One-Step Diffusion-Based Image Compression with Semantic Distillation

Naifu Xue[1], Zhaoyang Jia[2], Jiahao Li[3], Bin Li[3], Yuan Zhang[1], Yan Lu[3]

[1]Communication University of China, [2]University of Science and Technology of China, [3]Microsoft Research Asia

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Motivation:

Weakness of existing image codecs at low bitrates.

- Traditional / MSE-optimized codecs produce blurry textures.
- VAE-based generative codecs improve sharpness yet often introduce artifacts.
- Diffusion-based generative codecs raise perceptual realism, but they:
  - (i) occasionally generate content that deviates from the original;
  - (ii) Iterative sampling process leads to substantial computational overhead.

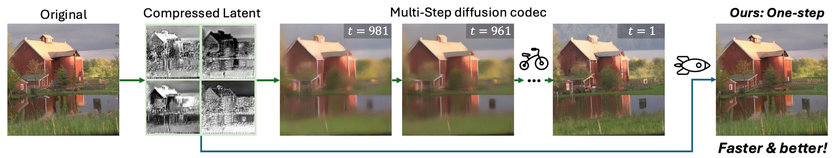


Figure 1: Multi-step diffusion v.s. one-step diffusion

Observation:

Multi-step sampling may not necessary for compression.

- In diffusion models, early steps generate coarse structure, while later steps progressively refine high-frequency details.
- An image codec already transmits most low-frequency content through its latent representation, the decoder only need to synthesize fine details.
- Besides, multi-step diffusion prohibits direct pixel-domain supervision.

Textual conditioning in diffusion models is suboptimal for compression.

- Caption generation depends on heavy vision–language models.
- Text prompts are coarse and non-spatial, unable to describe fine local details.

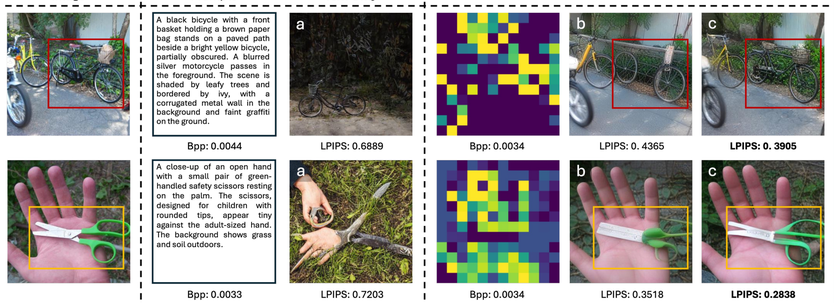


Figure 2: Reconstruction examples of different semantic guidance

Idea:

A carefully designed one-step diffusion is sufficient for image coding.

In this work, we:

- Use **one-step diffusion** as the generator in the image codec.
- Replace textual guidance with **hyperprior-based conditioning**.
- Enhance the guidance accuracy of the hyperprior through **semantic distillation**.
- Adopt **pixel–latent hybrid supervision** to improve reconstruction quality.

Finally achieving SOTA RD performance with fast decoding!

Method

Our Solution: OneDC (One-step Diffusion based Image Compression)

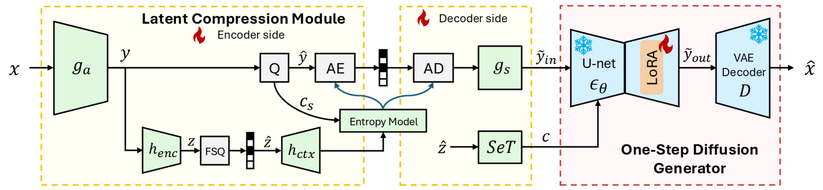


Figure 3: Framework of OneDC

Key components:

- **Latent compression module.** A learned compression framework, encoding spatial features and hyperprior side information for compact transmission.
- **One-step diffusion generator.** A diffusion U-Net that reconstructs images from compressed latents in a single denoising step.
- **Hyperprior-based conditioning.** The hyperprior feature replaces text embeddings to provide spatial semantic guidance via cross-attention layers in the U-Net.
  - As illustrated in Fig.2, the hyperprior provide more accurate description than text.

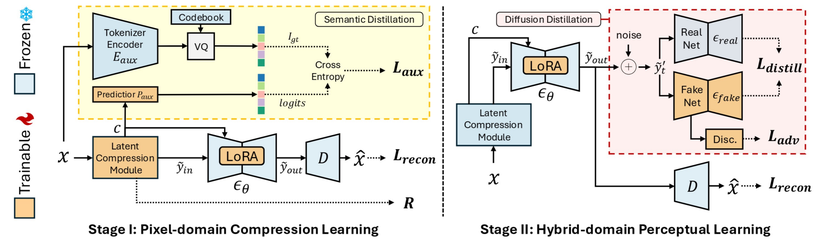


Figure 4: Training pipeline of OneDC

Training strategies:

Two-stage optimization.

- **Stage I:** Train the compression module and adapt the U-Net for the one-step image reconstruction task.
- **Stage II:** Fine-tune the one-step diffusion U-Net to enhance perceptual realism.

Semantic distillation of hyperprior (Stage I).

- Distill semantic priors from a pretrained VQ-tokenizer into the hyperprior codec, enriching its representation and guidance accuracy.
  - As shown in Fig.2, hyperprior + semantic distillation further improves fidelity.

$$L_{stageI} = L_{recon} + \lambda R + \alpha L_{aux}, \quad \text{where} \quad L_{recon} = L_1(x, \hat{x}) + L_{perceptual}(x, \hat{x})$$
$$I_{gt} = VQ(E_{aux}(x)), \quad L_{aux} = CE(I_{gt}, P_{aux}(c))$$

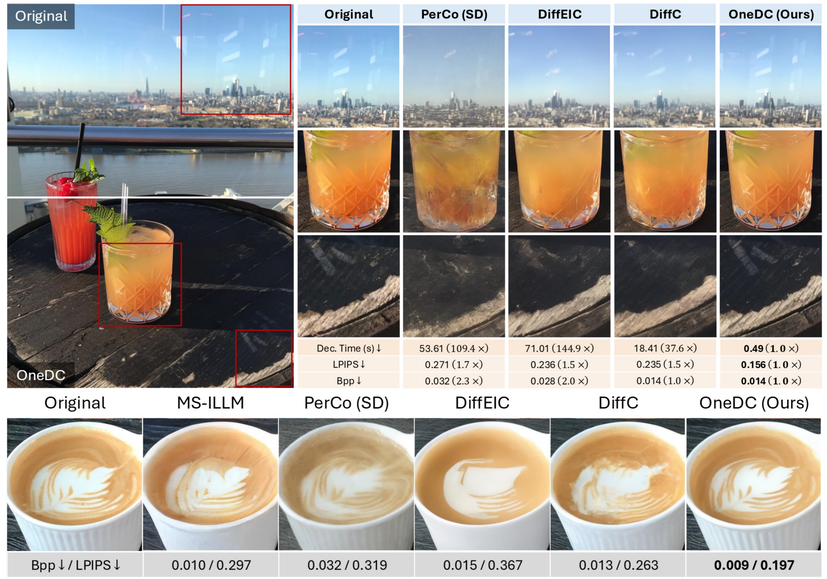
Hybrid-domain supervision of one-step diffusion (Stage II).

- Perform diffusion distillation in the latent domain, transferring generative priors from a multi-step teacher to the one-step student.
- Combine with pixel-domain supervision to maintain reconstruction fidelity.

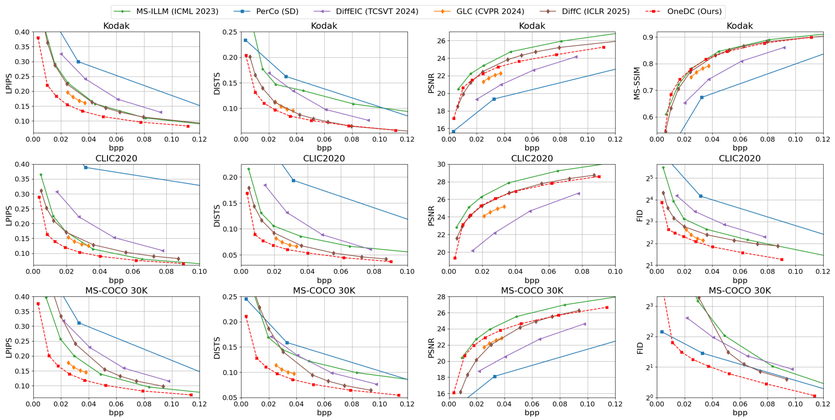
$$L_{stageII} = L_{distill} + \beta L_{recon} + \gamma L_{adv}, \quad \text{where:}$$
$$L_{distill} = \mathbb{E}_{t,\tilde{y}'_t}[\epsilon_{fake}(\tilde{y}'_t, t) - \epsilon_{real}(\tilde{y}'_t, t)], \quad L_{adv} = \mathbb{E}_{t,\tilde{y}'_t}[-Disc(\epsilon_{fake}(\tilde{y}'_t, t), t)]$$

Experiment

Qualitative Evaluation: OneDC delivers best visual with lowest bpp.



Quantitative Evaluation: OneDC achieves SOTA RD performance.



Ablation studies validate the effectiveness of the hyperprior guidance and Hybrid-domain training.

Table 1: Ablation studies with BD-Rate (%) ↓.

Settings	CLIC2020 DISTS	CLIC2020 FID
Semantic guidance		
No guidance	44.0	45.1
Text guidance	24.2	36.3
Hyperprior guidance	20.7	24.3
Hyperprior + Sem. Distil.→ **Ours**	0.00	0.00
Loss variation		
Pixel-domain only	11.4	51.8
Latent-domain only	60.7	37.1
Hybrid-domain → **Ours**	0.00	0.00

Complexity analysis validate OneDC is much faster than all multi-step diffusion codecs while preserving high quality.

Table 2: Comparison of coding time and BD-Rate (%) ↓.

Methods	Times (s) Enc.	Times (s) Dec.	MS-COCO 30K LPIPS	MS-COCO 30K DISTS	MS-COCO 30K FID
VAE-based					
MS-ILLM	0.14	0.17	138.3	253.0	478.4
Multi-step diffusion					
DiffEIC	0.32	12.4	305.0	239.1	341.0
PerCo (SD)	0.58	8.80	538.8	345.8	59.6
DiffC	3.9∼15.6	6.9∼10.8	234.0	196.1	690.9
One-step diffusion					
OneDC → **Ours**	0.15	0.34	0.00	0.00	0.00