2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICME57554.2024.10687549

Unifying Generation and Compression: Ultra-low bitrate Image Coding Via Multi-stage Transformer

Naifu Xue[1], Qi Mao[2*], Zijian Wang[1], Yuan Zhang[2] and Siwei Ma[3]
[1]School of Information and Communication Engineering,
Communication University of China, Beijing, China
[2]State Key Laboratory of Media Convergence and Communication,
Communication University of China, Beijing, China
[3]School of Electronics Engineering and Computer Science,
Peking University, Beijing, China
{aaronxuenf, qimao, wangzijian, yzhang}@cuc.edu.cn, {swma}@pku.edu.cn

***Abstract*—Recent progress in generative compression technology has significantly improved the perceptual quality of compressed data. However, these advancements primarily focus on producing high-frequency details, often overlooking the ability of generative models to capture the prior distribution of image content, thus impeding further bitrate reduction in extreme compression scenarios (< 0.05 bpp). Motivated by the capabilities of predictive language models for lossless compression, this paper introduces a novel *Unified Image Generation-Compression* (UIGC) paradigm, merging the processes of generation and compression. A key feature of the UIGC framework is the adoption of vector-quantized (VQ) image models for tokenization, alongside a multi-stage transformer designed to exploit spatial contextual information for modeling the prior distribution. As such, the dual-purpose framework effectively utilizes the learned prior for entropy estimation and assists in the regeneration of lost tokens. Extensive experiments demonstrate the superiority of the proposed UIGC framework over existing codecs in perceptual quality and human perception, particularly in ultra-low bitrate scenarios (≤ 0.03 bpp), pioneering a new direction in generative compression.**

***Index Terms*—Generative Compression, Extreme Compression, Image Generation, VQGANs, Transformer**

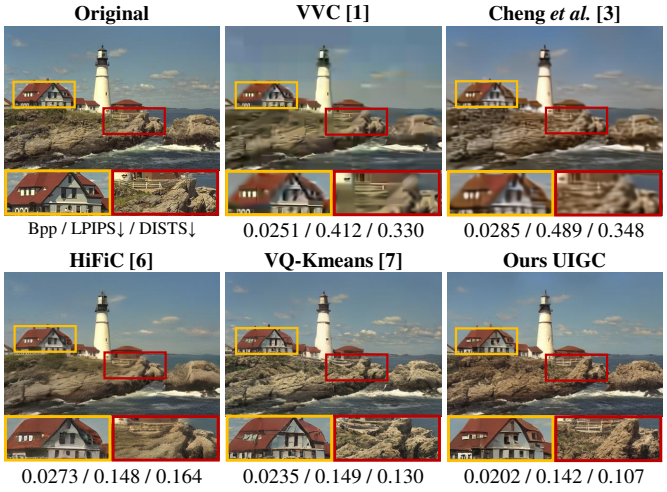


Fig. 1: **Qualitative comparisons between state-of-the-art image compression methods**, including traditional [1], learning-based [3], generative-based [6], [7], and Ours.

I. Introduction

Ultra-low bitrate compression presents a significant challenge in the field of image/video compression, particularly due to substantial information loss when faced with extremely limited network bandwidth, such as in satellite communications. Traditional block-based compression codecs, *e.g.*, VVC [1], are constrained to use large quantization steps in such scenarios, inevitably leading to noticeable blurring and blocking artifacts. Despite the superior rate-distortion (R-D) performance of learning-based compression techniques [2]–[5], these methods produce blurry images at low bitrates, due to the reliance on optimization of pixel-oriented distortion metrics.

*Corresponding author: Qi Mao, qimao@cuc.edu.cn. This work was supported in part by the National Natural Science Foundation of China under Grants 62201526; in part by the National Key Research and Development Project of China under Grant 2022YFF0902402; in part by the Fundamental Research Funds for the Central Universities (CUC23GZ007); and in part by the Public Computing Cloud at CUC, all of which are gratefully acknowledged.

To address challenges in ultra-low bitrate scenarios, generative compression methods [6]–[14] have employed generative models [15]–[17] to enhance the visual quality of decoded images, with a focus primarily on the generator's ability to *produce high-frequency details*. This paradigm follows two primary technical pathways: one involves training existing end-to-end image codecs using perceptual and adversarial losses [6], [8]–[10], and the other [7], [11]–[14] leverages specially designed encoders to compress images into more compact representations. However, despite their effectiveness, these methods tend to overlook *modeling the prior distribution of image content*, a critical aspect that differentiates image generation from image reconstruction task. In situations where significant information loss occurs due to extremely limited bandwidth, it is plausible to reconstruct some of the lost content by sampling from the prior distribution.

Meanwhile, the fundamental aspect of entropy estimation requires accurately determining the prior probability distribution of symbols, thereby boosting the efficiency of entropy

979-8-3503-9015-5/24/$31.00 ©2024 IEEE

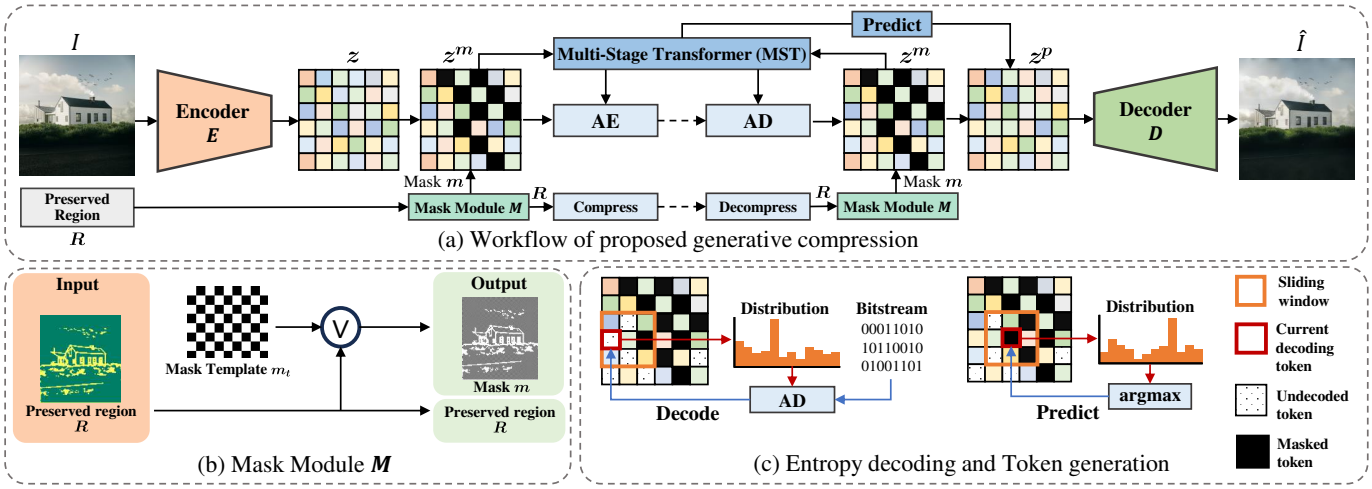


Fig. 2: **Overview of the proposed UIGC framework.** (a) the overall compression workflow: we adopt a multi-stage transformer, and AE/AD denote arithmetic encoder/decoder; (b) the mask mechanism using mask module $\boldsymbol{M}$, and $\vee$ denotes the logical "OR" operator; (c) entropy decoding and token generation on the decoder side.

coding. Consequently, the mathematical equivalence in estimating this *prior probability distribution*—between entropy minimization in lossless compression and $\log_2$-likelihood maximization in generation—poses a critical inquiry: *Is it possible to develop a method that models the prior distribution for both entropy estimation in compression and sampling in generation, all within a cohesive and unified framework?*

Recently, in the field of natural language process (NLP), Delétang *et al.* [18] have demonstrated that sequence generation models, such as large language models (LLMs), can be effectively used for lossless compression. Nevertheless, the extensive representation space of images presents a significant challenge in efficiently modeling the prior distribution. On the brighter side, advancements in Vector-Quantized Image Modeling (VIM) [19], [20] have made strides in compressing images into compact and discrete token representations using the vector-quantized (VQ) encoder. This development paves the way for transforming images into compact and discrete token representations via VIM, which enables the utilization of discrete generative models, similar to LLMs, for both entropy estimation and token generation.

In this work, we present a novel image compression paradigm, the Unified Image Generation-Compression (UIGC) framework, innovatively designed to facilitate *both entropy encoding of tokens* and *the prediction of lost tokens*. By converting images into discrete token representations using VIM [19], our UIGC codec focuses on accurate prior modeling of these tokens and the strategic discarding of nonessential tokens, leading to enhanced bitrate reduction while still producing perceptually pleasing images. Departing from the traditional autoregressive [19] and non-autoregressive [20] models typically used in NLP, we propose a **Multi-Stage Transformer** (MST) specifically tailored to image characteristics. The MST restructures the autoregressive order by dividing the token map into four groups, enabling most tokens to effectively utilize the surrounding context for prediction. Recognizing the crucial role of structural information in visual perception, we incorporate an edge-preserved checkerboard mask pattern, which selectively discards tokens while maintaining essential structural details. With MST's multi-stage order, the prediction of lost tokens is significantly enhanced, utilizing the surrounding content to ensure the generation of high-quality images.

To evaluate the efficiency of the proposed UIGC framework, we conduct experiments on the Kodak [21] and CLIC [22] datasets. The experimental results, both quantitative and qualitative, demonstrate that our method surpasses existing techniques in maintaining perceptual quality under ultra-low bitrate conditions ($\leq$ **0.03 bpp**). As shown in Fig. 1, our framework effectively reduces the bitrate while maintaining uncompromising image quality.

II. Unifying Generation and Compression

In this work, we aim to compress the image I at ultra-low bitrates while reconstructing the image $\hat{I}$ with a pleasing perceptual quality. In contrast to previous generative compression approaches that predominantly concentrate on the reconstruction of high-frequency details, our proposed UIGC framework shifts its emphasis toward modeling the prior distribution of image content for both *entropy estimation* and *content generation*. Fig. 2 presents the overview of the proposed method. In the following sections, we provide a detailed explanation of the image coding methodologies in Section II-A. Subsequently, we introduce the proposed Multi-Stage Transformer (MST) in Section II-B and the edge-preserved checkerboard mask in Section II-C to leverage spatial contextual dependencies in image token maps and effectively eliminate redundant tokens. This approach facilitates efficient prior modeling and bitrate savings while preserving image quality.

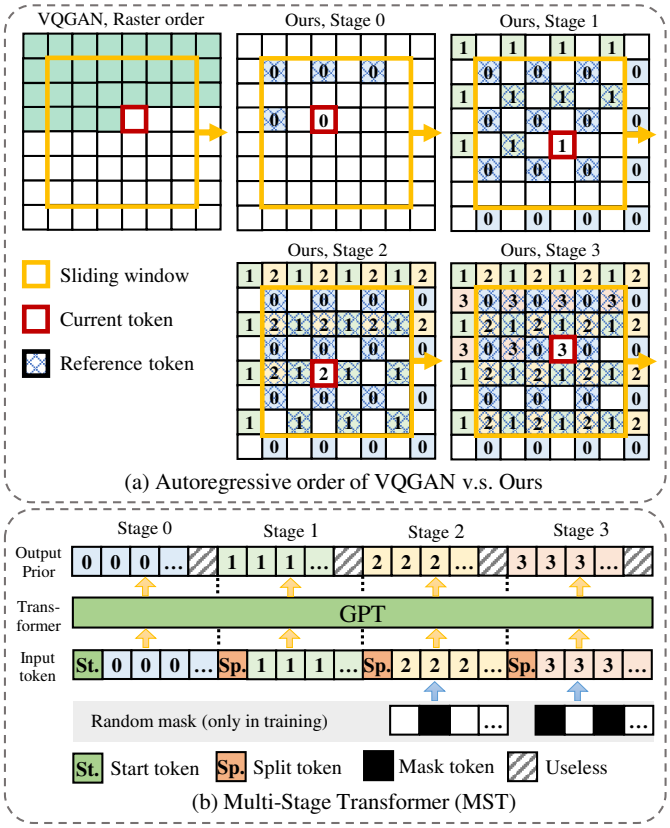


Fig. 3: **Our MST.** We implement a GPT-style transformer for the prior modeling at each stage.

A. Image Coding via UIGC

The primary goal of our UIGC framework is to leverage the prior distribution for both entropy estimation and content generation. To achieve this, we selectively compress essential tokens, discarding redundant ones to achieve bitrate savings. The discarded tokens are then generated directly on the decoder side. As illustrated in Fig. 2(a), on the encoder side, the VQ encoder $\boldsymbol{E}$ [19] transforms the given image I into a token representation $\boldsymbol{z} \in \mathbb{N}^{h \times w}$. Then, we discard redundant tokens using the mask mechanism in Section II-C, where the mask $\boldsymbol{m} \in \{0,1\}^{h \times w}$ replaces discarded tokens as [Mask]:

$$z^m_{i,j} = m_{i,j}\, z_{i,j} + (1 - m_{i,j})[\text{Mask}]. \quad i \in h, j \in w \tag{1}$$

Subsequently, the masked token map $\boldsymbol{z}^m$ is compressed by the arithmetic compression codec using the prior distribution of the MST, while the [Mask] tokens are skipped. Since the mask $\boldsymbol{m}$ is generated by our Mask Module $\boldsymbol{M}$ according to the preserved region $\boldsymbol{R}$, we further losslessly compress this region and transmit it to the decoder side.

On the decoder side, we restore the layout of the token map $\boldsymbol{z}^m$ using the decoded mask $\boldsymbol{m}$. As illustrated in Fig. 2(c), the unmasked tokens undergo decompression using the prior distribution from the MST. Simultaneously, the masked tokens are predicted with the highest probability from this prior distribution. Thus, the $\boldsymbol{z}^p$ is derived as:

$$z^p_{i,j} = m_{i,j}\, z^m_{i,j} + (1 - m_{i,j})\,(\text{argmax}\, p_{i,j}). \tag{2}$$

Finally, the VQ decoder $\boldsymbol{D}$ [19] utilizes $\boldsymbol{z}^p$ to reconstruct the decoded image $\hat{I}$.

B. Multi-Stage Transformer

As illustrated in Fig. 3(a), the transformer in VQGAN [19] employs a sliding window for raster order autoregressive encoding to manage memory usage. However, this design limits the current encoding position to consider only its *upper-left* context for prediction, potentially compromising the accuracy of prior modeling. Recognizing the spatial correlation dependencies in images, we introduce the MST inspired by the multi-stage grouping algorithm in [5] to enhance this accuracy by rearranging the autoregressive order. In particular, the token map is partitioned into four groups, and each group undergoes processing in raster order using a sliding window, as presented in Fig. 3(a).

Consequently, the MST is structured into four distinct stages. In Stage 0, tokens in Group 0 are sequentially encoded, with each token referencing only its upper-left content within the group. Following this, Stage 1 encodes tokens in Group 1, allowing each token to reference surrounding tokens in Group 0 and upper-left tokens in Group 1. This sequential encoding pattern persists in Stages 2 and 3, enabling each token to reference surrounding tokens in preceding groups and upper-left tokens in the current group. Each group of tokens in the sliding window is flattened in raster order and inputted into each stage of the transformer, as depicted in Fig. 3(b). We utilize a GPT-style transformer for autoregressive encoding, aiming to maximize the $\log_2$-likelihood of tokens defined as:

$$L_{\text{transformer}} = \mathbb{E}_{I \sim p(I)}[\sum_{k=0}^{h \times w} -\log_2 p(z_k | z^m_{\leq k})]. \tag{3}$$

During training, a random mask is applied to groups 2-3 to simulate lost tokens, and the transformer estimates the categorical distribution as the prior.

C. Edge-Preserved Checkerboard Mask

Another essential concern is the strategic discarding of redundant tokens to achieve bitrate savings without compromising image quality. In our proposed MST, tokens in Group 0 and Group 1 serve as anchors, providing surrounding references for all tokens in the subsequent groups. Hence, a *checkerboard pattern* is used as the mask template, retaining all tokens in both groups to ensure accurate prior modeling. Furthermore, tokens associated with the object structure are preserved, recognizing their crucial role in visual perception. As such, in this section, we propose an edge-preserving checkerboard mask mechanism as demonstrated in Fig. 2(b). First, we extract the object structure using the edge extractor [23], regarding it as the preserved region $\boldsymbol{R} \in \{0,1\}^{h \times w}$. Then, our proposed mask module $\boldsymbol{M}$ generates the final mask

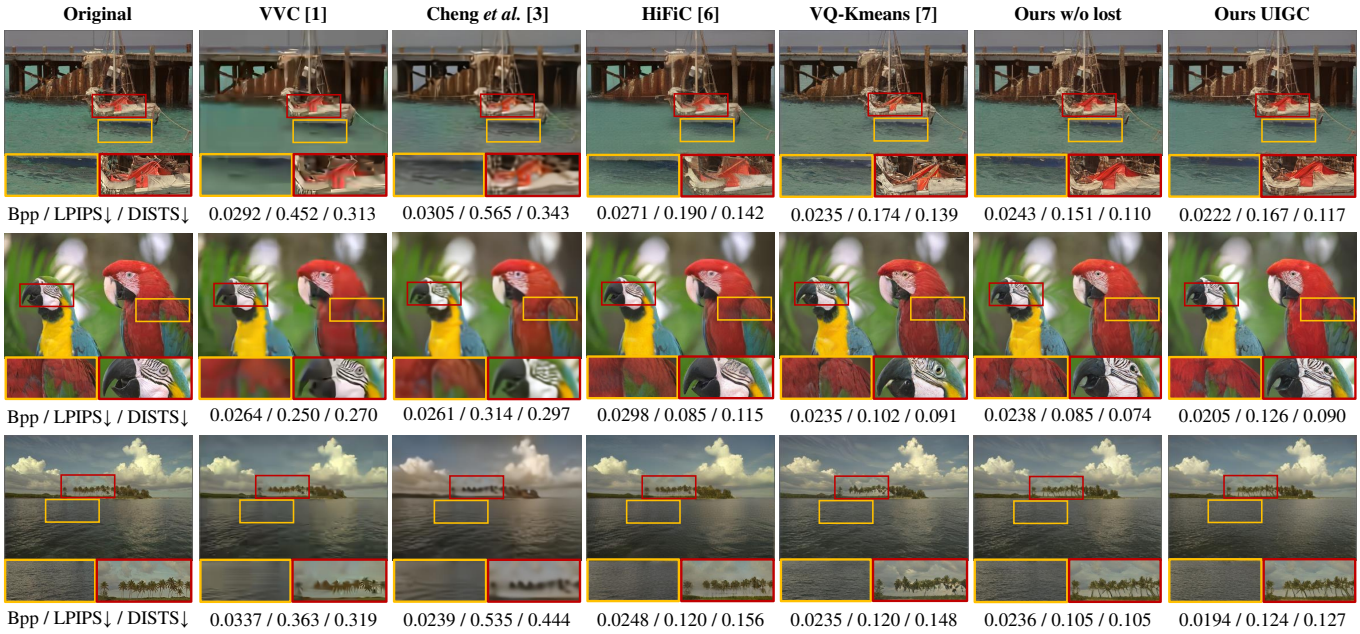
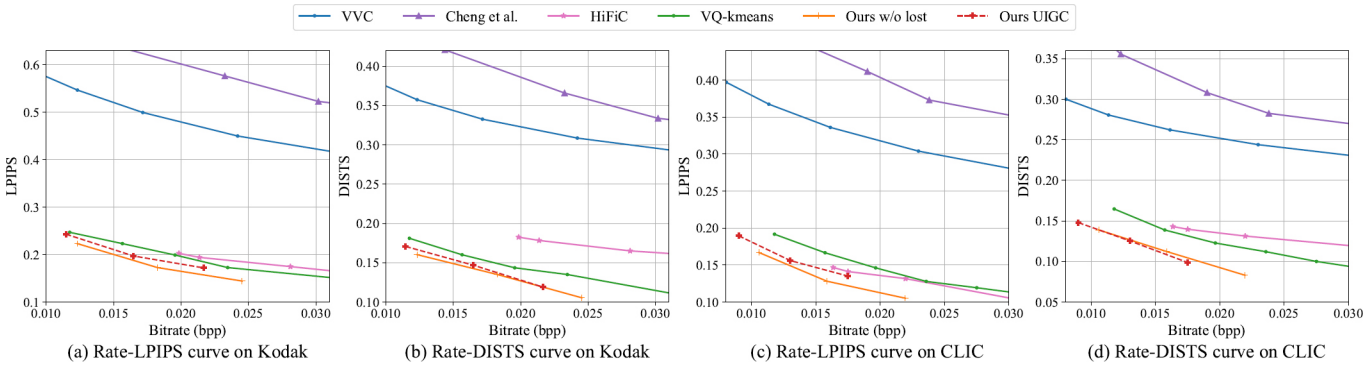


Fig. 4: **Qualitative comparisons results on the Kodak dataset [21].** In particular, ↓ indicates that lower is better.



(a) Rate-LPIPS curve on Kodak (b) Rate-DISTS curve on Kodak (c) Rate-LPIPS curve on CLIC (d) Rate-DISTS curve on CLIC

Fig. 5: **R-D curves on the Kodak [21] and the CLIC [22] datasets.** Ours w/o lost: ours method without token lost and generation.

$\boldsymbol{m} = \boldsymbol{m_t} \vee \boldsymbol{R}$, where the $\boldsymbol{m_t} \in \{0,1\}^{h \times w}$ represents the checkerboard template. The positions of Group 0 and Group 1 are set to 1 in the template $\boldsymbol{m_t}$ to retain the corresponding tokens. Our supplementary material provides further details.

III. Experiments

A. Implementation Details

We employ the architecture of encoder and decoder from VQGAN [19], and utilize the K-means clustering method detailed in [7] to fine-tune the officially provided pre-trained model with a codebook size of 16384, yielding models with codebook sizes of $\{16, 64, 256\}$ (denoted as VQ16, VQ64, and VQ256) suitable for ultra-low bitrates. During the VQ-codec training, we utilize the default settings and training losses as in [19]. For the MST, we set the size of the sliding window at 18×18. We train the proposed model on the ImageNet dataset [24]. To evaluate the performance of the proposed model, we use two widely recognized datasets in image compression: the Kodak [21] and the CLIC datasets [22].

TABLE I: **Average BD-LPIPS↓/DISTS↓ gains on the Kodak [21] and the CLIC [22] datasets.** Anchor: VVC [1].

Method	Kodak		CLIC	
	LPIPS	DISTS	LPIPS	DISTS
Cheng *et al.* [3]	0.041	0.044	0.043	0.054
HiFiC [6]	-0.260	-0.134	-0.176	-0.113
VQ-kmeans [7]	-0.288	-0.182	-0.170	-0.127
Ours w/o lost	-0.321	-0.199	-0.207	-0.151
Ours UIGC	-0.310	-0.195	-0.198	-0.150

B. Compression Performance Evaluation

Compared Methods. To assess the effectiveness of our proposed framework, we conduct a benchmark against both

TABLE II: **R-D performance of RT and MST on the Kodak dataset [21].** The codec is VQ16 and VQ256.

Method	Bpp	LPIPS↓	DISTS↓
VQ16 RT	0.0113	0.2540	0.1712
VQ16 MST	0.0115	**0.2429**	**0.1707**
VQ256 RT	0.0216	0.1895	0.1226
VQ256 MST	0.0217	**0.1720**	**0.1189**

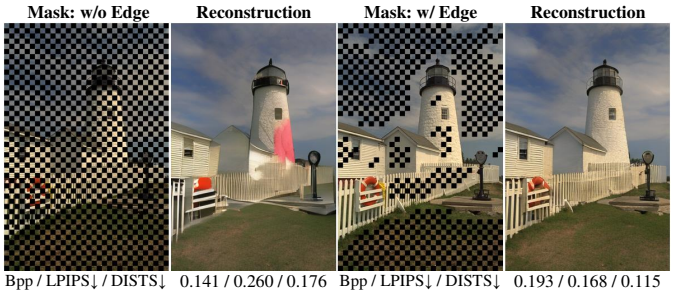


Fig. 6: **Visual comparisons between checkerboard mask with and without edge preservation.** Denote as Mask: w/ edge and Mask: w/o edge, respectively.

traditional standard and neural-based compression frameworks, including the latest VVC [1] codec, the typical end-to-end learning-based approach by Cheng *et al.* [3] (MS-SSIM optimized, for better perceptual quality), and generative image compression codecs, such as HiFiC [6], as well as the VQGAN-based codec VQ-Kmeans [7]. Furthermore, we develop an additional baseline denoted as "Ours w/o lost", which directly applies the MST for entropy estimation without lost tokens and prediction.

Quantitative Evaluation. Rather than relying on traditional objective quality assessments like PSNR and SSIM, we incorporate recent perceptual quality-based metrics, such as Learned Perceptual Image Patch Similarity (LPIPS) and Deep Image Structure and Texture Similarity (DISTS), as they offer closer alignment with human perception of images. Additionally, we employ bits per pixel (bpp) as a metric to evaluate the rate performance. We present the R-D performance in Fig. 5. We also evaluate the R-D performance improvement using VVC as an anchor with Bjontegaard-Delta metric [25]. In particular, we adopt BD-LPIPS and BD-DISTS metrics in Table I, which represent the average perceptual quality improvement under the equivalent bitrate. It can be clearly observed that our proposed UIGC exhibits superior R-D performance, delivering enhanced visual quality in ultra-low bitrate scenarios (bpp $\leq$ 0.03). Note that VQ-kmeans [7] utilizes the same VQ codec as ours. However, its R-D performance is inferior due to the absence of entropy estimation. This finding underscores the effectiveness of the UIGC method, which integrates both entropy estimation and content generation through the use of the prior distribution. While there is a slight performance drop compared to "Ours w/o lost" due to the generated content being slightly different from the real image, the UIGC framework further reduces

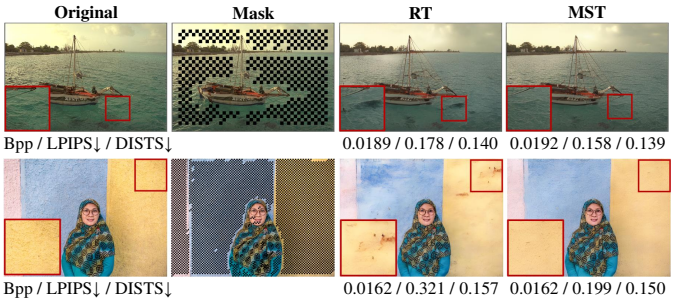


Fig. 7: **Visual Comparisons between MST and RT.** The RT method produces artifacts.

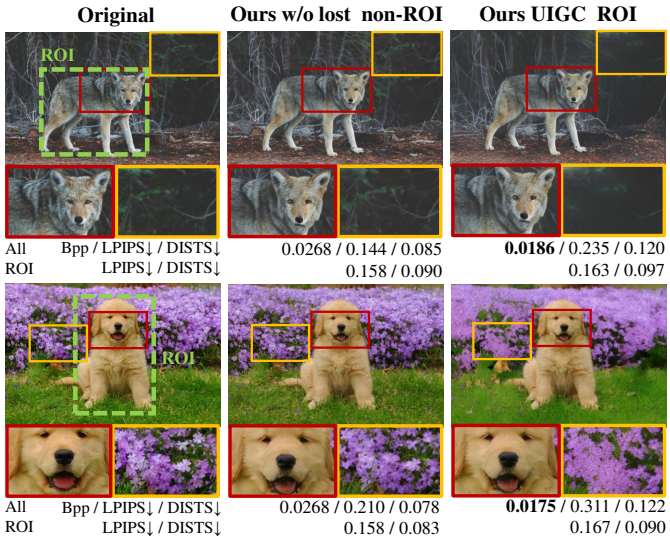


Fig. 8: **Examples of ROI coding.** The ROI area is marked by the green bounding box. "Ours UIGC ROI" ensures ROI quality while significantly lowering bitrate.

bitrate and achieves comparable perceptual quality as shown in Fig. 4.

Qualitative Evaluation. Fig. 4 shows the reconstruction results of various methods on the Kodak dataset [21], with the corresponding bpp, LPIPS, and DISTS. At ultra-low bitrates, VVC [1] and Cheng *et al.* [3] exhibit severe blurring. Although HiFiC [6] and VQ-Kmeans [7] have improved image quality, the detail is not satisfactory: HiFiC shows the grid artifact on the sea and the abnormal reconstruction of the parrot's eye; VQ-Kmeans produces distorted structure on the boat. In contrast, our methods exhibit more natural sea surface and object structures (*e.g.*, the boat, parrots, trees). Moreover, the UIGC further saves bitrate while maintaining almost the same visual quality as "Ours w/o lost", with only a negligible loss in water texture details.

C. Ablation Studies and Discussion

Efficiency of the MST. We conduct experiments to ascertain the advancements of the proposed MST over the Raster Transformer (RT) detailed in [19]. Table II indicates that

while RT and MST exhibit similar levels of entropy coding efficiency, the MST surpasses RT in the perceptual quality of the generated images. Fig. 7 presents examples of images generated by both MST and RT. The RT tends to create unnatural textures in areas like the sea and walls due to its predominant reliance on upper-left positional references, resulting in a lower visual quality compared to MST.

Mask Pattern. To evaluate the importance of tokens associated with the object structure, we test two mask patterns: checkerboard with and without edge preservation. Fig. 6 demonstrates that although excluding edge tokens contributes to further bitrate reduction, it simultaneously causes issues like distorted edges and abnormal content (for instance, the red region on the tower). This observation effectively confirms the essentiality of implementing an edge preservation mechanism in our approach.

Region of Interest Compression. Region of interest (ROI) coding, essential in multimedia applications, requires high-quality compression of selected regions while allowing for more aggressive compression in non-essential areas to reduce bitrate. Our proposed UIGC framework is adept at accommodating this need by selectively preserving tokens in the ROI. As shown in Fig. 8, UIGC's approach to ROI coding not only significantly reduces the bitrate but also maintains an aesthetically pleasing visual quality in the regions of interest.

IV. CONCLUSIONS

In this work, we propose a novel UIGC paradigm, specifically tailored for ultra-low bitrate image compression. This versatile framework adeptly utilizes the learned prior distribution for both entropy estimation and the regeneration of lost tokens. We further design the MST to boost prior modeling accuracy, and introduce an edge-preserving checkerboard mask pattern to discard unnecessary tokens for bitrate saving. Our experimental results validate the UIGC's superiority over existing codecs in visual quality, particularly in ultra-low bitrate ($\leq$ 0.03 bpp) scenarios. We believe that the UIGC scheme represents a significant advancement in generative compression, charting a new course for future developments.

REFERENCES

[1] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. 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.

[2] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[3] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 771–14 780.

[5] M. Lu, F. Chen, S. Pu, and Z. Ma, "High-efficiency lossy image coding through adaptive neighborhood information aggregation," *arXiv preprint arXiv:2204.11448*, 2022.

[6] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 11 913–11 924.

[7] Q. Mao, T. Yang, Y. Zhang, S. Pan, M. Wang, S. Wang, and S. Ma, "Extreme image compression using fine-tuned vqgans," *arXiv preprint arXiv:2307.08265*, 2023.

[8] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] S. Iwai, T. Miyazaki, Y. Sugaya, and S. Omachi, "Fidelity-controllable extreme image compression with generative adversarial networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 8235–8242.

[10] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-realism image compression with a conditional generator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 324–22 333.

[11] J. Chang, Q. Mao, Z. Zhao, S. Wang, S. Wang, H. Zhu, and S. Ma, "Layered conceptual image compression via deep semantic synthesis," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 694–698.

[12] J. Chang, Z. Zhao, C. Jia, S. Wang, L. Yang, Q. Mao, J. Zhang, and S. Ma, "Conceptual compression via deep structure and texture synthesis," *IEEE Transactions on Image Processing*, vol. 31, pp. 2809–2823, 2022.

[13] J. Chang, Z. Zhao, L. Yang, C. Jia, J. Zhang, and S. Ma, "Thousand to one: Semantic prior modeling for conceptual coding," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[14] J. Chang, J. Zhang, J. Li, S. Wang, Q. Mao, C. Jia, S. Ma, and W. Gao, "Semantic-aware visual decomposition for image coding," *International Journal of Computer Vision*, vol. 131, no. 9, pp. 2333–2355, 2023.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[18] G. Delétang, A. Ruoss, P.-A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau *et al.*, "Language modeling is compression," 2024.

[19] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 873–12 883.

[20] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 315–11 325.

[21] E. Kodak, "Kodak photocd dataset," 2013.

[22] T. George, S. Wenzhe, T. Radu, T. Lucas, B. Johannes, A. Eirikur, J. Nick, and M. Fabian, "Workshop and challenge on learned image compression (clic2020)," CVPR, 2020.

[23] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[25] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU SG16 Doc. VCEG-M33*, 2001.