Trajectory-Consistent Denoising Diffusion Codebook Models for Zero-Shot High-Fidelity Image Compression at Ultra-Low Bitrates

Fang Zhang and **Linli Xu**
University of Science and Technology of China
fangzhang@mail.ustc.edu.cn, linlixu@ustc.edu.cn

Abstract

Denoising Diffusion Codebook Models (DDCM) have emerged as a promising framework for zero-shot image compression by replacing stochastic sampling with discrete selection from a reproducible Gaussian codebook. By greedily picking noise vectors that best match the target image, DDCM encodes the generative trajectory into a compact sequence of indices for image compression. However, the target-guided noise selection provides only coarse guidance as the discrepancy between the denoiser's output and the target image is substantial. This discrepancy inevitably results in trajectory drift, necessitating massive codebooks to span the enlarged search space and extensive inference steps to gradually compensate for the deviation. In this paper, we propose the Trajectory-Consistent Diffusion Codebook Model (TC-DDCM) to address these inefficiencies. Rather than approximating the distant target, TC-DDCM navigates the generative process by strictly tracking a pre-computed inversion trajectory. This paradigm shift yields two critical advantages: 1) By minimizing the deviation to the proximal reference state along the trajectory, the search space is drastically shrunk, resulting in significantly smaller codebooks. 2) By aligning the generative steps with inversion methods (e.g., $T \approx 15$), TC-DDCM achieves optimal reconstruction in a few inference steps. Extensive experiments demonstrate that TC-DDCM significantly outperforms state-of-the-art zero-shot methods in rate-distortion-perception performance at ultra-low bitrates, making zero-shot diffusion-based compression practical.

1 Introduction

The advent of Denoising Diffusion Probabilistic Models (DDPMs) [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020; Song *et al.*, 2020] has revolutionized generative AI, delivering unprecedented quality in image generation [Liu *et al.*, 2023; Cai *et al.*, 2025], restoration [Raphaeli *et al.*, 2025; Zamir *et al.*, 2022], and editing [Kawar *et al.*, 2023]. Recently, these models have been successfully adapted for image compression. By leveraging extensive training [Yang and Mandt, 2023; Li *et al.*, 2023], fine-tuning [Careil *et al.*, 2024; Xue *et al.*, 2025], or structural modifications [Zhang *et al.*, 2025; Park *et al.*, 2025], recent diffusion-based methods have demonstrated significant advantages over other models such as manually crafted codecs [Bellard, 2018], Generative Adversarial Networks (GANs) [Iwai *et al.*, 2024; Muckley *et al.*, 2023; Mentzer *et al.*, 2020] and Variational Autoencoders (VAEs) [Theis *et al.*, 2017]. However, these approaches lack zero-shot capability on unseen data distributions and require training separate models for specific compression rates. In contrast, the Denoising Diffusion Codebook Model (DDCM) [Ohayon *et al.*, 2025] introduces a novel paradigm enabling flexible bitrate adjustment without additional training, achieving SOTA performance among zero-shot based methods [Xue *et al.*, 2025; Elata *et al.*, 2024]. Specifically, DDCM compresses a target image by sequentially selecting noise vectors that best match the target image from a fixed Gaussian codebook during the generative diffusion process. By recording the indices of these selected vectors, DDCM effectively converts the generative trajectory into a transmittable bitstream, the length of which scales linearly and log linearly with the number of inference steps and codebook size respectively.

However, as illustrated in Fig 1(a), standard DDCM and its variants (e.g., Turbo-DDCM) select noise vectors by minimizing the residual error relative to the target image x_0. Given the substantial distance between noisy intermediate state and the clean target, this target-guided noise selection strategy provides only coarse guidance, inevitably leading to *trajectory drift*—a phenomenon where the generative path gradually diverges from the target. Consequently, DDCM requires massive codebooks to span the enlarged search space and extensive inference steps to correct accumulated deviations, resulting in significant compression inefficiency. We provide a detailed quantitative analysis of this target-guided noise selection in Sec 5.5.

To address these challenges, we propose the **Trajectory-Consistent Diffusion Codebook Model (TC-DDCM)**. Our core insight is to shift the paradigm from approximating the distant target x_0 to **tracking a pre-computed inversion trajectory** $\mathcal{T}_{inv}$. As the inversion trajectory fits in the underlying generative distribution of diffusion models, TC-DDCM can naturally align the generative process with the reference tra-

jectory through noise selection. Specifically, our method operates in two distinct stages: *Inversion Trajectory Construction* and *Trajectory-Guided encoding*.

During the inversion trajectory construction stage, we introduce two variants for this construction: TC-DDCM (DDIM), which utilizes standard deterministic DDIM inversion [Ho *et al.*, 2020], and TC-DDCM (DDPM), which employs CycleDiffusion [Wu and De la Torre, 2023] as DDPM-based inversion. We leave the incorporation of other effective inversion methods for future work. Once the trajectory is established, our trajectory-guided noise selection mechanism minimizes the deviation between the predicted state and the proximal reference point at each step along the pre-computed inversion trajectory. As illustrated in Fig 1(b), this paradigm shift directly leads to two critical advantages:

- **Small Codebook Efficiency via Precise Guidance:** The minimal distance between the current prediction and the proximal reference point on the trajectory provides precise guidance (as shown in Tab 1) that drastically reduces the feasible search space for the noise vector, leading to significantly smaller codebooks.
- **Few-Step Inference via Inversion Alignment:** The inference steps T in TC-DDCM are strictly aligned with the chosen inversion method. This allows our framework to capitalize on efficient inversion techniques that achieve high-fidelity reconstruction performance in very few steps (e.g., $T \approx 20-30$). Consequently, TC-DDCM reduces the number of required denoising steps substantially.

Benefiting from these two advantages, TC-DDCM achieves a substantial reduction in Bits Per Pixel (BPP) while maintaining reconstruction performance comparable to the original DDCM. It is worth noting that the theoretical upper bound of TC-DDCM is determined by the degree of alignment with the inversion trajectory, rather than the reconstruction quality of the inversion itself. As long as the trajectory terminates at the target image and adheres to the generative distribution of DPMs, precise alignment with the trajectory leads to high-fidelity reconstruction.

Extensive experiments demonstrate that TC-DDCM can track inversion trajectories closely even with reduced codebook sizes and fewer inference steps. It surpasses other zero-shot methods in rate-distortion-perception performance especially at ultra-low bitrates (e.g. < 0.05 bpp) and achieves over $50\times$ faster coding latencies, making zero-shot diffusion-based image compression practical.

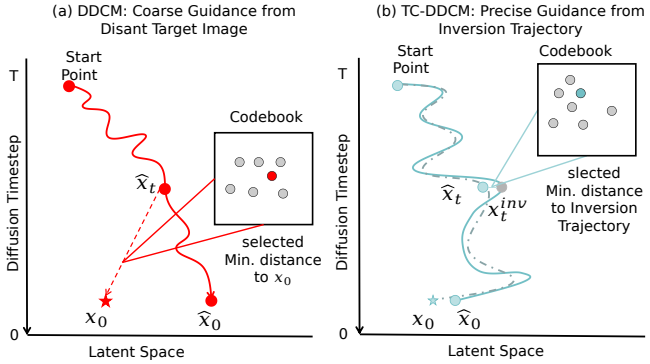

Figure 1: Comparison of noise selection mechanisms. (a) DDCM guides the generative process to jump directly to the distant target image, providing only coarse guidance. (b) TC-DDCM minimizes the distance between the generative process and the pre-computed proximal inversion trajectory, providing precise guidance.

2 Related Work

2.1 Diffusion Based Image Compression

Recently, diffusion-based generative compression has demonstrated impressive reconstruction quality, significantly outperforming traditional codecs in terms of perceptual fidelity. A predominant paradigm employs uniform coding strategies to quantize latent representations of Variational Autoencoders (VAEs) [Cemgil *et al.*, 2020]. Building on this, several methods [Körber *et al.*, 2024; Pan *et al.*, 2022; Lei *et al.*, 2023] leverage existing text-to-image diffusion models as powerful priors, conditioning the generation on both vector-quantized latents and textual descriptions. Furthermore, frameworks such as DiffEIC [Li *et al.*, 2024] and RDEIC [Li *et al.*, 2025] synergize compressive VAEs with pre-trained diffusion backbones to reconstruct photo-realistic images, particularly at low bitrates. However, the substantial computational cost of iterative denoising remains a critical bottleneck. To address this, recent works have attempted to achieve one-step generation [Guo *et al.*, 2025; Zhang *et al.*, 2025; Park *et al.*, 2025]. For instance, StableCodes [Zhang *et al.*, 2025] modifies the encoder architecture, while DiffO [Park *et al.*, 2025] employs a training strategy that models the residual between quantized and original latents.

2.2 Zero-shot Image Compression

Alternatively, zero-shot methods leverage pretrained diffusion models without training or fine-tuning. DiffC [Vonderfecht and Liu, 2025] utilizes Reverse Channel Coding (RCC) to achieve high-fidelity reconstruction. DDCM [Ohayon *et al.*, 2025] compresses images by greedily selecting noise vectors during the generation process to minimize the denoising error at each step. Turbo-DDCM [Vaisman *et al.*, 2025] further extends this by solving a closed-form sparse least squares problem to combine multiple noise vectors. Despite these advancements, all zero-shot methods are strictly limited by either requiring hundreds of denoising steps or excessive codebook sizes.

2.3 Image Inversion

Inversion approaches [Ho *et al.*, 2020] map an input image into one or multiple latent noise vectors that reconstruct the original image during the generative process and hold significant potential for image editing [Wallace *et al.*, 2023; Hu *et al.*, 2022] and super resolution [Yue *et al.*, 2025] tasks. Most existing inversion algorithms rely on DDIM sampling to establish a deterministic trajectory that maps an initial noise to the generated image. However, these methods inevitably introduce approximation errors at each step. These errors accumulate along the sampling chain, leading

to substantial deviations from the target, particularly when Classifier-Free Guidance (CFG) is employed. Several approaches have attempted to mitigate the artifacts induced by CFG by optimizaing prompt embeddings [Dong *et al.*, 2023; Mokady *et al.*, 2023] or incorporating affine coupling layers [Wallace *et al.*, 2023].

A distinct line of research focuses on DDPM-based inversion algorithms, which achieves perfect reconstruction by constructing a sequence of noise maps along the DDPM sampling trajectory. CycleDiffusion [Wu and De la Torre, 2023] constructs the auxiliary images that approximate the underlying generative process to extract noise sequences. The Edit Friendly DDPM inversion [Huberman-Spiegelglas *et al.*, 2024] modifies the construction of auxiliary images, thereby yielding a noise space that remains statistically independent and is amenable to editing.

3 Background

3.1 Denoising Diffusion Models

Denoising Diffusion Models (DDMs) are trained with a denoising objective to predict the noise that progressively corrupts a clean image x_0 into white Gaussian noise.

$$x_t = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\alpha_t = \sqrt{1-\beta_t}$, and $\{\beta_t\}_{t=1}^{T}$ is a monotonically increasing noising schedule and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

The reverse process of DDMs starts from a random noise vector $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it using the noise $\epsilon_\theta(x_t, t)$ predicted by the neural network:

$$x_{t-1} = \mu_\theta(x_t) + \sigma_t z_t, \quad z_t \sim \mathcal{N}(0, \mathbf{I}) \tag{2}$$

$$\mu_\theta(x_t) = \underbrace{\sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0|t}}_{\text{predicted } x_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(x_t,t)}_{\text{direction pointing to } x_t} \tag{3}$$

Here $\hat{x}_{0|t} = \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right)$ is the predicted x_0 given x_t[1]. The variance σ_t is determined by the hyperparameter η:

$$\sigma_t = \eta\beta_t\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \tag{4}$$

where $\eta \in [0, 1]$. $\eta = 1$ corresponds to the original DDPM sampling process, and $\eta = 0$ corresponds to the deterministic DDIM scheme.

3.2 Denoising Diffusion Codebook Models

Denoising Diffusion Codebook Models (DDCM) [Ohayon *et al.*, 2025] build upon the DDPM framework but modify the reverse process by replacing stochastic sampling with discrete selection. Specifically, instead of sampling z_t from a continuous Gaussian distribution, DDCM selects a vector from a fixed codebook $\mathcal{C}_t = [\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(K)}]$ containing K i.i.d. Gaussian vectors:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t\mathcal{C}_t(k_t) \tag{5}$$

[1]This is also referred to as the minimum mean-squared-error (MMSE) estimator.

where $k_t \in \{1, \ldots, K\}$. The process begins with a fixed starting point $x_T = \mathcal{C}_{T+1}(k_{T+1})$ and ends at $t = 1$. To ensure reconstruction fidelity, DDCM employs a greedy selection criterion that pulls the predicted state $\hat{x}_{0|t}$ close to the target. It computes the residual error between $\hat{x}_{0|t}$ and the ground-truth x_0, selecting the entry that maximizes the inner product:

$$\begin{aligned} k_t &= \underset{k\in\{1,\ldots,K\}}{\arg\min} \left\| \underbrace{(\mathcal{C}_t(k) + \hat{x}_{0|t})}_{\text{predicted state}} - \underbrace{x_0}_{\text{target}} \right\|^2 \\ &= \underset{k\in\{1,\ldots,K\}}{\arg\max} \langle \mathcal{C}_t(k),\, x_0 - \hat{x}_{0|t} \rangle \end{aligned} \tag{6}$$

The above equivalence is justified by the Johnson-Lindenstrauss Lemma [Dasgupta and Gupta, 2003] in high-dimensional spaces. By recording the indices $(k_t)_{t=2}^{T}$, DDCM converts the generative trajectory into a bitstream. During decoding, these indices retrieve the noise vectors to reconstruct the image via Eq 5. To further reduce quantization error, DDCM employs Matching Pursuit (MP) [Mallat and Zhang, 1993] to approximate the noise as a linear combination of M vectors. The resulting bitrate (BPP) is:

$$\text{BPP}_{\text{DDCM}} = \frac{(T-1)\left(\lceil\log_2(K)\rceil M + C(M-1)\right)}{\text{Pixels}} \tag{7}$$

3.3 Image Inversion

DDIM Inversion. DDIM inversion ($\eta = 0$) exploits the ODE consistency that can reverse the deterministic generation process from the initial noise x_T. By rearranging the sampling update, x_t can be derived from x_{t-1} as:

$$\begin{aligned} x_t &= \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}x_{t-1} + \sqrt{\bar{\alpha}_t}\left(\sqrt{\frac{1}{\bar{\alpha}_t}-1} - \sqrt{\frac{1}{\bar{\alpha}_{t-1}}-1}\right)\epsilon_\theta(x_t, t) \\ &\approx \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}x_{t-1} + \sqrt{\bar{\alpha}_t}\left(\sqrt{\frac{1}{\bar{\alpha}_t}-1} - \sqrt{\frac{1}{\bar{\alpha}_{t-1}}-1}\right)\epsilon_\theta(x_{t-1}, t-1) \end{aligned} \tag{8}$$

A critical limitation highlighted in Eq 8 is the dependency on $\epsilon_\theta(x_t, t)$, which is unknown at step $t-1$. DDIM inversion approximates this using $\epsilon_\theta(x_{t-1}, t-1)$. This **local linearization assumption** introduces approximation errors that accumulate across steps, leading to reconstruction degradation,especially in low-step regimes.

DDPM Inversion. DDPM inversion aims to find a specific noise sequence instead of a single noise map that perfectly reconstructs the image within the DDPM sampling process ($\eta = 1$). We adopt CycleDiffusion [Wu and De la Torre, 2023] as our baseline for DDPM inversion. This method constructs an auxiliary trajectory $\mathcal{T}_{ref} = \{x_t\}_{t=0}^{T}$ that adheres to the posterior distribution $q(x_{1:T}|x_0)$. Concretely, for each step $t = T, \ldots, 1$, they isolate the reconstruction noise ϵ_t^{rec} derived from the current x_t and the ground-truth x_0 (Eq 1), and substitute this ϵ_t^{rec} for $\epsilon_\theta(x_t, t)$ in Eq 2 to obtain x_{t-1}. Through this iterative cycle, a DDPM inversion trajectory $\mathcal{T}_{ref} = \{x_t\}_{t=0}^{T}$ is explicitly constructed. The detailed procedure is summarized in [Wu and De la Torre, 2023].

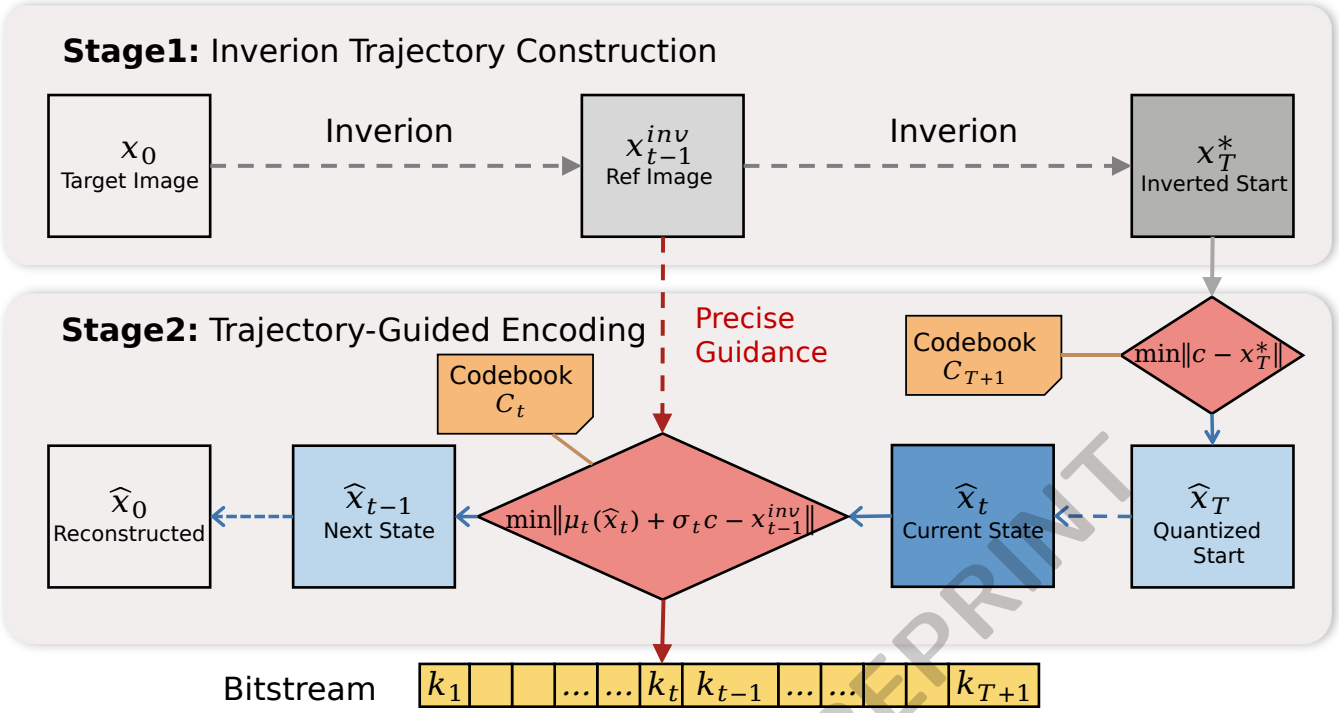

Figure 2: Overview of Trajectory-Consistent Diffusion Codebook Model (TC-DDCM). We first employ inversion methods to construct a reference trajectory from x_0 to x_T^*. During encoding, instead of greedily matching the target, we select codebook indices that minimize the deviation from this pre-computed trajectory.

4 Method

We propose the Trajectory-Consistent Diffusion Codebook Model (TC-DDCM), a unified framework that eliminates trajectory drift by strictly guiding the generative process along a pre-computed inversion trajectory. As illustrated in Figure 2, our approach comprises two phases: *Reference Trajectory Construction* and *Trajectory-Guided Encoding*.

4.1 Phase 1: Reference Trajectory Construction

In this phase, TC-DDCM constructs a deterministic inversion trajectory $\mathcal{T}_{ref}$ that serves as the reference path for the entire encoding process. This trajectory fits in the distribution of generative process in DPMs so that it is natural for TC-DDCM to align the generative process with the reference trajectory. The pre-computed trajectory $\mathcal{T}_{ref}$ provides a deterministic starting point x_T^* and a sequence of intermediate proximal reference points $\{x_t^{inv}\}_{t=T-1}^{1}$ ending at the target image x_0. We implement this phase using two distinct protocols:

- **TC-DDCM (DDIM):** Following DDIM [Song *et al.*, 2021], we obtain a deterministic inversion trajectory where the starting point x_T^* varies for each image.
- **TC-DDCM (DDPM):** We employ CycleDiffusion [Wu and De la Torre, 2023] as our DDPM-based inversion strategy. Crucially, to ensure determinism during construction, we initialize the inversion from a **fixed initial noise state** x_T^* and predefine the random seed at each step, ensuring the inversion trajectory unique yet reproducible for each image. Although DDPM inversion theoretically guarantees perfect reconstruction for arbitrary steps, the practical performance is limited by numerical precision. Empirically, we observe that it achieves near-lossless reconstruction at approximately $T = 10$ steps.

4.2 Phase 2: Trajectory-Guided Encoding

Initialization with Inverted Start

In contrast to the fixed Gaussian initialization used in DDCM, TC-DDCM initializes the generative process directly from the inverted starting point of the reference trajectory, $x_T^* \in \mathcal{T}_{ref}$. To transmit this starting point, we quantize it using the initial codebook $\mathcal{C}_{T+1}$:

$$c_{start} = \arg\min_{c \in \mathcal{C}_{T+1}} \|c - x_T^*\|^2 \tag{9}$$

Trajectory-Guided Noise Selection

The core innovation of TC-DDCM lies in its selection criterion. Recall that standard DDCM (Eq. 6) selects noise to minimize alignment with the distant target x_0. As illustrated in Fig 1, this creates a substantial optimization gap, providing only coarse guidance that exacerbates trajectory drift.

Instead, TC-DDCM employs a trajectory guided noise selection mechanism. At each timestep t, given the current state $\hat{x}_t$, the goal is to select a codebook vector $\mathcal{C}_t(k)$ such that the predicted next state $\hat{x}_{t-1}$ lands as close as possible to the proximal reference state $x_{t-1}^{inv} \in \mathcal{T}_{ref}$.

Mathematically, we substitute the random noise z_t in the sampler with the codebook entry and solve:

$$k_t = \arg\min_{k\in\{1,\dots,K\}} \left\| \underbrace{\mu_t(\hat{x}_t) + \sigma_t \mathcal{C}_t(k)}_{\text{Predicted State } \hat{x}_{t-1}} - \underbrace{x_{t-1}^{inv}}_{\text{Reference}} \right\|^2 \tag{10}$$
$$= \arg\max_{k\in\{1,\dots,K\}} \left\langle \mathcal{C}_t(k),\ \frac{x_{t-1}^{inv} - \mu_t(\hat{x}_t)}{\sigma_t} \right\rangle$$

where η is a hyperparameter controlling the variance schedule σ_t. Concretely, for the DDPM variant, we strictly synchronize η with the value used in the inversion trajectory, while for the DDIM variant, we set $\eta \in (0, 1]$ to modulate the stochasticity of the generation.

By minimizing the distance to the proximal reference point on this inversion trajectory, our optimization objective provides precise guidance. Even when quantization errors occur at step t, this trajectory-guided strategy actively pulls the latent state back towards the inversion path, ensuring the generative process remains strictly consistent with the trajectory that eventually converges at the target.

4.3 Coding Latency and Bitrate Analysis

The coding latency comprises two stages: encoding and decoding. Specifically, the encoding phase incurs a cost of $2T$ function evaluations (NFE), accounting for both the reference trajectory construction (T steps) and the trajectory-consistent encoding (T steps). The decoding phase requires T steps, identical to DDCM. Consequently, the total NFE for the coding process amounts to $T_{total} = 2T + T = 3T$ (compared to $2T$ in DDCM).

Although our method requires encoding the inverted starting point x_T^*, this adds only a single codebook index. Compared to DDCM, this overhead is negligible. The total BPP is calculated as follows:

$$\text{BPP} = \frac{T \cdot (\lceil \log_2(K) \rceil M + C \cdot M)}{\text{Pixel}} \tag{11}$$

5 Experiments

5.1 Implementation Details

We conduct comprehensive evaluations on three standard benchmarks: Kodak24 [Franzen, 1999], CLIC2020 [Toderici *et al.*, 2020], and DIV2K [Agustsson and Timofte, 2017]. For preprocessing, we apply a center crop of 512×512 for Kodak24 and CLIC2020, and 768×768 for DIV2K. We employ FID [Bińkowski *et al.*, 2018] and LPIPS [Zhang *et al.*, 2018] to evaluate perceptual quality, and PSNR for reconstruction fidelity. Following Mentzer et al. [Mentzer *et al.*, 2020], we compute FID on extracted image patches. Both encoding and decoding times are measured on a single NVIDIA A100 GPU.

We benchmark our approach against existing zero-shot diffusion-based methods, including PSC [Elata *et al.*, 2024], DDCM [Ohayon *et al.*, 2025], DiffC [Theis *et al.*, 2022], and Turbo-DDCM [Vaisman *et al.*, 2025]. Additionally, we compare with non-neural, fine-tuning-based, and training-based methods, specifically including BPG [Bellard, 2018], HiFiC [Mentzer *et al.*, 2020], PerCo (SD) [Körber *et al.*, 2024; Careil *et al.*, 2024], ILLM [Muckley *et al.*, 2023], two

Algorithm 1. Trajectory-Consistent DDCM (TC-DDCM)

Input: Image x_0, Codebooks $\{\mathcal{C}_t\}$
Output: Bitstream B

$\mathcal{T}_{ref} \leftarrow \text{Inversion}(x_0)$ {Construct reference trajectory}
$x_T^* \leftarrow \mathcal{T}_{ref}[T]$
$c_{start} \leftarrow \arg\min_{c\in\mathcal{C}_{T+1}} \|c - x_T^*\|^2$
$B \leftarrow \{c_{start}\}$; $\quad \hat{x}_T \leftarrow c_{start}$
for $t = T$ **down to** 1 **do**
  $k_t \leftarrow \arg\min_k \|\mu_t(\hat{x}_t) + \sigma_t \mathcal{C}_t(k) - \mathcal{T}_{ref}[t-1]\|^2$
  $\hat{x}_{t-1} \leftarrow \mu_t(\hat{x}_t) + \sigma_t \mathcal{C}_t(k_t)$
  Append k_t to B
end for
return B

configurations of CRDR [Iwai *et al.*, 2024], and the one-step diffusion method StableCodec [Zhang *et al.*, 2025].

All zero-shot methods utilize the same pre-trained diffusion backbones: the SD 2.1 Base [Rombach *et al.*, 2022] model for 512×512 images and the SD 2.1 model for 768×768 images. The bitrate of TC-DDCM is modulated by varying the inference timesteps T, codebook size K, hyperparameters of matching pursuits M and C. Please refer to the supplementary material for further details.

5.2 Rate-Distortion-Perception Performance

As illustrated in Fig 3, our TC-DDCM significantly outperforms all competing methods across PSNR and FID metrics in the ultra-low bitrate regime. In particular, when compared to DDCM and its variant, Turbo-DDCM, our approach demonstrates superiority and stability across all the metrics with well-preserved fidelity, and reaches extreme bitrates as low as 0.002 bpp. As the bitrate increases, TC-DDCM yields highly competitive results, surpassing the previous state-of-the-art zero-shot method, DiffC. It is worth noting that the DDPM-based variant consistently yields superior results compared to the DDIM-based variant across different bitrates, which may be due to inherent approximation errors in DDIM inversion.

5.3 Coding Latency

We compare encoding and decoding latencies against zero-shot baselines and one-step baseline StableCodec as shown in Fig 3. Our method achieves remarkable acceleration, performing approximately $50\times$ faster than the standard DDCM, and consistently outperforming DiffC in both encoding and decoding speeds across all bitrate settings. Compared to Turbo-DDCM, our method exhibits a slight higher encoding latency due to the computational overhead of constructing the inversion trajectory, yet maintaining comparable decoding speeds. Crucially, this additional encoding investment yields significant gains in Rate-Distortion-Perception performance. Notably, these two approaches are orthogonal: Turbo-DDCM focuses on optimizing the matching pursuit parameters (the number of atoms M and coefficient bits C), whereas TC-DDCM primarily optimizes the inference schedule (T) and codebook size (K). Since these optimization strategies are

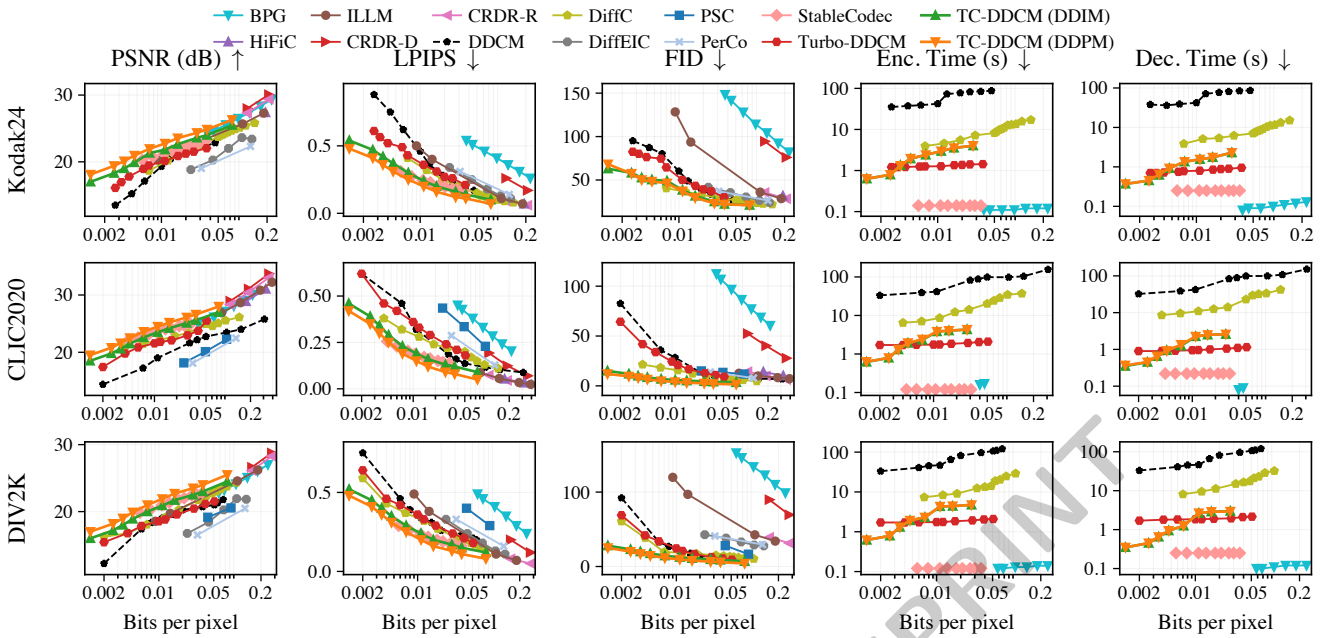

Figure 3: Quantitative evaluation on the Kodak24, CLIC2020, and DIV2K datasets. We compare rate-distortion-perception performance (PSNR, LPIPS, FID) and coding latency across various bitrates. Results for Turbo-DDCM, DDCM, DiffC, and StableCodec are obtained using their official implementations, while other baselines are sourced from the DDCM paper.

Step t	**Optimization Distance** (Target Magnitude, L_2 ↓)	**Alignment Precision (ERR)** **Error Reduction Ratio** $\in [0, 1]$ (↑)		
	DDCM / Ours	$K = 1024$ (**DDCM / Ours**)	$K = 2048$ (**DDCM / Ours**)	$K = 4096$ (**DDCM / Ours**)
$t = 20$	125.24 / **10.51**	0.009 / **0.061**	0.011 / **0.065**	0.014 / **0.067**
$t = 100$	67.88 / **7.45**	0.012 / **0.075**	0.015 / **0.079**	0.021 / **0.080**
$t = 200$	59.51 / **5.38**	0.036 / **0.089**	0.039 / **0.093**	0.047 / **0.093**

Table 1: Quantitative analysis of the optimization distance and alignment precision. We measure the **optimization distance** using L_2 distance (↓) and the **alignment precision** using Error Reduction Ratio (ERR↑).

parallel, combining them to achieve even higher efficiency remains a promising direction for future work.

5.4 Qualitative Comparison

Fig 4 presents the qualitative comparison against state-of-the-art zero-shot methods. Notably, TC-DDCM demonstrates superior capability in generating visually aligned details and realistic textures, even at ultra-low bitrates. As showcased in the first and second rows, our method faithfully reconstructs intricate structures, such as the murals on the wall and the model's teeth. In contrast, all other methods fail to maintain high realism and fidelity under such severe bitrate limitations. Specifically, DDCM yields blurry reconstructions with lost details, while DiffC and Turbo-DDCM suffer from distinct artifacts and structural deformations in complex regions.

5.5 Analysis on Noise Selection Mechanisms

To empirically validate the effectiveness of TC-DDCM, we conduct a comparative analysis between the standard target-guided strategy (DDCM) and our proposed trajectory-guided strategy. We focus on quantifying two critical aspects highlighted in our introduction: the magnitude of the optimization objective and the alignment precision of guidance.

First, we quantify the optimization distance by measuring the L_2 and cosine similarity of the objective: $\|\hat{x}_{0|t} - x_0\|_2$ for DDCM versus $\|\mu_t(\hat{x}_t) - x_{t-1}^{inv}\|_2$ for TC-DDCM. The results are averaged across all samples on the DIV2K dataset over 50 sampling steps. As reported in Tab 1, the optimization distance of DDCM is consistently larger than that of TC-DDCM. This validates our core insight: the distance between the current prediction and the proximal reference is significant smaller compared to the distant target image x_0, as illustrated in Fig 1.

To further evaluate the efficiency of noise selection strategies, we introduce the **Error Reduction Ratio (ERR)**, defined as:

$$\textbf{ERR} = \frac{\left\|\mu_t(\hat{x}_t) - x_{t-1}^{inv}\right\|_2 - \left\|\mu_t(\hat{x}_t) + \sigma_t \mathcal{C}_t(k_t) - x_{t-1}^{inv}\right\|_2}{\left\|\mu_t(\hat{x}_t) - x_{t-1}^{inv}\right\|_2} \tag{12}$$

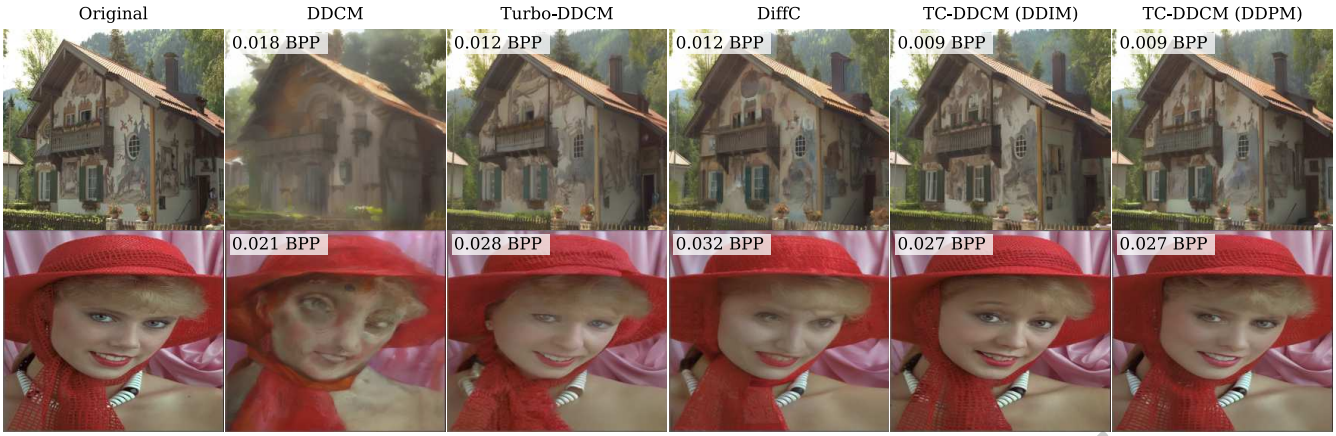

Figure 4: Qualitative results. Visualizations on the Kodak24 dataset (512×512). Our TC-DDCM achieves highly realistic reconstruction at lower BPP compared to state-of-the-art methods.

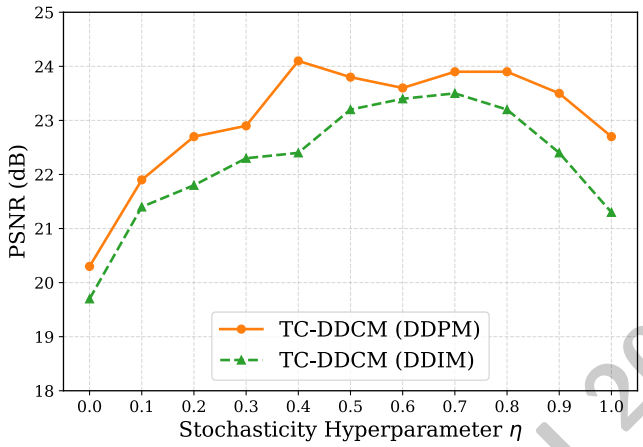

Figure 5: Ablation study of stochasticity hyperparameter η on Kodak24 dataset.

This metric quantifies the **compensatory capacity** of the selected codebook vector. It measures the proportion of the optimization discrepancy that is effectively bridged by the selected noise, serving as a direct indicator of how precisely the generative process is consistent with the target image or inversion trajectory.

As shown in Tab 1, the improvement in ERR, stemming from larger codebooks and extended inference steps, validates our hypothesis that the former expands the search space while the latter corrects cumulative errors, ultimately yielding more precise guidance. However, standard DDCM continues to exhibit persistently lower ERR scores under all settings. This indicates that even with an expanded search space and extended inference steps, the codebook vectors cannot compensate for the vast distance to the remote x_0. The limitation validates that the target-guided strategy is inherently capable of providing only coarse guidance. Conversely, TC-DDCM achieves significantly higher ERR scores. Since the inversion trajectory naturally fits in the underlying generative distribution of DPMs, our method can naturally align the generative process with the reference points through noise selection. This tight alignment proves that our trajectory-guided strategy provides such precise guidance that it alleviates the need for massive codebooks or excessive inference steps, enabling high-fidelity reconstruction even at ultra-low bitrates.

Ablation Study of Hyperparameter η

In diffusion-based generative processes, the hyperparameter $\eta \in [0, 1]$ controls the degree of stochasticity injected at each sampling step by modulating the variance schedule $\sigma_t(\eta)$. $\eta = 0$ corresponds to a deterministic DDIM process, while $\eta = 1$ corresponds to a fully stochastic DDPM process. To investigate its impact on our framework, we evaluate the PSNR performance of both TC-DDCM variants (DDIM and DDPM) across different η values, as shown in Fig 5. As depicted in Fig 5, the DDPM variant consistently outperforms the DDIM variant across all η values, attributed to the higher precision of its underlying reference trajectory. Specifically, when η is set too high (approaching 1), the injected stochasticity becomes excessive, where random noise dominates the signal. This causes the generative trajectory to deviate significantly from the reference path. Conversely, when η is set too low (approaching 0), the insufficient stochastic support fails to actively pull the generative process back towards the inversion trajectory, resulting in poor alignment.

6 Conclusion

In this paper, we introduce the Trajectory-Consistent Diffusion Codebook Model (TC-DDCM), a unified zero-shot compression framework designed to overcome the limitations of trajectory drift in existing DDCM methods. TC-DDCM reformulates the encoding paradigm from approximating a distant target to explicitly tracking a pre-computed inversion trajectory. Extensive experiments confirm that minimizing the deviation from the reference trajectory effectively shrinks the codebook size and minimizes the inference steps, achieving SOTA performance at ultra-low bitrates.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62276245).

References

[Agustsson and Timofte, 2017] 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*, pages 126–135, 2017.

[Bellard, 2018] Fabrice Bellard. BPG image format, 2018.

[Bińkowski *et al.*, 2018] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[Cai *et al.*, 2025] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18434–18443, 2025.

[Careil *et al.*, 2024] 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.

[Cemgil *et al.*, 2020] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The autoencoding variational autoencoder. *Advances in Neural Information Processing Systems*, 33:15077–15087, 2020.

[Dasgupta and Gupta, 2003] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[Dong *et al.*, 2023] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023.

[Elata *et al.*, 2024] Noam Elata, Tomer Michaeli, and Michael Elad. Zero-shot image compression with diffusion-based posterior sampling. *arXiv preprint arXiv:2407.09896*, 2024.

[Franzen, 1999] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 1999.

[Guo *et al.*, 2025] 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, 2025.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[Hu *et al.*, 2022] Xueqi Hu, Qiusheng Huang, Zhengyi Shi, Siyuan Li, Changxin Gao, Li Sun, and Qingli Li. Style transformer for image inversion and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11337–11346, 2022.

[Huberman-Spiegelglas *et al.*, 2024] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly DDPM noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024.

[Iwai *et al.*, 2024] Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2900–2909, 2024.

[Kawar *et al.*, 2023] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6007–6017, 2023.

[Körber *et al.*, 2024] Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. PerCo (SD): Open perceptual compression. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

[Lei *et al.*, 2023] 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.

[Li *et al.*, 2023] Binglin Li, Jie Liang, Haisheng Fu, and Jingning Han. ROI-based deep image compression with swin transformers, 2023.

[Li *et al.*, 2024] 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*, 2024.

[Li *et al.*, 2025] Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Ajmal Mian. Rdeic: Accelerating diffusion-based extreme image compression with relay residual diffusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[Liu *et al.*, 2023] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023.

[Mallat and Zhang, 1993] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[Mentzer *et al.*, 2020] 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.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

[Muckley *et al.*, 2023] 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*, pages 25426–25443. PMLR, 2023.

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

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

[Park *et al.*, 2025] Chanung Park, Joo Chan Lee, and Jong Hwan Ko. Diffo: Single-step diffusion for image compression at ultra-low bitrates. *arXiv preprint arXiv:2506.16572*, 2025.

[Raphaeli *et al.*, 2025] Ron Raphaeli, Sean Man, and Michael Elad. Silo: Solving inverse problems with latent operators, 2025.

[Rombach *et al.*, 2022] 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*, pages 10684–10695, 2022.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Song *et al.*, 2020] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[Theis *et al.*, 2017] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations*, 2017.

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

[Toderici *et al.*, 2020] George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Balle, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. Workshop and challenge on learned image compression (CLIC2020), 2020.

[Vaisman *et al.*, 2025] Amit Vaisman, Guy Ohayon, Hila Manor, Michael Elad, and Tomer Michaeli. Turbo-ddcm: Fast and flexible zero-shot diffusion-based image compression. *arXiv preprint arXiv:2511.06424*, 2025.

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

[Wallace *et al.*, 2023] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[Wu and De la Torre, 2023] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023.

[Xue *et al.*, 2025] Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. One-step diffusion-based image compression with semantic distillation, 2025.

[Yang and Mandt, 2023] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64971–64995. Curran Associates, Inc., 2023.

[Yue *et al.*, 2025] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23153–23163, 2025.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

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