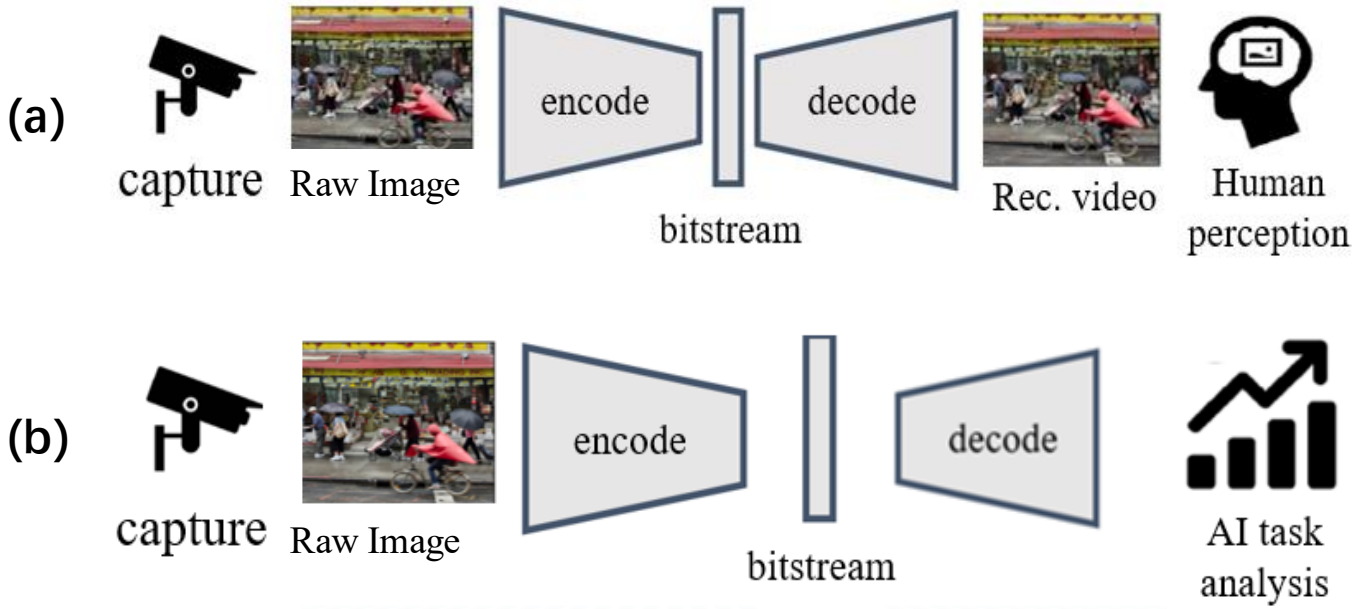
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Main Contribution

- In today's world, image data consumers include not only humans but also machine vision analysis, posing a crucial question of efficiently encoding data to serve both.
- We propose Diff-ICMH, a generative image compression framework aiming for ***harmonizing machine and human vision in image compression***.



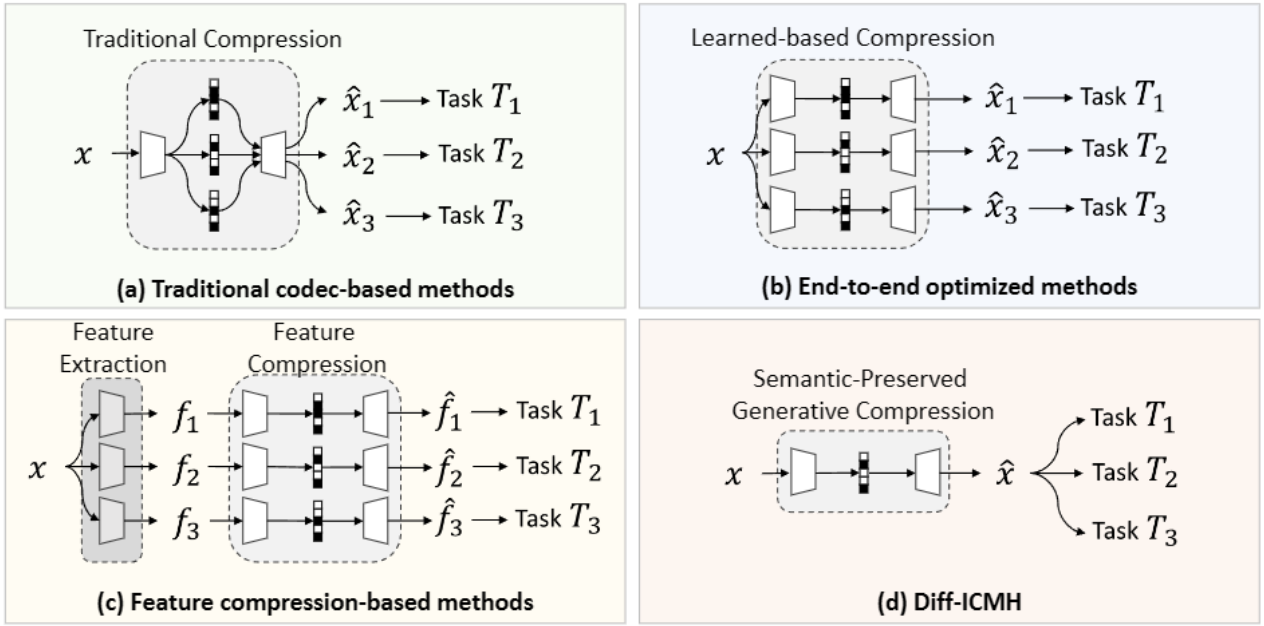
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Pipeline Comparison

- Previous methods need to compress independent bitstream for each task, lacking of generalization.
- In addition, compression methods specifically designed for machine vision are often not friendly to human visual perception.
- Diff-ICMH supports various downstream tasks and human visual perception simultaneously through only one codec and the corresponding code stream.



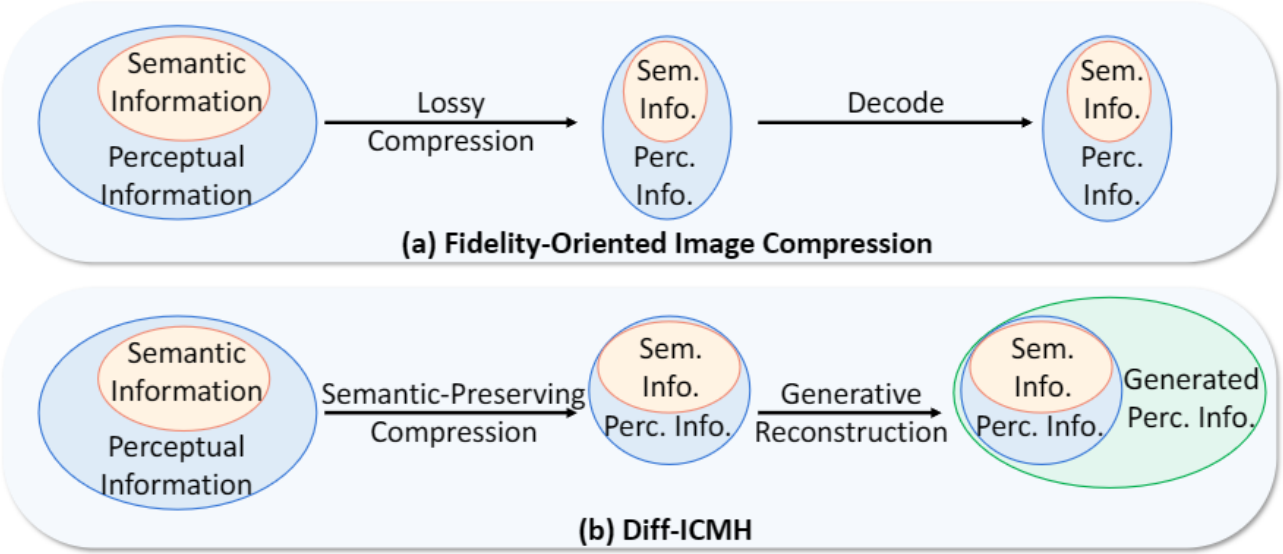
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Motivation

- An image's information consists of both ***semantics and texture***. Traditional compression methods, designed directly for signal fidelity without distinction, often cause substantial semantic loss and thus degraded performance in intelligent tasks.
- We aim **to *preserve semantic information during encoding process*** and ***generate texture information during decoding process.***



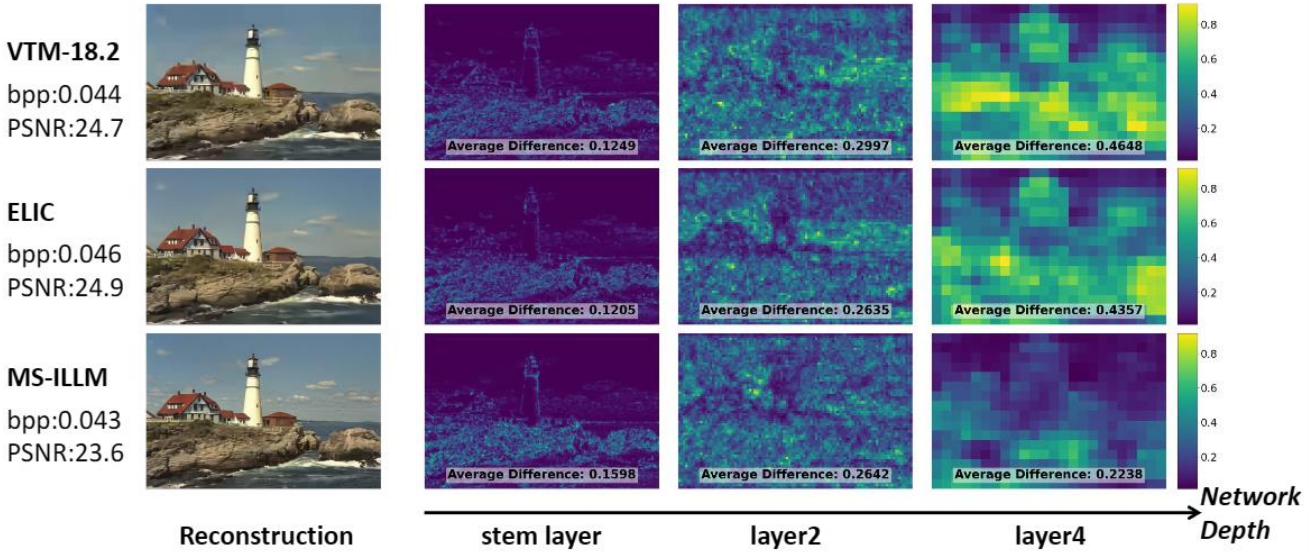
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Evidence of Motivation

- Fidelity-oriented compression methods suffer from substantial semantic mismatch in neural network's deeper layers.
- Perception-oriented generative compression methods can better ensure the consistency of semantic features.



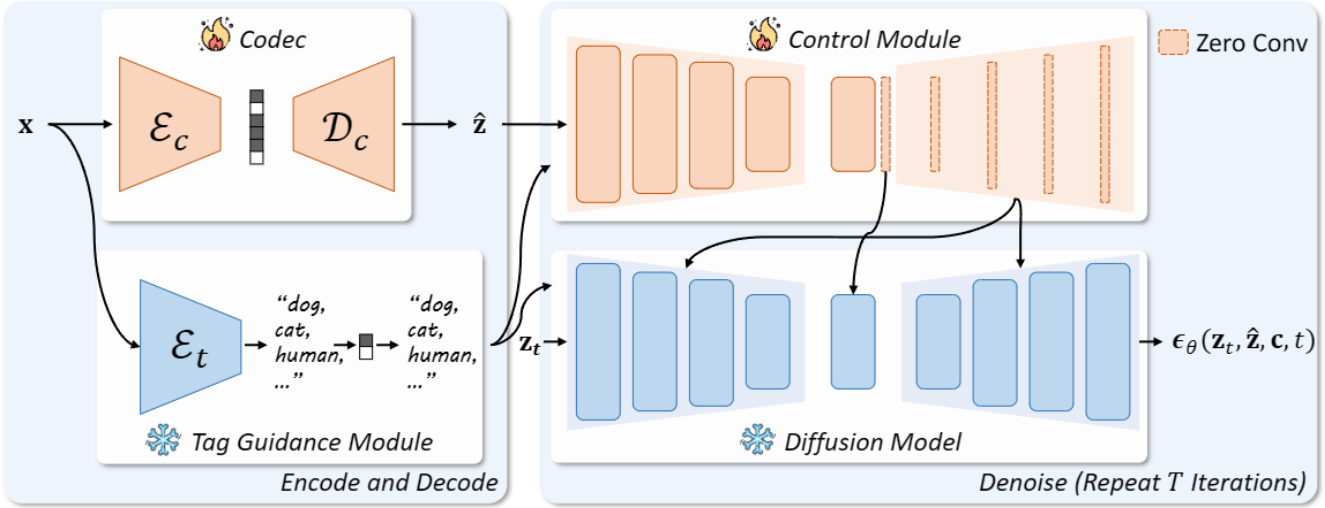
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Method

- Our framework consists of a latent-space codec, a tag guidance module, and a diffusion decoder.
- Input image is compressed and decoded into latent feature $\hat{\mathbf{z}}$.
- Tags of the image are extracted and lossless compressed.
- Both decoded latent and tags are input into the diffusion decoder as conditions for generation.



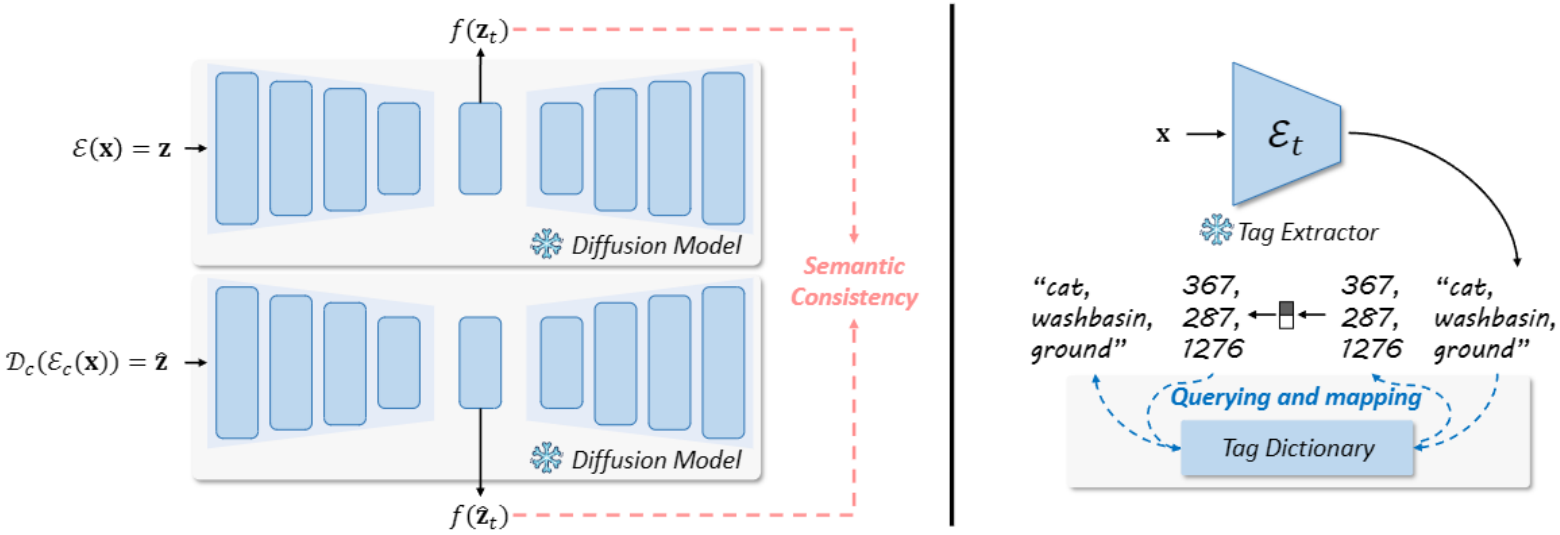
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Method

- We propose ***semantic consistency loss*** to better preserve semantics during compression process.
- We propose ***tag guidance module*** to better activate the generative prior of the pre-trained generative model.



(a) Semantic Consistency Loss (b) Tag Guidance Module

Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Method

➢ Training loss consists of ***rate, distance in latent space, diffusion loss, and semantic consistency loss***.

$$\mathcal{L}_{\text{rate}} = \mathcal{R}(\hat{\mathbf{y}}) + \mathcal{R}(\hat{\mathbf{z}}_h).$$

$$\mathcal{L}_{\text{dist}} = \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 = \|\mathcal{E}_{\text{VAE}}(\mathbf{x}) - \mathcal{D}_c(\hat{\mathbf{y}})\|_2^2,$$

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z},t,\mathbf{c},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \hat{\mathbf{z}}, \mathbf{c}, t)\|_2^2\right],$$

$$\mathcal{L}_{\text{sem}} = -\mathbb{E}_{\mathbf{z},\hat{\mathbf{z}}}\left[\frac{1}{N}\sum_{n=1}^{N}\text{sim}(f(\mathbf{z})_n, f(\hat{\mathbf{z}})_n)\right],$$

$$\mathcal{L}_{\text{final}} = \lambda_{\text{rate}}\mathcal{L}_{\text{rate}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}},$$

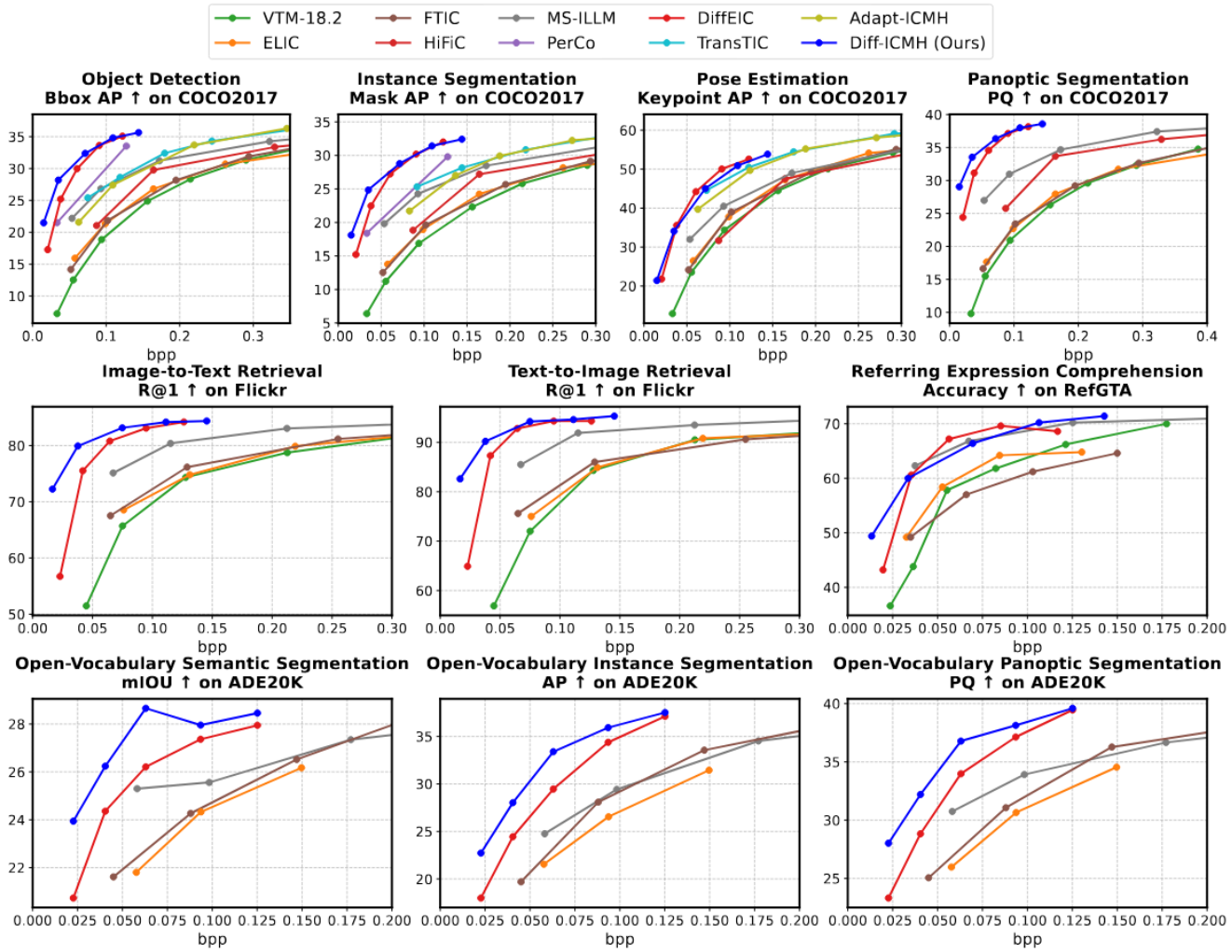
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Experiments

➢ Diff-ICMH achieves SOTA performance in ***machine-vision oriented compression***.



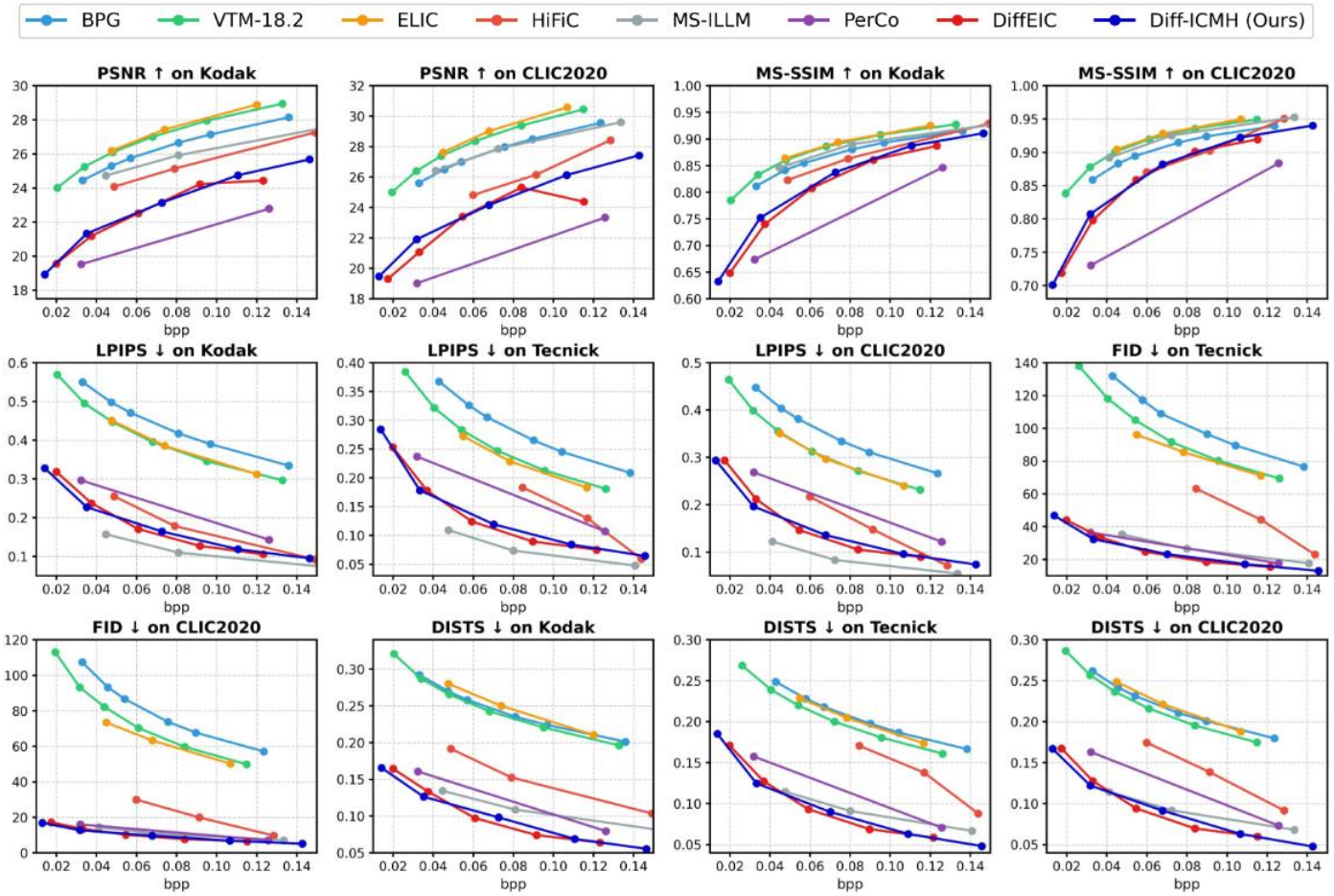
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Experiments

➢ Diff-ICMH achieves results comparable to SOTA in ***perception-oriented compression***.



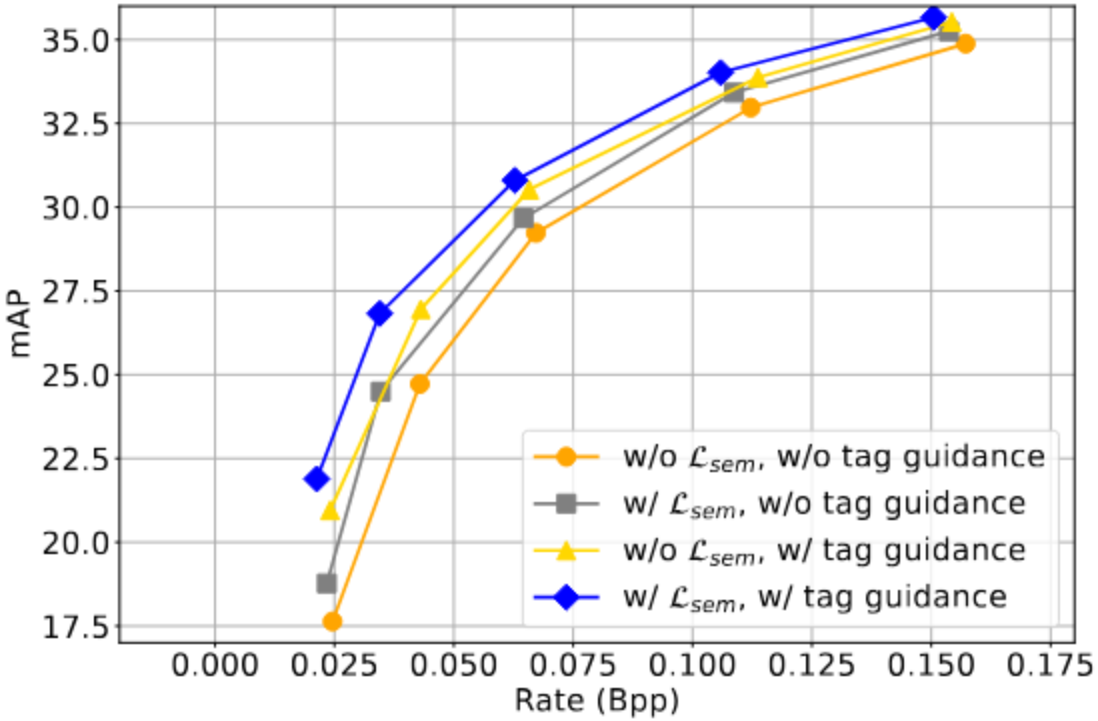
Diff-ICMH: Harmonizing Machine and Human Vision in Image Compression with Generative Prior

Ruoyu Feng* Yunpeng Qi* Jinming Liu Yixin Gao
Xin Li# Xin Jin Zhibo Chen#

EASTERN INSTITUTE OF TECHNOLOGY, NINGBO

❖ Experiments

➢ Ablation study demonstrate the effectiveness of our proposed ***semantic consistency loss and tag guidance module***.



Thank you

E-mail: ustcfry@mail.ustc.edu.cn