生成图像编码与扩散先验

Jianhui Chang
China Telecom Cloud Computing Research Institute
changjh1@chinatelecom.cn

摘要—随着生成技术的进步，视觉内容已经演变成自然图像与 AI 生成图像的复杂混合体，推动了对更高效编码技术的需求，这些技术优先考虑感知质量。传统编解码器和学习方法难以在高压缩比下保持主观质量，而现有的生成方法则面临视觉保真度和泛化能力方面的挑战。为此，我们提出了一种新的基于扩散先验的生成编码框架，以提高低比特率下的压缩性能。我们的方法采用预先优化的编码器来生成通用的压缩域表示，并通过轻量级适配器和注意力融合模块与预训练模型的内部特征相结合。该框架有效地利用了现有的预训练扩散模型，并能以极小的再训练成本高效地适应不同的预训练模型以满足新的需求。我们还引入了一种分布重新归一化方法，进一步提高重构保真度。广泛的实验表明，我们的方法：(1) 在低比特率下在视觉保真度方面优于现有方法；(2) 相比 H.266/VVC，压缩性能提高了高达 79%；(3) 为 AI 生成的内容提供了一种高效解决方案，并且能够适应更广泛的内容类型。

***Index Terms*—生成编码、图像压缩、潜扩散模型、AI 生成内容**

I. 介绍

随着生成技术的快速发展，视觉内容已经从主要是自然图像演变为自然图像和生成图像的多样化混合体，推动了对更高效图像编码技术的需求，以优化主观感知质量。基于变换的传统编码标准（例如，HEVC [1]，VVC [2]）和端到端学习的编码方法 [3], [4] 在保持高压缩比下的主观重建质量方面面临挑战，主要是由于数据退化以及外部先验知识利用不足。

近期的生成图像编码方法 [5], [6] 利用了生成模型捕捉复杂数据分布的强大能力，能够在超低比特率下实现高质量的重建。尽管前景广阔，当前主要基于 GANs [7] 的技术受限于其生成能力，特别是在保留现实纹理和结构细节方面。此外，目前的生成编码方法 [6], [8] 的泛化能力有限，因为压缩性能与模型的训练范围紧密相关。这些方法通常需要对特定场景数据（如例如，AI 生成内容或自然场景）进行端到端的率失真（RD）优化，使得训练资源消耗巨大。另外，在不同场景下往往需要重新训练整个模型，这会导致显著的开销并在设备间解码时带来挑战。

扩散模型最近在图像生成方面取得了显著的成功，特别是在文本到图像的任务中。它们能够产生视觉丰富且结构良好的图像，展示了从低级纹理到高级语义的大量关于视觉内容构建的先验知识。基于扩散的压缩领域的现有工作遵循不同的设计范式。一类研究 [9], [10] 利用扩散模型作为现有编解码器之上的后处理器，重点在于提高解码图像的质量而不是生成编码。另一类研究 [11], [12] 则使用以图像衍生表示为条件的扩散模型作为解码器。例如，文本条件的扩散解码器 [11], [12] 已被探索用于异常低比特率（<0.01 bpp）的情况，在保持高级语义一致性方面表现出色，但往往牺牲了视觉保真度。此外，同时进行的工作 PerCo [12] 利用向量量化代码簿来表示图像及其文本信息。然而，这些方法通常需要单独训练辅助编码器和不同大小的代码簿，并且需要对扩散模型进行微调以适应不同的目标速率和内容，限制了它们在适应各种压缩需求方面的灵活性和效率。

为此，本文提出了一种生成编码框架，利用强大的扩散先验来提高压缩性能，主要集中在改善低码率范围内的感知保真度，通常在 0.01 到 0.2 bpp 左右。我们使用预训练的潜扩散模型 [13] 作为生成解码器，同时通过一个与特定生成模型无关的前置端到端学习压缩任务来优化编码器。引入了一个轻量级适配器以确保有效的领域适应性，提供了一种灵活且兼容的解决方案，适用于不同的模型。为进一步提高重构保真度和压缩效率，我们采用跨注意力机制来更好地对齐压缩潜变量与内部特征，并使用分布归一化方法来减轻重构失真。本工作的主要贡献如下：

- 我们提出了一种生成编码框架，利用强大的扩散先

This work is partially sponsored by the National Key Research and Development Program of China (2024YFB4505704).

验来实现高效的以人类为中心的压缩。该框架包括一个轻量级适配器和一个注意力融合模块，能够有效进行领域适应，并确保与各种预训练潜在扩散模型兼容。

- 我们引入了一种交叉注意力机制，以优化压缩潜变量与预训练模型内部特征的融合，并采用一种分布重新归一化方法来提高重建保真度，这两种方法都进一步提升了压缩性能。
- 大量实验表明，我们的方法在压缩性能上比 VVC 提高了多达 79%，并在自然场景和使用不同预训练扩散模型的 AI 生成场景中展现了广泛适用性。

II. 提出的方法

所提出的生成式图像编码方法的主要框架如图 1所示。它包括几个关键组件：(1) 图像编码器 $\mathcal{E}$ 和熵模型 $\mathcal{H}$，它们协作将图像编码为具有速率约束的潜变量；(2) 配备了注意力融合模块的潜变量适配器 $\mathcal{F}$，其任务是将压缩后的潜变量转换为控制信号；以及 (3) 潜变量扩散模型 [13](LDM)$\mathcal{G}$，它作为真实图像重建的强大先验。我们的目标是利用预训练模型 $\mathcal{G}$ 的能力，通过轻量级的 $\mathcal{F}$ 和注意力融合将压缩潜变量与其内部先验相结合，从而实现通过压缩信号对重建过程进行准确引导。

在编码端，为了确保灵活性和通用性，使用预优化的编码器 $\mathcal{E}$ 将输入图像 $\boldsymbol{x}$ 转换为统一的压缩域潜在表示 $\boldsymbol{y}$，定义为 $\boldsymbol{y} = \mathcal{E}(\boldsymbol{x})$。转换后，潜在变量 $\boldsymbol{y}$ 经量化和熵编码生成比特流。在解码器一侧，解码的潜变量 $\hat{\boldsymbol{y}}$ 经过潜变量适配器 $\mathcal{F}$ 转换为一个特征集 $\boldsymbol{f}$。这个集合然后与去噪 U-Net 在潜扩散模型中的每个时间步 t 的中间表示 $\boldsymbol{c}_t$ 对齐并融合。这种编码信息的整合引导了生成过程，使得能够合成高保真度的重构图像 $\hat{\boldsymbol{x}}$。在潜扩散模型中，生成过程从先验高斯分布 $\mathcal{N}(0,1)$ 中采样的随机噪声 $\boldsymbol{z}$ 开始，并以压缩信息 $\hat{\boldsymbol{y}}$ 为条件，表示为：

$$\hat{\boldsymbol{x}} = \mathcal{G}(\boldsymbol{z}, \hat{\boldsymbol{y}}), \hat{\boldsymbol{x}} \sim P(\hat{\boldsymbol{x}} \mid \hat{\boldsymbol{y}}). \tag{1}$$

由于压缩信息 $\hat{\boldsymbol{y}}$ 提供了强烈的空间指导，目标分布 $P(\hat{\boldsymbol{x}})$ 主要由 $\hat{\boldsymbol{y}}$ 和内部扩散先验决定，使我们能够实现压缩任务的高保真重建结果。

A. 潜隐适配器和注意融合

为了将预训练模型适应计算机视觉中的不同任务或领域，人们探索了多种方法，尽管这些方法可能并不完全适合图像压缩。例如，ControlNet [14] 使用一个可训练的大型预训练主干副本生成额外的特征图，这会导致显著的计算开销。另一个代表性方法 T2I-Adapter [15]，控制生成图像的颜色和结构，但不能提供实现我们方法所需的高感知重构保真度所必需的粒度。

为了解决这些挑战，本文设计了一个潜在适配器和一个注意潜在融合模块以实现高视觉保真度的生成编码。该适配器将压缩域潜变量 $\hat{\boldsymbol{y}}$ 与 U-Net 的下采样和上采样模块中的特征对齐，定义为 $\boldsymbol{f} = \mathcal{F}(\hat{\boldsymbol{y}})$。每个特征提取模块包括一个卷积层和两个残差模块，在每一时间步 t 产生与相应的 U-Net 特征 $c_t^{(i)}$ 对齐的特征 f_i。

变换的潜码 $\boldsymbol{f}$ 和内部特征 $\boldsymbol{c}_t$ 然后被融合并整合到生成扩散过程中，用于图像重建。常用的加性融合方法 [14], [15] 假设空间对齐并忽略了上下文信息，限制了重建能力。因此，我们引入了一个注意力融合模块来提高融合精度，基于空间交叉注意力提升图像重建的保真度。如图 2所示，U-Net 中间特征 $c_t^{(i)}$ 和变换后的潜码 f_i 相加形成基础特征 $\hat{c}_t^{(i)} = c_t^{(i)} + f_i$，增强了空间相关性。从压缩信息中得到的转换潜变量 f_i 作为上下文向量。通过将它们的维度重塑为 $HW \times C$，我们实现了空间维度的线性化，使得每个特征向量都能够独立地对注意力机制做出贡献。三个 1×1 卷积层将 $\hat{c}_t^{(i)}$ 映射到 $\boldsymbol{Q}$，并将 f_i 重塑为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。用 FC 表示一个全连接层，融合特征 $\hat{\boldsymbol{f}}_i$ 计算如下：

$$\hat{\boldsymbol{f}}_i = \boldsymbol{V} + FC(Attn(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})), \tag{2}$$

$$Attn(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{C}}\right) \cdot \boldsymbol{V}. \tag{3}$$

交叉注意力机制通过捕获压缩信息与生成模型在每个空间位置的内部特征之间的相关性来促进融合。因此，它使潜在扩散模型能够获得源自目标图像的变换潜在变量更精确的空间细节，从而有效提高重建图像的准确性。

B. 优化策略

预训练的生成模型通常为视觉重建提供强大的先验信息，但重新训练它们成本高昂且有损这些先验信息的风险。为了避免这种情况，我们保持预训练的潜在扩散模型 $\mathcal{G}$ 冻结，并采用两阶段优化策略：首先通过一个辅助任务优化编码器 $\mathcal{E}$ 和熵估计模型 $\mathcal{H}$，然后冻结它们并对轻量级适配器 $\mathcal{F}$ 和注意力融合模块进行微调。

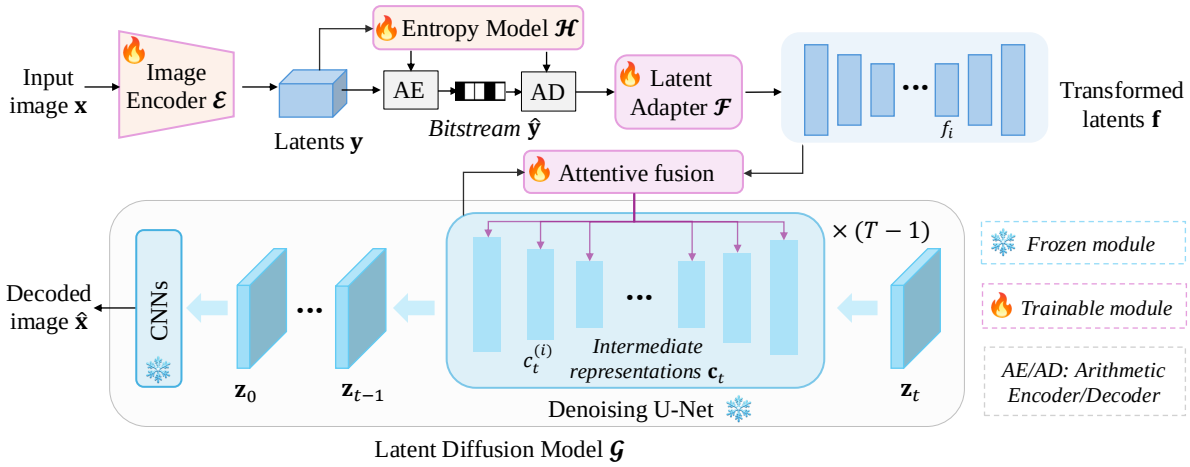


图 1: 提出的生成式图像编码框架概述。

(1) **编码器优化**。在第一阶段，使用了一个学习的图像压缩任务 [16] 进行预训练优化。这包括一个编码器 $\mathcal{E}$，一个辅助解码器和一个熵模型 $\mathcal{H}$。这个辅助解码器专门为了预训练优化而引入，在优化完成后将会被丢弃。为了实现统一的编码器和熵模型下的可变速率编码，我们采用了通道缩放和量化技术，具体细节见 [17]。对于率失真优化 [18], [19]，每个拉格朗日乘子 λ_s 都与一组速率级别 s 的量化参数相关联。速率约束 $\mathcal{L}^s_{rate}$ 由 $\mathcal{H}$ 确定。使用 MS-SSIM [20] 来衡量重构失真 $\mathcal{L}^s_{dist}$，预训练优化目标定义如下：

$$\mathcal{L}_{pretext} = \sum_{s=0}^{L-1} \mathcal{L}^s_{rate} + \lambda_s \mathcal{L}^s_{dist}. \tag{4}$$

在使用方程 (4) 进行训练时，辅助解码器通过随机梯度下降与 $\mathcal{E}$ 和 $\mathcal{H}$ 一起更新。预训练优化完成后，仅保留并固定了 $\mathcal{E}$ 和 $\mathcal{H}$ 的参数，以便在后续与其他模块一起进行训练时保持不变。

(2) **适配器优化**。第一阶段训练后，第二阶段旨在优化潜在适配器 $\mathcal{F}$ 以及注意融合模块，并保持 $\mathcal{E}$、$\mathcal{H}$ 和 $\mathcal{G}$ 固定。在扩散过程中，输入图像 $\boldsymbol{x}$ 的潜变量 $\boldsymbol{z}_0$ 在每个时间步长 t 中都会被添加噪声以产生 $\boldsymbol{z}_t$。逆向扩散过程通过使用 U-Net 噪声估计器 ϵ_θ 对 $\boldsymbol{z}_t$ 进行去噪，迭代地执行图像重建，条件是转换后的压缩特征集 $\boldsymbol{f}$。去噪损失函数定义如下：

$$\mathcal{L}_{adp} = \mathbb{E}_{\boldsymbol{z}_0,t,\boldsymbol{f},\epsilon\sim\mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{f})\|_2^2 \right]. \tag{5}$$

在训练阶段，输入图像由 $\mathcal{E}$ 编码，传输，并通过 $\mathcal{F}$ 转换为 $\boldsymbol{f}$，然后通过融合模块整合到去噪过程中。保持其他模块固定不变，$\mathcal{F}$ 和融合模块进行微调，使预训练模型 $\mathcal{G}$ 在完成后能够用于图像压缩任务。

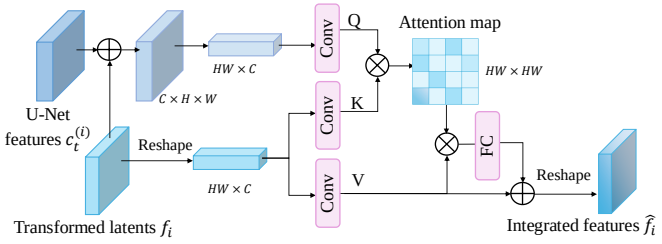


图 2: 基于空间交叉注意力的注意潜在融合。

C. 保真度增强

提出的生成编码方法成功重建了图像，同时保持了一致的视觉语义和结构。然而，颜色分布中的失真突显了提高保真的需求。在图像处理中，颜色矩——均值 (μ) 和标准差 (σ) ——是表征颜色分布的基本指标，其中均值捕捉亮度，而标准差通过像素值的变化反映对比度。风格迁移研究 [21] 表明，在 DNN 特征空间中调整实例归一化可以保留内容同时转移颜色样式。扩展这一见解，我们提出了一种方法来校正合成图像像素域的颜色偏差，通过对其通道的均值和标准差与原图进行对齐，从而提高重建保真度。

具体来说，为了最小化重建失真，重构图像的颜色分布参数应与原始图像的相一致。均值 (μ) 和标准差 (σ) 对每个通道进行计算，分别为 $\mu = \mu_R, \mu_G, \mu_B$ 和 $\sigma = \sigma_R, \sigma_G, \sigma_B$。为了高效传输，这些参数被量化为 $\hat{\mu}$ 和 $\hat{\sigma}$，步长为 Δ。在解码端，初始重建 $\hat{\boldsymbol{x}}$ 通过使用传输参数 $\hat{\mu}$ 和 $\hat{\sigma}$ 进行归一化以增强保真度，如下所示：

$$\hat{\boldsymbol{x}}_{norm} = \left(\frac{\hat{\boldsymbol{x}} - \hat{\mu}}{\hat{\sigma}}\right)\sigma + \mu, \tag{6}$$

其中 $\hat{\boldsymbol{x}}_{norm}$ 表示在提高颜色保真度后的最终解码图像。所提出的方法显著提高了重建图像的颜色准确性。此外，我们发现使用来自较小区域的统计信息可以导致更

图 3: VVC、Cheng-VBR、TCM-VBR、HiFiC、PerCo 和我们方法在 AIGC 数据集上的定性比较。

准确的颜色校正。该方法进一步扩展到块级校正，在此过程中统计参数被传输并应用于每个图像块，确保了几乎不增加比特率成本的情况下实现更精确的颜色对齐。

III. 实验

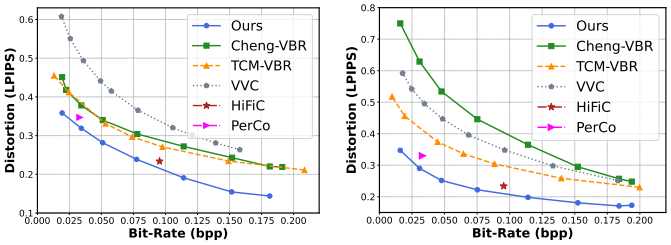
A. 实验设置

a) 数据集: 人工智能生成内容（AIGC）最近已成为关键的人工智能研究焦点。AIGC 数据的统计与生成模型先验很好地对齐，使其非常适合用于生成编码。因此，本研究利用了包含 1400 万张高质量、主要是动漫风格图像的 DiffusionDB 数据集 [22]，这些图像是从真实用户提示生成的。在实验中，随机选择了 50,634 张图像进行训练，并选择 200 张用于测试，分辨率为 512×512。此外，为了评估该方法与其他预训练扩散模型在自然场景中的兼容性，使用了 OpenImages 数据集 [1] 的 10 万张图像来微调潜在适配器。使用 Kodak 数据集 [2] 来评估真实世界图像的压缩性能。

b) 实现细节: 所有实验均使用 PyTorch 在四块 NVIDIA Tesla 32G-V100 GPU 上进行。采用 Adam 优化器，学习率为 $1e-5$。第一阶段模型遵循 Cheng 等人 [16] 的方法，具有 128 个潜在通道，在 600 个周期内训练，并使用拉格朗日乘子{50.0, 16.0, 3.0, 1.0, 0.5, 0.25, 0.1, 0.05, 0.01, 0.005}和大小为 96 的批处理，从预训练的 CompressAI[3] 参数开始。预训练的潜在扩散模型源自 Stable Diffusion v1.4[4]，输入为 512×512，潜在维度为 $4 \times 64 \times 64$。跨比特率共享的潜在适配器网络和融合模块在大小为 8 的批处理下进行了 10 个周期的微调。在推理过程中，图像使用 DDIM 确定性采样计划 [23] 生成，采样步骤为 10 步。测试结果是使用全局随机种子 42 获得的，确保了去噪过程中的潜在噪声 z 的确定性初始化。改变随机种子会产生一组重建图像，这些图像可能反映出特定压缩潜变量 [12] 内在的不确定性。LPIPS 值在不同随机初始化下的标准差在我们的结果中经验上小于 0.01，表明其对重建结果的影响可以忽略不计。为了提高保真度，使用了大小为 64×64 的块，并将参数量化为 6 位。编码这些参数平均增加了 0.01bpp。

c) 评估指标: 传统编码方法通常侧重于保持信号保真度，使用如 PSNR 和 SSIM 等指标来评估像素级失真。然而，这些指标往往难以反映感知质量。同样，数据集级别的度量标准如 FID 和 KID 评估分布差异但

[1] https://storage.googleapis.com/openimages/web/index.html
[2] https://r0k.us/graphics/kodak/
[3] https://interdigitalinc.github.io/CompressAI/
[4] https://huggingface.co/CompVis/stable-diffusion-v1-4



(a) 扩散数据库上的 R-D 结果。 (b) R-D 结果在柯达图像上。

图 4: Kodak 数据集和 AIGC 数据集，DiffusionDB 上的 R-D 比较。较低的 LPIPS 分数表示更高的保真度。

不足以评估图像级别上的感知保真度。因此，我们采用 Learned Perceptual Image Patch Similarity (LPIPS) [24]，该方法测量配对图像之间的特征域失真，更好地与人类感知相一致，并且在领域内被广泛认可 [5], [6]。此外，采用每像素比特 (bpp) 来评估速率性能。

B. 压缩性能比较

a) 对比方法: 所提出的方法与几种代表性基线进行了比较：(1)**VVC**，多功能视频编码标准 [2]，使用 VTM-11.0 进行内编码，在常见测试条件下；(2) **程-VBR**，一种基于深度学习的端到端压缩方法 [16]，在本研究中用于预训练优化；(3)**TCM-VBR**，最先进的方法之一 [4]，利用可学习量化尺度矩阵和与本工作类似的培训策略进行可变比特率编码；以及 (4) **高保真压缩**，一种生成压缩方法 [5]，在固定低比特率下训练。(5) **每核** [5] [12]，最新的基于扩散的生成编码方法之一。发布的检查点使用了 Stable Diffusion v2.1，比我们的更为先进。

b) 定量评估: 图 4表明，所提出的生成编码方法在使用 LPIPS 作为度量标准的 RD 性能方面显著优于 VVC、Cheng-VBR [16]、HiFiC、PerCo 和 TCM-VBR [4]。表 I进一步通过 Bjontegaard 度量标准 [25] 定量化了这一改进，显示与 VVC 相比比特率降低了 67.8%，与 Cheng-VBR 相比降低了 50.1%，与 TCM-VBR 相比在 AIGC 数据集上降低了 40.1%。就 BD-LPIPS 而言，所提出的方法提升了 27.0% 的重建质量超过 Cheng-VBR，提升了 24.5% 超过 TCM-VBR，并且提升了 37.8% 超过 VVC。TCM-VBR 在其变换和熵模型中结合了 CNN 和变压器以增强局部和全局特征捕捉，实现了比 Cheng-VBR 和 VVC 更好的 RD 性能。HiFiC 的性能受到 GANs 不佳的生成能力的限制，而 PerCo 引入了不有助于提

[5]https://github.com/Nikolai10/PerCo

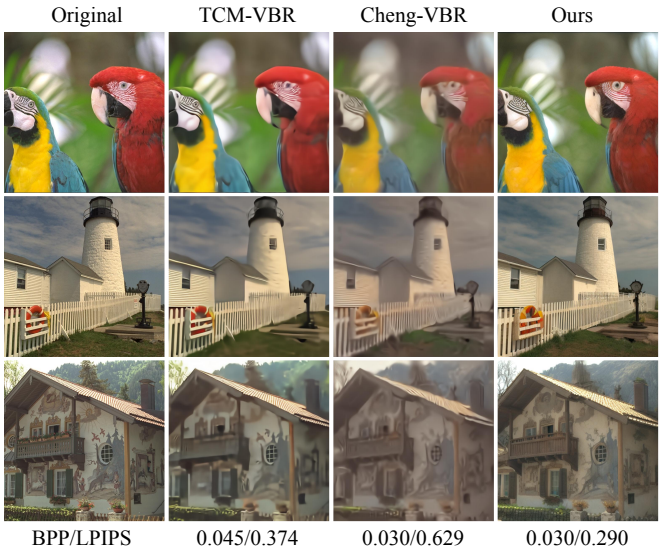


图 5: Kodak 数据集上的定性比较。

表 I: BD-Rate 和 BD-度量指标对于 VVC、Cheng-VBR 和 TCM-VBR 的结果。LPIPS 被用作失真度量。

Dataset	Metric	VVC	Cheng-VBR	TCM-VBR
人工智能生成内容	BD-Rate	-67.75%	-50.14%	-40.08%
	BD-LPIPS	-37.76%	-27.03%	-24.46%
柯达	BD-Rate	-79.24%	-81.62%	-69.93%
	BD-LPIPS	-39.42%	-45.80%	-29.34%

高视觉保真度的文本信息。相比之下，所提出的方法利用强大的扩散先验和注意融合机制，使从紧凑潜在表示中获得更感知真实的重建成为可能。这些结果突出了该方法在广泛压缩比范围（100× 至 2000×）内的强大 RD 性能，证明了其有效性。

c) 定性评估: 图 3展示了所提出的生成压缩方法与 VVC、Cheng-VBR [16]、TCM-VBR [4]、HiFiC [5] 和 PerCo [12] 的主观重建结果。VVC 表现出明显的块效应，而 TCM-VBR 和 Cheng-VBR 则受到过度平滑和模糊的影响。HiFiC 引入了显著的生成失真。特别是，像头发、草、叶子和山脉这样的纹理显得模糊，其他方法未能保持锐利边缘，导致视觉质量下降。虽然 PerCo 产生了看起来真实的图像结果，但它保留细粒度细节（例如、颜色、边缘、形状）的能力有限。相比之下，所提出的方法利用生成扩散先验不仅能够产生清晰且逼真的纹理边缘，还能在重建中实现更高的视觉保真度。这些结果显示了其优越的视觉质量，验证了它提高以人类为中心的压缩性能的有效性。

C. 泛化能力评估

本研究中使用的 LDM 以生成视觉艺术品而闻名，使得所提出的生成编码方法特别适用于 AIGC 艺术作品的编码。为了评估该方法与各种预训练模型的兼容

表 II: 所提方法模块和 VVC 参考软件的复杂性分析。

Method	VVC		我们的		
Module	编码器	Decoder	编码器	适配器	LDM
Inference Time (ms)	3940.6	187.0	59.2	2.6	76.5

性，我们应用了另一个来自 Civitai[6] 的预训练扩散模型 Realistic Vision V6.0 来压缩一般摄影内容。我们仅对潜在适配器 $\mathcal{F}$ 和相关融合模块进行了微调，使用方程 (5) 中的损失函数进行 2 个周期的训练，同时保持所有其他模块固定。所提出的方法主要与 Cheng-VBR [16] 和 TCM-VBR [4] 在 Kodak 数据集上进行了比较。值得注意的是，所提出的方法和 Cheng-VBR 使用了相同的编码器。

如图 4(b) 和表 I所示，所提出的方法显著优于 HiFiC、PerCo、Cheng-VBR 和 TCM-VBR，在 BD-rate 自然场景图像方面分别比 Cheng-VBR 和 TCM-VBR 提高了 81.6%和 69.9%的压缩效率。此外，我们的方法在 BD-LPIPS 测量下比 Cheng-VBR 和 TCM-VBR 分别提升了 45.8%和 29.3%的重建质量。对于图 5所示的 800× 压缩比率下的主观质量，端到端学习编码方法表现出显著的模糊和失真。相比之下，所提出的方法保持了高视觉保真度，生成了清晰、逼真的纹理。这些结果验证了所提出的生成编码方法在确保与各种预训练扩散模型兼容并在不同场景下实现强大性能方面的有效性。

D. 复杂度分析

表 II提供了所提出方法每个模块的推理时间，LDM 反映了单步推理。作为参考，也提供了 VVC 参考软件 VTM-11.0 在 CPU 平台上对 512×512 分辨率图像编码和解码的时间。值得注意的是，适配器对总体推理时间几乎没有贡献。因此，为不同的预训练模型微调一个轻量级适配器引入的计算开销可以忽略不计，提供了将各种生成扩散先验纳入其中的一种可适应且计算效率高的解决方案。我们方法报告的压缩性能基于 10 次 LDM 推理迭代。尽管更多迭代会增加推理时间，但加速采样研究表明减少复杂性和步骤数量也能产生质量图像，从而实现实用的生成编码部署。

IV. 结论

本文提出了一种生成编码框架，在低码率下实现了高质量的感知效果。通过利用预优化编码器、轻量级适配器和融合模块中的扩散先验，我们的方法确保了与各种预先训练好的扩散模型的兼容性。注意力特征融合和分布重归一化的结合进一步提升了重建保真度，提高了压缩效率。实验结果表明，所提出的方法比 H.266/VVC 高出 79%，展示了其在自然内容和 AI 生成内容上的有效性和灵活性。这些发现突显了该方法作为 AI 生成内容的有效解决方案及其在各种生成编码应用中的灵活应用潜力。

[6]https://civitai.com/models/4201/realistic-vision-v60-b1

致谢

我们感谢贾武、刘宏斌、杨浩和马思伟的深入讨论和计算支持。

参考文献

[1] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[3] Johannes Ballé, Philip A Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici, "Nonlinear Transform Coding," *IEEE Journal of Selected Topics in Signal Processing*, pp. 339–353, 2020.

[4] Jinming Liu, Heming Sun, and Jiro Katto, "Learned Image Compression with Mixed Transformer-CNN Architectures," in *CVPR*, 2023, pp. 14388–14397.

[5] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," in *NeurIPS*, 2020, pp. 11913–11924.

[6] Jianhui Chang, Jian Zhang, Jiguo Li, Shiqi Wang, Qi Mao, Chuanmin Jia, Siwei Ma, and Wen Gao, "Semantic-aware visual decomposition for image coding," *International Journal of Computer Vision*, vol. 131, no. 9, pp. 2333–2355, 2023.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in *NeurIPS*, 2014, pp. 2672–2680.

[8] Jianhui Chang, Zhenghui Zhao, Chuanmin Jia, Shiqi Wang, Lingbo Yang, Qi Mao, Jian Zhang, and Siwei Ma, "Conceptual Compression via Deep Structure and Texture Synthesis," *IEEE Transactions on Image Processing*, vol. 31, pp. 2809–2823, 2022.

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

[10] Noor Fathima Ghouse, Jens Petersen, Auke Wiggers, Tianlin Xu, and Guillaume Sautière, "A Residual Diffusion Model for High Perceptual Quality Codec Augmentation," *arXiv preprint arXiv:2301.05489*, 2023.

[11] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti, "Text+Sketch: Image Compression at Ultra Low Rates," in *ICML 2023 Workshop on Neural Compression*, 2023.

[12] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *ICLR*, 2023.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[15] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI*, 2024, vol. 38, pp. 4296–4304.

[16] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules," in *CVPR*, 2020, pp. 7939–7948.

[17] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai, "Asymmetric Gained Deep Image Compression with Continuous Rate Adaptation," in *CVPR*, 2021, pp. 10532–10541.

[18] Claude Elwood Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[19] Johannes Ballé, Valero Laparra, and Eero Simoncelli, "End-to-end optimized image compression," in *ICLR*, 2017.

[20] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. IEEE, 2003, vol. 2, pp. 1398–1402.

[21] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[22] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau, "DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models," in *ACL*, 2023, pp. 893–911.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," in *ICLR*, 2020.

[24] 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, pp. 586–595.

[25] Bjontegaard, Gisle, "Calculation of average PSNR differences between RD-curves," *ITU-T VCEG-M33*, 2001.