
Generative Latent Coding for Ultra-Low Bitrate Image Compression

CVPR 2024



Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, Yan Lu
University of Science and Technology of China, Microsoft Research Asia

Group Member: 蘇柏叡、許良亦、王致雅、陳晴川

23, Nov 2025

Outline

01

Introduction

02

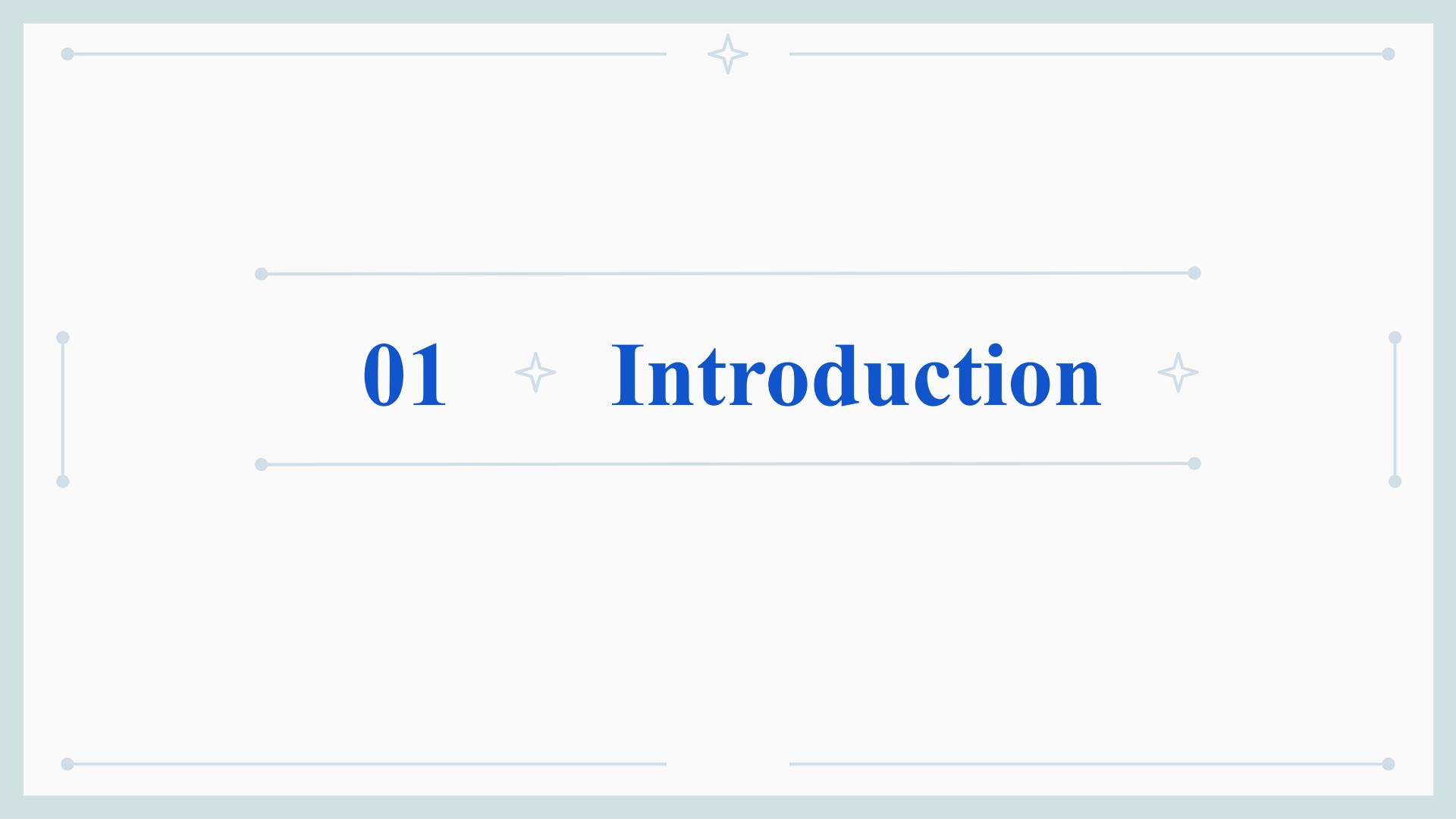
Method

03

Experiments

04

Conclusion



01 ✶ Introduction



Introduction

- **Task:** Ultra-Low Bitrate Image Compression
- **Definition:** Compress images at ultra-low bitrates in the generative latent space of a VQ-VAE, while keeping the **reconstruction visually natural** and semantically consistent.
- **Motivation:**
 - Pixel-space codecs, including most learned methods, break down at ultra-low bitrates, producing **blurry textures, artifacts**, or wrong semantics.
 - The VQ-VAE latent space is sparse and semantic, better aligned with perception, so we compress there and let the **generative decoder recover realistic details**.

Challenges in Image Compression

- **Extreme rate–distortion trade-off**

Achieving **visually acceptable reconstructions** when compressing images to only a few hundred bits imposes a very stringent rate–distortion constraint.

- **Perception–distortion mismatch**

Pixel-wise distortion measures such as MSE and PSNR **are poorly aligned** with **human visual perception**, particularly in the ultra-low bitrate regime.

- **Texture–semantics trade-off**

Under **severe bit budget limitations**, it is challenging to simultaneously preserve **sharp, realistic textures** and maintain accurate high-level semantic content.

Main Contributions of this Work

- Propose **Generative Latent Coding (GLC)**, a framework that performs ultra-low bitrate image compression in the generative latent space of a VQ-VAE instead of the pixel space.
- Design **specialized latent modeling** with a categorical hyper module and code-prediction-based supervision, enabling more efficient entropy coding of latents while preserving semantic consistency and perceptual quality.
- Demonstrate **SOTA performance** on standard image compression benchmarks under ultra-low bitrate settings, together with flexible latent-space applications such as image restoration and style modification.

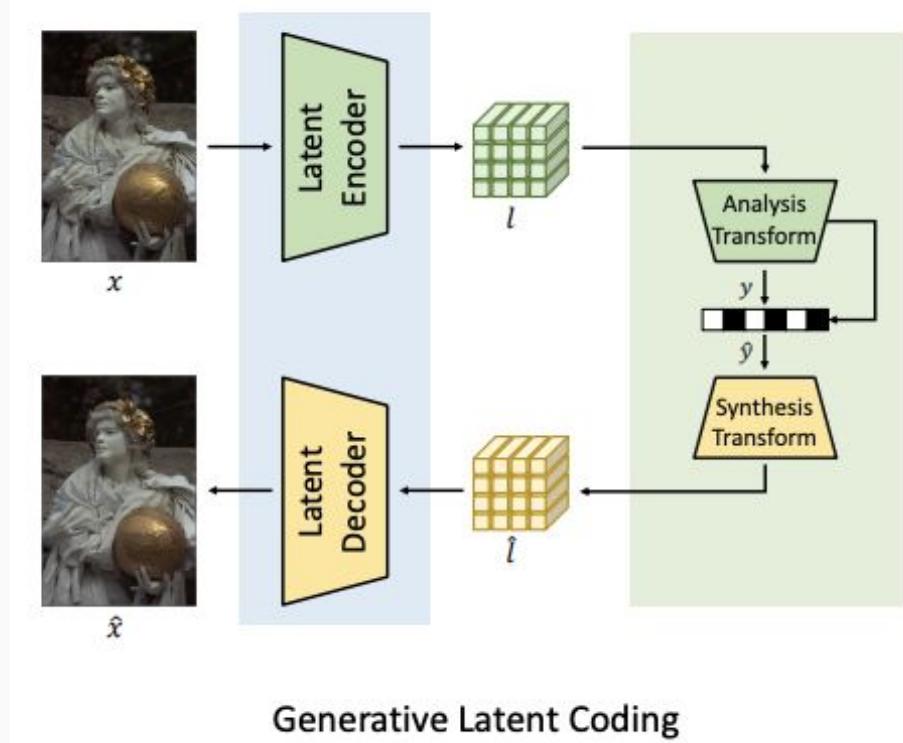
02



Method



Overview of the Generative Latent Coding (GLC) framework



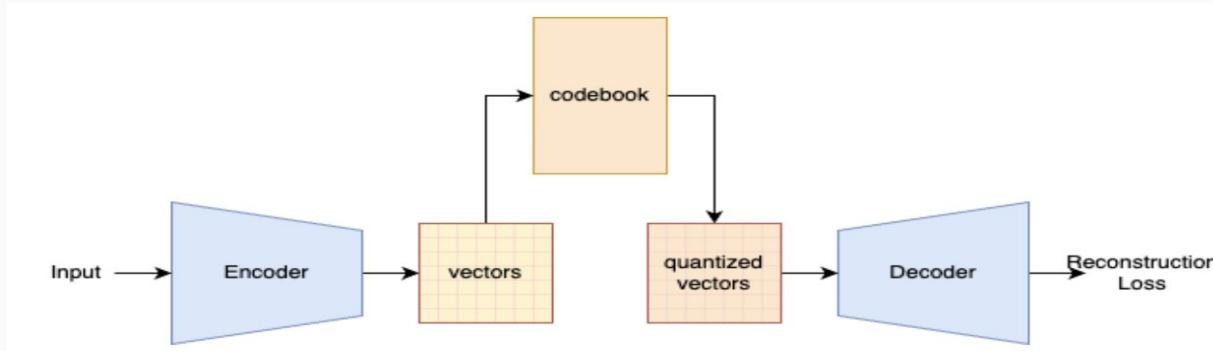
$$l = E(x), \quad y = g_a(l)$$

$$\hat{y} = Q(y)$$

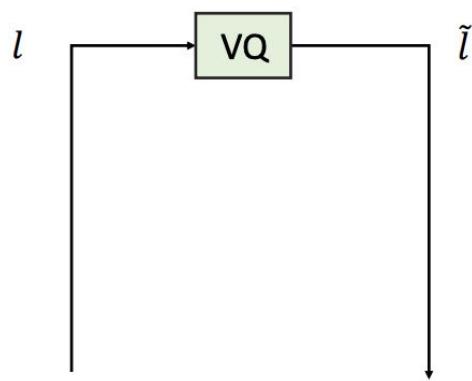
$$\hat{l} = g_s(\hat{y}), \quad \hat{x} = D(\hat{l})$$

Generative Latent Auto-Encoder

- Create a **human-perception-aligned** latent space
- Use Generative **VQ-VAE** as **encoder E** and **decoder D**
- Map images into visual semantic elements via **codebook C**
- advantages of this latent space:
 - captures semantic structure, not raw pixels
 - discrete and low-entropy, making it highly compressible
 - decoder restores realistic textures

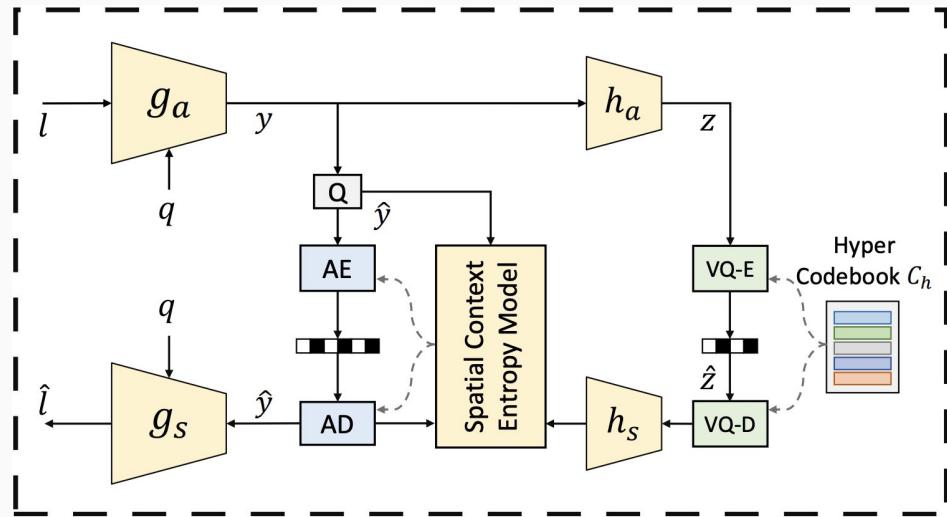


Transform Coding in Latent Space



indices-map coding

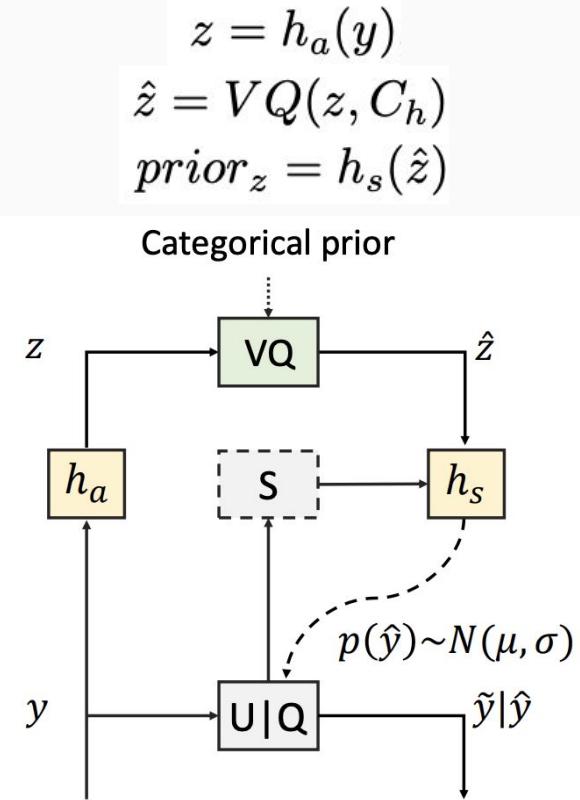
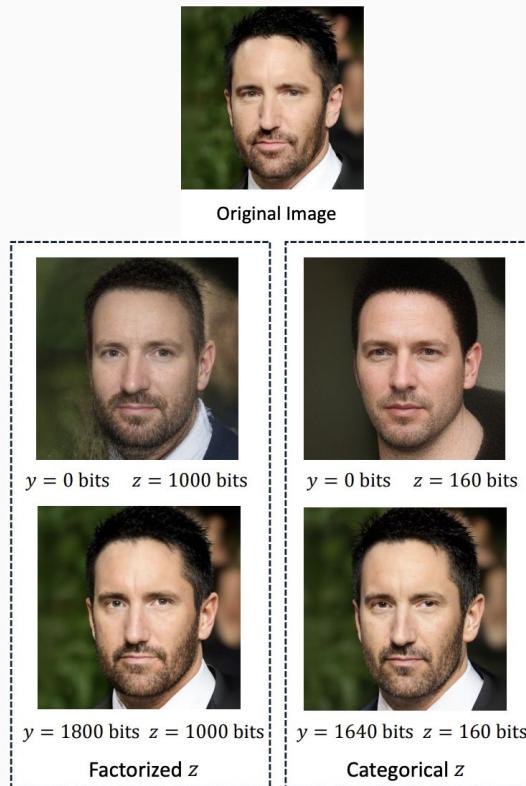
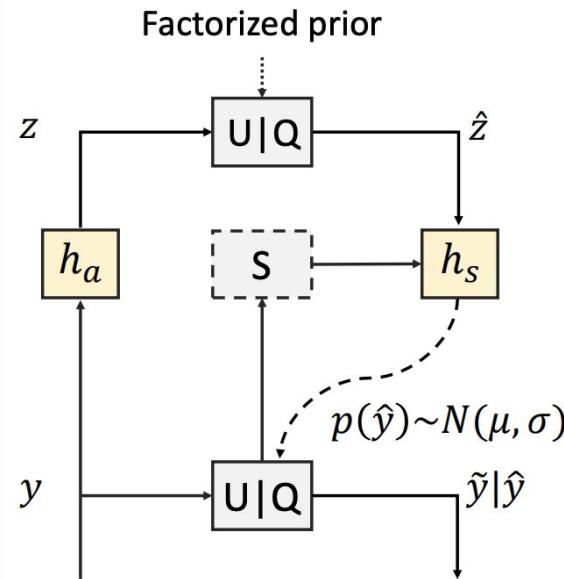
- no redundancy reduction
- fixed bitrate
- poor compression at ultra-low bitrate



transform coding

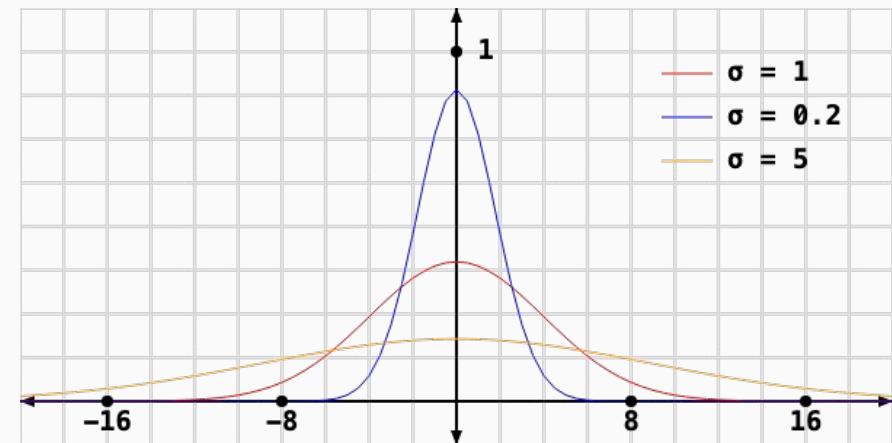
- allow continuous-valued latent representations
- Gaussian entropy modeling
- variable-rate control

Categorical Hyper Module

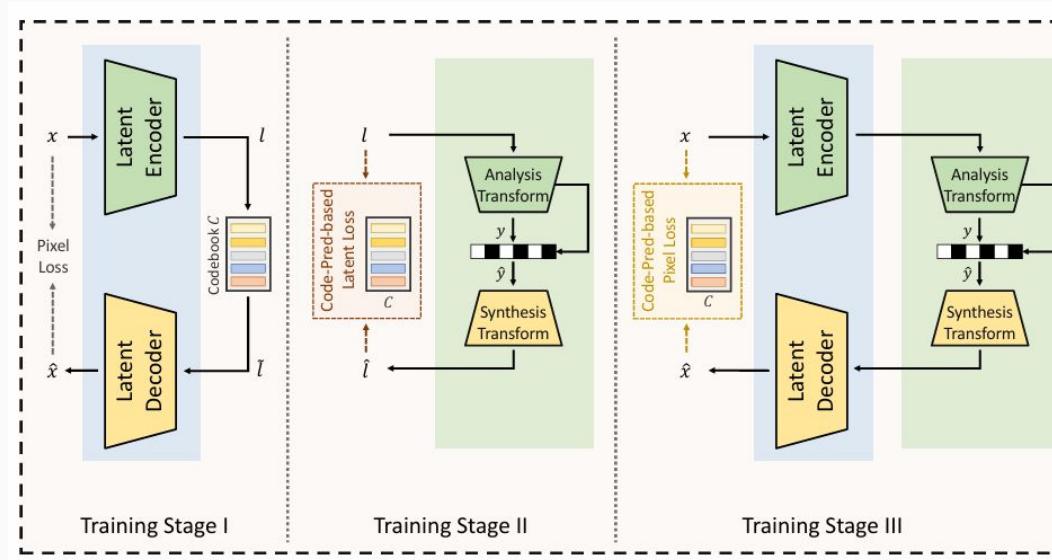


Rate-Variable Transformation

- Transform-coded latent y modeled with a Gaussian-based entropy model
- Bitrate controlled via scaling of distribution parameters
- Supports multiple bitrate operating points with one single model
- Avoids fixed-rate restriction of index-map coding
- Enables flexible trade-off between rate and quality

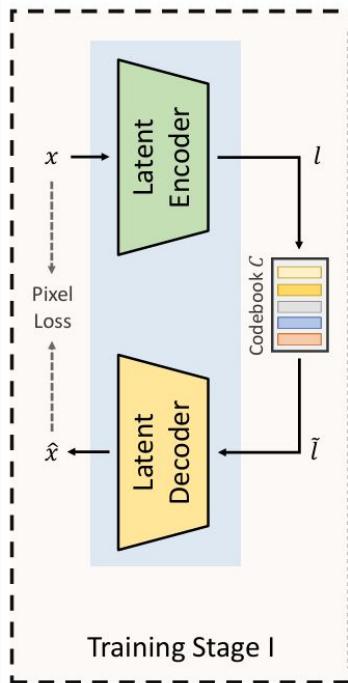


Progressive Training Strategy



- Stage I: Learn a perception-aligned latent space with VQ-VAE.
- Stage II: Train transform coding with semantic supervision.
- Stage III: Joint fine-tuning for best rate-distortion quality.

Progressive Training - Stage I



Pixel Loss:

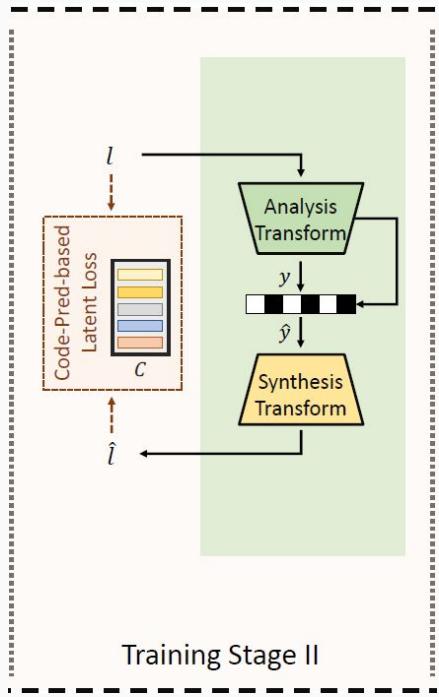
- **Reconstruction loss** $\|x - \hat{x}\|$
Ensures that the reconstructed image remains close to the original at the pixel level.
- **Perceptual Loss** $\mathcal{L}_{per}(x, \hat{x})$
Encourages the reconstruction to match the original in semantic and perceptual similarity using deep features.
- **Adversarial Loss** $\mathcal{L}_{adv}(x, \hat{x})$
Improves the realism and texture quality of the reconstructed image through adversarial training.

$$\mathcal{L}_{\text{Stage I}} = \|x - \hat{x}\| + \mathcal{L}_{per}(x, \hat{x}) + \lambda_{adv} \cdot \mathcal{L}_{adv}(x, \hat{x}) + \mathcal{L}_{codebook}$$

Codebook Loss $\mathcal{L}_{codebook}$

Stabilizes the VQ-VAE latent space by aligning encoder outputs with codebook embeddings.

Progressive Training - Stage II



- **Code-Prediction Loss**

Encourages the compressed latent \hat{l} to correctly predict the VQ code indices, preserving semantic information.

$$\mathcal{D}_{code}(l, \hat{l}) = \alpha \cdot CE(M_l, \hat{M}_{\hat{l}}) + \|l - \hat{l}\|_2^2$$

$CE(M_l, \hat{M}_{\hat{l}})$: Compares the true VQ code indices of the original latent l with the predicted code indices derived from the compressed latent \hat{l}

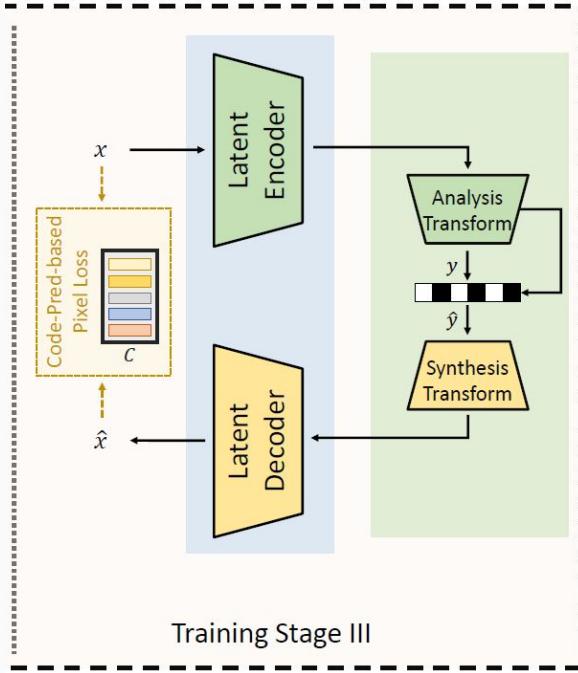
- **Rate Term**

Optimizes the trade-off between bitrate and distortion for transform coding.

$$\mathcal{L}_{\text{Stage II}} = \mathbf{E}_{x \sim p_X} [\mathcal{R}(\hat{y}) + \lambda \cdot \mathcal{D}_{code}(l, \hat{l})]$$

$R_{(\hat{y})}$: Represents the estimated bitrate of the quantized latent \hat{y} .

Progressive Training - Stage III



$$\mathcal{D}_{\text{Stage III}} = \|x - \hat{x}\| + \mathcal{L}_{\text{per}}(x, \hat{x})$$

$$+ \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{adv}}(x, \hat{x}) + \lambda_{\text{code}} \cdot \mathcal{D}_{\text{code}}(l^p, \hat{l}^p)$$

- **Pixel-Level Code-Prediction Loss**

Enforces semantic consistency by comparing the VQ-VAE latent codes of the input x and the reconstruction \hat{x} .

The rate-distortion trade-off supervision:

$$\mathcal{L}_{\text{Stage III}} = \mathbf{E}_{x \sim p_X} [\mathcal{R}(\hat{y}) + \lambda \cdot \mathcal{D}_{\text{Stage III}}]$$

03 Experiments

Datasets

- **Training**
 - Natural images ⇒ **ImageNet train set [1]**(Stage I), OpenImages test patches[2] (Stages II–III)
 - Facial images ⇒ **FFHQ 512×512 for all stages[3]** (GLC-face)
- **Evaluation datasets**
 - **CLIC 2020 test set[4]** (natural scenes, original resolution)
 - **CelebA-HQ 512×512[5]** (faces, ultra-low bitrate < 0.01 bpp)
 - Additional benchmarks in the supplementary: **Kodak[6], DIV2K[7], MS-COCO 30K[8]**
- **Evaluation metrics**
 - Bitrate: bits per pixel (bpp)
 - Perceptual metrics: **LPIPS, DISTS** (reference-based)
 - Generative quality metrics: **FID, KID** (distribution-level)

Implementation Details

- Overall GLC framework

- Generative VQ-VAE encodes images into a discrete latent space
- Transform + entropy coding are performed in latent space instead of pixel space

- Architectures

- Natural images (GLC-image): VQGAN-style auto-encoder, downsampling $\times 16$
- Faces (GLC-face): CodeFormer-style auto-encoder, downsampling $\times 32$

- Latent transform & entropy model

- Lightweight CNN-based transform in latent space
- Categorical hyper module + context model for probability estimation and entropy coding

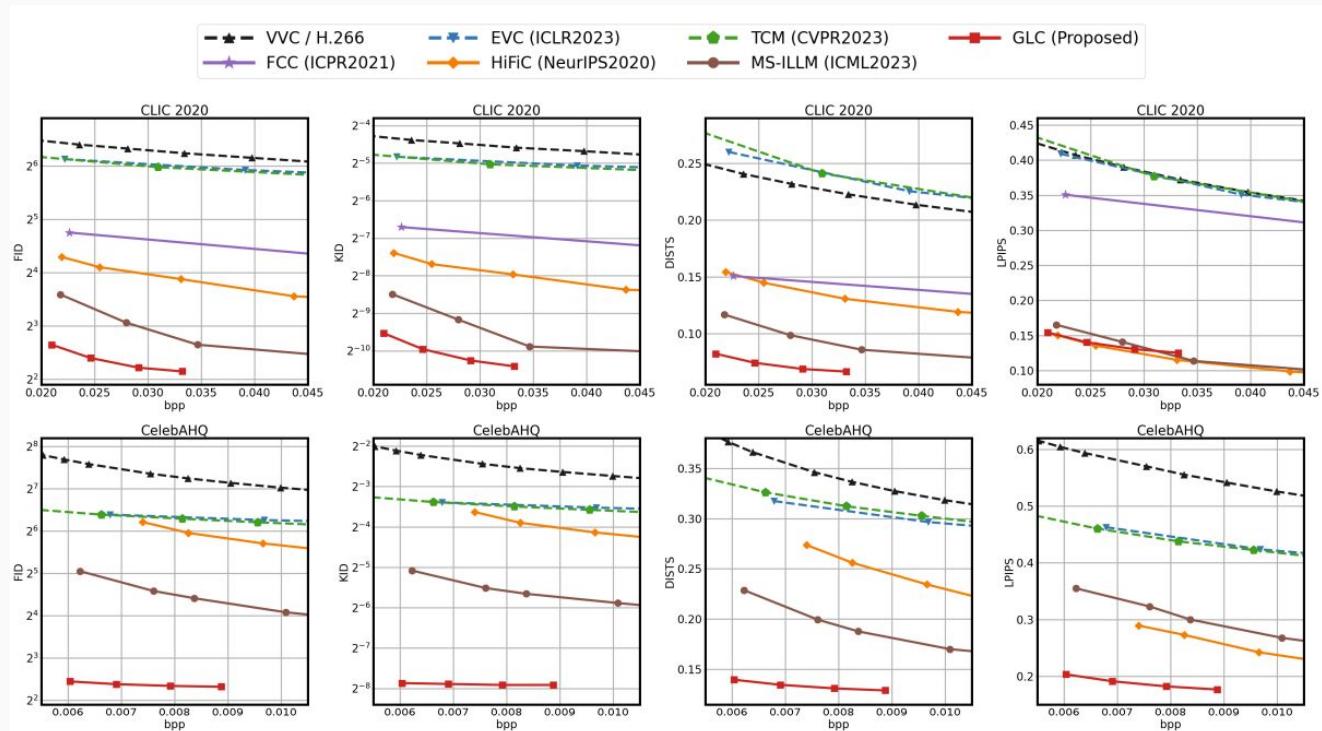
Three-stage training: first train a generative auto-encoder with L1, LPIPS, and GAN losses, then add rate-distortion loss and code-prediction-based supervision for joint fine-tuning.

Experiment Setup Details

- **Training setups:**

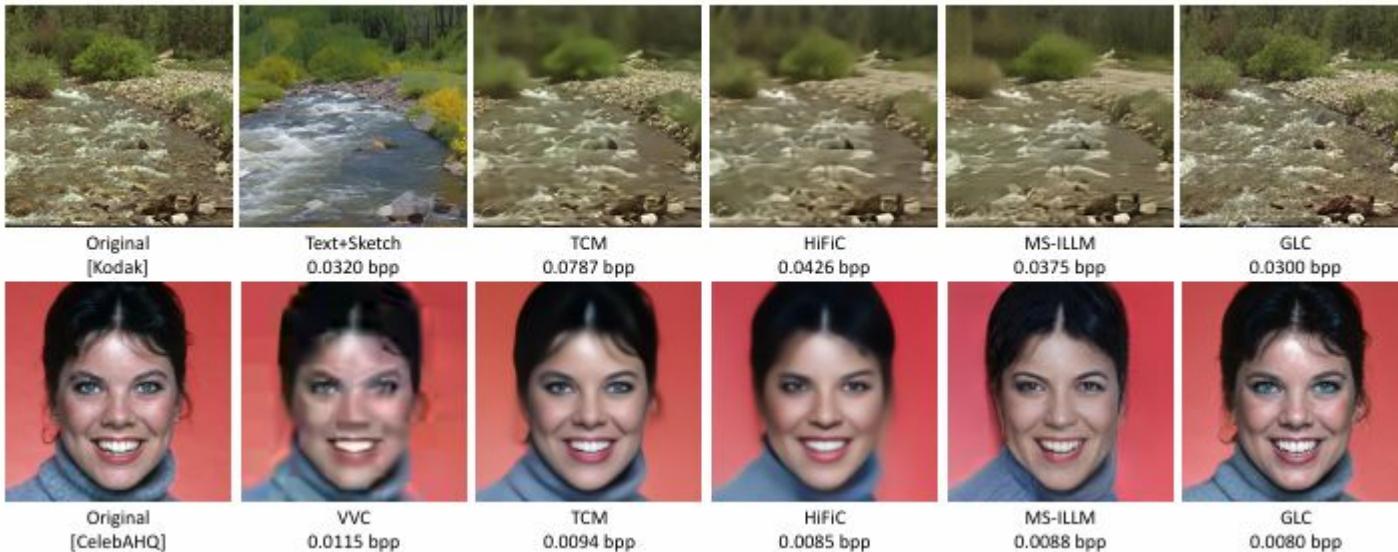
- Optimizer: AdamW, batch size 8, multiple λ (rate–distortion trade-off) per batch.
- Stage I: L1 + LPIPS + GAN + codebook loss to train the generative auto-encoder.
- Stage II/III: rate–distortion loss + code-prediction-based supervision.
- Stage III additionally uses pixel-level GAN and perceptual loss for joint fine-tuning.
- Stage I loss coefficients: the adversarial loss weight is set to $\lambda_{\text{adv}} = 0.8$ for the PatchGAN discriminator.
- Code-prediction loss coefficients: the latent code prediction loss uses $\alpha = 0.5$, and the Stage III pixel-level loss includes an additional term with $\lambda_{\text{code}} = 0.05$.

Experiment Results



[9] Jia, Z., Li, J., Li, B., Li, H., & Lu, Y. (2024). Generative latent coding for ultra-low bitrate image compression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26088–26098).

Qualitative Comparison



[9] Jia, Z., Li, J., Li, B., Li, H., & Lu, Y. (2024). Generative latent coding for ultra-low bitrate image compression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26088–26098).

Ablation Study

- Latent coding schemes

- Direct indices-map coding (VQ indices only): **+66% BD-Rate**
- Transform coding + factorized hyperprior: **+18% BD-Rate**
- Transform coding + categorical hyper module: **best performance**

- Effect of categorical hyper module

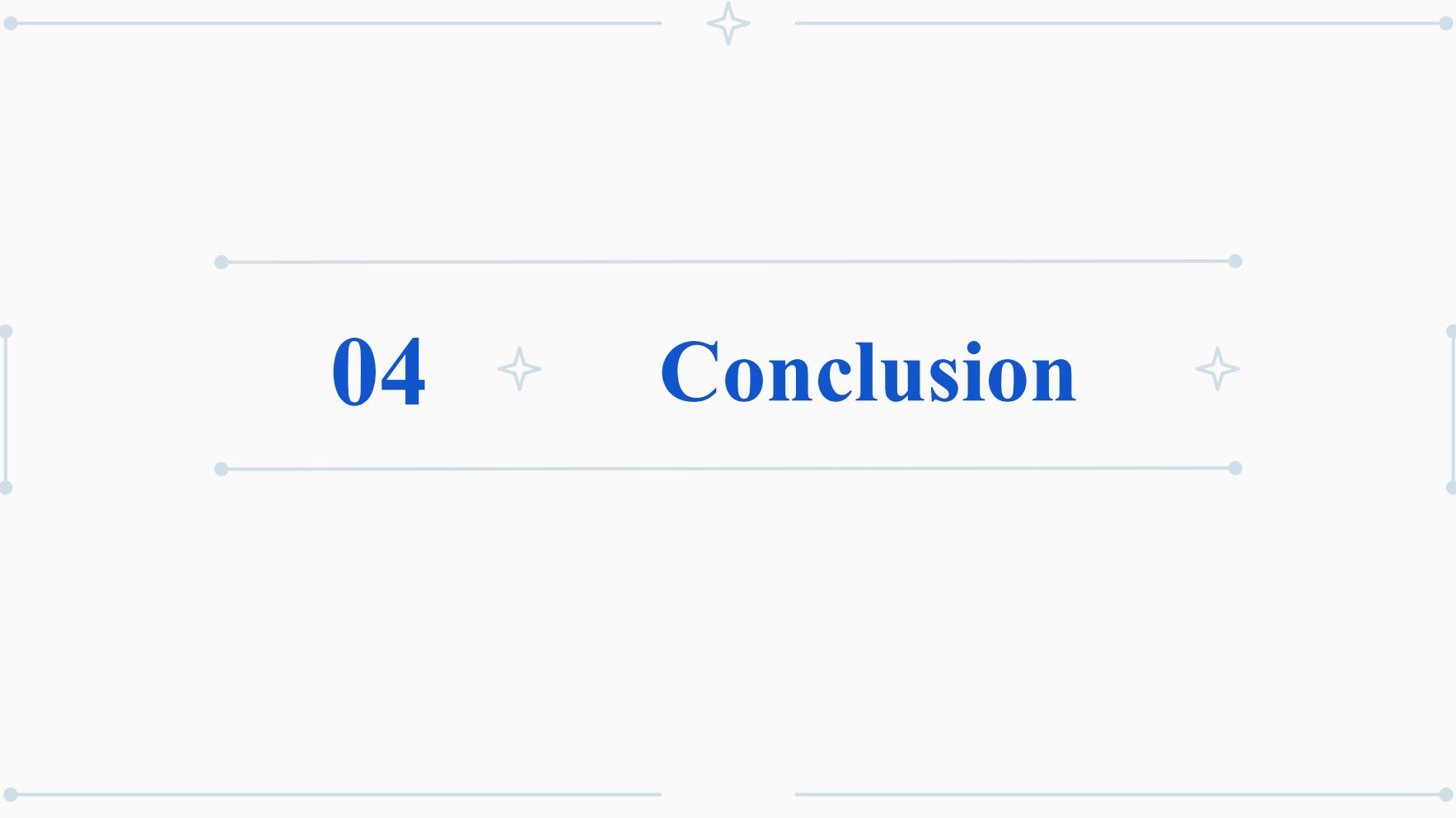
- Hyper codebook learns semantic hyper-latent representation
- Reduces bits spent on low-level details at ultra-low bitrates

- Code-prediction-based supervision

- Without code prediction: **+13% BD-Rate**
- Code predictor in inference path: **+61% BD-Rate**
- Code prediction used only as an auxiliary training loss: **best performance**

Latent coding scheme	Probability model of z	BD-Rate ↓
Indices-map coding	-	66.2%
Transform coding	Factorized prior Categorical prior	17.7% 0%

code prediction usage	BD-Rate ↓
w/o code pred.	13.1%
code pred. in network	60.7%
code pred. as supervision	0%



04 Conclusion

Applications

- **Unified generative latent space**
 - One GLC latent space can support multiple applications by changing encoders/decoders
- **Image restoration with GLC**
 - Train a restoration encoder that maps degraded images to the clean latent space and use the same GLC decoder to reconstruct restored images.
 - Compared to Restormer + compression codec, GLC-based pipeline achieves lower bpp and better FID/DISTS with fewer parameters
- **Stylization**
 - Replace the standard decoder with a stylization decoder
 - Same latent code can be decoded as a stylized image while preserving content structure

Conclusion

- Propose Generative Latent Coding for ultra-low bitrate image compression.
- Perform transform and entropy coding in a generative latent space instead of pixel space.
- Design a categorical hyper module to model semantic hyper-latent efficiently.
- Introduce code-prediction-based supervision that enriches semantics without runtime overhead.
- Achieve **state-of-the-art** perception–bitrate trade-off on CLIC 2020, CelebA-HQ, and other benchmarks.

Limitations

- Trained on specific datasets; generalization to other domains is limited
- Struggles on non-natural images such as screen content, text, or UI graphics
- Pixel-wise metrics (**PSNR, MS-SSIM**) remain relatively low at ultra-low bitrates

Reference

- [1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
- [2] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 128(7):1956–1981, 2020.
- [3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019.
- [4] George Toderici, Lucas Theis, Nick Johnston, Eirikur Agustsson, Fabian Mentzer, Johannes Ball'e, Wenzhe Shi, and Radu Timofte. Clic 2020: Challenge on learned image compression, 2020, 2020.
- [5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In International Conference on Learning Representations, 2018.
- [6] Kodak Lossless True Color Image Suite. <http://r0k.us/graphics/kodak/>.
- [7] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 126–135, 2017.
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.
- [9] Jia, Z., Li, J., Li, B., Li, H., & Lu, Y. (2024). Generative latent coding for ultra-low bitrate image compression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 26088–26098).