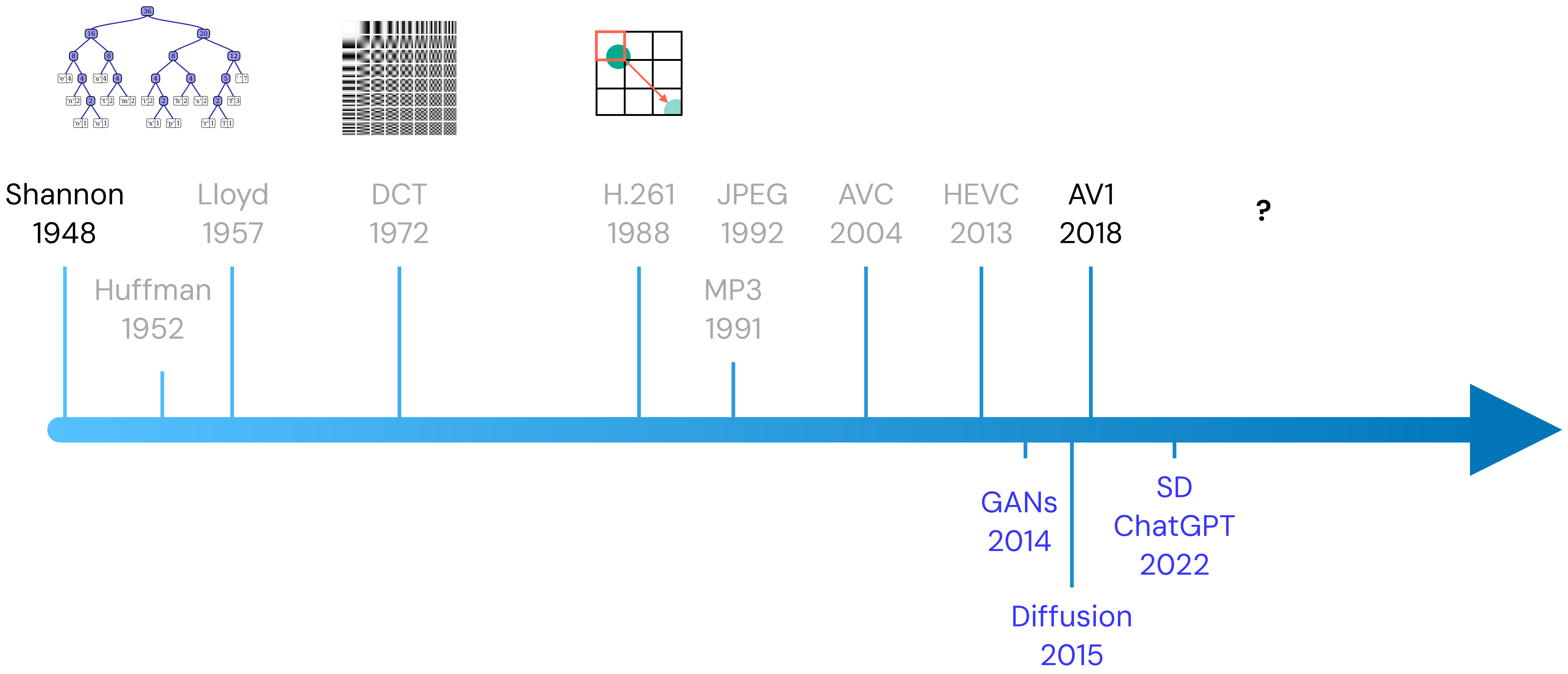
DCC 2025

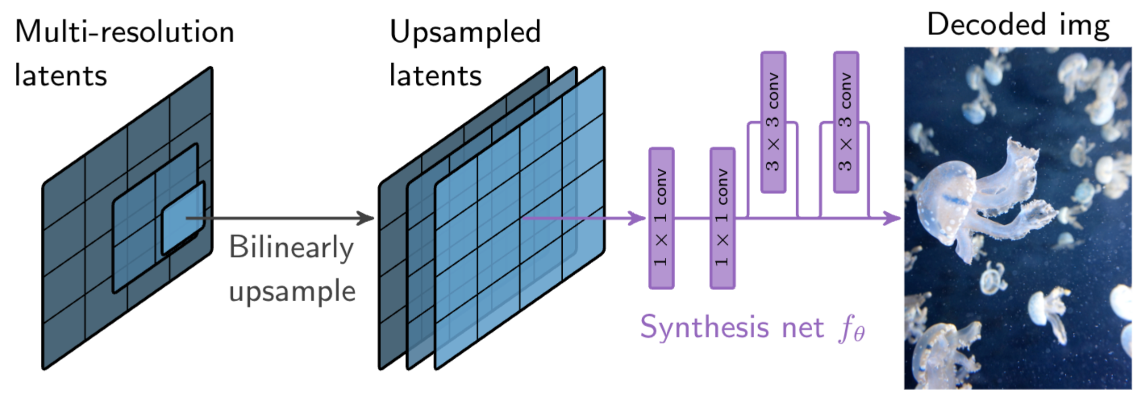
COMPRESSION IN THE AGE OF GENAI

PERFECT REALISM AT EVERY BITRATE

Lucas Theis, Mabyduck



Shannon
1948
Huffman
1952
Lloyd
1957
DCT
1972
H.261
1988
MP3
1991
JPEG
1992
AVC
2004
HEVC
2013
GANs
2014
Diffusion
2015
AV1
2018
SD
ChatGPT
2022
?



Low-complexity
neural compression

COOL-CHIC
(Ladune et al., 2023)

C3
(Kim et al., 2024;
Ballé et al., 2024)

Generative AI

Realism-distortion trade-offs

Reverse channel coding
(a.k.a. channel simulation)

Diffusion compression
(Ho et al., 2020; Theis et al., 2022;Yang et al., 2025)

Rate-distortion theory

Quantization

Transform coding


Input


JPEG
0.068 bpp


JPEG-XL
0.072 bpp


ELIC (MSE; based on He et al., 2022)
0.056 bpp


HFD (MSE + diffusion; Hoogeboom et al., 2023)
0.056 bpp


Input

Overview

Background
What is the cost of perfect realism?

Diffusion I: High-fidelity diffusion (HFD)
A transform coding approach

Diffusion II: DiffC
A novel approach using reverse channel coding

Realism

Accuracy of an ideal observer (± c)

Divergence

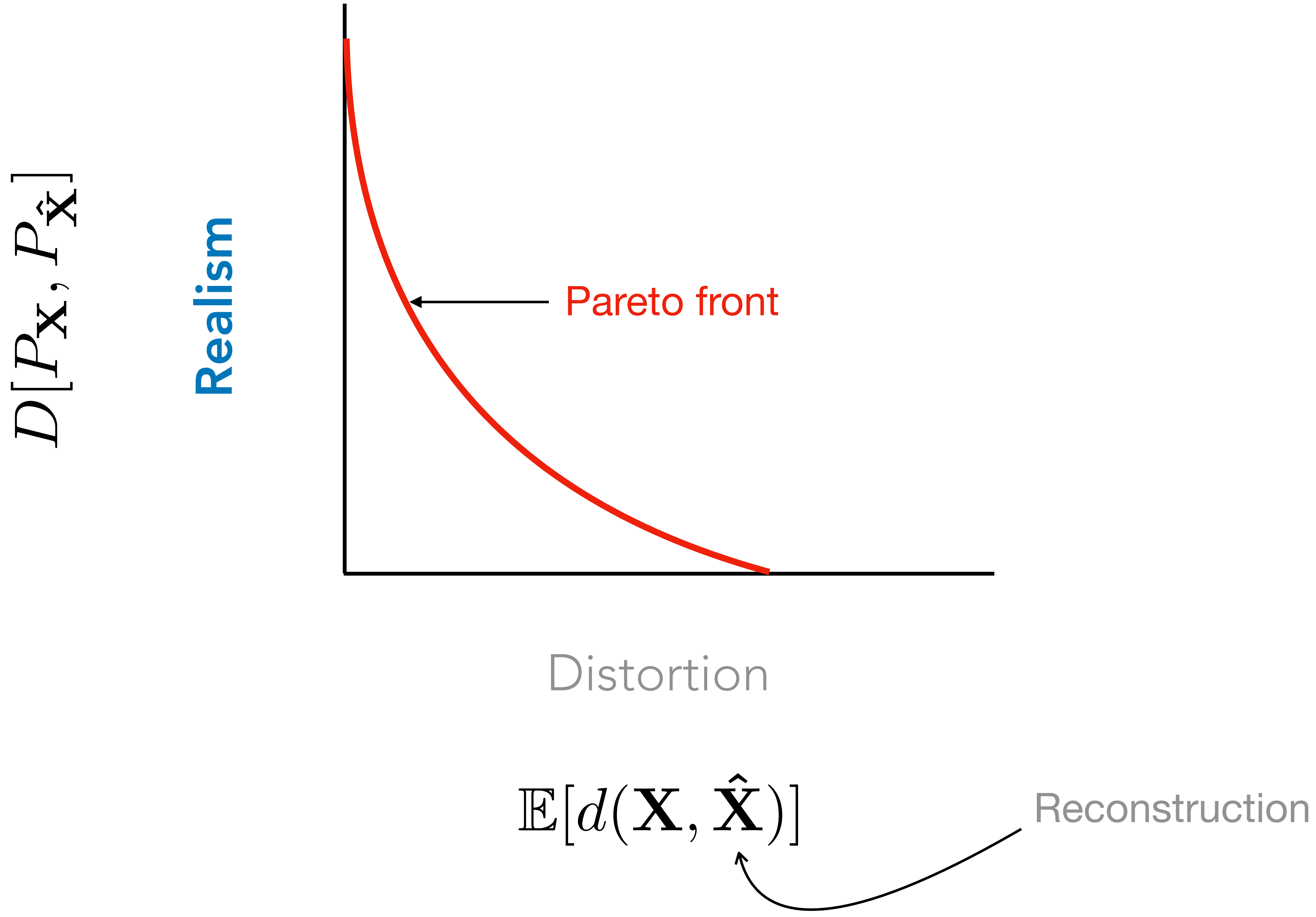
$$D[P_{\mathbf{x}}, P_{\hat{\mathbf{x}}}]$$

Image

Reconstructed image

E.g., M. Li et al., 2012

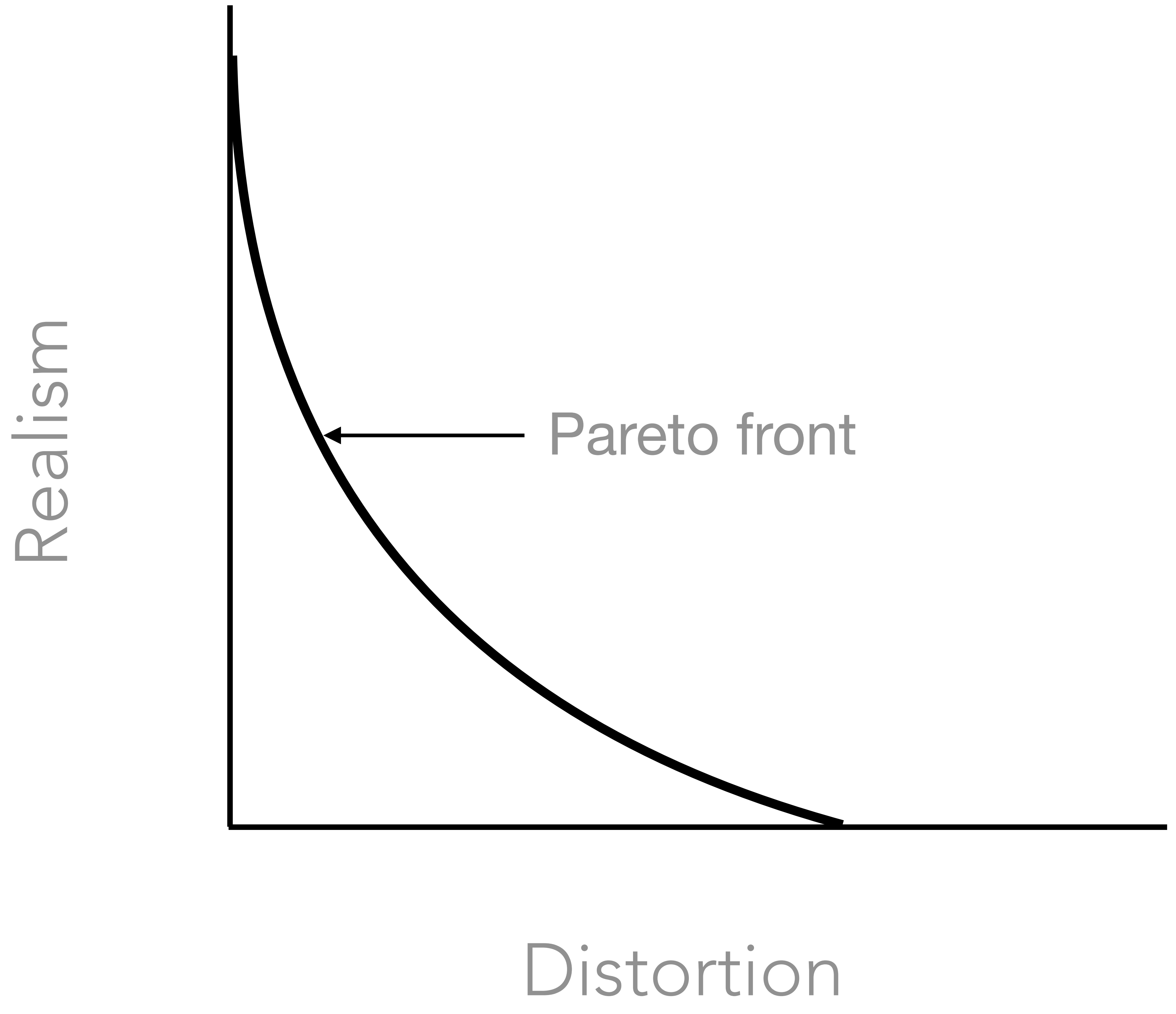
Realism-distortion trade-off



Blau & Michaeli (2018)

Realism-distortion trade-off

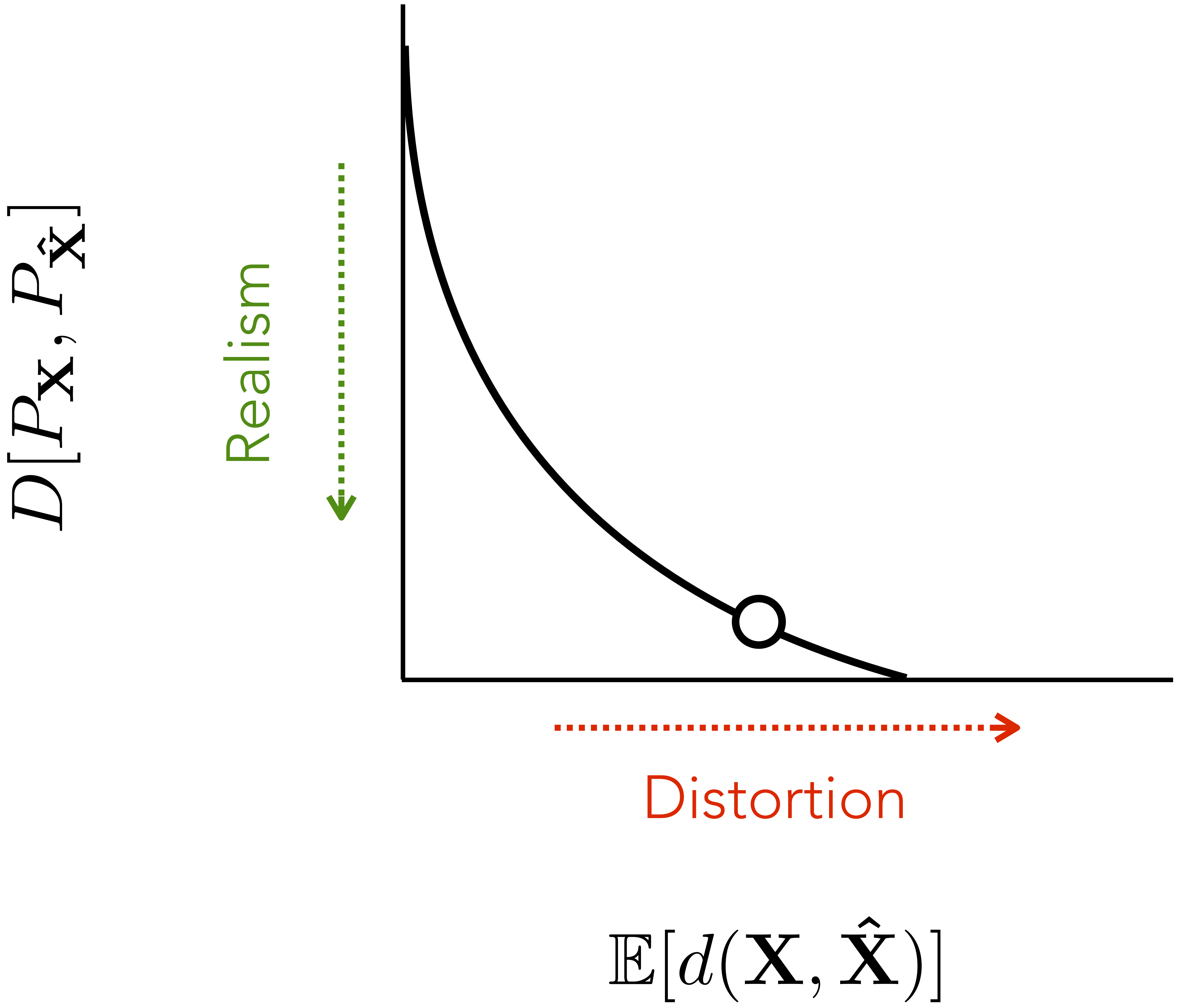
$$D[P_{\mathbf{X}}, P_{\hat{\mathbf{X}}}]$$



$$\mathbb{E}[d(\mathbf{X}, \hat{\mathbf{X}})]$$

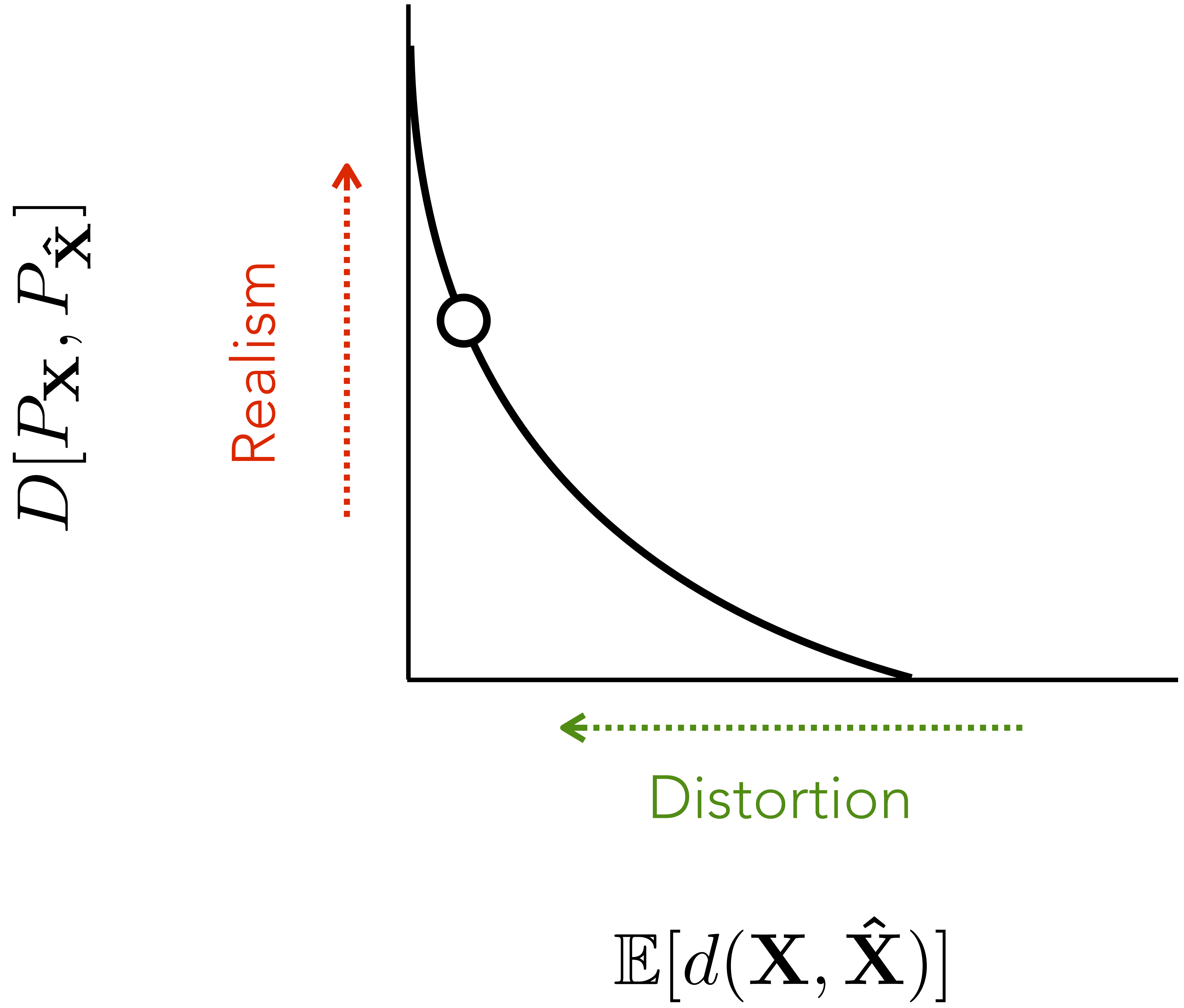
Blau & Michaeli (2018)

Realism-distortion trade-off



Blau & Michaeli (2018)

Realism-distortion trade-off



Blau & Michaeli (2018)

Perfect realism

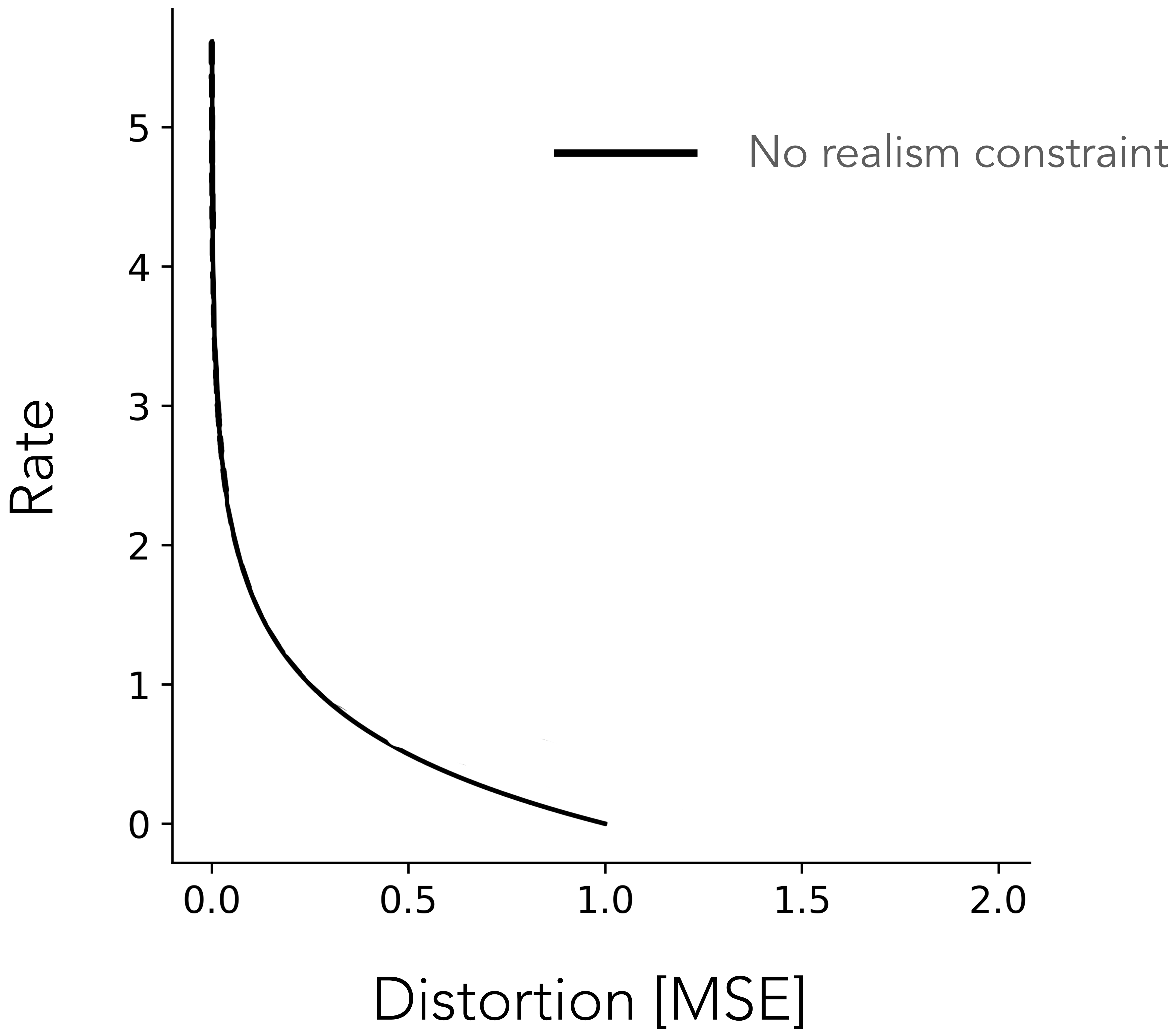
$$D[P_{\mathbf{X}}, P_{\hat{\mathbf{X}}}] = 0$$

Image

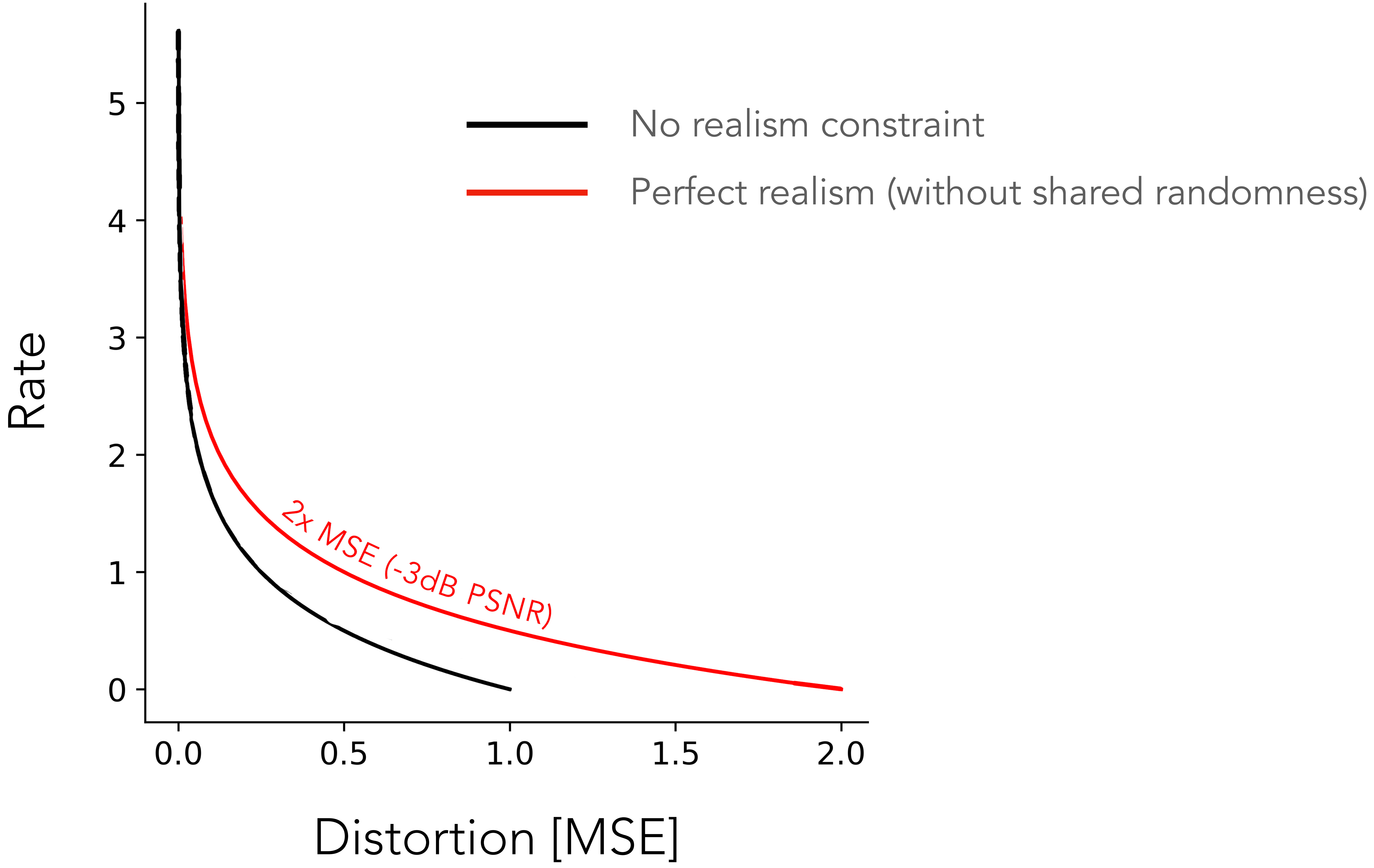
Reconstructed image

E.g., M. Li et al., 2012

The cost of perfect realism

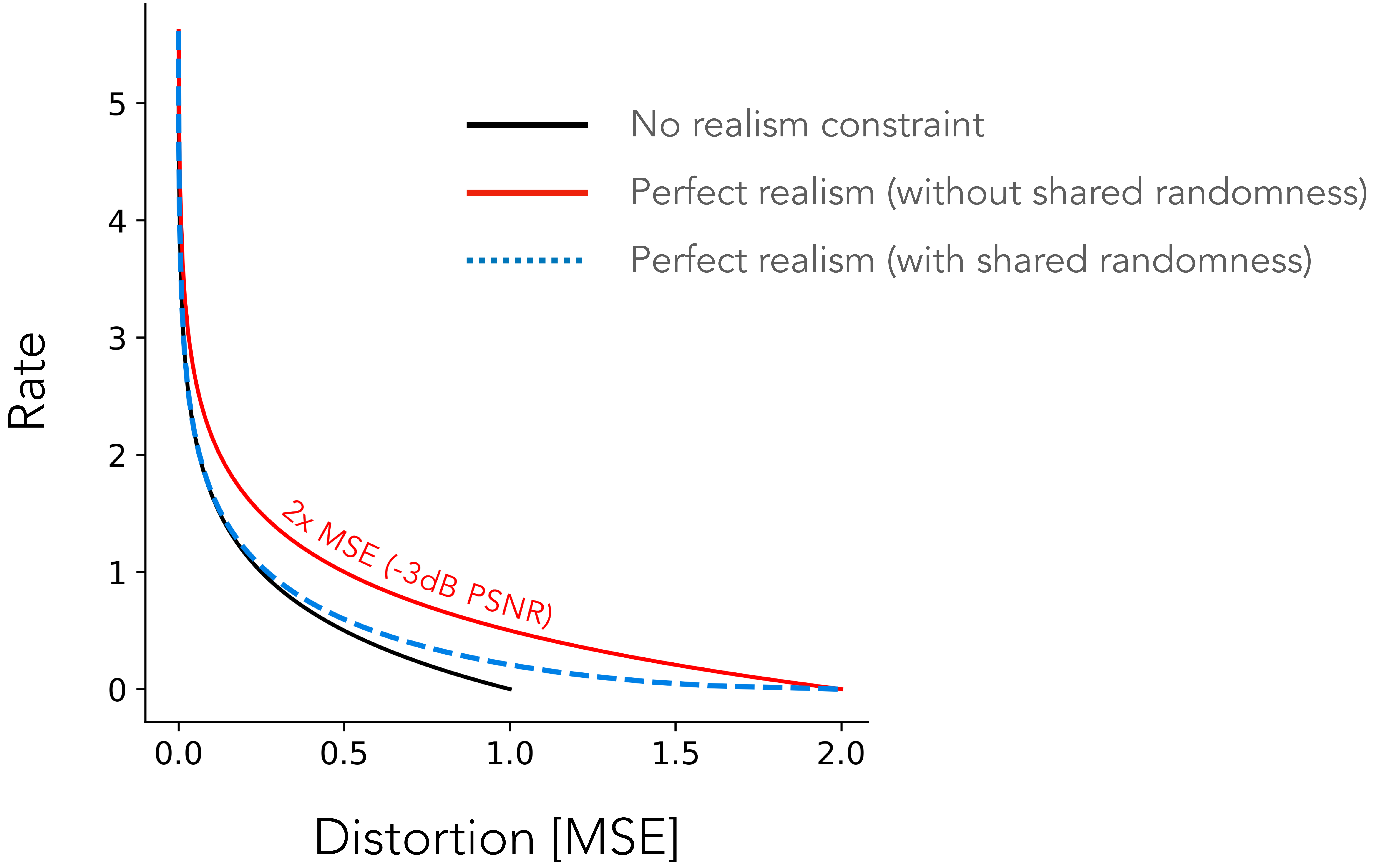


The cost of perfect realism



M. Li et al., 2012, Blau & Michaeli (2018; 2019), Theis & Agustsson (2021), Yan et al. (2021), Wagner (2022)

The cost of perfect realism



M. Li et al., 2012, Blau & Michaeli (2018; 2019), Theis & Agustsson (2021), Yan et al. (2021), Wagner (2022)

A mathematical theory of communication

systems; i.e., it must be possible to say of two systems represented by $P_1(x,y)$ and $P_2(x,y)$ that, according to our fidelity criterion, either (1) the first has higher fidelity, (2) the second has higher fidelity, or (3) they have equal fidelity. This means that a criterion of fidelity can be represented by a numerically valued function:

$$v\big(P(x,y)\big)$$

whose argument ranges over possible probability functions $P(x,y)$.

We will now show that under very general and reasonable assumptions the function $v\big(P(x,y)\big)$ can be written in a seemingly much more specialized form, namely as an average of a function $\rho(x,y)$ over the set of possible values of x and y:

Shannon, 1948

DIFFUSION I:

High-fidelity diffusion (HFD)

Approach

1) Train neural compressor with MSE

2) Train generative model conditioned on output

$\hat{\mathbf{X}}_{\mathrm{MSE}}$

$\hat{\mathbf{X}} \sim P_{\mathbf{X}|\hat{\mathbf{x}}_{\mathrm{MSE}}}$


ELIC (0.1674 bpp)


HFD (0.1674 bpp)

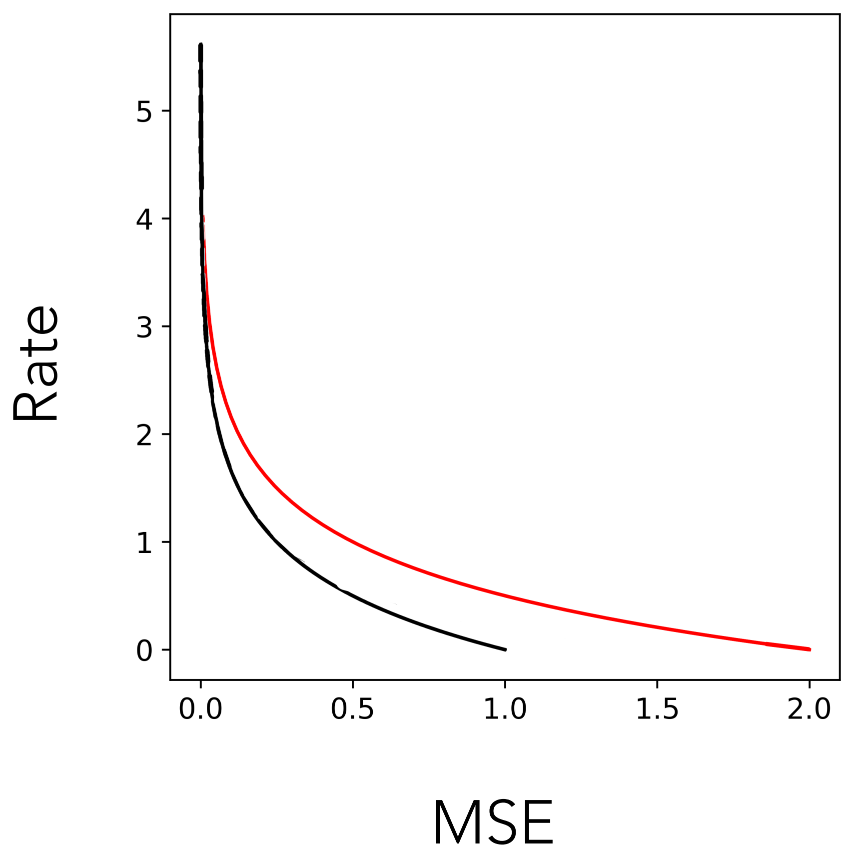
Justification

$$\mathbb{E}[\|\hat{\mathbf{X}} - \mathbf{X}\|^2] \qquad \mathbb{E}[\|\hat{\mathbf{X}}^{\mathrm{MSE}} - \mathbf{X}\|^2]$$

Output of generative model
(**realistic**)

Output of neural compressor
(**blurry**)

Justification



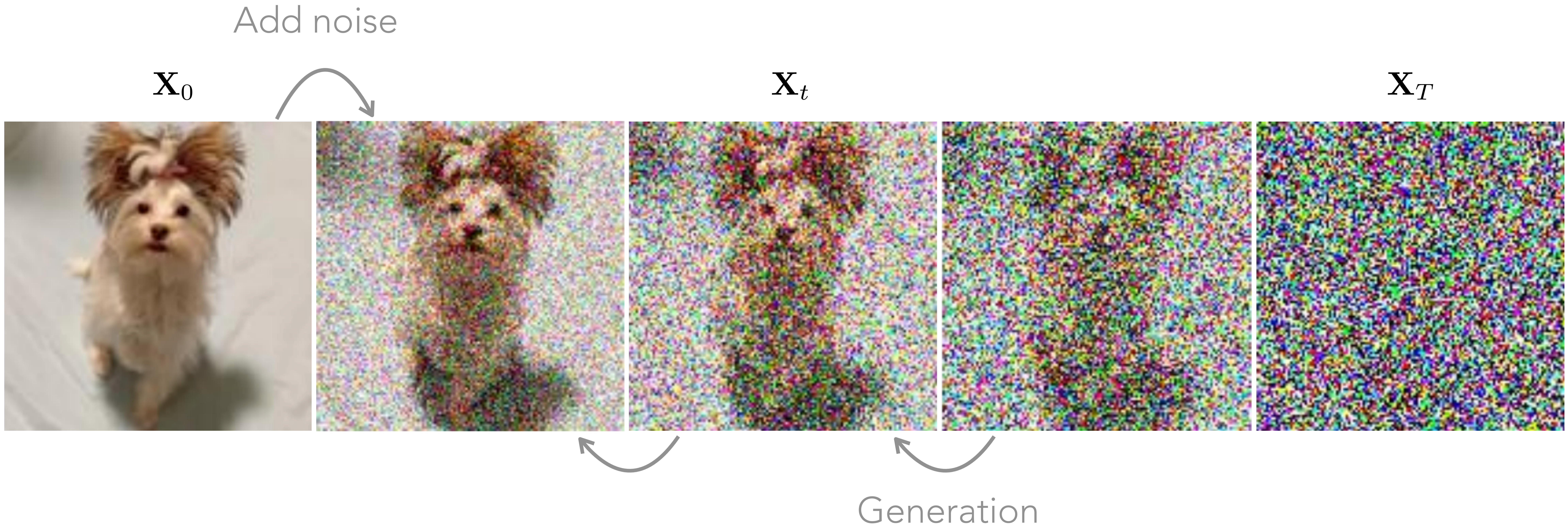
$$\mathbb{E}[\|\hat{\mathbf{X}} - \mathbf{X}\|^2] \leq 2\,\mathbb{E}[\|\hat{\mathbf{X}}^{\mathrm{MSE}} - \mathbf{X}\|^2]$$

Output of generative model
(**realistic**)

Output of neural compressor
(**blurry**)

Blau & Michaeli (2018)

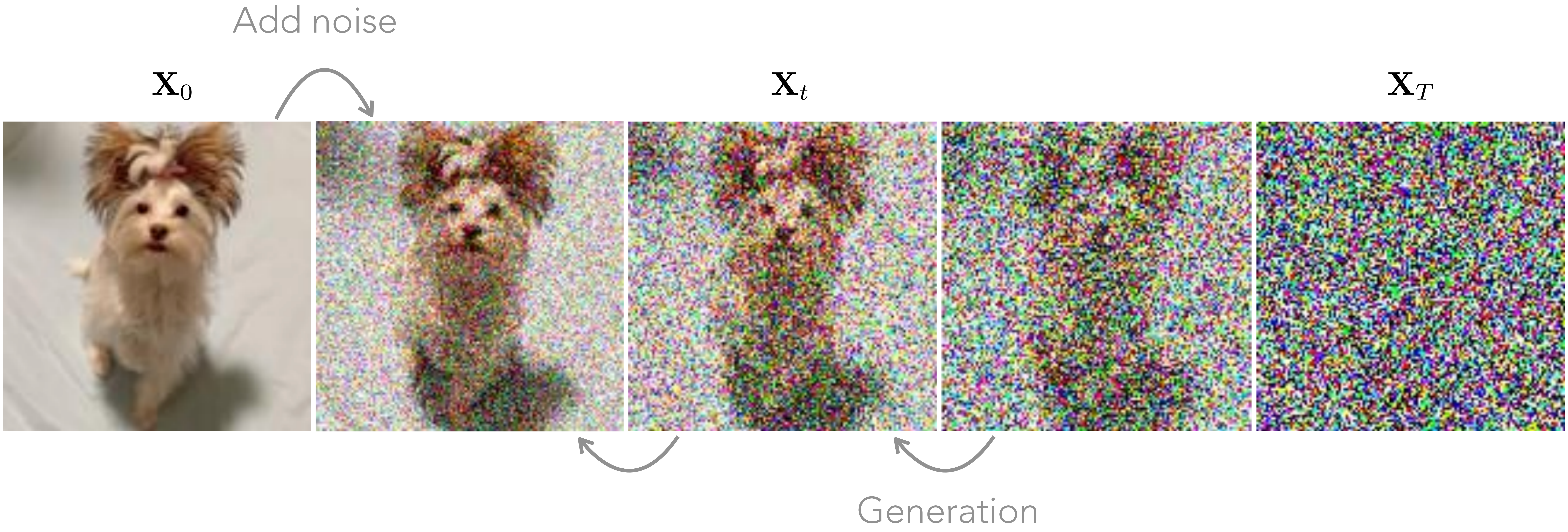
Diffusion



💡 $P_{\mathbf{X}_{t-\delta}|\mathbf{X}_t}$ is approximately Gaussian (e.g., Feller, 1949; Anderson, 1982).

Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021)

Diffusion



💡 Optimize $\mathbb{E}[\|\mathbf{X}_{t-\delta} - m_{\boldsymbol{\theta}}(\mathbf{X}_t)\|^2]$ so that $m_{\boldsymbol{\theta}}(\mathbf{X}_t) \approx \mathbb{E}[\mathbf{X}_{t-\delta} \mid \mathbf{X}_t]$.

Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021)

Overview

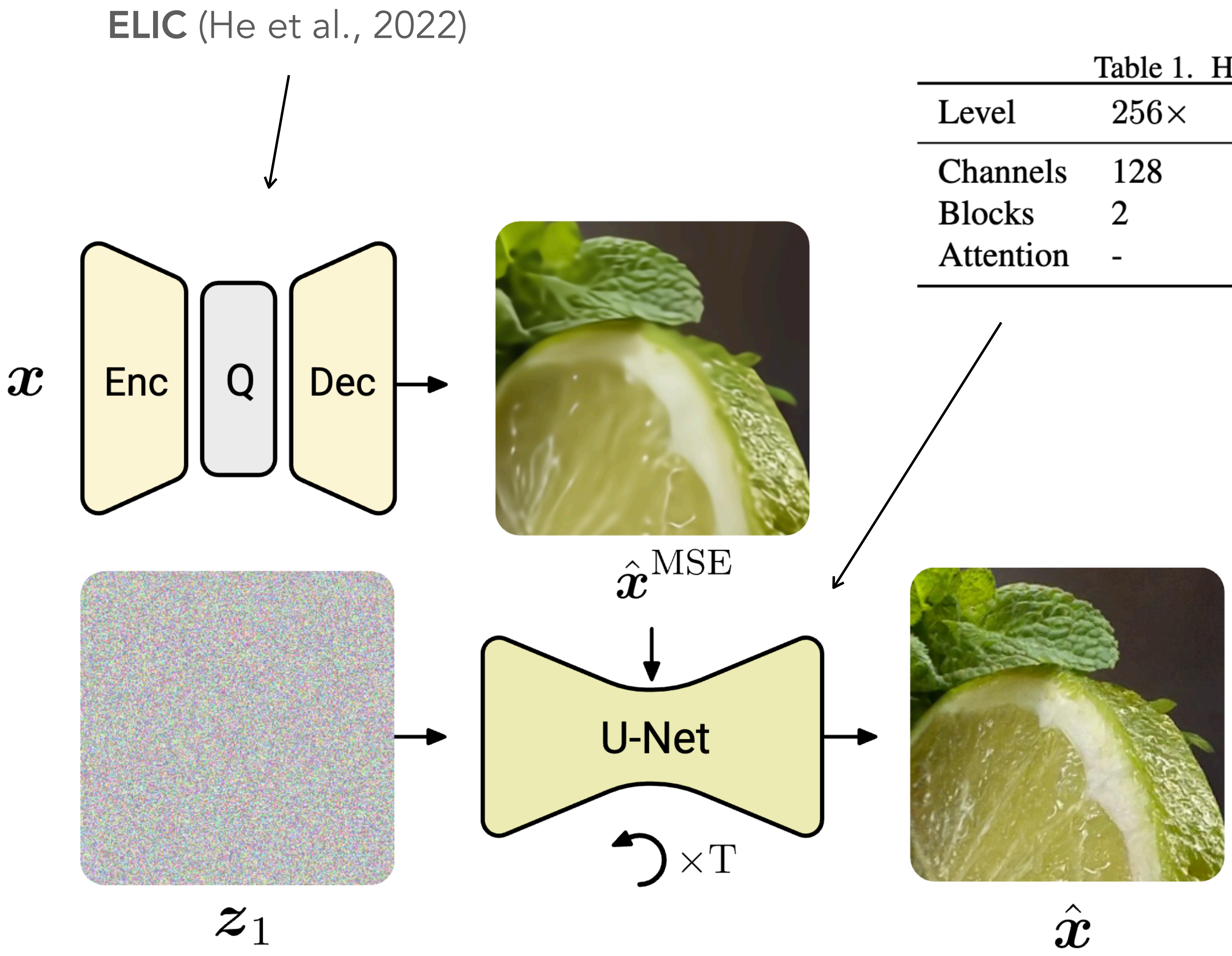
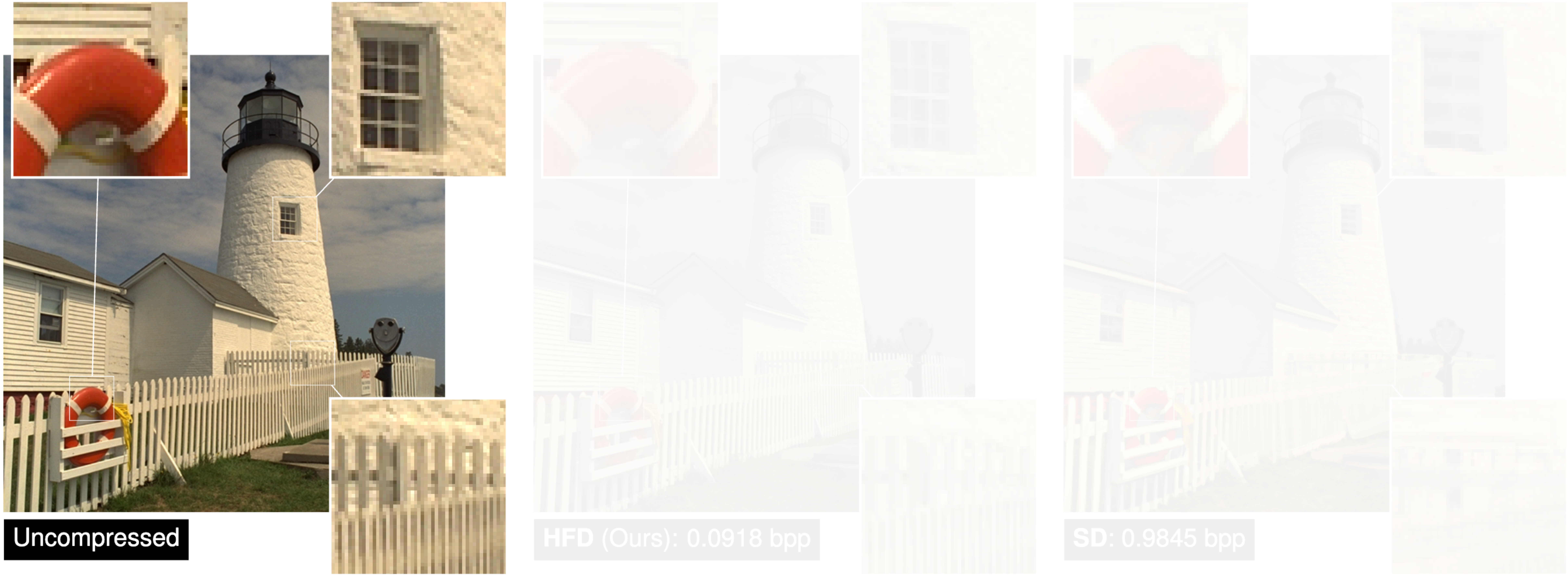


Table 1. HFD U-Net architecture

Level	256×	128×	64×	32×	16×
Channels	128	128	256	256	1024
Blocks	2	2	2	2	16
Attention	-	-	-	-	✓

JPEG + StableDiffusion



Rombach et al. (2022), Hoogeboom et al. (2023)


SD: 0.9845 bpp

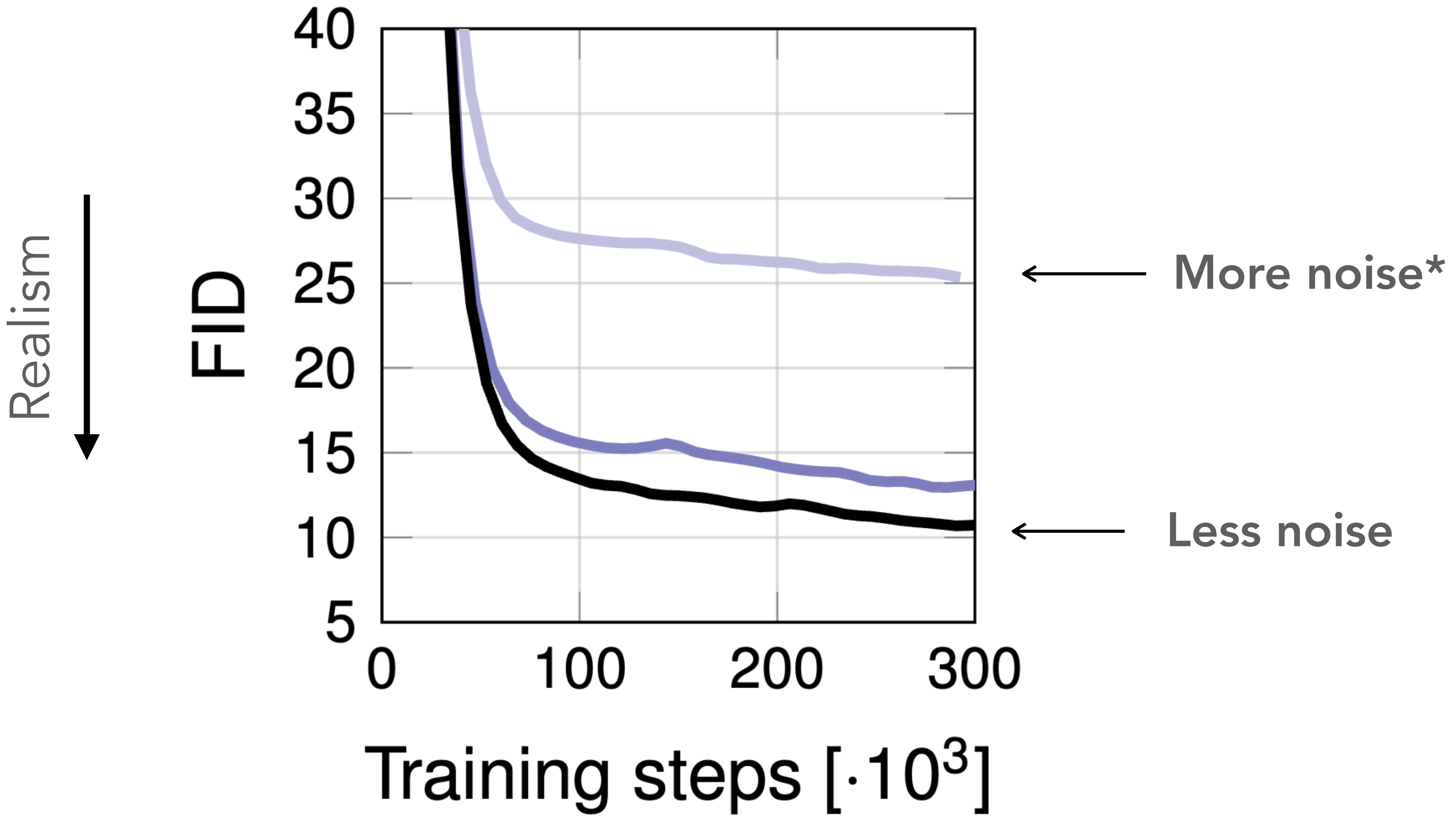

HFD (Ours): 0.0918 bpp

"A lighthouse in Maine behind a white fence with a red life buoy hanging on it."

Sharper, but less consistent

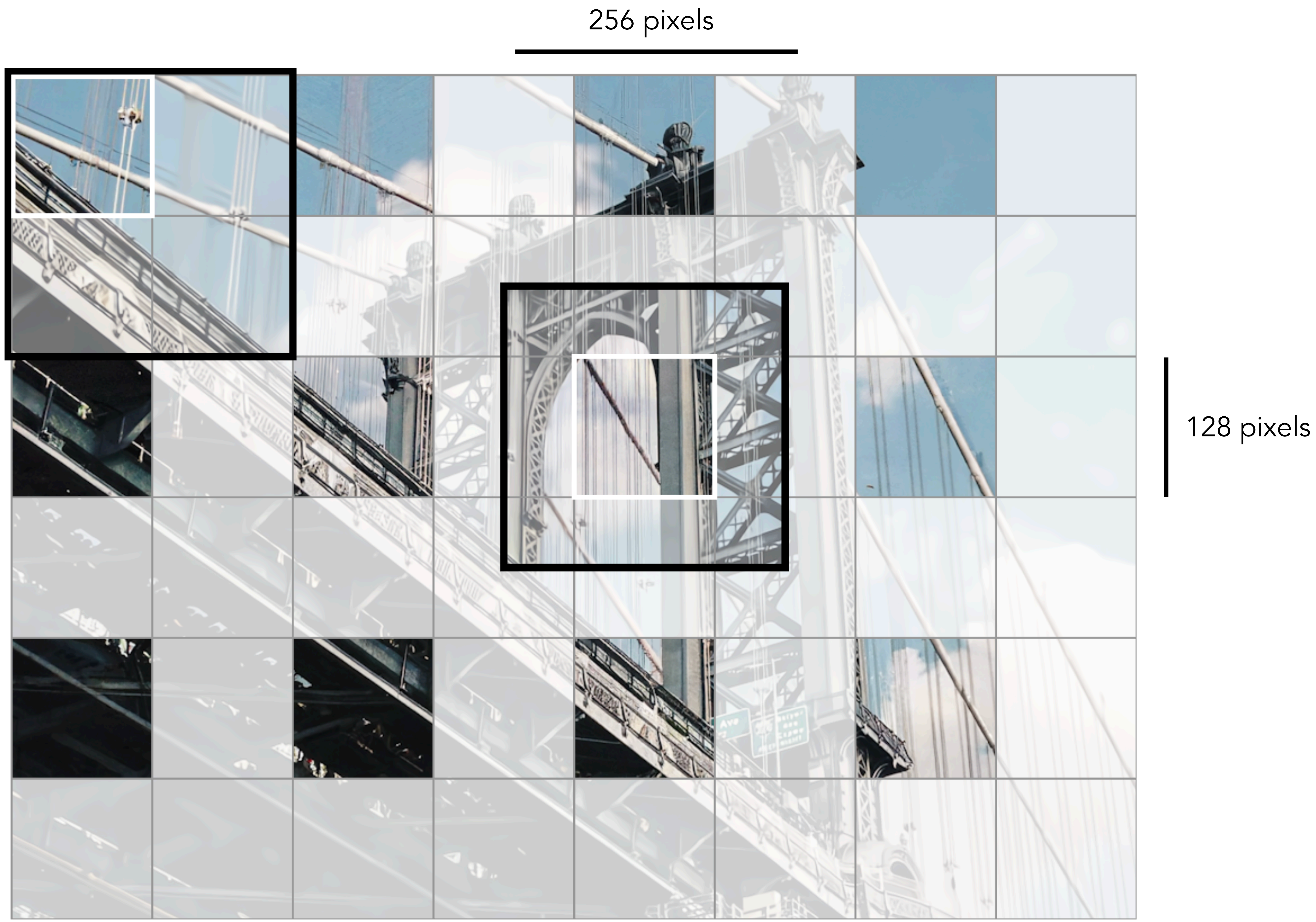
Rombach et al. (2022), Hoogeboom et al. (2023)

Noise schedule



* Recommended for text-to-image models (Chen et al., 2023; Hoogeboom et al., 2023a)

Patch-wise generation



HFD (Ours): 0.2639 bpp

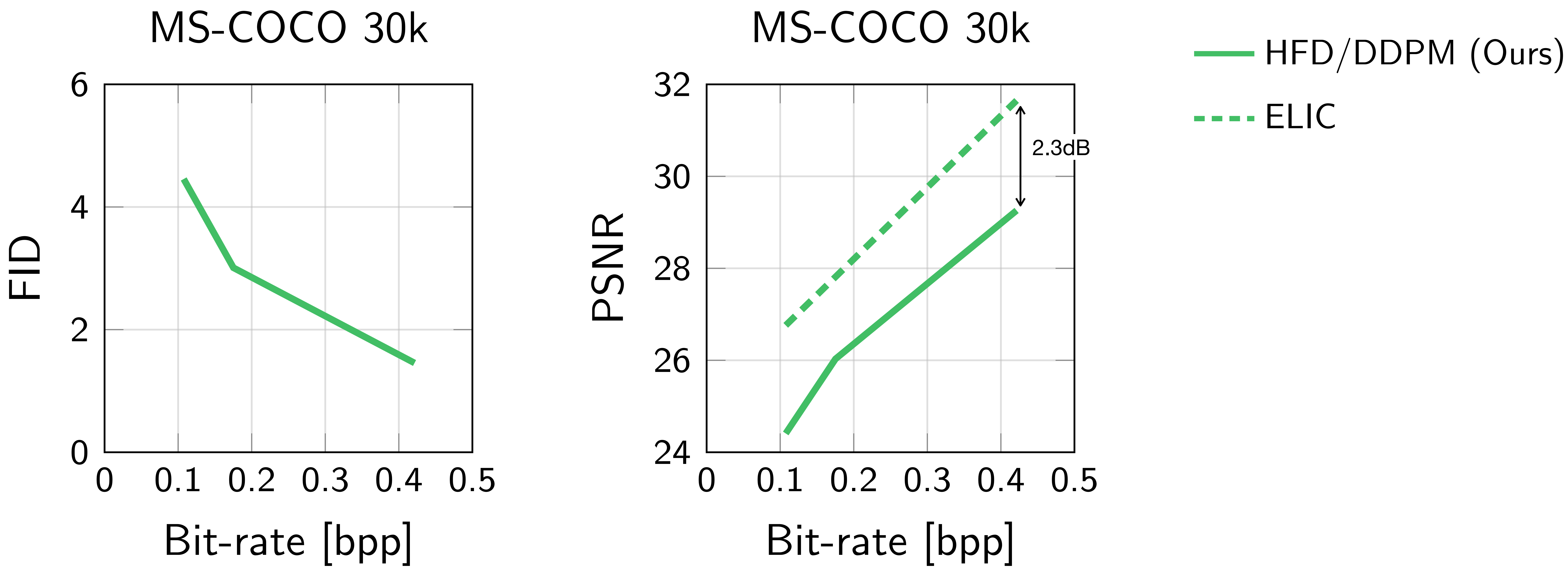
Yang & Mandt (2023): 0.2971 bpp

End-to-end trained

Yang & Mandt (2023): 0.2971 bpp

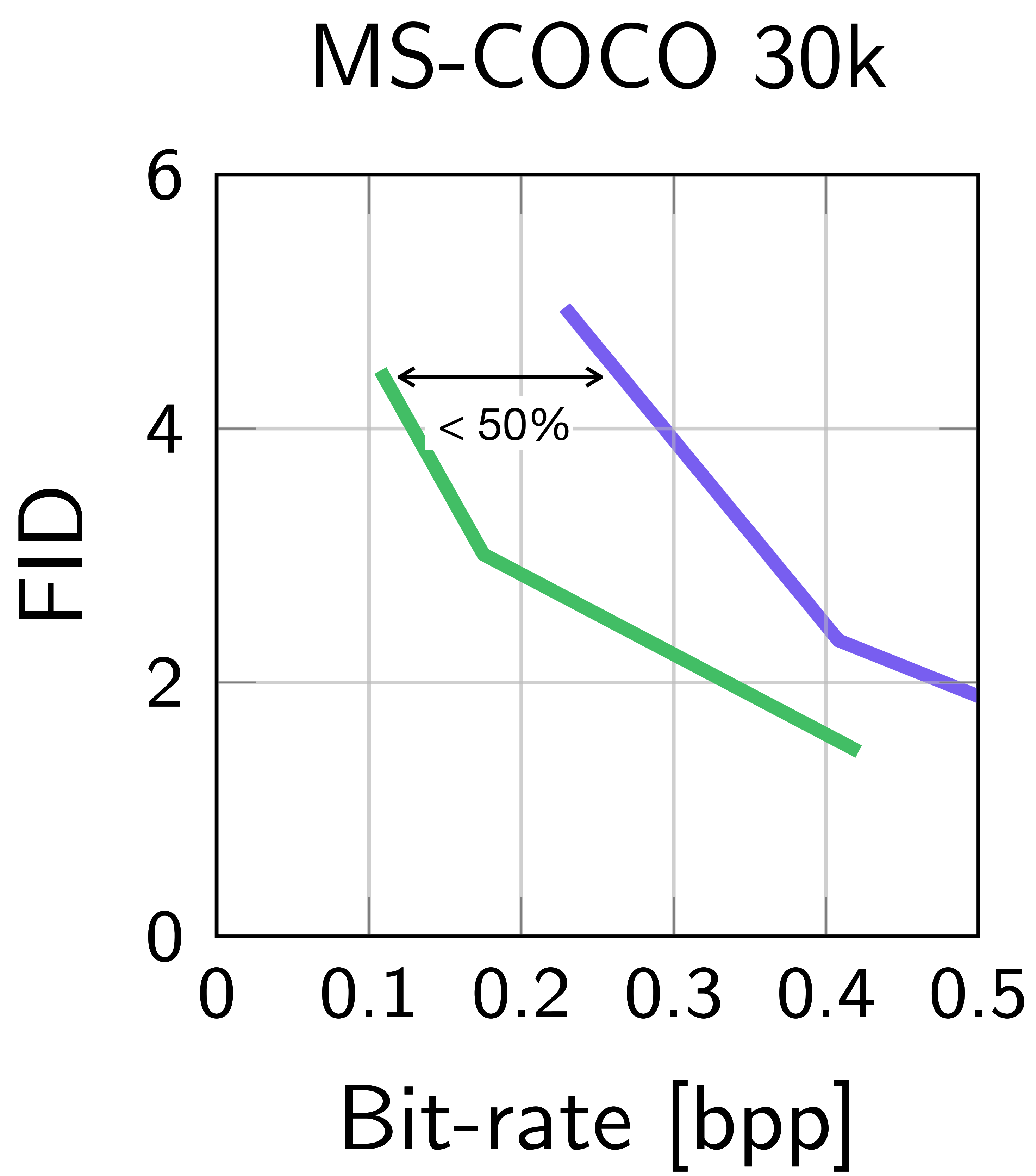
HFD (Ours): 0.2639 bpp

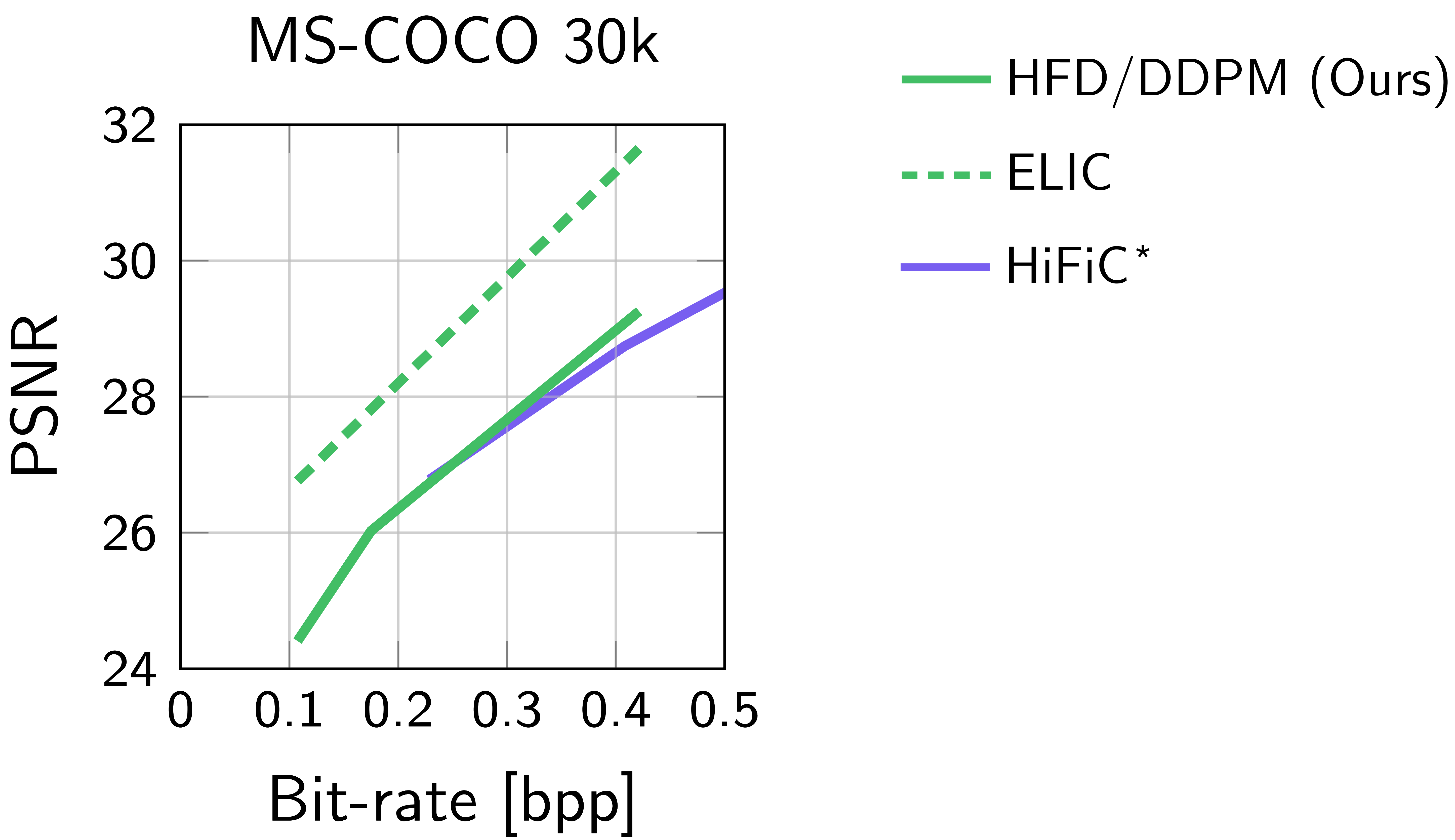
Realism Distortion



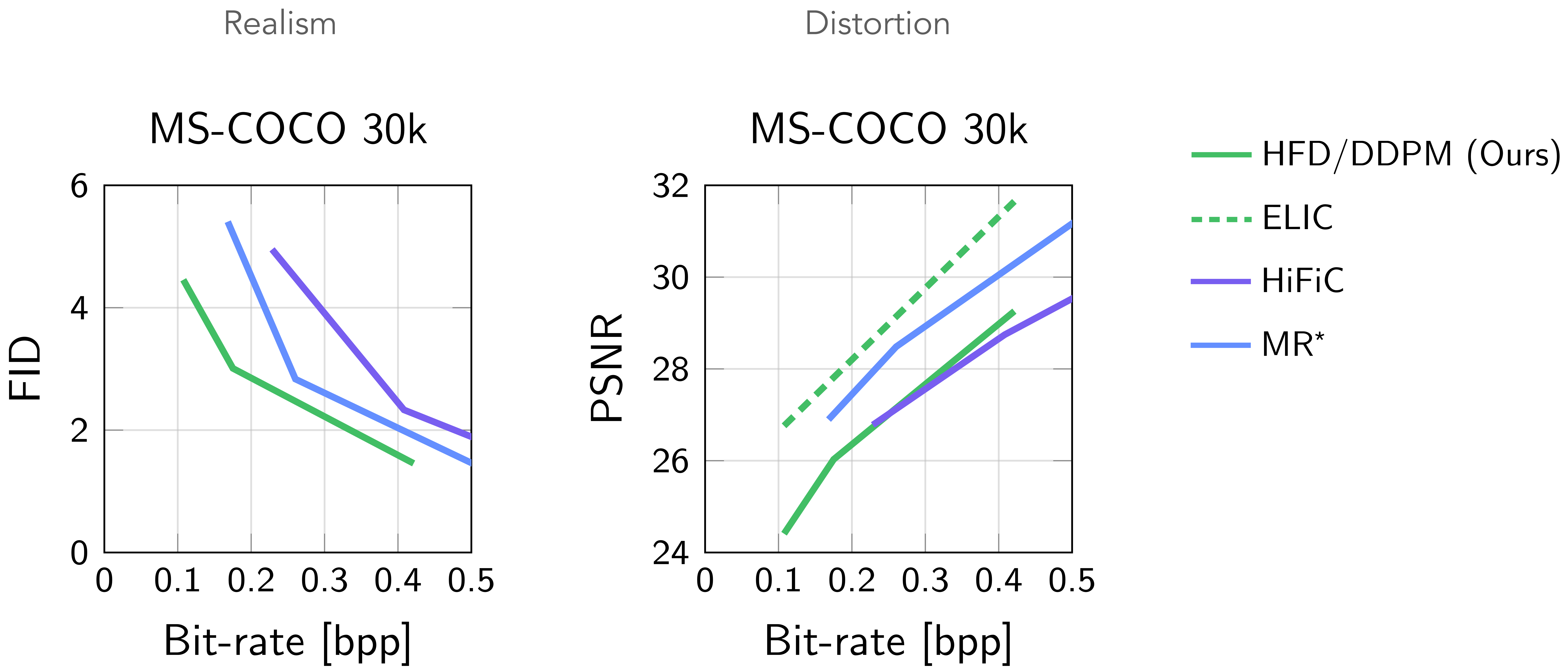
Realism

Distortion





* Mentzer et al., 2020



* Agustsson et al., 2023

HFD (Ours): 0.0538 (44.2%)

PQ-MIM [8]: 0.124 (100%)

PQ-MIM [8]: 0.124 (100%) (similar FID)

(El-Nouby et al., 2023)

HFD (Ours): 0.0538 (44.2%)

(similar FID)

(Hoogeboom et al., 2023)

DIFFUSION II:

DiffC

DiffC

HFD (transform coding)

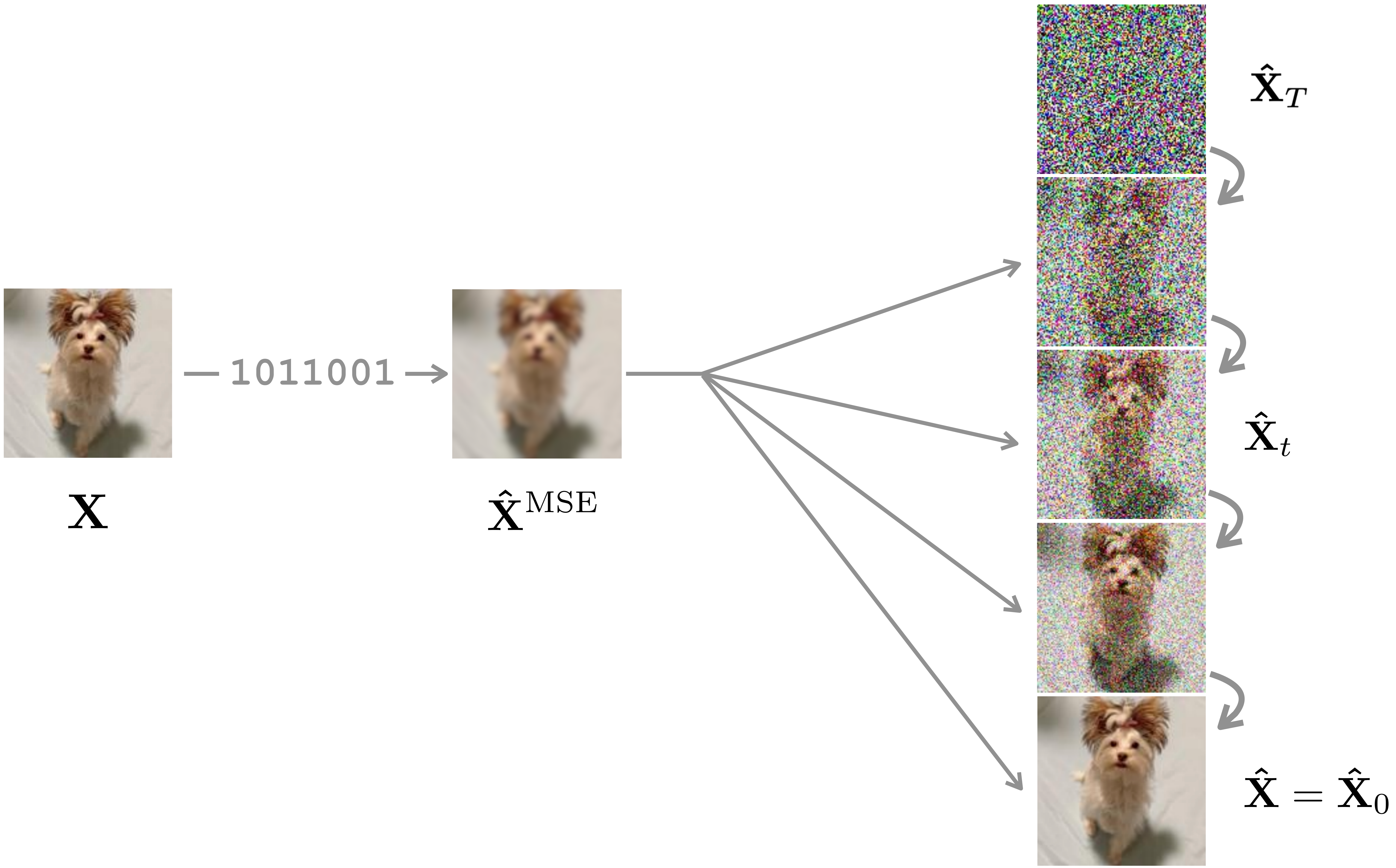
- Analysis transform
- Synthesis transform
- Conditional diffusion model
- Entropy model
- Quantization

x N (potentially different models for different bit-rates)

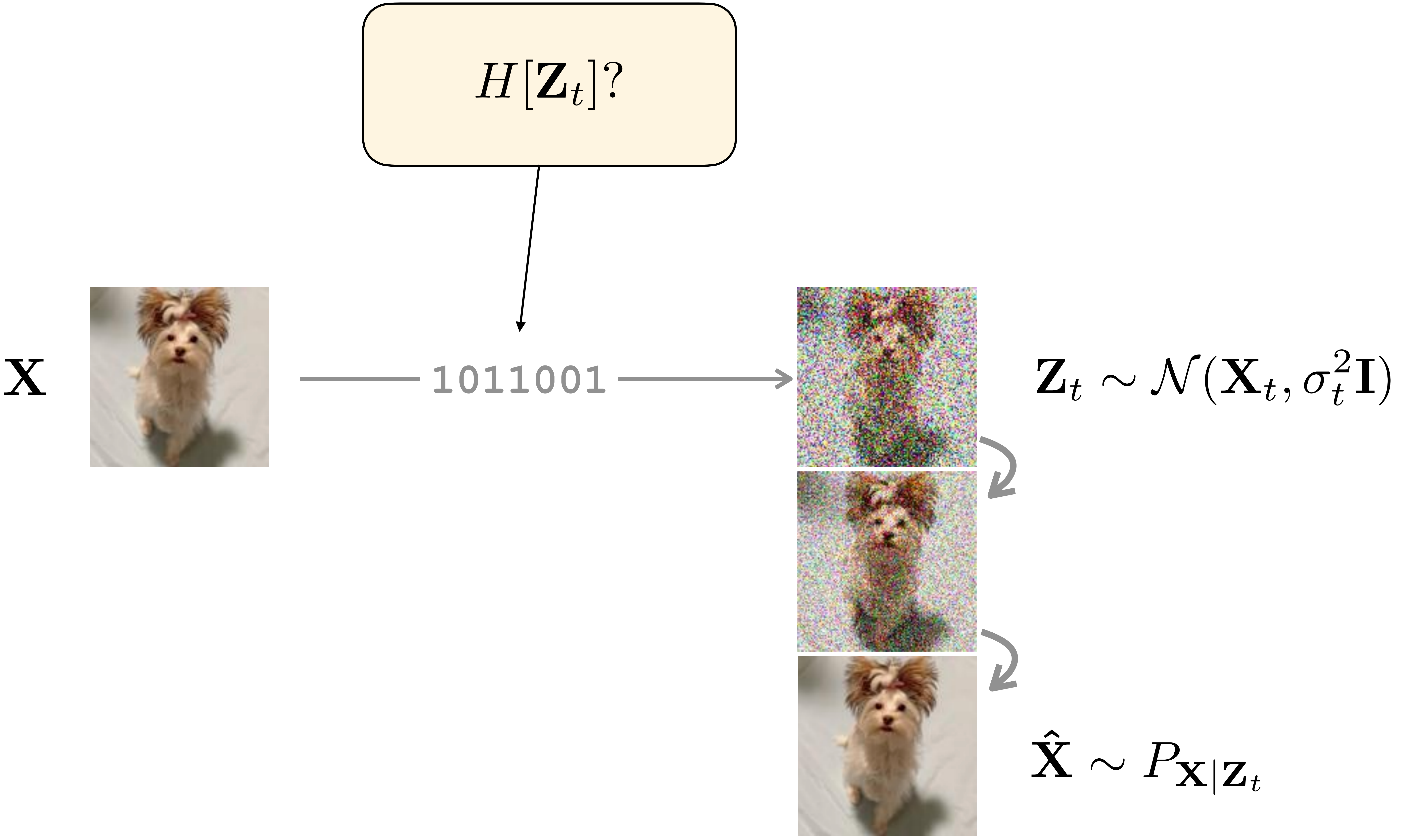
DiffC

- A single <u>unconditional diffusion model</u> for all bit-rates
- *Reverse channel coding*

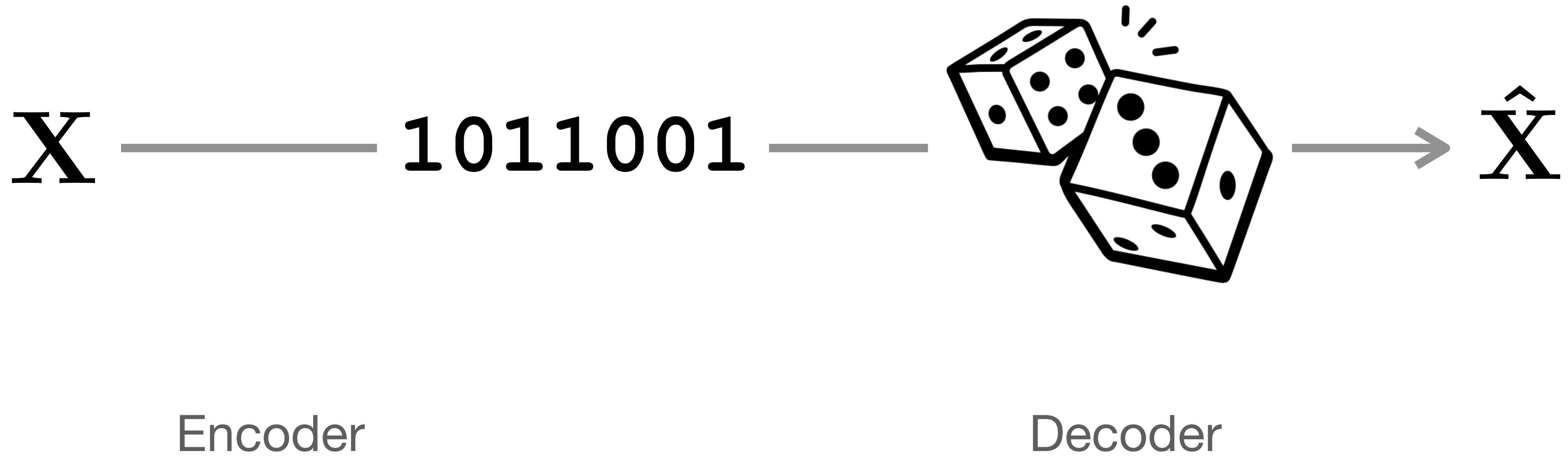
HFD


1011001
$\mathbf{X}$
$\hat{\mathbf{X}}^{\mathrm{MSE}}$
$\hat{\mathbf{X}}_T$
$\hat{\mathbf{X}}_t$
$\hat{\mathbf{X}} = \hat{\mathbf{X}}_0$

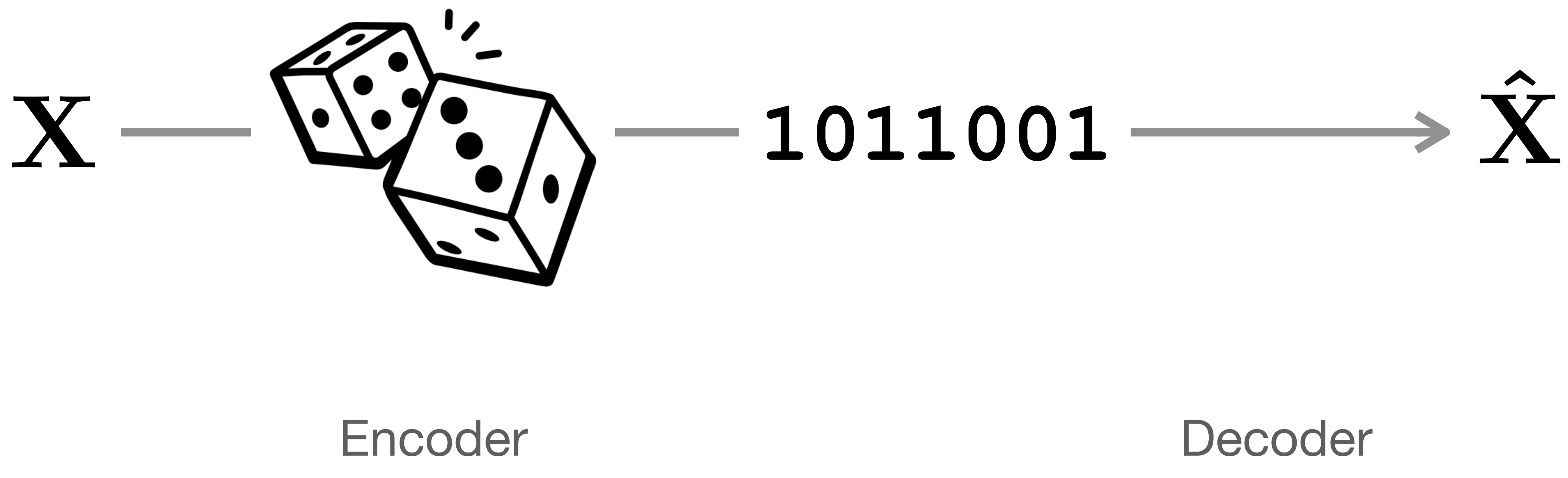
DiffC


$H[\mathbf{Z}_t]$?
$\mathbf{X}$
1011001
$\mathbf{Z}_t \sim \mathcal{N}(\mathbf{X}_t, \sigma_t^2 \mathbf{I})$
$\hat{\mathbf{X}} \sim P_{\mathbf{X}|\mathbf{Z}_t}$

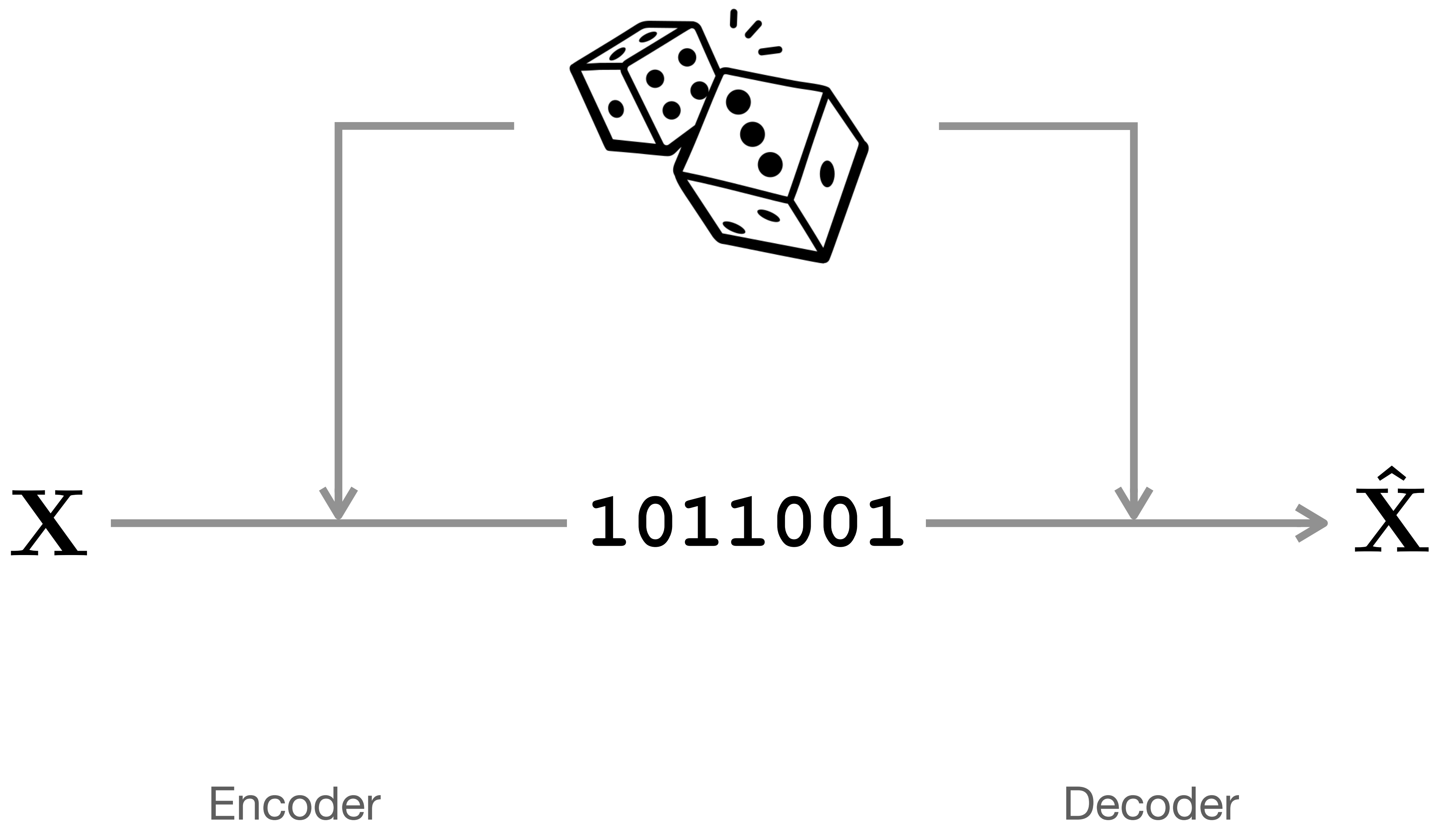
Decoder randomness



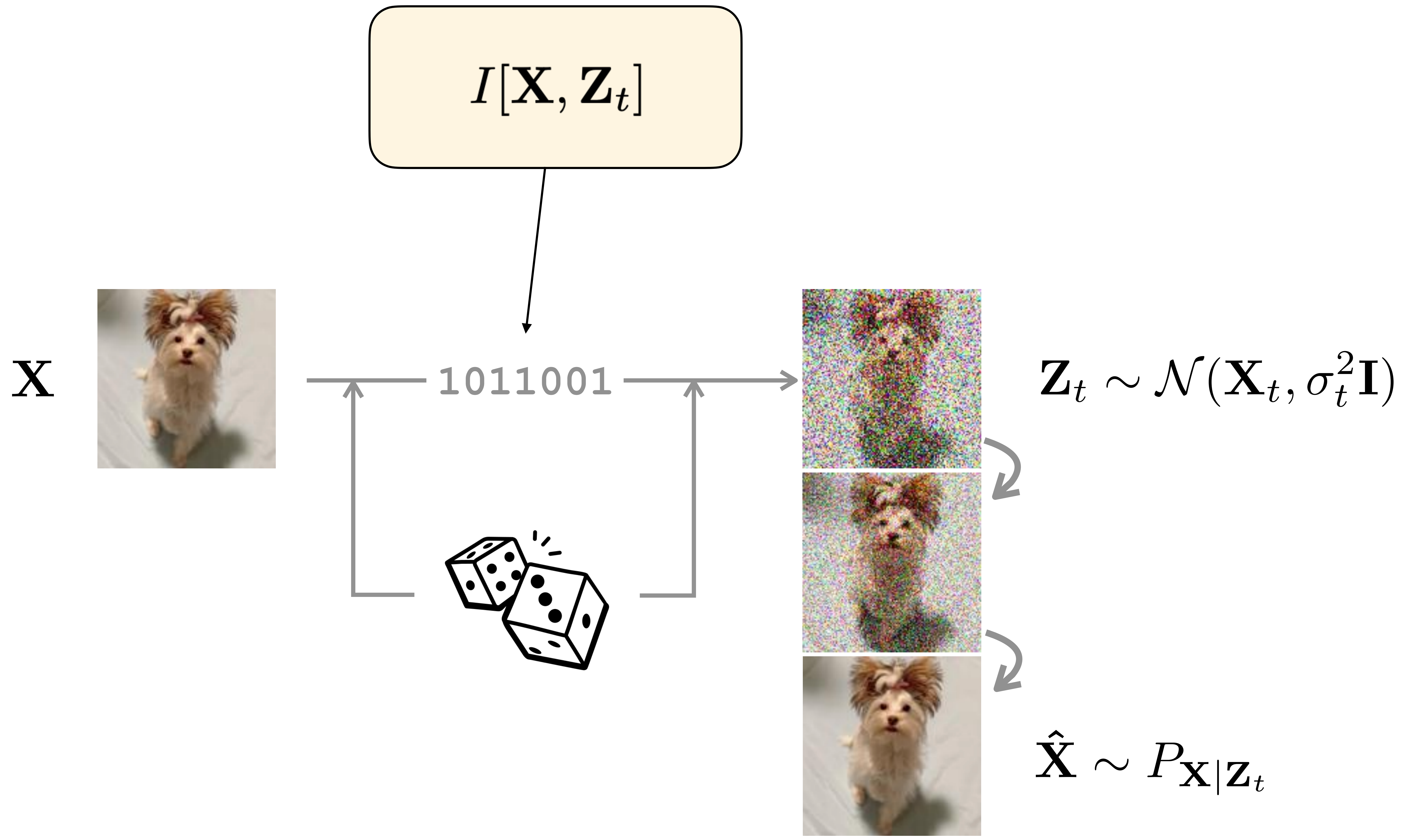
Encoder randomness



Shared randomness

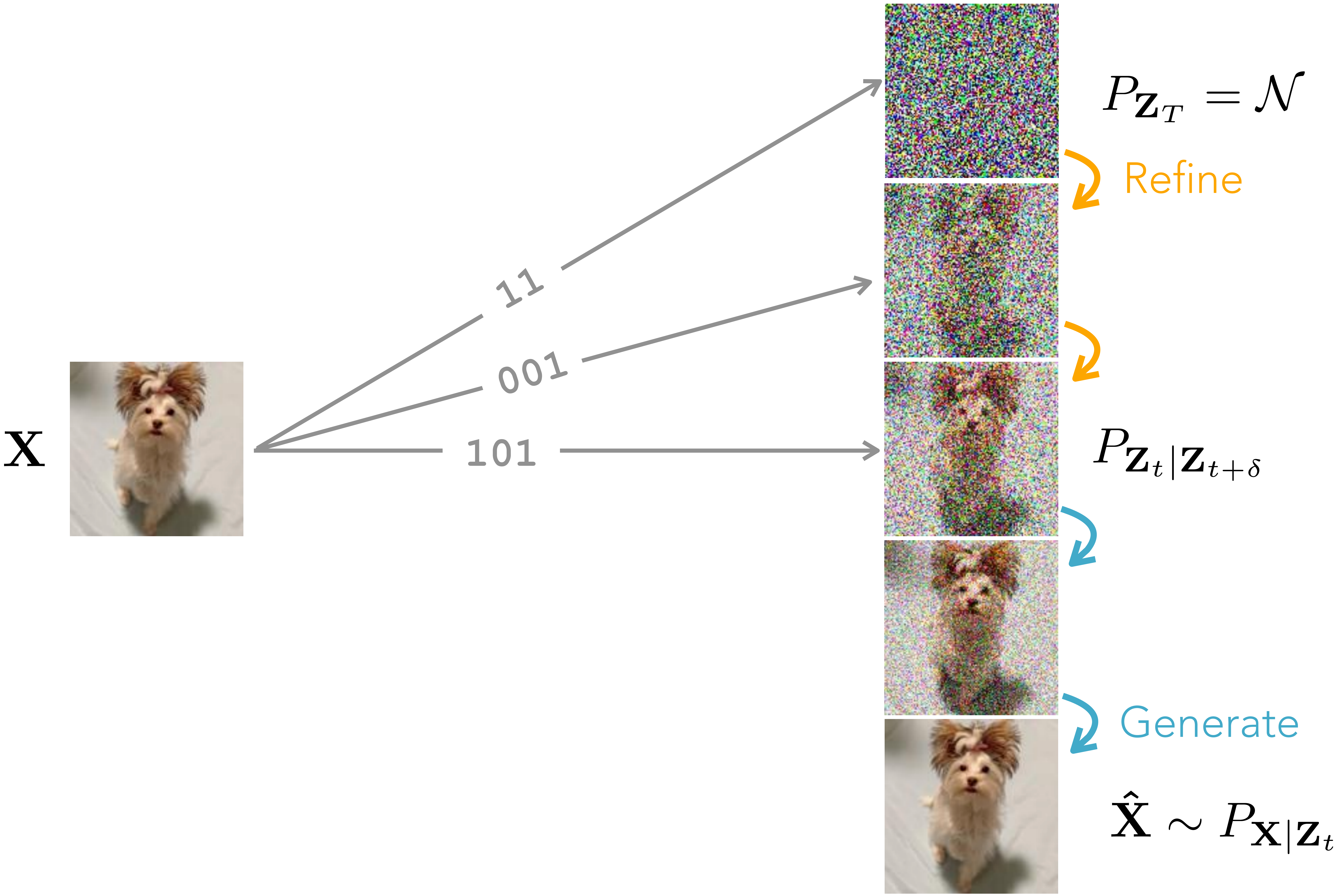


DiffC



$I[\mathbf{X}, \mathbf{Z}_t]$
$\mathbf{X}$
1011001
$\mathbf{Z}_t \sim \mathcal{N}(\mathbf{X}_t, \sigma_t^2 \mathbf{I})$
$\hat{\mathbf{X}} \sim P_{\mathbf{X}|\mathbf{Z}_t}$

DiffC



$\mathbf{X}$
11
001
101
$P_{\mathbf{Z}_T} = \mathcal{N}$
Refine
$P_{\mathbf{Z}_t|\mathbf{Z}_{t+\delta}}$
Generate
$\hat{\mathbf{X}} \sim P_{\mathbf{X}|\mathbf{Z}_t}$

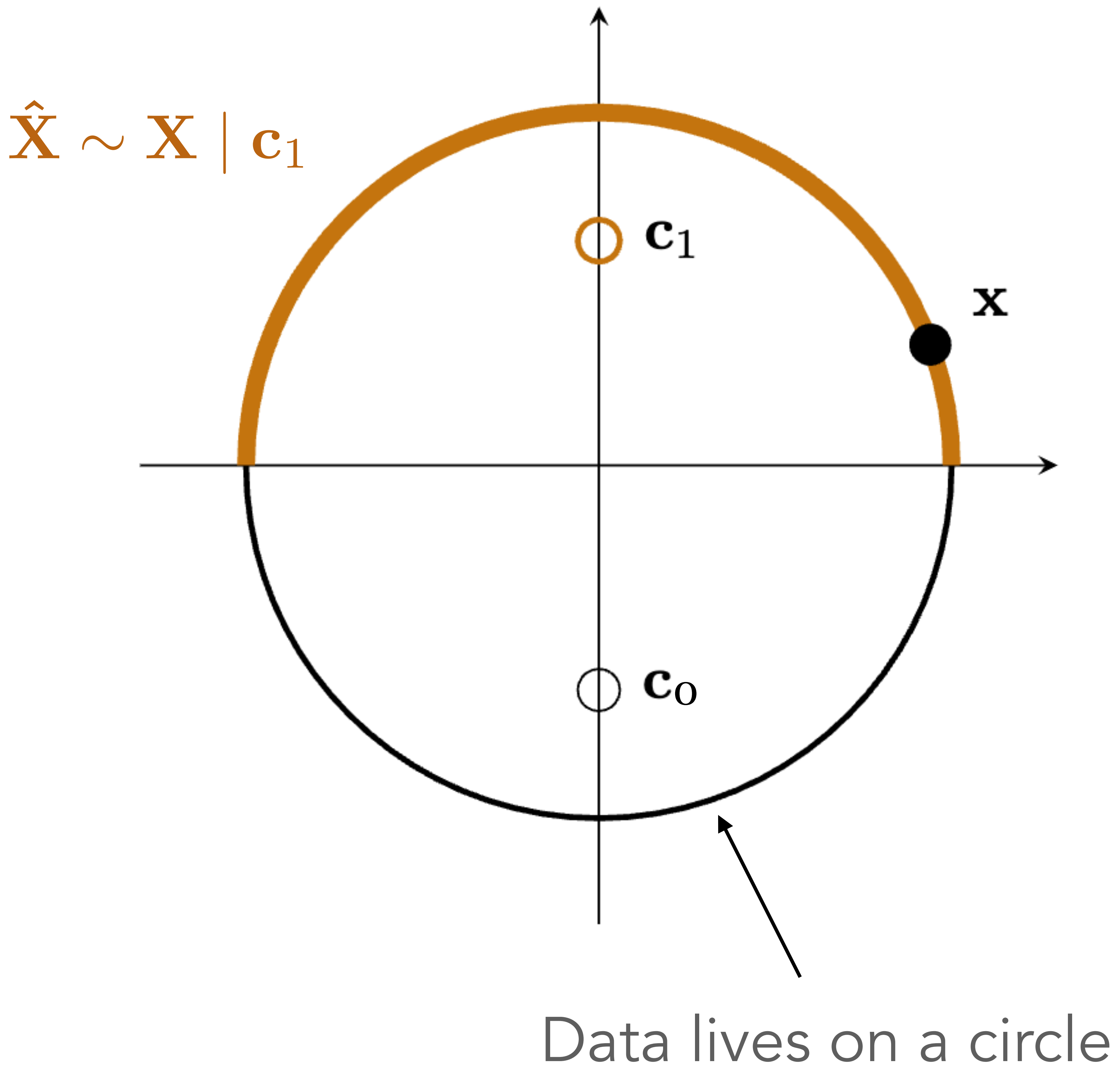
Universal quantization

Encoder Decoder

$$Z = \underbrace{\lfloor Y - U \rceil}_{K} + U \sim Y + U' \qquad U, U' \sim \mathrm{Uniform}([-1/2, 1/2))$$

Roberts (1962), Ziv (1985), Zamir & Feder (1995)

Toy example

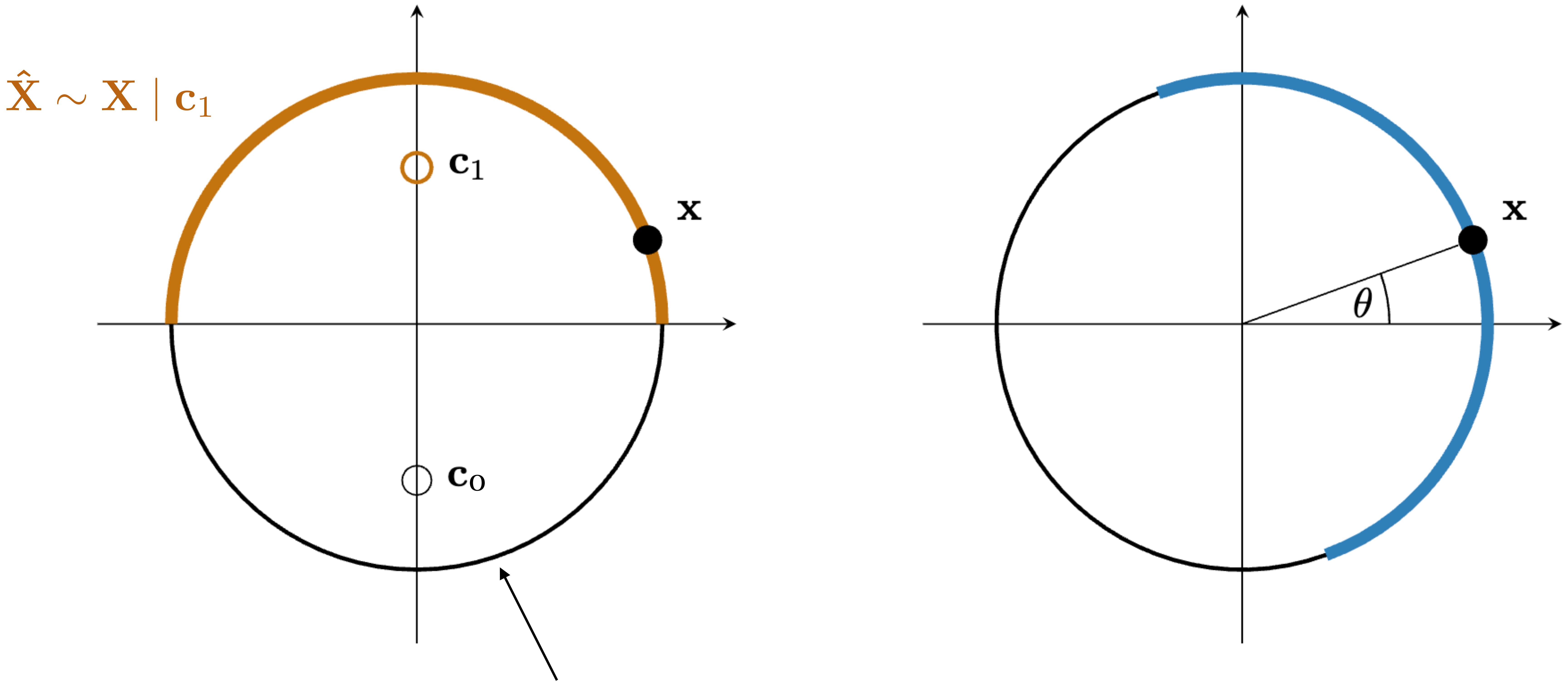


Theis & Agustsson (2020)

Toy example

$$K = \left\lfloor \frac{\Theta}{\pi} - U \right\rceil \mod 2$$

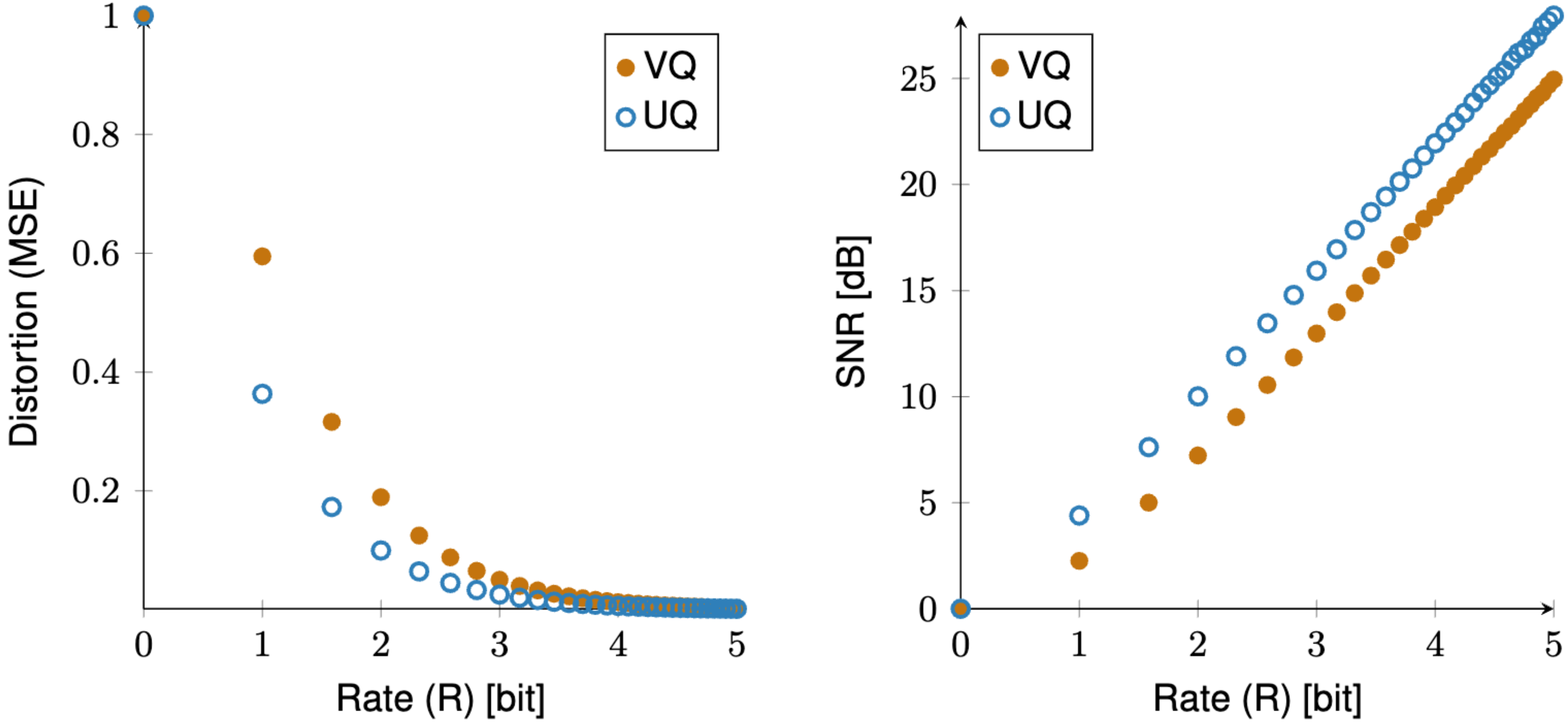
$$\hat{\Theta} = \pi(K + U)$$



Data lives on a circle

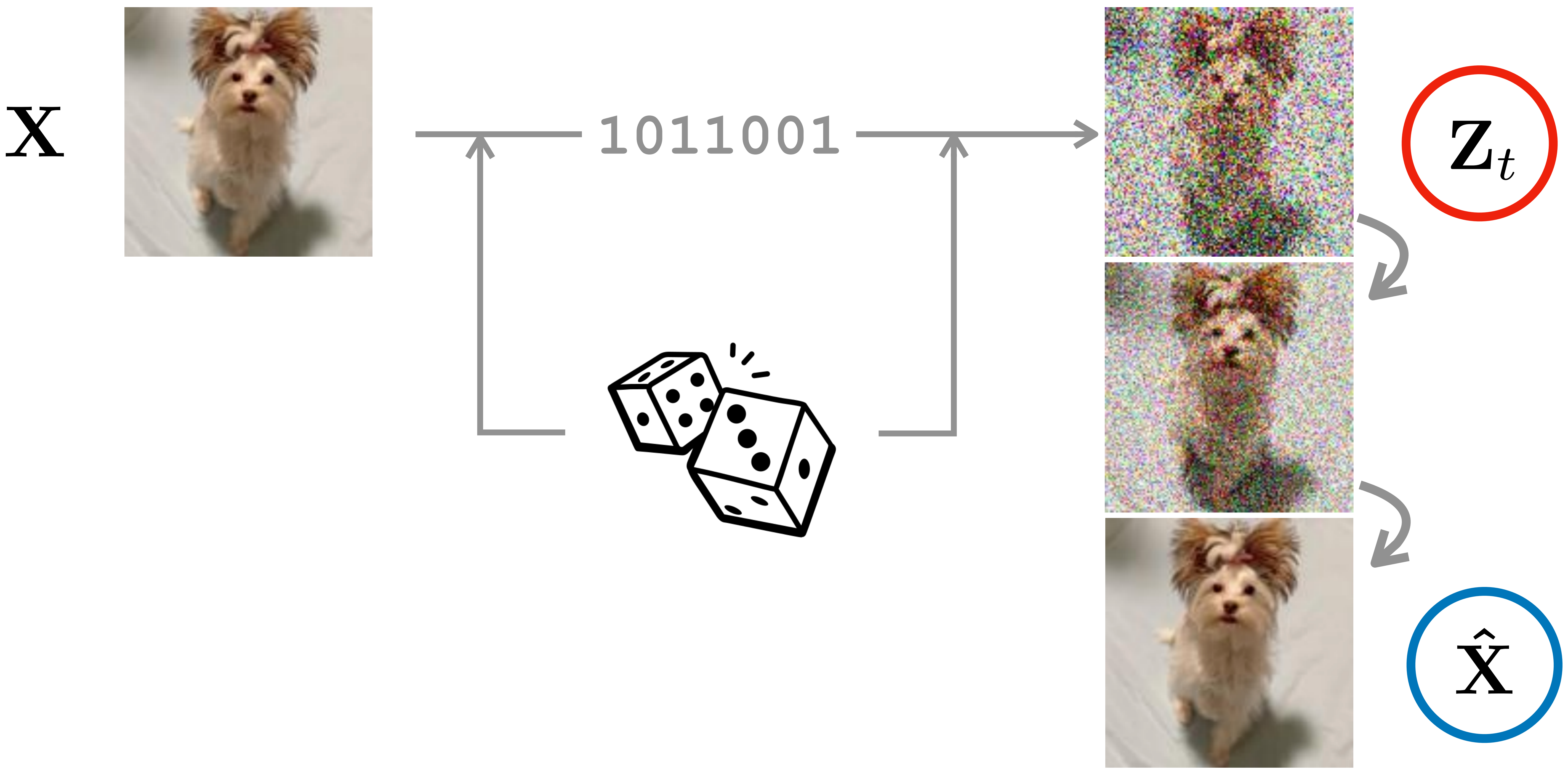
Theis & Agustsson (2020)

Toy example

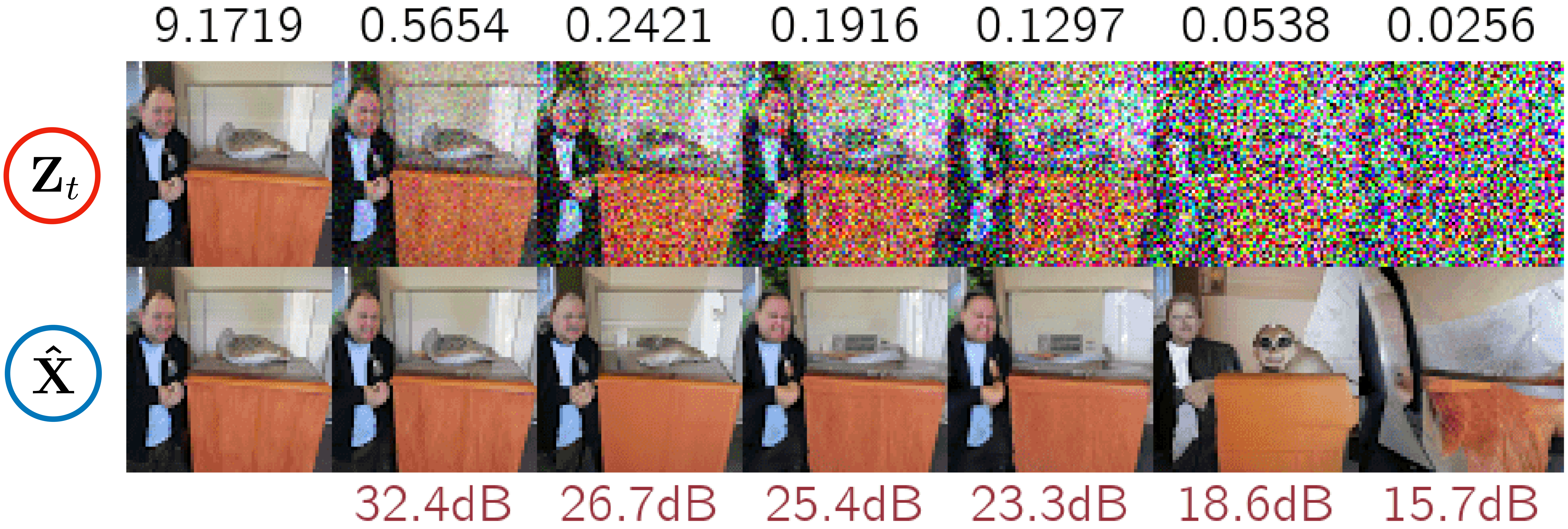


Theis & Agustsson (2020)

DiffC

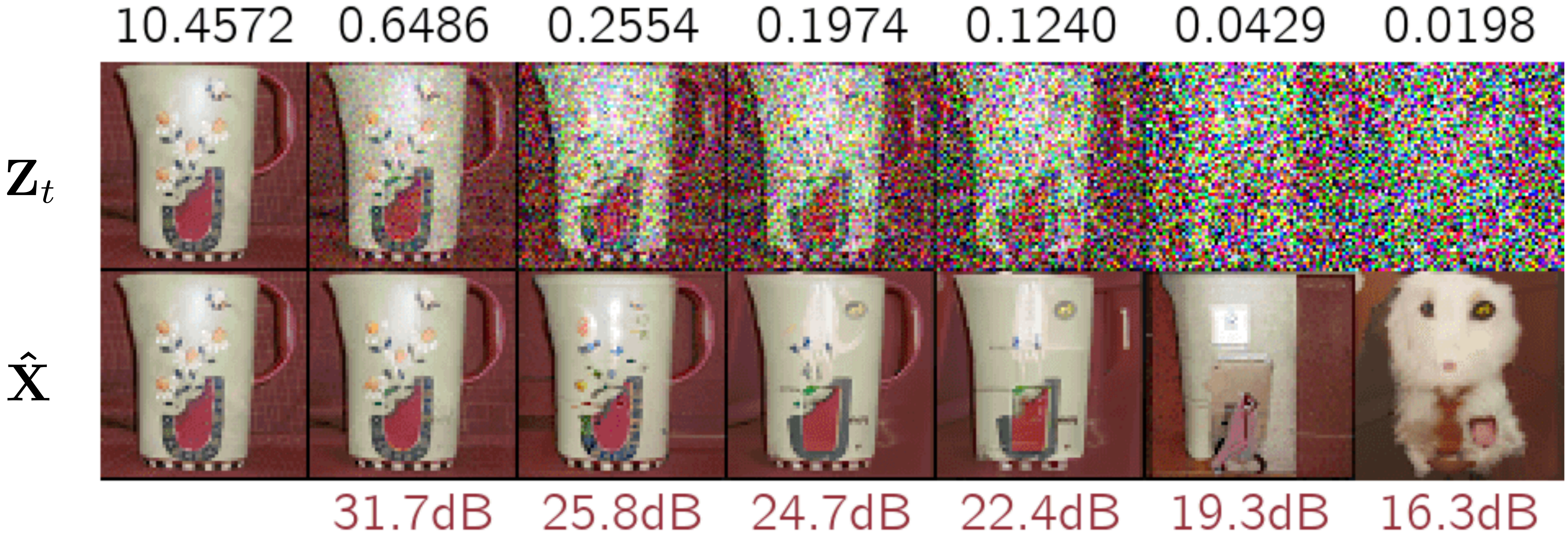


DiffC



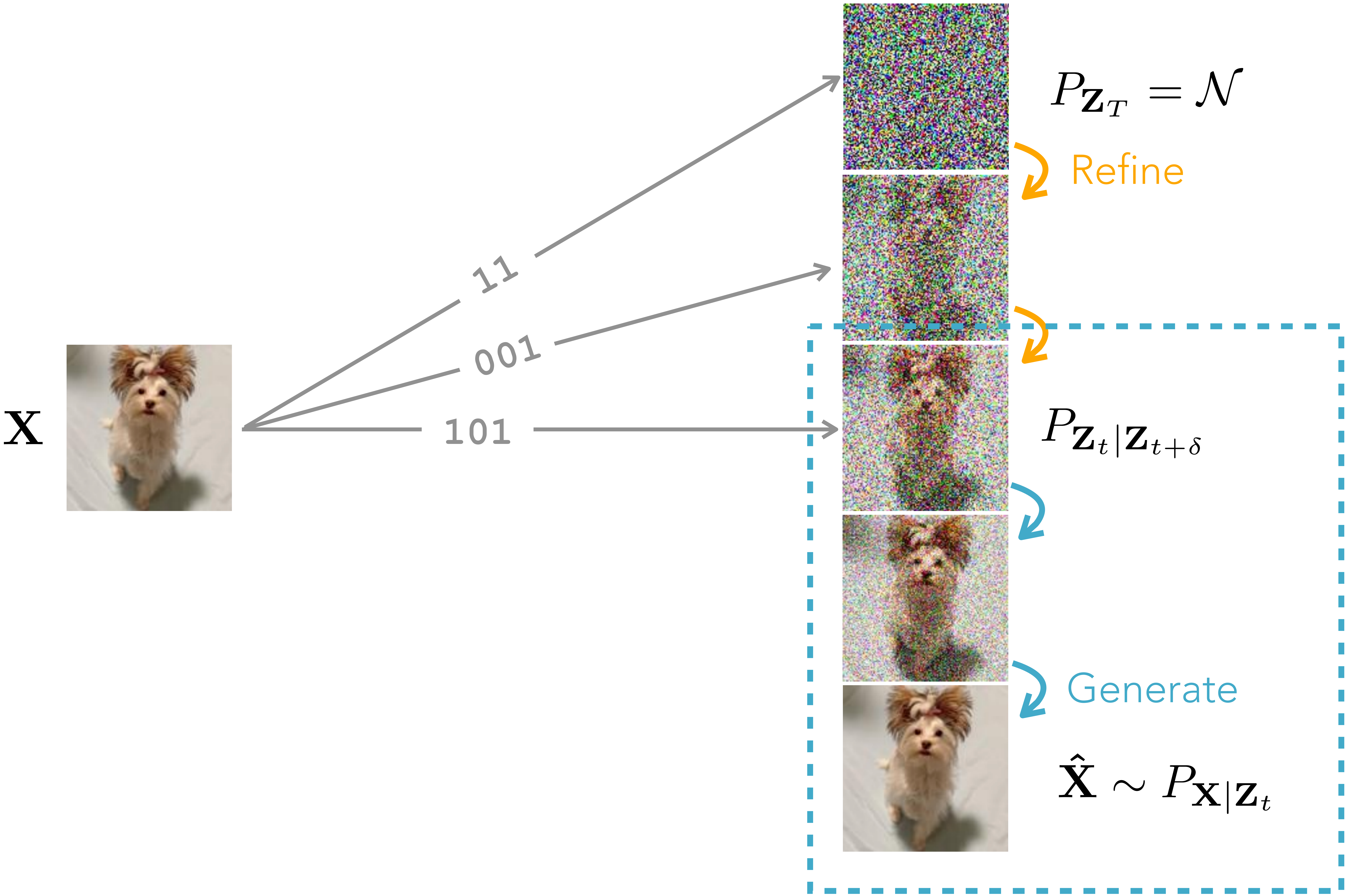
Theis et al. (2022)

DiffC



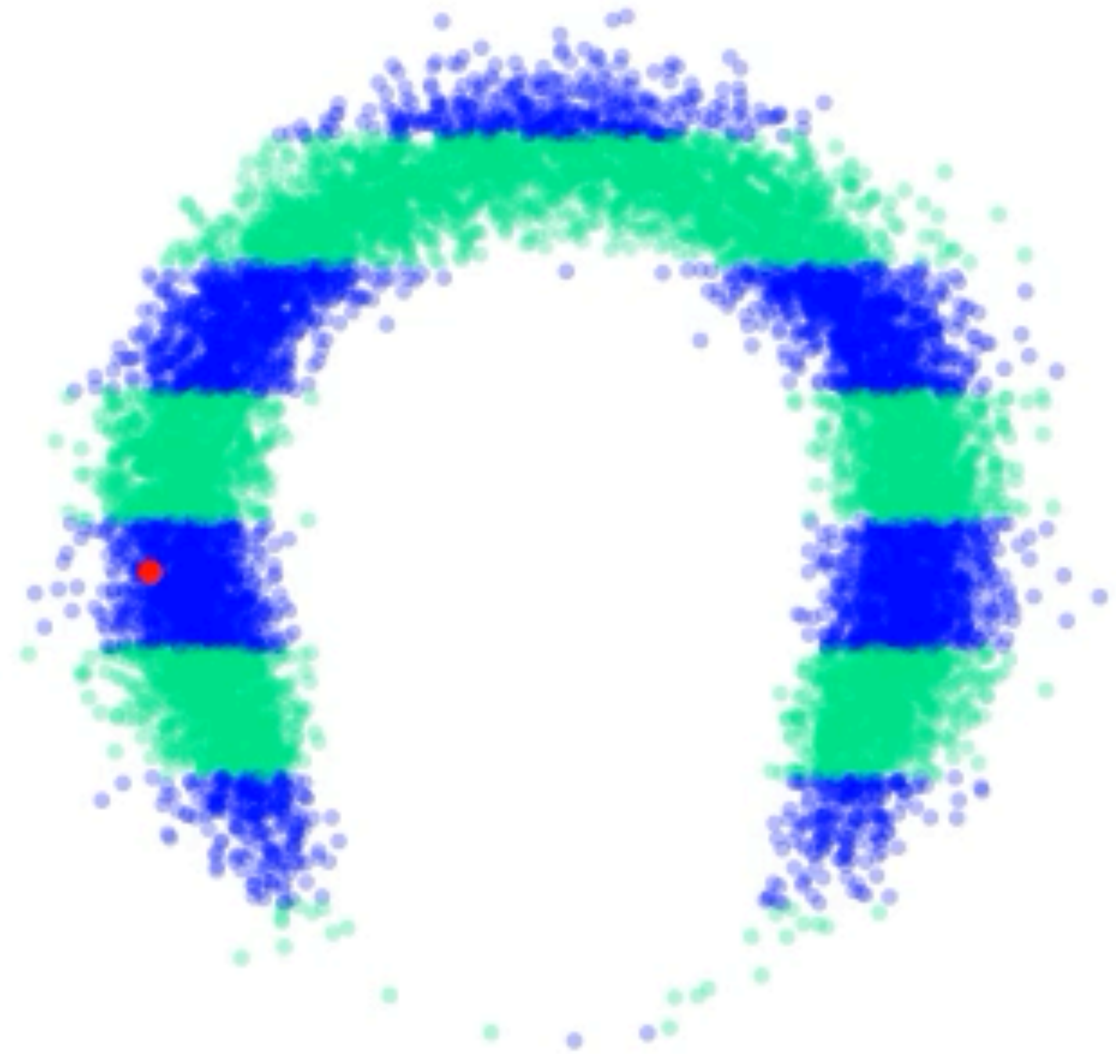
Theis et al. (2022)

DiffC

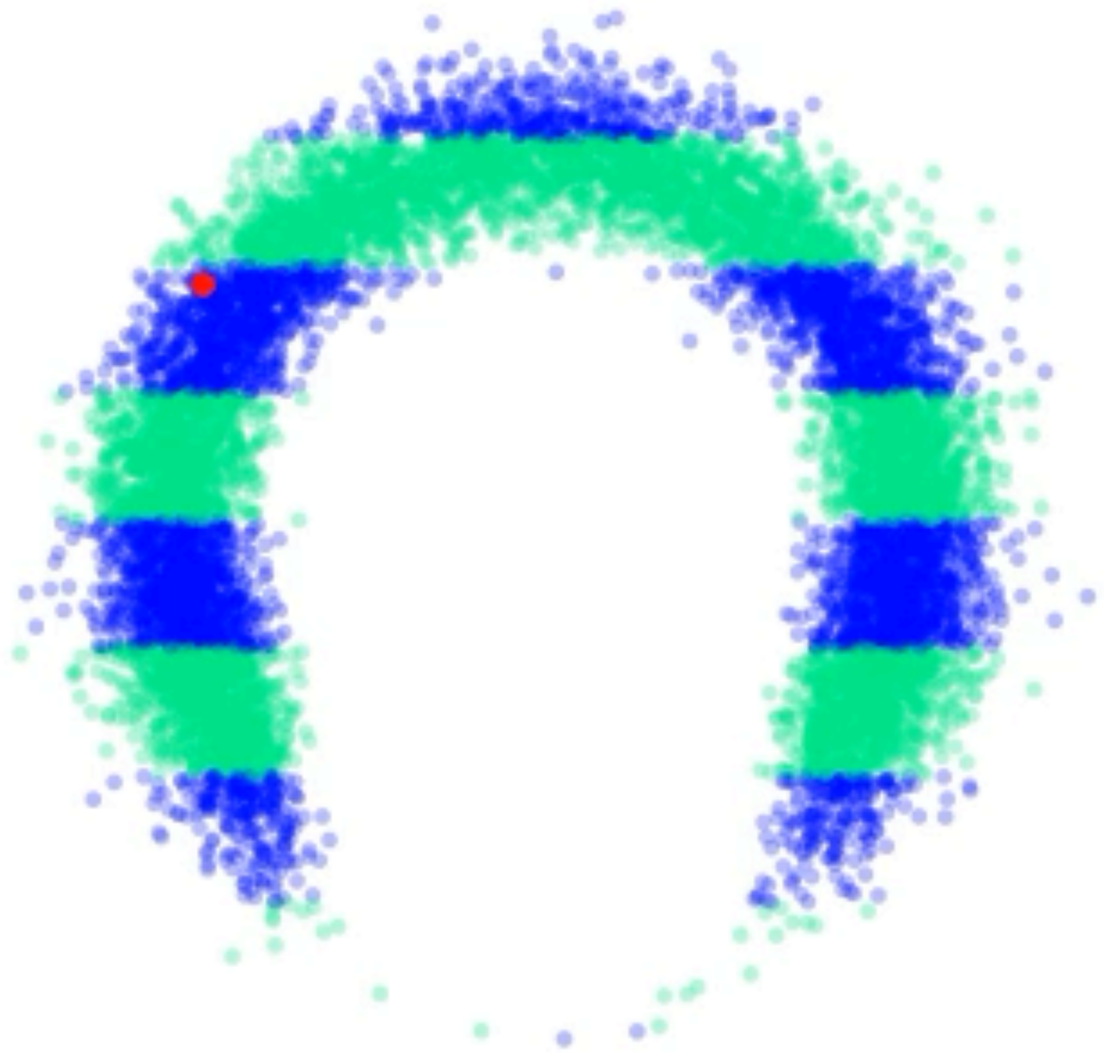


$\mathbf{X}$
11
001
101
$P_{\mathbf{Z}_T} = \mathcal{N}$
Refine
$P_{\mathbf{Z}_t|\mathbf{Z}_{t+\delta}}$
Generate
$\hat{\mathbf{X}} \sim P_{\mathbf{X}|\mathbf{Z}_t}$

DiffC-A (SDE)



DiffC-F (ODE)



DiffC-F vs DiffC-A

Theorem. Let $\mathbf{X} : \Omega \to \mathbb{R}^M$ have a smooth density p with finite

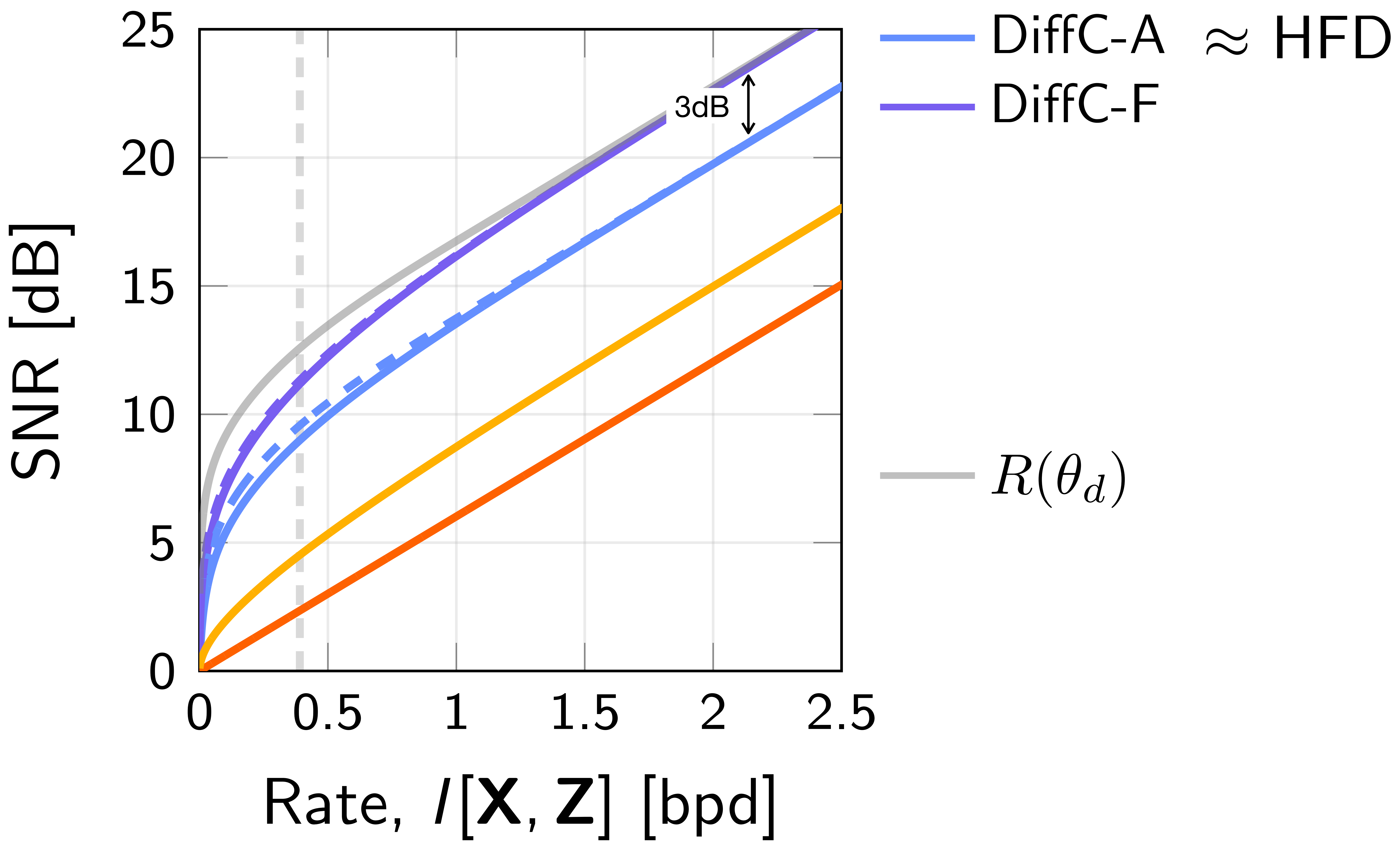
$$G = \mathbb{E}[\|\nabla \ln p(\mathbf{X})\|^2].$$

Let $\mathbf{Z}_t = \sqrt{1 - \sigma_t^2}\mathbf{X} + \sigma_t \mathbf{U}$ with $\mathbf{U} \sim \mathcal{N}(0, \mathbf{I})$. Let $\hat{\mathbf{X}}_A \sim P(\mathbf{X} \mid \mathbf{Z}_t)$ and let $\hat{\mathbf{X}}_F = \mathbf{Z}_0$ be the solution to the ODE with $\mathbf{Z}_t$ as initial condition. Then

$$\lim_{\sigma_t \to 0} \frac{\mathbb{E}[\|\hat{\mathbf{X}}_F - \mathbf{X}\|^2]}{\mathbb{E}[\|\hat{\mathbf{X}}_A - \mathbf{X}\|^2]} = \frac{1}{2}$$

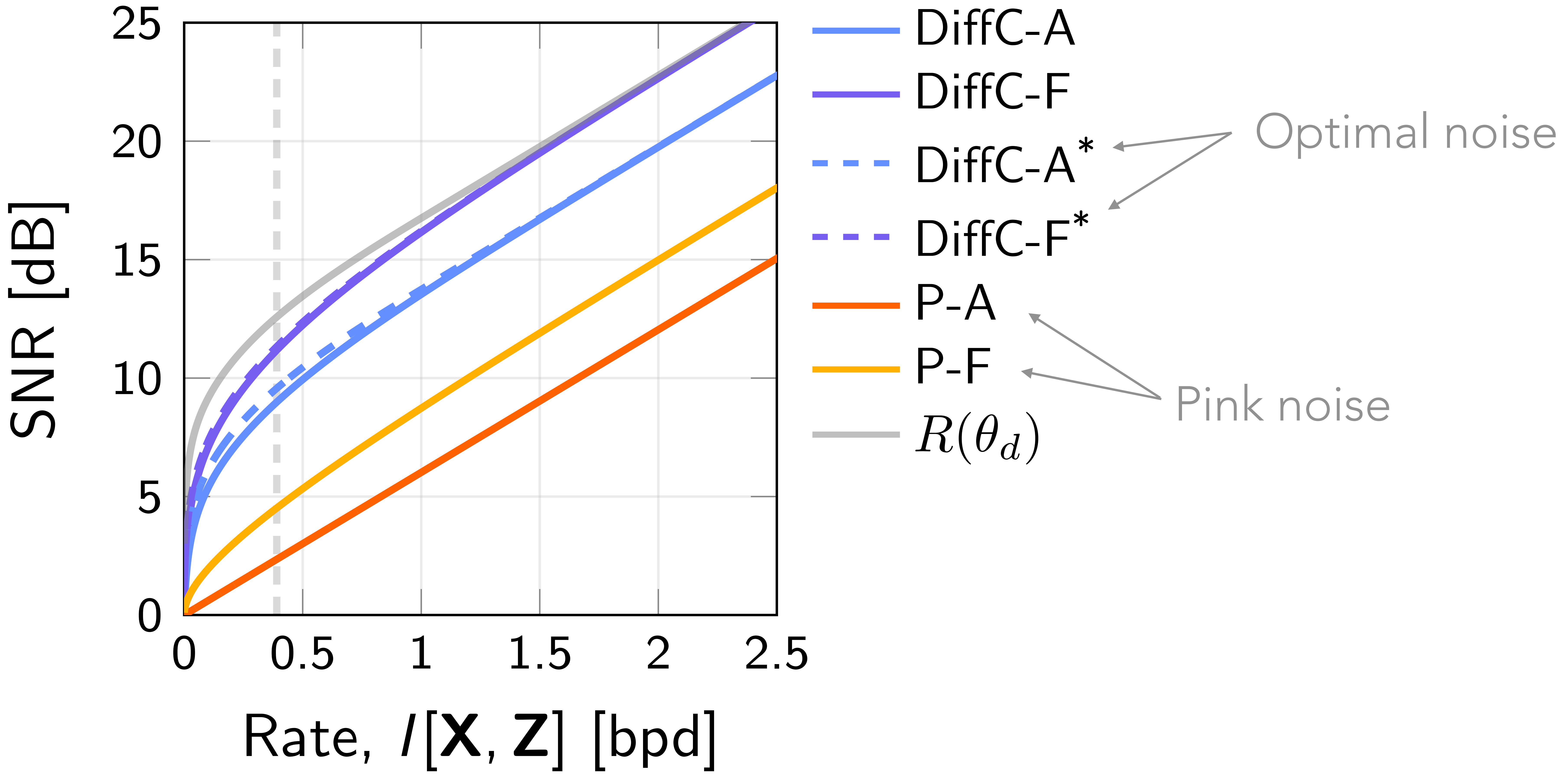
Theis et al. (2022)

Example: Multivariate Gaussian



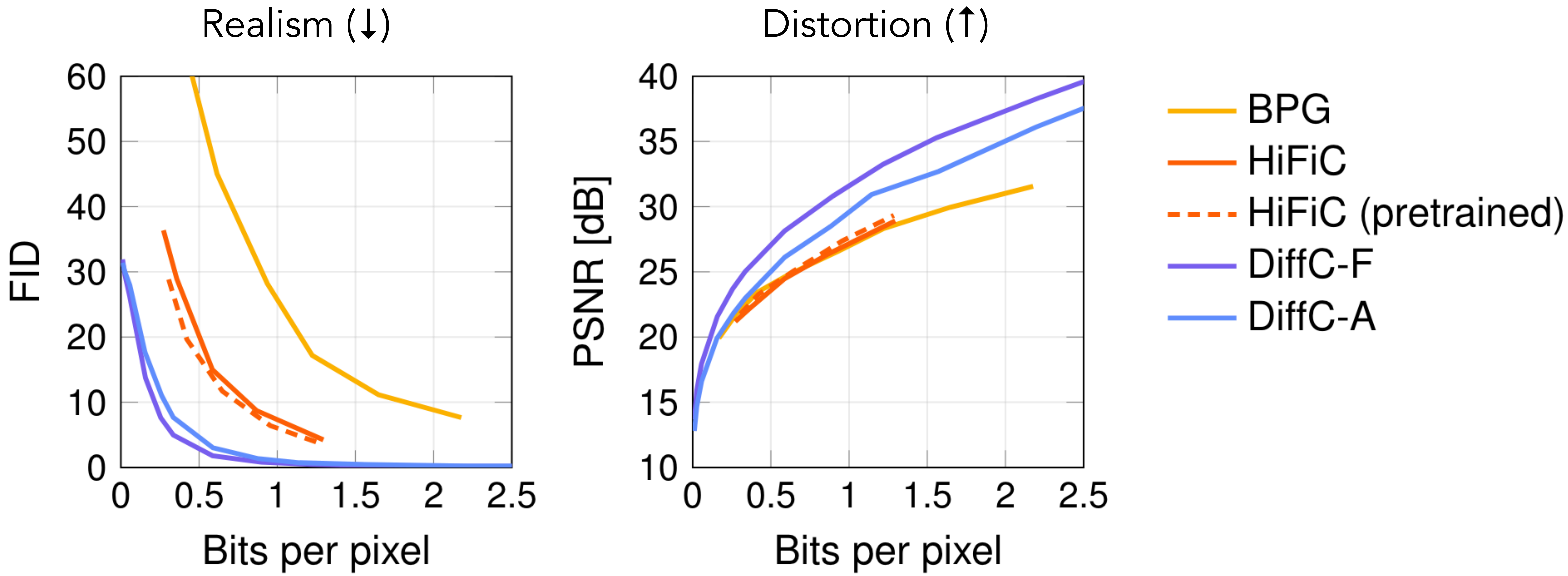
Theis et al. (2022)

Example: Multivariate Gaussian

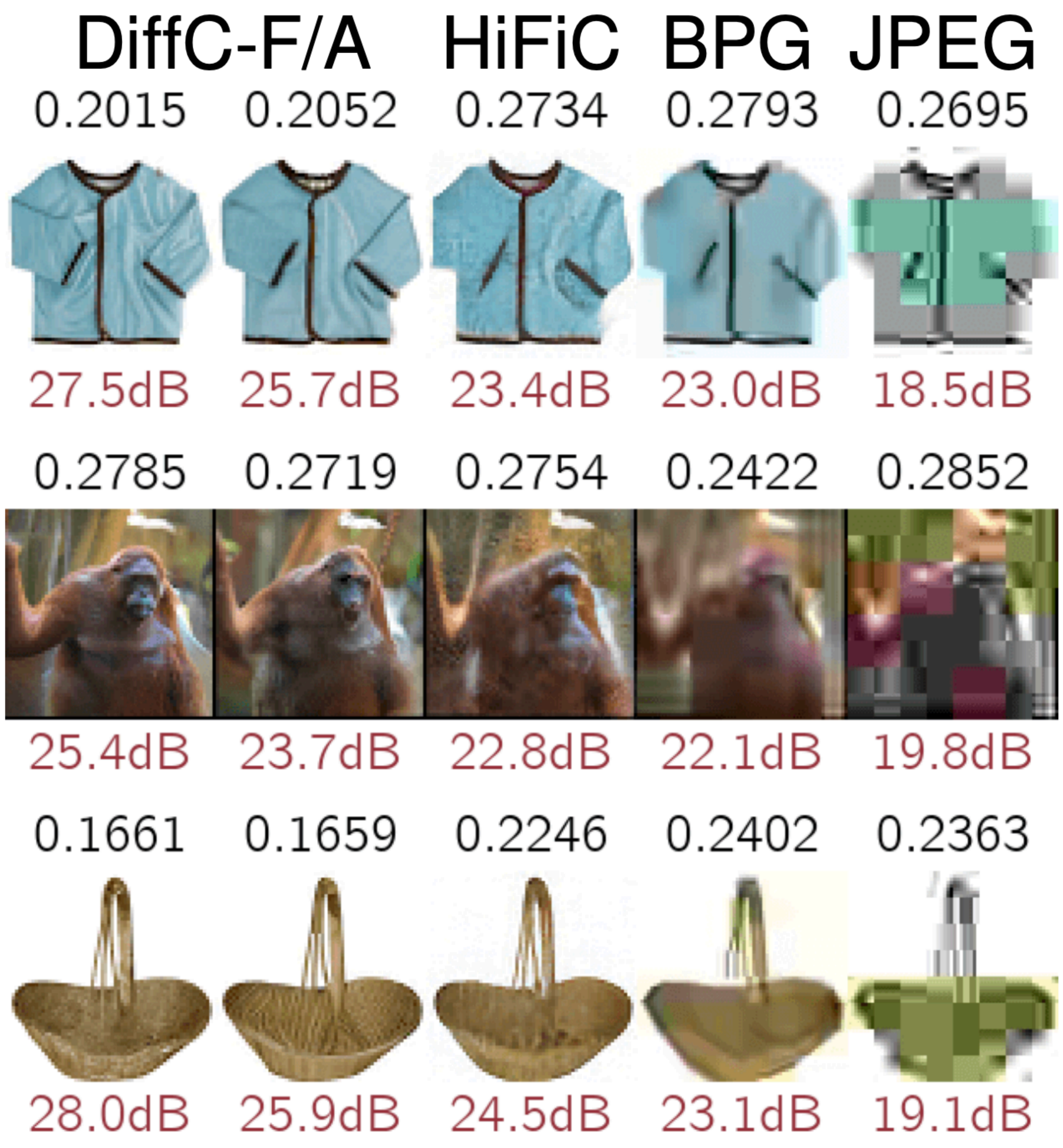


Theis et al. (2022)

ImageNet 64x64



Theis et al. (2022)



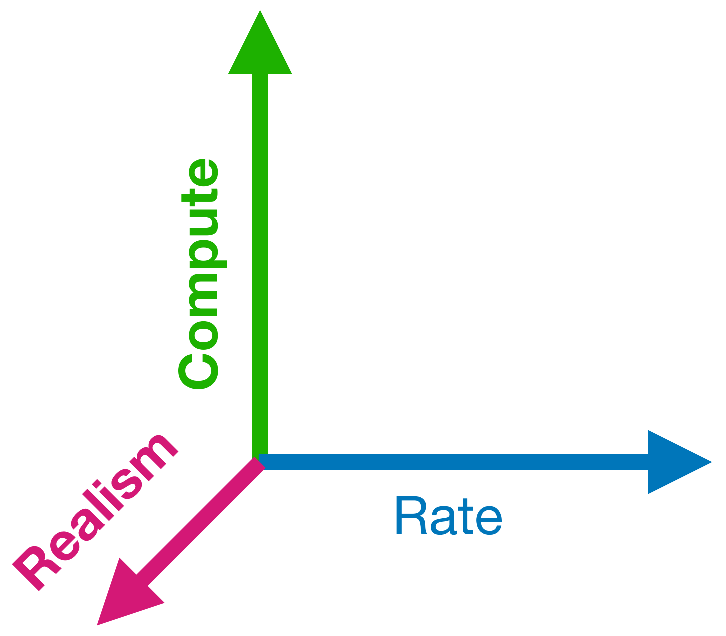
Theis et al. (2022)

Future research

- **What do good analysis transforms and distortions look like?**
  - Assuming perfect realism, what should the distortion measure?
  - Assuming perfect realism, what does a 1dB change in PSNR mean perceptually?

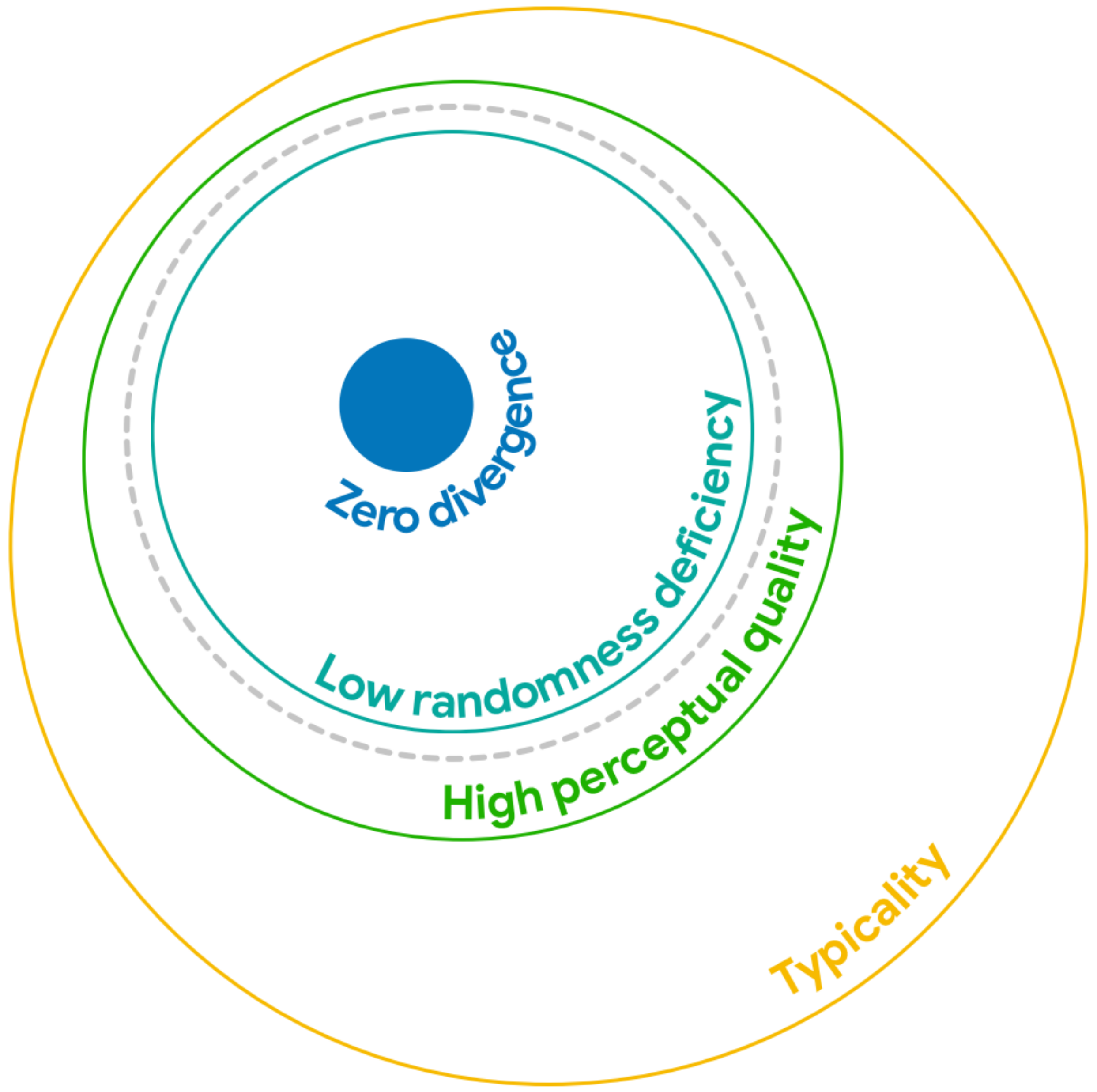
- **Make diffusion-based compression practical**
  - Fast Gaussian reverse channel coding
  - Dithered quantization instead of Gaussian noise
  - Rectified flows (or “flow matching”)
  - Distillation methods

- What are the limits of **low rate**, ***low complexity***, & **high realism**?



Realism revisited

What makes an image realistic?
ICML 2024



"Universal critic"

$$U(\mathbf{x}) = \log \sum_k \pi_k Q_k(\mathbf{x}) - \log P(\mathbf{x})$$

$$U(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \log \sum_k \pi_k \prod_n Q_k(\mathbf{x}_n) - \log \prod_n P(\mathbf{x}_n)$$

$$\rightarrow N D_{\mathrm{KL}}[Q \| P] \quad \text{where} \quad \mathbf{x}_n \sim Q$$

Thank you

Collaborators