

基于世界模型因果推断的图像压缩方法^{*}

匡浩玮¹, 杨文瀚², 郭宗明¹, 刘家瑛¹



¹(北京大学 王选计算机研究所, 北京 100080)

²(鹏城实验室, 广东 深圳 518055)

通讯作者: 刘家瑛, E-mail: liujiaying@pku.edu.cn

摘要: 图像压缩的目标是在有限码率约束下尽可能保留图像信息. 现有方法或将图像视为像素的组合进行率失真优化, 或借助生成模型先验合成感知质量良好的图像, 但均未对图像内部因果关联结构进行显式建模, 导致图像内部丰富的因果信息未能被有效利用. 本文提出一种基于世界模型因果推断的图像压缩方法, 将自然图像建模为具有严密因果结构约束的世界场景切片, 将承载场景因果关联结构的关键区域定义为因果锚点, 即场景的世界基元, 并从信息论视角分析将码率优先分配给因果锚点的理论最优性, 证明世界模型的预测能力越强, 向因果锚点倾斜分配码率所带来的压缩增益越显著. 基于此, 本文设计了以世界模型为推理引擎的编解码框架: 编码端通过链式推理定位因果锚点, 构建非均匀码率分配策略; 解码端以因果锚点为起点, 借助世界模型的因果推理能力完成高质量场景重建. 实验结果表明, 本方法在极低码率区间的各项指标上显著优于现有方法.

关键词: 图像压缩; 世界模型; 因果推断; 码率分配

中图法分类号: TP391

中文引用格式: 匡浩玮, 杨文瀚, 郭宗明, 刘家瑛. 基于世界模型因果推断的图像压缩方法. 软件学报. <http://www.jos.org.cn/1000-9825/7759.htm>

英文引用格式: Kuang HW, YANG WH, GUO ZM, Liu JY. Causal Inference-driven Image Compression via World Models. Ruan Jian Xue Bao/Journal of Software (in Chinese). <http://www.jos.org.cn/1000-9825/7759.htm>

Causal Inference-driven Image Compression via World Models

KUANG Hao-Wei, YANG Wen-Han, GUO Zong-Ming, LIU Jia-Ying

¹(Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China)

²(Pengcheng Laboratory, Shenzhen 518055, China)

Abstract: Image compression fundamentally seeks to maximize information preservation under limited bitrate constraints. Existing methods either treat images as collections of pixels for rate-distortion optimization, or leverage generative model priors to synthesize perceptually pleasing reconstructions, yet neither explicitly models the causal dependency structure underlying image regions, leaving rich intra-image causal information unexploited. This paper proposes a world-model-driven image compression framework that models natural images as structured slices of the physical world governed by causal constraints. We define causal anchors — the primitive world elements encoding the scene's causal dependency structure — and provide an information-theoretic justification for prioritizing bitrate allocation toward these anchors, proving that stronger world model predictive capacity yields greater compression gains from anchor-biased rate allocation. The encoder performs hierarchical causal decomposition via chain-of-thought reasoning to precisely localize causal anchors and construct a non-uniform rate allocation mask; the decoder takes causal anchors as the logical seed and leverages the world model's causal inference capability to perform chain-based scene reassembly for high-fidelity reconstruction. Experiments demonstrate that the proposed method substantially outperforms existing approaches across all metrics at ultra-low bitrates.

Key words: image compression; world model; causal inference; rate allocation

图像压缩是多媒体通信与存储领域的基础问题, 其目标是在有限码率约束下尽可能保留图像信息. 随着移

^{*} 基金项目: 北京市科技计划课题(Z251100008425023); 鹏城实验室交叉前沿研究项目(PCL2025QYB013)

收稿时间: 2026-04-27; 修改时间: 2026-06-18; 采用时间: 2026-08-20; jos 在线出版时间: 2026-09-14

移动互联网与流媒体的高速发展,图像传输对更低码率下高质量重建的需求日益迫切,推动了图像压缩技术的持续演进.过去几十年间,以 JPEG^[1]、BPG^[2]为代表的传统方法基于离散余弦变换与熵编码,在行业中广泛应用.随着深度学习技术的迅猛发展,基于神经网络的端到端可学习图像压缩方法被研究者们广泛探索.Ballé等人^[3]提出了端到端图像压缩框架,以可训练的深度网络替代手工设计的变换与熵编码模块,在率失真性能上取得了明显提升.Minnen等人^[4]进一步引入自回归上下文模型,将隐变量的概率分布建模为已解码隐变量的条件分布,大幅提升了熵编码效率.此后,Transformer架构的引入进一步提高了压缩性能,Liu等人^[5]提出的混合Transformer-CNN架构成为后续工作广泛采用的压缩骨干.然而,上述方法均以率失真为核心优化目标,将图像视为像素信号进行端到端编解码,未能对图像内部区域间的结构关联关系进行显式建模,在极低码率下感知质量下降明显.

为突破率失真框架在极低码率下的感知质量瓶颈,生成式图像压缩方法应运而生.其核心思路是:在极低码率下无法精确重建所有像素,因而转向利用生成先验补全丢失的信息,将解码过程重新定义为有条件的图像生成问题.Agustsson等人^[6]率先将生成对抗网络引入图像压缩,在极低码率下合成感知质量良好的重建图像.之后,Careil等人^[7]以扩散模型为解码器,以极少量比特传输语义条件,在 0.1 bpp 以下展现出超越传统方法的感知质量;Jia等人^[8]进一步引入全局语义条件引导扩散模型生成,提升了重建图像的语义一致性.上述方法虽取得了较好的感知质量,但其生成先验是统计意义上的图像分布,描述的是"这类场景通常长什么样",而非"这张图像的内部区域之间存在怎样的物理约束".因此,生成式方法在重建结果中常出现光照方向不一致、透视关系错误等问题,在高保真需求的应用场景中难以满足要求.此外,生成式方法继承了均匀码率分配的假设,对所有区域一视同仁,未区分哪些信息应精确保留、哪些信息可由先验推断.

上述问题在于现有方法对图像的建模假设存在局限.值得注意的是,人类视觉系统对图像的理解方式与此并不一致.认知神经科学研究表明,大脑并不被动地接收光信号,而是持续地对世界状态进行预测,并用感知输入修正预测误差.在 Friston^[9]提出的自由能原理框架下,感知是一种贝叶斯推断过程:大脑维护关于外部世界的内部生成模型,通过最小化预测误差(自由能)来推断感知输入的成因.海马体研究进一步表明^{[10][11]}海马体支持场景的主动构建而非被动存储:人类在记忆场景时会主动向边界外延伸,形成超出实际感知范围的扩展内部表示,并能将场景细节整合为连贯的空间表示加以重建.受上述研究启发,本文将这一机制类比为一种因果感知的信息压缩——优先保留对场景逻辑具有较大作用的关键信息,其余内容则借助内部世界模型推断.

本文将图像建模为一个具有因果结构约束的世界场景切片,而非像素的随机集合.图像内部各区域通过物理规律相互关联,例如光源位置决定阴影走向,消失点约束透视关系等.这种区域间的因果关联意味着,场景中存在少量对全局影响较大的关键区域,即世界基元,亦称为世界因果关系的锚点.其余区域的信息在锚点已知条件下高度可预测,构成关联残差.以图 1 输入图像所示的室内场景为例,灯的光源位置与白色方块的轮廓边缘,是对全场景具有约束力的因果锚点:光源决定了场景中所有物体的阴影方向与明暗分布,方块的几何轮廓支配了其在地面的遮挡关系;而地面的阴影走向、墙面的光晕渐变等,则是在锚点确定后高度可推断的关联残差.均匀码率分配未考虑区域间的信息不对称性,对可推断区域分配了不必要的码率;而若能将有限码率更多分配于因果锚点,并借助世界模型从锚点推断其余内容,有助于在极低码率下获得物理合理重建结果.

世界模型的兴起与演进为上述思路提供了实现基础.世界模型泛指能够对物理世界状态进行建模与预测的生成模型^[12],近年来随着大规模多模态预训练的发展,逐渐具备了跨模态语义理解与高质量图像生成的综合能力.与普通生成模型仅依赖统计相关性不同,世界模型能够建立区域间的因果关联关系,从局部的因果锚点出发推断全局场景的合理状态,这种因果推断能力是普通生成模型所不具备的.将世界模型引入图像压缩,使得解码端从统计分布采样转变为以因果锚点为条件的推断,有助于提升重建结果的合理性.

基于上述动机,本文提出一种基于世界模型因果推断的图像压缩方法.首先从信息论视角建立因果分解下的码率分配理论,将图像信息严格分解为因果锚点信息与关联残差信息,证明世界模型的预测能力越强,向因果锚点倾斜分配码率所带来的压缩增益越显著,为非均匀码率分配策略提供了理论依据.在此基础上,设计了以世界模型为推理引擎的完整编解码框架:编码端通过多轮链式推理对场景因果结构进行逐层解构,定位因果锚点

并构建非均匀码率分配策略,同时引入降质补偿机制对语义驱动的锚点识别进行校正;解码端以因果锚点为逻辑起点,首先驱动世界模型对场景进行因果逆推,还原完整的场景逻辑结构,再借助多路机制协同的扩散重建完成从稀疏因果锚点到全局自治场景的高质量重建.图1对比了传统端到端图像压缩、生成式图像压缩与本文方法三类范式的核心区别.实验结果表明,本方法在更低码率区间的各项指标上显著优于现有方法.

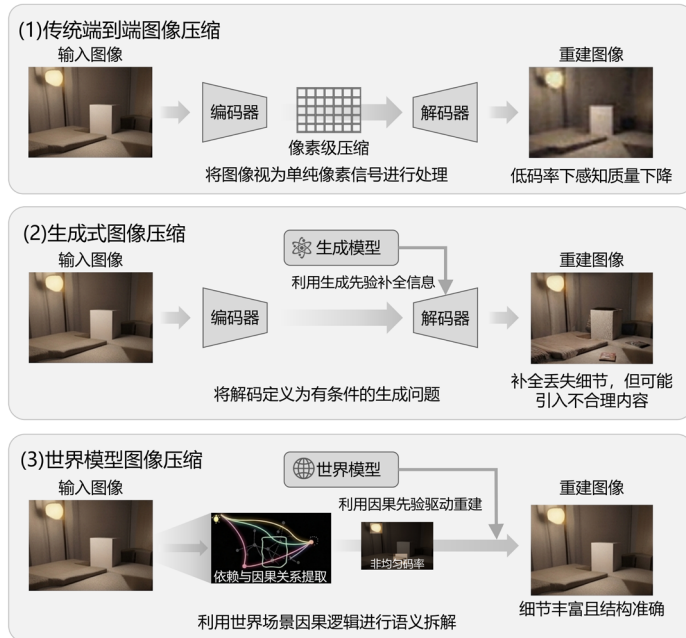


图1 传统端到端压缩、生成式压缩与世界模型压缩的范式对比

本文的主要贡献如下:

(1) 本文将自然图像建模为具有因果结构约束的场景切片,定义对全局场景具有决定性约束的世界基元为因果锚点,从信息论视角分析了将码率向因果锚点倾斜的优势,并从理论上说明了为何更强的世界模型能够带来更大的因果锚点优先分配收益,为本文方法提供了信息论依据.

(2) 本文设计了一种基于因果锚点提取的非均匀图像编码方法,以世界模型为推理引擎,通过链式思维对场景语义进行逐层解构,定位对全局场景具有关键影响的因果锚点区域,将有限码率集中分配于因果锚点.

(3) 以因果锚点为逻辑起点,本文设计了一种基于世界模型因果链条的场景重构方法,通过对因果锚点与极低码率压缩全图的联合推理沿因果链条还原场景逻辑结构,以此为条件驱动世界模型在自身因果先验引导下完成从稀疏因果锚点到全局自治场景的高质量重建.

本文其余部分组织如下:第1节对相关工作进行介绍,主要包括基于率失真优化的图像压缩、基于生成先验的图像压缩和世界模型;第2节介绍了基于因果分解的锚点优先码率分配理论基础;第3节详述本文设计的基于世界模型因果推断的图像压缩方法;第4节通过实验验证方法的有效性;第5节总结全文.

1 相关工作

1.1 基于率失真优化的端到端图像压缩

基于深度学习的端到端图像压缩框架由 Ballé 等人^[3]开创,其提出的变分自编码器架构将编解码器端到端统一优化,以可训练的非线性变换替代手工设计的离散余弦变换,在率失真性能上超越了传统编码标准.Minnen 等人^[4]在此基础上引入自回归上下文模型,将隐变量的概率分布建模为已解码邻域的条件分布,显著提升了熵编码效率.Cheng 等人^[13]则进一步引入注意力机制与离散高斯混合似然模型,更好地建模隐变量分布.

此后,研究者们从计算效率与架构设计两个维度持续推进.He 等人^[14]提出的方法通过分析深度压缩模型中的能量集中特性,设计了非均匀通道分组的空间-通道联合上下文自适应编码模型,在压缩性能与解码速度之间取得了权衡.Liu 等人^[5]提出了混合 Transformer-CNN 的图像压缩架构,引入基于 Swin Transformer 的注意力模块,通过局部与全局信息的协同建模进一步提升了潜在表示的紧致性.Wang 等人^[15]提出的方法通过条件编码与可变码率控制实现了灵活的率失真调节.Kuang 等人^[16]提出的方法则通过在编码过程中引入跨粒度在线优化机制,在实际编码中实现了更好的性能.此外,还有一些图像增强的工作可以被用于图像压缩的后处理模块来恢复图像细节,Yang 等人^[17]通过在线检索外部高频先验并与内部重建联合优化,Ding 等人^[18]则基于能量函数设计稳定的对抗辅助损失以生成更清晰的高分辨率图像,均为压缩后图像的细节恢复提供了有益思路.

在非均匀码率分配方向,Li 等人^[19]提出内容加权的图像压缩,通过卷积网络学习图像的重要性映射,对视觉显著区域分配更多码率,验证了非均匀分配策略在提升感知质量上的有效性.上述方法均以率失真为核心优化目标,以像素为处理单元,未能建立对图像内部物理因果关联关系的显式建模.Jia 等人^[20]对压缩领域的研究进展进行了系统梳理,指出融合深度网络模型的视频处理与压缩技术正在经历深刻变革,亟需建立视频大数据高效紧凑表示的理论与方法.

1.2 基于生成先验的图像压缩

生成式图像压缩以极低码率下的感知质量为核心目标,将解码端重新定义为有条件的图像生成问题.Agustsson 等人^[6]率先将生成对抗网络引入图像压缩,提出了一种结合编码器、生成器解码器与多尺度判别器的框架,在极低码率下合成感知质量良好的重建图像.Mentzer 等人^[21]系统研究了归一化层、生成器与判别器架构设计、训练策略以及感知损失的组合,在中低码率区间实现了与输入感知上高度相似的高质量重建.之后,Muckley 等人^[22]在 Mentzer 等人^[21]工作的基础上改进了判别器架构,引入基于多类分类的非二元判别器,进一步增强了重建图像的统计真实性.

随着扩散模型的研究,也有一些研究者尝试利用扩散模型进行图像压缩.Careil 等人^[7]以扩散模型为解码器,以向量量化图像表示与全局文字描述为条件,在极低码率下实现了感知上真实的重建,将生成式压缩的可用码率区间大幅向下延伸.Jia 等人^[8]则提出将变换编码从像素空间转移到向量化变分自编码器的潜在空间,利用语义对齐的潜在表示在极低码率下同时实现高保真与高感知质量.此外,还有一些工作引入文本或草图作为语义条件引导扩散生成.Lei 等人^[23]将图像标题与边缘草图作为条件控制网络的输入,实现了极低码率下的语义引导重建.Zhang 等人^[24]提出以单步扩散策略替代多步去噪过程,以压缩潜在特征为起点直接完成解码,在极低码率下实现了高效重建,有效缓解了扩散模型推理效率不足的问题.然而,上述方法的共同局限在于,其生成先验是统计意义上的图像分布,缺乏对场景内部物理因果关系的建模,在重建结果中常出现光照方向不一致、透视关系错误等物理不合理现象,且继承了均匀码率分配的基本假设.

近期亦有研究尝试利用大模型先验引导图像压缩.Li 等人^[25]提出利用大型多模态模型生成图像的全局语义描述与局部块-文本匹配,驱动扩散模型解码重建,但其码率分配仍基于均匀网格与 CLIP 相似度排序,未建模区域间的物理因果关联关系.Liu 等人^[26]通过大语言模型与空间定位模型在编码端进行语义分解,识别任务相关的语义显著区域以指导码率分配,但其显著性仍停留在任务语义层面,未涉及区域间的因果结构.Sargent 等人^[27]提出将视觉语言模型作为感知评判者用于训练阶段的偏好对齐,但该视觉语言模型不参与码率分配决策,压缩流程本身仍是均匀编码加扩散解码.上述方法将大模型作为描述源或评判者使用,提供的是统计意义上的"这类场景通常长什么样"或"哪种重建更符合人类偏好",而非区域间的物理因果关联关系.不同于上述工作,本文将世界模型作为因果推理引擎,通过链式推理建模区域间的物理因果关联,码率分配由因果分析驱动,解码端以因果锚点为条件进行因果推断,在物理合理性上具备优势.

1.3 世界模型

世界模型是指通过大规模数据训练、内化了物理世界运行规律的生成理解模型,具备从局部观测推断全局状态的能力.Ha 和 Schmidhuber^[12]提出了世界模型的基本框架,将环境的压缩潜在表征与未来状态预测相

结合,首次验证了智能体可以在自身生成的“梦境”中进行策略训练.

在表征空间与建模目标的演进上,世界模型经历了从像素空间到潜空间再到语义空间的持续深化.早期工作直接在像素空间进行状态预测,计算代价高昂且难以捕捉语义层面的因果结构.Hafner 等人^[28]提出的 Dreamer 系列采用的循环状态空间模型将高维观测压缩为低维潜表征,并将确定性隐状态与随机隐变量解耦,在保留时序依赖的同时显式建模环境随机性.随着语言模型的发展,部分研究进一步在语义符号空间中构建世界模型,以自然语言描述状态转移与因果关系,增强了推理的可解释性与跨域泛化能力.在预测目标上,建模对象从单一状态转移扩展至奖励、价值与完整交互轨迹:Schrittwieser 等人^[29]提出的 MuZero 摒弃像素重建,转而学习面向价值预测的抽象潜表征以支持蒙特卡洛树搜索,在围棋、国际象棋及 Atari 游戏上均超越人类水平;Yang 等人^[30]则将真实交互经验统一为视频预测框架,使智能体得以在生成场景中进行策略评估.在应用层面,世界模型的因果推断能力在具身智能中也展现出突出的价值.Hu 等人^[31]将场景理解、运动预测与轨迹规划统一于世界模型表征体系,使规划模块能够在模型内部预演候选轨迹并评估安全性;Bruce 等人^[32]提出的 Genie 则仅凭单张图像生成可交互的虚拟场景,为具身智能训练环境的低成本构建提供了新思路.

近年来,大型多模态模型的发展极大地拓展了世界模型的能力边界.Deng 等人^[33]提出的统一多模态基础模型,基于混合 Transformer 专家架构,在万亿量级交错多模态数据上进行预训练,展现出自由形式图像操作、未来帧预测、三维空间操作与世界导航等世界建模能力.与普通生成模型仅依赖统计相关性不同,世界模型能够建立区域间的物理因果关联关系,从局部的因果锚点出发推断全局场景的合理状态,这为将其引入图像压缩提供了关键的能力基础.

2 基于因果分解的锚点优先码率分配

本章从信息论视角建立因果分解下的码率分配理论,证明以下结论:在将图像分解为因果锚点与关联残差的基础上,世界模型的预测能力越强,将码率优先分配给因果锚点所带来的压缩性能增益越显著.

设图像 x 具有如下因果分解:

$$x = (x_c, x_r),$$

其中 x_c 为图像中的因果锚点,即场景的世界基元,承载图像的主要因果关联结构; x_r 为关联残差,表示因果锚点已知条件下的剩余信息.由熵的链式法则有

$$H(x) = H(x_c) + H(x_r | x_c).$$

由于关联残差的不确定性高度受制于因果锚点的约束,在因果锚点已知的条件下,关联残差的条件熵远小于其边缘熵:

$$H(x_r | x_c) \ll H(x_r).$$

等价地,因果锚点与关联残差之间具有较大的互信息:

$$I(x_c; x_r) = H(x_r) - H(x_r | x_c).$$

设世界模型为 W ,其核心能力在于根据因果锚点信息逼近关联残差的真实条件分布:

$$W \approx p(x_r | x_c).$$

第一步:建立非均匀码率分配的优化目标

设总码率为 R ,将其分配给因果锚点与关联残差两部分:

$$R = R_c + R_r.$$

记因果锚点的失真函数为 $D_c(R_c)$, 关联残差在重建因果锚点 $\hat{x}_c$ 条件下的失真函数为 $D_r(R_r | \hat{x}_c)$. 由 $R_r = R - R_c$, 总失真可写为

$$D_{\text{total}}(R_c) = D_c(R_c) + \mathbb{E}_{\hat{x}_c} [D_r(R - R_c | \hat{x}_c)].$$

残差项对 R_c 的关联包含两条路径: 其一为码率守恒 $R_r = R - R_c$, 其二为重建因果锚点 $\hat{x}_c$ 的质量由 D_c 决定, 进而影响世界模型的条件推断精度. 对 $D_{\text{total}}(R_c)$ 关于 R_c 求导并对残差项应用链式法则, 得

$$\frac{dD_{\text{total}}}{dR_c} = \frac{dD_c}{dR_c} + \frac{\partial \mathbb{E}[D_r]}{\partial R_r} \cdot \frac{dR_r}{dR_c} + \frac{\partial \mathbb{E}[D_r]}{\partial D_c} \cdot \frac{dD_c}{dR_c},$$

其中第一项为 R_c 变化对 D_c 的直接影响, 第二项为 R_r 随 R_c 增大而减小所带来的残差失真上升, 第三项为因果锚点质量提升通过世界模型条件推断间接降低残差失真的效果. 由 $R_r = R - R_c$, 有 $dR_r / dR_c = -1$.

定义关联系数 η 为

$$\eta := \frac{\partial \mathbb{E}[D_r]}{\partial D_c} \geq 0,$$

其非负性来自于: 当 D_c 减小时, 重建因果锚点 $\hat{x}_c$ 与真实因果锚点 x_c 更为接近, 世界模型的条件推断精度提升, 残差失真 $\mathbb{E}[D_r]$ 随之减小, 二者同向变化. 代入上式得

$$\frac{dD_{\text{total}}}{dR_c} = \frac{dD_c}{dR_c} (1 + \eta) - \frac{\partial \mathbb{E}[D_r]}{\partial R_r}.$$

对于不对因果结构进行分解的均匀码率分配方法, 码率变化对因果锚点与关联残差的边际影响近似相等, 即

$$\frac{dD_c}{dR_c} \approx \frac{\partial \mathbb{E}[D_r]}{\partial R_r}.$$

将此代入, 得到向因果锚点倾斜分配码率时的边际总失真改善率:

$$\left. \frac{dD_{\text{total}}}{dR_c} \right|_* \approx \frac{dD_c}{dR_c} \cdot \eta.$$

由于 $dD_c / dR_c < 0$, 当 $\eta > 0$ 时 $dD_{\text{total}} / dR_c|_* < 0$, 即向因果锚点倾斜分配码率使总失真下降, 且下降幅度正比于关联系数 η .

第二步: 建立关联系数与互信息的关系

定义残差条件熵膨胀为因果锚点降质导致关联残差剩余不确定性的增量:

$$\Delta H(D_c) := H(x_r | \hat{x}_c) - H(x_r | x_c) \geq 0,$$

其中 $\hat{x}_c$ 对应失真 D_c 下的重建因果锚点. 将关联链 $D_c \mapsto \Delta H \mapsto \mathbb{E}[D_r]$ 按链式法则展开, 关联系数可分解为

$$\eta = \frac{\partial \mathbb{E}[D_r]}{\partial \Delta H} \cdot \frac{\partial \Delta H}{\partial D_c}.$$

该分解中两个因子分别具有如下单调性:

(i) $\partial \mathbb{E}[D_r] / \partial \Delta H > 0$, 由率失真函数的基本性质, 在固定残差码率 R_r 下, 待编码信息的条件熵上升必然导致重建失真上升.

(ii) $\partial \Delta H / \partial D_c$ 正比于 $I(x_c; x_r)$, 因果锚点与残差耦合越紧(互信息越大), 同等因果锚点失真对残差约束的破坏越严重:

$$\frac{\partial \Delta H}{\partial D_c} \propto I(x_c; x_r).$$

综合两因子, 得到关联系数与互信息的定性关系:

$$\eta \propto I(x_c; x_r).$$

第三步: 建立互信息与世界模型预测能力的关系

由 $I(x_c; x_r) = H(x_r) - H(x_r | x_c)$, 其中 $H(x_r)$ 为数据内在属性, 与模型无关; $H(x_r | x_c)$ 则反映了在因果锚点已知条件下关联残差的剩余不确定性, 取决于世界模型对场景因果规律的掌握程度. 世界模型的预测能力越强, 其所能达到的条件熵 $H(x_r | x_c)$ 越低, 互信息 $I(x_c; x_r)$ 越高, 关联系数 η 越大.

综合三步推导, 得到本章的核心结论:

$$\left. \frac{dD_{\text{total}}}{dR_c} \right|_* \approx \frac{dD_c}{dR_c} \cdot \eta, \quad \eta \propto I(x_c; x_r), \quad \text{世界模型越强} \Rightarrow I(x_c; x_r) \uparrow.$$

世界模型的预测能力越强, $I(x_c; x_r)$ 越大, 关联系数 η 越大, 将码率更多分配给因果锚点所带来的总失真下降越显著. 这一结论从理论层面为本文提出的非均匀码率分配策略提供了严格的信息论依据, 也表明世界模型的预测能力越强, 基于因果分解的压缩方法的压缩增益越显著.

3 基于世界模型因果推断的图像压缩方法

上一章从信息论视角证明了世界模型的预测能力越强, 将码率优先分配给因果锚点所带来的压缩增益越显著. 基于这一理论结论, 本章介绍了一种基于世界模型因果推断的图像压缩方法, 将理论中的因果锚点与关联残差分解转化为可操作的编解码流程. 该方法的流程示意图如图 2 所示.

本方法采用字节跳动发布的统一生成理解模型 BAGE^[33] 作为世界模型推理引擎. 该模型基于混合 Transformer 架构, 通过大规模多模态数据训练内化了丰富的物理世界因果规律, 在统一框架下同时具备跨模态语义理解、因果推断与高质量图像生成能力. 该模型对中文具有优异的理解与生成支持, 有利于在码流中实现更为紧致的语义信息表达.

编码端通过多轮链式推理对输入图像进行逐层因果解构, 将图像的语义内容逐步提炼为精确定位的因果锚点集合, 依次经过三个阶段. 第一阶段为世界场景语义建模, 对输入图像进行全局语义理解与场景实例分割, 对图像做整体语义与布局理解. 第二阶段为世界因果链路感知, 在跨实例层面评估实例间因果关联关系并兼顾场景级重要性, 之后在实例内部对每个实例的子区域进行三维因果打分, 逐步建立从场景到实例再到子区域的完整因果重要性序列. 第三阶段为因果锚点空间定位, 在完整图像的全局坐标系下对各因果锚点进行空间定位, 确定每个锚点的归一化边界框, 将语义层面的因果角色与图像空间位置显式绑定, 为后续分层编码的区域划分提供空间依据.



因果感知的分层压缩阶段首先针对链式推理无法感知实际压缩降质的固有局限,对极低码率压缩全图与原图进行区域级降质分析,将降质程度过高的区域补充纳入因果锚点范围,对语义驱动的锚点识别加以补充.在此基础上构建非均匀码率分配策略,对因果锚点区域以高码率精确保护,对其余区域以极低码率压缩为关联残差,同时提炼生成以紧致中文表达的场景语义骨架,将三路信息共同传输至解码端.

解码端接收分层编码传输的三路码流后,首先以场景语义骨架、因果锚点与关联残差为多模态输入,驱动世界模型进行因果逆推,沿因果链条还原场景的完整逻辑结构,输出世界因果关系描述.继而以世界因果关系描述与场景语义骨架 S 为语义条件,以关联残差重建图加噪构建扩散起始点,借助因果锚点区域的不确定性重注入在每步去噪迭代中锁定锚点区域.同时引入隐空间流场引导机制对全局结构施加梯度约束,多路机制协同驱动世界模型完成高质量场景重建.下文将对各部分流程展开详细介绍.

3.1 因果锚点的链式分析与定位

本节介绍编码端如何以世界模型为推理引擎,通过多轮链式调用将输入图像逐步解构为精确定位的因果锚点集合,并生成用于解码端重建的场景语义骨架.编码端的因果分析采用链式推理架构,每一步的输出作为下一步的上下文输入注入世界模型,使得每一次分析都在上一次的基础上继续完善.这种思维链传递机制确保每一步的世界模型调用都建立在前序步骤充分语义理解的基础上,避免独立分析带来的语义偏差.

整个链式推理流程依次经过世界场景语义建模、世界因果链路感知与因果锚点空间定位三个层面,由六个阶段组成.每个阶段在调用世界模型时均以原始输入图像作为视觉基础,同时将上一阶段的输出作为上下文输入注入,确保每一步的分析都建立在前序步骤充分语义理解的基础上.整个过程中每一步的输出作为下一步的上下文输入注入世界模型,形成层层递进的因果分析链,最终汇总全部链式分析结果生成场景语义骨架 S .具体细节将在下面分三个小节介绍.

3.1.1 世界场景语义建模

接收输入图像后,编码端通过全局描述与实例枚举两个阶段建立整体语义与空间布局理解,两者共享同一张输入图像,通过上下文注入形成递进式理解.

在全局描述阶段,世界模型以原始输入图像为唯一输入,从场景类型、空间分区、关键主体、大面积均匀区域、覆盖性核查五个角度建立空间心智地图,输出一段 3-6 句的场景描述文本,涵盖场景类型、主要主体位置与背景区域分布,为后续实例枚举提供空间理解基础.

在实例枚举阶段,世界模型以原始输入图像与全局描述阶段输出的场景描述为输入,对图像进行实例级分割.语义实例定义为任何可独立识别的区域,包括前景对象及大尺度背景表面,前景对象与背景面平等对待.同类且共处同一区域的重复元素强制合并为单一集体实例,最终实例数量约束为 4 至 8 个.完成枚举后输出每个实例的名称与语义类别.

完成实例枚举后,以各实例的语义描述为提示调用实例分割模型 SAM^[34],以原始输入图像与实例名称为输入,为每个实例生成不规则像素掩码,将语义层面的实例划分落实到像素级空间位置,为后续因果分析与分层编码提供区域基础.

3.1.2 世界因果链路感知

以实例枚举结果为输入,编码端通过两轮世界模型调用分析各实例及子区域的因果关系,分别从场景层面与实例层面展开,逐步建立从场景到实例再到子区域的完整因果重要性序列.

第一轮调用在场景层面进行,以原始输入图像与实例枚举阶段输出的实例列表为输入,从视觉重要性与物理因果约束强度两个角度评估每个实例:视觉重要性判断该实例是否为视觉焦点并归入主层、次层或背景层;物理因果约束强度判断该实例是否通过现实因果约束决定其他实例的可见外观,例如光源决定阴影方向与明暗分布、桌面决定其上物体的位置与稳定性.综合两者以因果得分量化约束强度,光源、地平线、主体轮廓等对全局场景具有决定性约束的实例获得更高因果得分,叠加视觉重要性分数得到最终重要性分数.

第二轮调用在实例层面对每个实例独立执行,将场景级评估结果与当前实例信息一并注入世界模型,识别 2-6 个语义子区域.子区域定义为实例内具有独立视觉特征与语义作用的部分,如人的脸部、躯干、四肢,树木的树冠、树干、根部等,而非属性类型.随后从因果推断视角对每个子区域进行三维打分:不可推断性衡量该区域被遮挡后世界模型从剩余可见子区域推断其外观的难度,独特标记、表情、标志等获得高分;引导性衡量仅以该区域为参考时世界模型重建其余子区域的能力,可铺贴整面的纹理样本等获得高分;物理因果性衡量该区域通过物理规律约束其他子区域外观的强度,如关节约束肢体方向、局部光源决定阴影等.每部分分数范围为 1-10 分,要求在实例内覆盖全部子区域.三个维度共同刻画每个子区域的因果重要程度,为后续锚点角色分配提供依据.

3.1.3 因果锚点空间定位

以三维因果打分结果为输入,世界模型对每个实例的子区域计算因果分数,按如下公式加权聚合三个维度:

$$S_{causal} = \frac{0.45(u-1) + 0.40(g-1) + 0.15(p-1)}{9},$$

其中 S_{causal} 表示因果得分, u 、 g 、 p 分别对应不可推断性、引导性与物理因果性的原始评分.按锚点分数降序排列,将每个实例内评分最高的前两个子区域设为锚点,其余设为关联残差.随后在完整图像的全局坐标系下为每个锚点子区域分配归一化边界框,将语义上的因果区域落实到图像的具体空间位置.

完成空间定位后,将锚点子区域的像素掩码对齐为边长为 256 像素的矩形块,以便以坐标形式紧致地表达锚点位置信息.对齐过程中可能引入实例间的区域重叠,对于交并比大于 0.3 的两块锚点区域,将因果得分更低的区域降级为关联残差,并限制总锚点面积不超过图像面积的 30%,确定所有因果锚点的位置.

3.2 面向因果感知的分层压缩

链式因果分析完成后,编码端进入压缩阶段,首先通过降质补偿对锚点集合进行检查确认,随后基于最终锚点集合执行分层编码,输出三路码流.

3.2.1 降质补偿锚点扩充

由于链式因果分析阶段识别的因果锚点完全基于语义理解,无法感知实际压缩过程中各区域的降质情况.某些在语义上并不突出的区域,在极低码率压缩后可能出现严重失真.为此,本方法在压缩前引入降质补偿机制,对因果锚点集合进行扩充.

具体来说,编码端对输入图像进行极低码率预压缩,得到全图的降质参考图.随后驱动世界模型对原图与降

质参考图进行逐区域对比分析,评估每个非锚点区域的压缩损伤程度,输出压缩敏感度评分.对于压缩敏感度较高的区域,在原有三维因果打分的基础上引入压缩敏感度作为第四维,按如下公式重新计算锚点分数:

$$S_{causal} = \frac{0.35(u-1)+0.30(g-1)+0.10(p-1)+0.25(cs-1)}{9},$$

其中 cs 为压缩敏感度评分.若重新计算后的锚点分数超过当前锚点列表中的最低锚点分数,则将该区域补充纳入锚点列表.

3.2.2 锚点优先分层编码

最后,基于最终锚点列表构建非均匀码率分配策略.本方法采用 Xue 等人^[35]提出的双路潜空间融合压缩方法为压缩网络.对于因果锚点区域,本方法以高码率进行压缩,从而确保因果结构的精确传输.对于关联残差,将全图下采样至 256×256 后以极低码率压缩,保留全局空间布局的粗粒度参考.

同时,编码时还汇总全部链式分析结果,驱动世界模型生成一段不超过 30 个汉字的中文场景语义骨架,涵盖场景类型、主要因果结构与关键锚点类型.得益于世界模型对中文的优异支持,中文语义描述能够以更紧致的码流承载等量的场景语义信息.最终将因果锚点码流、关联残差码流与场景语义骨架三路信息共同传输至解码端.

3.3 基于世界模型的链式因果重组与重建

解码端接收三路信息——因果锚点码流、关联残差码流与场景语义骨架,以世界模型为核心推理引擎,分为因果逆推与扩散重建两个阶段.

3.3.1 因果逆推场景重组

解码端首先对接收到的三路信息进行因果逆推分析,将稀疏的因果锚点扩展为对全局场景的完整因果表示.具体地,以提供全局空间布局的关联残差重建图与锚点图作为视觉输入,结合场景语义骨架 S ,驱动世界模型对场景进行因果分析.这里,为了让锚点图注入时同时表征锚点所在位置,将所有锚点图粘贴到了同一张灰色图上,保留锚点位置精确像素,使世界模型在获得锚点的纹理、材质与边缘细节的同时确定锚点所在位置.

在世界模型因果逆推的推理过程中,需要调用其以链式思维方式在给定提示词的引导下按照以下五个维度进行分析:

- 1) 空间布局分析,基于关联残差重建图理解主要结构元素及其空间关系;
- 2) 锚点纹理与材质分析,从锚点加灰图中精确感知锚点区域的表面纹理、材质属性、边缘锐度与颜色梯度;
- 3) 非锚点内容推断,基于全局语义与空间上下文推断每个非锚点区域的合理内容,而非仅依赖局部邻域;
- 4) 光照与氛围分析,推断全场景的光线方向、色温与整体色调;
- 5) 物理约束推断,基于对称性、重力、材质连续性、纹理周期性与透视一致性等物理规律进行全局自洽性校验.

完成上述分析后,再次调用世界模型将上述推理过程浓缩为 2-6 句完整的场景重组表示,这一表示与场景语义骨架共同构成后续扩散重建的完整语义条件.

3.3.2 世界先验驱动重建

以因果逆推场景重组的输出为条件,解码端在世界模型上执行扩散过程完成场景重建.重建前,解码端首先构建一张信息分层的预处理图像作为视觉条件输入.该图像的锚点区域保留精确的高质量像素,其它区域使用关联残差的高通滤波版本,保留结构边缘并移除低频颜色,以避免低频颜色干扰扩散生成过程导致重建图像模糊.

重建过程以关联残差重建图加噪后的隐变量作为扩散起始点,从而保留全局空间布局信息,同时赋予扩散过程足够的生成自由度.在每步去噪迭代中,本方法还借鉴了 Askari 等人^[36]提出的隐空间流引导思路,对扩散过程中的隐变量施加来自关联残差重建图的变分自编码器雅可比梯度约束.通过在早期去噪阶段提供空间布局的梯度引导,并在晚期去噪阶段关闭该引导,实现了在保证重建整体结构与关联残差图对齐的同时避免将过多

的压缩失真带入输出中.同时对因果锚点区域与关联残差区域之间的过渡边界区域单独施加边界连续性约束,使待生成的关联残差区域与因果锚点区域平滑过渡,最终输出因果自洽的高质量重建图像.

4 实验分析

4.1 实验设定

实验使用 CLIC 2020 专业测试集^[37]与 DIV2K 验证数据集^[38]进行评估.CLIC 2020 专业测试集包含 41 张 2K 分辨率高质量图像,涵盖多样化的自然场景与人工场景,能够有效评估方法在高分辨率图像上的压缩性能.DIV2K 验证数据集包含 100 张高分辨率自然图像,涵盖人像、风景、建筑、动植物等多种场景类型,在场景多样性与内容复杂度上更加全面,能够更充分地检验方法在不同视觉场景下的泛化性能.

本文在极低码率区间设置码率点进行评估,码率范围约为 0.001 到 0.003 比特每像素.码率约束通过联合调节因果锚点的区域大小、最大覆盖面积以及压缩网络所加载的码率档位来实现.

本方法不需要补充任何额外的训练,全部模块加载已有的预训练参数.其中世界模型引擎 BAGEL^[33]加载的预训练参数为 Xie 等人^[39]添加重建损失微调后的参数,该微调策略有助于提升 BAGEL 模型在图像保真上的性能.压缩骨干采用 Xue 等人^[35]提供的预训练模型.所有实验推理过程在单张 NVIDIA A800 GPU 上完成.

本文从保真度、感知相似度与无参考质量三个维度选取了 9 个评价指标对重建图像质量进行了全面的评估.在保真度维度,传统的 PSNR 与 SSIM 在极低码率下区分度极低,难以反映方法间的实际差异,因此本文采用 VIF^[36]与 GMSD^[41],其中 VIF 基于信息论建模视觉信息保留量,对极低码率下保留的有效结构信息敏感;GMSD 基于梯度幅值衡量结构保真度,对极低码率下图像的锐度退化具有较好的区分能力.在感知相似度维度,采用 LPIPS-A^[42]、DISTS^[43]、FSIM^[44]与 PIEAPP^[45].该维度在极低码率下成为用户体验的主导因素,上述四个指标从不同视角互补:LPIPS-A 基于 AlexNet^[46]深度特征衡量感知距离,与人类感知判断高度相关;DISTS 联合建模结构与纹理相似性,对生成式压缩方法常见的纹理失真敏感;FSIM 基于相位一致性与梯度特征,结构感知特性互补于 LPIPS;PIEAPP 基于成对偏好学习,直接对齐人类主观判断.在无参考质量维度,采用 NIQE^[47]、MUSIQ^[48],其中 NIQE 基于自然图像统计建模无参考质量;MUSIQ 基于多尺度 Transformer 预测感知质量,对极低码率下的退化更为敏感.FID^[49]则度量重建图像分布与原图分布的整体差异,适合分布级质量评估.上述指标覆盖了从像素级保真到感知相似再到分布质量的多层次评估需求,能够全面反映极低码率下本方法重建图像的综合质量.

4.2 客观结果分析

本文进行了丰富的对比实验.包括传统编码标准 BPG^[2]、VTM^[50],面向保真度的率失真优化压缩方法 ELIC^[14],面向感知质量的图像压缩方法 HiFiC^[21]、MS-ILLM^[22]、MRIC^[51]、CDC^[52]、CGDM^[53]、CGDM+^[54]与 CRDR^[55],面向极低码率的概念压缩方法 GLC-Image^[8]、DLF^[35]、PerCo^[7]与 PICS^[23].上述方法覆盖了不同的压缩策略与码率范围,能够全面衡量本文方法在极低码率区间的竞争力.本文方法与各对比方法在 CLIC 2020 专业测试集和 DIV2K 验证数据集上的率失真曲线分别如图 3 和图 4 所示.

值得注意的是,各方法的码率覆盖范围存在显著差异.BPG、VTM、ELIC、HiFiC、CGDM 等方法的码率范围集中在 0.02 bpp 以上,而本文方法的码率区间为 0.001~0.003 bpp,远低于上述方法的可用码率范围.为便于观察极低码率区间的细节,在每个指标的率失真曲线图下方均附有 0~0.02 bpp 范围的局部放大图.可以看到,本文方法在感知相似度度量、保真度度量和无参考质量度量等各个维度上均优于对比方法,验证了基于世界模型因果推断的压缩框架在极低码率场景下的有效性.

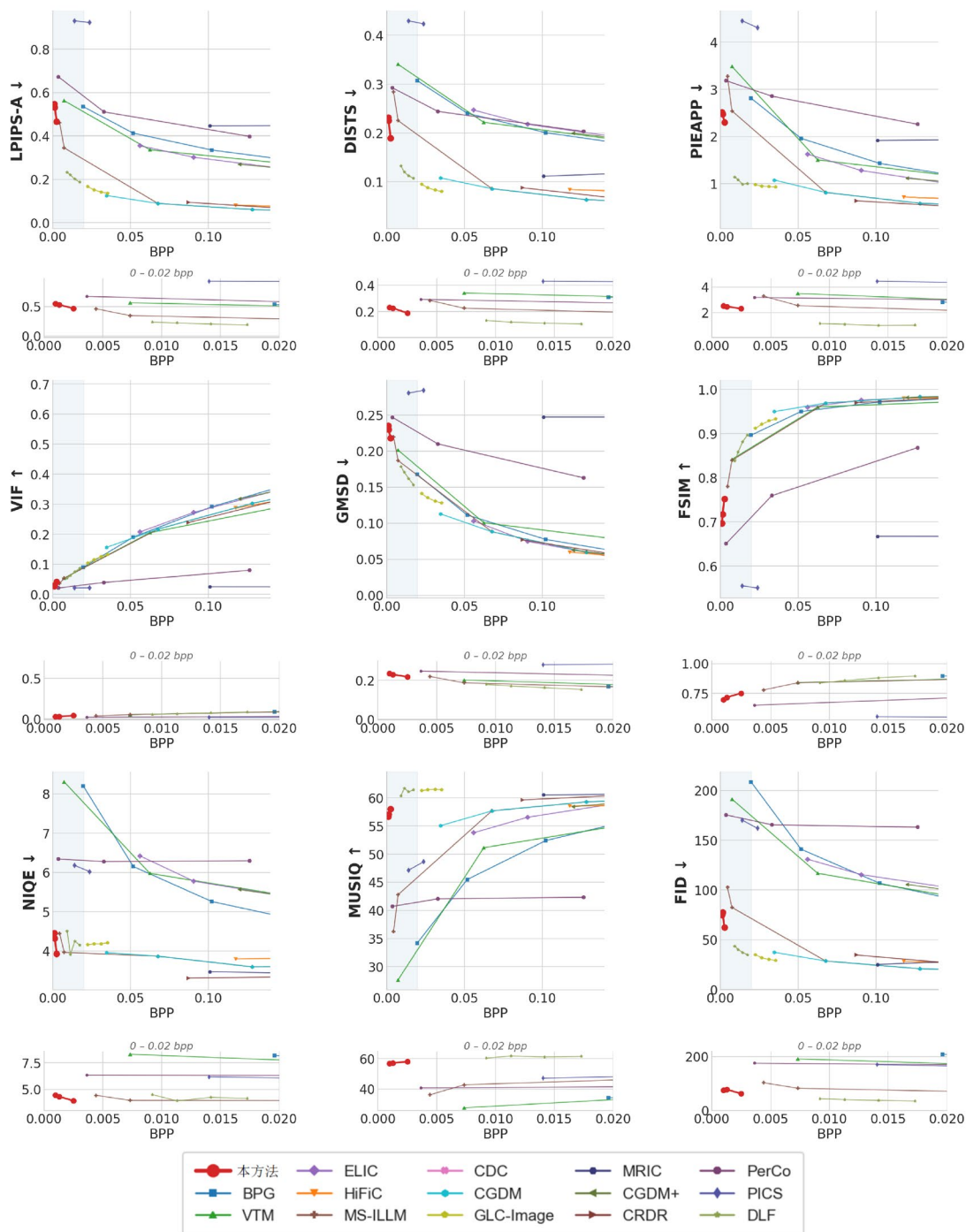


图 3 不同方法在 CLIC 2020 数据集上的率失真曲线对比图 (彩印)

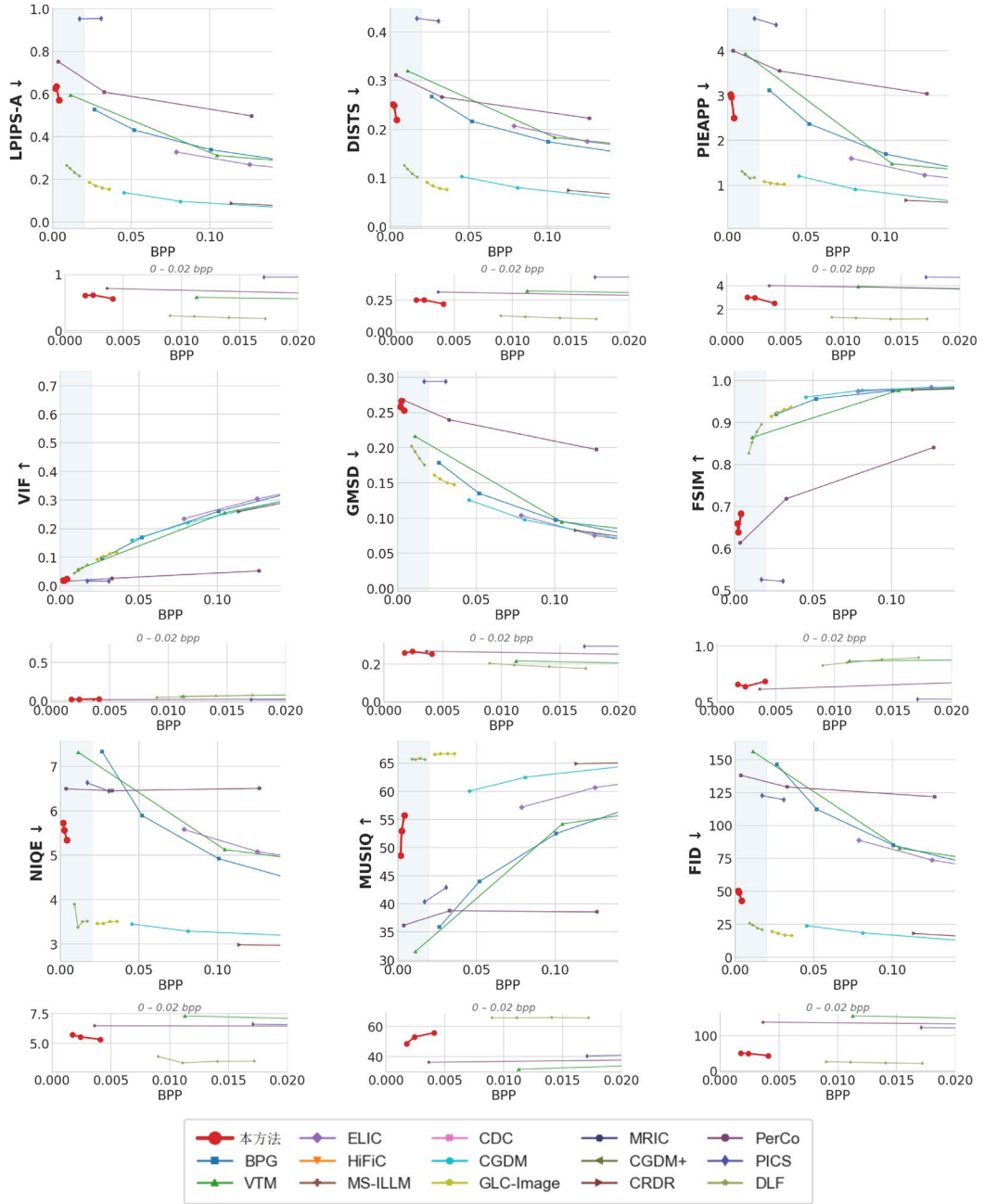


图4 不同方法在DIV2K验证数据集上的率失真曲线对比图(彩印)

4.3 主观结果分析

同时,为了更好地展示本文方法在极低码率下实现的人眼感知质量,本文展示了部分与本文方法处于同码率范围的方法重建图像.重建图像的主观结果如图5所示.

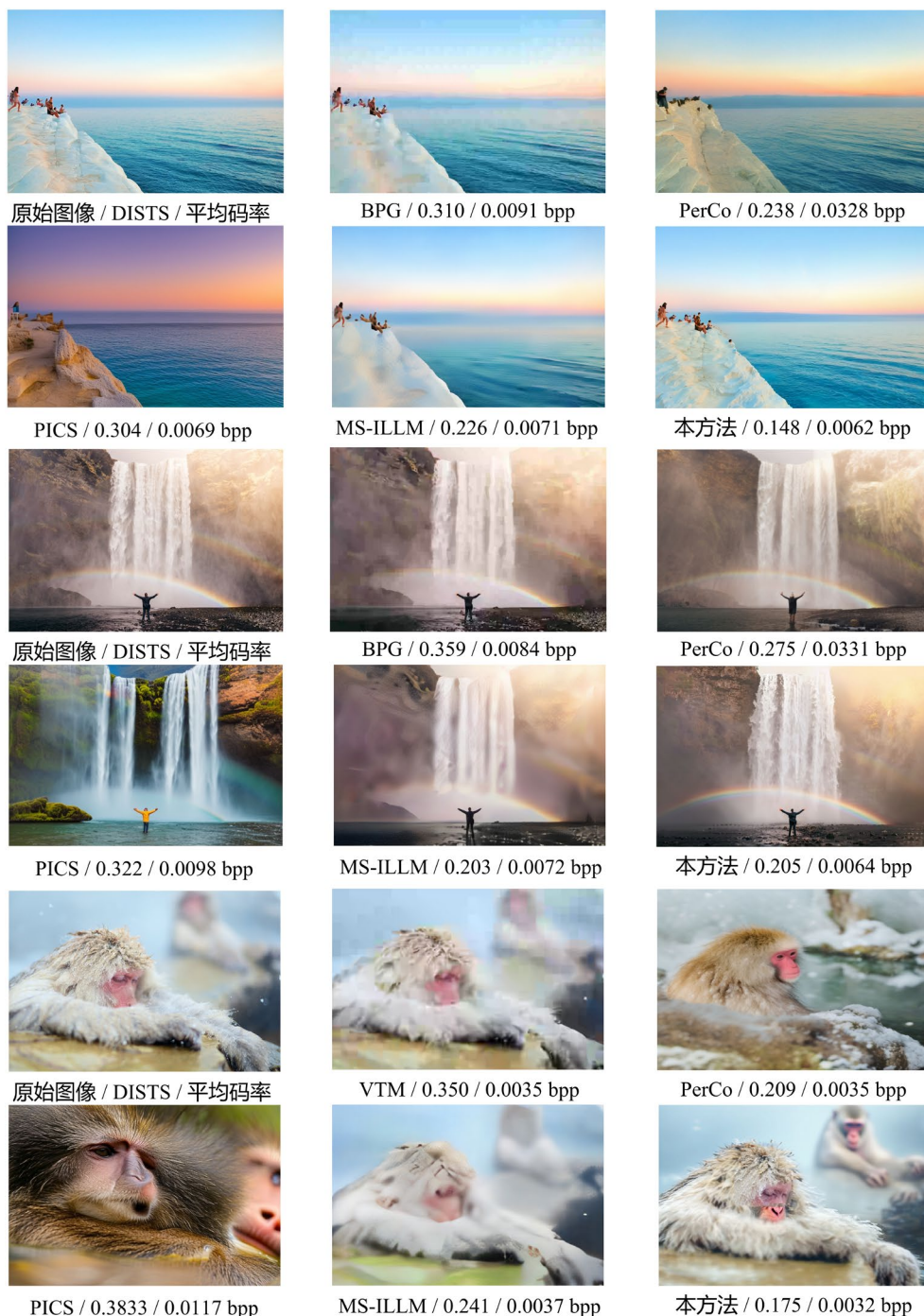


图5 不同方法重建图像的主观对比图(彩印)

可以看到,VTM 等传统编码方法在极低码率下会产生严重的块状伪影与细节丢失,人物面部等关键区域甚至完全无法辨认.PerCo、PICS、MS-ILLM、DLF 等面向感知质量的压缩方法和概念压缩方法虽能在一定程度上生成视觉上较为平滑的结果,但往往引入物理不合理的幻觉内容,或出现色调严重偏移、结构失真等问题.相比之下,本文方法在更低的码率下重建的图像具有更丰富的视觉细节与更合理的场景结构,人物轮廓、动物毛

发纹理与自然场景的色彩均得到了自然还原,说明世界模型因果先验有助于在极低码率下提升重建的物理合理性.

4.4 消融实验分析

为验证本文方法各核心模块的有效性,设计了五组消融实验,五组消融实验的率失真曲线图如图 6 所示.

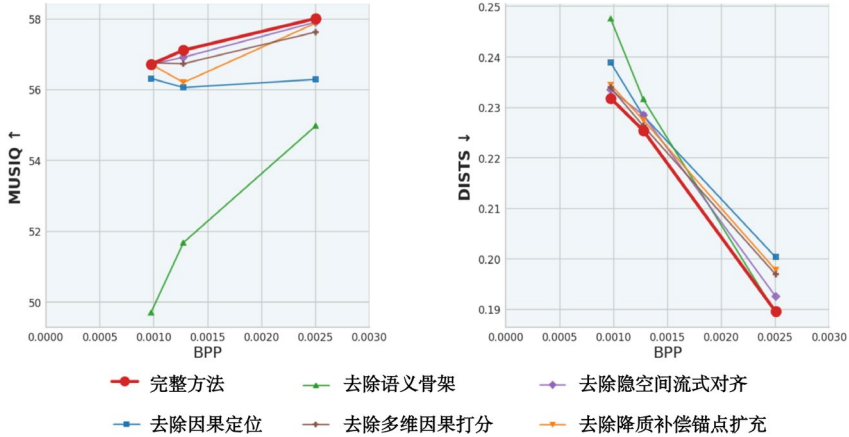


图 6 消融实验的率失真曲线 (彩印)

首先验证的是因果锚点定位的有效性.如图 6 中蓝色曲线所示,将世界模型链式推理定位的因果锚点替换为随机选取的同等面积区域,各项指标出现一定程度的下降.这表明因果锚点的定位质量对最终重建效果具有显著影响:随机选取的区域不具备对全局场景的因果约束力,世界模型难以从中推断出物理自洽的完整场景,验证了链式推理精准定位因果锚点的必要性.

之后验证语义骨架的有效性.如图 6 中绿色曲线所示,去除场景语义骨架的传输与引导后,性能指标出现了明显下降.这表明语义骨架对重建质量具有较大影响:缺乏语义骨架引导时,解码端的世界模型失去了对场景整体逻辑结构的文字约束,因果逆推过程缺乏有效的语义锚定,导致重建图像在感知自然度与结构相似性两个维度上均出现显著退化.这一结果表明语义骨架对因果逆推具有重要作用.

隐空间流场对齐模块的有效性方面,如图 6 中紫色曲线所示,去除隐空间流场对齐模块后,各项指标出现小幅下降.这表明隐空间流场对齐模块通过在隐空间中对全局结构进行对齐,对重建质量起到了稳定的提升作用,但本文框架的整体性能主要依赖于因果锚点的精准定位与语义骨架的全局引导,隐空间流场对齐模块在此基础上进一步细化了重建结果的结构一致性.

降质补偿锚点扩充机制的有效性方面,如图 6 中橙色曲线所示.去除降质补偿机制后,各项指标出现一定程度的下降,且退化集中在较低码率档位.这表明降质补偿机制对纹理复杂但语义不显著的区域具有不可替代的保护作用:这些区域在链式推理中不会被选为锚点,但在极低码率压缩后失真严重,降质补偿机制通过引入压缩敏感度作为补充维度,有效弥补了仅靠语义驱动锚点识别的盲区.

最后验证多维因果打分的有效性.如图 6 中棕色曲线所示,仅使用不可推断性单维打分后,方法性能出现小幅退化.去除多维度因果打分后,世界模型仅能进行单维因果重要性评估,无法从多视角综合分析区域间的因果关联关系,导致锚点选取质量下降.

5 总 结

本文提出了一种基于世界模型因果推断的图像压缩方法.受认知神经科学中人类场景感知机制的启发,本文将自然图像建模为具有因果结构约束的世界场景切片,将对全局场景具有关键影响的区域定义为因果锚点,即场景的世界基元.之后,本文从信息论视角证明了将码率优先分配给因果锚点的优越性,并在此基础上设计了以世界模型为推理引擎的图像压缩方法:编码端通过多轮链式推理定位因果锚点,构建非均匀码率分配策略;解

码端以因果锚点为逻辑起点,通过因果逆推场景重组与扩散重建,完成从稀疏因果锚点到全局自洽场景的高质量重建.实验结果表明,本文方法在极低码率区间内的各项评价指标上均优于现有方法,验证了基于因果先验的图像压缩的有效性.

References

- [1] Wallace GK. The JPEG still picture compression standard. *Communications of the ACM*, 1991,34(4):30-44.
- [2] Bellard F. BPG image format. Technical Report, 2017. (<https://bellard.org/bpg/>)
- [3] Ballé J, Laparra V, Simoncelli EP. End-to-end optimized image compression. In: *Proc. of the Int'l Conf. on Learning Representations (ICLR)*. Toulon, France, 2017. 1-27.
- [4] Minnen D, Toderici G, Singh S, Hwang SJ, Covell M. Image-dependent local entropy models for learned image compression. In: *Proc. of the IEEE Int'l Conf. on Image Processing (ICIP)*. Athens, Greece: IEEE, 2018. 430-434.
- [5] Liu J, Sun H, Katto J. Learned image compression with mixed transformer-CNN architectures. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, Canada: IEEE, 2023. 14388-14397.
- [6] Agustsson E, Tschannen M, Mentzer F, Timofte R, Van Gool L. Generative adversarial networks for extreme learned image compression. In: *Proc. of the IEEE/CVF Int'l Conf. on Computer Vision (ICCV)*. Seoul, South Korea: IEEE, 2019. 221-231.
- [7] Careil M, Muckley MJ, Verbeek J, Lathuilière S. Towards image compression with perfect realism at ultra-low bitrates. In: *Proc. of the 12th Int'l Conf. on Learning Representations (ICLR)*. Kigali, Rwanda, 2023. 1-21.
- [8] Jia Z, Li J, Li B, Li H, Lu Y. Generative latent coding for ultra-low bitrate image compression. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2024. 26088-26098.
- [9] Friston K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 2010,11(2):127-138.
- [10] O'Keefe J, Nadel L. *The hippocampus as a cognitive map*. Oxford: Oxford University Press, 1978. 1-598.
- [11] Maguire EA, Intraub H, Mullally SL. Scenes, spaces, and memory traces: what does the hippocampus do? *The Neuroscientist*, 2016,22(5):432-439.
- [12] Ha D, Schmidhuber J. Recurrent World Models Facilitate Policy Evolution. In: *Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems (NeurIPS)*. Montréal, Canada: Curran Associates, 2018. 2455-2467.
- [13] Cheng Z, Sun H, Takeuchi M, Katto J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020. 7936-7945.
- [14] He D, Yang Z, Peng W, Ma R, Qin H, Wang Y. ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2022. 5708-5717.
- [15] Wang GH, Li J, Li B, Lu Y. EVC: Towards real-time neural image compression with mask decay. In: *Proc. of the Int'l Conf. on Learning Representations (ICLR)*. Kigali, Rwanda, 2023. 1-23.
- [16] Kuang H, Yang W, Guo Z, Liu J. Cross-granularity online optimization with masked compensated information for learned image compression. In: *Proc. of the IEEE/CVF Int'l Conf. on Computer Vision (ICCV)*. Honolulu, HI, USA: IEEE, 2025. 16514-16523.
- [17] Yang WH, Liu JY, Xia SF, Guo ZM. Data-driven external compensation guided deep networks for image super-resolution. *Ruan Jian Xue Bao/Journal of Software*, 2018,29(4):900-913 (in Chinese).
- [18] Ding L, Ding SF, Zhang J, Zhang ZC. Single image super-resolution reconstruction based on VGG energy loss. *Ruan Jian Xue Bao/Journal of Software*, 2021,32(11):3659-3668 (in Chinese).
- [19] Li M, Zuo W, Gu S, Zhao D, Zhang D. Learning convolutional networks for content-weighted image compression. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA: IEEE, 2018. 3214-3223.
- [20] Jia CM, Ma HC, Yang WH, Ren WQ, Pan JS, Liu D, Liu JY, Ma SW. Video processing and compression technologies. *Journal of Image and Graphics*, 2021,26(6):1179-1200 (in Chinese).
- [21] Mentzer F, Toderici GD, Tschannen M, Agustsson E. High-fidelity generative image compression. In: *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada: Curran Associates, 2020. 11913-11924.

- [22] Muckley MJ, El-Nouby A, Ullrich K, Jegou H, Verbeek J. Improving statistical fidelity for neural image compression with implicit local likelihood models. In: Proc. of the 40th Int'l Conf. on Machine Learning (ICML). Honolulu, HI, USA: PMLR, 2023. 25426-25443.
- [23] Lei E, Uslu YB, Hassani H, Saeedi Bidokhti S. Text + sketch: Image compression at ultra-low rates. In: Proc. of the ICML 2023 Workshop on Neural Compression: From Information Theory to Applications (NCW). Honolulu, HI, USA, 2023. 1-10.
- [24] Zhang Z, Qiao X, Li Z, An N, Kong H. Toward extreme image compression with one-step diffusion and quantization semantics. Pattern Recognition and Artificial Intelligence, 2026,39(2):157-169 (in Chinese).
- [25] Li C, Lu G, Feng D, Wu H, Zhang Z, Liu X, Zhai G, Lin W, Zhang W. MISC: Ultra-low bitrate image semantic compression driven by large multimodal model. IEEE Trans. on Image Processing, 2025,34:335-349.
- [26] Liu J, Wei Y, Lin J, Zhao S, Sun H, Chen Z, Zeng W, Jin X. Semantics disentanglement and composition for universal image coding with efficiently LLM reasoning and generative diffusion. IEEE Trans. on Circuits and Systems for Video Technology, 2026.
- [27] Sargent K, Gao R, Henzler P, Herrmann C, Holynski A, Fei-Fei L, Wu J, Zhang JY. VLIC: Vision-language models as perceptual judges for human-aligned image compression. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Denver, CO, USA: IEEE, 2026. 10462-10471.
- [28] Hafner D, Lillicrap T, Norouzi M, Ba J. Mastering atari with discrete world models. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). (Virtual), 2021. 1-15.
- [29] Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, Lillicrap T, Silver D. Mastering atari, go, chess and shogi by planning with a learned model. Nature, 2020,588(7839):604-609.
- [30] Yang S, Du Y, Ghasemipour SKS, Tompson J, Kaelbling LP, Schuurmans D, Abbeel P. Learning interactive real-world simulators. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). Vienna, Austria, 2024. 1-14.
- [31] Hu Y, Yang J, Chen L, Li K, Sima C, Zhu X, Chai S, Du S, Lin T, Wang W, Lu L, Jia X, Liu Q, Dai J, Qiao Y, Li H. Planning-oriented autonomous driving. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 17853-17862.
- [32] Bruce J, Dennis MD, Edwards A, Parker-Holder J, Shi Y, Hughes E, Lai M, Mavalankar A, Steigerwald R, Apps C, Aytar Y, Bechtle S, Behbahani FMP, Chan SCY, Heess N, Gonzalez L, Osindero S, Ozair S, Reed SE, Zhang J, Zolna K, Clune J, de Freitas N, Singh S, Rocktäschel T. Genie: Generative interactive environments. In: Proc. of the 41st Int'l Conf. on Machine Learning (ICML). Vienna, Austria: PMLR, 2024. 4603-4623.
- [33] Deng C, Zhu D, Li K, Gou C, Li F, Wang Z, Zhong S, Yu W, Nie X, Song Z, Shi G, Fan H. Emerging properties in unified multimodal pretraining. arXiv:2505.14683, 2025.
- [34] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick RB. Segment anything. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3992-4003.
- [35] Xue N, Jia Z, Li J, Li B, Zhang Y, Lu Y. DLF: Extreme image compression with dual-generative latent fusion. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Honolulu, HI, USA: IEEE, 2025. 19227-19236.
- [36] Askari H, Luo Y, Sun H, Roosta F. Latent refinement via flow matching for training-free linear inverse problem solving. In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems (NeurIPS). San Diego, CA, USA: Curran Associates, 2025. 108127-108170.
- [37] Toderici G, Shi W, Timofte R, Theis L, Ballé J, Agustsson E, Johnston N, Mentzer F. Workshop and challenge on learned image compression. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA: IEEE, 2020.
- [38] Agustsson E, Timofte R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA: IEEE, 2017. 1122-1131.
- [39] Xie J, Darrell T, Zettlemoyer L, Wang X. Reconstruction alignment improves unified multimodal models. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). Rio de Janeiro, Brazil, 2026.
- [40] Sheikh HR, Bovik AC. Image information and visual quality. IEEE Trans. on Image Processing, 2006,15(2):430-444.
- [41] Xue W, Zhang L, Mou X, Bovik AC. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Trans. on Image Processing, 2014,23(2):684-695.

- [42] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018. 586-595.
- [43] Ding K, Ma K, Wang S, Simoncelli EP. Image quality assessment: Unifying structure and texture similarity. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022,44(5):2567-2581.
- [44] Zhang L, Zhang L, Mou X, Zhang D. FSIM: A feature similarity index for image quality assessment. IEEE Trans. on Image Processing, 2011,20(8):2378-2386.
- [45] Prashnani E, Cai H, Mostofi Y, Sen P. PIEAPP: Perceptual image-error assessment through pairwise preference. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018. 1808-1817.
- [46] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems (NIPS). Lake Tahoe, NV, USA: Curran Associates, 2012. 1106-1114.
- [47] Mittal A, Soundararajan R, Bovik AC. Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters, 2013,20(3):209-212.
- [48] Ke J, Wang Q, Wang Y, Milanfar P, Yang F. MUSIQ: Multi-scale image quality transformer. In: Proc. of the IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montréal, Canada: IEEE, 2021. 5128-5137.
- [49] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems (NIPS). Long Beach, CA, USA: Curran Associates, 2017. 6626-6637.
- [50] Bross B, Wang YK, Ye Y, Liu S, Chen J, Sullivan GJ, Ohm JR. Overview of the versatile video coding (VVC) standard and its applications. IEEE Trans. on Circuits and Systems for Video Technology, 2021,31(10):3736-3764.
- [51] Agustsson E, Minnen D, Toderici G, Mentzer F. Multi-realism image compression with a conditional generator. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 22324-22333.
- [52] Yang R, Mandt S. Lossy image compression with conditional diffusion models. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems (NeurIPS). New Orleans, LA, USA: Curran Associates, 2023. 64971-64995.
- [53] Kuang H, Ma Y, Yang W, Guo Z, Liu J. Consistency guided diffusion model with neural syntax for perceptual image compression. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia (MM). Melbourne, Australia: ACM, 2024. 1622-1631.
- [54] Kuang H, Yang W, Guo Z, Liu J. Syntax-driven multi-realism image compression with consistency guided diffusion model. IEEE Trans. on Circuits and Systems for Video Technology, 2026,36(5):6211-6224.
- [55] Iwai S, Miyazaki T, Omachi S. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In: Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, 2024. 2888-2897.

附中文参考文献

- [17] 杨文瀚,刘家瑛,夏思烽,郭宗明.数据外补偿的深度网络超分辨率重建.软件学报,2018,29(4):900-913.
- [18] 丁玲,丁世飞,张健,张子晨.使用 VGG 能量损失的单图像超分辨率重建.软件学报,2021,32(11):3659-3668.
- [20] 贾川民,马海川,杨文瀚,任文琦,潘金山,刘东,刘家瑛,马思伟.视频处理与压缩技术.中国图象图形学报, 2021,26(6):1179-1200.
- [24] 张洲弘,乔欣,李智远,安宁,孔贺.基于单步扩散和量化语义的极端图像压缩方法.模式识别与人工智能,2026,39(02):157-169.