Steering One-Step Diffusion Model with Fidelity-Rich Decoder for Fast Image Compression

Zheng Chen[1*], **Mingde Zhou**[1*], **Jinpei Guo**[2], **Jiale Yuan**[1], **Yifei Ji**[1], **Yulun Zhang**[1†]

[1]Shanghai Jiao Tong University,
[2]Carnegie Mellon University

Abstract

Diffusion-based image compression has demonstrated impressive perceptual performance. However, it suffers from two critical drawbacks: **(1)** excessive decoding latency due to multi-step sampling, and **(2)** poor fidelity resulting from over-reliance on generative priors. To address these issues, we propose SODEC, a novel single-step diffusion image compression model. We argue that in image compression, a sufficiently informative latent renders multi-step refinement unnecessary. Based on this insight, we leverage a pre-trained VAE-based model to produce latents with rich information, and replace the iterative denoising process with a single-step decoding. Meanwhile, to improve fidelity, we introduce the fidelity guidance module, encouraging output that is faithful to the original image. Furthermore, we design the rate annealing training strategy to enable effective training under extremely low bitrates. Extensive experiments show that SODEC significantly outperforms existing methods, achieving superior rate-distortion-perception performance. Moreover, compared to previous diffusion-based compression models, SODEC improves decoding speed by more than 20×.

Code — https://github.com/zhengchen1999/SODEC
Extended version — https://arxiv.org/abs/2508.04979

Introduction

The rising cost of data storage and transmission underscores the importance of image compression. Traditional codecs such as JPEG2000 (Taubman, Marcellin, and Rabbani 2002) and VVC (Bross et al. 2021) perform reliably at medium to high bitrates. However, when the bitrate drops to low levels (*e.g.*, <0.1 bpp), they tend to produce block artifacts, blurring, and structural distortions. Achieving a balance between distortion and perceptual quality under low bitrate constraints remains a challenging problem.

In recent years, learning-based image compression models built upon variational autoencoders (VAEs) (Kingma, Welling et al. 2019) have surpassed traditional methods in the rate-distortion trade-off (Ballé et al. 2018; Cheng et al. 2020; Van Den Oord, Vinyals et al. 2017). Benefiting from advances in probabilistic modeling, such as hyperpriors (Ballé et al. 2018; Minnen, Ballé, and Toderici 2018), these approaches typically excel in distortion-oriented metrics like PSNR and MS-SSIM. Moreover, to better align with human perception, subsequent works further incorporate perception-oriented objectives, leading to a more comprehensive rate-distortion-perception framework (Blau and Michaeli 2019; Mentzer et al. 2020; Muckley et al. 2023; Agustsson et al. 2023; He et al. 2022b). These methods achieve a more realistic reconstruction by employing distortion and perceptual losses to enhance realism. However, VAE-based methods struggle to reconstruct details when operating at extremely low bitrates, resulting in poor perceptual quality. In other words, while the reconstructions may appear "technically correct", they often lack realism.

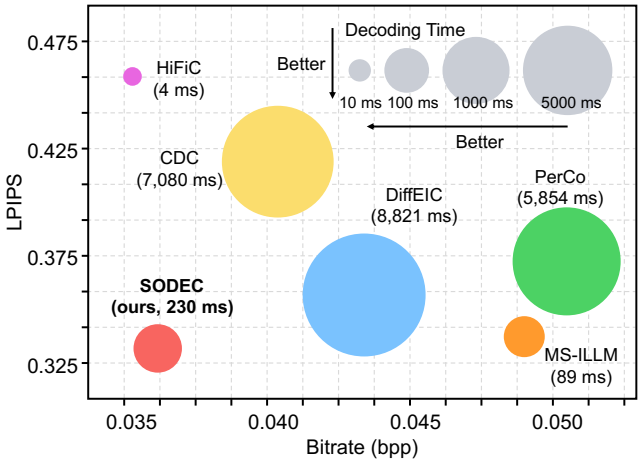

Figure 1: LPIPS-bitrate-latency comparison on DIV2K-Val. Decoding time is measured on 512×512 images using one A6000 GPU. Our method achieves the best perceptual quality (*i.e.*, LPIPS). Meanwhile, compared to the multi-step diffusion-based method DiffEIC (Li et al. 2024), our method offers a 38× speedup in decoding time.

In contrast, diffusion models (Ho, Jain, and Abbeel 2020) have recently demonstrated remarkable performance in the rate-perception trade-off, due to their powerful generative priors. Specifically, in diffusion-based methods, the encoder produces a compact latent representation, while decoding is reformulated as a multi-step conditional denoising process (Theis et al. 2022; Lei et al. 2023b). Guided by con-

*These authors contributed equally.
†Corresponding author: Yulun Zhang, yulun100@gmail.com

ditional signals derived from the bitstream, the diffusion model iteratively refines a noisy latent (Yang and Mandt 2023; Vonderfecht and Liu 2025; Careil et al. 2024; Ghouse et al. 2023; Relic et al. 2024). Thus, diffusion-based models can synthesize highly realistic textures and details, even under extreme compression. Moreover, some approaches integrate global (*e.g.*, text prompts) or local (*e.g.*, quantized features) guidance to constrain the generative process (Pan, Zhou, and Tian 2022; Careil et al. 2023; Li et al. 2024).

However, such models face two critical challenges: **(1) High latency.** The multi-step denoising process incurs substantial decoding latency and computational cost. This limits their applicability in real-time or resource-constrained scenarios. **(2) Low fidelity.** The generative nature of diffusion models makes them heavily reliant on pre-trained priors rather than the input itself. This leads to reconstructions that deviate from the original content, compromising fidelity.

To address these challenges, we propose SODEC (steering one-step diffusion model with fidelity-rich decoder), a novel image compression model designed for low-bitrate scenarios. Our SODEC is designed around efficient decoding and high-fidelity guidance. **(1) Single-step decoding.** To mitigate the high latency of multi-step diffusion, we replace the iterative denoising process with a single-step process. Benefits to the informative latent representations produced by the pre-trained VAE-based compression model, single-step decoding is sufficient to realize high-quality reconstruction. **(2) Fidelity guidance module.** To compensate for the potential fidelity loss, we employ a pre-trained VAE-based compression model to produce a high-fidelity preliminary reconstruction. This reconstruction serves as explicit visual guidance to the diffusion model, encouraging outputs faithful to the source image. **(3) Rate annealing training strategy.** To ensure effective training at extremely low bitrates, we adopt a three-stage optimization. The model is first pre-trained at higher bitrates to learn informative representations. Then, we gradually anneal the model to the target bitrate, selectively preserving essential information.

Benefits to above designs, SODEC achieves impressive performance in terms of rate-distortion-perception trade-off. Furthermore, due to the single-step and lightweight conditioning, our SODEC achieves excellent decoding efficiency. As shown in Fig. 1, compared to multi-step diffusion paradigms (*e.g.*, PerCo (Careil et al. 2023), DiffEIC (Li et al. 2024)), SODEC delivers over 20× speedup.

Our contributions are summarized as follows:

- We propose SODEC, a single-step diffusion image compression model that significantly accelerates decoding while preserving high perceptual-fidelity quality.
- We introduce the fidelity guidance module, a diffusion guidance mechanism conditioned on high-fidelity reconstruction, effectively improving content fidelity.
- We develop the rate annealing training strategy, a three-stage optimization scheme that enables the model to retain critical information at extremely low bitrates.
- SODEC achieves state-of-the-art performance in the rate-distortion-perception trade-off, while delivering significantly improved decoding efficiency.

Related Work

VAE-based Compression Model

Compressing images at extremely low bitrates is a challenge where traditional methods like JPEG2000 (Taubman, Marcellin, and Rabbani 2002) and VVC (Bross et al. 2021) often produce severe blurring and artifacts. Recently, learned compression based on Variational Autoencoders (VAEs) (Kingma and Welling 2014) has surpassed traditional codecs in rate-distortion performance (Ballé et al. 2018; Cheng et al. 2020; Wang et al. 2022; Minnen and Singh 2020; He et al. 2022a), largely due to innovations like the hyperprior model. This architecture is refined with sophisticated context models and quantization strategies, such as the hierarchical prior model (Minnen, Ballé, and Toderici 2018) and VQ-VAE (Van Den Oord, Vinyals et al. 2017), achieving state-of-the-art performance on distortion-oriented metrics like PSNR and MS-SSIM. Subsequently, to enhance visual realism, perception-oriented models (Tschannen, Agustsson, and Lucic 2018; Blau and Michaeli 2019; Agustsson et al. 2019; Mentzer et al. 2020; Muckley et al. 2023) are introduced to optimize the rate-distortion-perception. However, these models still tend to produce artifacts and lack detail at extremely low bitrates.

Diffusion-based Compression Model

Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022) excel at high-quality image synthesis by framing generation as an efficient, latent-space noise prediction task. Recent compression works adapt these models for image compression by treating it as a conditional denoising problem (Saharia et al. 2021; Xia et al. 2025; Liu et al. 2024). Typically, an encoder transforms the source image into a compact latent representation that conditions the reverse diffusion process, enabling reconstructions with high perceptual quality. This paradigm is demonstrated by foundational methods like CDC (Yang and Mandt 2023), which conditions on a learned latent. More sophisticated strategies, *e.g.*, DiffC (Vonderfecht and Liu 2025), use reverse-channel coding to steer a pre-trained diffusion model without fine-tuning.

Moreover, diffusion-based compression models employ various guidance signals to enhance reconstruction quality. For example, some approaches compress images into a purely semantic space (Lei et al. 2023a; Bachard, Bordin, and Maugey 2024; Pan, Zhou, and Tian 2022). For instance, Pan et al. (Pan, Zhou, and Tian 2022) encode an image into a textual embedding that subsequently guides a pre-trained text-to-image model. Other works (Careil et al. 2023; Guo et al. 2025; Li et al. 2024) utilize more sophisticated conditioning. For example, PerCo applies both pre-extracted text prompts for global context and quantized visual features for local details. In contrast, DiffEIC derives its guidance internally, extracting a global context vector from the hyperprior and injecting it into the diffusion process.

However, these methods share two primary limitations: **(1)** the substantial latency from their multi-step diffusion process, and **(2)** the tendency to sacrifice fidelity for perceptual realism due to the diffusion prior.

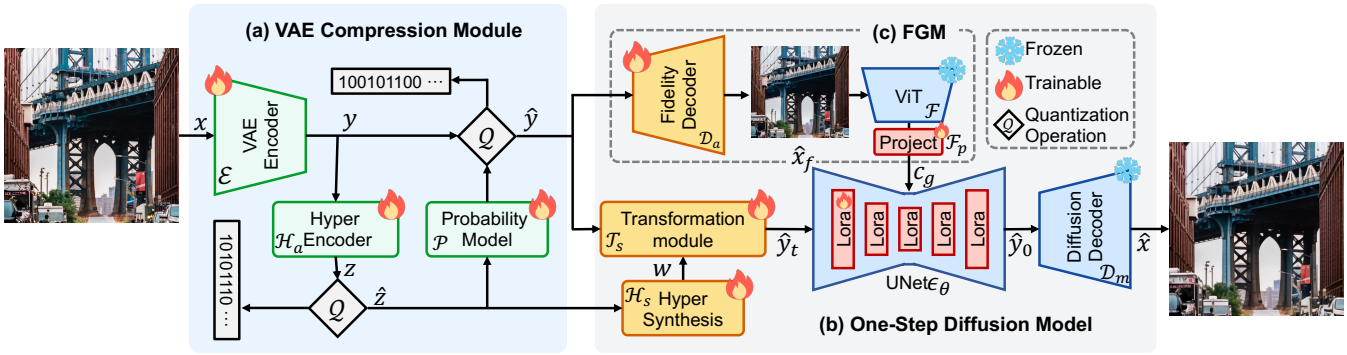

Figure 2: Overview of SODEC. (a) VAE compression module: A pre-trained VAE-based compression model is used to generate the informative latent representation. (b) One-step diffusion model: The latent is mapped to the diffusion space via the transformation module, followed by single-step denoising to produce the reconstructed output. (c) Fidelity guidance module (FGM): A high-fidelity preliminary reconstruction is generated using the VAE-based compression model. Then, the pre-trained ViT is used to extract visual features as the guidance for the diffusion model.

Methodology

In this section, we provide an overview of our proposed model, SODEC, as illustrated in Fig. 2. The section begins with the VAE Compression Module. Subsequently, we elaborate on the core component of SODEC: the one-step diffusion model and the fidelity guidance module. Finally, we detail our rate annealing training strategy.

SODEC Overview

The overview of SODEC is illustrated in Fig. 2. The framework begins with a VAE Compression Module that downsamples a raw image $x \in \mathcal{R}^{(H\times W\times 3)}$ for 16 times to a compact latent representation $y \in \mathcal{R}^{(H/16\times W/16\times C)}$, where C is the latent channels (usually 220). After the entropy coding, restored $\hat{y}$ and $\hat{z}$ are passed into a transformation module $\mathcal{T}_s$ and converted into a content variable $\hat{y}_t \in \mathcal{R}^{(64\times 64\times 4)}$ suitable for diffusion process. Then, we apply the one-step diffusion model to speed up the denoising time compared to the previous multi-step diffusion model (Careil et al. 2023; Li et al. 2024). The one-step diffusion model will then be used to generate the denoised content variable $\hat{y}_0$.

Simultaneously, we utilize a pre-trained fidelity-rich decoder $\mathcal{D}_a$ and further fine-tune it for high fidelity. To achieve this goal, we introduce an alignment loss $\mathcal{L}_{align}$ that consists of pixel-wise loss that constrains $\mathcal{D}_a$ to consistently produce high-fidelity images. After $\mathcal{D}_a$ decodes the latent representation $\hat{y}$ into the raw image $\hat{x}_f$, we use the pre-trained ViT model (Liu et al. 2021) to capture the high-fidelity feature information. Then we linearly project it into the embedding space, getting the condition guidance c_g.

To achieve the best performance, we also introduce the rate annealing training strategy. This strategy first pretrains a complete VAE model with a higher bitrate than our final aim. This VAE model comprises a rich representation in the latent space. Then, we lift the rate penalty by applying a larger trade-off parameter λ in the loss function. Thus, the model can "distill" from the rich representation and selectively discard non-essential information. This strategy is proven to achieve better performance than directly training.

VAE Compression Module

The proposed SODEC employs a VAE-based compression backbone to efficiently encode the input image into a bitstream. This module is comprised of the encoder $\mathcal{E}$, hyperencoder $\mathcal{H}_a$, and probability model $\mathcal{P}$.

Given an input image x, the encoder $\mathcal{E}$ produces a compact latent representation $y{=}\mathcal{E}(x)$. Hyperencoder $\mathcal{H}_a$ then extracts the hyper-latent $z{=}\mathcal{H}_a(y)$. Next, both of these representations y and z are quantized into $\hat{y}{=}\mathcal{Q}(y)$, $\hat{z}{=}\mathcal{Q}(z)$, where $\mathcal{Q}(\cdot)$ represents the quantization operation. Finally, the learned probability model $\mathcal{P}$ conditions on the quantized hyperprior $\hat{z}$ to predict the parameters (μ, σ) of a Gaussian distribution, which models the probability of latent representation $\hat{y}$ for efficient entropy coding.

For the compression model, we pre-train HiFiC (Mentzer et al. 2020) and use its learned weights to initialize our compression backbone $\mathcal{E}$, $\mathcal{H}_a$, and $\mathcal{P}$. In addition, we use the pre-trained VAE decoder to initialize the decoder $\mathcal{D}_a$ in the fidelity guidance module. We apply $\mathcal{D}_a$ to generate the high-fidelity preliminary reconstruction $\hat{x}_f$:

$$\hat{x}_f = \mathcal{D}_a(\mathcal{Q}(\mathcal{E}(x))). \tag{1}$$

This model is optimized using the rate-distortion function:

$$\mathcal{L}_{EG} = \mathbb{E}_{x\sim p_x}\left[\lambda \cdot r(\hat{y}, \hat{z}) + d(x, \hat{x}_f)\right], \tag{2}$$

where $r(\cdot)$ denotes the rate and λ is the hyperparameter to control rate penalty, and $d(x, \hat{x}_f)$ represents the distortion:

$$d(x, \hat{x}_f) = k_M \cdot \mathrm{MSE}(x, \hat{x}_f) + k_P \cdot d_P(x, \hat{x}_f), \tag{3}$$

where k_M and k_P are hyperparameters. We choose LPIPS for the "perception distortion" d_p (in all the subsequent training, we also choose LPIPS by default).

It is worth noting that, in this pre-training stage, we adopt a smaller λ (*i.e.*smaller rate penalty) to train a stronger VAE encoder-decoder pair with higher bitrates. This is beneficial for our subsequent training. More details will be shown in the "Rate Annealing Training Strategy" section.

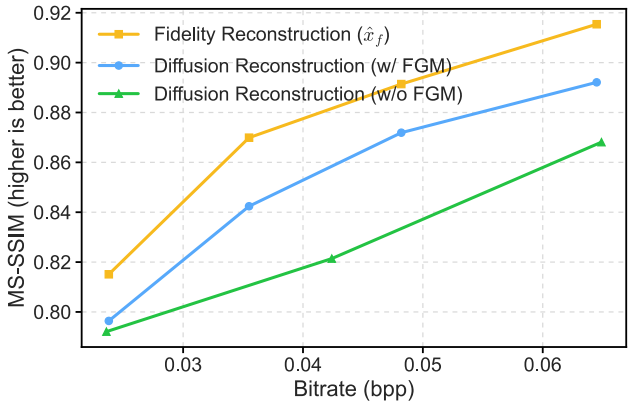

Figure 3: Fidelity comparison (*i.e.*, MS-SSIM) on DIV2K-Val. We compare MS-SSIM (with GT) under different bitrates for the fidelity reconstruction and the diffusion outputs with (w/) and without (w/o) the fidelity guidance module (FGM). The use of FGM improves reconstruction fidelity.

One-Step Diffusion Model

Given $\hat{y}$ and $\hat{z}$ from the bitstream, we propose a transformation module to convert them into a content variable $\hat{y}_t$, which is suitable for diffusion denoising.

First, we use a hyper synthesis network $\mathcal{H}_s$ to extract global information w from the hyperprior $\hat{z}$, where $w{=}\mathcal{H}_s(\hat{z})$. Then, we then merge w and $\hat{y}$ and convert them into content variables $\hat{y}_t{=}\mathcal{T}_s(\hat{y}, w)$, where $\mathcal{T}_s$ denotes the transformation module. Here, $\hat{y}_t$ is conceptually analogous to a noisy latent at the timestep t from the forward process:

$$\hat{y}_t = \sqrt{\bar{\alpha}_t}\hat{y}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{4}$$

For standard diffusion models, they perform a multi-step diffusion process to predict a clear version of a noisy latent. However, these processes are extremely slow and are the most time-consuming steps during the image reconstruction process. Thus, to speed up the diffusion process, we introduce the one-step diffusion model, based on Stable Diffusion 2.1 (Rombach et al. 2022). In this diffusion model, a noise estimator with a UNet architecture ϵ_θ is used to predict the clear, denoised version of the content variable $\hat{y}_0$:

$$\hat{y}_0 = \frac{\hat{y}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\hat{y}_t, t, c_g)}{\sqrt{\bar{\alpha}_t}}, \tag{5}$$

where c_g is the condition guidance. We describe the details of c_g in the next section. Finally, a pre-trained diffusion decoder $\mathcal{D}_m$ reconstructs the output image $\hat{x}$ from the denoised content variable $\hat{y}_0$, where $\hat{x}{=}\mathcal{D}_m(\hat{y}_0)$. In SODEC, we set the timestep t as 999. Meanwhile, to adapt the diffusion model to image compression tasks, we adopt LoRA (Hu et al. 2022) to fine-tune the diffusion model.

Fidelity Guidance Module

The powerful generative prior of diffusion models enables the synthesis of high-perceptual-quality images. However, it often comes at the cost of reconstruction fidelity. To address this limitation, we propose the fidelity guidance module that injects high-fidelity information into the diffusion process.

As shown in Fig. 3, the preliminary reconstruction $\hat{x}_f$ is highly faithful to the original image, although it may lack perceptual richness. Conversely, while the diffusion model excels at synthesizing realistic textures, it lacks explicit knowledge of the source image. Therefore, we can apply the high-fidelity reconstruction $\hat{x}_f$ as a strong conditional guide, to steer the diffusion generator to reconstruct details that are plausible and consistent with the original content. Thus, we can achieve both good fidelity and perception results.

Specifically, the module first utilizes a pre-trained fidelity-rich decoder, $\mathcal{D}_a$, to generate the high-fidelity preliminary reconstruction $\hat{x}_f$ from the compressed latent $\hat{y}$:

$$\hat{x}_f = \mathcal{D}_a(\hat{y}), \tag{6}$$

where $\mathcal{D}_a$ comes from the pre-trained HiFiC encoder-decoder pair, as shown in Eq. (2).

Subsequently, a pre-trained ViT Transformer (Dosovitskiy et al. 2021), denoted as the feature extractor $\mathcal{F}$, is employed to capture deep visual features from this intermediate image. These features are then mapped into the conditioning space of the diffusion model by a projection network $\mathcal{F}_p$ to produce the final guidance condition c_g:

$$c_g = \mathcal{F}_p(\mathcal{F}(\hat{x}_f)), \tag{7}$$

where the resulting condition $c_g{\in}\mathcal{R}^{L\times D}$ consists of a sequence of L embedding vectors of dimension D. In our model, L and D are chosen as 77 and 1024.

This high-fidelity guidance c_g, which encapsulates rich high-fidelity structural information from the source, is then injected into the diffusion denoising model ϵ_θ through cross-attention to steer the diffusion process.

Table 2 in the ablation study demonstrates that this guidance mechanism can effectively steer the generative process, ensuring the final output is both perceptually realistic and highly faithful to the original content.

Rate Annealing Training Strategy

We propose a three-stage training strategy for our SODEC, illustrated in Fig. 2. This idea is based on the motivation that selecting from a rich representation and discarding non-essential information is easier than recreating detailed information. Thus, we decide to first train a VAE model with higher bitrates and then increase the rate penalty to force the model to discard and choose the most useful information.

Stage 1: High-Bitrate VAE Pre-training. Our strategy begins by pre-training HiFiC (Mentzer et al. 2020) model, which serves as the core compression component. In this stage, the model is trained end-to-end on the rate-distortion function as shown in Eq. (2), *i.e.*, $\mathcal{L}_{EG} = \mathbb{E}_{x\sim p_x}[\lambda \cdot r(y) + d(x, \hat{x}_f)]$. We intentionally use a small value for the Lagrange multiplier λ to place a lower penalty on the bitrate. This encourages the model to learn a rich and comprehensive latent representation by prioritizing high-fidelity reconstructions. This pre-training phase is conducted on the HiFiC. After this stage, we obtain the high-bitrate version of networks $\mathcal{E}$, $\mathcal{H}_a$, $\mathcal{P}$, and $\mathcal{D}_a$.

Stage 2: Diffusion Path Warm-up. In the second stage, we transfer the learned weights of the VAE components $(\mathcal{E}, \mathcal{H}_a, \mathcal{P}, \mathcal{D}_a)$ into our SODEC architecture, as shown in Fig. 2. The entire VAE encoding module $(\mathcal{E}, \mathcal{H}_a, \mathcal{P})$ is frozen. The gradient flow is shown as follows:

$$\begin{aligned} \hat{x}_f &= \mathcal{D}_a(\text{sg}(\mathcal{Q}(\mathcal{E}(x)))), \\ w &= \mathcal{H}_s(\text{sg}(\hat{z})), \end{aligned} \tag{8}$$

where "sg" denotes the stop-gradient operation, which cuts off the backpropagation of the gradient for this path.

Training is focused exclusively on the diffusion-based generator and path. Specifically, we freeze the well pre-trained model ViT and diffusion decoder $\mathcal{D}_m$ and fine-tune the UNet in diffusion using LoRA. Moreover, we train the following networks with full parameter updating: hyper synthesis network $\mathcal{H}_s$, transformation module $\mathcal{T}_s$, fidelity guidance decoder $\mathcal{D}_a$, and linear projection network $\mathcal{F}_p$. The optimization objective for this stage only includes a distortion loss between the output $\hat{x}$ and the original image x:

$$\mathcal{L} = \mathbb{E}_{x \sim p_x} \left[d(x, \hat{x}) \right], \tag{9}$$

where $d(\cdot)$ is the same as Eq. (3). Particularly, we do not apply a rate penalty nor an alignment loss $\mathcal{L}_{align}$, because the VAE module is frozen and the latent representation $\hat{y}$ is not distorted. This training stage aims to teach the one-step diffusion generator to effectively map the fixed latent representations to high-quality reconstructions.

Stage 3: Joint Training with Rate Annealing. In this stage, we perform end-to-end optimization of the entire framework. The pre-trained ViT and the final VAE decoder $\mathcal{D}_m$ remain frozen, while all other networks are trained with full parameters, except the U-Net, which continues to be fine-tuned via LoRA. As the VAE encoder is updated, the latent representation $\hat{y}$ can become distorted. To ensure that the fidelity decoder $\mathcal{D}_a$ continues to produce high-fidelity reconstructions, we introduce an alignment loss, $\mathcal{L}_{align}$. From experiments, we find the MSE loss to be most effective:

$$\mathcal{L}_{align} = \mathbb{E}\left[\|x - \hat{x}_f\|_2^2 \right], \quad \text{where} \quad \hat{x}_f = \mathcal{D}_a(\hat{y}). \tag{10}$$

The experimental details of $\mathcal{L}_{align}$ are provided in the ablation. Then, the training objective becomes:

$$\mathcal{L}_{\text{overall}} = d(x, \hat{x}) + \lambda \cdot r(\hat{y}, \hat{z}) + \alpha \cdot \mathcal{L}_{align}. \tag{11}$$

This objective is to fully leverage the generative power of the diffusion model under the guidance of fidelity-rich features to achieve an optimal rate-distortion-perception trade-off.

Finally, the model is fine-tuned with a GAN-based objective $\mathcal{L}_g$ to enhance the synthesis of rich details while maintaining fidelity. Therefore, the overall loss for this final fine-tuning stage can be written as:

$$\mathcal{L}_{\text{finetune}} = d(x, \hat{x}) + \lambda \cdot r(\hat{y}, \hat{z}) + \alpha \cdot \mathcal{L}_{align} + \beta \cdot \mathcal{L}_g, \tag{12}$$

where the hyperparameter β is used to control the penalty of the GAN loss. Detailed training hyperparameter settings are provided in the implementation details of the main paper and the supplementary material.

Experiments

Experimental Settings

Datasets. Our SODEC model is trained using random 512×512 patches extracted from the LSDIR dataset. To evaluate performance, we benchmark SODEC on three standard datasets: Kodak (Eastman Kodak Company 1999), DIV2K Validation dataset (denoted as DIV2K-Val), and CLIC2020 test set (Toderici et al. 2020). We center-crop all images in the validation datasets to 512×512 resolution to facilitate a consistent and fair comparison.

Metrics. Finally, the compression rate is measured in bits per pixel (bpp). For reconstruction fidelity, we report the PSNR and the MS-SSIM (Wang, Simoncelli, and Bovik 2003). To measure perceptual similarity to the ground truth, we employ LPIPS (Zhang et al. 2018) and DISTS (Ding et al. 2020). Furthermore, to evaluate the realism of the generated images in a reference-free setting, we adopt the no-reference metrics NIQE (Mittal, Soundararajan, and Bovik 2012) and CLIPIQA (Wang, Chan, and Loy 2023). The compression rate is measured in bits per pixel (bpp).

Implementation Details. We choose the HiFiC model (Mentzer et al. 2020) without a discriminator as the VAE compression module. We utilize Stable Diffusion 2.1 (Rombach et al. 2022) and set the timestep t as 999 to perform one-step diffusion. We set the batch size to 2 and use the AdamW optimizer with β_1=0.9 and β_2=0.999. We conduct our experiments on 2 NVIDIA RTX A6000 GPUs. More settings are provided in the supplementary material.

Main Results

We conduct extensive experiments to validate the effectiveness of our one-step diffusion model, SODEC, in the ultra-low bitrate regime. To provide a comprehensive analysis, we benchmark our method against several state-of-the-art generative compression models, covering dominant VAE-based, generative tokenizer paradigms, and multi-step diffusion. Specifically, we compare against MS-ILLM (Muckley et al. 2023) and HiFiC (Mentzer et al. 2020), which are leading VAE-based methods. For multi-step diffusion approaches, we compare with CDC (Yang and Mandt 2023). We also include the current diffusion-based models: PerCo (Körber et al. 2024) and DiffEIC (Li et al. 2024).

Quantitative Evaluation. As shown in the visualized results in Fig. 4, our proposed SODEC establishes a new state-of-the-art across all evaluated metrics. Our model achieves superior perceptual quality, outperforming other diffusion-based compression models like PerCo (Körber et al. 2024) and DiffEIC (Li et al. 2024). Moreover, our SODEC also excels in reconstruction fidelity (*e.g.*, MS-SSIM).

Qualitative Evaluation. We present visual comparisons on three datasets in Fig. 5. SODEC achieves reconstructions closer to the original images. In contrast, existing methods often suffer from missing details or content inconsistencies under extreme compression. More visual comparisons are provided in the supplementary material.

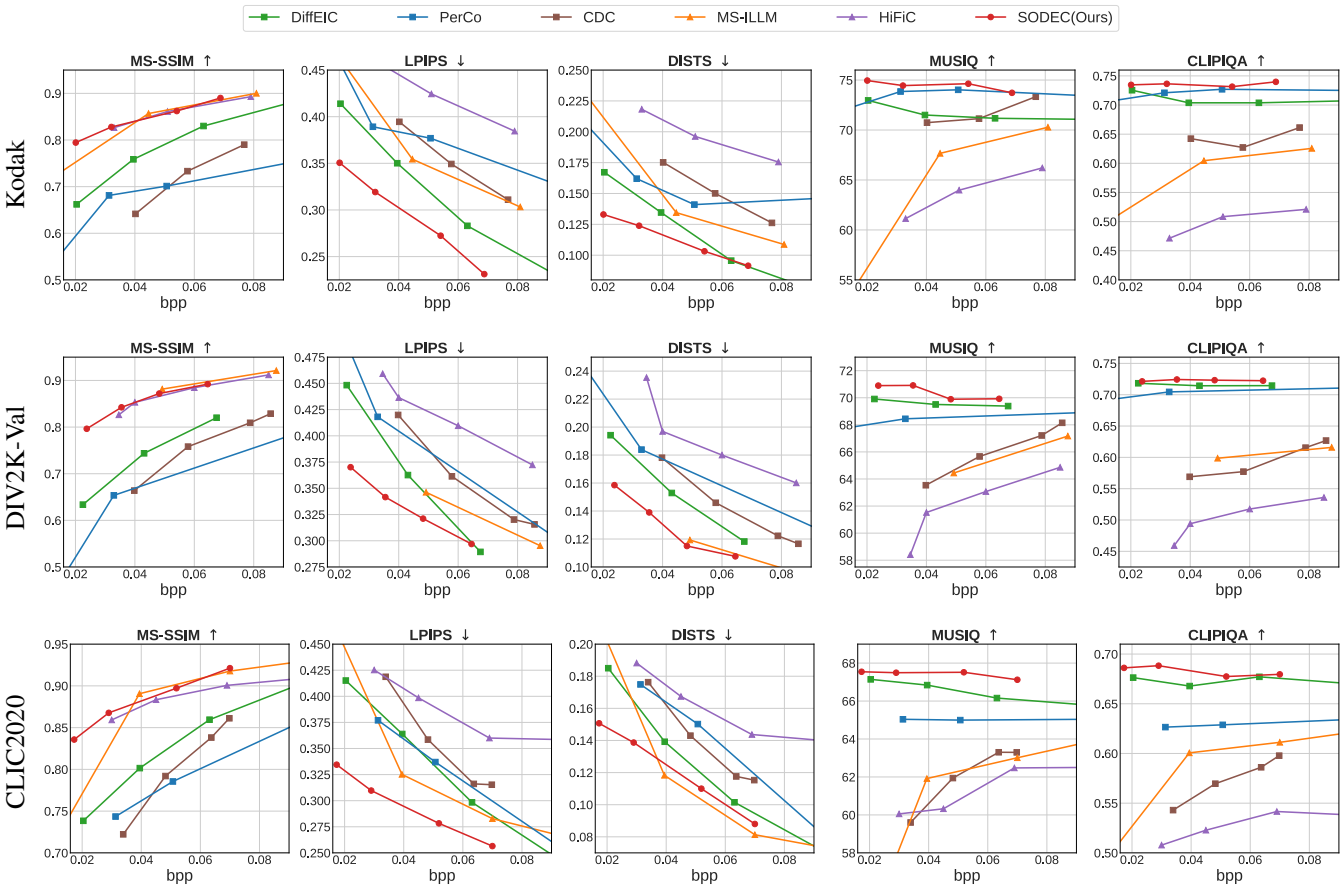

Figure 4: Quantitative comparison with state-of-the-art methods on the Kodak, DIV2K-Val, and CLIC2020 datasets.

Model	Total Time (ms)	Enc. Time (ms) ↓	Dec. Time (ms) ↓	bpp ↓
HiFiC	9.3	5.4	3.9	0.0310
MS-ILLM	9.3	54.5	84.4	0.0395
PerCo	6,242.2	1,540.0	4,702.2	0.0313
DiffEIC	7,827.5	266.4	7,561.1	0.0391
SODEC	232.9	5.0	227.9	0.0314

Table 1: Inference efficiency comparison on the DIV2K-Val dataset. Total, encoding, and decoding times are measured on one A6000 GPU with the 512×512 image.

Inference Efficiency. Moreover, we compare the inference time in Tab. 1. The runtime is tested on one A6000 CPU with the 512×512 image. Our single-step diffusion model, SODEC, offers a substantial advantage in latency. Compared to the multi-step diffusion-based method, PerCo (Körber et al. 2024), our SODEC is 26× faster.

Ablation Study

We conduct our ablation study on LSDIR (train) and DIV2K-Val (test). By default, the models are trained for 50K steps in the pre-training process, and 40K steps in the SODEC end-to-end training process for a fair comparison.

Guidance Strategy	MS-SSIM ↑	LPIPS ↓	bpp ↓
(i) No Guidance	0.8212	0.3625	0.0424
(ii) Text Prompt Guidance	0.8185	0.3631	0.0412
(iii) Hyperprior Guidance	0.8258	0.3527	0.0385
(iv) Aux. Fidelity Guidance (ours)	0.8481	0.3351	0.0368

Table 2: Ablation on the fidelity guidance module.

Fidelity Guidance Module. We conduct an ablation study to validate the effectiveness of our proposed fidelity guidance module. We compare three settings: **(i)** no explicit guidance; **(ii)** text prompt guidance (used by PerCo); **(iii)** semantic features guidance extracted from the hyperprior (used by DiffEIC); and **(iv)** our fidelity guidance module.

As shown in Tab. 2, the baseline model without guidance has poor performance. While using text prompts (case ii) or guidance from the hyperprior (case iii) yields some gains, its impact on reconstruction fidelity is limited. In contrast, our proposed fidelity guidance module leads to a substantial improvement in reconstruction accuracy. Crucially, this significant gain in fidelity is achieved with almost no degradation in perceptual quality as measured by LPIPS. This demonstrates that our guidance mechanism achieves a superior balance between realism and fidelity.

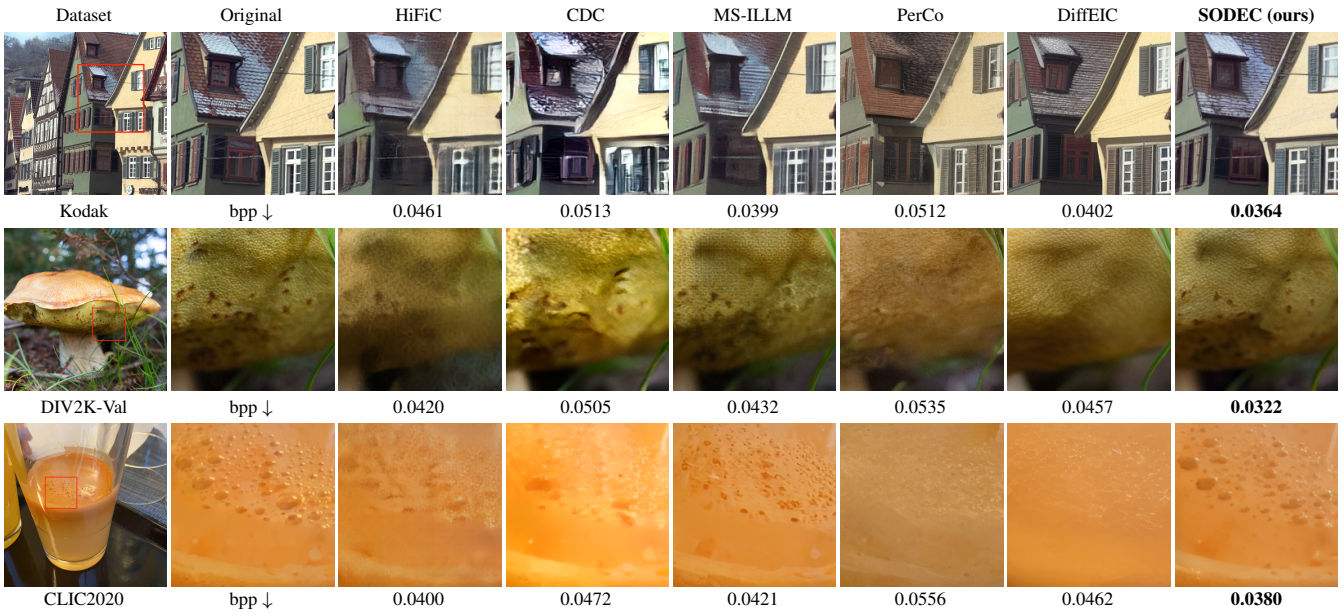

Figure 5: Qualitative comparison with state-of-the-art methods on the Kodak, DIV2K-Val, and CLIC2020 datasets.

Alignment Loss Config.	MS-SSIM ↑	LPIPS ↓	bpp ↓
(i) No Alignment Loss	0.7490	0.4210	0.0203
(ii) MSE + LPIPS	0.7481	0.3961	0.0199
(iii) Merged into Main Loss	0.7984	0.4023	0.0232
(iv) MSE only (ours)	0.7948	0.3827	0.0227

Table 3: Ablation on the setting of alignment loss ($\mathcal{L}_{align}$).

Training Strategy	MS-SSIM ↑	LPIPS ↓	bpp ↓
(i) Frozen VAE Module	0.8512	0.3761	0.0695
(ii) Joint Training (Matched bpp)	0.8621	0.3750	0.0678
(iii) Low-to-High bpp Curriculum	0.8643	0.3451	0.0593
(iv) Rate Annealing (ours)	0.8951	0.3113	0.0604

Table 4: Ablation study on different training strategies.

Alignment Loss. To ensure the preliminary reconstruction $\hat{x}_f$ remains high-fidelity even as the latent representation $\hat{y}$ gets distorted during fine-tuning, we introduce an alignment loss $\mathcal{L}_align$ to constrain it. We investigate four distinct formulations for this fidelity-preservation mechanism: **(i)** no alignment loss, where decoder $\mathcal{D}_a$ receives no direct gradient supervision; **(ii)** a composite loss of perceptual (LPIPS) and distortion (MSE); **(iii)** no separate $\mathcal{L}_{align}$ term ($\mathcal{L}_{align}$=0); and **(iv)** a distortion-only (MSE) loss.

As summarized in Tab. 3, our results validate the need for an explicit alignment loss, as its omission (case i) significantly degrades performance. While a composite loss (case ii) provides no significant improvement in fidelity, merging the constraint into the main loss (case iii) enhances fidelity but at the expense of perceptual quality. In contrast, a dedicated, distortion-only alignment loss (case iv) substantially boosts fidelity over the composite loss (case ii) with a negligible impact on perception compared with case (ii).

Rate Annealing Training Strategy. To validate the efficacy of our proposed rate annealing training strategy, we conduct a comparative analysis of four distinct training schemes: **(i)** training with the entire VAE compression module frozen, thereby excluding it from the optimization process; **(ii)** apply a joint training approach, but we manually tune the Lagrange multiplier λ to ensure the final bitrate is close to the original values; **(iii)** a low-to-high bpp curriculum, where the rate penalty is progressively relaxed; and **(iv)** our proposed high-to-low bpp Rate Annealing strategy.

The results are presented in Tab. 4. It is evident that our rate annealing training strategy significantly outperforms all other training schemes. For a given reconstruction quality, our method achieves an average bitrate saving of over 30%. Conversely, at an equivalent bitrate, our proposed method provides substantially better reconstruction quality. This demonstrates the effectiveness of our approach, which allows the model to first learn a rich feature representation in a less constrained, high-bitrate regime before distilling it into a more efficient, low-bitrate representation.

Conclusion

In this paper, we address the challenges of high latency and poor fidelity in existing diffusion-based compression models. We propose SODEC, a novel model that demonstrates the effectiveness of single-step diffusion for image compression. We introduce the fidelity guidance module to improve reconstruction fidelity. The module provides explicit structural guidance through high-fidelity preliminary reconstruction. Furthermore, we introduce the rate annealing training strategy that enables effective optimization at extremely low bitrates. Extensive experiments demonstrate that our SODEC achieves excellent rate-distortion-perception performance. Compared with multi-step diffusion approaches, SODEC offers more than 20× decoding speedup.

Acknowledgments

This work was supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and the Fundamental Research Funds for the Central Universities.

References

Agustsson, E.; Minnen, D.; Toderici, G.; and Mentzer, F. 2023. Multi-realism image compression with a conditional generator. In *CVPR*.

Agustsson, E.; Tschannen, M.; Mentzer, F.; Timofte, R.; and Gool, L. V. 2019. Generative adversarial networks for extreme learned image compression. In *ICCV*.

Bachard, T.; Bordin, T.; and Maugey, T. 2024. CoCliCo: Extremely low bitrate image compression based on CLIP semantic and tiny color map. In *PCS*.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*.

Blau, Y.; and Michaeli, T. 2019. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *ICML*.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021. Overview of the versatile video coding (VVC) standard and its applications. *TCSVT*.

Careil, M.; Muckley, M. J.; Verbeek, J.; and Lathuilière, S. 2023. Towards image compression with perfect realism at ultra-low bitrates. In *ICLRns*.

Careil, M.; Muckley, M. J.; Verbeek, J.; and Lathuilière, S. 2024. Towards Image Compression with Perfect Realism at Ultra-Low Bitrates. In *ICLR*.

Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *CVPR*.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *TPAMI*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Eastman Kodak Company. 1999. Kodak Lossless True Color Image Suite. http://r0k.us/graphics/kodak/. Accessed: 2024-05-28.

Ghouse, N. F.; Petersen, J.; Wiggers, A.; Xu, T.; and Sautière, G. 2023. A Residual Diffusion Model for High Perceptual Quality Codec Augmentation. *arXiv preprint arXiv:2301.05489*.

Guo, J.; Ji, Y.; Chen, Z.; Liu, K.; Liu, M.; Rao, W.; Li, W.; Guo, Y.; and Zhang, Y. 2025. OSCAR: One-Step Diffusion Codec Across Multiple Bit-rates. *arXiv preprint arXiv:2505.16091*.

He, D.; Yang, Z.; Peng, W.; Ma, R.; Qin, H.; and Wang, Y. 2022a. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *CVPR*.

He, D.; Yang, Z.; Yu, H.; Xu, T.; Luo, J.; Chen, Y.; Gao, C.; Shi, X.; Qin, H.; and Wang, Y. 2022b. Po-elic: Perception-oriented efficient learned image coding. In *CVPR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*.

Körber, N.; Kromer, E.; Siebert, A.; Hauke, S.; Mueller-Gritschneder, D.; and Schuller, B. 2024. Perco (sd): Open perceptual compression. *arXiv preprint arXiv:2409.20255*.

Lei, E.; Uslu, Y. B.; Hassani, H.; and Bidokhti, S. S. 2023a. Text+ sketch: Image compression at ultra low rates. In *ICMLW*.

Lei, E.; Uslu, Y. B.; Hassani, H.; and Saeedi Bidokhti, S. 2023b. Text + Sketch: Image Compression at Ultra Low Rates. In *ICMLW*.

Li, Z.; Zhou, Y.; Wei, H.; Ge, C.; and Jiang, J. 2024. Towards extreme image compression with latent feature guidance and diffusion prior. *TCSVT*.

Liu, L.; Zhou, Y.; Liu, Y.; Ma, S.; and Gao, W. 2024. Extreme Generative Image Compression by Learning Text Embedding from Diffusion Models. In *CVPR*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

Mentzer, F.; Toderici, G. D.; Tschannen, M.; and Agustsson, E. 2020. High-fidelity generative image compression. In *NeurIPS*.

Minnen, D.; Ballé, J.; and Toderici, G. D. 2018. Joint autoregressive and hierarchical priors for learned image compression. In *NeurIPS*.

Minnen, D.; and Singh, S. 2020. Channel-wise autoregressive entropy models for learned image compression. In *ICIP*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *SPL*.

Muckley, M. J.; El-Nouby, A.; Ullrich, K.; Jégou, H.; and Verbeek, J. 2023. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *ICML*.

Pan, Z.; Zhou, X.; and Tian, H. 2022. Extreme generative image compression by learning text embedding from diffusion models. *arXiv preprint arXiv:2211.07793*.

Relic, L.; Azevedo, R.; Gross, M.; and Schroers, C. 2024. Lossy image compression with foundation diffusion models. In *ECCV*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2021. Image Super-Resolution via Iterative Refinement. *arXiv preprint arXiv:2104.07636*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Taubman, D. S.; Marcellin, M. W.; and Rabbani, M. 2002. JPEG2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*.

Theis, L.; Salimans, T.; Hoffman, M. D.; and Mentzer, F. 2022. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*.

Toderici, G.; Theis, L.; Johnston, N.; Agustsson, E.; Mentzer, F.; Ballé, J.; Shi, W.; and Timofte, R. 2020. CLIC 2020: Challenge on Learned Image Compression. In *CVPRW*.

Tschannen, M.; Agustsson, E.; and Lucic, M. 2018. Deep generative models for distribution-preserving lossy compression. *NeurIPS*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NeurIPS*.

Vonderfecht, J.; and Liu, F. 2025. Lossy compression with pretrained diffusion models. *arXiv preprint arXiv:2501.09815*.

Wang, D.; Yang, W.; Hu, Y.; and Liu, J. 2022. Neural data-dependent transform for learned image compression. In *CVPR*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *AAAI*.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*.

Xia, Y.; Zhou, Y.; Wang, J.; An, B.; Wang, H.; Wang, Y.; and Chen, B. 2025. DiffPC: Diffusion-based High Perceptual Fidelity Image Compression with Semantic Refinement. In *ICLR*.

Yang, R.; and Mandt, S. 2023. Lossy image compression with conditional diffusion models. In *NeurIPS*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.