PerCoV2: Improved Ultra-Low Bit-Rate Perceptual Image Compression with Implicit Hierarchical Masked Image Modeling

Nikolai Körber[1,2] Eduard Kromer[2] Andreas Siebert[2]
Sascha Hauke[2] Daniel Müller-Gritschneder[3] Björn Schuller[1]

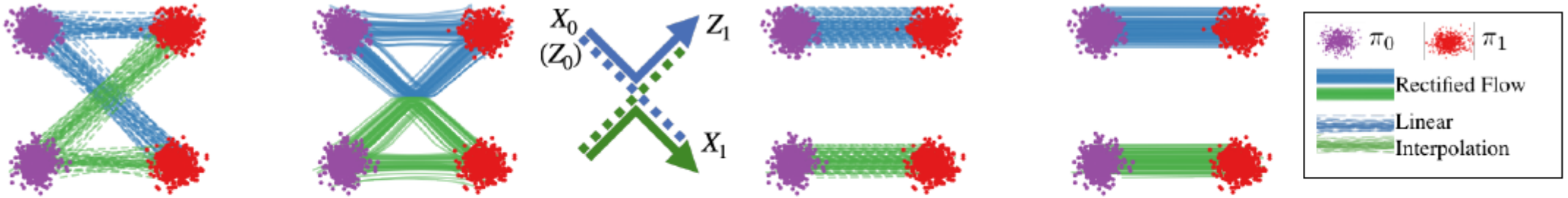
[1]Technical University of Munich
[2]University of Applied Sciences Landshut [3]TU Wien

arXiv:2503.09368

Presenter: Wen Si
2025.11.16

Background: Rectified Flow

- Solve ODE: $\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)), \quad \phi_0(x) = x;$
- Conditional flow matching: $\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)} \left\| v_t(x) - u_t(x|x_1) \right\|^2,$
- Reparameterize: $\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p(x_0)} \left\| v_t(\phi_t(x_0)) - (x_1 - (1 - \sigma_{\min})x_0) \right\|^2.$
- Reflow to make paths straight, distill at last step
- Boost sampling speed



(a) Linear interpolation $X_t = tX_1 + (1-t)X_0$

(b) Rectified flow Z_t induced by (X_0, X_1)

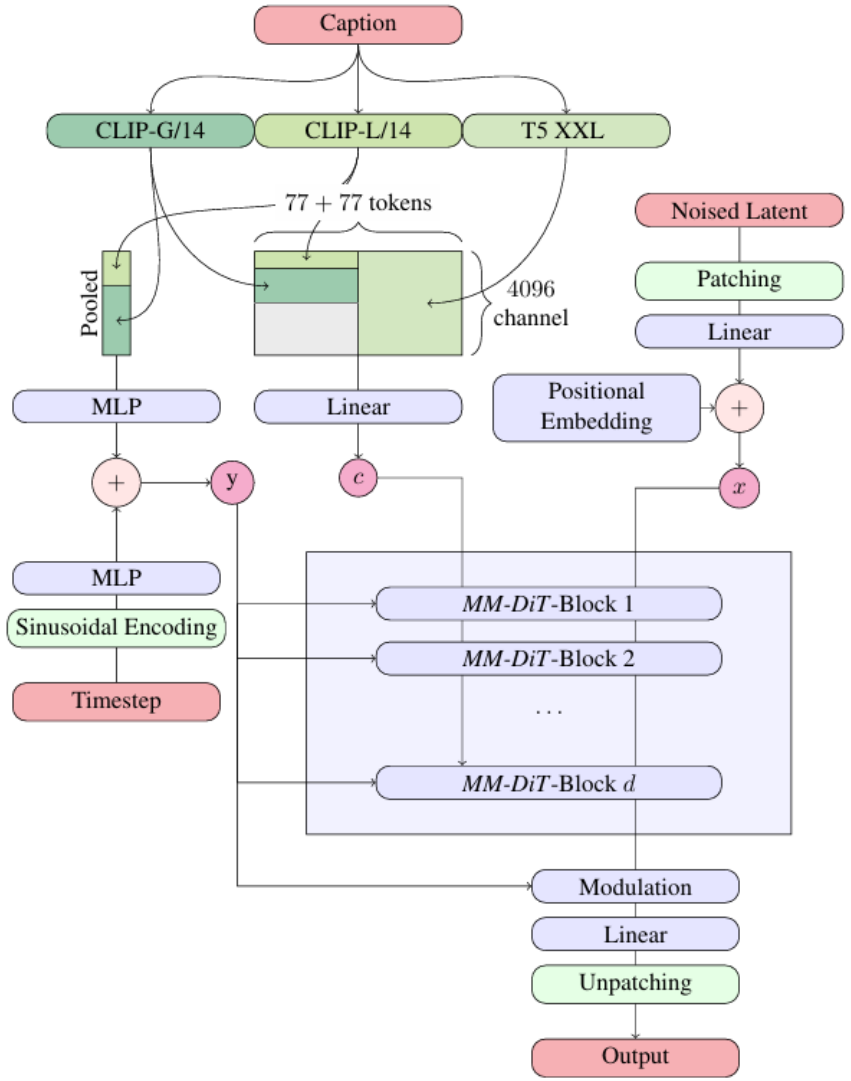
(c) Linear interpolation $Z_t = tZ_1 + (1-t)Z_0$

(d) Rectified flow Z'_t induced by (Z_0, Z_1)

Liu et al. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. (ICLR 2023 oral)

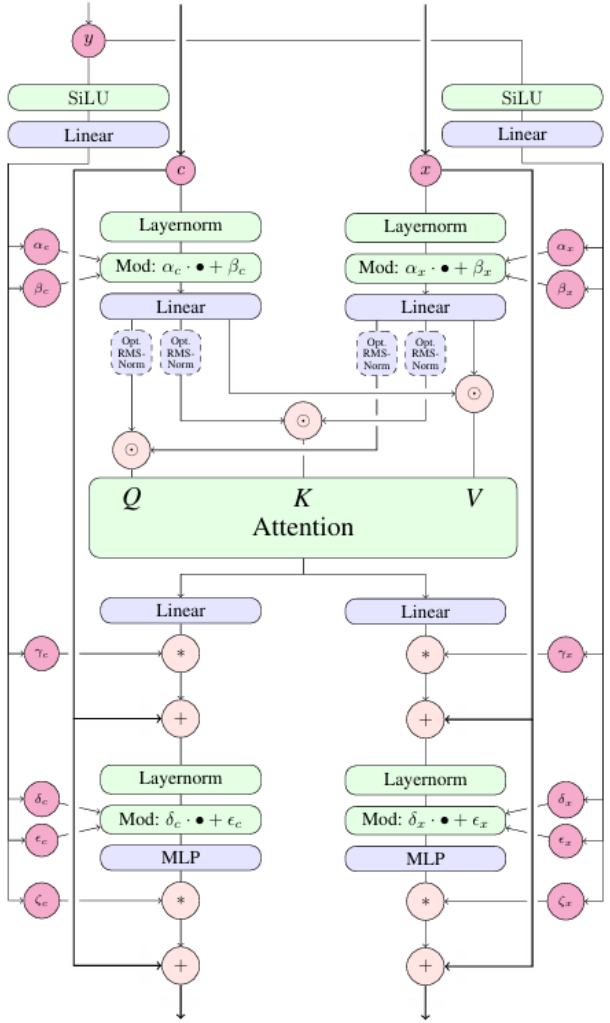
Background: Stable Diffusion 3 (SD3.0)

- Overview: Text-to-image framework
- Text:
  - SDXL dense embedding
  - Conditioning at every timestep
- Image:
  - Latent 2×2 patching
  - Positional embedding
- Use separate weights for 2 modalities



Background: Stable Diffusion 3 (SD3.0): MM-DiT

- Unet → Diffusion Transformer (DiT)
- Enable more condition inputs (timestep, label)
- Adaptive Layernorm: scale + shift, gate
  - Controllable training
- Concat for Q, K, V each
- Better multimodal alignment



(b) One *MM-DiT* block

Esser et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. (arXiv:2403.03206)

Background: Rate-distortion Theory

- Rate: number of bits per data sample
- Distortion: preferably modeled on human perception / aesthetics

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases} \qquad d(x, \hat{x}) = (x - \hat{x})^2$$

- Problem formulation (MSE): $\inf_{Q_{Y|X}(y|x)} I_Q(Y; X) \text{ subject to } D_Q \leq D^*.$

$$D_Q = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{X,Y}(x, y)(x - y)^2 \, dx \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Q_{Y|X}(y \mid x) P_X(x)(x - y)^2 \, dx \, dy.$$

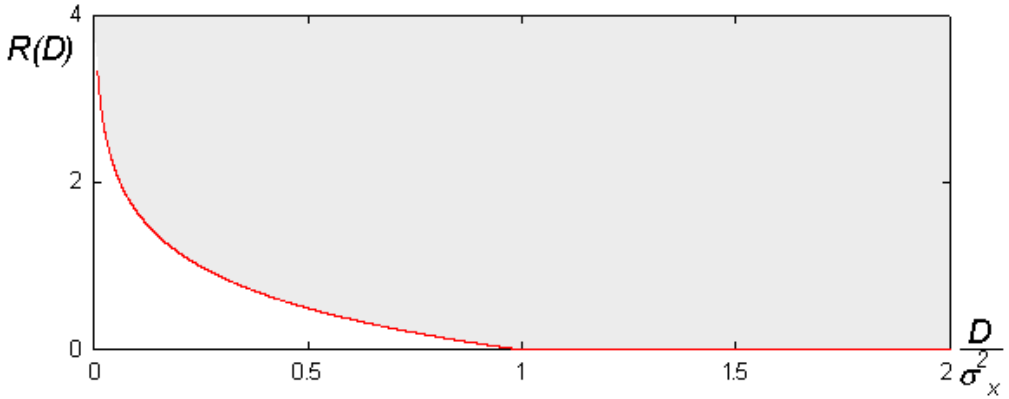
$$I(Y; X) = H(Y) - H(Y \mid X) \ = -\int_{-\infty}^{\infty} P_Y(y) \log_2(P_Y(y)) \, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Q_{Y|X}(y \mid x) P_X(x) \log_2(Q_{Y|X}(y \mid x)) \, dx \, dy.$$

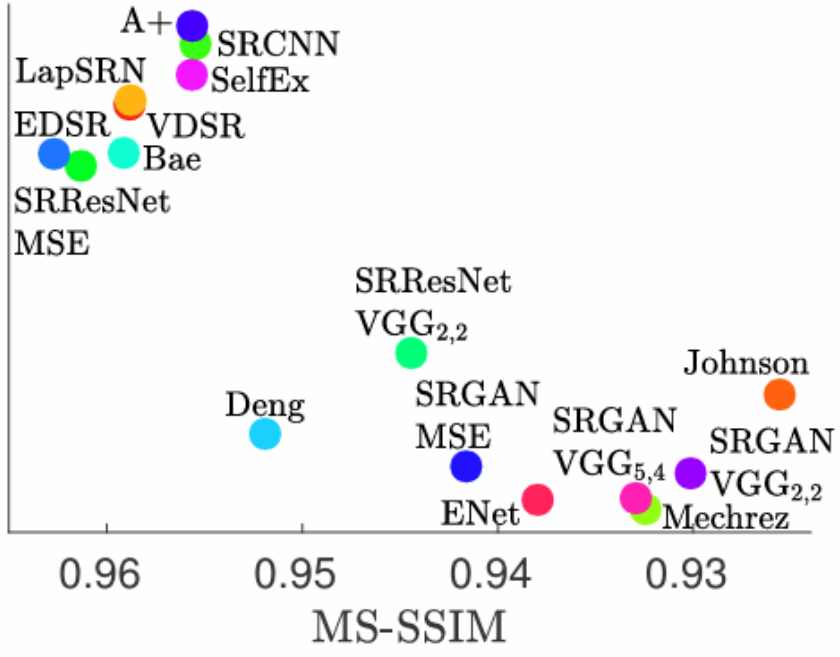
C. E. Shannon. A mathematical theory of communication. (1948)

Background: Rate-distortion-perception Trade-off

- Gaussian random variable: $R(D) \geq h(X) - h(D)$

$$R(D) = \begin{cases} \frac{1}{2}\log_2(\sigma_x^2/D), & \text{if } 0 \leq D \leq \sigma_x^2 \\ 0, & \text{if } D > \sigma_x^2. \end{cases}$$



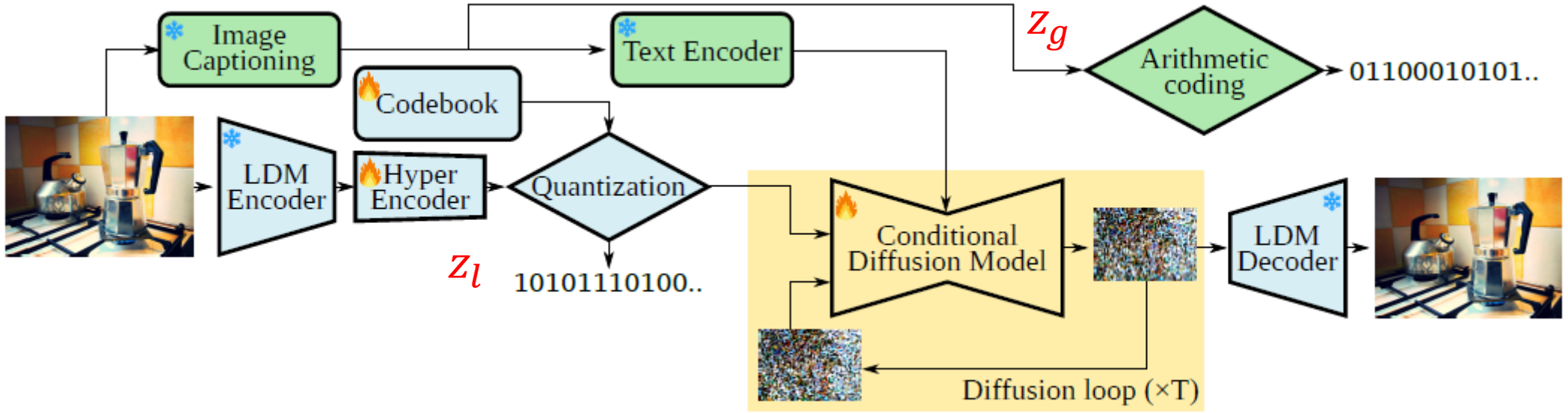


- No compressor/algorithm can go below red line
- RDP Trade-off: for visual quality, distortion must be sacrificed

C. E. Shannon. A mathematical theory of communication. (1948)
Blau et al. The Perception-Distortion Tradeoff. (CVPR 2018)

Background: Perceptual Compression (PerCo)

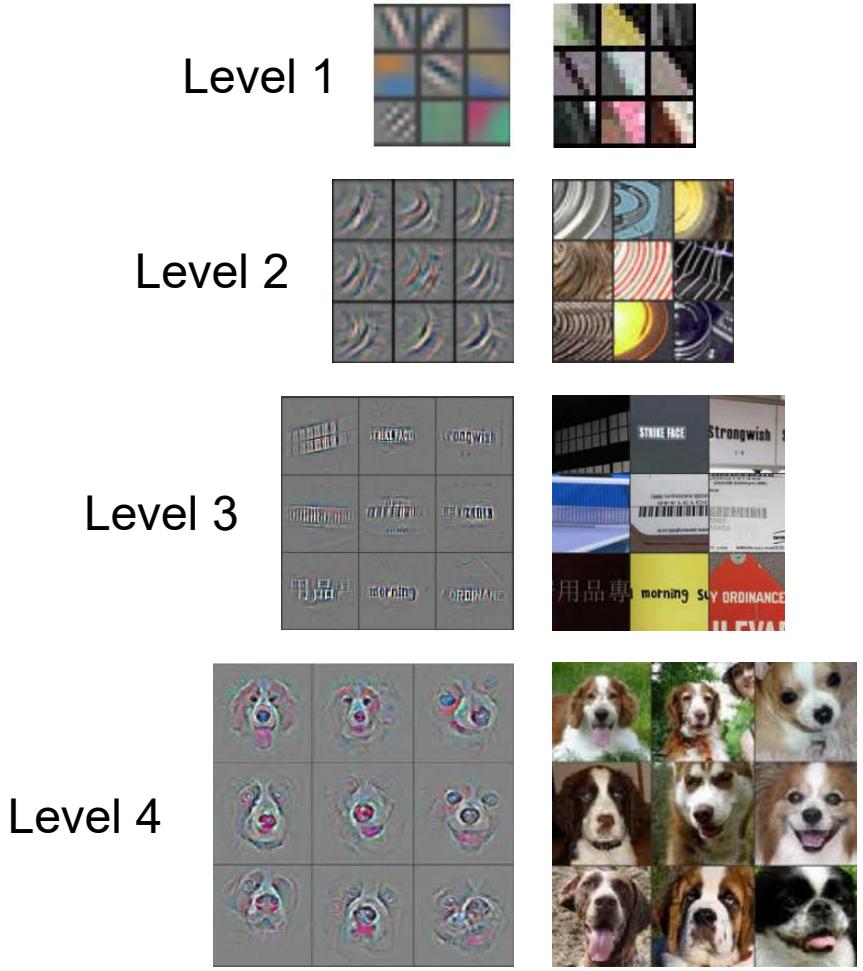
- Training objective: Rate + Distortion $\mathcal{L}_{\mathrm{RD}} = \mathbb{E}_{P_{\boldsymbol{x}}}[\mathbb{E}_{P_{\boldsymbol{z}|\boldsymbol{x}}}\mathcal{L}_{\mathrm{R}}(\boldsymbol{z}) + \lambda\mathcal{L}_{\mathrm{D}}(\hat{\boldsymbol{x}}(\boldsymbol{z}), \boldsymbol{x})],$
- Distortion loss: $\mathcal{L}^t_{\mathrm{Diff}} = \mathbb{E}_{P_{\boldsymbol{x}}}\mathbb{E}_{P_{\boldsymbol{z},\boldsymbol{x}_t|\boldsymbol{x}}}\|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}(\boldsymbol{x}_t, \boldsymbol{z})\|_2^2 \propto \mathbb{E}_{P_{\boldsymbol{x}}}\mathbb{E}_{P_{\boldsymbol{z},\boldsymbol{x}_t|\boldsymbol{x}}} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{z}, t)\|_2^2.$
- Hyper Encoder: latent → hyper-latent $\boldsymbol{H}_s$ w/ smaller spatial resolution
- Codebook trained with VQ loss: $\mathcal{L}_{\mathrm{VQ}} = \mathbb{E}_{\boldsymbol{h}_s}\left[\|sg(\boldsymbol{h}_s) - \boldsymbol{z}_q\|_2^2 + \|sg(\boldsymbol{z}_q) - \boldsymbol{h}_s\|_2^2\right],$



Background: Hierarchical Representations

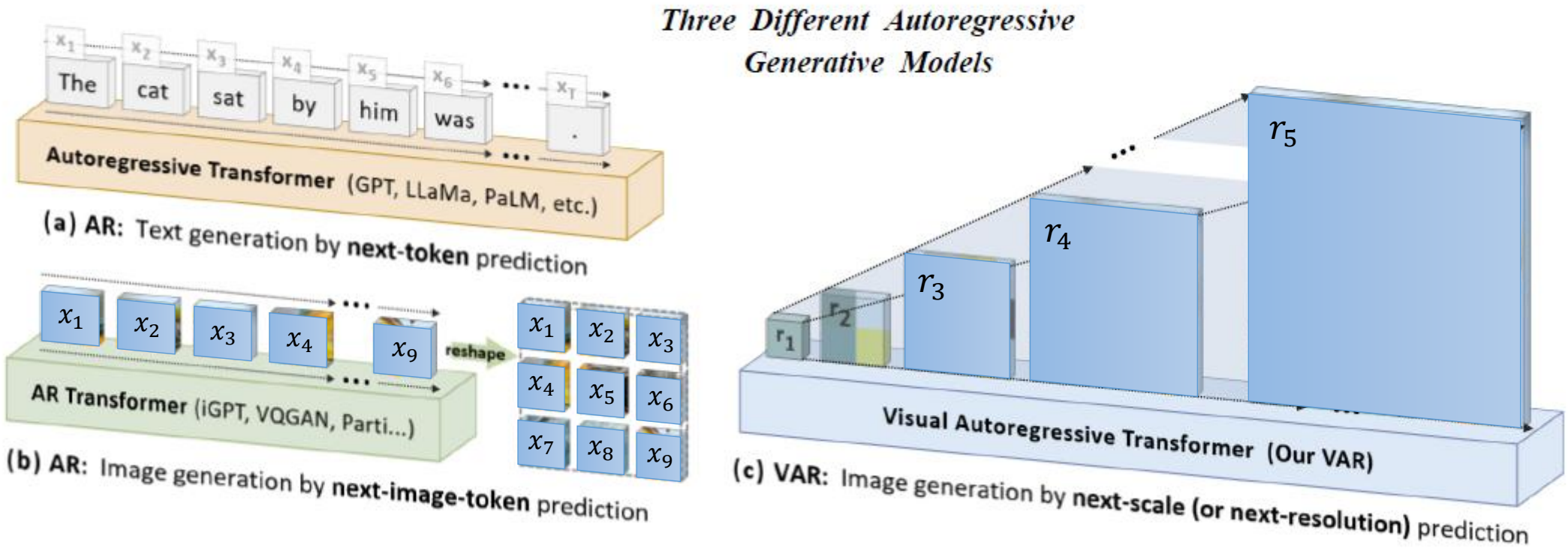
- Break down problems to different scales
- Sharable features across levels
- Deeper layers → Higher levels

Area	CV	NLP
Low-level Feature	Pixels, Edges	Words, Phrases
Mid-level Feature	Patterns	Syntactic
High-level Feature	Objects	Semantic



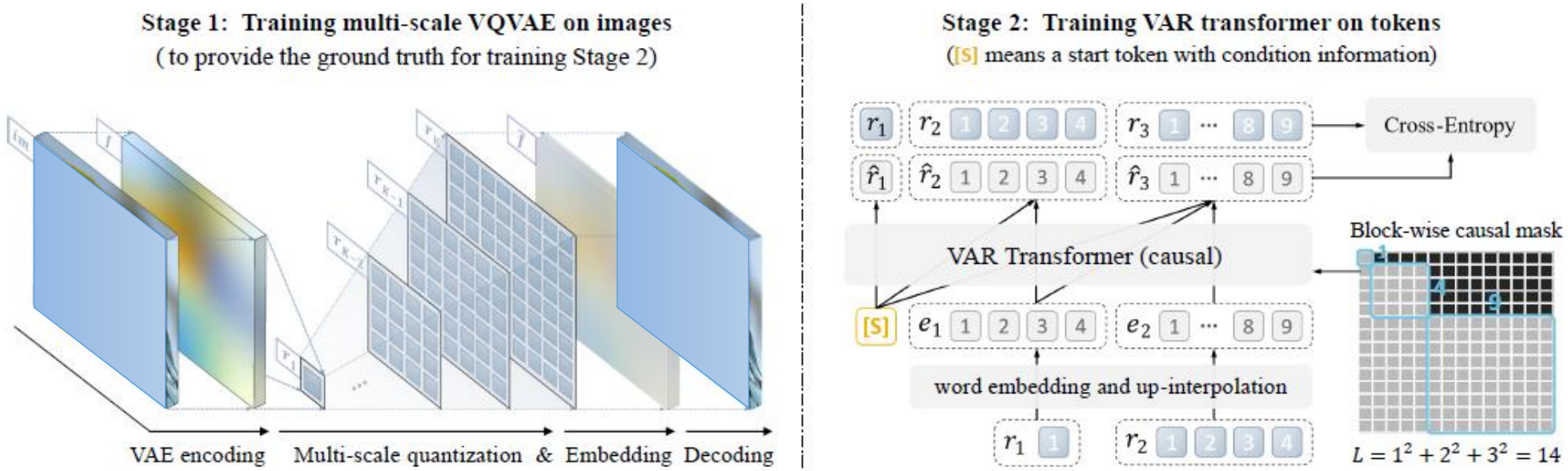
Background: Visual Autoregressive Modeling (VAR)

- Problem: Image data is non-sequential
- Modify AR: predict next token → predict next scale



Tian et al. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. (NeurIPS 2024 best paper)

Background: Visual Autoregressive Modeling (VAR)



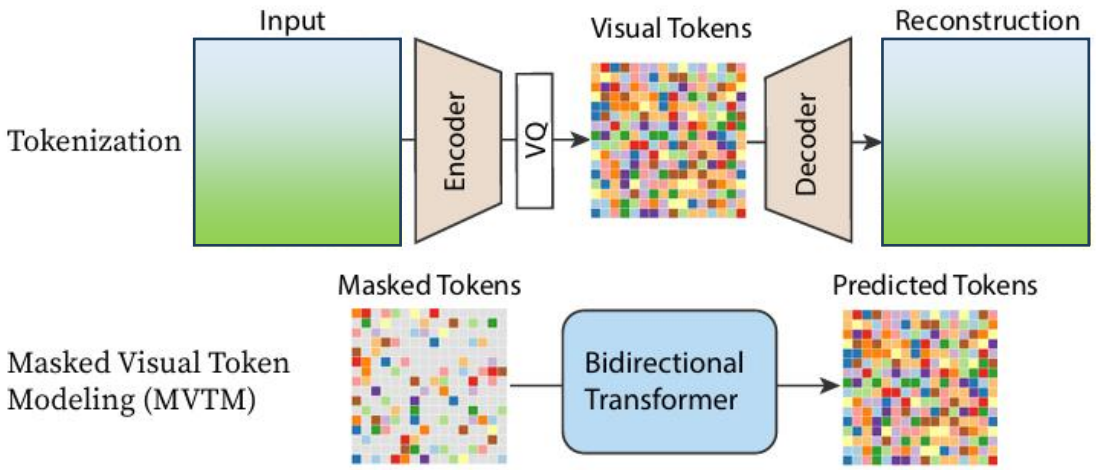
- Complexity of AR: $\sum_{i=1}^{n^2} i^2 = \frac{1}{6} n^2 (n^2+1)(2n^2+1) \sim \mathcal{O}(n^6).$

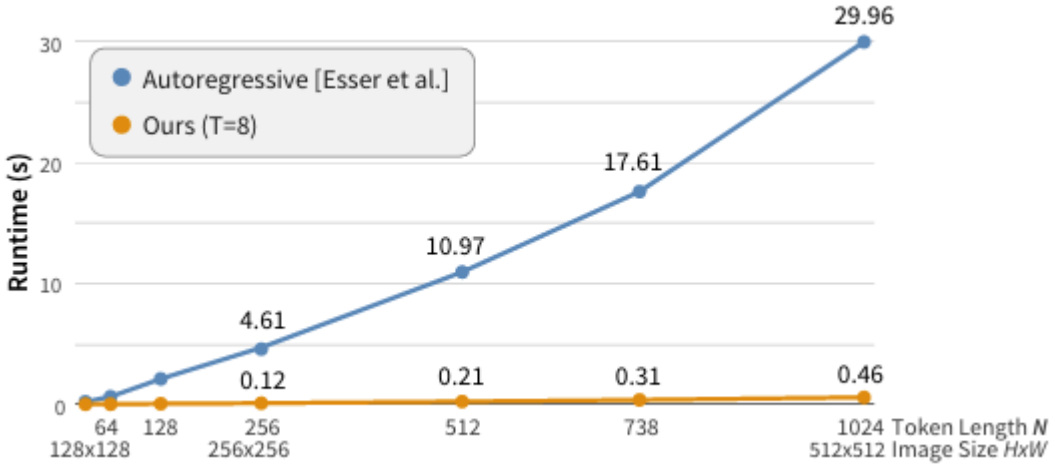
- Complexity of VAR: $\sum_{i=1}^{k} n_i^2 = \sum_{i=1}^{k} a^{2\cdot(k-1)} = \frac{a^{2k}-1}{a^2-1} \sum_{k=1}^{\log_a(n)+1} \left(\frac{a^{2k}-1}{a^2-1}\right)^2 \sim \mathcal{O}(n^4).$

Tian et al. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. (NeurIPS 2024 best paper)

Background: Masked Generative Image Transformer (MaskGIT)

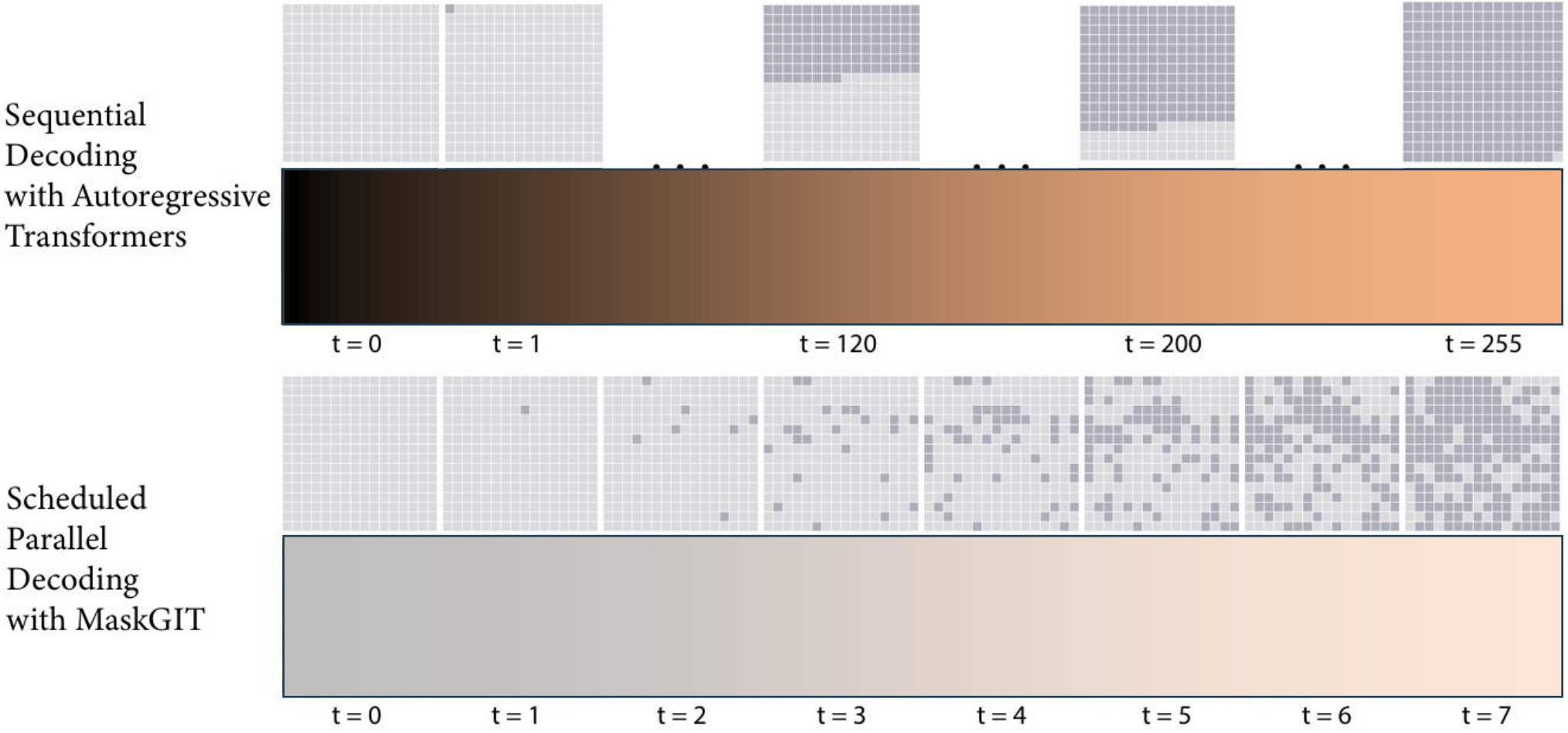
- Training: similar proxy task to the mask prediction in BERT
- Inference: non-autoregressive, only keep confident ones
- Much faster than VQGAN





Chang et al. MaskGIT: Masked Generative Image Transformer. (CVPR 2022)

Background: Masked Generative Image Transformer (MaskGIT)



Chang et al. MaskGIT: Masked Generative Image Transformer. (CVPR 2022)

Method: Motivations

- **Problems with PerCo:**
  - Relies on a proprietary LDM based on GLIDE (publicly unavailable)
  - Unclear if proprietary LDM works better than off-the-shelf models

- **Further Study: PerCo (SD)**
  - Ultra-low to extreme bitrate (0.003 ~ 0.03 bpp)
  - Using SD v2.1, similar to GLIDE
  - Relatively lightweight encoders and denoising network
  - Lower resolution (768×768 → 512×512)

Method: PerCo (SD)

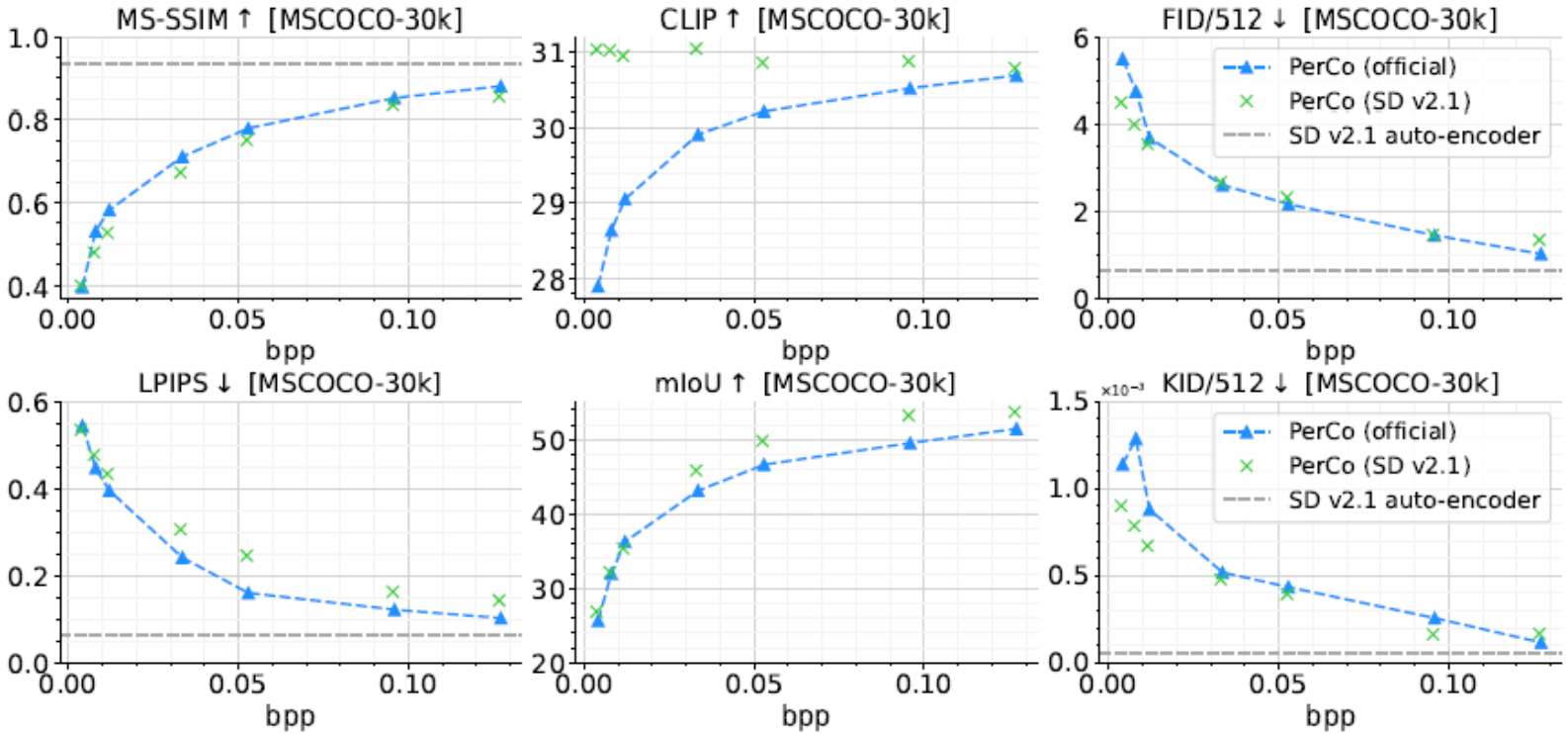
北京大学 PEKING UNIVERSITY

- Training steps: 150k iterations, 50% of PerCo.
- Peak learning rate: small learning rates (1e-5).
- U-Net finetuning: Tune whole U-Net due to resolution / distribution mismatch.
- Extended kernel: Init with zeros, encouraging the model to gradually incorporate additional conditional information (z_l).
- VQ-module: l_2-normalize codes, ensure stable training.

Table 1: Comparison of the design decisions: PerCo (official) vs. PerCo (SD). Key deviations are highlighted in gray and discussed in the main text.

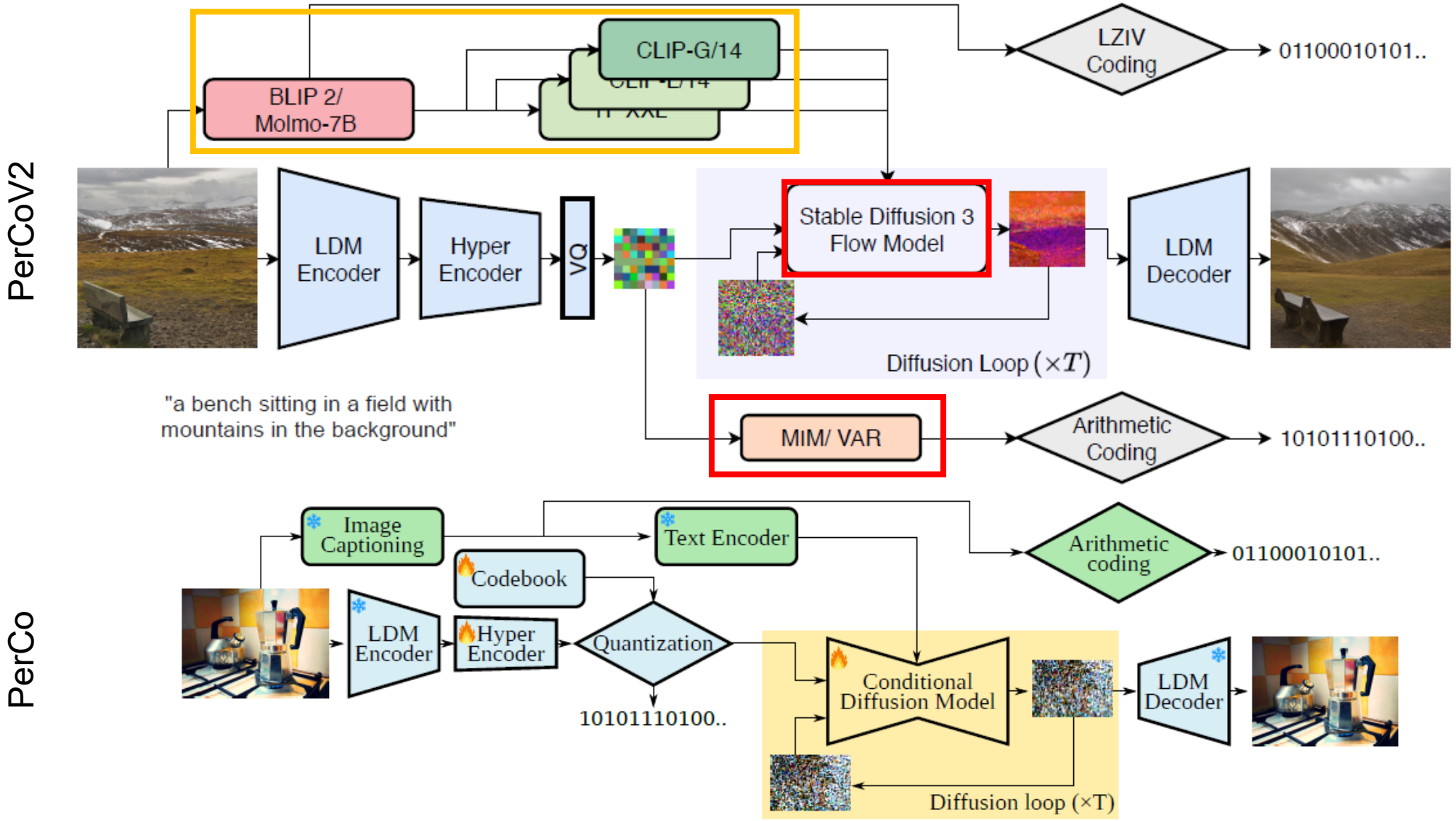
	PerCo (official)	PerCo (SD)
Training		
Training dataset	OpenImagesV6 [21] (9M)	OpenImagesV6 [21] (9M)
Optimizer	AdamW [27]	AdamW [27]
Training steps	5 epochs/ $\approx$ 300k	150k (50%)
Peak learning rate	$1e-4$	$1e-5$
Weight decay	0.01	0.01
Linear warm-up	10k	10k
Batch size	160 (w/o LPIPS), 40 w/ LPIPS	80 w/ LPIPS
U-Net finetuning	linear layers (15%)	all layers
LPIPS auxilliary loss	bit-rates $>$ 0.05bpp	all bit-rates
Text conditioning	drop in 10%	drop in 10%
Finetuning grid	50 steps	1000 steps (unchanged)
Extended kernel	random initialization	zero initialization
VQ-module	improved VQ [48]	improved VQ [48] + cosine similarity
Inference		
Scheduler	DDIM [40]	DDIM [40]
Denoising steps	5 for $>$ 0.05bpp, else 20	20
CFG [16]	3.0	3.0

Method: PerCo (SD)

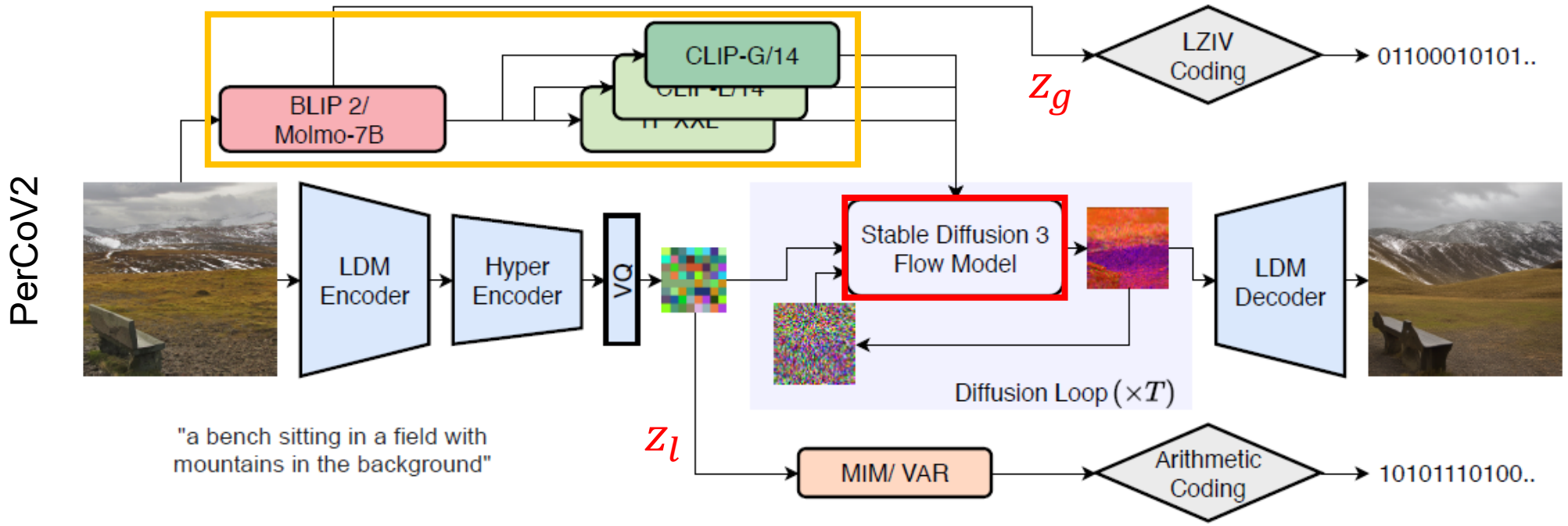


- Comparable or slightly better than official PerCo
- Conclusion: latent space design and the LDM capacity are crucial

Method: Overall Pipeline (Comparison)

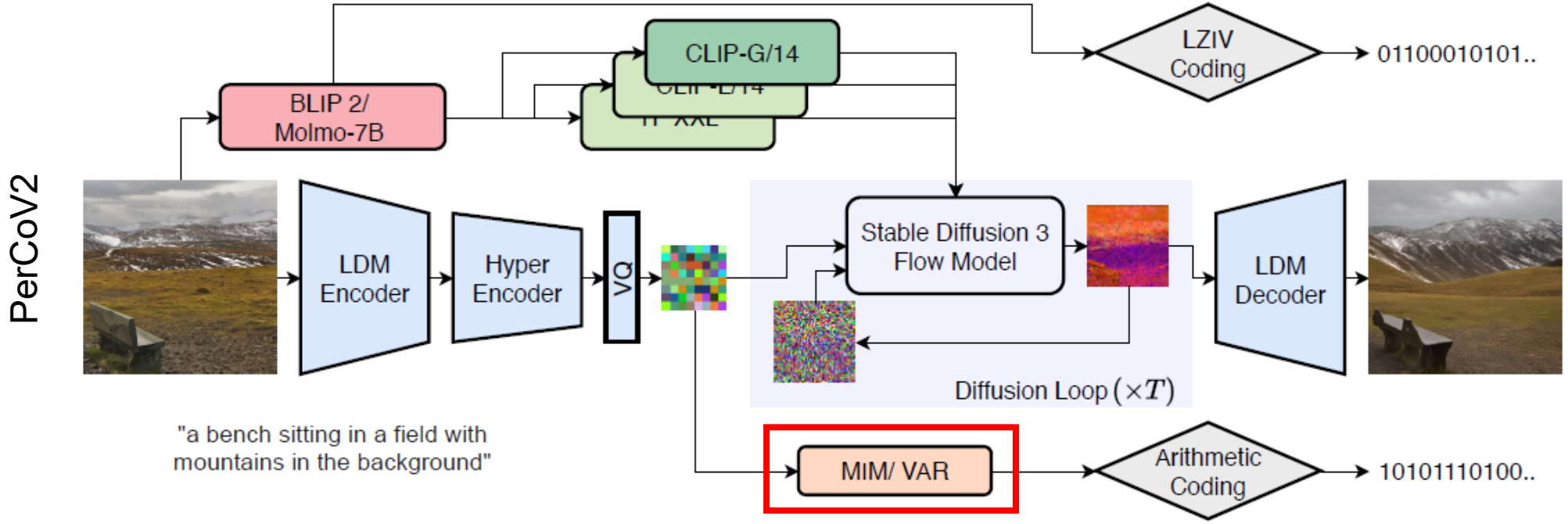


Method: Overall Pipeline



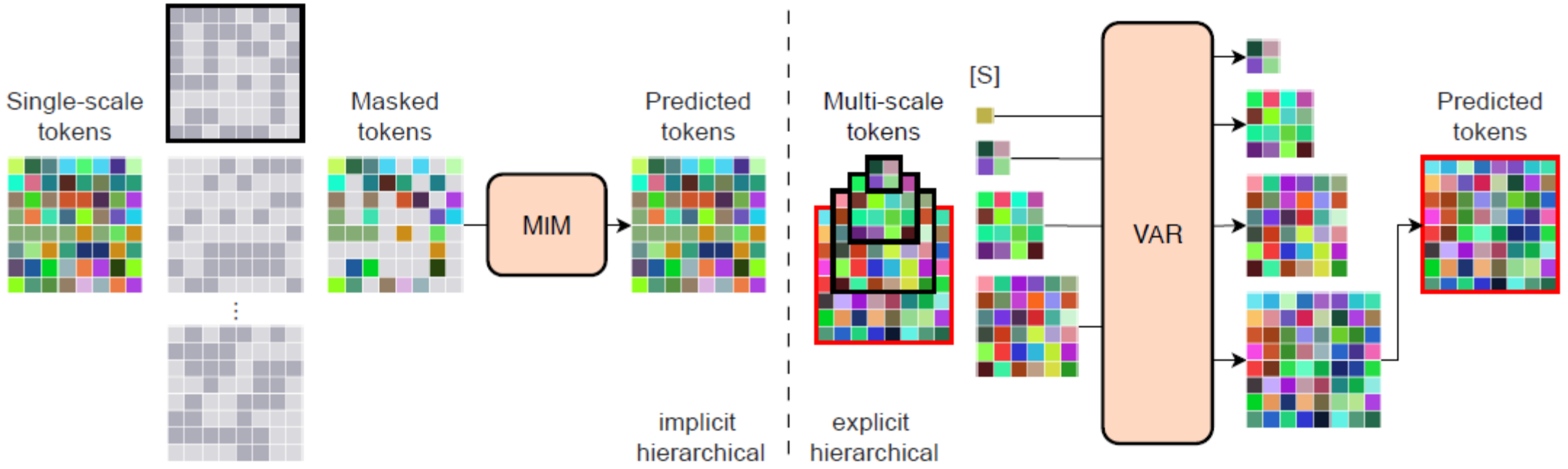
- Stable Diffusion 3 baseline:
  - BLIP 2/Molmo-7B Image Captioning; same text encoders as SD3
  - Flow model replacing UNet

Method: Overall Pipeline



- Continuous Entropy models (originally assumed uniform entropy):
  - Masked image model (MIM)
  - Visual autoregressive model (VAR)

Method: Hierarchical Masked Image Modeling



$$p(q) = \prod_{k=1}^{K} p(q_k \mid \mathcal{C}_k),$$

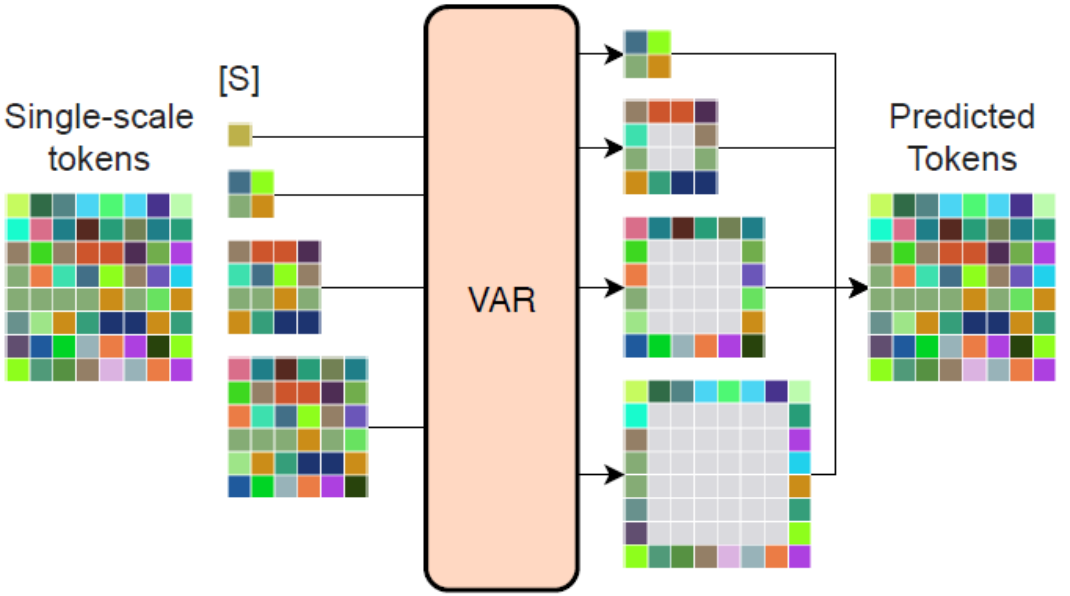
- MIM learns from token subsets
- VAR learns from given token map

Method: Hierarchical Masked Image Modeling

- **Implicit Hierarchical VAR**
  - Uniform Independent distributed:

$$r(z_l) = \frac{hw \log_2 V}{HW},$$

  - Reality: does not hold



  - MIM/VAR: Combining MIM and VAR
  - Learn more compact information

Method: Training / Optimization

- **First stage**
  - Conditional flow matching objective:

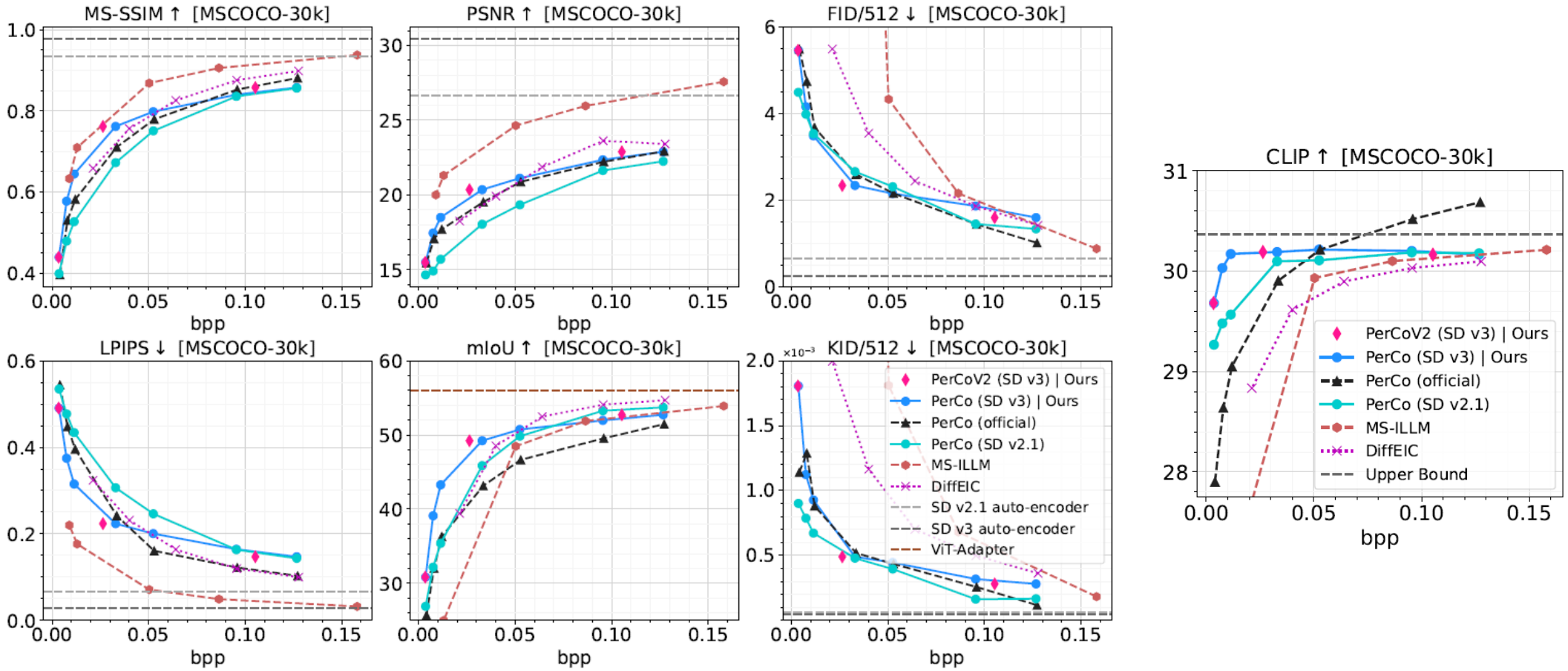
$$\mathcal{L}_{\text{CFM+}}(\Theta) = \mathbb{E}_{t,q(x_1),p(x_0)} \left\| v_t(\phi_t(x_0), z) - (x_1 - (1 - \sigma_{\min})x_0) \right\|^2 .$$

  - $z = (z_l, z_g)$: local and global latent
  - $\Theta = (\theta_1, \theta_2)$: parameters of flow model and hyper encoder
  - Drop z_g in 10% training iterations:

$$v_t' = v_t(\phi_t(x_0), (z_l, \emptyset)) + \lambda\big(v_t(\phi_t(x_0), (z_l, z_g)) - v_t(\phi_t(x_0), (z_l, \emptyset))\big).$$

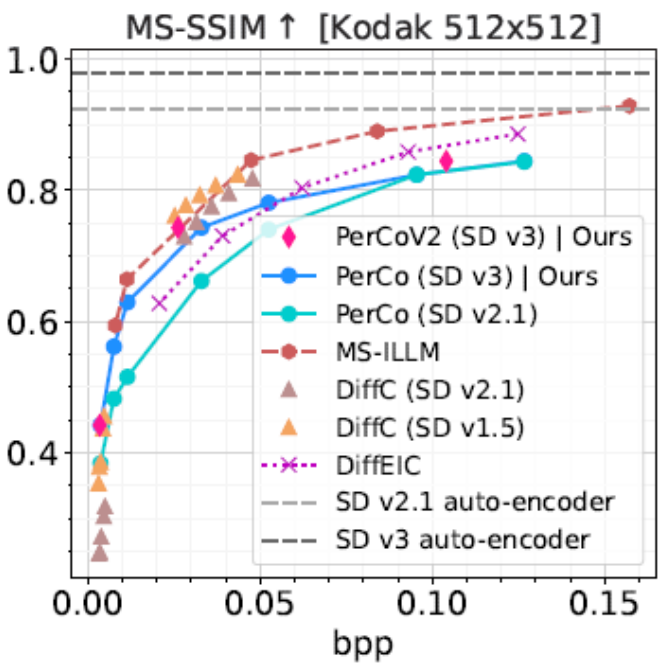
- **Second Stage**
  - Train MIM/VAR on hyper encoder representations

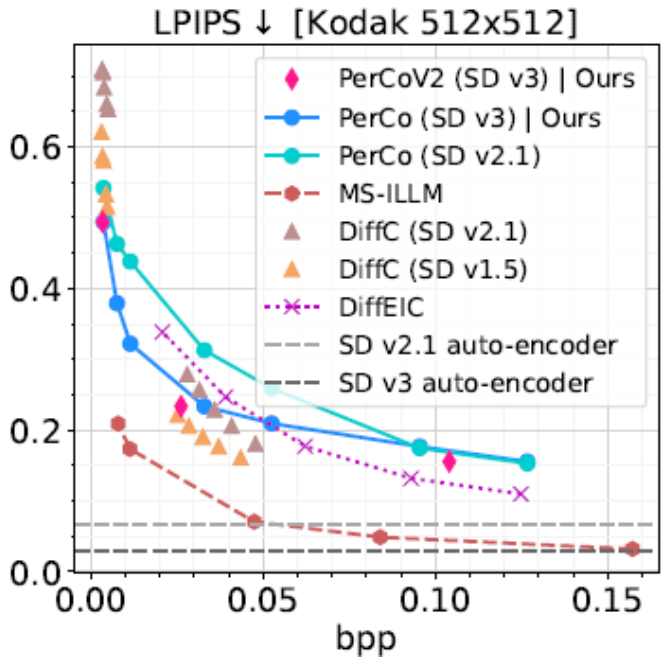
Experiments: Quantitative Comparison

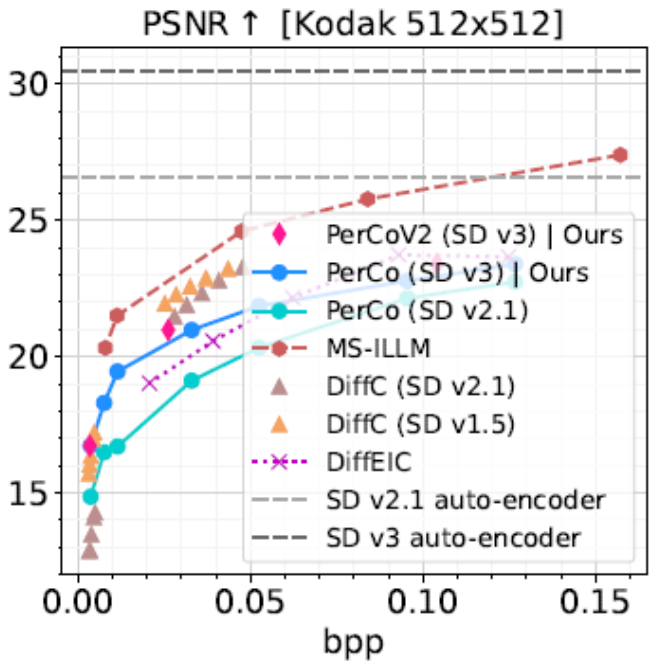


- Excels at ultra-low bitrate on MSCOCO-30k benchmark (FID, KID, mIoU)
- Not so effective at higher bitrates

Experiments: Quantitative Comparison

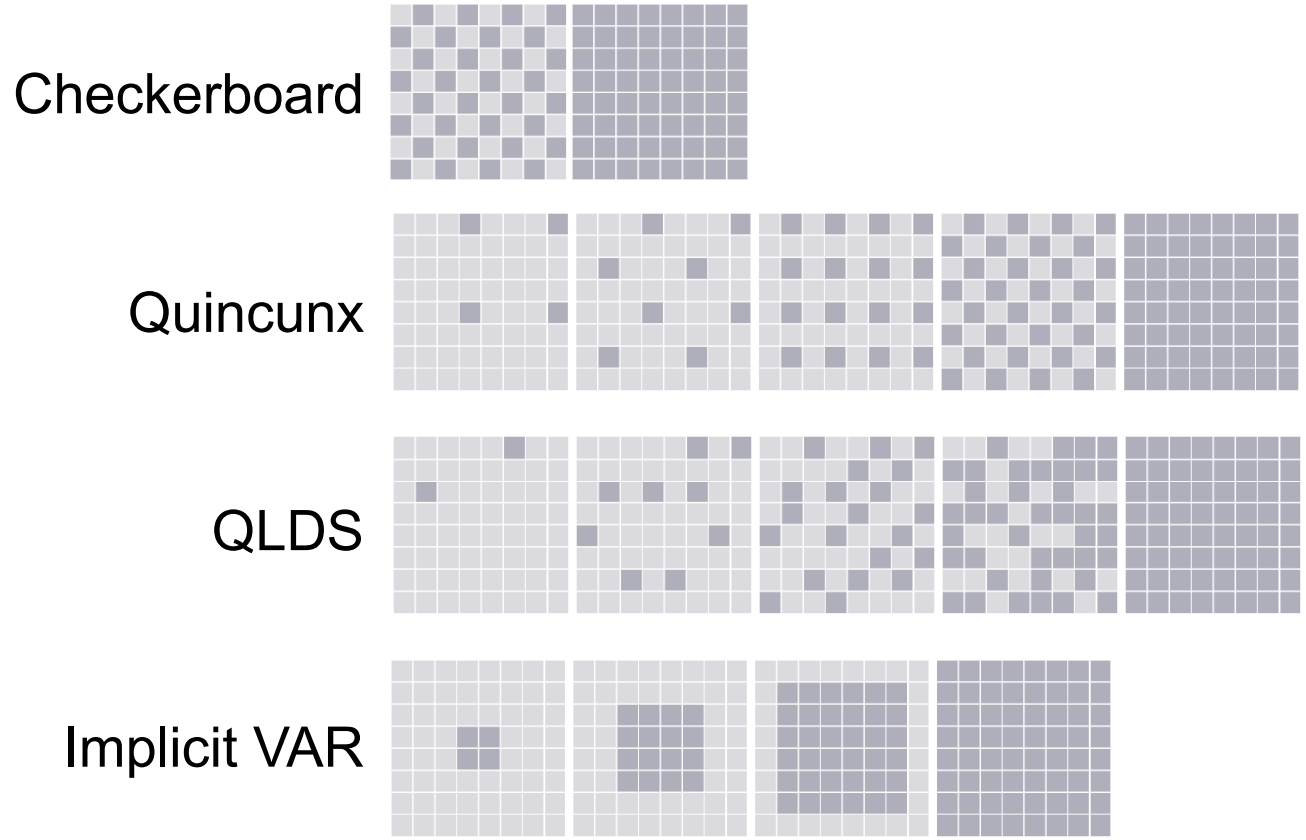




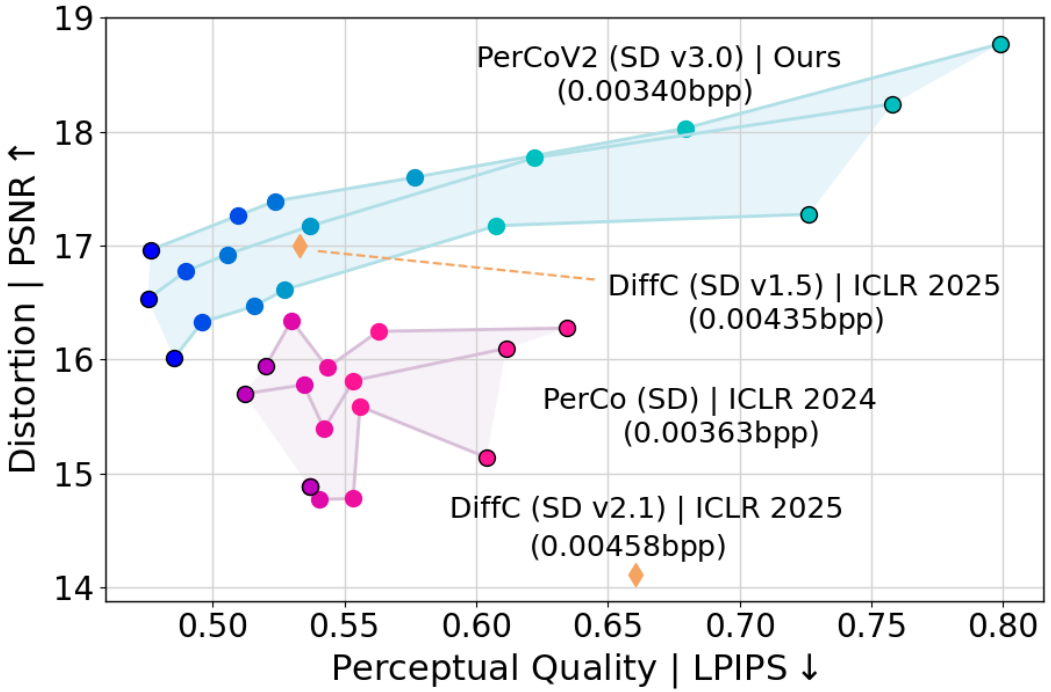


Experiments: Hierarchical Entropy Modeling Methods

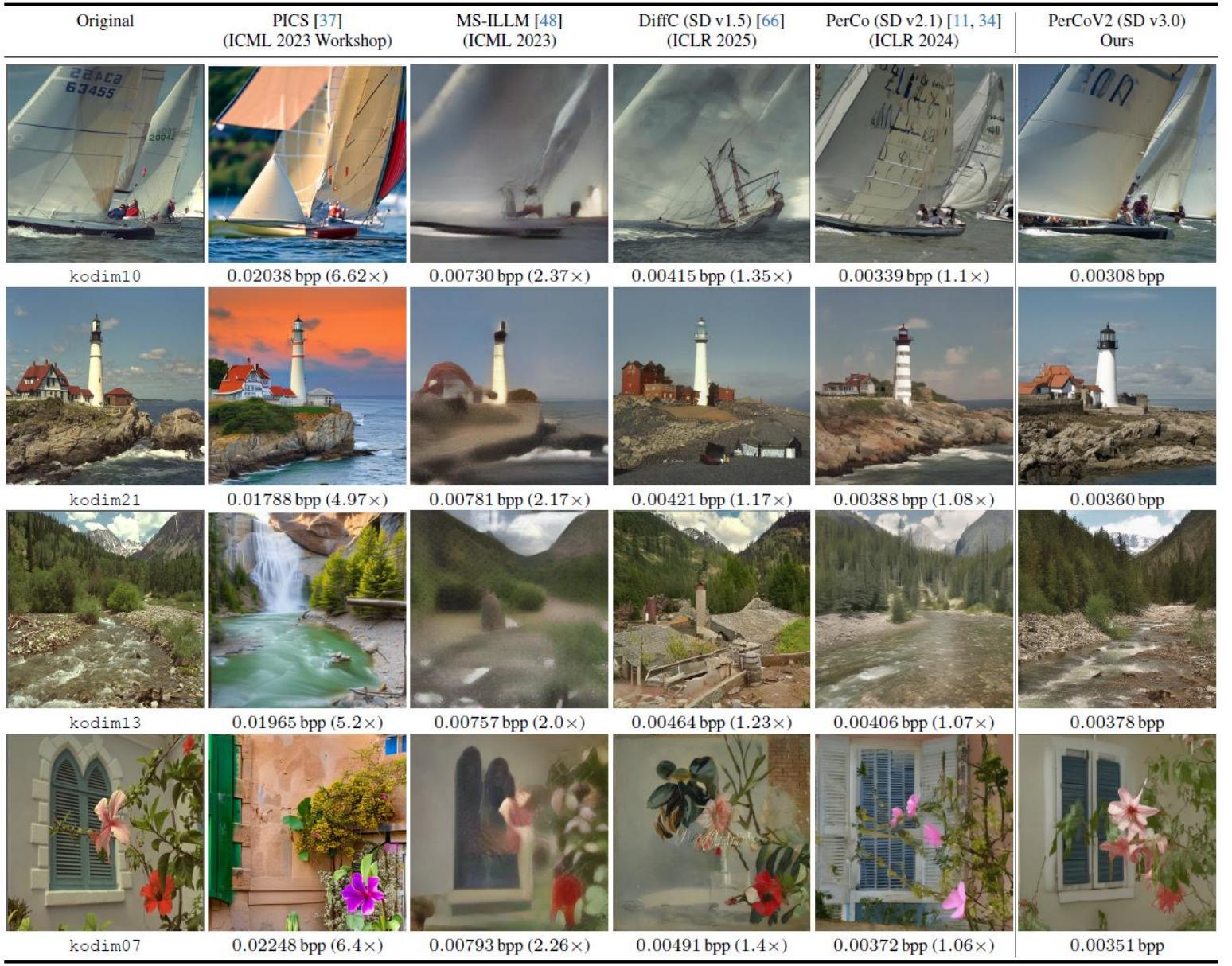
Method	bpp	Savings (%)
Ultra-Low Bit-Rate Setting		
Baseline	0.00363	–
Implicit Hierarchical VAR	0.00348	4.13
Checkerboard [23]	0.00342	5.79
Quincunx [17]	0.00340	6.34
QLDS ($S = 5$) [45]	0.00340	6.34
Extreme-Low Bit-Rate Setting		
Baseline	0.03293	–
Checkerboard [23]	0.02854	13.33
Quincunx [17]	0.02697	18.10
QLDS ($S = 5$) [45]	0.02667	19.01
QLDS ($S = 12$) [45]	0.02616	20.56



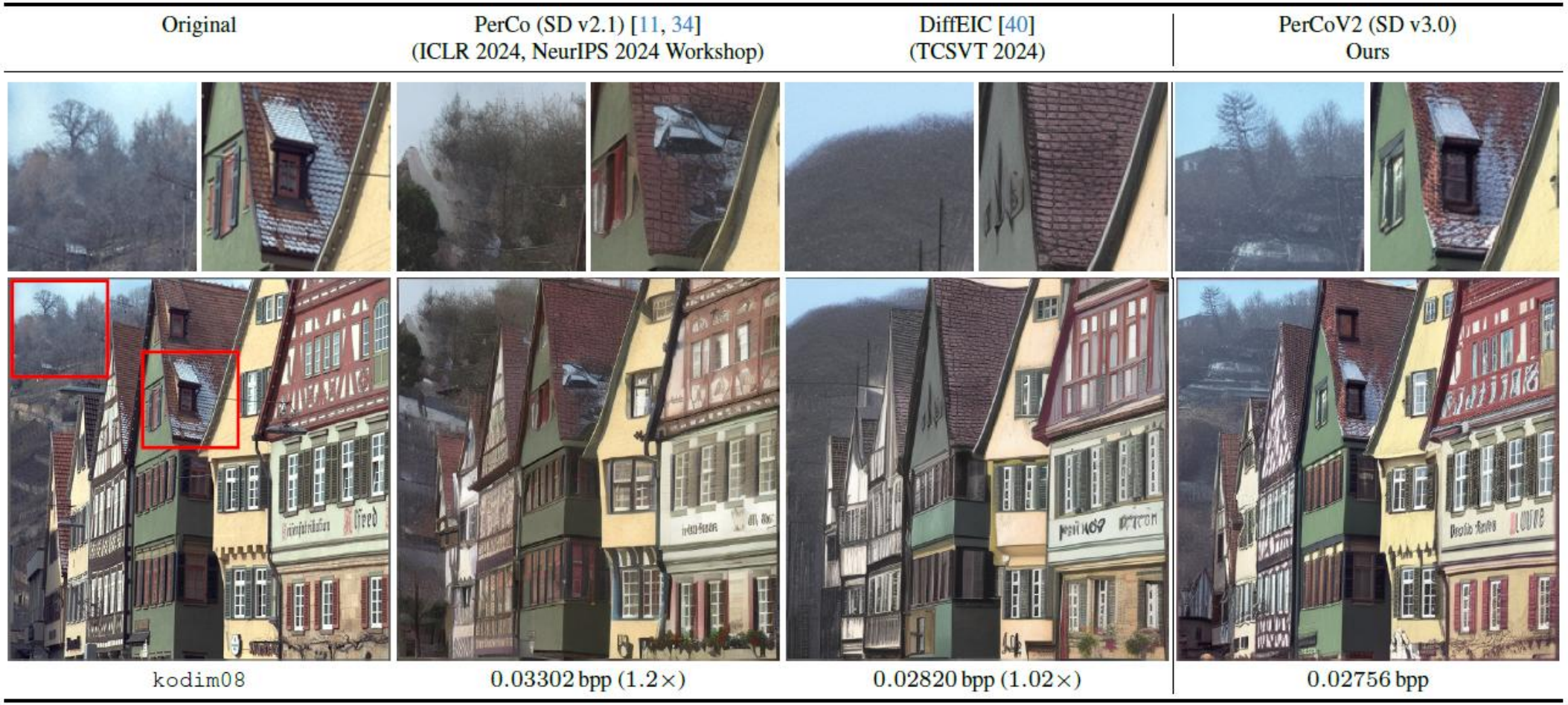
Experiments: Distortion-Perception Trade-off



Experiments: Subjective Quality



Experiments: Subjective Quality



Experiments: Different Text Configurations

Original

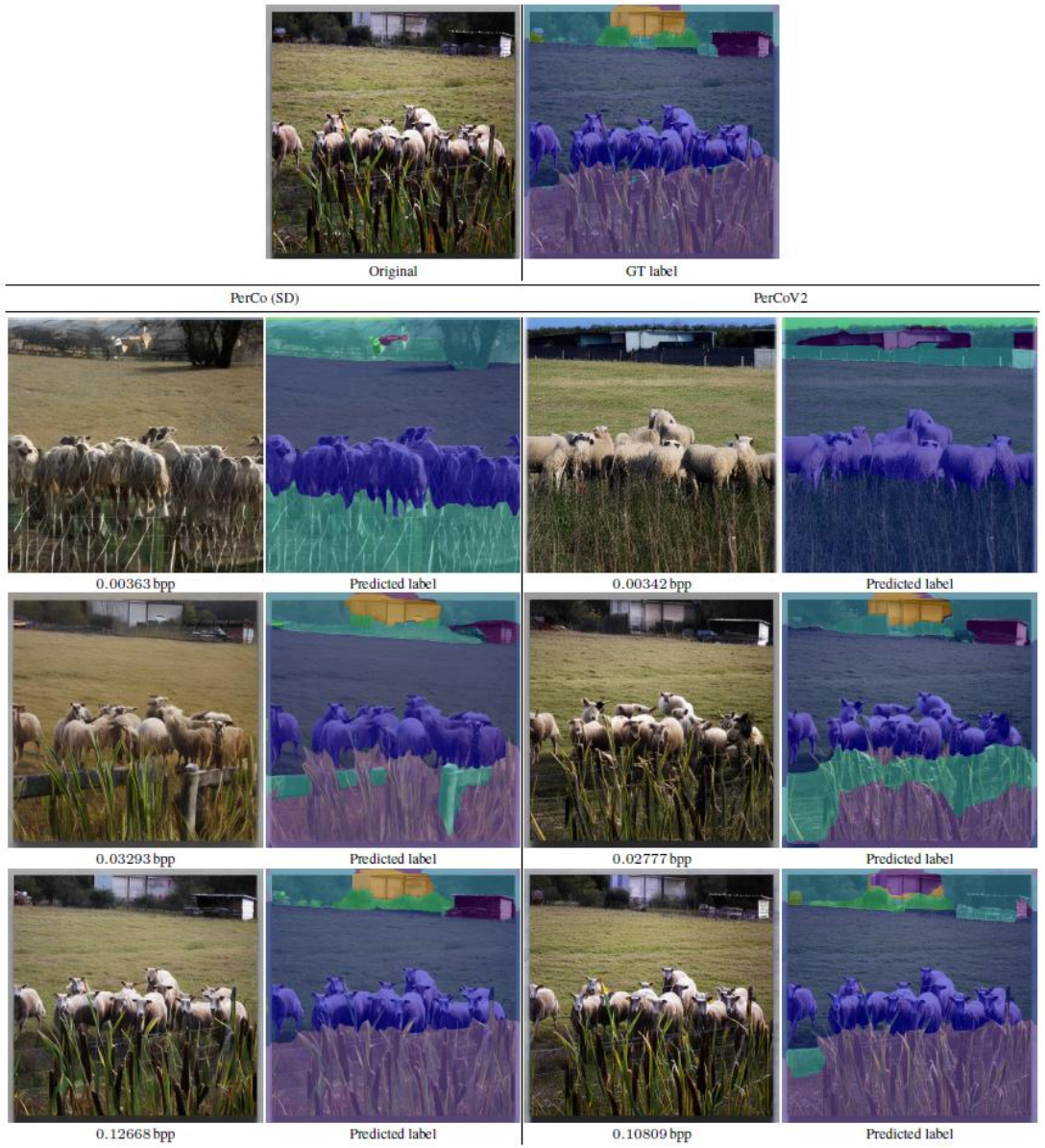
no text
Spatial bpp: 0.00171, Text bpp: 0.0

"a white fence with a lighthouse behind it" (BLIP 2)
Spatial bpp: 0.00171, Text bpp: 0.0014

"an old castle"
Spatial bpp: 0.00171, Text bpp: 0.0006

Experiments: Comparison to DiffC

Experiments: Semantic Alignment



Conclusion and discussion

北京大学 PEKING UNIVERSITY

- Proposes a novel and open framework for ultra-low bitrate PerCo
- Improves entropy coding efficiency
- Higher perceptual quality

- Discussion / Limitations:
  - Not effective on higher bitrates
  - Scheduling/VAR usage to be improved
  - Unsuitable for highly sensitive data
  - Code still not released!

Thanks for listening!

Presenter: Wen Si
2025.11.16