arXiv:2601.10373v1 [cs.CV] 15 Jan 2026

通过频率感知扩散先验细化实现高效低码率图像压缩

夏一冲[*1,2], **周一民**[*1], **王金鹏** [3], **陈斌** [†3]清华大学深圳国际研究生院，清华大学[2]鹏城实验室[3]哈尔滨工业大学（深圳）
xiayc23,zhouym24,wjp20@mails.tsinghua.edu.cn,
chenbin2021@hit.edu.cn

摘要

近期，基于扩散的生成先验取得了长足进展，使得在极低比特率下实现视觉上逼真的图像压缩成为可能。然而，由于训练范式碎片化，现有方法存在采样过程缓慢和比特分配欠佳的问题。本文提出一种新型压缩框架——通过一致性先验细化加速基于**扩散**的图像压缩（一**致性**先验**细化**，DiffCR），以实现高效且高保真的图像重建。DiffCR 的核心是频率感知跳跃估计（FaSE）模块，该模块细化预训练潜空间扩散模型的 ϵ-预测先验，并通过频率解耦注意力（FDA）使其与不同时间步的压缩潜变量对齐。此外，轻量级一致性估计器在保留扩散采样的语义轨迹的同时，实现快速**两步解码** 。无需更新骨干扩散模型，DiffCR 相较于最先进水平的基于扩散的压缩基线，实现了显著的码率节省（LPIPS 的 BD-rate 降低 27.2%，峰值信噪比的 BD-rate 降低 65.1%），速度提升超过 $10\times$ 倍。

引言

近年来，图像压缩技术取得了显著进展。然而，在低比特率下压缩高分辨率图像仍然面临重大挑战。传统的图像压缩标准（Bellard 2014；Wallace 1992）依赖于人工设计的变换方法。然而，在低比特率（$\leq$ 0.1每像素比特数）场景下，往往会出现明显的块效应和色彩失真问题。尽管基于学习的端到端神经压缩模型（Cheng 等人，2020；Ball´e、Laparra 和 Simoncelli 2016；Ball´e 等人，2018；He 等人，2022）在性能方面优于传统方法，但其由失真驱动的优化目标，在带宽受限时仍会导致重建图像中的纹理和细节大量丢失。（Blau 和 Michaeli 2019；Agustsson 等人，2023）将这一现象概括为比特率、失真和真实感之间的三重权衡，揭示了内在的

这与极端压缩必然导致语义信息丢失和感知质量下降相矛盾。

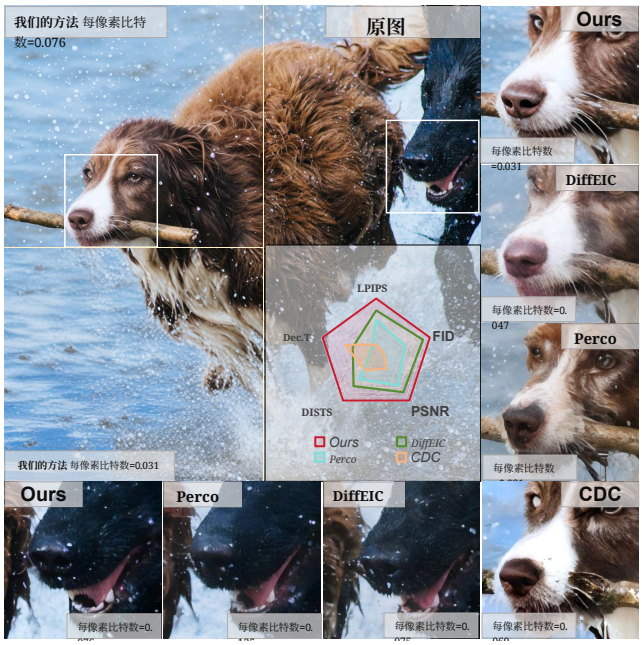

图 1：DiffEIC、Perco、HiFiC、MS-ILLM、CDC 与我们提出的方法之间的定性和定量比较。DiffCR（我们的方法）不仅具有真实感，还能忠实地还原原始图像的细节。右下角的雷达图展示了各种方法在 CLIC20 数据集上的性能比较。放大后观看效果更佳。

为了解决这一问题，近期研究（Mentzer 等人 2020；Muckley 等人 2023；Lu 等 2024；Qin 等人 2023）引入了生成对抗网络（GAN），以优化感知质量。在此基础上，Lee 等人 2024 在编码端和解码端均引入语义信息，以增强图像修复效果。受 GAN 结构和单一感知损失的限制，这些方法在统计保真度方面仍有很大的改进空间。同时，一些基于神经表示和高斯泼溅的 3D 生成方法（Liu 等人 2025a,b）也已应用于视频和图像的有损压缩任务。

[*]这些作者贡献相同。
[†]通讯作者 版权所有 © 2026，人工智能促进协会 (www.aaai.org)。保留所有权利。

相比之下，扩散模型（Dhariwal 和 Nichol 2021；Rombach 等人 2022）通过渐进式去噪过程展现出卓越的生成质量。其重建图像的统计特征更接近自然图像的分布。Yang 和 Mandt 2024、Hoogeboom 等人 2023 首次将去噪扩散概率模型（Ho、Jain 和 Abbeel 2020）框架应用于图像压缩领域，并取得了突破性进展。然而，该方法需要对扩散模型组件进行完整训练，且存在计算成本高、泛化能力受训练数据限制等问题。

近期，基于潜在扩散模型（LDMs）（Rombach 等人 2022）的多项研究提出了新颖的概念，以克服生成对抗网络框架中存在的问题。例如，Lei 等人 2023 致力于通过编码草图和文本语义来操控预训练潜在扩散模型，并利用扩散模型的采样过程进行图像解码。鉴于联合优化语义嵌入与用于压缩目标的传输表示存在困难，他们投入了大量资源来迭代优化语义嵌入并进行对齐。Careil 等人 2024 使用经过充分训练的潜在扩散模型，并通过可训练的码本对条件图像进行编码。然而，这种方法有两个缺点。首先，它需要在一个由数百万张图像构成的数据集上对模型进行精细优化。其次，独立训练压缩器和扩散模型会削弱压缩器的学习能力。Li 等人，2024a 使用预训练的基础潜在扩散模型。然而，其采用的联合训练方法使压缩器难以利用扩散先验学习紧凑表示。Ke 等人 2025 利用多模态大语言模型（MLLM）获取高质量的文本语义，并在超低比特率下取得了显著的性能提升。然而，这也带来了极高的编码延迟。

总体而言，现有基于扩散的图像压缩方案存在两个关键缺陷。（1）训练范式和潜在表示未对齐，导致比特分配欠佳。在当前研究中，压缩器与扩散模型要么分别训练（Yang 和 Mandt 2024；Careil 等人 2024），要么扩散损失与压缩器损失彼此未对齐（Li 等人，2024a；Ke 等人 2025）。已有研究尝试采用代理损失，使压缩器在训练期间能够捕获扩散先验（Xia 等人 2025），但仍未找到最紧凑的压缩表示。（2）基于扩散的方法依赖标准扩散模型，在采样时面临挑战，因为生成逼真图像需要多次迭代。这不仅会导致显著的解码延迟，还会妨碍其与应用于最终图像输出的感知损失无缝集成，进而降低解码保真度。

为解决上述关键问题，我们提出了基于一致性先验细化的加速**扩散**图像压缩方法（DiffCR）。借鉴基础扩散模型中 ϵ-Prediction 的训练范式，我们设计了一种一致性细化估计器（CRE），以主动校正并增强扩散先验，使模型能够兼顾生成图像的真实感与紧凑压缩表征。具体而言，我们设计了频率感知跳跃估计（FaSE），使扩散先验在不同时间步与压缩潜变量对齐；并采用频率解耦注意力（FDA），捕捉二者在不同频率下的上下文关系。此外，我们通过一致性损失，使 FaSE 采样轨迹与原始 ODE轨迹保持一致。与传统的一致性蒸馏方法不同，我们的方法仅用原始去噪网络 $1/100$ 的参数量（8M 对比 800M），便能通过两步实现高保真解码。此外，DiffCR 利用混合语义流进一步增强并稳定生成结果。得益于这些特性，DiffCR 仅需 0.4 秒，便能以极低比特率（$\leq$0.05 每像素比特数）解码出高度逼真的图像，如图 1 所示。

我们的贡献可总结如下：

- 我们提出用于精炼的频率感知跳跃估计（FaSE），以对齐预训练扩散先验与压缩潜变量。这不仅有助于压缩器获得更紧凑的压缩表示，还使我们的模型能够通过一致性约束仅用两步解码出高质量结果。

- 我们设计了频率解耦注意力（FDA），以同步不同时间步生成的扩散先验与压缩潜变量。FDA 在傅里叶域中将二者解耦，并通过时间掩码在不同时间步对其进行调制。
- 基于这些关键要点，我们提出了 DiffCR。DiffCR 不仅提高了编码效率，还利用语义分支稳定并提升解码性能。我们在多个高质量数据集上验证了 DiffCR，证明其具有卓越的感知性能。与一种同样基于 SD 2.1 的先进压缩方法（Li 等人，2024a）相比，DiffCR 的 BD-rate 分别节省了 27.2%（LPIPS）、32.8%（FID）和 65.1%（峰值信噪比）。

背景

学习式潜变量压缩

可学习图像压缩（Ball´e、Laparra 和 Simon- celli 2016；Ball´e 等人，2018；Minnen、Ball´e 和 Toderici，2018；He 等人，2022；Cheng 等人，2020；Qin 等人，2024）包括通过编码器提取图像表示，使用熵模型估计这些表征的概率分布，随后进行熵编码。依据 Shannon 的信息论（Shannon，1948），其训练范式可以统一为以下形式：

$$\mathcal{L} = R_{\boldsymbol{x}} + \lambda D(\boldsymbol{x}, \mathcal{M}(\boldsymbol{x})) \tag{1}$$

这里，$\mathcal{M}(\cdot)$ 表示压缩器，$R_{\boldsymbol{x}}$ 表示图像 $\boldsymbol{x}$的比特率。函数 $D(\cdot)$通常选择像素级失真度量，例如均方误差（MSE）或

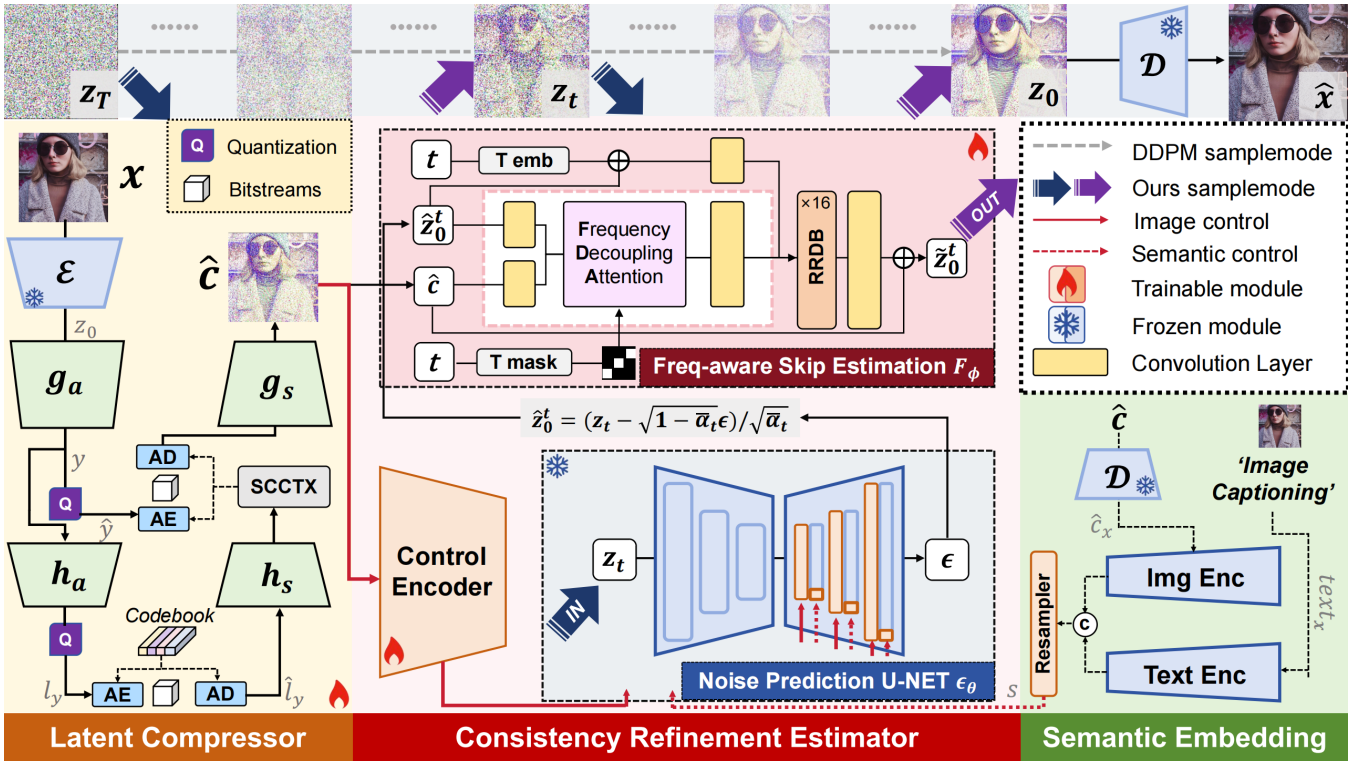

图 2：所提出的 DiffCR 示意图。RRDB 表示残差中的残差密集块（Wang 等人，2018）。DiffCR 利用两个分支引导预训练去噪网络进行生成，并通过 FaSE 模块加强先验校正与细化，使其与压缩器的重建目标保持一致。不同于常规采样过程（灰色虚线所示），DiffCR 支持两步采样方法（蓝紫色箭头所示）。

多尺度结构相似性（Wang、Simoncelli 和 Bovik，2003）。在估计某些生成式压缩任务时（Careil 等人 2024；Mao 等人，2024；Li 等人，2024a；Yang 和 Mandt 2024），压缩器不再对图像进行编码，而是使用预训练分析编码器提取的潜在表示。在这种情况下，训练目标替换为：

$$\mathcal{L} = R_{\mathcal{E}(\boldsymbol{x})} + \lambda D(\mathcal{E}(\boldsymbol{x}), \mathcal{M}(\mathcal{E}(\boldsymbol{x}))) \quad (2)$$

这里，$\mathcal{E}(\cdot)$ 表示预训练分析编码器。这种方法无需为压缩器设计复杂的特征提取模块，同时确保解码后的特征位于预训练生成模型的编码空间中，从而提升最终重建图像的质量。尽管这种方法减轻了压缩器的负担，但也可能导致特征域失真与图像域失真之间不一致，进而使压缩器偏离最优率失真权衡方案。

潜在扩散模型

扩散模型（DMs），也称为基于分数的生成模型（Ho、Jain 和 Abbeel 2020年；Song 等人，2021年），是一类生成模型，它逐步向数据中注入高斯噪声，然后通过反向去噪过程从噪声中生成样本。为了使扩散模型能够在计算资源有限的情况下进行训练，同时保持生成质量，潜在扩散模型（LDMs）（Rombach 等人 2022年）通过编码器 $\mathcal{E}$ 将图像 $\boldsymbol{x}$ 编码为潜在表示$\boldsymbol{z_0}$，并使用解码器 $\mathcal{D}$对其进行重建。

在潜在扩散模型的前向过程中，会逐步向干净潜在特征 $\boldsymbol{z_0}$ 添加高斯噪声，每一步的噪声强度由调度表 β_t 控制。该过程可表述为：

$$\boldsymbol{z_t} = \sqrt{\bar{\alpha}_t}\boldsymbol{z_0} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad t \in \{1, 2, \ldots, T\}, \quad (3)$$

其中，$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 为标准高斯噪声。此处，$\alpha_t = 1 - \beta_t$ 且 $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$。随着 t 增大，受扰动的$\boldsymbol{z_t}$ 逐渐趋近于高斯分布。对潜在扩散模型进行建模的反向过程等价于使用扩散损失训练噪声预测网络 $\boldsymbol{\epsilon_\theta}$：

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\boldsymbol{z_0},t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon_\theta}\left(\boldsymbol{z_t}, \boldsymbol{c}, t\right) \right\|_2^2. \quad (4)$$

推理时，潜在扩散模型使用预训练去噪网络 $\boldsymbol{\epsilon_\theta}(\boldsymbol{z_t}, \boldsymbol{c}, t)$ 和文本条件 $\boldsymbol{c}$ 预测噪声，生成潜在表示 $\boldsymbol{z_{t-1}}$，从而逐步获得最终潜在表示 $\boldsymbol{z_0}$（如图 2 中灰色虚线所示）。Stable Diffusion 作为预训练潜在扩散模型，其先验已被证明可广泛应用于低级图像任务（Lin 等人，2023；Wang 等人，2024；Li 等人，2024b）。

一致性模型

一致性模型（一致性模型）（Song 等人，2023年）使用一致性映射，将 ODE 轨迹上的任意点直接映射回其起点，从而实现快速的少步生成。

形式化地，该映射定义为 $\boldsymbol{f}:(\boldsymbol{x}_t, t) \longmapsto \boldsymbol{x}_\epsilon$其中 ϵ 是一个固定的小正数。其一个关键性质是自一致性：

$$\boldsymbol{f}(\boldsymbol{x}_t, t) = \boldsymbol{f}(\boldsymbol{x}_{t'}, t'), \forall t, t' \in [\epsilon, T]. \tag{5}$$

为确保 $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}, \epsilon) = \boldsymbol{x}$，模型参数化如下：

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = c_{\text{skip}}(t)\boldsymbol{x} + c_{\text{out}}(t)\boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{x}, t), \tag{6}$$

其中，$c_{\text{skip}}(0) = 1$，$c_{\text{out}}(0) = 0$。训练过程中，使用目标模型 $\boldsymbol{\theta}^-$ 强制实施自一致性——该模型是 $\boldsymbol{\theta}$ 的指数移动平均（EMA）：$\boldsymbol{\theta}^- \leftarrow \mu\boldsymbol{\theta}^- + (1-\mu)\boldsymbol{\theta}$。一致性损失为：

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-; \Phi) = \mathbb{E}_{\boldsymbol{x}, t}\left[d\left(\boldsymbol{f}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t_{n+1}}, t_{n+1}\right), \boldsymbol{f}_{\boldsymbol{\theta}^-}\left(\hat{\boldsymbol{x}}^{\Psi}_{t_n}, t_n\right)\right)\right]. \tag{7}$$

这里，$d(\cdot, \cdot)$ 是距离度量。Ψ 表示一个单步常微分方程求解器，用于估计 $\boldsymbol{x}_{t_n}$（来自 $\boldsymbol{x}_{t_{n+1}}$）。（Luo 等人，2023）将其应用于预训练潜在扩散模型，并结合跳步调度，确保非相邻步骤 $t_{n+k} \to t_n$ 之间的一致性，而不是 $t_{n+1} \to t_n$。

方法

总体框架

图 2 展示了 DiffCR 的总体框架。有关详细的算法流程，请参阅附录。图像 $\boldsymbol{x}$首先经过编码器 $\mathcal{E}(\cdot)$，以获得潜在扩散模型的潜在表示 $\boldsymbol{z_0}$，随后将其输入潜在压缩器进行有损压缩，生成比特流及其解码输出 $\hat{}\boldsymbol{c}$。$\hat{}\boldsymbol{c}$ 通过图像控制分支引导去噪网络。同时，我们将失真图像 $\mathcal{D}(\hat{}\boldsymbol{c})$ 提供的语义信息与来自 $\boldsymbol{x}$ 的文本语义进行融合，并将其嵌入去噪网络中。

接下来，DiffCR 通过频率感知跳跃估计，对去噪网络的输出 $\boldsymbol{\epsilon}$ 和图像级控制 $\hat{}\boldsymbol{c}$ 进行校正和细化，将任意时刻的任意输入（$\hat{}\boldsymbol{z}_0^t$）映射回初始状态 $\boldsymbol{z_0}$（$\tilde{\boldsymbol{z}}_0^t$）。与去噪扩散概率模型的采样路径不同，DiffCR 支持两步快速采样。最终采样结果经由解码器 $\mathcal{D}(\cdot)$ 处理，得到最终解码输出 $\hat{}\boldsymbol{x}$。

潜在压缩的一致性细化

频率感知跳跃估计 基础扩散模型通过将条件融入去噪模型来优化式 (4)，逐步提升生成图像的真实感。然而，这一目标无法作为压缩器的重建目标，表现为模式崩溃（Li 等人，2024a）。相反，现有方法通过将潜在表示式 (2) 对齐来训练压缩器，但潜在表示中的信息源自编码器-解码器 $\mathcal{E}$ 和 $\mathcal{D}$的先验，而非去噪网络 $\boldsymbol{\epsilon}_\theta$。FaSE 旨在建立一致性估计器 $\boldsymbol{f}_{\phi,\theta}(\boldsymbol{z_t}, \hat{}\boldsymbol{c}, t)$，将扩散先验和压缩潜在条件转换为相同的预测目标 $\boldsymbol{z_0}$：

$$\boldsymbol{f}_{\phi,\theta}(\boldsymbol{z_t}, \hat{\boldsymbol{c}}, t) = c_{\text{skip}}(t)\boldsymbol{z_t} + c_{\text{out}}(t)\left(\boldsymbol{F}_\phi(\boldsymbol{\epsilon}_\theta(\boldsymbol{z_t}), \hat{\boldsymbol{c}}, t)\right). \tag{8}$$

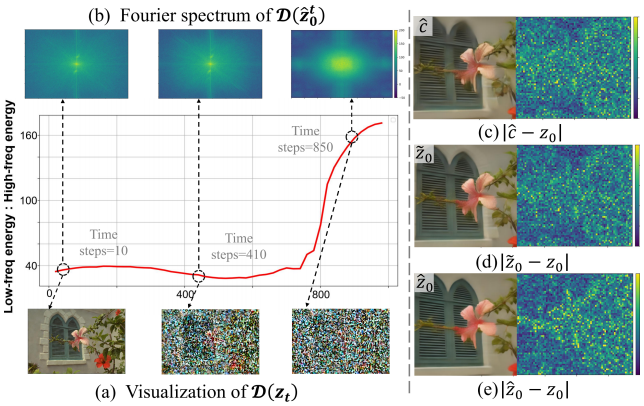

图3：（左）扩散模型在不同时间步下重建的频率能量变化趋势。（右）压缩器、去噪网络和 FaSE 在进行 $\boldsymbol{z_0}$ 预测时的定量性能比较。

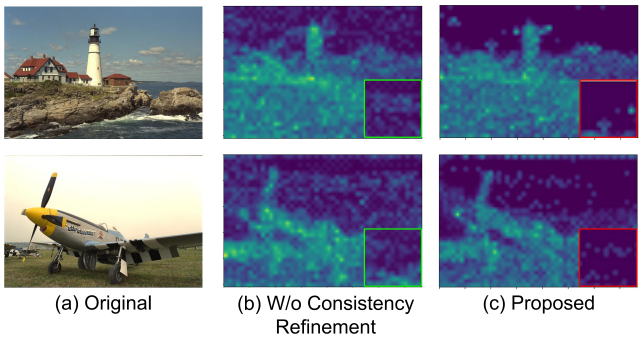

图4：Kodak 上比特率分配的可视化。

其中，$\boldsymbol{F}_\phi(\cdot, \cdot, \cdot)$的目标是将 $\boldsymbol{\epsilon}$-Prediction 先验与图像条件 $\hat{}\boldsymbol{c}$对齐，通过一致性损失预测 PF-ODE 的解（Luo 等人，2023）：

$$\mathcal{L}_{\mathcal{C}}(\boldsymbol{\phi}, \boldsymbol{\phi}^-; \Psi) = \mathbb{E}_{\boldsymbol{z}, \hat{\boldsymbol{c}}, n}\left[d\left(\boldsymbol{F}_{\boldsymbol{\phi}}\left(\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_{t_{n+k}}), \hat{\boldsymbol{c}}, t_{n+1}\right), \boldsymbol{F}_{\boldsymbol{\phi}^-}\left(\boldsymbol{\epsilon}_\theta(\hat{\boldsymbol{z}}^{\Psi}_{t_n}), \hat{\boldsymbol{c}}, t_n\right)\right)\right]. \tag{9}$$

其中，Ψ 表示 DDIM求解器（Song、Meng和 Ermon（2020）），$d(\cdot, \cdot)$ 表示均方误差，且 $k = 20$。与一致性蒸馏不同，$\boldsymbol{F}_\phi(\cdot, \cdot, \cdot)$ 以零初始化开始，因此需要使用 $\boldsymbol{z}_0$ 预测损失来稳定预测过程：

$$\mathcal{L}_{\boldsymbol{F}} = w(t) \cdot \|\boldsymbol{F}_\phi(\boldsymbol{\epsilon}_\theta(\boldsymbol{z_t}), \hat{\boldsymbol{c}}, t) - \boldsymbol{z_0}\|_2^2. \tag{10}$$

Here, $w(t) = \frac{1}{2}\left(\frac{\bar{\alpha}t-1}{1-\bar{\alpha}t-1} - \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right)$.

We refer to the predicted $\boldsymbol{z_0}$ 通过 $\boldsymbol{\epsilon}$-预测（公式（3））获得的结果为$\hat{}\boldsymbol{z_0}$，而由 FaSE 得出的预测结果为 $\tilde{\boldsymbol{z}}_0$。图 3 右侧展示了三种截然不同的 $\boldsymbol{z_0}$ 预测模式。$\hat{}\boldsymbol{c}$ 提供了最粗略的预测结果（图 3(c)），其中预测差异均匀分布于整个图像中。尽管$\hat{}\boldsymbol{z_0}$ 利用先验细化了许多纹理（图 3(e)），但在特定细节处存在明显的不一致性

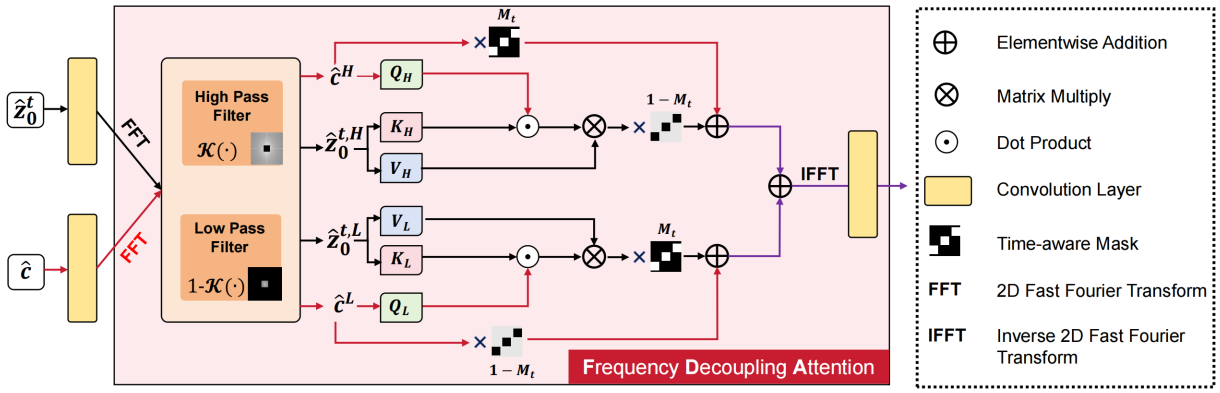

图5：FDA模块示意图，其中 $Q.K.V.$ 对应注意力机制中的查询、键和值。

和失真。相比之下，FaSE 能够有效对齐$\hat{\boldsymbol{c}}$ 与$\hat{\boldsymbol{z}}_0$，兼顾细节一致性与纹理丰富度（图 3(d)）。更重要的是，压缩器与 FaSE 的联合训练使其能够主动捕捉扩散先验在不同区域的关注，从而提取更紧凑的压缩表示。如图 4 所示，与不采用一致性细化相比，压缩器倾向于将比特浪费在自然图像中平坦且高度重复的模式上（例如天空、草地等），如图 4(b) 所示。然而，这些常见纹理可以由扩散先验轻松合成。相比之下，FaSE 能够完美捕捉这些先验信号，从而显著优化压缩器的比特率分配。

频率解耦注意力 扩散模型在不同时间步恢复的信号在频域中呈现出非均匀性。具体而言，在较早的时间步，去噪网络倾向于着重恢复低频信号，而高频信号则更多在较晚的时间步合成（Qian等，2024年；Yuan等，2025）。我们的实验也证实了这一观察结果，如图3左侧所示。这意味着，在设计 $\boldsymbol{F}_\phi(\cdot,\ \cdot,\ \cdot)$时，我们需要考虑不同时间步 t下、不同频域中$\hat{\boldsymbol{z}}_0^t$与$\hat{\boldsymbol{c}}$的对齐。为此，我们引入频率解耦注意力（FDA），如图5所示。具体而言，FDA首先使用快速傅里叶变换 $FFT(\cdot)$ 将$\hat{\boldsymbol{z}}_0^t$和$\hat{\boldsymbol{c}}$变换到傅里叶域，然后通过高通滤波器和低通滤波器分别提取两者的高频和低频分量：

$$\begin{aligned}\hat{\boldsymbol{c}}^{\boldsymbol{H}} &= FFT(\hat{\boldsymbol{c}}) \odot \mathcal{K}, \hat{\boldsymbol{c}}^{\boldsymbol{L}} = FFT(\hat{\boldsymbol{c}}) \odot (1-\mathcal{K}),\\ \hat{\boldsymbol{z}}_0^{\boldsymbol{t},\boldsymbol{H}} &= FFT(\hat{\boldsymbol{z}}_0^{\boldsymbol{t}}) \odot \mathcal{K}, \hat{\boldsymbol{z}}_0^{\boldsymbol{t},\boldsymbol{L}} = FFT(\hat{\boldsymbol{z}}_0^{\boldsymbol{t}}) \odot (1-\mathcal{K}).\end{aligned} \tag{11}$$

随后，我们采用两个独立的交叉注意力机制，分别匹配高频和低频分量：

$$\begin{aligned}Crossattn^H(\hat{\boldsymbol{z}}_0^{\boldsymbol{t},\boldsymbol{H}}, \hat{\boldsymbol{c}}^{\boldsymbol{H}}) &= \text{Softmax}\left(\frac{Q^H K^H}{\sqrt{d}}\right)\cdot V^H,\\ Crossattn^L(\hat{\boldsymbol{z}}_0^{\boldsymbol{t},\boldsymbol{L}}, \hat{\boldsymbol{c}}^{\boldsymbol{L}}) &= \text{Softmax}\left(\frac{Q^L K^L}{\sqrt{d}}\right)\cdot V^L.\end{aligned} \tag{12}$$

随后，我们使用线性时间掩码调节先验$\hat{\boldsymbol{z}}_0^t$在预测中的贡献，具体做法如下：

遵循最佳实践和以往的研究经验，随着时间步逐渐趋近于零，混合内容中先验高频成分的占比逐渐增加。以高频分量为例：

$$\hat{\boldsymbol{z}}_0^{\boldsymbol{t},\boldsymbol{H},\boldsymbol{o}} = Cattn^H(\hat{\boldsymbol{z}}_0^{\boldsymbol{t},\boldsymbol{H}}, \hat{\boldsymbol{c}}^{\boldsymbol{H}}) * M_t + \hat{\boldsymbol{c}}^{\boldsymbol{H}} * (1-M_t). \tag{13}$$

其中 $M_t = 1 - t/T$， T 为扩散模型中的总时间步数。我们在表 2 中展示了针对该模块的消融实验，结果表明，与直接使用交叉注意力相比，FDA 通过时间注意力机制和频率解耦实现了显著的性能提升。

采用混合控制的控制基础扩散

受（Xia 等人 2025）的启发，我们整合了由压缩器重建的图像控制分支，以及通过 CLIP（Radford等，2021年）提取的语义控制分支。我们通过控制编码器（Zhang、Rao和Agrawala，2023年）和交叉注意力将它们注入去噪网络（图 2）。图像控制增强了边缘和纹理生成，从而提升保真度；语义控制则稳定生成过程，并纠正去噪导致的颜色偏移。

对于图像控制，端到端压缩器执行潜空间压缩与重建。特征 $\boldsymbol{z}_0$ 通过 g_a下采样至 $\boldsymbol{y}$ 。基于 VQ 的类别型超先验模型（Jia 等人，2024）提取并压缩超先验，随后将其与 SCCTX 熵模型（He 等人，2022）结合，用于 $\hat{\boldsymbol{y}}$的熵估计。最后， g_s 对 $\hat{\boldsymbol{y}}$ 进行解码，以获得图像级控制 $\hat{\boldsymbol{c}}$。

对于语义控制，图像描述模型（Wang 等人，2022）提取图像文本描述（传输成本 ≤ 0.0001 每像素比特数）。两个 CLIP 编码器分别提取图像语义和文本语义；这些嵌入通过重采样层（Ye 等人，2023）进行调制，并注入去噪网络中。与单独的文本控制相比，这种多模态语义无需传输冗余即可提升解码性能。

训练策略

我们的模型分两个阶段进行训练。在第一阶段的训练中，压缩器、控制模块和跳跃估计

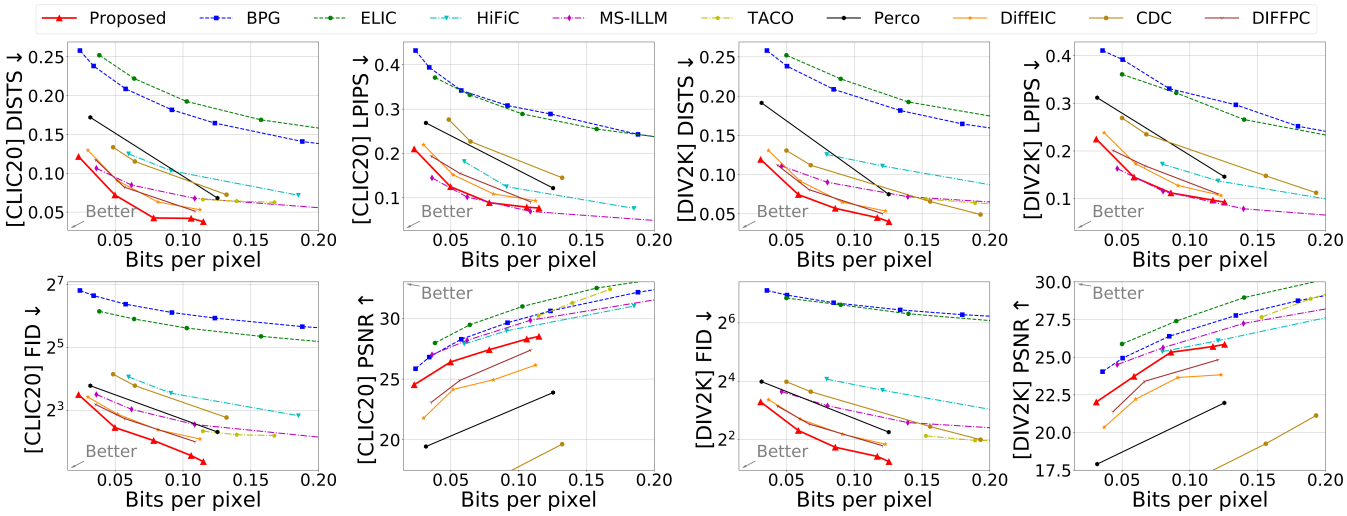

图6：在 CLIC20 和 DIV2K 数据集上的定性比较。灰色箭头表示性能更优的方向。

模块进行联合训练。在优化压缩器提取表征的能力时，也增强了控制模块和跳跃估计的引导能力。对于压缩器，我们需要使用率失真损失来平衡失真与压缩率：

$$\mathcal{L}_{\mathrm{LC}} = \lambda_1 \left\| \boldsymbol{z}_0 - \hat{\boldsymbol{c}} \right\|_2^2 + \lambda_2 R(\hat{\boldsymbol{y}}) + \mathcal{L}_{\mathrm{codebook}}, \quad (14)$$

其中，λ_1 和 λ_2 用于平衡失真和比特率，$\mathcal{L}_{\mathrm{codebook}}$ 的定义如下：

$$\mathcal{L}_{\mathrm{codebook}} = \left\| sg\left(\boldsymbol{l}_y\right) - \hat{\boldsymbol{l}}_y \right\|_2^2 + \beta \left\| sg\left(\hat{\boldsymbol{l}}_y\right) - \boldsymbol{l}_y \right\|_2^2, \quad (15)$$

其中，sg(·)表示停止梯度算子，且 $\beta = 0.25$。

对于跳跃估计模块 $\boldsymbol{F}_\phi(\cdot,\ \cdot,\ \cdot)$，我们需要在保留其用于 $\boldsymbol{z}_0$ 预测的功能的同时，维持其自一致性：

$$\mathcal{L}_{\mathrm{FaSE}} = \lambda_1 \mathcal{L}_{\boldsymbol{F}} + \mathcal{L}_{\mathcal{C}}\left(\boldsymbol{\phi}, \boldsymbol{\phi}^{-}; \Psi\right). \quad (16)$$

由于 $\mathcal{L}_{\boldsymbol{F}}$ 能够有效促进压缩器失真的优化，我们选择与压缩器失真设置一致的超参数。

对于控制模块，我们需要优化原始扩散损失公式（4），以将其整合到去噪网络中：

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\boldsymbol{z_0},\hat{\boldsymbol{c}},s,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{z_t}, \hat{\boldsymbol{c}}, s, t\right) \right\|_2^2. \quad (17)$$

第一阶段的总损失为：

$$\mathcal{L}_{1\mathrm{st}} = \mathcal{L}_{\mathrm{LC}} + \mathcal{L}_{\mathrm{FaSE}} + \lambda_3 \mathcal{L}_{\mathrm{diff}}. \quad (18)$$

在第二阶段，我们主要致力于训练$\boldsymbol{F}_\phi(\cdot,\ \cdot,\ \cdot)$，使其学习采样过程的模式。我们冻结压缩器的参数，并使用 $\boldsymbol{F}_\phi(\cdot,\ \cdot,\ \cdot)$进行两步采样。我们将采样结果映射回图像域，计算其与原始图像之间的重建损失。此外，我们还必须确保控制模块和 $\boldsymbol{F}_\phi(\cdot,\ \cdot,\ \cdot)$获得适当的引导。具体而言：

$$\mathcal{L}_{2\mathrm{st}} = \mathcal{L}_{\mathrm{per}}(\mathcal{D}(\hat{\boldsymbol{z}}_{\mathbf{0}}), \boldsymbol{x}) + \mathcal{L}_{\mathrm{FaSE}} + \mathcal{L}_{\mathrm{diff}}. \quad (19)$$

$\mathcal{L}_{\mathrm{per}}$ 表示感知损失，在这种情况下，我们使用 LPIPS（Zhang 等人，2018年）。

实验

实验设置

数据集。在验证阶段，我们参考了（Muckley 等人，2023年；Yang 和 Mandt 2024年），并使用了四个广受认可的图像压缩基准数据集：CLIC20（George Toderici，2020年）、DIV2K（Timofte 等人，2017年）、Kodak（Company，2013年）和 Tecnick （Asuni 和 Giachetti，2013 年）。CLIC20 包含 428 张高清图像。为确保在 DIV2K 上进行更可靠的统计保真度验证，我们选取了 DIV2K-train 中最初的 400 张图像和 DIV2K-valid 中的 100 张图像，共计 500 张图像用于测试。

基线。 我们将提出的 DiffCR 与多种具有代表性的图像压缩方法进行比较，涵盖传统压缩标准 BPG（Bellard 2014）；基于变分自编码器的压缩方法 ELIC（He 等人，2022）；生成对抗网络驱动的压缩技术，包括 HiFiC（Mentzer et al. 2020）、MS-ILLM（Muckley et al. 2023）和 TACO（Lee et al. 2024）；以及基于扩散模型的方法，如 CDC（Yang 和 Mandt 2024）、PerCo（Careil 等人 2024）、DiffEIC（Li 等人，2024a）、DiffPC（Xia 等人 2025）和 ResULIC（Ke 等人 2025）。更多详情请参阅附录。对于某些功能未公开的基线方法（Ke 等人 2025），我们仅比较其在 Kodak 上公开的数据点。

指标。我们采用多种广受认可的指标来衡量重建图像的视觉质量，包括 LPIPS（Zhang 等，2018年）、DISTS（Ding 等，2020年）、FID（Heusel 等，2017年）和 KID（Bi´nkowski 等，2018年）等感知指标。此外，我们还使用 PSNR 和多尺度结构相似性（Wang、Simon- celli 和 Bovik，2003年）等失真指标来评估重建的保真度。除非另有说明，LPIPS 均使用 Alex 网络计算；我们还报告了使用 VGG 网络计算的 LPIPS-V 指标验证结果。对于 Kodak 和 Tecnick 数据集，由于数据不足以计算统计保真度

模型	编码时间（秒）	解码时间（秒）	BD-rate(%)
			DISTS
ELIC	0.009	0.008	1017.3
HiFiC	0.012	0.062	376.04
MS-ILLM	0.069	0.068	125.25
CDC	0.007	3.081	116.47
Perco	0.131	7.455	90.757
DiffEIC	0.402	6.618	28.062
DiffPC	0.089	7.325	26.831
ResULIC	181.16*	1.021	16.904
Ours	0.076	0.481	0

表1：Kodak数据集上的编码和解码时间。BD-rate 越高，性能越差。* 所报告的数值取自（Ke 等，2025）的论文。

指标，因此我们不在这些数据集上验证 FID 和 KID 指标。更多细节请参阅附录。

主要结果

图6展示了 DiffCR 与基线方法在 CLIC20 和 DIV2K 上的定量比较。显然，我们提出的方法在感知质量方面表现出色：在 LPIPS 指标上达到与当前最先进方法 MS-ILLM 相当的水平，并始终优于所有扩散基线。特别是，在附录中，我们展示了 KID 和 LPIPS-V 两项指标，证明 DiffCR 持续优于所有基线。值得注意的是，感知指标的提升并非以增加失真为代价：在峰值信噪比方面，DiffCR 显著优于其他扩散基线，在 DIV2K 数据集上甚至展现出与 HiFiC 相当的竞争力。这清楚地表明，从压缩的角度来看，DiffCR 实现了更优的失真—感知权衡，而非仅仅依赖扩散模型先验来增强图像。补充材料中，我们还提供了定性视觉对比，以进一步阐明 DiffCR 的性能优势。

复杂度比较 表1展示了 Kodak数据集上 DiffCR 与基线方法的编码和解码延迟以及性能比较。所采用的性能指标为 BD-rate（Bjontegaard 2001）。以 DiffCR 作为参考模型时，较高的 BD-rate 表明模型需要更多比特才能达到与参考模型相同的指标。DiffCR 不仅具备令人满意的编码速度，而且由于仅需两步采样，其解码速度比采用 50 个采样步数的典型扩散基线快$10\times$，同时性能仍优于（Ke 等人 2025）的 4 步解码 $2\times$。

语义嵌入	CRE	FDA	两阶段	BD-rate(%)	
				LPIPS	FID
✔	✔	✔		44.5	51.1
✔			✔	38.24	41.21
✔	✔		✔	13.52	15.64
	✔	✔	✔	8.32	10.11
✔	✔	✔	✔	0	0

表 2：在 CLIC20 数据集上对各模块进行的消融研究

消融实验

我们在 CLIC20 上进行了消融实验，以 DiffCR 为基线，展示不同消融设置下的 BD-rate，如表2所示。

一致性精炼估计器（CRE）。在图 4 中，我们展示了 CRE 在优化比特率分配方面的优势。如表 2 所示，移除 CRE 会导致性能下降，最高达 40%。

频率解耦注意力（FDA）。我们用交叉注意力层替换了 FDA 模块。如前文所述，标准交叉注意力难以捕捉扩散先验和压缩潜变量在不同时间步下的复杂上下文，导致性能下降约 14%。

第二阶段训练（两阶段）。在第二阶段直接训练模型，使其从图像域失真中学习采样模式。省略这一训练阶段会导致性能下降约 47%。

语义嵌入（语义嵌入）。将混合语义嵌入替换为纯文本语义嵌入，会导致性能下降 9%，并使解码图像出现轻微的颜色偏差。

我们还讨论了模型对不同图像描述模型提取文本的敏感性，并针对采样步数进行了消融实验。更多细节请参阅附录。

结论

本文提出了 DiffCR，这是一种高保真、低比特率图像压缩框架，利用一致性细化扩散先验。通过大量实验，我们证明了频率感知跳跃预测模块在优化比特率分配、平衡感知质量与失真之间的权衡方面发挥了重要作用。DiffCR 在低比特率场景下具有令人满意的感知质量和解码效率。此外，灵活的模型架构展现出广阔的实际应用前景。

致谢

本研究部分受国家自然科学基金（项目编号：624 B2088、62301189、62571298）和深圳市科技计划（项目编号：KJZD20240903103702004、SYSPG20241211173609009）资助。

参考文献

Agustsson, E.; Minnen, D.; Toderici, G.; and Mentzer, F. 2023年. 使用条件生成器的多真实感图像压缩. 收录于 *IEEE/CVF* 计算机视觉与模式识别会议论文集 , 22324–22333. Asuni, N.; and Giachetti, A. 2013年. TESTIMAGES：用于显示和算法测试的大型数据档案库. 《图形工具期刊》, 17(4): 113–125. Ball´e, J.; Laparra, V.; and Simoncelli, E. P. 2016. 端到端优化的图像压缩. *arXiv*预印本 *arXiv:1611.01704*. Ball´e, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018年. 使用尺度超先验的变分图像压缩. *arXiv*预印本 *arXiv:1802.01436*. Bellard, F. 2014. BPG图像格式. https://bellard.org/ bpg/. Bi´nkowski, M.; S utherland, D. J.; Arbel, M.; and Gret- ton, A. 2018年. 揭开MMD生成对抗网络的神秘面纱. *arXiv*预印本 *arXiv:1801.01401*. Bjontegaard, G. 2001. 率失真曲线之间平均峰值信噪比差值的计算. 国际电信联盟第*16*研究组文件 *VCEG-M33*. Blau, Y.; and Michaeli, T. 2019. 重新思考有损压缩：率失真感知权衡. 收录于国际机器学习会议, 675–685. PMLR. Careil, M.; Muckley, M. J.; Verbeek, J.; and Lathuili`ere, S. 2024年. 探索在超低比特率下实现完美真实感的图像压缩. 收录于第十二届国际学习表征会议. Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020年. 采用离散化高斯混合似然和注意力模块的学习式图像压缩. 收录于*IEEE/CVF*计算机视觉与模式识别会议论文集, 7939–7948. Company, E. K. 2013年. 柯达无损真彩色图像套件. Dhariwal, P.; and Nichol, A. 2021年. 扩散模型在图像合成方面胜过生成对抗网络. 神经信息处理系统进展, 34: 8780–8794. Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020年. 图像质量评估：统一结构与纹理相似性. *IEEE*模式分析与机器智能汇刊, 44(5): 2567–2581. George Toderici, W. S. 2020年. 学习图像压缩研讨会与挑战赛（CLIC2020）. He, D.; Yang, Z.; Peng, W.; Ma, R.; Qin, H.; and Wang, Y. 2022年. ELIC：采用不均匀分组空间通道上下文自适应编码的高效学习式图像压缩. 收录于*IEEE/CVF*计算机视觉与模式识别会议论文集, 5718–5727. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017年. 通过两时间尺度更新规则训练的 GAN 收敛到局部纳什均衡。神经信息处理系统进展，30. Ho, J.; Jain, A.; and Abbeel, P. 2020年. 去噪扩散概率模型。神经信息处理系统进展，33: 6840–6851. Hoogeboom, E.; Agustsson, E.; Mentzer, F.; Versari, L.; Toderici, G.; and Theis, L. 2023年. 基于分数的生成模型实现高保真图像压缩。*arXiv*预印本 *arXiv:2305.18231*. Jia, Z.; Li, J.; Li, B.; Li, H.; and Lu, Y. 2024年. 面向超低比特率图像压缩的生成式潜在编码。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，26088–26098. Ke, A.; Zhang, X.; Chen, T.; Lu, M.; Zhou, C.; Gu, J.; and Ma, Z. 2025. 采用语义残差编码与压缩感知扩散的超低码率图像压缩。*arXiv*预印本 *arXiv:2505.08281*. Lee, H.; Kim, M.; Kim, J.-H.; Kim, S.; Oh, D.; and Lee, J. 2024年. 通过文本引导编码实现像素级和感知保真度的神经图像压缩。载于国际机器学习会议. Lei, E.; Uslu, Y. B.; Hassani, H.; and Bidokhti, S. S. 2023年. Text+ sketch：超低码率图像压缩。*arXiv*预印本 *arXiv:2307.01944*. Li, Z.; Zhou, Y.; Wei, H.; Ge, C.; and Jiang, J. 2024a. 利用潜在特征引导和扩散先验实现极限图像压缩。*IEEE*电路与视频技术汇刊. Li, Z.; Zhou, Y.; Wei, H.; Ge, C.; and Mian, A. 2024b. 基于扩散的极限图像压缩与压缩特征初始化。*arXiv*预印本 *arXiv:2410.02640*. Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023年. Diffbir：基于生成式扩散先验的盲图像复原。*arXiv*预印本 *arXiv:2308.15070*. Liu, X.; Chen, B.; Liu, Z.; Wang, Y.; and Xia, S.-T. 2025a. 使用熵约束三维高斯进行二维视频压缩的探索。载于第十三届国际学习表征会议. Liu, X.; Chen, J.; Chen, B.; Liu, Z.; An, B.; Xia, S.-T.; and Wang, Z. 2025b. 一种基于混合自回归模型、适用于低复杂度解码的高效隐式神经表示图像编解码器。*IEEE*多媒体汇刊. Lu, L.; Xie, Y.; Jiang, W.; Wang, W.; Lin, X.; and Wang, Y. 2024年. HybridFlow：将连续性融入掩码码本，实现极低比特率图像压缩。载于第*32*届*ACM*国际多媒体会议论文集，3010–3018. Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023年. 潜在一致性模型：通过少步推理合成高分辨率图像。*arXiv*预印本 *arXiv:2310.04378*. Mao, Q.; Yang, T.; Zhang, Y.; Wang, Z.; Wang, M.; Wang, S.; Jin, L.; and Ma, S. 2024年. 极限图像压缩

采用微调后的向量量化生成对抗网络。载于*2024*年数据压缩会议（*DCC*），203–212. IEEE. Mentzer, F.; Toderici, G. D.; Tschannen, M.; and Agustsson, E. 2020年. 高保真生成式图像压缩。神经信息处理系统进展，33: 11913–11924. Minnen, D.; Ball´e, J.; and Toderici, G. D. 2018年. 用于学习式图像压缩的联合自回归先验与分层先验。神经信息处理系统进展，31. Muckley, M. J.; El-Nouby, A.; Ullrich, K.; J´egou, H.; and Verbeek, J. 2023年. 利用隐式局部似然模型提升神经图像压缩的统计保真度。载于国际机器学习会议，25426–25443. PMLR. Qian, Y.; Cai, Q.; Pan, Y.; Li, Y.; Yao, T.; Sun, Q.; and Mei, T. 2024年. 在频域中采用移动平均采样增强扩散模型。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，8911–8920. Qin, S.; Chen, B.; Huang, Y.; An, B.; Dai, T.; and Xia, S.-T. 2023年. 利用协同跨模态辅助信息进行感知图像压缩。*arXiv*预印本 *arXiv:2311.13847*. Qin, S.; Wang, J.; Zhou, Y.; Chen, B.; Luo, T.; An, B.; Dai, T.; Xia, S.; and Wang, Y. 2024年. MambaVC：采用选择性状态空间的学习式视觉压缩。*arXiv*预印本 *arXiv:2405.15413*. Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021年. 从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议，8748–8763. PMLR. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022年. 利用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，10684–10695. Shannon, C. E. 1948年. 通信的数学理论。贝尔系统技术期刊，27(3): 379–423. Song, J.; Meng, C.; and Ermon, S. 2020年. 去噪扩散隐式模型。*arXiv*预印本 *arXiv:2010.02502*. Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023年. 一致性模型。载于国际机器学习会议，32211–32252. PMLR. Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021年. 通过随机微分方程实现基于分数的生成建模。载于国际学习表征会议. Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; Zhang, L.; Lim, B.; et al. 2017年. NTIRE 2017单幅图像超分辨率挑战赛：方法与结果。载于*IEEE*计算机视觉与模式识别会议（*CVPR*）研讨会. Wallace, G. K. 1992 年 . JPEG 静态图像压缩标准。*IEEE*消费电子汇刊，38(1): xviii–xxxiv. Wang, J.; Yue, Z.;Zhou,S.; Chan,K. C.; and Loy, C. C. 2024年. 利用扩散先验进行真实世界图像超分辨率。国际计算机视觉期刊。

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai,S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022年. OFA：通过简单的序列到序列学习框架统一架构、任务和模态。*CoRR*, abs/2202.03052.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018年. ESRGAN：增强型超分辨率生成对抗网络。载于欧洲计算机视觉会议（*ECCV*）研讨会论文集, 0–0. Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003年. 用于图像质量评估的多尺度结构相似性。载于第三十七届阿西洛马会议：信号、系统与计算机，*2003*年, 第2卷，1398–1402。IEEE。Xia, Y.; Zhou, Y.; Wang, J.; An, B.; Wang, H.; Wang, Y.; and Chen, B. 2025. DiffPC：通过语义细化实现基于扩散的高感知保真度图像压缩。载于第十三届国际学习表征会议。Yang, R.; and Mandt, S. 2024年. 使用条件扩散模型进行有损图像压缩。神经信息处理系统进展, 36. Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023年. IP-Adapter：兼容文本的图像提示适配器，用于文生图扩散模型。*arXiv*预印本 *arXiv:2308.06721*。Yuan, Y.; Guo, Y.; Wang, C.; Zhang, W.; Xu, H.; and Zhang, L. 2025. FreqPrior：通过频率滤波高斯噪声改进视频扩散模型。*arXiv*预印本 *arXiv:2502.03496*。Zhang, L.; Rao, A.; and Agrawala, M. 2023年. 为文生图扩散模型添加条件控制。载于*IEEE/CVF*国际计算机视觉会议论文集, 3836–3847。Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018年. 深度特征作为感知度量的惊人有效性。载于*IEEE*计算机视觉与模式识别会议论文集, 586–595。