arXiv:2512.15270v1 [eess.IV] 17 Dec 2025

基于预训练扩散模型的用于图像压缩的生成式预处理

Mengxi Guo, ShijieZhao*, Junlin Li, Li Zhang字节跳动，深圳，中国 字节跳动，美国加利福尼亚州圣迭戈，92122{guomengxi.qolab, zhaoshijie.0526, lijunlin.li, lizhang.idm}@bytedance.com

摘要

预处理是一种成熟的压缩优化技术，但现有方法主要针对率失真（R-D）进行优化，并受限于像素级保真度。本研究首次将大规模预训练扩散模型用于压缩预处理，开创了向率感知（R-P）优化迈进的新方向。我们提出了一种两阶段框架：首先，使用一致性分数恒等蒸馏（CiD），将多步 Stable Diffusion 2.1 蒸馏为紧凑的一步图像到图像模型。其次，在率感知损失和可微编解码器替代模型的引导下，对蒸馏模型的注意力模块进行参数高效微调。我们的方法可与标准编解码器无缝集成，无需进行任何修改，并利用模型强大的生成式先验增强纹理、减轻伪影。实验表明，该方法在率感知方面取得显著提升：在 Kodak 数据集上，DISTS 的 BD 码率最高降低 30.13%，且主观视觉质量更优。

引言

预处理是一种广泛应用于优化压缩的技术，能够显著提升图像和视频编解码器的率失真（R-D）性能。近年来，深度学习在压缩领域的普及推动了基于学习的预处理方法迅速发展。这些工作大致可分为两类：一类旨在提升压缩算法本身的率失真性能，另一类则致力于提高压缩数据在下游机器视觉任务中的表现。

在第一类方法中，一种常见范式是采用多层卷积神经网络（CNN）作为预处理模块。该模块通常与传统编解码器的可微替代模型进行端到端训练，并以率失真优化为指导。尽管这些方法取得了成功，但仍存在两个主要局限。首先，它们严重依赖像素级损失函数（例如均方误差或 L1）作为主要优化目标 [1, 2], ，而与主观质量相关性更高的损失（如对抗损失或感知损失 [3], ）往往只起辅助作用。尽管这种策略对于防止图像发生剧烈变化是必要的，但也迫使人们在码率、失真和感知之间做出艰难权衡。这一限制可由率失真感知权衡 [4], 解释，它从根本上限制了预处理的全部潜力。其次，现有研究通常采用从头开始训练的小型模型

*通讯作者

，未能利用现代大规模预训练基础模型的强大能力 [5, 6, 7] ，而这些模型已在各种生成和复原任务中展现出卓越性能。然而，这些模型在图像压缩预处理任务中的应用仍未得到探索。

受这些观察结果的启发，我们首次针对图像压缩预处理任务引入了预训练基础模型。当前最先进的预训练模型主要是文本到图像（T2I）扩散模型 [6,7, 5, 8]。然而，T2I 合成与压缩预处理的目标之间存在根本性的不匹配，主要体现在 T2I 模型的模型大小过大、推理速度缓慢，以及其细节保真度通常有限。为解决这些问题，我们提出了一种新颖的生成式预处理框架。首先，为构建轻量且高效的模型，我们使用变分分数蒸馏（VSD） [9] 对大型预训练 T2I 模型进行蒸馏，以获得紧凑的单步模型。该蒸馏模型的参数显著减少，计算开销也更低，但其生成能力经过了有意削弱。

随后，我们对蒸馏模型进行微调，使其适应压缩预处理任务。鉴于大多数当前最先进的生成模型都基于 UNet 中的 Transformer 块构建，我们重点研究了如何高效地微调这类结构。实验发现，仅更新模型中一小部分参数，就能使模型具备压缩优化能力，同时不会牺牲其从原始预训练模型继承的丰富语义理解和纹理合成能力。在微调过程中，我们设计了可微 BGP（diff-BPG），用于模拟真实世界编解码器的失真和码率波动，并采用率感知损失函数来管理处理强度与比特率之间的权衡。最后，与现有的基于 CNN 的预处理方法相比，我们的生成式方法不仅在 Kodak 数据集的客观指标上实现了 30.13% 的 BD 码率降低，还在主观评估中展现出更优的性能。

方法

我们提出的方法的整体架构如图1所示，包含两个主要阶段：用于创建紧凑型单步生成器的蒸馏阶段，以及用于使生成器适应压缩预处理任务的率失真感知微调阶段。

扩散模型

我们的方法基于预训练的 Stable Diffusion 2.1 [5], ，该模型在文本到图像合成方面展现了最先进的性能。扩散模型是一类概率生成模型，通过逆转逐步加噪的过程来学习生成数据。该过程包含两个主要阶段：正向扩散过程和反向去噪过程。

前向过程 q 会在一系列 T 时间步内，逐渐向初始数据样本 x_0 添加高斯噪声。在每个时间步 t，数据 x_t 通过以下方式生成：

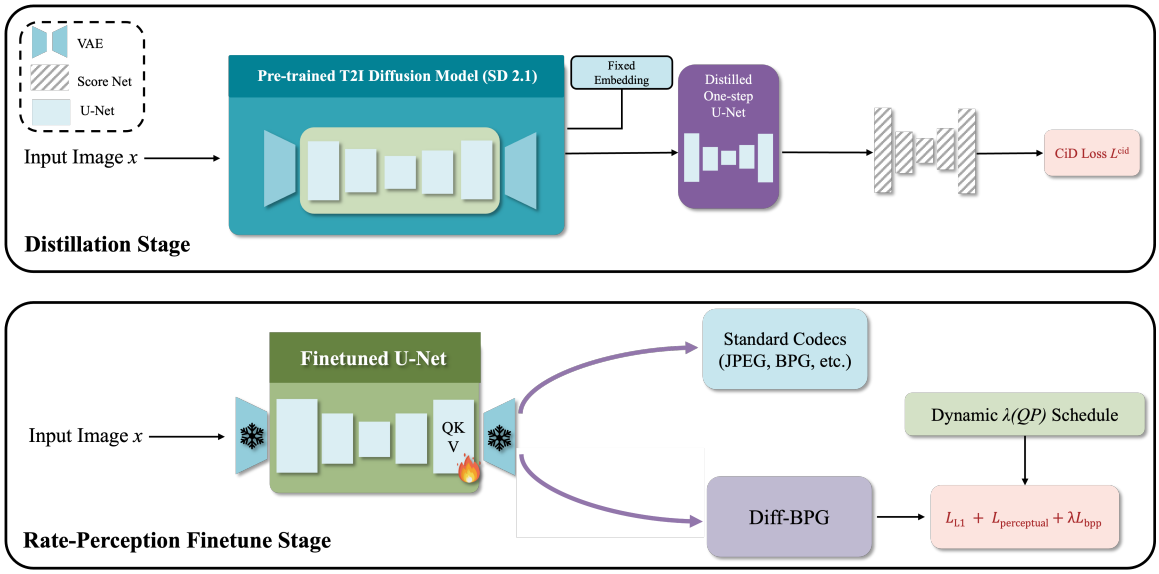


图1：我们提出的两阶段框架概览。**（顶部）蒸馏阶段：**我们使用一致性分数恒等蒸馏（CiD），将预训练的多步 Stable Diffusion 2.1 模型蒸馏为紧凑的单步 U-Net。VAE 和文本条件（由固定嵌入替代）保持冻结。**（底部）码率-感知微调阶段：**对蒸馏后的单步生成器进行微调，以完成预处理任务。输入图像经微调后的 U-Net （ VAE 保持冻结）处理后，再通过可微分的 BPG （ Diff-BPG ）替代模型。复合损失在动态的基于QP的调度下平衡 L1、感知项和比特率项，引导对 U-Net 注意力模块的优化。

从以前一个样本 x_{t-1} 为条件的高斯分布中采样：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

其中 $\{\beta_t\}_{t=1}^T$ 是预定义的方差调度。

反向过程旨在通过逆转正向加噪过程来学习数据分布。具体而言，通过训练一个神经网络（通常为记作 ϵ_θ 的 U-Net架构），预测在时间步 t 被添加到含噪数据 x_t 中的噪声 ϵ 。

为提高计算效率，我们在较低维度的潜在空间中进行操作。我们预先训练一个由编码器 $\mathcal{E}$ 和解码器 $\mathcal{D}$组成的自编码器，将图像 x 压缩为潜在表示 $z = \mathcal{E}(x)$。随后，将上述扩散过程应用于该潜变量 z。

蒸馏

标准 Stable Diffusion 2.1 模型的迭代特性导致压缩预处理的计算成本高得难以承受。为了构建高效的单步生成器，我们必须调整蒸馏过程本身。标准的噪声到图像

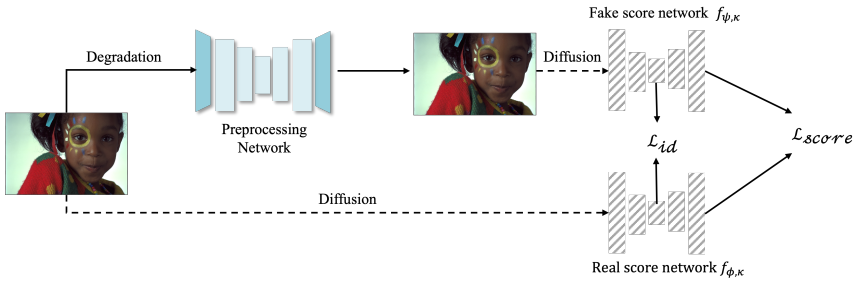


图 2：蒸馏阶段详解。

蒸馏并不适用于我们的预处理任务，因为它忽略了图像到图像翻译中固有的强输入—输出相关性。因此，我们调整了一种专为图像到图像任务设计的框架——一致性分数恒等式蒸馏（CiD） [9], ，将蒸馏重新表述为一个有引导的图像翻译问题。

CiD 框架如图 2 所示，旨在解决变分分数蒸馏（VSD）中的不稳定问题。VSD 依赖生成的潜变量 z_g ，其质量会不断波动。CiD 使用预训练教师模型充当**真实分数网络**，指导学生模型（即**虚假分数网络**），从而稳定训练。至关重要的是，它还引入了一个高保真潜变量锚点，我们将其定义为原始未处理图像的潜在表示 z_h 。最终训练目标由基于这一概念得出的两个核心损失项组合而成。

第一项是**恒等损失**， $\mathcal{L}_{\text{id}}$，它是 CiD 的主要驱动力。它用稳定的锚点 z_h取代不稳定的目标 z_g 。该损失通过使真实分数网络 $f_{\phi,\kappa}$ 与虚假分数网络 $f_{\psi,\kappa}$ 之间的分数差异与指向高保真锚点 z_h的方向保持一致，确保一致性：

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{t,\epsilon,c,z_t}\left[w(t)\left\langle f_{\phi,\kappa}(z_t,t,c) - f_{\psi,\kappa}(z_t,t,c), f_{\phi,\kappa}(z_t,t,c) - z_h\right\rangle\right]. \tag{2}$$

第二项是**分数差异损失**， $\mathcal{L}_{\text{score}}$。它的作用与原始 VSD 目标类似，但其形式经过设计，可规避无法计算虚假分数网络梯度 $\nabla_\theta\psi$的问题。按照先前工作的记法，将其记为 $\mathcal{L}_1$ 。随后，这两个损失与一个经验加权因子 ξ 结合，构成最终的 CiD 目标：

$$\mathcal{L}_{\text{cid}} = \mathcal{L}_{\text{id}} - \xi\mathcal{L}_{\text{score}}. \tag{3}$$

该蒸馏框架兼具强大的生成先验和强有力的图像条件约束，可生成高度高效、适用于我们任务的单步模型。最终架构由一个紧凑的 U-Net 和原始变分自编码器组成，移除了文本编码器并以静态嵌入取而代之，从而将推理时间缩短至原始模型的约 1%。

率-感知微调

虽然蒸馏后的模型计算效率很高，但它仍以生成模型的方式运行，无法直接用于压缩预处理。因此，

我们引入一个微调阶段，将其能力从生成适配到预处理，如图1下方所示。扩散模型的核心架构由 U-Net 和变分自编码器组成。按照将变分自编码器用作固定特征提取器的标准做法，我们冻结预训练变分自编码器的权重，仅对 U-Net 进行微调。

鉴于 U-Net 的基于注意力的架构，我们采用参数高效微调策略。我们发现，只更新每个 Transformer 块内注意力模块的参数，就足以取得出色的性能。具体而言，对于给定的注意力模块，其参数由查询 (W_Q)、键 (W_K) 和值 (W_V) 的投影矩阵组成。注意力机制的形式如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

其中，$Q = XW_Q$、$K = XW_K$、$V = XW_V$ 是从输入序列 X 投影得到的查询、键和值矩阵。只更新这些投影矩阵，可以让模型学习预处理所需的特定任务映射，同时保留基础模型中丰富的预训练特征。

可微分 BPG 编解码器 为了实现端到端训练，我们需要为目标编解码器构建一个可微的替代方案。受 [10], 启发，我们设计了一个可微分 BPG（diff-BPG）模块，用于模拟 BPG 压缩的关键阶段。为减少空间冗余，我们在模拟中加入帧内预测步骤，该步骤基于相邻像素使用一组候选模式（例如平面模式、直流模式、角度模式）。最佳预测模式的选择原本是对候选模式的率失真代价执行不可微的 argmin 运算；为实现梯度传播，我们使用 soft-argmin 近似对其进行松弛 [11]。另一个主要挑战是量化中的不可微舍入运算，我们通过采用基于傅里叶级数的近似来解决：

$$\hat{y} = y - \frac{1}{2\pi}\sum_{n=1}^{N}\frac{(-1)^{n+1}}{n}\sin(2\pi n y), \tag{5}$$

其中，y 是量化除法后的值，$\hat{}y$ 是其可微的舍入对应值。最后，熵编码阶段由学习式熵模型 [12] 取代，该模型可对每像素比特数（bpp）进行可微估计。

损失函数 我们的微调过程由一个复合损失函数引导，该函数旨在平衡码率与感知。总损失 $\mathcal{L}_{\text{total}}$ 定义如下：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{L1}} + w_p\mathcal{L}_{\text{perceptual}} + \lambda\mathcal{L}_{\text{bpp}}, \tag{6}$$

其中，$\mathcal{L}_{\text{L1}}$ 是原始图像与预处理重建图像之间的 L1 损失，有助于保留精细细节。
$\mathcal{L}_{\text{perceptual}}$ 是 DISTS 损失 [13], ，这是一种已知与人类感知质量高度相关的指标。$\mathcal{L}_{\text{bpp}}$ 是估计的

表 1：与未进行预处理的锚定编码器相比，BDBR（%）的降幅。数值越低越好，表示在感知质量相同的情况下码率节省越多。我们的方法在所有编解码器和数据集上均持续优于 TDP 基线。

数据集	指标	JPEG		WebP		BPG	
		TDP	我们的方法	TDP	我们的方法	TDP	Ours
CLIC	LPIPS	-1.62	**-5.74**	-3.65	**-8.30**	-11.18	**-23.16**
	DISTS	-3.99	**-9.74**	-7.88	**-21.43**	-11.2	**-27.68**
	TOPIQ-fr	-0.23	**-7.88**	-6.25	**-23.80**	-6.48	**-22.38**
Kodak	LPIPS	-1.33	**-6.14**	-3.37	**-7.43**	-6.17	**-12.57**
	DISTS	-4.27	**-13.69**	-8.30	**-25.22**	-12.62	**-30.13**
	TOPIQ-fr	-2.94	**-7.22**	-1.61	**-16.40**	-7.27	**-15.64**

比特率，来自我们的 diff-BPG 模块。项 w_p 是感知损失的固定权重。超参数 λ 控制感知与码率之间的权衡。我们根据训练期间使用的量化参数（QP），按照经验推导的计划动态调整 λ：

$$\lambda(\mathrm{QP}) = e^{(w_1 \cdot \mathrm{QP} + w_2)}, \tag{7}$$

其中我们设置 $w_1 = -0.115$和 $w_2 = 1.145$。这使模型能够在广泛的压缩级别范围内学习有效的率失真（R-P）权衡。

实验

实现细节

我们的训练过程分为两个不同阶段：蒸馏和微调。在蒸馏阶段，我们以官方 SD2.1 为模型初始化权重。蒸馏使用由 LSDIR [14]和 FFHQ [15]组成的混合数据集进行。在整个阶段中，我们采用固定学习率 1×10^{-5}。在随后的率失真-感知微调阶段，我们使用 DIV2K 和 Flickr2K [16] 数据集作为主要训练数据，并从中随机提取 256×256 图像块。

为增强模型对图像中具有挑战性区域的关注，我们引入了一种基于恰可察觉失真（JND）的数据筛选步骤 [17]。我们对图像块进行筛选，仅保留 JND 分数大于 0.8 的图像块。该策略优先选择纹理复杂度较高的图像块，例如背景区域；在人类视觉系统对细微变化不太敏感的区域，生成式预处理能够发挥最大效用。在微调过程中，我们的 diff-BPG 模块采用在 $[10, 50]$ 范围内随机波动的量化参数（QP）。微调使用 Adam 优化器进行优化，初始学习率为1×10^{-3}。我们采用余弦退火调度，将学习率逐渐降至最终值 1×10^{-8}。整个框架基于 PyTorch 并使用 CUDA 实现

支持。所有实验均在一台配备 8 个 NVIDIA V100 GPU 的服务器上进行。

在评估中，我们采用标准基准数据集：Kodak 数据集和 CLIC 专业验证数据集。在测试阶段，我们使用多种真实世界的图像压缩编解码器评估预处理图像的性能：JPEG、WebP 和 BPG。我们采用每像素比特数（bpp）衡量压缩过程中产生的编码成本。为评估感知质量，我们使用一系列成熟的指标，包括 LPIPS [18], DISTS [13], 和 TOPIQ-fr [19]。最终性能增益通过计算 Bjøntegaard Delta 比特率（BDBR）进行综合量化，该指标衡量在感知质量相同的情况下平均节省的比特率。

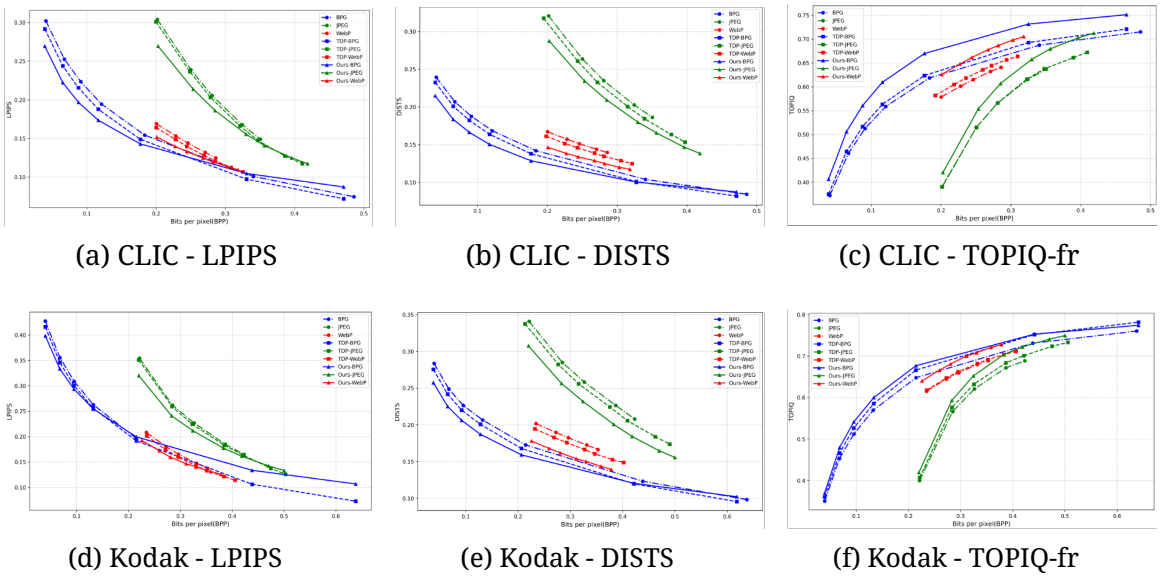
实验结果

本节将全面评估我们提出的生成式预处理框架。我们针对两种基线进行了广泛的定量和定性比较：未经过任何预处理的标准编解码器（称为锚点），以及一种先进的预处理方法 TDP [2]。

BDBR 性能。为量化性能提升，我们计算了三种广泛使用的感知指标（LPIPS、DISTS 和 TOPIQ-fr）下的 Bjontegaard Delta 比特率（BDBR）节省量。表 1 所示结果证明了我们方法的优越性。在 Kodak 和 CLIC 验证数据集上，与 TDP 基线相比，我们的方法在所有评估的编解码器上均实现了显著更高的比特率降低幅度。例如，搭配 BPG 编解码器时，我们的方法在 CLIC 和 Kodak 上分别实现了高达 **-27.68%** 和 **-30.13%** 的 DISTS BDBR 节省量，表现优于 TDP。这些幅度大且稳定的提升凸显了利用预训练生成模型进行面向感知的压缩预处理的巨大潜力。

率感知曲线。为进行更详细的分析，我们绘制了所有方法在两个数据集上的率感知（R-P）曲线，如图 3 所示。这些曲线直观展示了感知质量与比特率之间的权衡；曲线位置越靠左上方，性能越优。在关键的低至中等比特率范围内（例如，$bpp < 0.5$），对于所有指标和数据集，我们方法的曲线始终优于锚点和 TDP 基线。这表明，我们的生成式方法能够以相同甚至更低的比特率实现显著更高的感知质量，这与 BDBR 结果所体现的优势一致。

然而，我们观察到，在较高比特率（例如 $bpp > 0.5$）下，性能会出现交叉，此时我们的方法有时会逊于 TDP 或锚定编码器。这一现象是我们模型生成式本质所带来的固有权衡。由于我们的模型是在生成式基础模型上进行微调的，因此经过优化，能够合成合理且具有感知美感的纹理；与传统方法相比，这可能会更大程度地修改原始图像内容。在锚定编码器已经能够保留大部分细节的高比特率下，我们的方法所进行的修改



(a) CLIC - LPIPS (b) CLIC - DISTS (c) CLIC - TOPIQ-fr

(d) Kodak - LPIPS (e) Kodak - DISTS (f) Kodak - TOPIQ-fr

图 3：CLIC 验证数据集和 Kodak 数据集上的率感知（R-P）曲线。这些图比较了不同编解码器（JPEG、WebP、BPG）和预处理方法（TDP、本文方法）下感知质量（LPIPS、DISTS、TOPIQ-fr）与比特率（bpp）之间的关系。无论采用哪种指标或数据集，我们的方法都能以更低的比特率持续实现更高的感知质量。

虽然在感知上具有真实感，但可能与原始源图像存在更大偏差，从而受到 LPIPS 和 DISTS 等基于相似性的感知指标的惩罚。这种权衡——生成式方法通过创造逼真的细节，在低比特率范围内表现出色——是一个已知特征，并且已在此前的生成式压缩工作中得到观察 [20]。

定性比较 除了定量指标外，我们还提供了定性比较，以直观展示生成式预处理方法的优势。如图 4 所示，我们展示了来自 CLIC 验证数据集的一幅代表性图像中的放大区域，该图像已被压缩至相近的比特率。直接使用 WebP 压缩的图像呈现出明显的伪影，例如块效应和模糊，在纹理区域尤为突出。虽然 TDP 缓解了部分伪影，但往往会过度平滑精细细节，导致自然纹理丢失。相比之下，我们的方法能够有效去除压缩伪影，同时恢复并增强合理的高频细节。继承自预训练生成模型的丰富纹理合成能力，使我们的方法能够生成兼具锐利度与自然感、在视觉上令人愉悦的结果，展现出相较于其他方法的明显感知优势。

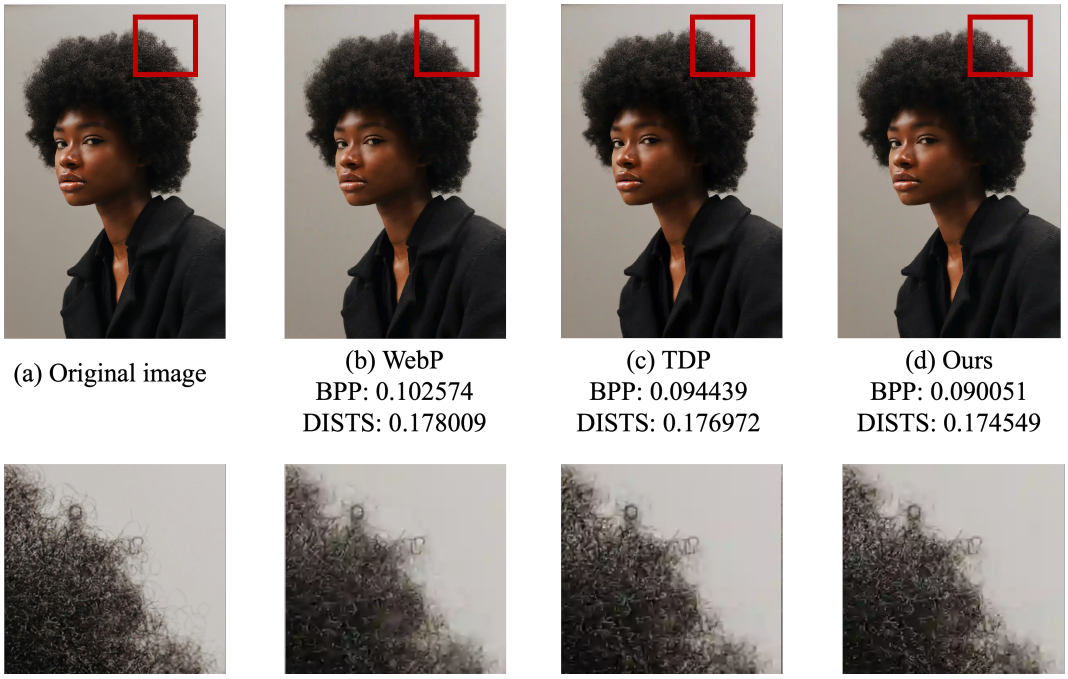


图 4：在 CLIC 验证数据集中的一个样本上进行的定性比较。与锚点 WebP 和 TDP 相比，我们的方法在有效去除压缩伪影的同时，保留并增强了细粒度纹理，从而获得更出色的视觉质量。

结论

本文提出了一种新颖的率失真-感知优化压缩预处理框架，首次成功将大规模预训练生成模型应用于该任务。我们的方法结合了高效的模型蒸馏与针对性的率失真-感知微调策略，有效弥合了生成模型与传统压缩流程之间的差距。实验表明，我们的方法能够提供更高的视觉质量，并以更优的感知评分实现显著的比特率节省。

参考文献

[1] F. Ye、B. Liu、L. Li 和 D. Liu，“面向 JPEG 压缩的率失真优化深度预处理”，*IEEE*《电路与系统视频技术汇刊》，2025年。[2] F. Zhao、M. Guo、S. Zhao、J. Li、L. Zhang 和 X. Xie，“用于用户生成内容视频压缩的三动态预处理框架”，载于*ICASSP 2024-2024 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2024年，第7865–7869页。[3]I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio，“生成对抗网络”，《*ACM*通讯》，第63卷，第11期，第139–144页，2020年。

[4]Y. Blau 和 T. Michaeli，“重新思考有损压缩：码率–失真–感知权衡”，载于国际机器学习会议。PMLR，2019年，第675–685页。[5] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，“使用潜在扩散模型的高分辨率图像合成”，2021年。[6] A. Ramesh、P. Dhariwal、A. Nichol、C. Chu 和 M. Chen，“利用CLIP潜变量进行分层文本条件图像生成”，*arXiv*预印本 *arXiv:2204.06125*，第1卷，第2期，第3页，2022年。[7] C. Saharia、W. Chan、S. Saxena、L. Li、J. Whang、E. L. Denton、K. Ghasemipour、R. Gontijo Lopes、B. Karagol Ayan、T. Salimans 等，“具有深度语言理解的照片级真实感文本生成图像扩散模型”，神经信息处理系统进展，第35卷，第36 479–36 494页，2022年。[8] W. Peebles 和 S. Xie，“使用Transformer的可扩展扩散模型”，*arXiv*预印本 *arXiv:2212.09748*，2022年。[9] Y. Wang、S. Zhao、K. Chen、K. Zhang、J. Li 和 L. Zhang，“Gendr：闪电生成式细节复原器”，*arXiv*预印本 *arXiv:2503.06790*，2025年。[10] F. Zhao、M. Guo、S. Zhao、J. Li、L. Zhang 和 X. Xie，“面向压缩环境下视频机器视觉的预处理框架”，载于*2024*年数据压缩会议（*DCC*）。IEEE，2024年，第312–321页。[11] A. Kendall、H. Martirosyan、S. Dasgupta、P. Henry、R. Kennedy、A. Bachrach 和 A. Bry，“端到端学习几何与上下文以进行深度立体回归”，载于*IEEE*国际计算机视觉会议论文集，2017年，第66–75页。[12] J. Ball´e、D. Minnen、S. Singh、S. J. Hwang 和 N. Johnston，“具有尺度超先验的变分图像压缩”，*arXiv*预印本 *arXiv:1802.01436*，2018年。[13]K. Ding、K. Ma、S. Wang 和 E. P. Simoncelli，“图像质量评价：统一结构相似性与纹理相似性”，$CoRR$，第abs/2004.07728卷，2020年。 [在线]。可访问：https://arxiv.org/abs/2004.07728[14] Y. Li、K. Zhang、J. Liang、J. Cao、C. Liu、R. Gong、Y. Zhang、H. Tang、Y. Liu、D. Demandolx 等，“LSDIR：用于图像复原的大规模数据集”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第1775–1787页。[15] T. Karras、S. Laine 和 T. Aila，“用于生成对抗网络的基于风格的生成器架构”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第4401–4410页。[16] B. Lim、S. Son、H. Kim、S. Nah 和 K. M. Lee，“用于单图像超分辨率的增强型深度残差网络”，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）研讨会，2017年7月。[17] J. Wu、L. Li、W. Dong、G. Shi、W. Lin 和 C.-C. J. Kuo，“针对模式复杂度图像的增强型恰可察觉差异模型”，*IEEE*图像处理汇刊，第26卷，第6期，第2682–2693页，2017年。[18] R. Zhang、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang，“深层特征作为感知度量的非凡有效性”，载于$CVPR$，2018年。[19] C. Chen、J. Mo、J. Hou、H. Wu、L. Liao、W. Sun、Q. Yan 和 W. Lin，“TOPIQ：一种从语义到失真的图像质量评价自顶向下方法”，*IEEE*图像处理汇刊，第33卷，第2404–2418页，2024年。[20] F. Mentzer、G. D. Toderici、M. Tschannen 和 E. Agustsson，“高保真生成式图像压缩”，神经信息处理系统进展，第33卷，第11 913–11 924页，2020年。