathbf{z}_c; \mathcal{P}), \qquad \widehat{\mathbf{x}} = \mathcal{D}_g(\widehat{\mathbf{z}}_0), \tag{1}$$

where $\mathcal{G}_\theta$ is a one-step latent mapping implemented by a DiT, and $\mathcal{D}_g$ maps the predicted latent back to image space through the generative VAE decoder, kept frozen throughout all stages.

A pretrained diffusion model, however, is designed for stochastic, iterative generation rather than deterministic decoding from a transmitted representation. We therefore adapt the generative process progressively. Stage I learns conditional transport from a stochastic neighborhood centered on the codec state, grounding the generative trajectory in transmitted evidence. Stage II distills this codec-conditioned transport into a one-step map. Stage III then contracts the remaining source randomness while jointly optimizing the codec and generator, continuously aligning the training-time source with the deterministic codec state used at deployment. Fig. 2 illustrates the visual interface and Stage-I adaptation, while Fig. 3 summarizes Stages II and III. Appendix B summarizes the notation.

3.2 BITSTREAM-DERIVED VISUAL EVIDENCE

The decoded bitstream provides the common source for both generation initialization and visual guidance. Entropy decoding and synthesis recover a feature $\mathbf{f}_c$, from which PrismCodec constructs the codec state $\mathbf{z}_c$ and an internal evidence image $\widetilde{\mathbf{x}}$. These two branches expose the same transmitted evidence through different interfaces: $\mathbf{z}_c$ determines where the generative process starts, whereas $\widetilde{\mathbf{x}}$ provides an image-domain view from which pretrained visual representations can be extracted.

Codec state. A learned projection $\mathcal{P}_c$ maps the decoded feature to

$$\mathbf{z}_c = \mathcal{P}_c(\mathbf{f}_c). \tag{2}$$

The projection places $\mathbf{z}_c$ in coordinates compatible with the generative VAE latent in spatial dimensions and numerical scale, without constraining it to a Gaussian distribution. In Stage I, $\mathbf{z}_c$ centers the stochastic source; after Stage III, it becomes the deterministic input to the one-step decoder. We denote the estimated bits per pixel (bpp) by $\mathcal{R}$.

Shared visual interface. In parallel, a lightweight evidence decoder $\mathcal{D}_e$ renders the same decoded feature into an image-domain representation,

$$\widetilde{\mathbf{x}} = \mathcal{D}_e(\mathbf{f}_c). \tag{3}$$

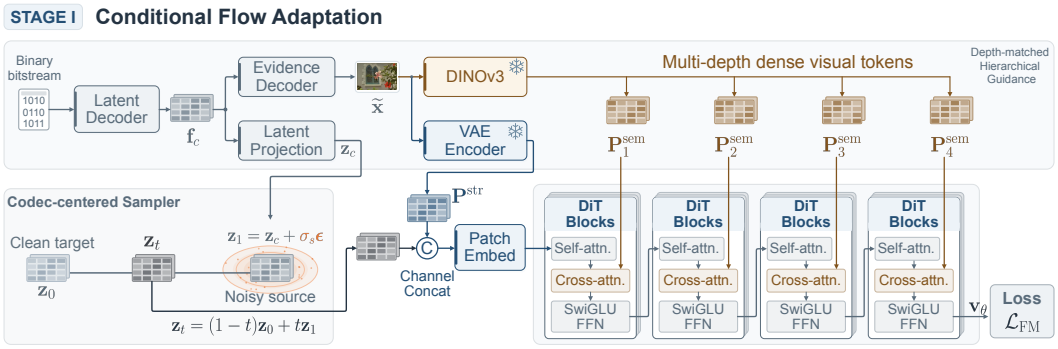

Figure 2: **Stage I: visual conditioning and codec-centered flow adaptation.** The decoded feature $\mathbf{f}_c$ yields a codec state $\mathbf{z}_c$ and an evidence image $\widetilde{\mathbf{x}}$ for two frozen encoders. DINOv3 supplies multi-depth dense semantic priors $\mathbf{P}^{\text{sem}}$ to successive DiT groups through depth-matched cross-attention. The VAE encoder supplies a spatially aligned structural prior $\mathbf{P}^{\text{str}}$, channel-concatenated with $\mathbf{z}_t$. The codec-centered sampler interpolates between the clean latent and a noisy source centered on $\mathbf{z}_c$.

Because $\widetilde{\mathbf{x}}$ is computed entirely from the received bitstream, it requires no additional transmission. More importantly, it provides a common image-domain interface through which the pretrained visual encoders interpret the same decoded evidence in the domain for which they were originally trained. We lightly supervise $\widetilde{\mathbf{x}}$ with an auxiliary objective $\mathcal{L}_{\text{ev}}$ so that it retains source layout, appearance, and semantic content. The objective combines pixel reconstruction with DINOv3 feature similarity. Together, $\mathbf{z}_c$ and $\widetilde{\mathbf{x}}$ expose the same decoded representation through the two roles required by the generative decoder: source initialization and visual guidance.

3.3 Hierarchical Visual Priors

The evidence image is interpreted through two complementary pretrained representation spaces. Multi-depth DINOv3 features provide dense semantic guidance, while the diffusion-native VAE provides a reference in the generator's native latent space (Shi et al., 2026b; Gong et al., 2026). Both priors are derived from $\widetilde{\mathbf{x}}$ at the decoder and therefore introduce no additional conditioning bits.

Multi-depth dense semantic prior. We extract spatial patch tokens from L selected depths of a frozen DINOv3 transformer, excluding class and register tokens. After per-depth normalization and learned projection, they form

$$\mathbf{P}^{\text{sem}} = \mathcal{H}(\widetilde{\mathbf{x}}) = \left(\mathbf{P}_1^{\text{sem}}, \ldots, \mathbf{P}_L^{\text{sem}}\right), \tag{4}$$

where each $\mathbf{P}_\ell^{\text{sem}}$ is a dense spatial token sequence projected to the DiT hidden dimension. Rather than collapsing the selected DINOv3 depths into a single condition, we preserve their ordering and assign successive semantic levels to successive groups of DiT blocks through cross-attention. This depth-matched design exposes the generator to complementary levels of visual abstraction throughout the network.

Diffusion-native structural prior. The semantic branch does not directly provide a reference in the native latent coordinates used by the generator. We therefore encode the same evidence image with the frozen native VAE encoder. Let $\mathcal{S}$ denote the encoder with its fixed latent calibration. We define

$$\mathbf{P}^{\text{str}} = \mathcal{S}(\widetilde{\mathbf{x}}), \qquad \mathbf{z}_0 = \mathcal{S}(\mathbf{x}), \tag{5}$$

where $\mathbf{P}^{\text{str}}$ is the structural prior and $\mathbf{z}_0$ is the clean latent used as the training target. Because $\mathbf{P}^{\text{str}}$ lies in the same latent coordinates as the generative state, it provides spatially aligned guidance complementary to the semantic tokens. Its role is distinct from that of $\mathbf{z}_c$: the codec state defines the source from which transport begins, whereas $\mathbf{P}^{\text{str}}$ acts as a structural reference throughout the generative mapping.

Complementary conditioning. The complete visual condition is $\mathcal{P} = (\mathbf{P}^{\text{str}}, \mathbf{P}^{\text{sem}})$. The two priors enter the DiT through distinct interfaces: $\mathbf{P}^{\text{str}}$ is concatenated with the evolving latent before patch embedding, whereas the assigned semantic tokens are injected through cross-attention. We denote the resulting predictor by $\mathbf{u}_\theta(\mathbf{z}, t, r; \mathcal{P})$, where t and r denote the current and destination flow times, respectively. Both conditioning paths remain active throughout adaptation and deployment.

3.4 Learning a One-Step Generative Decoder

The pretrained generator and the deployed codec differ along two coupled axes: the former starts from a stochastic source and follows iterative transport, whereas the latter must start from a bitstream-determined state and decode in one deterministic step. Rather than imposing both changes at once, we adapt the generative process progressively. Stage I re-centers the source on the codec state while retaining stochasticity for flow learning; Stage II distills the resulting codec-conditioned transport into a one-step map; and Stage III contracts the remaining source randomness while jointly adapting the complete compression system to its deployed deterministic input.

3.4.1 Stage I: Conditional Flow Adaptation

We first ground the generative source in the transmitted representation while retaining stochasticity to facilitate flow learning. Specifically, we define a codec-centered source

$$\mathbf{z}_1 = \mathbf{z}_c + \sigma_s \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

where $\sigma_s > 0$ controls the source-noise scale. Generation therefore starts from a stochastic neighborhood around $\mathbf{z}_c$, rather than from zero-centered noise. This follows the general intuition that an informative reconstruction can serve as the center of a generative source (Ohayon et al., 2024).

We pair this source with the clean latent $\mathbf{z}_0$ through the linear interpolation

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \qquad t \sim \mathcal{U}(0,1), \tag{7}$$

and train the DiT velocity predictor $\mathbf{v}_\theta(\mathbf{z}_t, t; \mathcal{P})$ with

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\|\mathbf{v}_\theta(\mathbf{z}_t, t; \mathcal{P}) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2\Big], \qquad \mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{rate}}^{\mathrm{I}}\mathcal{R} + \lambda_{\mathrm{ev}}^{\mathrm{I}}\mathcal{L}_{\mathrm{ev}}. \tag{8}$$

We jointly optimize the codec, latent projection, evidence decoder, semantic projections, and DiT, while keeping the pretrained DINOv3 and VAE encoders frozen. The resulting model defines an iterative conditional transport from the codec-centered source to the clean latent and serves as the teacher for Stage II. Sampling uses the learned velocity directly without CFG.

3.4.2 Stage II: One-Step Flow-Map Distillation

Stage I grounds the generative trajectory in the codec state, but the learned transport remains iterative. We therefore distill it into a direct map from the same source distribution to the clean endpoint. Importantly, the student distills the codec-conditioned transport learned in Stage I, rather than the original pretrained generative trajectory. We adopt AYF's Eulerian map distillation (AYF-EMD) (Sabour et al., 2025). For distillation, the student predictor $\mathbf{u}_\theta(\mathbf{z}, t, r; \mathcal{P})$ estimates the average velocity over the interval $[r, t]$ for $r < t$. At $r = t$, the same parameterization represents the instantaneous velocity used in Stage I:

$$\mathbf{v}_\theta(\mathbf{z}, t; \mathcal{P}) \equiv \mathbf{u}_\theta(\mathbf{z}, t, t; \mathcal{P}).$$

The Stage-I model is frozen as the teacher $\mathbf{v}_T$, while the student is initialized from its weights. The codec and all condition-producing modules are also frozen during this stage.

Distillation reuses the codec-centered source and paired interpolation in Eqs. (6)–(7), with the same visual priors $\mathcal{P}$ shared by teacher and student. Following the average-velocity parameterization (Geng et al., 2025), we fix the destination time to $r = 0$ and train the student to represent transport directly to the clean endpoint. At $(t, r) = (1, 0)$, the resulting one-step decoder is

$$\mathcal{G}_\theta(\mathbf{z}; \mathcal{P}) = \mathbf{z} - \mathbf{u}_\theta(\mathbf{z}, 1, 0; \mathcal{P}). \tag{9}$$

The student is trained with the AYF-EMD regression objective

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{map}} = \mathbb{E}\Big[w(t,0)\,\|\mathbf{u}_\theta(\mathbf{z}_t, t, 0; \mathcal{P}) - \mathrm{sg}(\mathbf{u}_{\mathrm{tgt}})\|_2^2\Big], \tag{10}$$

where $w(t, 0) \geq 0$ is the regression weight, $\mathbf{u}_{\mathrm{tgt}}$ is the AYF-EMD target constructed from the frozen teacher and a detached copy of the current student, and sg denotes stop-gradient. Only the student is updated in Stage II, so the codec state and visual conditions remain fixed throughout distillation.

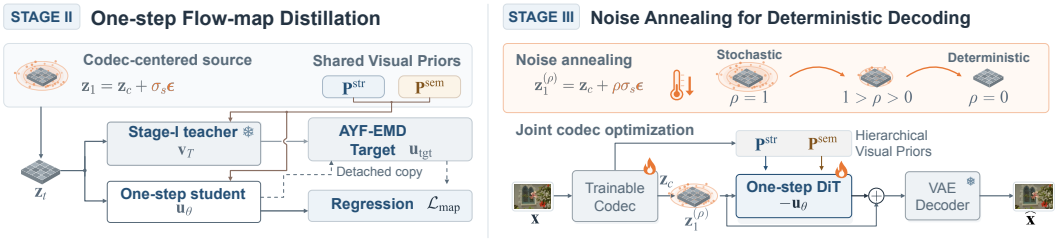

Figure 3: **Stages II and III: one-step distillation and deterministic source adaptation.** *Left:* The frozen Stage-I teacher and student share paired interpolants $\mathbf{z}_t$ and visual priors $\mathcal{P}$; separate conditioning arrows show the priors entering both models. The teacher and a detached copy of the current student form the AYF-EMD target $\mathbf{u}_{\text{tgt}}$ for regression $\mathcal{L}_{\text{map}}$, with destination time $r = 0$ fixed throughout distillation. *Right:* End-to-end optimization anneals $\rho : 1 \to 0$, contracting the stochastic source onto the deterministic codec state $\mathbf{z}_c$. Flames mark the jointly trained codec and DiT; the VAE decoder remains frozen. The residual sum combines $\mathbf{z}_1^{(\rho)}$ with the negative DiT prediction $-\mathbf{u}_\theta$.

3.4.3 Stage III: Noise Annealing for Deterministic Decoding

The distilled student already provides one-step decoding, but its source remains stochastic because Stage II is trained from $\mathbf{z}_c + \sigma_s\boldsymbol{\epsilon}$. Deployment, in contrast, should depend only on the received bitstream. We remove this train–deployment mismatch during end-to-end optimization by gradually annealing the source perturbation:

$$\mathbf{z}_1^{(\rho)} = \mathbf{z}_c + \rho\sigma_s\boldsymbol{\epsilon}, \qquad \widehat{\mathbf{z}}_0 = \mathcal{G}_\theta(\mathbf{z}_1^{(\rho)}; \mathcal{P}), \tag{11}$$

where ρ decreases from 1 to 0 over training. At $\rho = 1$, the starting distribution matches that used for flow adaptation and distillation; as $\rho \to 0$, it contracts to the codec state $\mathbf{z}_c$. This gradual transition avoids replacing the stochastic distilled source with the deterministic deployment input abruptly, while allowing the codec and one-step decoder to adapt jointly along the contraction. The final decoder therefore recovers Eq. (1) and requires no stochastic input at deployment.

During this stage, we jointly optimize the codec, evidence path, semantic projections, and one-step DiT; DINOv3 and both the VAE encoder and decoder remain frozen. The end-to-end objective is

$$\mathcal{L}_{\text{III}} = \lambda_{\text{rate}}^{\text{III}}\mathcal{R} + \mathcal{L}_{\text{img}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}^{\text{DINO}} + \lambda_{\text{ev}}^{\text{III}}\mathcal{L}_{\text{ev}}. \tag{12}$$

Here, $\mathcal{L}_{\text{img}}$ combines MSE, MS-SSIM, LPIPS, and DISTS (Wang et al., 2003; Zhang et al., 2018; Ding et al., 2022), while $\mathcal{L}_{\text{adv}}^{\text{DINO}}$ uses lightweight discriminator heads on frozen multi-depth DINOv3 features (Kumari et al., 2022). The evidence objective continues to regularize $\widetilde{\mathbf{x}}$ as the codec representation evolves during joint optimization. Construction and training details, loss definitions, and supporting analyses are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

4.1.1 Model and Training

Following Yao et al. (2025), we pretrain a class-conditional LightningDiT on ImageNet-1K (Deng et al., 2009) for 100 epochs, and then add visual cross-attention layers for Stage I adaptation. Following DiT-IC (Shi et al., 2026a), we train on high-quality images from CLIC 2020 Professional (Toderici et al., 2020), MLIC-Train-100K (Jiang & Wang, 2023), and LSDIR (Li et al., 2023), retaining images whose height and width both exceed 512 pixels. We use aspect-ratio buckets centered on a nominal 768×768 resolution, spanning ratios from 1:2 to 2:1. Before Stage I, we pretrain the codec and evidence path for 20k updates using the evidence MSE and DINO losses to obtain a stable initialization, with the same weights as Stage I: $\lambda_{\text{ev}}^{\text{I}} = 1$ and $\lambda_{\text{ev-sem}} = 0.5$. We run Stages I and II once at a high bitrate to obtain a one-step initialization. Stage III first optimizes the highest-rate codec, followed by end-to-end fine-tuning toward lower rates with 50k updates per rate. All stages use AdamW on 16 A100 GPUs with a global batch size of 16. For the initial high-rate training, Stages I and III each run for 50k updates with learning rates of 10^{-4} and 5×10^{-5}, respectively, while Stage II runs

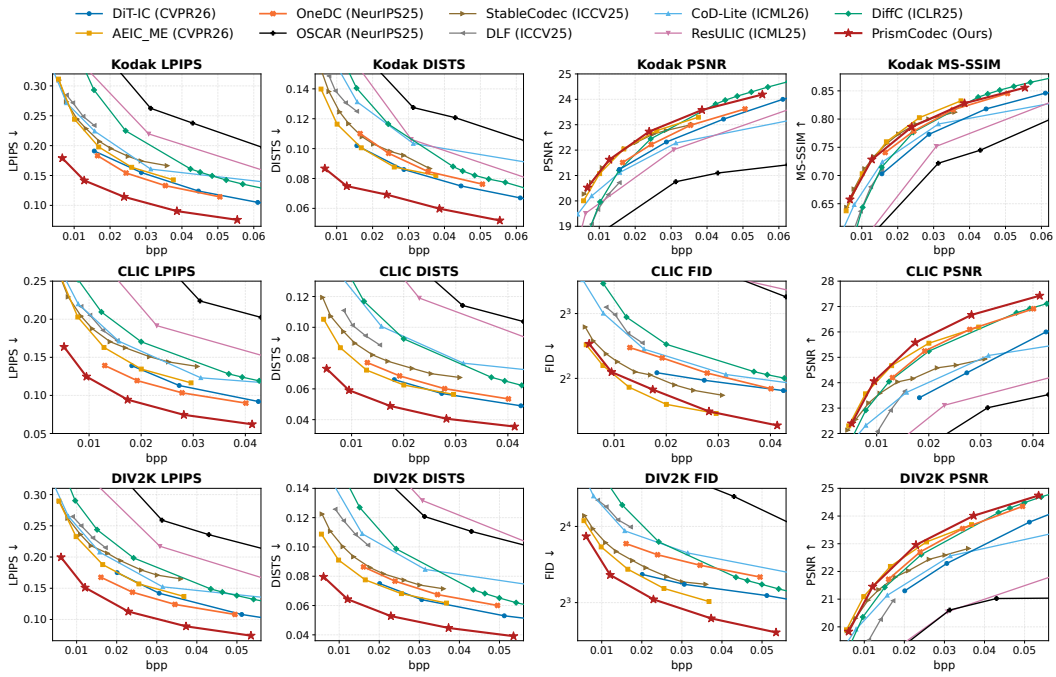

Figure 4: Rate–distortion and rate–perception comparisons on Kodak, CLIC2020, and DIV2K.

for 20k updates at 5×10^{-5}. We train five models with $\lambda_{\text{rate}}^{\text{III}} \in \{1, 2, 4, 8, 16\}$ and set $\sigma_s = 0.3$. For Stage III, we use $\lambda_{\text{mse}} = 1.0$, $\lambda_{\text{lpips}} = 1.0$, $\lambda_{\text{dists}} = 0.5$, and $\lambda_{\text{ms}} = 0.1$ for the final-image objective, with $\lambda_{\text{adv}} = 0.5$ and $\lambda_{\text{ev}}^{\text{III}} = 0.2$ under $\lambda_{\text{ev-sem}} = 0.5$. For the initial Stage III run, ρ follows a half-cosine decay from 1 to 0 over the first 10k updates and remains at 0 for the remaining 40k updates. We enable the adversarial loss after 30k Stage III updates. For inference, large images are decoded with overlapping latent tiles corresponding to 1024×1024 pixels, with a 256-pixel overlap. The evidence image is resized to 512×512 for DINOv3.

4.1.2 Datasets, Metrics, and Baselines

We evaluate on Kodak (Eastman Kodak, 1993) (24 images), the CLIC2020 test set (Toderici et al., 2020) (428 images), and the DIV2K validation set (Agustsson & Timofte, 2017) (100 images). We report PSNR and MS-SSIM (Wang et al., 2003) for distortion, LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022) for perceptual fidelity, and FID (Heusel et al., 2017) for distributional realism. The four paired metrics are computed on full-resolution RGB reconstructions and averaged over images; FID uses 256×256 patches on CLIC2020 and DIV2K, following Mentzer et al. (2020). We omit FID on Kodak given its small sample size. We compare with DiffC (Vonderfecht & Liu, 2025), ResULIC (Ke et al., 2025), DLF (Xue et al., 2025a), OSCAR (Guo et al., 2025), StableCodec (Zhang et al., 2025), OneDC (Xue et al., 2025b), AEIC-ME (Zhang et al., 2026a), DiT-IC (Shi et al., 2026a), and CoD-Lite (Jia et al., 2026a). We summarize bitrate efficiency with BD-rate (Bjøntegaard, 2001), using StableCodec as the reference. Negative BD-rate indicates bitrate savings at equal quality.

4.2 Main Results

4.2.1 Quantitative Evaluation

The main comparison tests whether grounding both the generative source and visual guidance in the received bitstream translates into a better rate–perception tradeoff without systematically sacrificing source fidelity. Fig. 4 reports the rate–distortion and rate–perception curves in the ultra-low-bitrate regime, while Table 1 summarizes BD-rates relative to StableCodec.

Perceptual fidelity and realism. Across all three datasets, PrismCodec shifts the LPIPS and DISTS curves favorably relative to prior generative codecs. On CLIC2020, it reduces bitrate relative to StableCodec by 71.46% at matched LPIPS and 77.08% at matched DISTS; the corresponding savings on DIV2K are 69.54% and 74.64%. The same operating points also yield FID-based savings of

Table 1: Detailed comparisons of BD-rate (%) relative to StableCodec. –: no overlapping range.

Method	Kodak				CLIC2020				DIV2K			
	LPIPS	DISTS	PSNR	MS-SSIM	LPIPS	DISTS	FID	PSNR	LPIPS	DISTS	FID	PSNR
DiffC (Vonderfecht & Liu, 2025)	84.91	74.71	34.58	33.51	48.49	101.71	181.51	−7.28	40.40	101.89	117.75	4.03
ResULIC (Ke et al., 2025)	105.82	76.51	72.91	91.57	118.90	274.69	2117.98	153.66	113.46	376.60	–	208.08
StableCodec (Zhang et al., 2025)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DLF (Xue et al., 2025a)	21.21	29.35	108.54	57.23	18.32	42.11	151.63	75.98	21.38	70.13	125.20	81.32
OneDC (Xue et al., 2025b)	−36.63	−5.46	23.88	10.74	−56.28	−28.94	87.88	−34.31	−52.46	−9.98	65.34	−17.47
OSCAR (Guo et al., 2025)	237.66	204.80	308.30	161.49	369.67	444.09	713.43	423.85	256.87	351.99	470.60	269.90
DiT-IC (Shi et al., 2026a)	−27.79	−28.08	40.25	43.46	−39.39	–	46.73	51.02	−31.57	−32.62	−13.07	30.47
AEIC-ME (Zhang et al., 2026a)	−9.10	−26.03	1.37	−2.42	−12.92	−35.81	−36.78	−29.67	−17.22	−34.14	−20.26	−20.65
CoD-Lite (Jia et al., 2026a)	0.97	62.40	55.68	28.09	10.35	98.48	107.12	36.73	−3.38	77.42	102.18	23.37
PrismCodec	**−75.49**	**−81.33**	**−5.82**	**0.40**	**−71.46**	**−77.08**	**−30.63**	**−30.45**	**−69.54**	**−74.64**	**−46.96**	**−17.21**

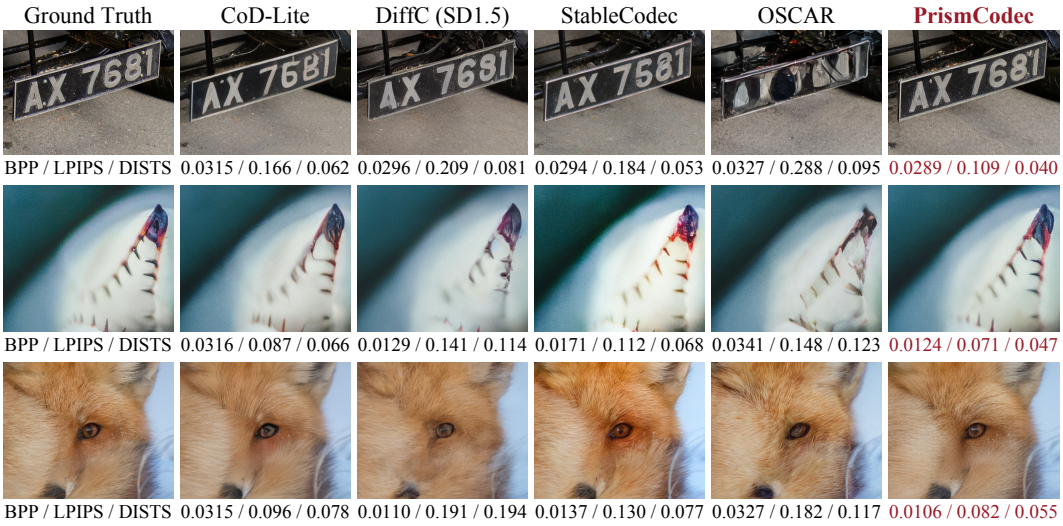

Figure 5: **Qualitative comparisons on DIV2K:** 0847, 0857, and 0862 (top to bottom).

30.63% and 46.96% on CLIC2020 and DIV2K, respectively, indicating that the gains in paired perceptual fidelity are accompanied by improved distributional realism. Appendix D provides complementary KID comparisons.

Source fidelity. The perceptual gains do not come with a systematic loss in distortion performance. PrismCodec achieves PSNR-based bitrate savings of 5.82%, 30.45%, and 17.21% on Kodak, CLIC2020, and DIV2K, respectively. MS-SSIM remains close to StableCodec, with a 0.40% bitrate increase on Kodak and reductions of 1.48% and 0.90% on CLIC2020 and DIV2K. Taken together, these results support the central premise of PrismCodec: bitstream-grounded generation can improve perceptual reconstruction while retaining source-dependent fidelity.

4.2.2 Qualitative Evaluation

Fig. 5 examines whether the quantitative gains correspond to source-specific visual details rather than only aggregate metric improvements. At lower bitrates than the compared baselines, PrismCodec better preserves plate lettering (DIV2K 0847), leaf-tip contours and serrations (DIV2K 0857), and the fox's eye, nose, and fur patterns (DIV2K 0862), whereas several baselines smooth or alter these details. These examples are qualitatively consistent with the intended role of bitstream-derived visual evidence in constraining generative reconstruction. Appendix F provides additional comparisons.

4.3 Efficiency Evaluation

We report deployed parameters, encoding and decoding latency, and computation (kMACs/pixel and FLOPs) in Table 2. Measurements use synthetic RGB inputs at 2048×2048, batch size one, and an NVIDIA A100 GPU; after warm-up, latency is averaged over 30 CUDA-synchronized runs. PrismCodec decodes a 2K image in 2.10 s, compared with 9.74 s for StableCodec and 3.22 s for OSCAR, while remaining in the same order of total computation as these two generative baselines.

Table 2: Encoding and decoding efficiency at 2048×2048.

Method	Params (B)	Latency (s)		kMACs/pixel			FLOPs (T)		
		Enc.	Dec.	Enc.	Dec.	Total	Enc.	Dec.	Total
StableCodec	1.07	4.65	9.74	2,965.0	10,343.7	13,308.7	24.87	86.77	111.64
AEIC-ME	0.93	0.08	5.16	204.0	5,959.2	6,163.2	1.71	49.99	51.70
DiT-IC	1.05	0.50	0.83	2,759.9	4,528.5	7,288.4	23.15	37.99	61.14
DLF	1.23	0.41	1.00	2,727.5	5,250.2	7,977.7	22.88	44.04	66.92
OneDC	1.40	0.65	1.51	2,888.1	10,292.4	13,180.5	24.23	86.34	110.57
OSCAR	1.05	0.89	3.22	3,150.1	10,890.4	14,040.5	26.42	91.36	117.78
PrismCodec	1.89	0.62	2.10	1,352.7	13,571.8	14,924.5	11.35	113.85	125.20

4.4 Ablation Studies

We organize the ablations around the three claims underlying PrismCodec: (i) semantic and structural visual priors provide complementary guidance, with multi-depth semantics contributing beyond a single representation level; (ii) transmitted evidence should determine the generative source as well as its conditions; and (iii) explicit one-step distillation and subsequent source contraction are both important for adapting stochastic generation to deterministic decoding. All results report LPIPS- and DISTS-based BD-rate (%) on DIV2K relative to StableCodec. Each variant trains its own high-rate model and uses matched fine-tuning schedules to obtain the five rate points in Sec. 4.1.

Table 3: Ablations of visual priors.

Variant	LPIPS (%)	DISTS (%)
Single semantic depth	-59.64	-68.37
Structural only	-51.67	-58.32
Semantic only	-42.89	-47.24
Full PrismCodec	$\mathbf{-69.54}$	$\mathbf{-74.64}$

Table 4: Ablations of source and training.

Variant	LPIPS (%)	DISTS (%)
Zero-centered noise	-37.81	-46.80
w/o annealing	-55.02	-54.41
w/o distillation	-54.74	-58.04
Full PrismCodec	$\mathbf{-69.54}$	$\mathbf{-74.64}$

Hierarchical visual priors. Table 3 separates two aspects of the proposed conditioning design. First, *Single semantic depth* repeats the final selected DINOv3 layer across all four DiT groups, increasing LPIPS- and DISTS-based BD-rate by 9.90 and 6.27 percentage points relative to the full model. This isolates the benefit of preserving multiple representation depths rather than collapsing semantic guidance to one level. Second, *Structural only* removes semantic guidance and increases the two BD-rates by 17.87 and 16.32 points, while *Semantic only* removes structural input conditioning and increases them by 26.65 and 27.40 points. The degradation in both single-branch variants supports the complementary roles of dense semantic guidance and diffusion-native structural guidance.

Bitstream-grounded source. With both priors retained, *Zero-centered noise* replaces the codec-centered source with a Gaussian-noise starting state and performs worse than both the stochastic codec-centered variant (*w/o annealing*) and the full model (Table 4). Thus, the gain from $\mathbf{z}_c$ is not merely duplicated by the conditioning paths: exposing the bitstream through the generative source provides an additional and complementary form of grounding.

Progressive adaptation to deterministic decoding. Table 4 evaluates the two transitions from the Stage-I iterative flow to deterministic one-step decoding. *w/o annealing* keeps $\rho = 1$ through Stage III and inference, using $\mathbf{z}_c + 0.3\boldsymbol{\epsilon}$ as the source, and increases LPIPS- and DISTS-based BD-rates by 14.52 and 20.23 points. This shows that the stochastic distilled source should be contracted to the deterministic codec state before deployment. *w/o distillation* skips Stage II and directly fine-tunes the Stage-I model with the full model's annealing schedule, increasing the two BD-rates by 14.80 and 16.60 points. Together, these results support both flow-map distillation and source-noise annealing in the transition to deterministic one-step decoding.

5 Conclusion

We presented PrismCodec, which recasts a pretrained DiT as a bitstream-grounded one-step decoder for extreme image compression. The compressed representation determines both the generative starting state and hierarchical semantic and structural guidance. Through codec-centered flow adaptation, flow-map distillation, and source-noise annealing, PrismCodecprogressively transforms stochastic iterative generation into deterministic one-step decoding. The resulting model reconstructs directly from the received bitstream without additional conditioning bits, yielding an efficient and effective rate–perception tradeoff.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and feedback on the presentation of the methodology and experiments. All AI-assisted suggestions were reviewed, verified, and revised as appropriate by the authors, who take responsibility for the final manuscript.

ETHICS STATEMENT

The authors have reviewed and complied with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Details necessary to reproduce our experiments, including training and inference settings and evaluation protocols, are documented in the main paper and Appendix.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

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.

Eirikur Agustsson et al. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJxdQ3jeg.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkcQFMZRb.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

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

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. Accessed 29 June 2026.

Hao Cao, Chengbin Liang, Wenqi Guo, Zhijin Qin, and Jungong Han. Progic: Progressive and lightweight generative image compression with residual vector quantization. *arXiv preprint arXiv:2603.02897*, 2026.

Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2024.

Jianhui Chang. Generative image coding with diffusion prior. *arXiv preprint arXiv:2509.13768*, 2025.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, 2025a.

Yunuo Chen, Qian Li, Bing He, Donghui Feng, Ronghua Wu, Qi Wang, Li Song, Guo Lu, and Wenjun Zhang. S2cformer: Reorienting learned image compression from spatial interaction to channel aggregation, 2025b. URL https://arxiv.org/abs/2502.00700.

Yunuo Chen, Zezheng Lyu, Bing He, Ning Cao, Gang Chen, Guo Lu, and Wenjun Zhang. Knowledge distillation for learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4996–5006, 2025c.

Yunuo Chen, Bing He, Zezheng Lyu, Hongwei Hu, Qunshan Gu, Yuan Tian, and Guo Lu. Adaptive learned image compression with graph neural networks. *arXiv preprint arXiv:2603.25316*, 2026a.

Yunuo Chen, Zezheng Lyu, Bing He, Hongwei Hu, Qi Wang, Yuan Tian, Li Song, Wenjun Zhang, and Guo Lu. Content-aware mamba for learned image compression. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=WwDNiisZQm.

Yunuo Chen, Chuqin Zhou, Jiangchuan Li, Xiaoyue Ling, Bing He, Jincheng Dai, Li Song, and Guo Lu. Next-frame decoding for ultra-low-bitrate image compression with video diffusion priors. *arXiv preprint arXiv:2603.15129*, 2026c. URL https://arxiv.org/abs/2603.15129.

Zheng Chen, Mingde Zhou, Jinpei Guo, Jiale Yuan, Yifei Ji, and Yulun Zhang. Steering one-step diffusion model with fidelity-rich decoder for fast image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3192–3200, 2026d.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7939–7948, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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*, 2022.

Zhenyu Du, Yanbo Gao, Shuai Li, Yiyang Li, Hui Yuan, and Mao Ye. A noise constrained diffusion (nc-diffusion) framework for high-fidelity image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992). http://r0k.us/graphics/kodak/, 1993. Accessed 29 June 2026.

Alaaeldin El-Nouby, Matthew J Muckley, Karen Ullrich, Ivan Laptev, Jakob Verbeek, and Hervé Jégou. Image compression with product quantized masked image modeling. *arXiv preprint arXiv:2212.07372*, 2022.

Noam Elata, Tomer Michaeli, and Michael Elad. Psc: Posterior sampling-based compression. *arXiv preprint arXiv:2407.09896*, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. URL https://arxiv.org/abs/2505.13447.

Noor Fathima Ghouse, Jens Petersen, Auke Wiggers, Tianlin Xu, and Guillaume Sautiere. A residual diffusion model for high perceptual quality codec augmentation. *arXiv preprint arXiv:2301.05489*, 2023.

Kaixiong Gong, Xin Cai, Bin Lin, Hao Wang, Yunlong Lin, Mingzhe Zheng, Bohao Li, Jian-Wei Zhang, Miles Yang, Zhao Zhong, et al. Twins: Learn to predict unified representations with focal loss. *arXiv preprint arXiv:2607.22531*, 2026.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jinpei Guo, Yifei Ji, Zheng Chen, Kai Liu, Min Liu, Wang Rao, Wenbo Li, Yong Guo, and Yulun Zhang. Oscar: One-step diffusion codec across multiple bit-rates. *arXiv preprint arXiv:2505.16091*, 2025.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14771–14780, 2021.

Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5718–5727, 2022.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yinhuan Huang, Hao Cao, Pu Chen, Wenqi Guo, and Zhijin Qin. Flowcodec: One-step flow prior for generative image compression. *arXiv preprint arXiv:2606.21030*, 2026.

Shoma Iwai, Tomo Miyazaki, Yoshihiro Sugaya, and Shinichiro Omachi. Fidelity-controllable extreme image compression with generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 8235–8242. IEEE, 2021.

Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26088–26098, 2024.

Zhaoyang Jia, Naifu Xue, Zihan Zheng, Jiahao Li, Bin Li, Xiaoyi Zhang, Zongyu Guo, Yuan Zhang, Houqiang Li, and Yan Lu. Cod-lite: Real-time diffusion-based generative image compression. *arXiv preprint arXiv:2604.12525*, 2026a.

Zhaoyang Jia, Zihan Zheng, Naifu Xue, Jiahao Li, Bin Li, Zongyu Guo, Xiaoyi Zhang, Houqiang Li, and Yan Lu. Cod: A diffusion foundation model for image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38420–38429, 2026b.

Shiyin Jiang, Wei Long, Minghao Han, Zhenghao Chen, Ce Zhu, and Shuhang Gu. Differentiable vector quantization for rate-distortion optimization of generative image compression. *arXiv preprint arXiv:2604.10546*, 2026.

Wei Jiang and Ronggang Wang. MLIC++: Linear complexity multi-reference entropy modeling for learned image compression. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023. URL `https://openreview.net/forum?id=hxIpcSoz2t`.

Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, and Ronggang Wang. Mlic: Multi-reference entropy model for learned image compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7618–7627, 2023.

Luoxu Jin and Hiroshi Watanabe. Perceptual image compression via stable diffusion at low bitrate. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 626–629. IEEE, 2024.

Anle Ke, Xu Zhang, Tong Chen, Ming Lu, Chao Zhou, Jiawen Gu, and Zhan Ma. Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. *arXiv preprint arXiv:2505.08281*, 2025.

Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. Egic: enhanced low-bit-rate generative image compression guided by semantic segmentation. In *European Conference on Computer Vision*, pp. 202–220. Springer, 2024a.

Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. Perco (sd): Open perceptual compression. *arXiv preprint arXiv:2409.20255*, 2024b.

Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. Percov2: improved ultra-low bit-rate perceptual image compression with implicit hierarchical masked image modeling. *arXiv preprint arXiv:2503.09368*, 2025.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10651–10662, 2022.

Hagyeong Lee, Minkyu Kim, Jun-Hyuk Kim, Seungeon Kim, Dokwan Oh, and Jaeho Lee. Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity. *arXiv preprint arXiv:2403.02944*, 2024.

Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text+ sketch: Image compression at ultra low rates. *arXiv preprint arXiv:2307.01944*, 2023.

Anqi Li, Feng Li, Yuxi Liu, Runmin Cong, Yao Zhao, and Huihui Bai. Once-for-all: Controllable generative image compression with dynamic granularity adaptation. *arXiv preprint arXiv:2406.00758*, 2024a.

Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware transformer for learned image compression. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1775–1787. IEEE, 2023.

Yuqi Li, Haotian Zhang, Li Li, and Dong Liu. Learned image compression with hierarchical progressive context modeling. *arXiv preprint arXiv:2507.19125*, 2025a.

Yuqi Li, Haotian Zhang, Li Li, Dong Liu, and Feng Wu. Scaling learned image compression models up to 1 billion. *arXiv preprint arXiv:2508.09075*, 2025b.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Towards extreme image compression with latent feature guidance and diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024c.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Ajmal Saeed Mian. Rdeic: Accelerating diffusion-based extreme image compression with relay residual diffusion. *arXiv preprint arXiv:2410.02640*, 2024d.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Jinming Liu, Heming Sun, and Jiro Katto. 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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Jingbo Lu, Leheng Zhang, Xingyu Zhou, Mu Li, Wen Li, and Shuhang Gu. Learned image compression with dictionary-based entropy model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12850–12859, 2025.

Lei Lu, Yanyue Xie, Wei Jiang, Wei Wang, Xue Lin, and Yanzhi Wang. Hybridflow: Infusing continuity into masked codebook for extreme low-bitrate image compression. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3010–3018, 2024.

Yiyang Ma, Wenhan Yang, and Jiaying Liu. Correcting diffusion-based perceptual image compression with privileged end-to-end decoder. *International conference on machine learning (ICML)*, 2024.

Qi Mao, Tinghan Yang, Yinuo Zhang, Zijian Wang, Meng Wang, Shiqi Wang, Libiao Jin, and Siwei Ma. Extreme image compression using fine-tuned vqgans. In *2024 Data Compression Conference (DCC)*, pp. 203–212. IEEE, 2024.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020.

David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3339–3343. IEEE, 2020.

David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 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*, volume 38, pp. 4296–4304, 2024.

Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pp. 25426–25443. PMLR, 2023.

Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum MSE photo-realistic image restoration. *arXiv preprint arXiv:2410.00418*, 2024. URL https://arxiv.org/abs/2410.00418.

Guy Ohayon, Hila Manor, Tomer Michaeli, and Michael Elad. Compressed image generation with denoising diffusion codebook models. *arXiv preprint arXiv:2502.01189*, 2025.

Zhihong Pan, Xin Zhou, and Hao Tian. Extreme generative image compression by learning text embedding from diffusion models. *arXiv preprint arXiv:2211.07793*, 2022.

Chanung Park, Joo Chan Lee, and Jong Hwan Ko. Single-step diffusion for image compression at ultra-low bitrates. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6391–6400. IEEE, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Yichen Qian, Ming Lin, Xiuyu Sun, Zhiyu Tan, and Rong Jin. Entroformer: A transformer-based entropy model for learned image compression. In *International Conference on Learning Representations*, May 2022.

Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. Lossy image compression with foundation diffusion models. In *European Conference on Computer Vision*, pp. 303–319. Springer, 2024.

Lucas Relic, Roberto Azevedo, Yang Zhang, Markus Gross, and Christopher Schroers. Bridging the gap between gaussian diffusion models and universal quantization for image compression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2449–2458. IEEE, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your flow: Scaling continuous-time flow map distillation. *arXiv preprint arXiv:2506.14603*, 2025. URL https://arxiv.org/abs/2506.14603.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Xihua Sheng, Lingyu Zhu, Tianyu Zhang, Dong Liu, Shiqi Wang, and Jing Wang. Cadc: Content adaptive diffusion-based generative image compression. *arXiv preprint arXiv:2602.21591*, 2026.

Junqi Shi, Ming Lu, Xingchen Li, Anle Ke, Ruiqi Zhang, and Zhan Ma. Dit-ic: Aligned diffusion transformer for efficient image compression. *arXiv preprint arXiv:2603.13162*, 2026a.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, volume 2026, pp. 154506–154537, 2026b.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Juan Song, Lijie Yang, and Mingtao Feng. Extremely low-bitrate image compression semantically disentangled by lmms from a human perception perspective. *arXiv preprint arXiv:2503.00399*, 2025.

Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Ballé, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. Workshop and challenge on learned image compression (clic2020). In *Proc. CVPR Workshops*, 2020.

Amit Vaisman, Guy Ohayon, Hila Manor, Michael Elad, and Tomer Michaeli. Turbo-DDCM: Fast and flexible zero-shot diffusion-based image compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eIF1QvC94Z.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Jeremy Vonderfecht and Feng Liu. Lossy compression with pretrained diffusion models. *International Conference on Learning Representations*, 2025.

Guo-Hua Wang, Jiahao Li, Bin Li, and Yan Lu. Evc: Towards real-time neural image compression with mask decay. *arXiv preprint arXiv:2302.05071*, 2023.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216. URL https://ece.uwaterloo.ca/~z70wang/publications/msssim.pdf.

Siqi Wu, Yinda Chen, Dong Liu, and Zhihai He. Conditional latent coding with learnable synthesized reference for deep image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12863–12871, 2025a.

Zhuojie Wu, Heming Du, Shuyun Wang, Ming Lu, Haiyang Sun, Yandong Guo, and Xin Yu. Cmamba: Learned image compression with state space models. *arXiv preprint arXiv:2502.04988*, 2025b.

Yichong Xia, Yimin Zhou, Jinpeng Wang, Baoyi An, Haoqian Wang, Yaowei Wang, and Bin Chen. Diffpc: Diffusion-based high perceptual fidelity image compression with semantic refinement. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tongda Xu, Ziran Zhu, Dailan He, Yanghao Li, Lina Guo, Yuanyuan Wang, Zhe Wang, Hongwei Qin, Yan Wang, Jingjing Liu, and Ya-Qin Zhang. Idempotence and perceptual image compression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Cy5v64DqEF.

Tongda Xu, Jiahao Li, Bin Li, Yan Wang, Ya-Qin Zhang, and Yan Lu. Picd: Versatile perceptual image compression with diffusion rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28436–28445, 2025.

Naifu Xue, Qi Mao, Zijian Wang, Yuan Zhang, and Siwei Ma. Unifying generation and compression: Ultra-low bitrate image coding via multi-stage transformer. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. Dlf: Extreme image compression with dual-generative latent fusion. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2025a.

Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. One-step diffusion-based image compression with semantic distillation. *Advances in Neural Information Processing Systems*, 2025b.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36:64971–64995, 2023.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems. In *European Conference on Computer Vision*, pp. 343–362. Springer, 2024.

Fanhu Zeng, Hao Tang, Yihua Shao, Siyu Chen, Ling Shao, and Yan Wang. Mambaic: State space models for high-performance learned image compression. *arXiv preprint arXiv:2503.12461*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

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.

Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. Stablecodec: Taming one-step diffusion for extreme image compression. *arXiv preprint arXiv:2506.21977*, 2025.

Tianyu Zhang, Dong Liu, and Chang Wen Chen. Ultra-low bitrate perceptual image compression with shallow encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12118–12128, 2026a.

Ziyuan Zhang, Yichong Xia, Bin Chen, Tianwei Zhang, Hao Wang, and Han Qiu. Autoregressive-based progressive coding for ultra-low bitrate image compression. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Chuqin Zhou, Xiaoyue Ling, Yunuo Chen, Jincheng Dai, Guo Lu, and Wenjun Zhang. Dual-representation image compression at ultra-low bitrates via explicit semantics and implicit textures. *arXiv preprint arXiv:2602.05213*, 2026.

MingSheng Zhou and MingMing Kong. Glic: General format learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10815–10824, 2025.

A Extended Related Work

This appendix expands the positioning in Sec. 2. We first place PrismCodec within learned and generative image compression, and then compare the closest pretrained-prior codecs along the same two questions that motivate our method: *how bitstream-derived evidence enters the generator* and *how stochastic generation is adapted to one-step reconstruction*.

A.1 Conventional Learned Compression

Learned codecs have advanced through stronger entropy models (Minnen & Singh, 2020; He et al., 2021; 2022; Qian et al., 2022; Jiang et al., 2023; Jiang & Wang, 2023; Cheng et al., 2020), more expressive transforms (Li et al., 2024b; Liu et al., 2023; Zeng et al., 2025; Wu et al., 2025b; Chen et al., 2025b; 2026b;a), and specialized, scalable, or distilled designs for efficiency (Zhou & Kong, 2025; Wang et al., 2023; Wu et al., 2025a; Lu et al., 2025; Li et al., 2025a;b; Chen et al., 2025c). These approaches improve the coded representation itself; PrismCodec instead focuses on how a pretrained generative prior should decode the limited evidence that remains after extreme compression.

A.2 Broader Generative Compression

Generative codecs introduce learned priors to recover visually plausible content at low rates. Existing families use adversarial priors (Goodfellow et al., 2020; Agustsson et al., 2019; Mentzer et al., 2020; Muckley et al., 2023; Agustsson et al., 2023; Körber et al., 2024a; Iwai et al., 2021), discrete representations (Van Den Oord et al., 2017; Esser et al., 2021; El-Nouby et al., 2022; Mao et al., 2024; Xue et al., 2024; Jia et al., 2024), or complementary representations (Lu et al., 2024; Xue et al., 2025a; Zhou et al., 2026); related work also studies bitrate control, progressive decoding, and rate–distortion-optimized quantization (Li et al., 2024a; Zhang et al., 2026b; Cao et al., 2026; Jiang et al., 2026).

Diffusion-based codecs reconstruct from compressed information through conditional denoising (Theis et al., 2022; Yang & Mandt, 2023; Hoogeboom et al., 2023), residual generation (Ghouse et al., 2023; Li et al., 2024d), or pretrained generative priors (Podell et al., 2023; Black Forest Labs, 2024; Relic et al., 2024; Jin & Watanabe, 2024). Other methods learn latent guidance and reconstruction corrections (Li et al., 2024c; Ma et al., 2024; Xia et al., 2025; Xu et al., 2025; Ke et al., 2025; Chang, 2025), or use frozen priors for zero-shot coding (Vonderfecht & Liu, 2025; Xu et al., 2024; Elata et al., 2024; Ohayon et al., 2025; Vaisman et al., 2026); recent work further connects diffusion noise with quantization (Relic et al., 2025; Du et al., 2026). CoD pretrains diffusion specifically for compression (Jia et al., 2026b), and CoD-Lite combines this with a lightweight one-step convolutional decoder (Jia et al., 2026a). Accelerated sampling and distillation (Song et al., 2020; Salimans & Ho, 2022; Sauer et al., 2024; Yin et al., 2024b;a) also motivate one-step codecs (Guo et al., 2025; Park et al., 2026), while FlowCodec uses LoRA-adapted flow priors for near-terminal one-step latent transport with an empty prompt (Huang et al., 2026; Hu et al., 2022).

A second line of work concerns the conditioning interface. Text-guided codecs transmit descriptions or learn language-aligned representations (Lei et al., 2023; Careil et al., 2024; Körber et al., 2024b; 2025; Lee et al., 2024; Pan et al., 2022; Song et al., 2025), and CADC derives text from reconstructions (Sheng et al., 2026). Image conditions are also widely used for controlled generation and restoration (Zhang et al., 2023; Mou et al., 2024; Zavadski et al., 2024; Chen et al., 2025a), while REPA aligns diffusion hidden states with pretrained visual features during training (Yu et al., 2024).

For the closest compression methods, Table 5 summarizes how the transmitted representation determines the generator input and conditions, together with the mechanism used for one-step adaptation. The categories below concern the explicit conditioning pathway; all methods also receive image-specific information from their coded representations.

A.3 Fixed Text Conditions

StableCodec (Zhang et al., 2025) adapts SD-Turbo with a fixed prompt and a codec-produced latent input. Its auxiliary encoder supplies source features during compression, and an auxiliary decoder predicts a structural residual that is added after denoising and before VAE decoding. The fixed prompt therefore coexists with substantial image-specific guidance. PrismCodec separates the same

Table 5: **Comparison of bitstream grounding and one-step adaptation.** The middle column summarizes how image-specific coded evidence enters the generator; the right column summarizes how each method obtains fast or one-step decoding. Fixed text conditions can coexist with image-specific latent or anchor guidance.

Method	Generator input and conditioning	One-step adaptation
StableCodec	Codec latent and structural residual branch; fixed text.	SD-Turbo adaptation with bitrate pruning.
NeFIC	Anchor tokens and a codec-decoded bypass latent.	Next-frame adaptation, then bypass training.
OneDC	Codec latent and tokenizer-supervised hyperprior tokens.	One-step diffusion with semantic and perceptual distillation.
DiT-IC	Compressed latent, text-aligned context, and variance-derived pseudo-timesteps.	Reconstruction flow and encoder-latent self-alignment.
AEIC	Codec latent and structural residual; cross-attention removed.	Pruned SD-Turbo decoder and feature distillation for shallow encoders.
SODEC	Transformed codec latent and ViT features of an auxiliary image.	Fixed-timestep SD2.1 adaptation with rate annealing.
PrismCodec	Codec-centered source, multi-depth semantic tokens, and a VAE structural condition.	Conditional flow adaptation, flow-map distillation, and source contraction.

broad grounding problem into distinct roles: a codec-centered source determines where generation starts, while a frozen VAE latent and multi-depth visual tokens provide structural and semantic guidance through separate DiT interfaces. StableCodec also tightens the rate penalty during training; PrismCodec's source contraction instead changes the generator's input distribution as it approaches deterministic deployment.

NeFIC (Chen et al., 2026c) also retains a fixed prompt, using a decoded anchor for conditioning and a codec-derived bypass latent as the input to its video-diffusion generator.

A.4 Learned Semantic Surrogates: OneDC and DiT-IC

OneDC (Xue et al., 2025b) decodes a semantic token sequence from the quantized hyperprior and uses it as cross-attention context. A pretrained generative tokenizer supplies training labels through an auxiliary prediction task; the tokenizer and predictor are unnecessary at inference. The resulting hyperprior tokens retain spatial locality. PrismCodec instead computes semantic conditions from an evidence image at decoding time, preserves patch-token sets from multiple pretrained depths, and assigns them to successive DiT block groups. This makes the semantic hierarchy an explicit inference-time condition. OneDC additionally uses diffusion distillation for perceptual refinement; its tokenizer supervision and refinement therefore address different roles from PrismCodec's Stage-II flow-map distillation.

DiT-IC (Shi et al., 2026a) projects compressed features into the text conditioning space and uses contrastive alignment with text embeddings during training; inference uses the latent condition alone. Its adaptation also includes a variance-derived spatial pseudo-timestep and self-distillation that aligns reconstructed latents with frozen encoder outputs. PrismCodec preserves multiple visual representation depths and adds a separate diffusion-native VAE structural condition. Its one-step adaptation also follows a different path: Stage I first learns continuous conditional transport around the codec state, and Stage II distills the resulting finite-time map. The source perturbation scale therefore defines a Gaussian neighborhood of the transmitted codec state, whereas DiT-IC's pseudo-timestep controls spatially varying reconstruction strength.

A.5 Removing the Cross-Attention Pathway: AEIC

AEIC (Zhang et al., 2026a) simplifies the SD-Turbo denoiser by removing the text encoder, timestep embeddings, and cross-attention layers. Its codec synthesis transform produces both a denoising input and a structural residual, so reconstruction still depends on image-specific compressed information. Dual-side feature distillation transfers knowledge from a moderate-encoder model to shallow-encoder

variants. This targets encoder efficiency and feature preservation. PrismCodec retains explicit visual cross-attention and adapts a time-dependent conditional flow before distilling its map in Stage II. The two uses of distillation therefore solve different problems: AEIC targets encoder efficiency and feature transfer, whereas PrismCodec compresses an already bitstream-grounded generative transport into one step.

A.6 Auxiliary Reconstruction Guidance: SODEC

SODEC (Chen et al., 2026d) is closely related to our use of a decoded image as a visual interface. Its fidelity decoder reconstructs an auxiliary image from compressed latents; a frozen ViT and a learned projection produce a cross-attention condition. A separate transform supplies the input to an SD2.1 U-Net evaluated at a fixed timestep. Both methods therefore derive visual guidance and the generative input from the bitstream. PrismCodec further separates the conditioning role into two pretrained representation spaces: semantic features remain distinct across DINOv3 depths, while a diffusion-native VAE latent provides spatially aligned structural guidance. The shared evidence image gives both encoders a common image-domain view of the decoded evidence, and its auxiliary objective includes paired semantic supervision (Appendix C.2).

The adaptation mechanisms also differ. SODEC pretrains a high-bitrate codec, warms up the diffusion path, and then tightens the rate penalty during joint optimization. PrismCodec instead learns codec-centered conditional transport, distills that transport into a one-step map, and then contracts the stochastic source onto the deterministic codec state. SODEC's rate annealing changes the coding constraint; PrismCodec's source-noise annealing changes the generator's input distribution.

A.7 Visual Priors and Progressive Flow Adaptation

PrismCodec's semantic branch uses dense DINOv3 features (Siméoni et al., 2025), while its structural branch uses the frozen VAE encoder. Both are explicit inference-time conditions derived from the same bitstream-dependent evidence image. This differs from representation-alignment methods such as REPA (Yu et al., 2024), which use pretrained visual features as training supervision for diffusion hidden states. More generally, using semantic features in an auxiliary loss or discriminator is not equivalent to supplying those features as decoder-side visual context.

The adaptation axis is likewise compositional. Stage I builds on flow matching and rectified flow (Lipman et al., 2023; Liu et al., 2022); Stage II uses established AYF-EMD distillation with an interval-average velocity parameterization (Sabour et al., 2025; Geng et al., 2025). PMRF (Ohayon et al., 2024) provides a related precedent for centering a stochastic flow source on an informative deterministic estimate. PrismCodec instead centers the source on an entropy-decoded codec state, without a posterior-mean assumption or inherited PMRF optimality guarantee, and couples that source with hierarchical visual conditions. Stage III then specializes the distilled map to the deterministic codec state used at deployment (Sections 3.4.1–3.4.3).

B NOTATION

Tables 6 and 7 collect the symbols that define the two central parts of PrismCodec: bitstream-derived visual grounding and progressive adaptation to deterministic one-step decoding. Bold symbols denote tensor-valued states, features, predictions, or ordered collections. Superscripts str and sem identify the two visual-prior types, while I–III denote training stages. Latent subscripts 0 and 1 denote flow endpoints. The Stage-III multiplier ρ controls source contraction independently of the flow times.

Table 6: Core notation used in the main text and pipeline figures.

Symbol	Meaning
$\mathbf{x}, \mathbf{b}$	Source image and transmitted binary bitstream.
$\mathbf{f}_c, \mathbf{z}_c$	Shared decoded feature and projected codec-centered latent state.
$\widetilde{\mathbf{x}}$	Internal evidence image produced by the visual-evidence decoder.
$\mathbf{z}_0$	Clean target latent from the frozen VAE encoder.
$\mathbf{z}, \mathbf{z}_t$	Generic generative latent and paired training interpolant at time t.
$\mathbf{z}_1, \mathbf{z}_1^{(\rho)}$	Perturbed codec states in Stages I/II and Stage III.
$\widehat{\mathbf{z}}_0, \widehat{\mathbf{x}}$	Predicted generative latent and final reconstruction.
$\mathcal{P}_c, \mathcal{D}_e$	Latent projection and visual-evidence decoder acting on $\mathbf{f}_c$.
$\mathcal{G}_\theta, \mathcal{D}_g$	One-step latent map and frozen, calibrated generative VAE decoder.
$\mathcal{S}, \mathcal{H}$	Structural mapping and hierarchical semantic mapping.
$\mathbf{P}^{\mathrm{str}}$	Structural prior: a VAE latent with the same shape as $\mathbf{z}$.
$\mathbf{P}_\ell^{\mathrm{sem}}, \mathbf{P}^{\mathrm{sem}}$	Semantic patch-token matrix at level ℓ and the ordered hierarchy of all L levels.
$\mathcal{P}$	Joint visual condition $(\mathbf{P}^{\mathrm{str}}, \mathbf{P}^{\mathrm{sem}})$, fixed along flow time.
L, ℓ	Number of semantic levels and level index $1 \le \ell \le L$.
t, r	Current and destination flow times; $r = t$ in Stage I and $r = 0$ in Stages II/III.
$\boldsymbol{\epsilon}, \mathbf{0}, \mathbf{I}$	Standard Gaussian noise, zero vector, and identity of the latent dimension.
σ_s	Per-coordinate source perturbation standard deviation.
ρ	Scheduled Stage-III perturbation multiplier, reduced from 1 to 0.
$\mathbf{v}_\theta, \mathbf{v}_T$	Trainable instantaneous velocity and frozen teacher velocity.
$\mathbf{u}_\theta, \theta$	DiT predictor and its parameters; other module parameters are implicit.
$\mathcal{R}$	Expected entropy-model rate estimate, in bits per pixel (bpp).
$\mathcal{L}_{\mathrm{FM}}, \mathcal{L}_{\mathrm{map}}$	Stage-I flow-matching loss and Stage-II regression loss.
$\mathcal{L}_{\mathrm{ev}}, \mathcal{L}_{\mathrm{img}}$	Paired evidence loss and final-image reconstruction loss.
$\mathcal{L}_{\mathrm{adv}}^{\mathrm{DINO}}$	Generator-side adversarial loss on DINOv3 features.
$\mathcal{L}_{\mathrm{I}}, \mathcal{L}_{\mathrm{II}}, \mathcal{L}_{\mathrm{III}}$	Total objectives of the three training stages.
λ with subscripts	Non-negative weights of the named loss terms; stage superscripts distinguish their settings.

Table 7: Additional notation for construction, distillation, and analysis.

Symbol	Meaning
$\widehat{\mathbf{y}}, \widehat{\mathbf{h}}, p$	Quantized main/hyperlatents and learned discrete probability masses.
$\mathbb{E}, \mathcal{N}, \mathcal{U}$	Expectation, Gaussian distribution (or its explicitly evaluated density), and uniform distribution.
$i_\ell, \Phi_{\mathrm{DINO}}^{(i_\ell)}$	Zero-based DINOv3 block index and frozen feature extractor at that index.
$\mathrm{Patch}, \mathrm{LN}_\ell, \mathcal{A}_\ell$	Spatial-token selection, learned layer normalization, and learned per-depth MLP projection.
$N_\ell, n, \mathbf{F}_{\ell,n}$	Patch count, patch index, and unit-normalized token for paired semantic supervision.
$\mathcal{L}_{\mathrm{ev}}^{\mathrm{DINO}}$	Per-image paired DINOv3 feature term inside the expected evidence loss.
$K, k, \gamma(k)$	DiT block count, one-based block index, and assigned semantic level.
$\mathbf{H}_0, \mathbf{H}_k$	Patch-embedded input and output of DiT block k.
$\overline{\mathbf{H}}_k, \widetilde{\mathbf{H}}_k, \mathbf{e}_{t,r}$	Intermediate block token sequences and time conditioning.
$\mathrm{SA}, \mathrm{CA}, \mathrm{FFN}$	Self-attention, cross-attention, and feed-forward network.
$\mathrm{PatchEmbed}, [\cdot\,;\,\cdot]$	DiT patch embedding and ordered channel concatenation.
$\mathbf{v}_{\mathrm{opt}}$	Conditional-mean velocity minimizing the fixed-construction flow loss.
$w(t, 0)$	Non-negative Stage-II regression weight for the interval $[0, t]$.
$\mathbf{u}_{\mathrm{tgt}}, \mathrm{sg}$	AYF-EMD regression target and stop-gradient operator.
$q_1, q_{1,\rho}, \delta_{\mathbf{z}_c}$	Stage-I/II source density, Stage-III source law, and point mass at $\mathbf{z}_c$.
d, M, W_2	Latent coordinate count, fixed-decoder Lipschitz constant, and 2-Wasserstein distance.
$\mathcal{J}, j, \lvert\mathcal{J}\rvert$	Discriminator feature-index set, one of its indices, and its size.
$D_{\omega,j}, \mathrm{Head}_{\omega,j}, \omega$	Feature-level discriminator, trainable head, and head parameters.
$\ell_D^+, \ell_D^-, \ell_G$	Real-image, fake-image, and generator classification-loss functions.
$\mathcal{L}_D^{\mathrm{DINO}}$	Discriminator-side objective, separate from the generator loss.

C Additional Method Details

C.1 Codec State and Rate Estimate

Entropy decoding and synthesis recover the shared feature $\mathbf{f}_c$, from which PrismCodec exposes the transmitted evidence through two complementary paths. The latent projection $\mathcal{P}_c$ produces $\mathbf{z}_c = \mathcal{P}_c(\mathbf{f}_c)$, which determines the generative starting state, while the evidence decoder $\mathcal{D}_e$ produces $\widetilde{\mathbf{x}} = \mathcal{D}_e(\mathbf{f}_c)$, the shared image-domain interface used to construct the semantic and structural priors. Both paths are therefore derived from the same received bitstream.

Let $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{h}}$ denote the quantized main latent and hyperlatent. For an RGB image of height H and width W, the expected rate estimate in bits per pixel (bpp) is

$$\mathcal{R} = \mathbb{E}\left[\frac{-\log_2 p(\widehat{\mathbf{y}} \mid \widehat{\mathbf{h}}) - \log_2 p(\widehat{\mathbf{h}})}{HW}\right]. \tag{13}$$

Here, p denotes discrete probability masses from the learned entropy models, including the context factorization of the main latent. The expectation averages training images and the codec's training-time quantization, with each image's estimated code length normalized by its own pixel count. Measured bpp uses the actual serialized stream length, including coding overhead and required metadata, divided by the original image's pixel count before averaging over images.

C.2 Evidence-Preservation Objective

Let $\mathbf{F}_\ell(\cdot)$ be the spatial patch-token matrix from the ℓ-th selected DINOv3 depth, with each token normalized to unit ℓ_2 norm. Write N_ℓ for its token count and $\mathbf{F}_{\ell,n}$ for token n. The stop-gradient operator sg preserves its argument's value while blocking optimization gradients. For an image pair, the semantic term is

$$\mathcal{L}_{\mathrm{ev}}^{\mathrm{DINO}} = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{N_\ell}\sum_{n=1}^{N_\ell}\left[1 - \left\langle \mathbf{F}_{\ell,n}(\widetilde{\mathbf{x}}), \mathrm{sg}\big(\mathbf{F}_{\ell,n}(\mathbf{x})\big)\right\rangle\right]. \tag{14}$$

The inner product is cosine similarity because the token features are normalized. The source-image features are targets; their gradients are stopped. DINOv3 parameters are frozen on both branches, while gradients through its evidence-image input are retained to train $\mathcal{D}_e$ and, when enabled, the upstream codec. The total evidence loss averages the paired terms:

$$\mathcal{L}_{\mathrm{ev}} = \mathbb{E}\big[\mathrm{MSE}(\widetilde{\mathbf{x}}, \mathbf{x}) + \lambda_{\text{ev-sem}}\mathcal{L}_{\mathrm{ev}}^{\mathrm{DINO}}\big]. \tag{15}$$

Here, MSE averages squared differences over all pixels and RGB channels, and $\lambda_{\text{ev-sem}}$ balances paired pixel fidelity and semantic consistency. Stages I and III apply their respective evidence-loss weights, as specified in the corresponding objectives. Importantly, this objective supervises the *interface* from which the visual priors are extracted; it does not directly optimize the final generative reconstruction.

C.3 Prior Construction and Injection

The semantic branch preserves the depth structure of the pretrained visual representation instead of collapsing it into one condition. Let i_ℓ be the zero-based DINOv3 block index selected for semantic level $\ell \in \{1, \ldots, L\}$. The frozen feature extractor $\Phi_{\mathrm{DINO}}^{(i_\ell)}$ includes image preprocessing and the applicable pretrained normalization; Patch retains only spatial tokens. The projected semantic tokens are

$$\mathbf{P}_\ell^{\mathrm{sem}} = \mathcal{A}_\ell\Big(\mathrm{LN}_\ell\Big(\mathrm{Patch}\Big(\Phi_{\mathrm{DINO}}^{(i_\ell)}(\widetilde{\mathbf{x}})\Big)\Big)\Big), \tag{16}$$

where LN_ℓ is a learned layer normalization and $\mathcal{A}_\ell$ is a learned MLP projection into the DiT hidden dimension. These operations preserve token order. The implementation uses indices $(5, 11, 17, 23)$, corresponding to the 6th, 12th, 18th, and 24th transformer blocks. The last selected level uses DINOv3's pretrained final normalization. Evidence crops are resized to 512×512 and normalized with the visual encoder's input statistics; class and register tokens are excluded. These operations are collectively denoted by $\mathcal{H}$ in the main text.

The codec state $\mathbf{z}_c$, structural prior $\mathbf{P}^{\mathrm{str}}$, and clean target $\mathbf{z}_0$ share the frozen generative VAE's latent coordinates. The DiT concatenates the evolving state with $\mathbf{P}^{\mathrm{str}}$ along the channel dimension and predicts a latent update with the same shape as the evolving state.

For the structural mapping $\mathcal{S}$, images in $[0, 1]$ are converted to the VAE input range $[-1, 1]$. The encoder's posterior mode is calibrated by subtracting the checkpoint's per-channel latent mean and dividing by its per-channel standard deviation. $\mathcal{D}_g$ reverses this calibration before VAE decoding and converts the decoded image back to $[0, 1]$. The fixed latent calibration is distinct from the source-noise scale σ_s.

For K DiT blocks divided into L equal consecutive groups, with K divisible by L, block $k \in \{1, \ldots, K\}$ receives semantic level

$$\gamma(k) = 1 + \left\lfloor \frac{(k-1)L}{K} \right\rfloor, \tag{17}$$

where $\lfloor \cdot \rfloor$ is the floor operation. With $L = 4$, the four semantic levels guide four consecutive DiT groups, as shown in Fig. 2.

Let $\mathbf{H}_0 = \mathrm{PatchEmbed}([\mathbf{P}^{\mathrm{str}}; \mathbf{z}])$ be the initial token sequence; $[\,\cdot\,;\,\cdot\,]$ concatenates channels in the displayed order. $\mathbf{H}_k$ is the output of block k, so block k receives $\mathbf{H}_{k-1}$. Let $\mathbf{e}_{t,r}$ denote its time conditioning. Suppressing normalization and modulation details, the block has the form

$$\overline{\mathbf{H}}_k = \mathbf{H}_{k-1} + \mathrm{SA}_k(\mathbf{H}_{k-1}; \mathbf{e}_{t,r}), \tag{18}$$

$$\widetilde{\mathbf{H}}_k = \overline{\mathbf{H}}_k + \mathrm{CA}_k\left(\overline{\mathbf{H}}_k, \mathbf{P}^{\mathrm{sem}}_{\gamma(k)}\right), \tag{19}$$

$$\mathbf{H}_k = \widetilde{\mathbf{H}}_k + \mathrm{FFN}_k(\widetilde{\mathbf{H}}_k; \mathbf{e}_{t,r}). \tag{20}$$

Here, SA, CA, and FFN denote self-attention, cross-attention, and the feed-forward network. The barred and tilded token sequences are the intermediate outputs after self- and cross-attention, respectively. The destination-time embedding is multiplied by $t - r$ before combination with the current-time embedding; its contribution therefore vanishes when $r = t$ in Stage I. In Stage II, $r = 0$ is fixed and only t varies; Stage III and deployment use $(t, r) = (1, 0)$. The structural prior is therefore present from the generator input, whereas the semantic hierarchy supplies depth-dependent cross-attention context throughout the DiT. This realizes the complementary roles described in the main text without introducing any separately transmitted condition.

C.4 Conditional Flow Targets

This subsection clarifies what Stage I learns when the source is centered on the codec state. Hold the codec and condition-producing parameters fixed, and let the training construction induce a joint law of $(\mathbf{z}_0, \mathbf{z}_1, \mathcal{P})$. For a fixed t and $\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1$, least-squares regression has the population optimum

$$\mathbf{v}_{\mathrm{opt}}(\mathbf{z}, t; \mathcal{P}) = \mathbb{E}[\mathbf{z}_1 - \mathbf{z}_0 \mid \mathbf{z}_t = \mathbf{z}, \mathcal{P}]\,. \tag{21}$$

This follows by conditioning the squared error on $(\mathbf{z}_t, \mathcal{P})$: the cross term between its conditional mean and zero-mean residual vanishes. Under the regularity conditions for flow matching, this field generates the corresponding conditional marginal path (Lipman et al., 2023). Individual linear training interpolants need not be its ODE trajectories; the conditional expectation averages over different endpoint pairs.

The source in Eq. (6) has conditional density $q_1(\mathbf{z} \mid \mathbf{z}_c) = \mathcal{N}(\mathbf{z}; \mathbf{z}_c, \sigma_s^2 \mathbf{I})$, where the right-hand side is the Gaussian density evaluated at $\mathbf{z}$, with mean $\mathbf{z}_c$ and covariance $\sigma_s^2 \mathbf{I}$. Its per-coordinate variance is σ_s^2. This conditioning differs from conditioning on the visual priors alone:

$$q_1(\mathbf{z} \mid \mathcal{P}) = \mathbb{E}_{\mathbf{z}_c \mid \mathcal{P}}\left[\mathcal{N}(\mathbf{z}; \mathbf{z}_c, \sigma_s^2 \mathbf{I})\right]. \tag{22}$$

The distinction above formalizes why the codec-centered source is not redundant with the visual condition: $\mathcal{P}$ need not identify the complete bitstream or the exact codec state that centers the source. At the same time, the conditional-flow interpretation should not be overextended. It does not guarantee recovery of every paired clean latent or exact matching of the full image posterior conditioned on a bitstream, and the fixed-construction regression argument does not establish convergence when the codec and visual interface are optimized jointly.

C.5 Codec-Conditioned AYF-EMD Distillation

We follow the AYF-EMD objective (Sabour et al., 2025) to distill the Stage-I codec-conditioned flow into a one-step decoder, as summarized in Eq. (10). Teacher and student share the paired interpolants $\mathbf{z}_t$ and visual priors $\mathcal{P}$ associated with the codec-centered source $\mathbf{z}_1 = \mathbf{z}_c + \sigma_s \boldsymbol{\epsilon}$. The source-noise scale remains fixed throughout Stage II; noise annealing starts in Stage III. We specialize the average-velocity parameterization by fixing the destination time to $r = 0$, so the student learns transport to the clean endpoint and yields the one-step map in Eq. (9) at $t = 1$. The frozen Stage-I DiT serves as the teacher and provides the student's initialization. The codec and all condition-producing modules, including the evidence decoder, remain frozen; only the student DiT is updated.

C.6 Source Contraction and Its Scope

Stage III contracts the stochastic source used for flow learning and distillation onto the deterministic codec state used at deployment. The following fixed-parameter analysis makes this source transition explicit. Fix the codec, condition-producing networks, and generative-decoder parameters, and let d be the number of coordinates in the starting latent. For $\rho > 0$, the conditional source is $q_{1,\rho}(\cdot \mid \mathbf{z}_c) = \mathcal{N}(\mathbf{z}_c, \rho^2 \sigma_s^2 \mathbf{I})$; at $\rho = 0$ it is the point mass $\delta_{\mathbf{z}_c}$, rather than a non-degenerate Gaussian density. Here, $q_{1,\rho}$ denotes the conditional source law and $\delta_{\mathbf{z}_c}$ the point mass at $\mathbf{z}_c$. Its covariance is $\rho^2 \sigma_s^2 \mathbf{I}$. Writing W_2 for the 2-Wasserstein distance, every coupling with this point mass has a constant second component, hence

$$W_2^2(q_{1,\rho}(\cdot \mid \mathbf{z}_c), \delta_{\mathbf{z}_c}) = \mathbb{E}\|\rho \sigma_s \boldsymbol{\epsilon}\|_2^2 = d\rho^2 \sigma_s^2. \tag{23}$$

Thus, at a fixed parameter snapshot, the source distribution approaches the deployed codec state continuously in W_2 as $\rho \to 0$. This statement concerns the input distribution only; it neither requires nor implies convergence in total variation to the point mass.

For a fixed condition, suppose the complete map $\mathbf{z} \mapsto \mathcal{D}_g(\mathcal{G}_\theta(\mathbf{z}; \mathcal{P}))$ is M-Lipschitz in its starting latent, meaning its output distance is at most M times its input distance. Then

$$\mathbb{E}\|\mathcal{D}_g(\mathcal{G}_\theta(\mathbf{z}_c + \rho \sigma_s \boldsymbol{\epsilon}; \mathcal{P})) - \mathcal{D}_g(\mathcal{G}_\theta(\mathbf{z}_c; \mathcal{P}))\|_2^2 \le M^2 d \rho^2 \sigma_s^2. \tag{24}$$

The bound compares perturbed and unperturbed outputs of the same fixed decoder and therefore formalizes only the continuity of the deployment transition; it is not a bound on reconstruction error to the source image. During end-to-end training, both the source center and DiT parameters can change, so the fixed-snapshot statements do not establish optimizer convergence, monotonic distortion reduction, or preservation of the teacher's output distribution. The deployed decoder is intentionally deterministic and returns one reconstruction per bitstream rather than sampling the complete conditional image posterior.

C.7 Final-Image and Adversarial Objectives

Stage III jointly optimizes the codec and one-step decoder for the final rate–distortion–perception objective, both during source-noise annealing and after the source becomes deterministic. The final-image term combines complementary full-reference criteria:

$$\begin{aligned}\mathcal{L}_{\text{img}} = \mathbb{E}\big[&\lambda_{\text{mse}} \,\text{MSE}(\widehat{\mathbf{x}}, \mathbf{x}) + \lambda_{\text{ms}} \big(1 - \text{MS-SSIM}(\widehat{\mathbf{x}}, \mathbf{x})\big) \\ &+ \lambda_{\text{lpips}} \,\text{LPIPS}(\widehat{\mathbf{x}}, \mathbf{x}) + \lambda_{\text{dists}} \,\text{DISTS}(\widehat{\mathbf{x}}, \mathbf{x})\big].\end{aligned} \tag{25}$$

The weights balance pixel accuracy with perceptual and structural similarity (Wang et al., 2003; Zhang et al., 2018; Ding et al., 2022). As in the evidence objective, MSE averages squared differences over all pixels and RGB channels. The expectation averages image pairs and the current Stage-III source perturbation. The rate term is the estimated bpp defined in Appendix C.1.

For adversarial supervision, we adapt the vision-aided discriminator formulation (Kumari et al., 2022). Let $\mathcal{J}$ be the selected set of DINOv3 block indices for the discriminator, and let $\text{Head}_{\omega,j}$ be its trainable head at index $j \in \mathcal{J}$. These indices need not be the semantic-conditioning indices i_ℓ. The discriminator at each level is

$$D_{\omega,j}(\mathbf{x}) = \text{Head}_{\omega,j}\left(\Phi_{\text{DINO}}^{(j)}(\mathbf{x})\right), \tag{26}$$

where ω comprises the head parameters and the DINOv3 backbone remains frozen. Let ℓ_D^+, ℓ_D^-, and ℓ_G denote the real-image, fake-image, and generator classification losses, respectively. Then

$$\mathcal{L}_D^{\mathrm{DINO}} = \mathbb{E}\left[\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\left\{\ell_D^+\big(D_{\omega,j}(\mathbf{x})\big) + \ell_D^-\big(D_{\omega,j}(\mathrm{sg}(\widehat{\mathbf{x}}))\big)\right\}\right], \tag{27}$$

$$\mathcal{L}_{\mathrm{adv}}^{\mathrm{DINO}} = \mathbb{E}\left[\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\ell_G\big(D_{\omega,j}(\widehat{\mathbf{x}})\big)\right]. \tag{28}$$

Here, $|\mathcal{J}|$ is the number of discriminator feature levels. For the discriminator update, the reconstructed image is detached. For the generator update, discriminator parameters are held fixed while gradients through the discriminator and DINOv3 input reach the reconstructed image. These distribution-level objectives differ from the paired evidence loss, which directly compares each intermediate evidence image with its source.

C.8 Training and Gradient Paths

Table 8 makes explicit how the progressive adaptation stages differ in what they optimize. A frozen visual encoder can still transmit gradients to its input in Stages I and III. The generative VAE decoder remains frozen throughout, but Stage III image-loss gradients pass through $\mathcal{D}_g$ to the upstream codec and DiT. Stage II isolates the flow-map distillation problem: the codec, state projection, evidence decoder, and semantic projections are frozen, and detaching $\mathbf{z}_c$ together with the complete condition $\mathcal{P}$ confines map-loss gradients to the student. No evidence-loss update is applied in this stage.

Table 8: **Optimization roles across the three adaptation stages.** Stage II isolates one-step flow-map distillation: only the student is updated, using the frozen Stage-I teacher and fixed bitstream-derived conditions.

Component	Stage I	Stage II	Stage III
Codec and state projection	Train	Freeze	Train
Visual-evidence decoder	Train	Freeze	Train
Semantic projections	Train	Freeze	Train
DiT / average-velocity student	Train	Train	Train
DINOv3 and VAE encoders	Freeze	Freeze	Freeze
Generative VAE decoder	Not used	Not used	Freeze
DINOv3 discriminator heads	Not used	Not used	Train

D ADDITIONAL KID EVALUATION

The main experiments use FID as a distribution-level complement to paired fidelity metrics. We additionally report Kernel Inception Distance (KID) (Bińkowski et al., 2018) on the DIV2K validation and CLIC2020 test sets to provide a second view of distributional realism. For PrismCodec, KID is computed from the same 256×256 patches used for FID in Sec. 4.1, so the comparison changes the distributional metric rather than the evaluation samples. Fig. 6 reports the measured rate–KID curves.

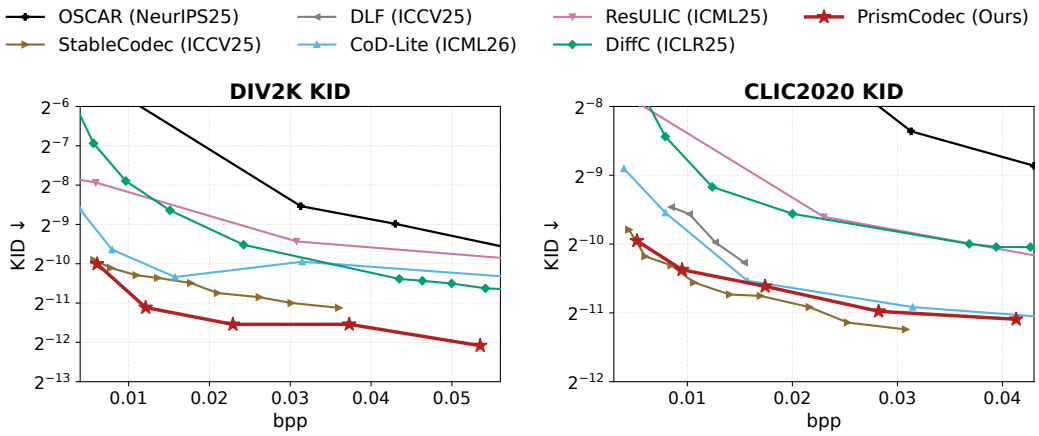

Figure 6: **Rate–KID comparisons on DIV2K and CLIC2020.** Markers denote measured operating points; lower KID is better. The vertical axes use a base-2 logarithmic scale, and the bitrate windows match those used in the main rate–perception comparisons. PrismCodec is shown in red.

E LIMITATIONS AND FUTURE WORK

PrismCodecprioritizes rate–perception performance and effective adaptation of a pretrained generative prior, rather than minimizing model size. As a result, the current decoder remains relatively large compared with lightweight compression models. Moreover, high-resolution images are decoded with overlapping latent tiles to keep memory consumption manageable. These computational costs accompany the substantial rate–perception gains obtained by the current design, but leave clear opportunities for improving deployment efficiency. Future work will explore more compact generative decoders and architectures capable of processing full-resolution images directly without tiled inference, while preserving the benefits of bitstream-grounded one-step decoding.

F Additional Qualitative Comparisons

Fig. 7 extends the main comparison with six additional DIV2K examples.

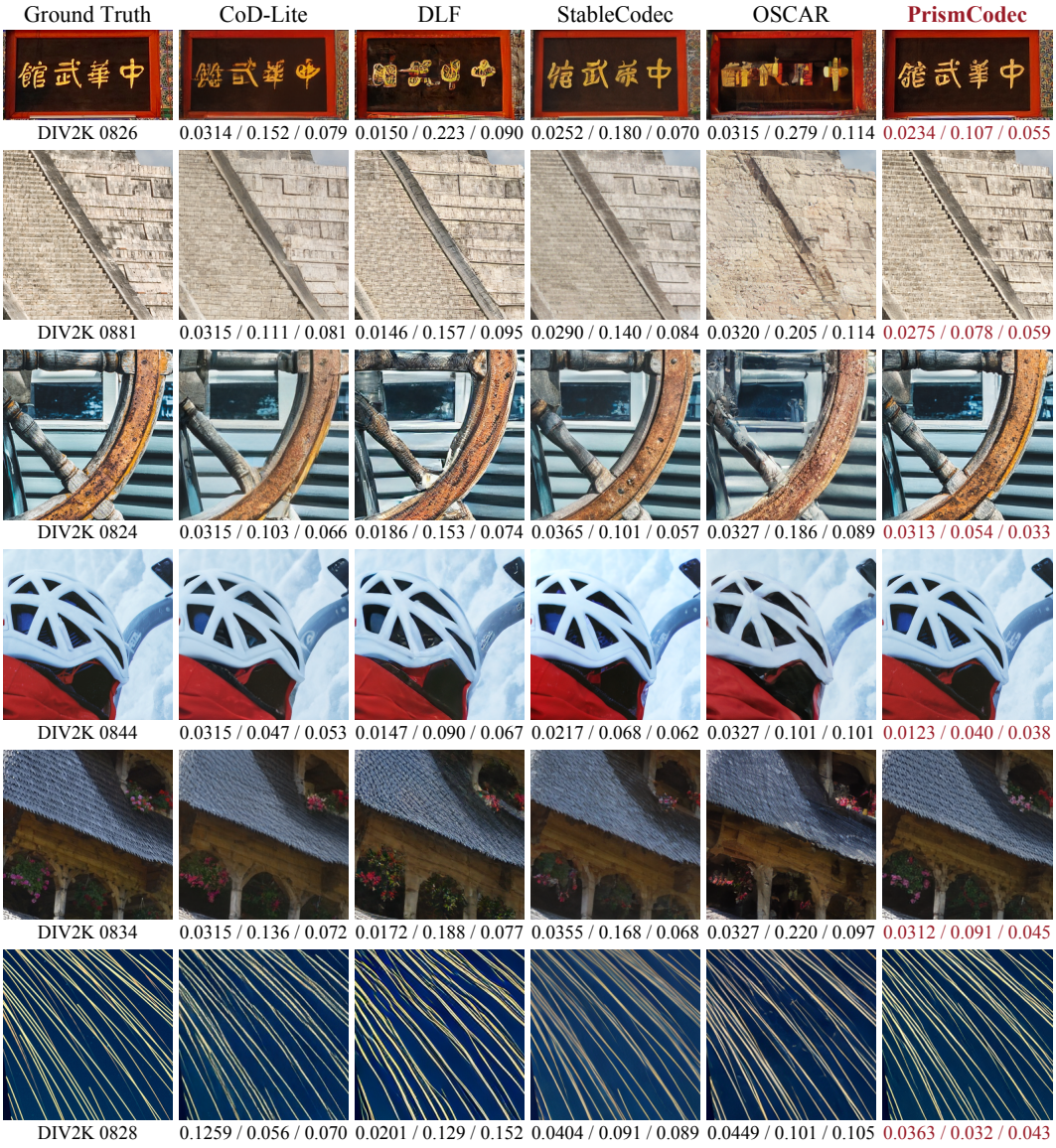

Figure 7: **Additional qualitative comparisons on DIV2K.** Rows show 0826, 0881, 0824, 0844, 0834, and 0828. Each method uses the same native crop. Labels give full-image bpp / LPIPS / DISTS; lower values are better. These examples examine signage, stepped boundaries, wood grain and spoke geometry, small manufactured structures, architectural ornament, and thin light trails, respectively.