SQ-VAE:

Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization

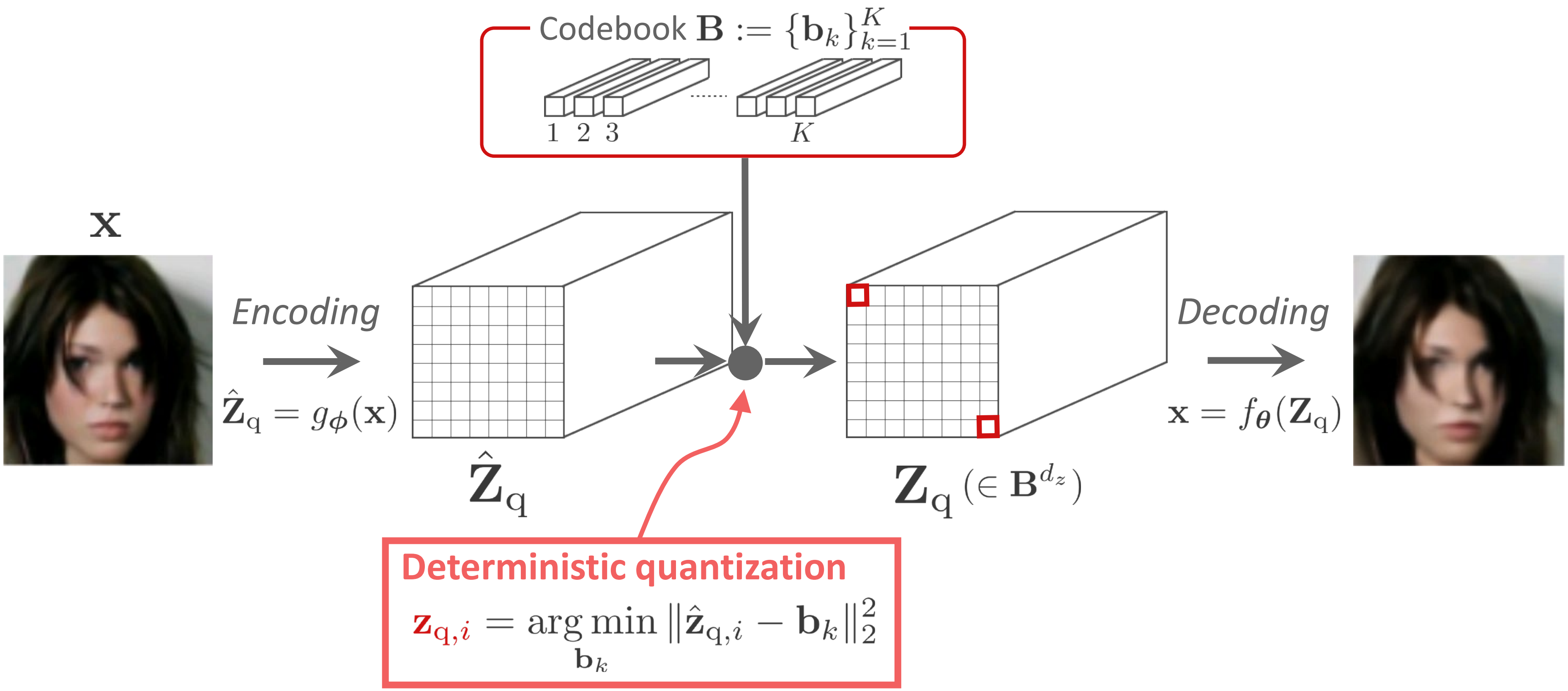
Yuhta Takida[1], Takashi Shibuya[1], WeiHsiang Liao[1], Chieh-Hsin Lai[1], Junki Ohmura[1], Toshimitsu Uesaka[1], Naoki Murata[1], Shusuke Takahashi[1], Toshiyuki Kumakura[2], Yuki Mitsufuji[1]

[1]Sony Group Corporation,

[2]Sony Corporation of America

SONY

VQ-VAE



$$\mathcal{L}_{\mathrm{VQ}} = \underbrace{-\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}})}_{\text{reconstruction}} + \underbrace{1.0 \times \|\mathrm{sg}[g_{\phi}(\mathbf{x})] - \mathbf{Z}_{\mathrm{q}}\|_F^2 + \beta \times \|g_{\phi}(\mathbf{x}) - \mathrm{sg}[\mathbf{Z}_{\mathrm{q}}]\|_F^2}_{\text{codebook + commitment losses}}$$

VQ-VAE

Heuristics:

- Stop gradient operator
- EMA update only for commitment loss
- Codebook reset (optional)

Hyperparameters:

- Coefficients for balancing loss functions
- Weighting for EMA update
- Parameters for codebook reset (optional)

Problems:

- Often suffer from "codebook collapse"
  (only few codebook elements are used)
- Need to tune "codebook size" as well
  (dimension and number of codebook elements)

Summary

Questions

- ✓ Can we eliminate common heuristics from VQ-VAE training?
- ✓ Can we reduce # of hyperparameters?
- ✓ Can we enhance codebook usage (circumvent "codebook collapse")?

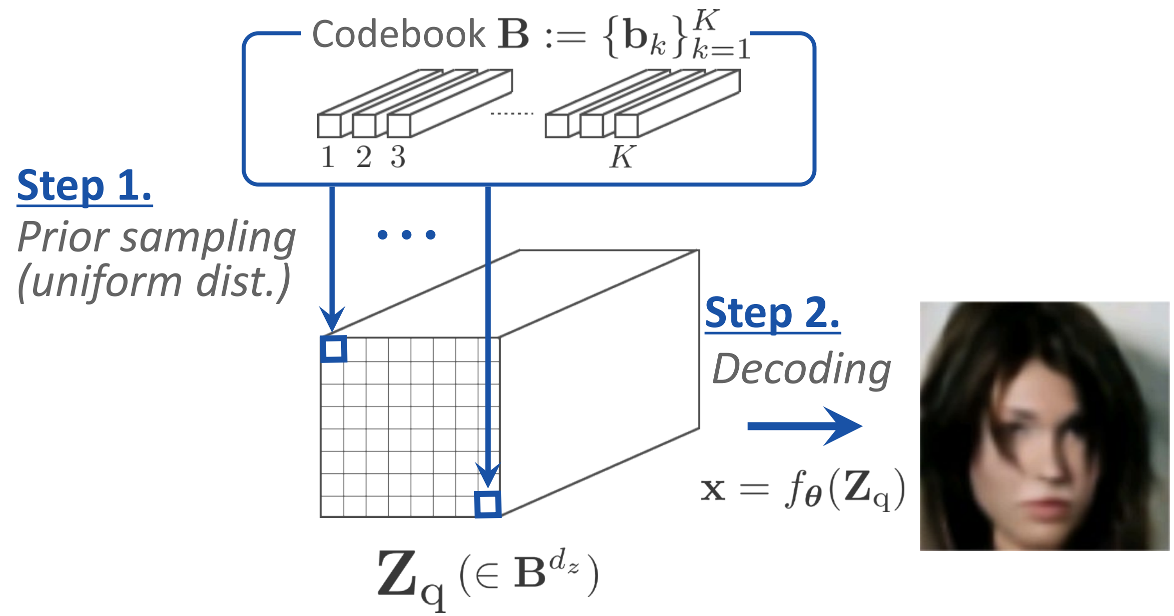
Our work

- Formulated VAE equipped with learnable codebook as
  Stochastically **Q**uantized-**V**ariational **A**uto**E**ncoder (SQ-VAE)
- Derived two variants of SQ-VAE:
  - Gaussian SQ-VAE for continuous distribution
  - von Mises-Fisher (vMF) SQ-VAE for categorical distribution

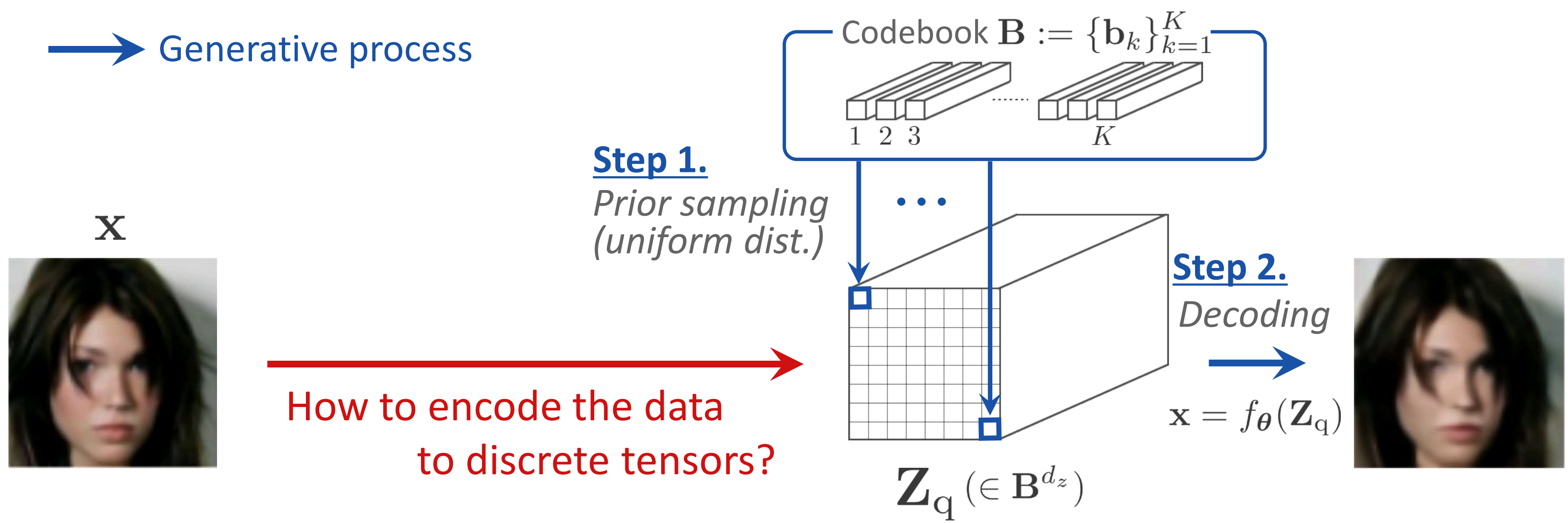
SQ-VAE addresses the above questions naturally

Probabilistic processes in SQ-VAE

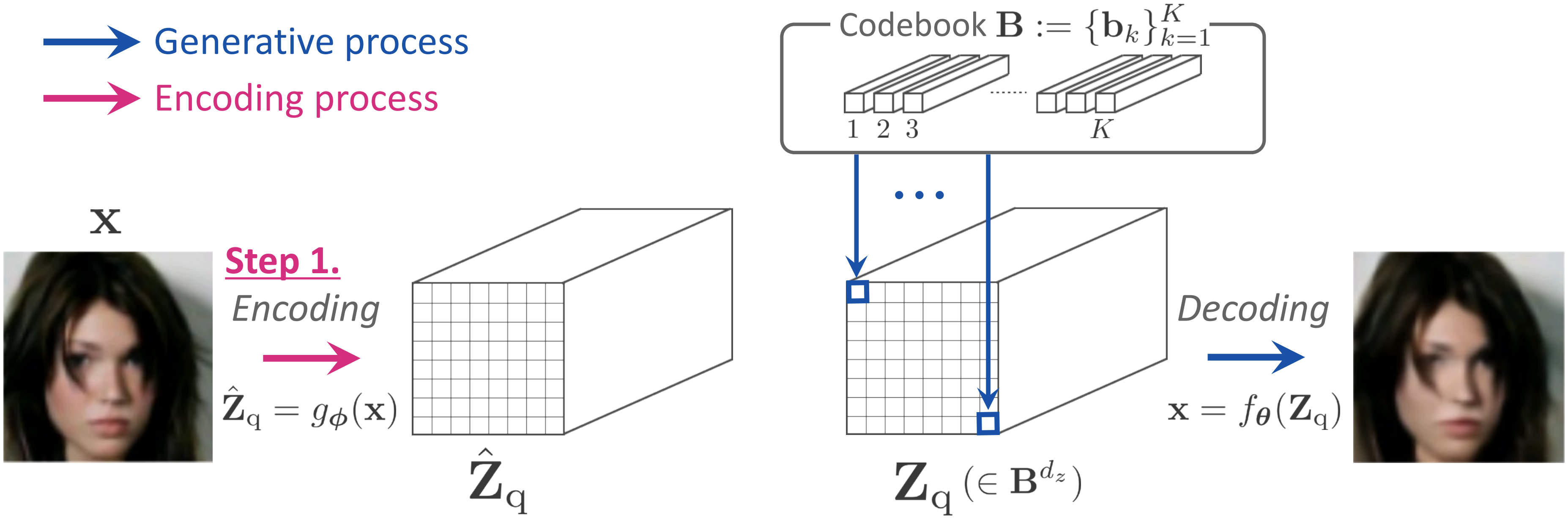
→ Generative process



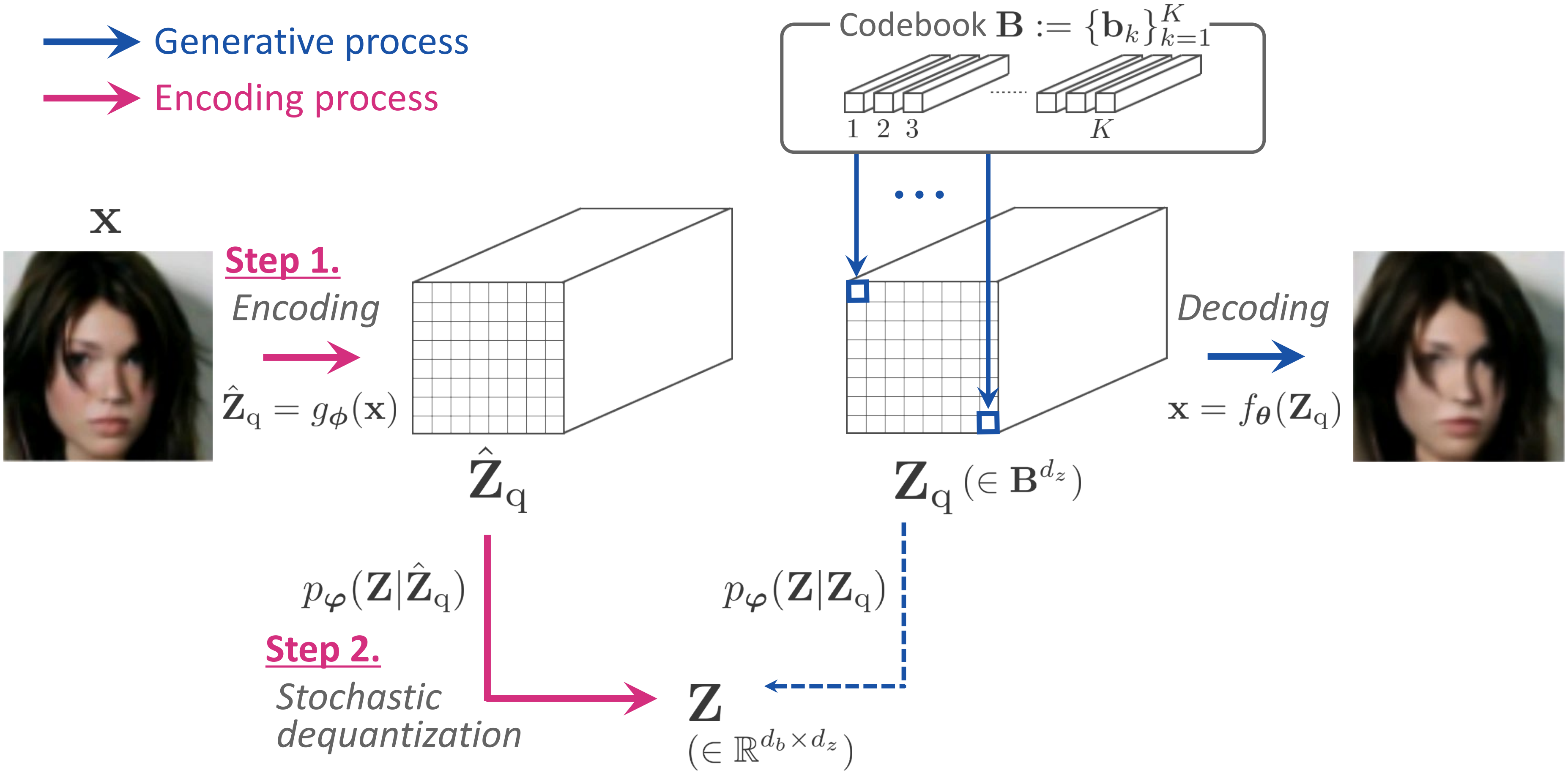
Probabilistic processes in SQ-VAE



Probabilistic processes in SQ-VAE

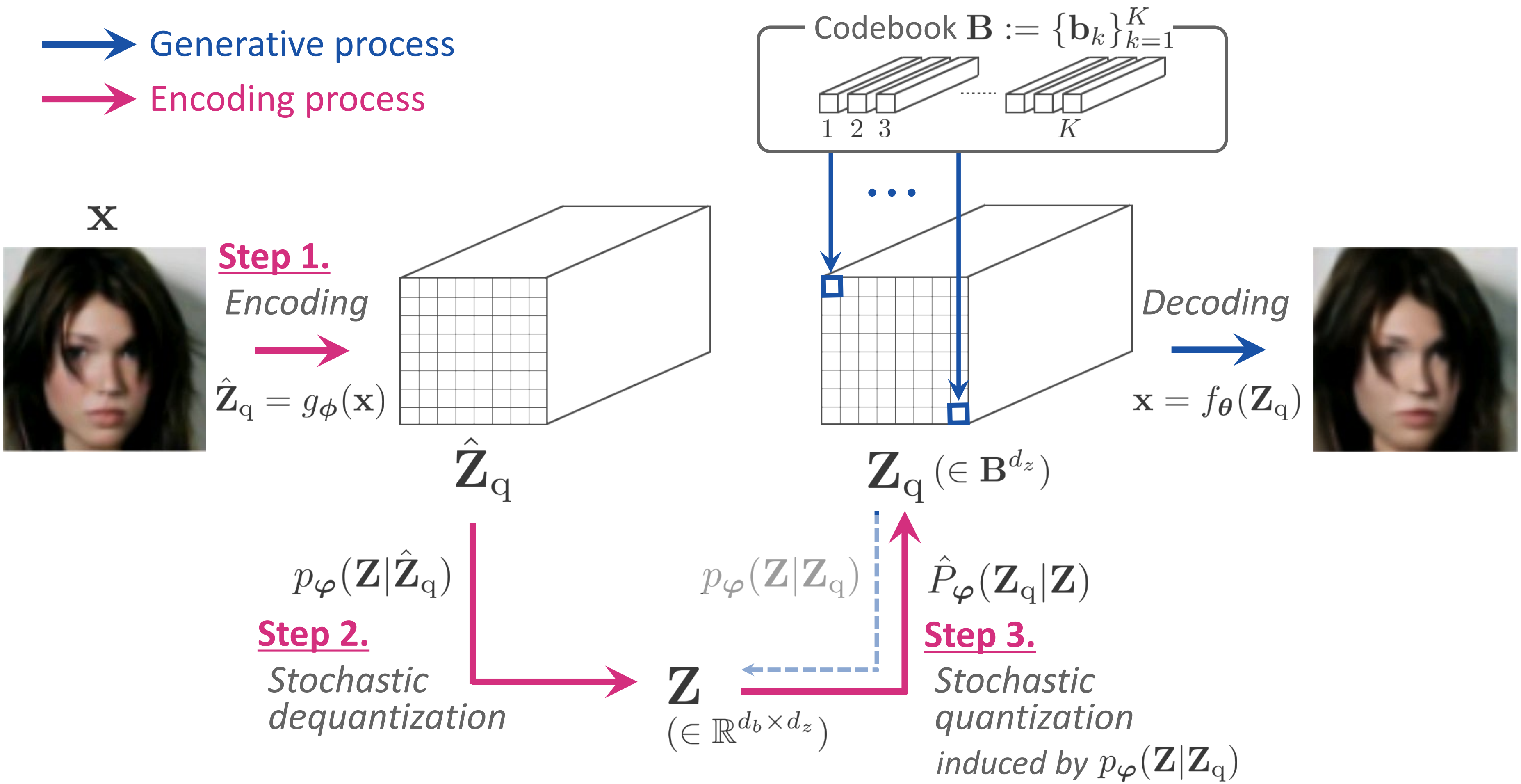


Probabilistic processes in SQ-VAE

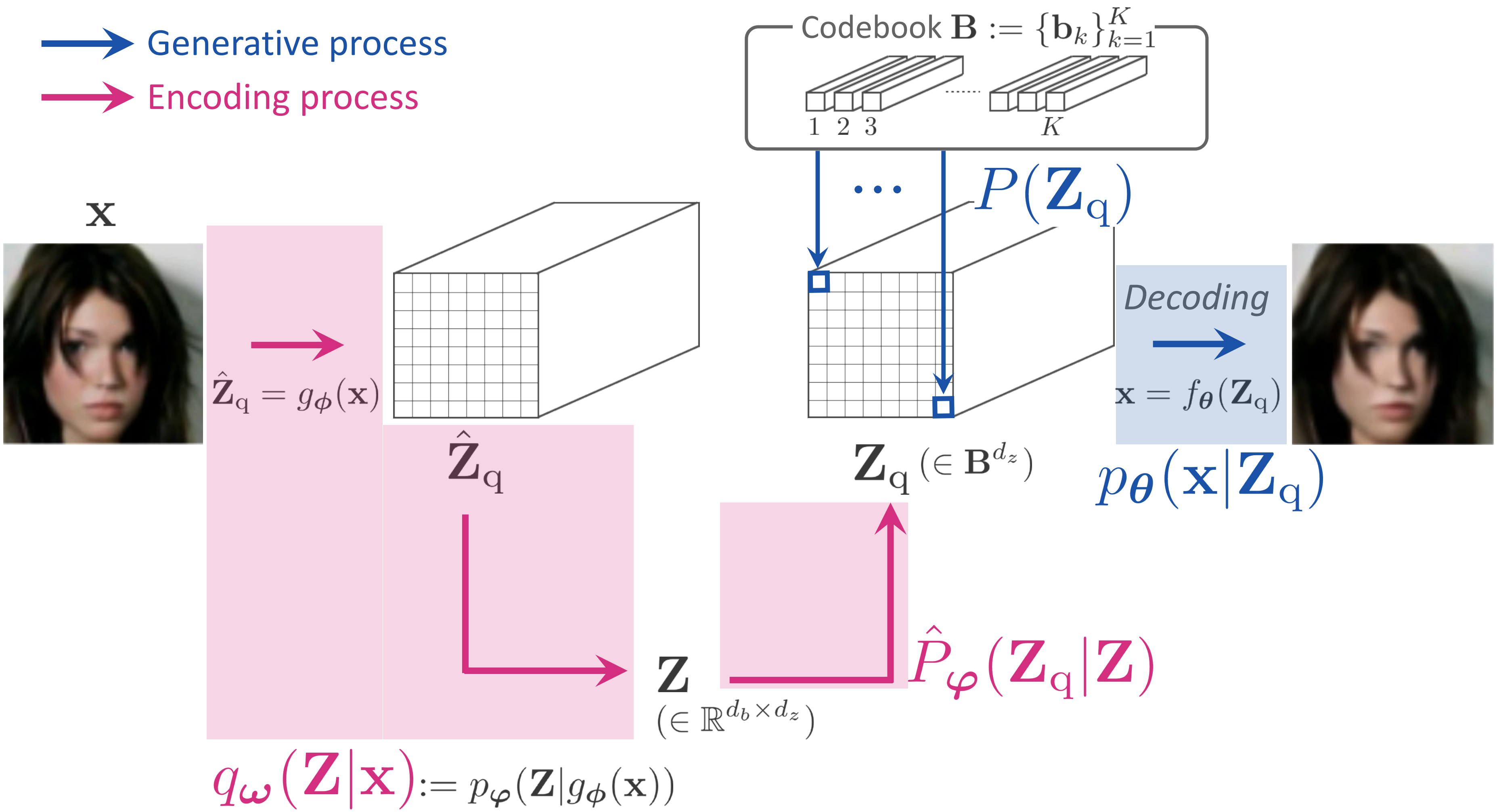


SONY

Probabilistic processes in SQ-VAE



Probabilistic processes in SQ-VAE



SONY

ELBO for generic SQ-VAE

Decoder/encoder distributions

- Overall decoding: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}})P(\mathbf{Z}_{\mathrm{q}})$ *prior sampling -> decode*
- Overall encoding: $q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})$ *encode -> dequantize -> quantize*

Objective function

ELBO

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) \geq -\mathcal{L}_{\mathrm{SQ}}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\omega},\mathbf{B}) + \text{const.},$$

$$\mathcal{L}_{\mathrm{SQ}}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\omega},\mathbf{B}) := \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}})p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})P(\mathbf{Z}_{\mathrm{q}})}{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\right]$$

$$=^{+} \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}}) + \log\frac{p_{\boldsymbol{\varphi}}(\mathbf{Z}|\mathbf{Z}_{\mathrm{q}})}{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})}\right] + \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})}H(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z}))$$

Reconstruction — Discrepancy b/w $\mathbf{Z}_{\mathrm{q}}$ and $\hat{\mathbf{Z}}_{\mathrm{q}}$ — Entropy regularization of codebook

SONY

Formulation of Gaussian SQ-VAE

Probabilistic processes (modeled by Gaussian w/ isotropic covariances)

- Gaussian decoder: $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{Z}_{\mathrm{q}}) = \mathcal{N}(f_{\boldsymbol{\theta}}(\mathbf{Z}_{\mathrm{q}}), \sigma^2\mathbf{I})$
- Gaussian dequantization: $p_{\boldsymbol{\varphi}}(\mathbf{z}_i|\mathbf{Z}_{\mathrm{q}}) = \mathcal{N}(\mathbf{z}_{\mathrm{q},i}, \sigma_{\boldsymbol{\varphi}}^2\mathbf{I})$

(inducing) → Quantization: $\hat{P}_{\boldsymbol{\varphi}}(\mathbf{z}_{\mathrm{q},i} = \mathbf{b}_k|\mathbf{Z}) = \mathrm{softmax}_k\left(\left\{-\frac{\|\mathbf{z}_j - \mathbf{b}_k\|_2^2}{2\sigma_{\boldsymbol{\varphi}}^2}\right\}_{j=1}^{k}\right)$

Objective function

Balanced with trainable parameters

$$\mathcal{L}_{\mathcal{N}\text{-SQ}}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\omega},\mathbf{B}) := \mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})}\left[\frac{1}{2\sigma^2}\|\mathbf{x} - f_{\boldsymbol{\theta}}(\mathbf{Z}_{\mathrm{q}})\|_2^2 + \frac{1}{2\sigma_{\boldsymbol{\varphi}}^2}\|\mathbf{Z} - \mathbf{Z}_{\mathrm{q}}\|_F^2\right]$$

$$-\mathbb{E}_{q_{\boldsymbol{\omega}}(\mathbf{Z}|\mathbf{x})}\left[H\left(\hat{P}_{\boldsymbol{\varphi}}(\mathbf{Z}_{\mathrm{q}}|\mathbf{Z})\right)\right] + \frac{D}{2}\log\sigma^2 + \mathrm{const.}$$

Approximated by Gumbel-softmax trick

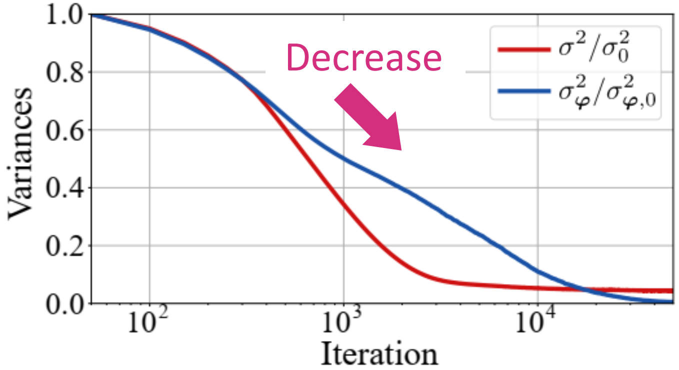
✓ *Any common heuristics (e.g., stop-gradient, EMA) are no longer needed*

✓ *# of hyperparameters is reduced to only one (for Gumbel soft-max trick)*

SONY

The effect of "Self-annealing"

From stochastic to deterministic quantization

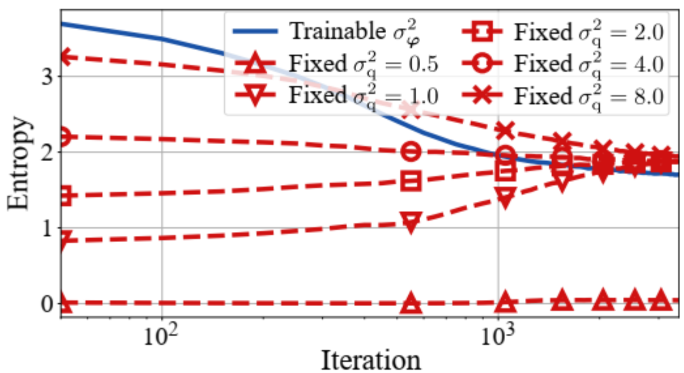


σ_φ^2 decreases along with σ^2 (see also Proposition 1)

$$\hat{P}_\varphi(\mathbf{z}_{q,i} = \mathbf{b}_k|\mathbf{Z}) \propto -\frac{\|\mathbf{z}_j - \mathbf{b}_k\|_2^2}{2\sigma_\varphi^2}$$

$\begin{cases} \sigma_\varphi^2 \to \infty \text{ makes } \hat{P}_\varphi(\mathbf{z}_{q,i} = \mathbf{b}_k|\mathbf{Z}) \text{ uniform dist.} \\ \sigma_\varphi^2 \to 0 \text{ induces deterministic quantization} \end{cases}$

The variational property reduces stochasticity of quantization



Gets close to deterministic quantization as iteration progresses

Zero entropy value means the quantization is deterministic

✓ *The self-annealing effect is expected to enhance codebook usage*

SONY

Gaussian SQ-VAE on vision dataset

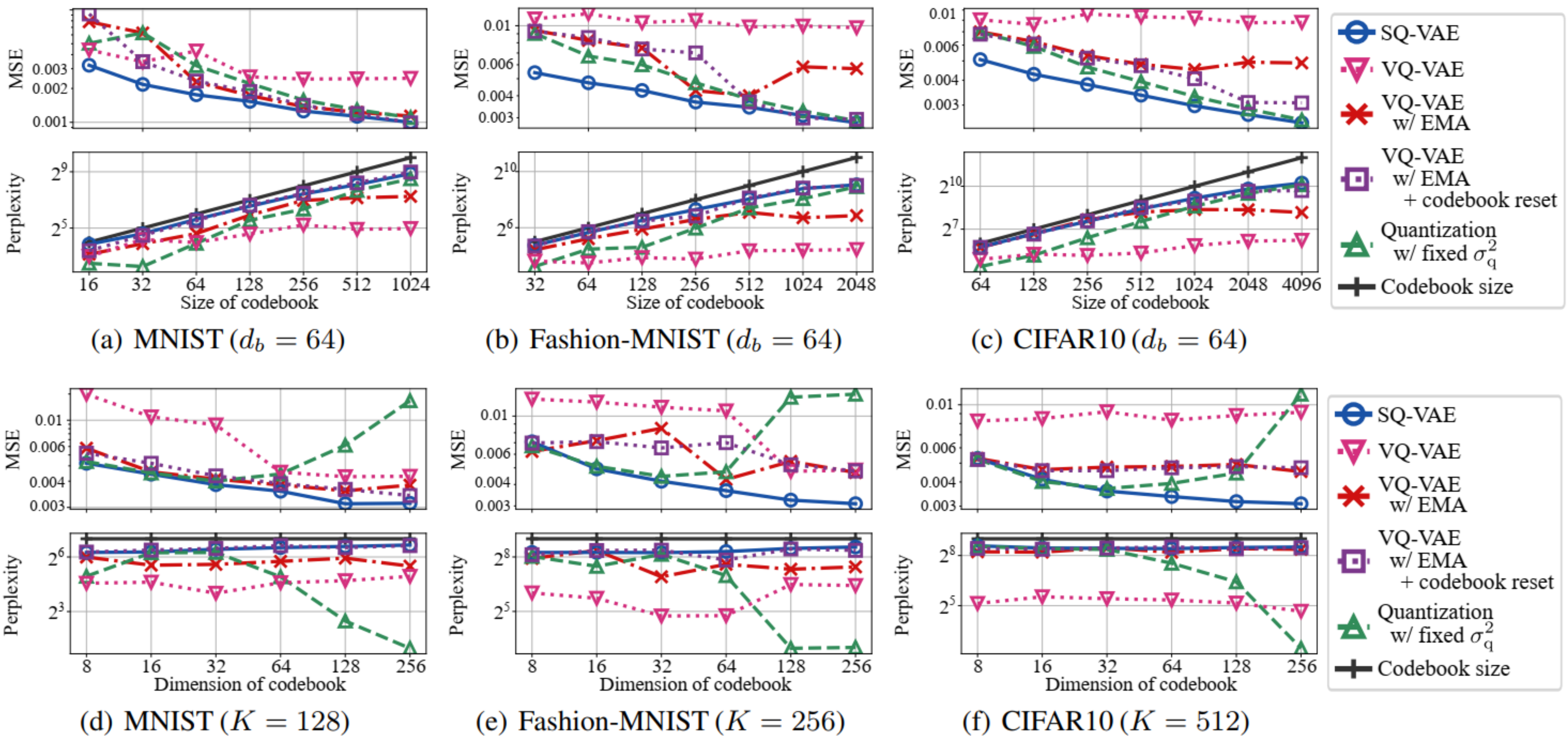


Figure 5. Empirical studies on the impact of codebook capacity examined on MNIST Fashion-MNIST and CIFAR10. (a)–(c) The size K is swept with the dimension d_b fixed to 64. (d)–(f) Various d_b values are tested with the size K fixed as 128, 256, and 512, respectively. The black lines with "+" marks indicate the upper bounds of the perplexities, i.e., K. All the y-axes are in log-scale.

SONY

Gaussian SQ-VAE on CelebA

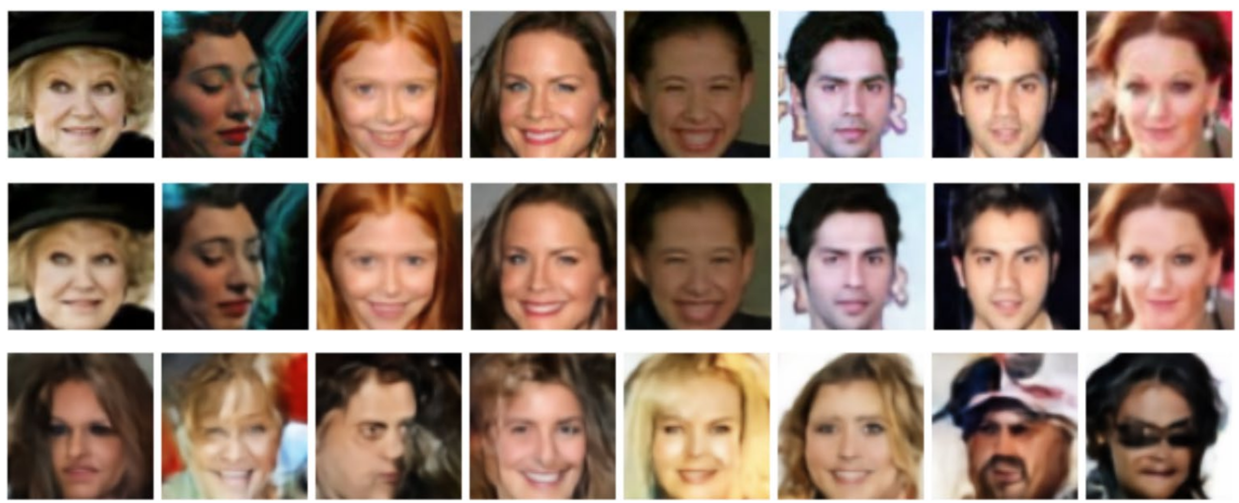
Table 2. Evaluation on CelebA. The MSE ($\times 10^3$) and reconstructed FID (rFID) are evaluated using the test set. The codebook capacity for the discrete latent space are set to $(n_b, k) = (64, 512)$. The Roman numerals for Gaussian SQ-VAEs correspond to those in Table 1. We also show the FID of samples generated with a prior learned with PixelCNN.

Model	Reconstruction		Generation	Latent manipulation (FID)			
	MSE	rFID	(FID)	Neighbor-3	Neighbor-5	Neighbor-10	Interpolation
VAE	4.79 ± 0.01	40.3 ± 0.3	–	–	–	–	–
VQ-VAE w/ EMA	1.33 ± 0.41	18.5 ± 5.1	42.0 ± 11.5	31.9 ± 14.8	42.8 ± 20.7	70.7 ± 35.4	28.2 ± 6.4
VQ-VAE w/ EMA+codebook reset	1.62 ± 0.36	22.0 ± 5.9	51.8 ± 10.8	39.7 ± 12.0	52.7 ± 14.7	83.2 ± 20.4	32.6 ± 7.1
Quantization w/ fixed σ_q^2	1.09 ± 0.01	15.9 ± 0.1	38.2 ± 0.9	20.0 ± 0.4	26.4 ± 0.8	41.5 ± 2.1	18.6 ± 0.3
Gaussian SQ-VAE (I)	$\mathbf{0.96} \pm 0.01$	14.8 ± 0.3	28.2 ± 0.9	17.8 ± 0.1	21.9 ± 0.1	$\mathbf{33.1} \pm 0.3$	$\mathbf{17.6} \pm 0.6$
Gaussian SQ-VAE (II)	0.98 ± 0.01	14.3 ± 0.2	$\mathbf{27.7} \pm 1.1$	17.8 ± 0.2	22.2 ± 0.4	34.0 ± 0.9	$\mathbf{17.6} \pm 0.1$
Gaussian SQ-VAE (III)	$\mathbf{0.96} \pm 0.00$	$\mathbf{13.9} \pm 0.1$	28.1 ± 0.3	$\mathbf{17.3} \pm 0.2$	$\mathbf{21.6} \pm 0.3$	33.5 ± 0.6	18.5 ± 0.4

Ground truth

Reconstruction

Random sampling
(w/ learned PixelCNN)

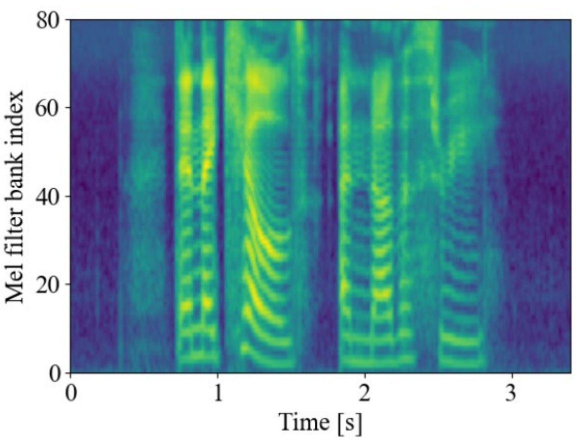


SONY

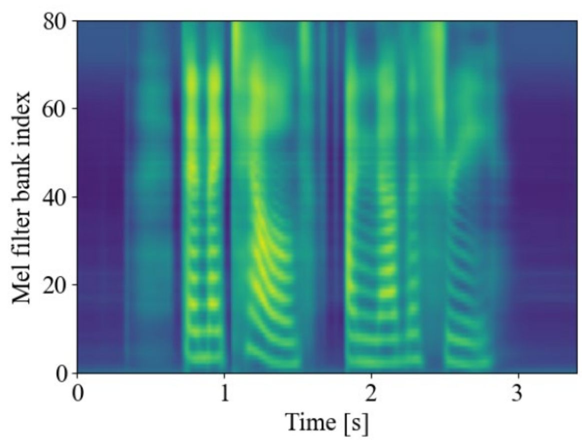
Gaussian SQ-VAE on speech dataset

Table 3. Evaluation on VCTK and ZeroSpeech 2019. The MSE (dB^2) of sample reconstruction is evaluated using the test set. We do not apply SQ-VAE (II) in this evaluation because of the variable length property of speech data and the different manipulations of speech signals between training and inference (See Appendix E.2).

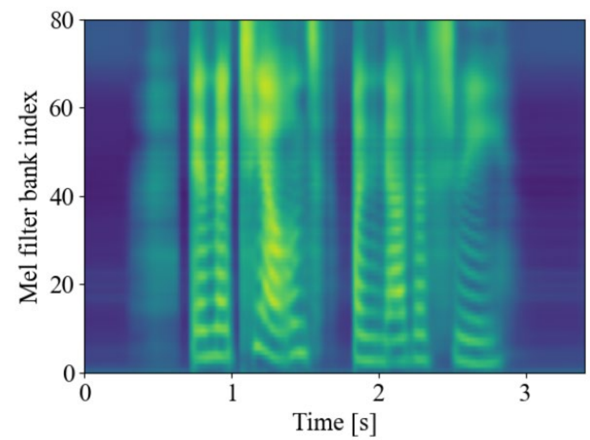
Model	MSE (dB^2)	
	VCTK	ZeroSpeech 2019
VQ-VAE w/ EMA	29.59 ± 0.25	34.33 ± 1.57
Gaussian SQ-VAE (I)	25.52 ± 0.08	33.17 ± 1.11
Gaussian SQ-VAE (III)	25.94 ± 0.22	34.35 ± 1.07
Gaussian SQ-VAE (IV)	$\mathbf{24.68} \pm 0.21$	$\mathbf{32.32} \pm 0.88$



Ground truth



Reconstruction by SQ-VAE (IV)



Reconstruction by VQ-VAE w/EMA

SONY

Recap

Formulation of SQ-VAE naturally

- Eliminates common heuristics
  such as EMA, stop-gradient and codebook reset
- Reduces # of hyperparameters to one
  (a temperature parameter for Gumbel softmax trick)
- Enhances codebook usage
  thanks to "self-annealing" effect in our quantization scheme

Code link: https://github.com/sony/sqvae

SONY