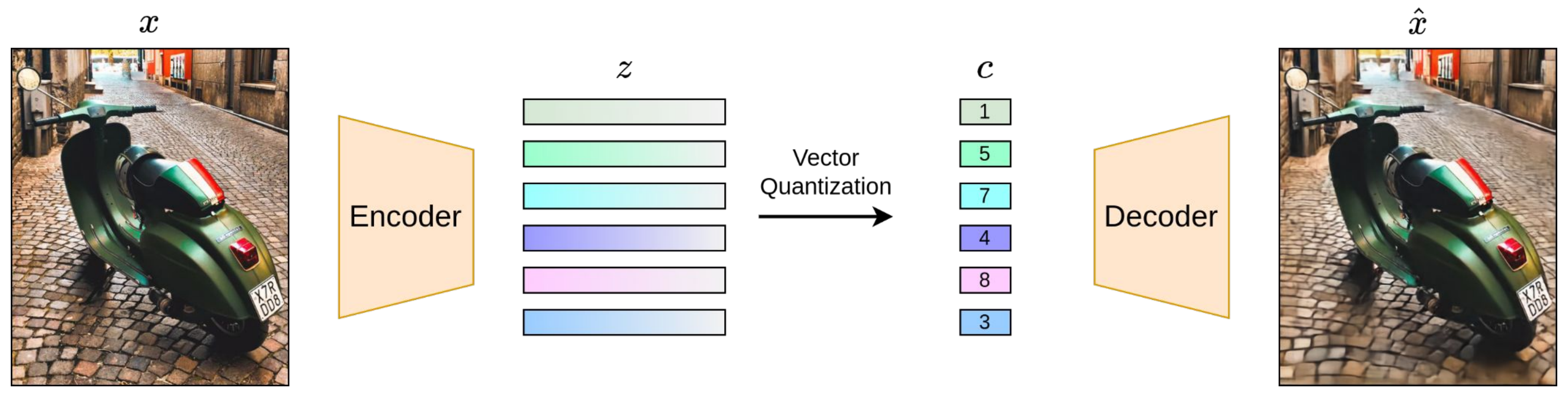
Variable Rate Allocation for Vector-Quantized Autoencoders

Federico Baldassarre[K], Alaaeldin El-Nouby[M], and Hervé Jégou[M]
[K] KTH - Royal Institute of Technology, Stockholm, Sweden
[M] FAIR lab at Meta AI, Paris, France

Vector-Quantized Autoencoders for Image Compression

VQ-VAE: encoder-decoder transformer model with a quantized latent space. Each *PxP* patch is projected to a latent vector $z \in \mathbb{R}^d$ that gets quantized to the closest entry in a learned codebook.



Compression: store the discrete quantization indices of each patch. For a vocabulary of size *V*, the encoding cost is $\log_2 V$ bits per patch. Further rate reduction using an entropy coding is possible but not considered in this work.

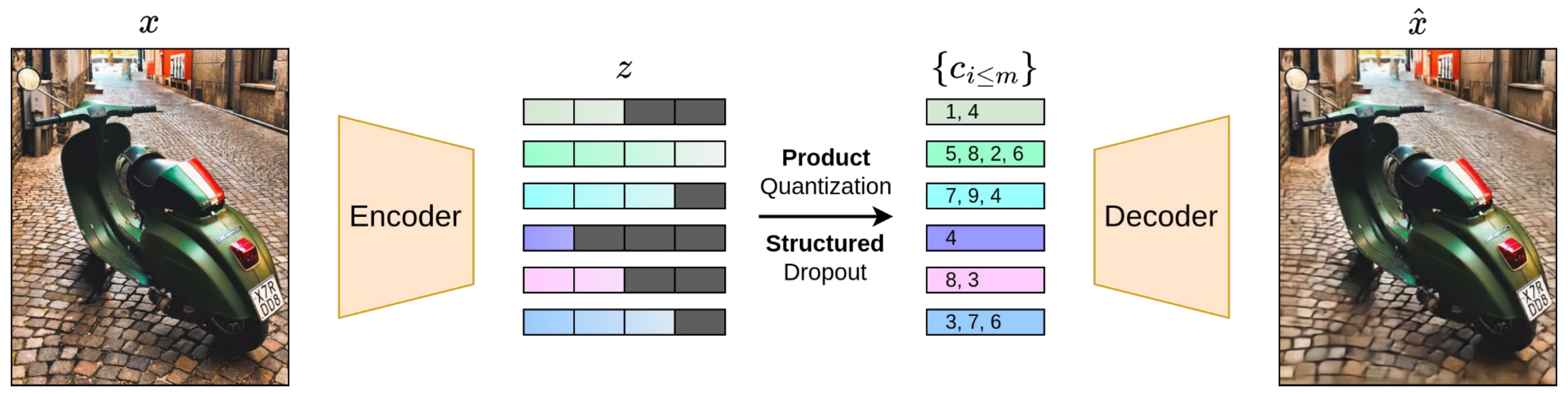
Limitations:

- Learning large codebooks to model the diversity of natural images
- Local rate control: each patch can be either encoded ($\log_2 V$ bits) or skipped (0 bits) with an all-or-nothing operation.

Product Quantization and Structured Dropout

Product quantization: for each *PxP* patch, split the latent vector into *d/S*-dimensional chunks and quantize each subvector independently using *S* codebooks of size *V*. Therefore, each patch is encoded as a list $\{c_i\}^S$ of discrete codes which can be stored using $S \log_2 V$ bits.

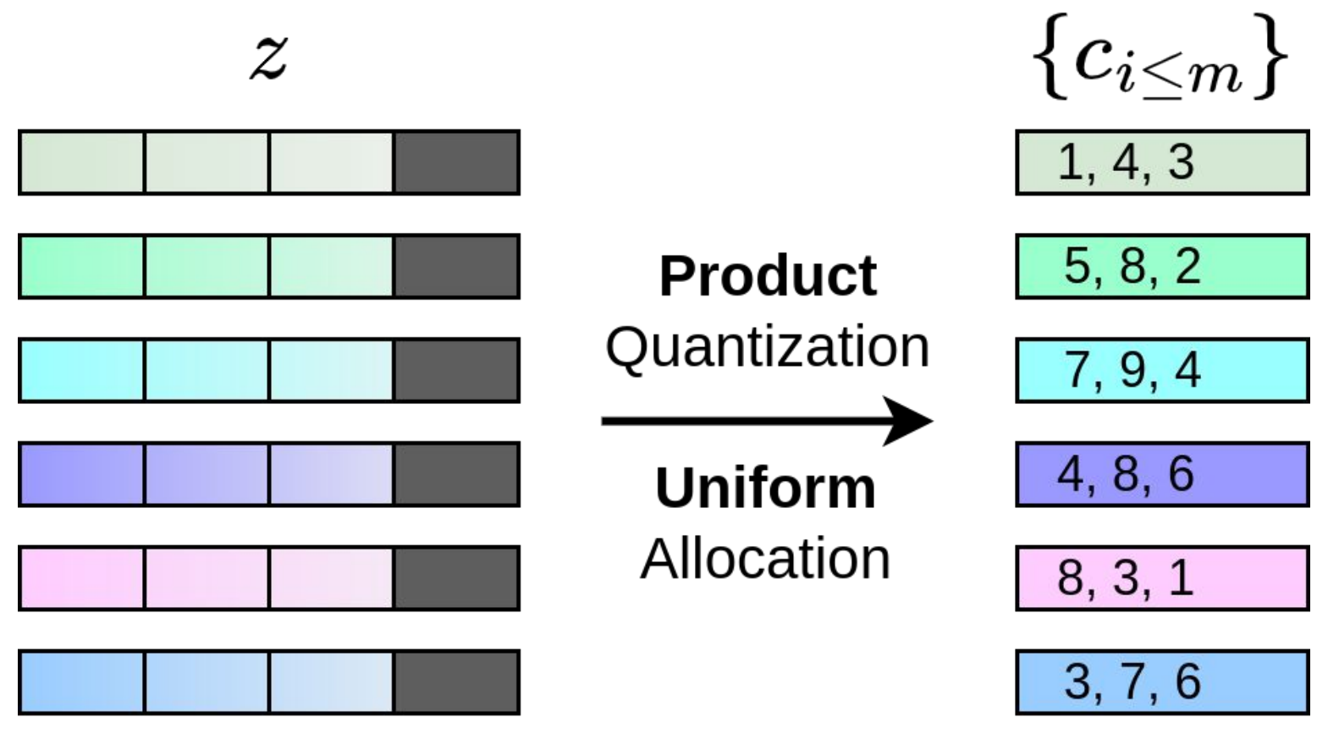
Structured dropout: during training, randomly keep the first *m* codes of each patch from the head of the list and drop the rest. Also, modulate the MSE reconstruction loss for each patch accordingly. The number of codes to keep is sampled at random to ensure uniform training at all compression rates.



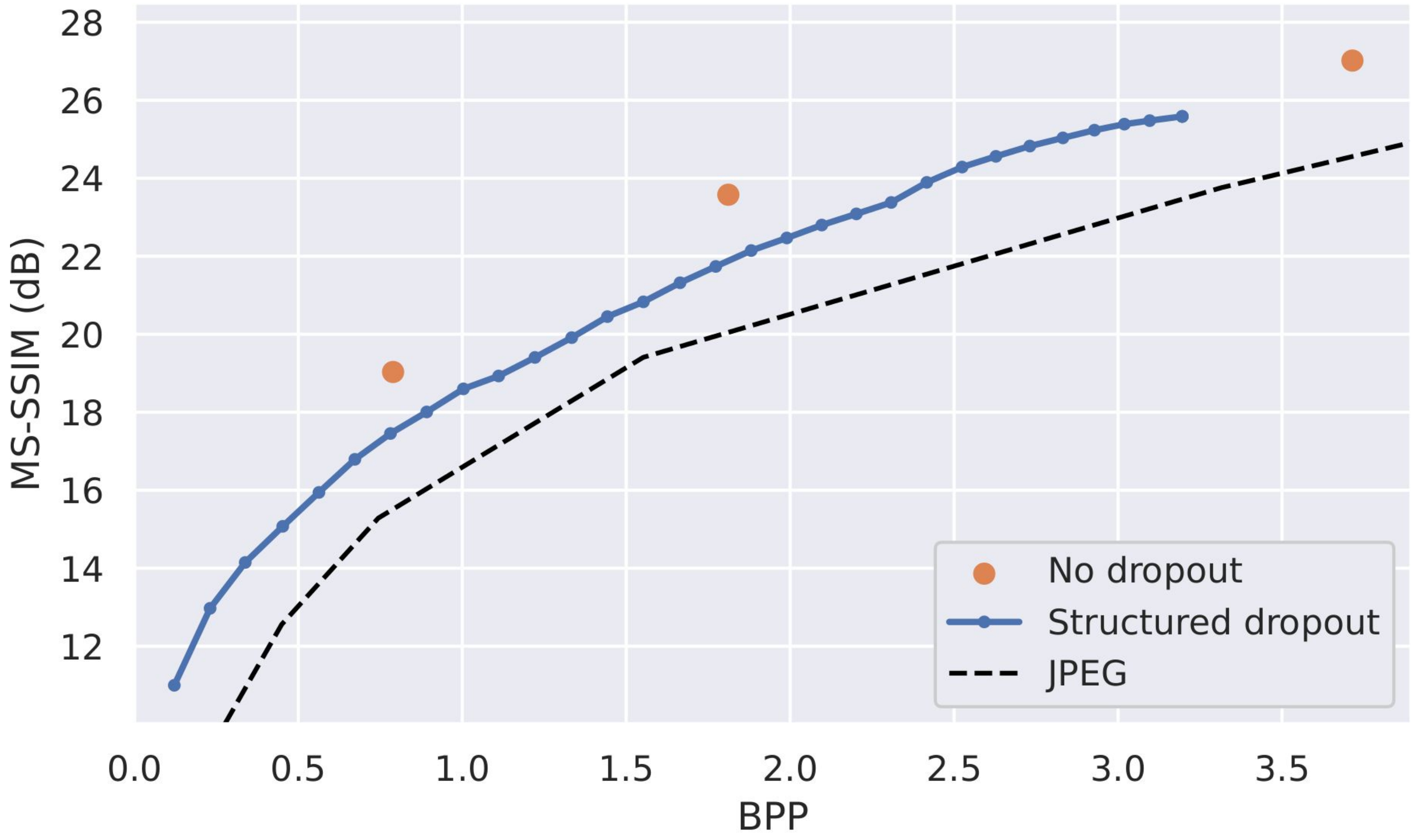
Compression cost: $m \log_2 V + \log_2 S$ bits per patch

Variable Rate Allocation

Vanilla PQ-VAE models ● ● ●
Given a number of codebooks *S* of size *V*, a vanilla PQ-VAE can only compress images at the fixed cost of $S \log_2 V$ bits per patch. To target a different bit rate, multiple models need to be trained and deployed.

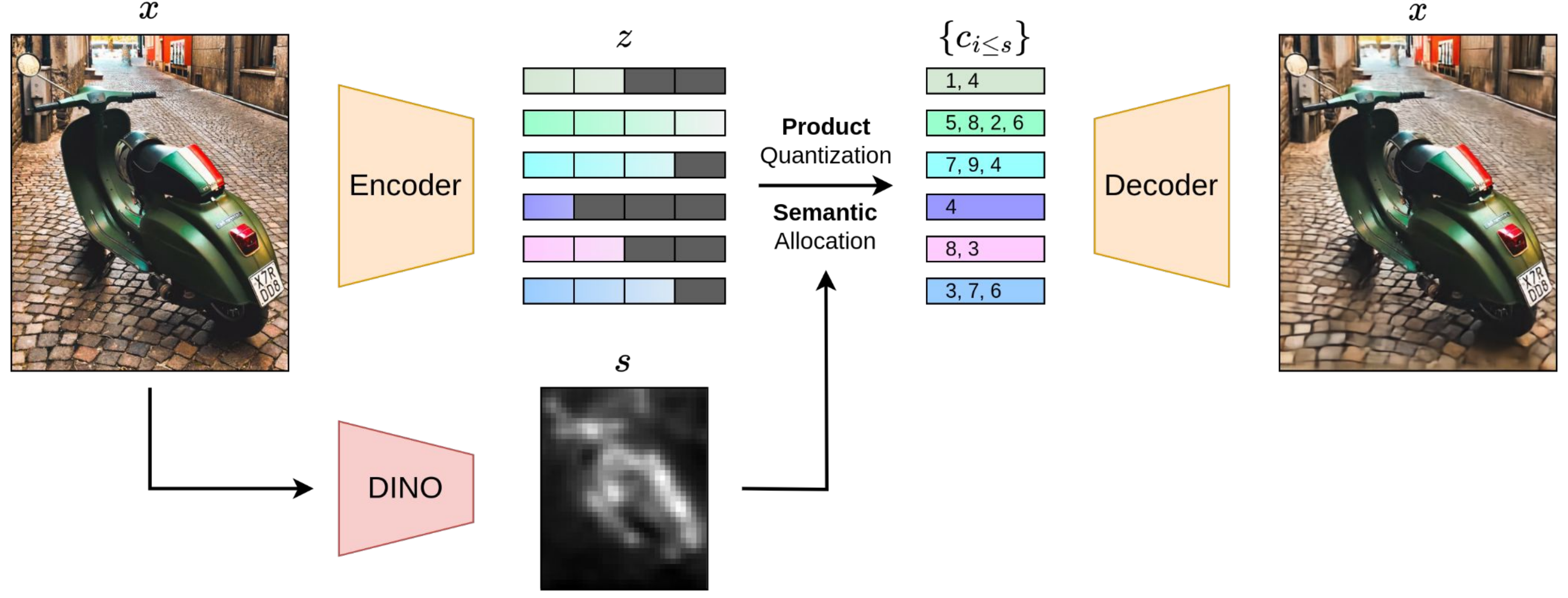


Uniform rate allocation ●—
A single PQ-VAE model trained with structured dropout can compress images at multiple operating points of the rate-distortion curve. A simple strategy is to select $m \in \{1 \dots S\}$ and keep only the first *m* codes for all image patches, effectively allocating the rate budget in a uniform way.



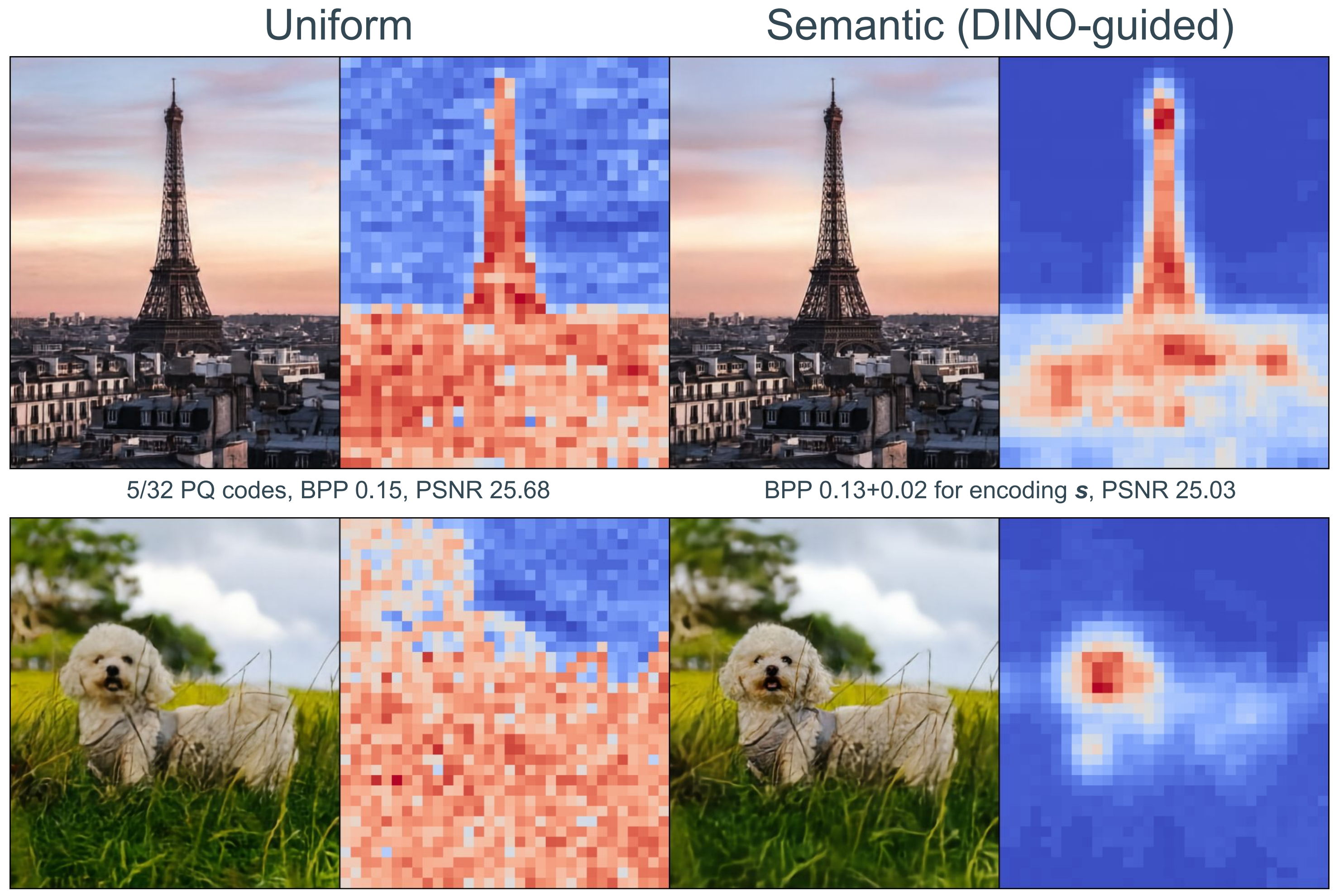
Semantic Rate Allocation

We can also allocate the rate budget according to the semantic content of a patch. We use the attention map of a pre-trained DINO model as a proxy for saliency: important regions are assigned more codes while background details are sacrificed to save bits.



Compression Evaluation

Qualitative comparison of rate allocation strategies



5/32 PQ codes, BPP 0.15, PSNR 25.68

BPP 0.13+0.02 for encoding ***s***, PSNR 25.03

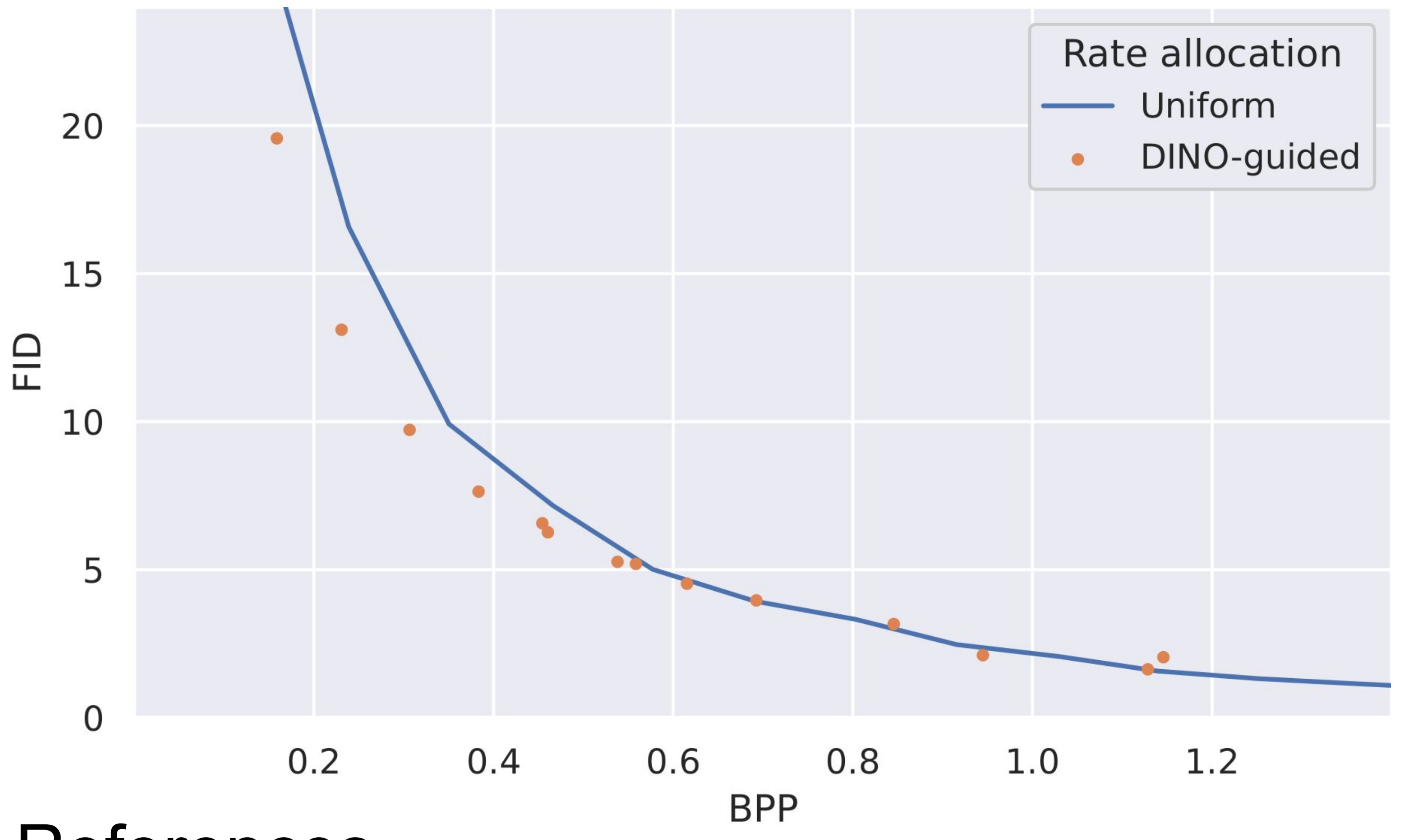
12/32 PQ codes, BPP 0.32, PSNR 27.41

BPP 0.27+0.02 for encoding ***s***, PSNR 28.76

The heatmaps indicate effective bits per patch. Salient regions are encoded with more details, background elements degrade gracefully. Observe: PSNR distortion metric does not reflect perceived quality.

Perceptual quality of semantically-compressed images

DINO-guided rate allocation results better perceptual quality as indicated by a low Fréchet Inception Distance, especially at low BPP.



References

van den Oord et al. *“Neural Discrete Representation Learning”* (NeurIPS 2017)

Jégou et al. *“Product Quantization for Nearest Neighbor Search”* (IEEE PAMI 2011)

El-Nouby et al. *“Image Compression with Product Quantized Masked Image Modeling”* (TMLR 2023)