# PORT: Perception-Oriented Image Compression with Real-Time Decoding

Wei Jiang[1], Bohao Feng[2], Wenqiang Wang[2], Bo Huang[2], Lin Ding[2], Ronggang Wang[1,3,4*]

[1]Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology
Shenzhen Graduate School, Peking University

[2]Alibaba [3]Pengcheng Laboratory [4]Migu Culture Technology Co., Ltd

jiangwei@stu.pku.edu.cn rgwang@pkusz.edu.cn

## Abstract

This white paper presents PORT, a Perception-Oriented image compression framework with Real-Time decoding. PORT is our approach for the image compression track at CLIC 2025. To enhance perceptual quality, we incorporate both semantic and patch-wise adversarial losses to generate realistic textures, and employ a region-of-interest (ROI) mask to guide bit allocation across different regions. To accelerate decoding, PORT builds upon the DCVC-RT architecture, while introducing more advanced entropy models to capture long-range correlations. Our team is PKUSZ-AliMerlin.

## 1 Introduction

Learned image compression [1] has become an active research area in recent years. Several models [2–7] have already outperformed the latest non-learned codec, VVC. However, most existing methods primarily focus on optimizing non-perceptual distortion metrics. To improve perceptual quality, some approaches [8–10] employ Generative Adversarial Networks (GANs) [11] to synthesize perceptually convincing textures. In addition, learned perceptual metrics such as VGG-based [12] or LPIPS [13] losses are often used to stabilize training and improve convergence.

Since subjective quality standards vary across users, content-sensitive images (e.g., faces and documents) require authenticity preservation rather than the generation of vivid yet unrealistic details. To this end, building upon the generative compression framework of [9], we propose PORT: Perception-Oriented Image Compression with Real-Time Decoding, developed for the CLIC 2025 Image Track. PORT employs semantic- and patch-wise discriminators to guide texture synthesis, thereby enhancing semantic and spatial consistency between generated and reference images. Furthermore, a region-of-interest (ROI) mask is used to adaptively control pixel-level distortion weights, enabling flexible rate allocation.

For efficient decoding, PORT adopts the architecture of DCVC-RT, while incorporating a more advanced entropy model to capture long-range correlations. Finally, to meet specific bit-rate constraints, our model is trained with a variable-rate compression strategy inspired by [14]. Our team is PKUSZ-AliMerlin.

* Corresponding author.

## 2 Method

### *2.1 Overview*

In this section, we briefly introduce our solution PORT to the CLIC 2025 Image Track. The overall architecture of PORT follows DCVC-RT [15] and Ma et al. [16], while the gain and inverse gain units [14] are adopted for variable-rate control.

The optimization process consists of two stages. In the first stage, the model is trained with a mean squared error (MSE) objective, together with the gain and inverse gain units. In the second stage, the model is further optimized for perceptual quality.

For perceptual optimization, we employ both patch-wise and semantic adversarial losses [11] to guide the decoder in generating realistic textures at low bitrates. To preserve sharpness, we additionally use the Charbonnier loss. The LPIPS loss [13] and style loss [17] are incorporated to enhance perceptual fidelity, while the Laplacian loss [18] is used to mitigate color variations. Following [16], a region-of-interest (ROI) mask is introduced to allocate more bits to semantically important regions.

The overall loss function is formulated as

$$\begin{aligned}\mathcal{L} = \lambda_r \times \mathcal{R} + \lambda_{mse} \times (\delta \odot \mathcal{D}_{mse}) + \lambda_{L1}^{ROI} \times (\delta \odot \mathcal{D}_{L1}) + \\ \lambda_{L1}^{non-ROI} \times [(1-\delta) \odot \mathcal{D}_{L1}] + \lambda_{lpips} \times \mathcal{D}_{lpips} + \\ \lambda_{sty} \times \mathcal{D}_{sty} + \lambda_{lap} \times \mathcal{D}_{lap} + \lambda_{adv} \times \mathcal{D}_{adv},\end{aligned} \tag{1}$$

where $\mathcal{R}$ is the rate, $\mathcal{D}_{mse}$ is the MSE distortion, $\mathcal{D}_{L1}$ is the L1 distortion, $\mathcal{D}_{lpips}$ is the VGG-16 [12] Lpips [13] distortion, $\mathcal{D}_{sty}$ is the style loss [17], $\mathcal{D}_{lap}$ is the Laplician distortion [18], $\mathcal{D}_{adv}$ is the BCE adversarial distortion, $\delta$ is the ROI map. We use $\{\lambda_r, \lambda_{mse}, \lambda_{L1}^{ROI}, \lambda_{L1}^{non-ROI}, \lambda_{lpips}, \lambda_{sty}, \lambda_{lap}, \lambda_{adv}\}$ to adjust the weight of each loss.

### *2.2 ROI-Weighted Distortion and Bit Allocation*

In our method, following Ma et al. [16], we employ a saliency map as the ROI mask, since saliency detection naturally separates an image into focused regions and background, which aligns well with our allocation strategy. Specifically, we adopt RMformer [19] to generate ROI maps. The RMformer is kept fixed during both training and testing. The ROI detection process is formulated as

$$Mask_{2D} = \text{sigmoid}(\text{RMformer}(\boldsymbol{x})), \tag{2}$$

where $\boldsymbol{x}$ is the input image, and $Mask_{2D}$ is the detected ROI map.

To further smooth boundaries and facilitate gradual bit allocation, the ROI map is processed with average pooling:

$$RM_{2D} = \text{AvgPool}(Mask_{2D}), \tag{3}$$

where $RM_{2D}$ denotes the smoothed ROI map. Additionally, a scaling factor $\alpha$ is introduced to control the relative importance of ROI regions during rate allocation.

To avoid background texture degradation, a fixed number of channels are reserved to retain sufficient information for non-ROI regions.

Finally, bit allocation is performed on the encoder side as follows:

$$\boldsymbol{y} = g_a(\boldsymbol{x}),\ \tilde{\boldsymbol{y}}\ = \boldsymbol{y} \odot \frac{RM_{2D} + \alpha}{1 + \alpha},\ \overline{\boldsymbol{y}} = \boldsymbol{y}_{\text{0-63}}, |, \tilde{\boldsymbol{y}}_{\text{64-320}},\ \hat{\boldsymbol{y}}\ = Q(\overline{\boldsymbol{y}}),\ \hat{\boldsymbol{x}} = g_s(\hat{\boldsymbol{y}}), \tag{4}$$

where $g_a$ and $g_s$ denote the analysis and synthesis transforms, respectively. Since L1 and MSE losses are pixel-wise, we directly use the ROI map to weight pixel-level distortions, thereby guiding the model to allocate more bits to ROI regions.

### *2.3 Adversial Training*

To generate realistic textures, we optimize the model with adversarial losses, which include both semantic-level and patch-wise components. Following prior works [9, 16], we employ the binary cross-entropy (BCE) loss for adversarial training.

For the semantic adversarial loss, we leverage CLIP [20] to extract semantic priors $\boldsymbol{s}$ from the input image, which are used as conditional inputs for discrimination:

$$\mathcal{D}^s_{adv} = \mathbb{E}\big[-\log(D^s(\hat{\boldsymbol{x}} \mid \boldsymbol{s}))\big],\ \mathcal{D}^s_{disc}\ = \mathbb{E}\big[-\log(1 - D^s(\hat{\boldsymbol{x}} \mid \boldsymbol{s}))\big] + \mathbb{E}\big[-\log(D^s(\boldsymbol{x} \mid \boldsymbol{s}))\big], \tag{5}$$

where $\hat{\boldsymbol{x}}$ denotes the reconstructed image and $\boldsymbol{x}$ the original image.

For patch-wise discrimination, the loss function is defined as:

$$\mathcal{D}^p_{adv} = \mathbb{E}\big[-\log(D^p(\hat{\boldsymbol{x}} \mid \hat{\boldsymbol{y}}))\big],\ \mathcal{D}^p_{disc}\ = \mathbb{E}\big[-\log(1 - D^p(\hat{\boldsymbol{x}} \mid \hat{\boldsymbol{y}}))\big] + \mathbb{E}\big[-\log(D^p(\boldsymbol{x} \mid \hat{\boldsymbol{y}}))\big], \tag{6}$$

where $\hat{\boldsymbol{y}}$ denotes the latent representation. Here, $D^s$ and $D^p$ denote the semantic and patch-wise discriminators, respectively. The discriminator objectives are $\mathcal{D}^s_{disc}$ and $\mathcal{D}^p_{disc}$, while the generator is optimized with the corresponding adversarial losses $\mathcal{D}^s_{adv}$ and $\mathcal{D}^p_{adv}$. The overall adversarial loss is defined as the weighted sum of the two terms:

$$\mathcal{D}adv = \lambda^s_{adv}\mathcal{D}^s_{adv} + \lambda^p_{adv}\mathcal{D}^p_{adv}, \tag{7}$$

where $\lambda^s_{adv}$ and $\lambda^p_{adv}$ are hyperparameters that balance the contributions of the semantic and patch-wise adversarial objectives.

### *2.4 Variable Rate Adaptation*

The gain units and inverse gain units are employed for continuous rate adaptation. Specifically, the model is trained to support $n$ target bitrates. The gain units, denoted as $M_G \in \mathbb{R}^{c \times n}$, adjust the quantization step of each channel, where $c$ is the number of channels. The inverse gain is implemented as $1/M_G$ in our model.

With gain adaptation, the process in Equation 4 is modified as follows:

$$
\begin{aligned}
\boldsymbol{y} &= g_a(\boldsymbol{x}), \\
\tilde{\boldsymbol{y}} &= \boldsymbol{y} \odot \frac{RM_{2D} + \alpha}{1 + \alpha}, \\
\overline{\boldsymbol{y}} &= \boldsymbol{y}_{0-47} || \tilde{\boldsymbol{y}}_{ch48-191}, \\
\hat{\boldsymbol{y}} &= Q(\overline{\boldsymbol{y}} \odot M_G), \\
\hat{\boldsymbol{x}} &= g_s(\hat{\boldsymbol{y}} \odot \frac{1}{M_G}),
\end{aligned}
\tag{8}
$$

where $g_a$ and $g_s$ denote the analysis and synthesis transforms, respectively, $Q$ is the quantization operation, and $\odot$ represents element-wise multiplication. The gain units $M_G$ scale the latent representation before quantization, while the inverse gain compensates the scaling during synthesis to reconstruct the image at the desired bitrate.

### *2.5 Entropy Model*

The entropy model is equipped with a quadtree context module [21], a global context module, and a hyperprior module. For each subtree in the quadtree context, we employ a depth-wise convolutional block [15] to capture local context, and an efficient attention block [5, 22] to extract global context.

## 3 Conclusion

In this white paper, we have presented PORT, a Perception-Oriented Image Compression framework with Real-Time Decoding for the CLIC 2025 Image Track. PORT integrates semantic- and patch-wise adversarial learning, ROI-guided bit allocation, variable-rate gain units, and an advanced entropy model with quadtree and global context modules. Extensive design choices, including model architecture, perceptual and reconstruction losses, ensure both high perceptual quality and efficient decoding. Our approach demonstrates the effectiveness of combining perceptual optimization with real-time performance, providing a practical solution for content-aware image compression.

## References


[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *Int. Conf. on Learning Representations*, 2018.

[2] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro 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.

[3] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," *arXiv preprint arXiv:2203.10886*, 2022.

[4] Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, and Ronggang Wang, "Mlic: Multi-reference entropy model for learned image compression," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7618–7627.

[5] Wei Jiang and Ronggang Wang, "Mlic++: Linear complexity multi-reference entropy modeling for learned image compression," *arXiv preprint arXiv:2307.15421*, 2023.

[6] Wei Jiang, Peirong Ning, and Ronggang Wang, "Slic: Self-conditioned adaptive transform with large-scale receptive fields for learned image compression," *arXiv preprint arXiv:2304.09571*, 2023.

[7] Wei Jiang, Yongqi Zhai, Jiayu Yang, Feng Gao, and Ronggang Wang, "Mlicv2: Enhanced multi-reference entropy modeling for learned image compression," *arXiv preprint arXiv:2504.19119*, 2025.

[8] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 221–231.

[9] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11913–11924, 2020.

[10] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer, "Multi-realism image compression with a conditional generator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22324–22333.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[14] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10532–10541.

[15] Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu, "Towards practical real-time neural video compression," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12543–12552.

[16] Yi Ma, Yongqi Zhai, Chunhui Yang, Jiayu Yang, Ruofan Wang, Jing Zhou, Kai Li, Ying Chen, and Ronggang Wang, "Variable rate roi image compression optimized for visual quality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 1936–1940.

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[18] Simon Niklaus and Feng Liu, "Context-aware synthesis for video frame interpolation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1701–1710.

[19] Xinhao Deng, Pingping Zhang, Wei Liu, and Huchuan Lu, "Recurrent multi-scale transformer for high-resolution salient object detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7413–7423.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[21] Jiahao Li, Bin Li, and Yan Lu, "Neural video compression with diverse contexts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22616–22626.

[22] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li, "Efficient attention: Attention with linear complexities," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3531–3539.