Recent Advances in Diffusion-Based Generative Compression

Yibo Yang

Learn to Compress workshop @ ISIT 2025
June 26, 2025

Diffusion everywhere

Image

Imagen3

Video

Veo

Protein

Generative AI imagines new protein structures

"FrameDiff" is a computational tool that uses generative AI to craft new protein structures, with the aim of accelerating drug development and improving gene therapy.

FrameDiff

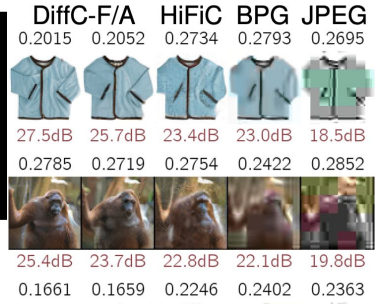
Music

Udio

Video games

Gamesgen

Compression?



Overview

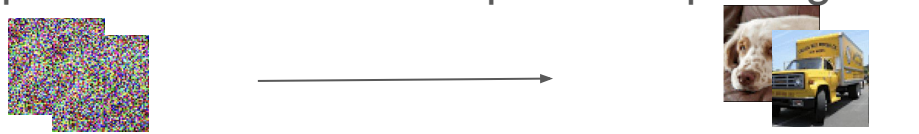
1. Background: diffusion
2. “Generative” lossy compression
3. Recent advances
4. Discussions

Background: diffusion

Borrowed from Arnaud Doucet's NeurIPS 2024 talk (https://neurips.cc/virtual/2024/invited-talk/101133)

Generative modeling and mass transport

The core problem is to find a map Φ transporting one distribution into another:

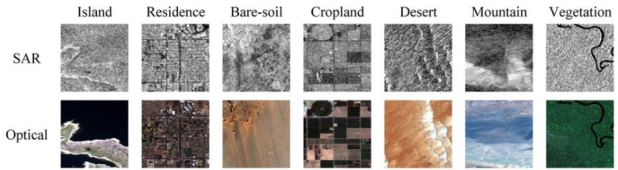


$$X_0 \sim P_{source}\ , \qquad X_1 = \Phi(X_0) \sim P_{target}$$

- Example: generative modeling:

$$P_{source} = \mathcal{N}(0, \mathrm{I}), \qquad P_{target} = P_{data}\ .$$

- Example: unpaired data-to-data translation / opt. transport:



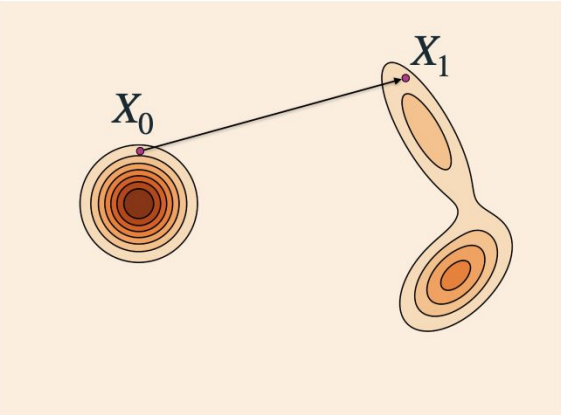
SAR-to-optical [Wang et al., 2024]

Image transfer [Gushchin et al., 2024]

SOTA methods are based on *iterative refinement*

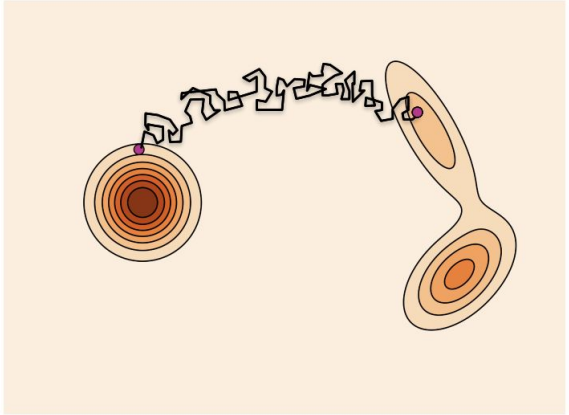
Direct mapping:
GAN, VAE, etc.

$$X_1 = \Phi(X_0)$$



ODE/SDE-based mapping:
diffusion, flow/bridge matching, etc.

$$(X_t)_{0 \le t \le 1} \quad X_{t+h} \leftarrow \Phi_{t+h|t}(X_t)$$



Source: https://neurips.cc/virtual/2024/tutorial/99531

Diffusion

- Forward process: $\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}B_t$

X_0 X_1

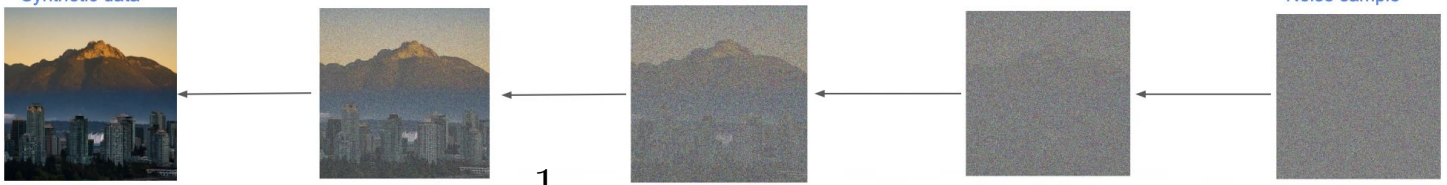
Diffusion

- Forward process: $\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}B_t$

X_0 X_1

- Reverse process: time reversal $(\bar{X}_t)_{t\in[0,1]}$ with $\bar{X}_t \overset{d}{=} X_{1-t}$

$\bar{X}_1 = X_0$ $\bar{X}_0 = X_1$



$$\mathrm{d}\bar{X}_t = \{\frac{1}{2}\bar{X}_t\mathrm{d}t + \nabla \log p_{1-t}(\bar{X}_t)\}\mathrm{d}t + \mathrm{d}\bar{B}_t$$

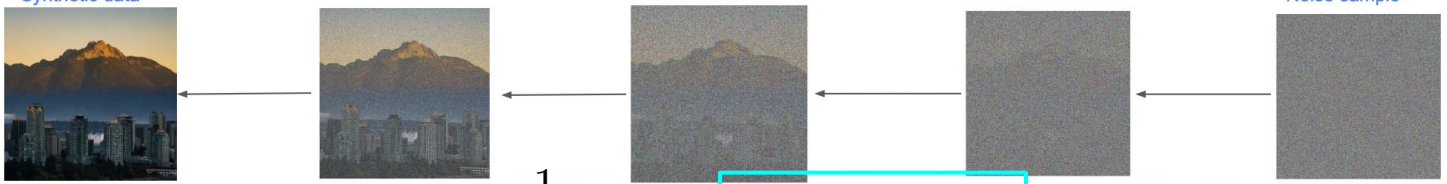
Diffusion

- Forward process: $\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}B_t$

X_0 X_1

- Reverse process: time reversal $(\bar{X}_t)_{t\in[0,1]}$ with $\bar{X}_t \overset{d}{=} X_{1-t}$

$\bar{X}_1 = X_0$ $\bar{X}_0 = X_1$



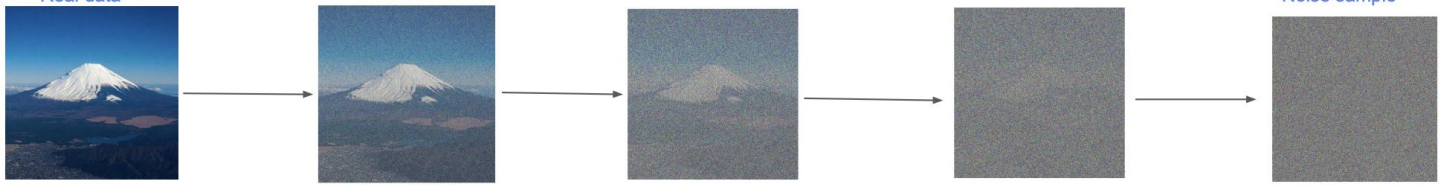
$$\mathrm{d}\bar{X}_t = \{\frac{1}{2}\bar{X}_t\mathrm{d}t + \nabla \log p_{1-t}(\bar{X}_t)\}\mathrm{d}t + \mathrm{d}\bar{B}_t$$

Intractable

Diffusion

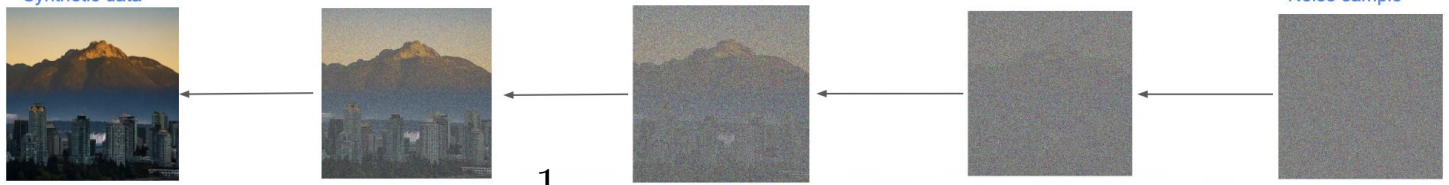
- Forward process: $\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}B_t$

X_0 Real data — X_1 Noise sample



- Reverse process: time reversal $(\bar{X}_t)_{t\in[0,1]}$ with $\bar{X}_t \overset{d}{=} X_{1-t}$

$\bar{X}_1 = X_0$ Synthetic data — $\bar{X}_0 = X_1$ Noise sample



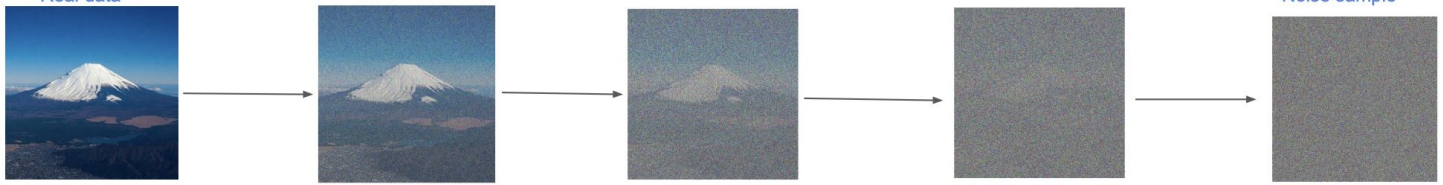
$$\mathrm{d}\bar{X}_t = \{\frac{1}{2}\bar{X}_t\mathrm{d}t + \nabla \log p_{1-t}(\bar{X}_t)\}\mathrm{d}t + \mathrm{d}\bar{B}_t$$

Tweedie's identity: $\nabla \log p_t(x_t) = \frac{1}{\sigma_t^2}(\alpha_t \mathbb{E}[X_0|X_t = x_t] - x_t)$ Denoiser

Diffusion

- Forward process: $\mathrm{d}X_t = -\frac{1}{2}X_t\mathrm{d}t + \mathrm{d}B_t$

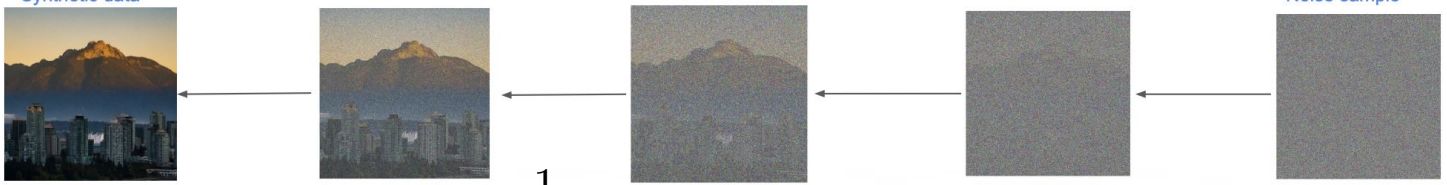
X_0 Real data



X_1 Noise sample

- Reverse process: time reversal $(\bar{X}_t)_{t\in[0,1]}$ with $\bar{X}_t \overset{d}{=} X_{1-t}$

$\bar{X}_1 = X_0$ Synthetic data



$\bar{X}_0 = X_1$ Noise sample

$$\mathrm{d}\bar{X}_t = \{\frac{1}{2}\bar{X}_t\mathrm{d}t + \nabla \log p_{1-t}(\bar{X}_t)\}\mathrm{d}t + \mathrm{d}\bar{B}_t$$

Train by regression: $\mathbb{E}[X_0|X_t = x_t] \approx \hat{x}_\theta(x_t), \hat{x}_\theta = \arg\min_{\phi_\theta} \mathbb{E}[\|X_0, \phi_\theta(X_t)\|^2]$

Conditional diffusion

- Given (X_0, Y) pairs, e.g., (image, text).
- A **conditional diffusion model** can be trained to approximate $P_{X_0|Y}$.
  (also see classifier (free) guidance [Dhariwal and Nichol 2021, Ho and Salimans 2022]).

Conditional generative process (score fun → conditional score fun):

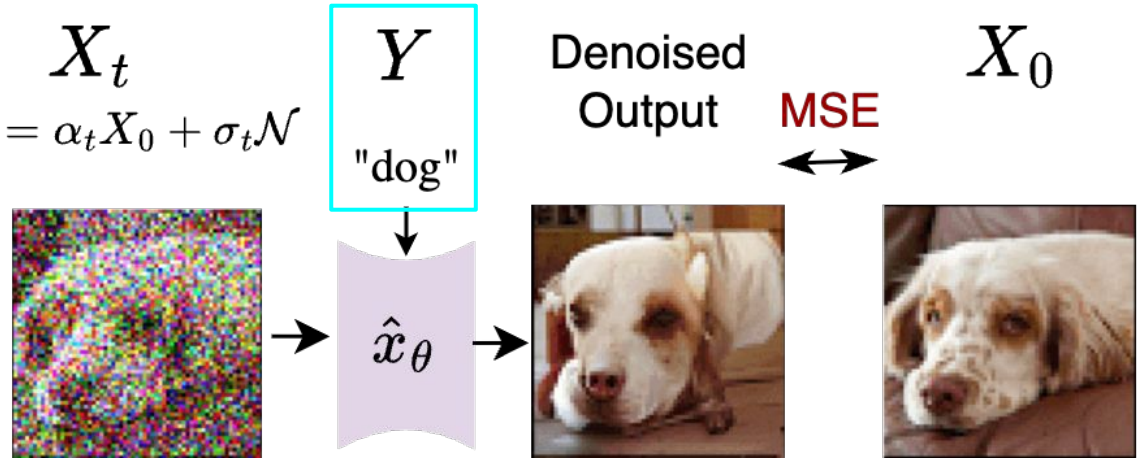
$$\mathrm{d}\bar{X}_t = \{\frac{1}{2}\bar{X}_t \mathrm{d}t + \nabla \log p_{1-t}(\bar{X}_t|y)\}\mathrm{d}t + \mathrm{d}\bar{B}_t$$

Tweedie's: $$\nabla \log p_t(x_t|y) = \frac{1}{\sigma_t^2}(\alpha_t \mathbb{E}[X_0|X_t = x_t, Y = y] - x_t)$$

Conditional diffusion

Training: $\mathbb{E}[X_0|X_t = x_t, Y = y] \approx \hat{x}_\theta(x_t, y), \hat{x}_\theta = \arg\min_{\phi_\theta} \mathbb{E}[\|X_0, \phi_\theta(X_t, Y)\|^2]$

Everything remains the same, except the denoising net receives y as extra input.



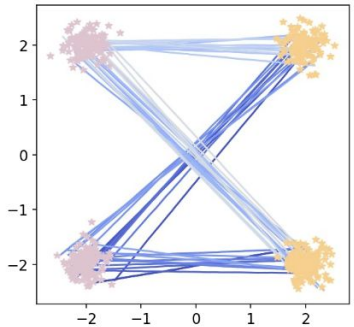
Generalizing Diffusion Models

(Albergo et al. 2022, Lipman et al. 2022, Liu et al. 2022)

- $X_0 \sim p_0, X_1 \sim p_1$
- **Deterministic** interpolation path

non-Markov

$$X_t = (1-t)\,X_0 + t\,X_1 \implies \mathrm{d}X_t = (X_1 - X_0)\mathrm{d}t = \frac{X_1 - X_t}{1-t}\mathrm{d}t$$



samples from p_0

samples from p_1

Generalizing Diffusion Models

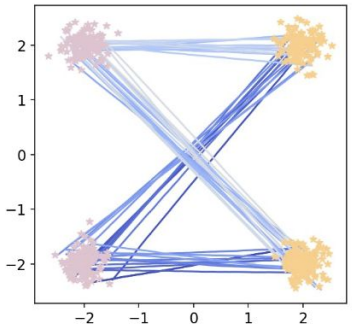
(Peluchetti 2021, Albergo et al. 2022, Liu et al. 2022)

- $X_0 \sim p_0, X_1 \sim p_1, Z \sim \mathcal{N}(0, I)$
- **Stochastic** interpolation path

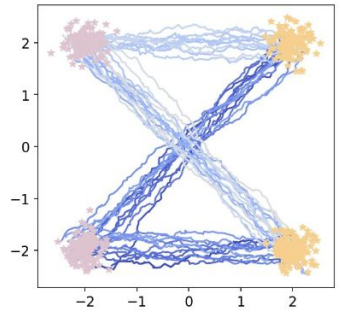
$$X_t = (1-t)\,X_0 + t\,X_1 + \sqrt{\epsilon t(1-t)}Z \implies \underbrace{\mathrm{d}X_t = \frac{X_1 - X_t}{1-t}\mathrm{d}t + \sqrt{\epsilon}\mathrm{d}B_t}_{\text{Brownian Bridge}}$$

non-Markov

noise



$\epsilon = 0$



$\epsilon > 0$

samples from p_0

samples from p_1

Markovian Projection (Gyöngy, 1986)

Removing the dependency on the future

- For $X_0 \sim p_0, X_1 \sim p_1$

non-Markov

$$\mathrm{d}X_t = \frac{X_1 - X_t}{1-t}\mathrm{d}t + \sqrt{\epsilon}\mathrm{d}B_t \quad \text{and} \quad \mathrm{d}X_t = \frac{\mathbb{E}[X_1|X_t] - X_t}{1-t}\mathrm{d}t + \sqrt{\epsilon}\mathrm{d}B_t$$

Markov

$$\approx \hat{X}_{\theta^*}(t, X_t)$$

Denoiser

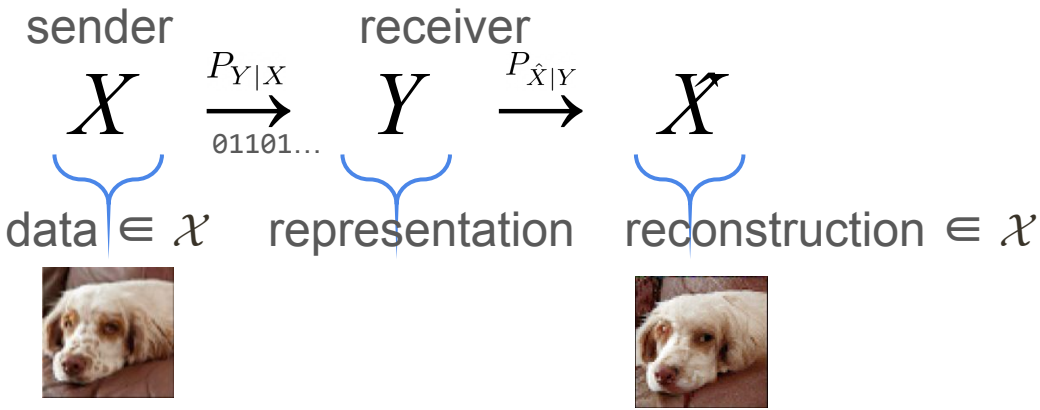
$$\theta^* = \operatorname{argmin} \mathbb{E}[||X_1 - \hat{X}_\theta(t, X_t)||^2]$$

Both ODE/SDE have **same marginals**!

- If $\epsilon > 0$ - **Bridge Matching** (Peluchetti 2021, Albergo et al. 2022, Liu et al. 2022)
- If $\epsilon = 0$ - **Flow Matching** (Lipman et al. 2022, Liu et al. 2022, DelBracio et al. 2023)

Generative lossy compression

Setup:

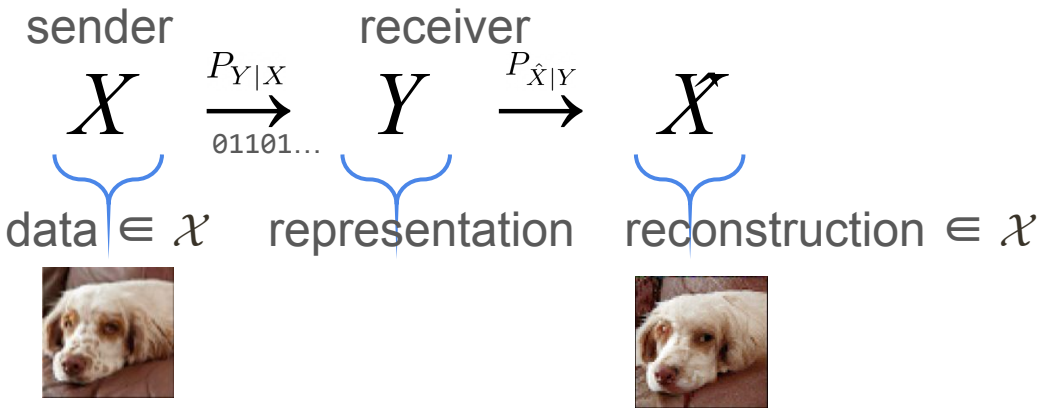


Many possibilities for Y, e.g.,

- Learned embedding [Mentzer et al., 2020], [Yang and Mandt, 2023]
- Text / caption [Lei et al., 2023], [Careil et al., 2023]
- Reconstruction produced by another codec [Hoogeboom et al., 2023], [Ghouse et al., 2023]

Generative lossy compression

Setup:

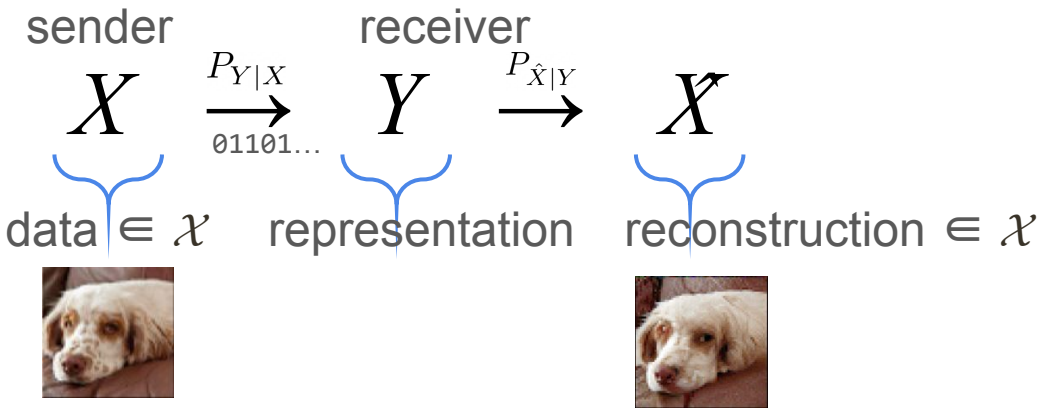


We want to optimize this system for:

- Low *distortion* $\mathbb{E}[\rho(X, \hat{X})]$
  for some distortion function $\rho: \mathcal{X} \times \mathcal{X} \to [0, \infty)$, e.g. squared error.
- Low *bit-rate*: $H[Y]$ (or $I(X, Y)$ if channel simulation).

Generative lossy compression

Setup:



We want to optimize this system for:

- Low *distortion* $\mathbb{E}[\rho(X, \hat{X})]$
  for some distortion function $\rho: \mathcal{X} \times \mathcal{X} \to [0, \infty)$, e.g. squared error.
- Low *bit-rate*: $H[Y]$ (or $I(X, Y)$ if channel simulation).
- High *realism* (low divergence): $d(P_X, P_{\hat{X}})$ [Blau and Michaeli, 2019]

This talk will focus on the case of "perfect realism", i.e., $P_X = P_{\hat{X}}$.

Generative lossy compression: some theory

Rate-distortion-perception theory [Blau and Michaeli, 2019]:

- Realism generally comes at the expense of rate-distortion performance.
- In particular, define the realism-constrained R-D function:

$$R(D,P) = \min_{P_{\hat{X}|X}} I(X,\hat{X})$$
$$\text{s.t. } \mathbb{E}[\rho(X,\hat{X})] \le D, d(P_X, P_{\hat{X}}) \le P$$

Theorem 2 *When using the squared-error distortion, the function $R(\cdot,0)$ (rate-distortion at perfect perceptual quality) is bounded by*

$$R(D,0) \le R(\tfrac{1}{2}D,\infty). \quad (8)$$

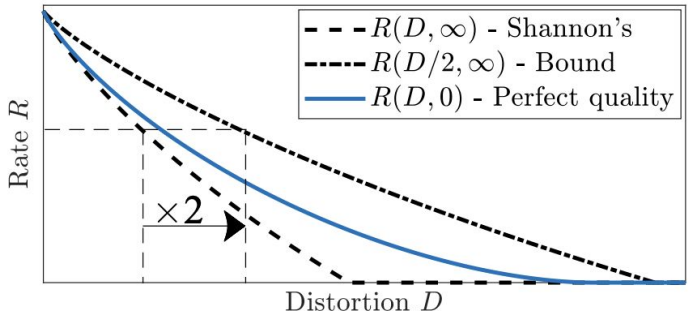


Figure 5. **Illustration of Theorem 2.** When using the MSE distortion, the rate-distortion curve for compression with perfect perceptual quality (blue) is higher than Shannon's rate-distortion function (black dashed line) but is necessarily lower than the $2\times$ scaled version of Shannon's function (dotted line).

Generative lossy compression: some theory

Assuming no common randomness (and squared distortion as before):

- Perfect realism → exactly a two-fold increase in squared error [Yan et al., 2021].

$$R(D,0) = R(\frac{1}{2}D, \infty)$$

- A common choice for $P_{X|Y}$ is to train a *conditional generative model* for $P_{X|Y}$.
  - It can be shown that the resulting X' satisfies:
    - $P_X = P_{X'}$, i.e., perfect realism, assuming ideal modeling;
    - Its distortion is bounded by 2 x distortion of the MMSE estimator [Blau and Michaeli, 2019; Hoogeboom et al., 2023].

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

    - If Y is R-D optimal with distortion D, then this architecture is R-D-P optimal with perfect realism and distortion $2D$ [Yan et al., 2021].

Generative lossy compression: some theory

With common randomness: better performance can be achieved [Theis and Agustsson, 2021], [Zhang et al., 2021, e.g., encoding/transmitting information using diffusion itself [Theis et al., 2022].

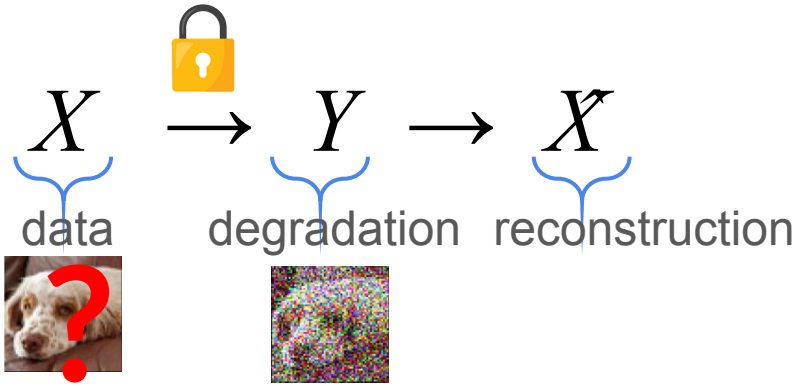
- $P_X = P_{X'}$, i.e., perfect realism, assuming ideal modeling;
- Its distortion is bounded by 2 x distortion of the MMSE estimator [Blau and Michaeli, 2019; Hoogeboom et al., 2023].

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

- If Y is R-D optimal with distortion D, then this architecture is R-D-P optimal with perfect realism and distortion $2D$ [Yan et al., 2021].

Relation to inverse-problem solving (“image restoration”)

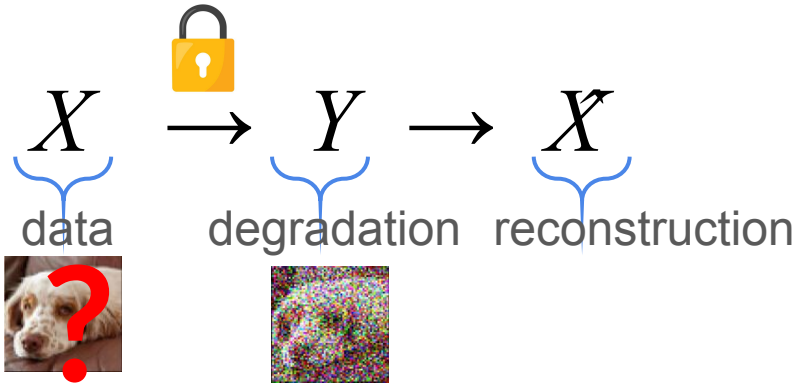
In image restoration, we assume a fixed degradation process $X \rightarrow Y$:



- Given $Y = y$, want to estimate X consistent with y while being “image-like”/realistic.

Relation to inverse-problem solving ("image restoration")

In image restoration, we assume a fixed degradation process $X \to Y$:



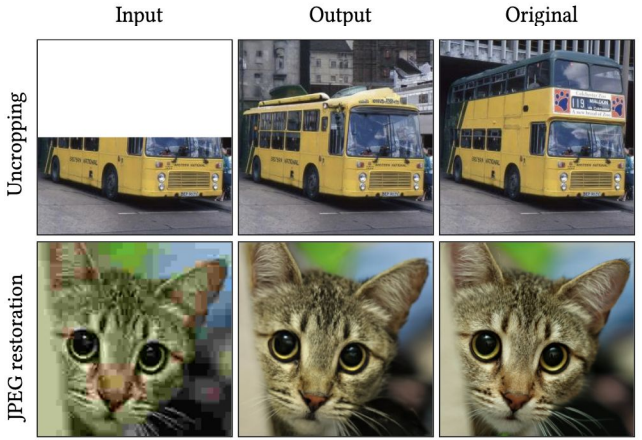
- Given $Y = y$, want to estimate X consistent with y while being "image-like"/realistic.
- Typically solved by sampling from $P_{X|Y=y}$ or computing $\mathbf{E}[X|Y=y]$.
- If $X \to Y$ is an existing codec (e.g., JPEG), the problem then becomes compression artifact removal ("JPEG restoration").

Relation to inverse-problem solving ("image restoration")

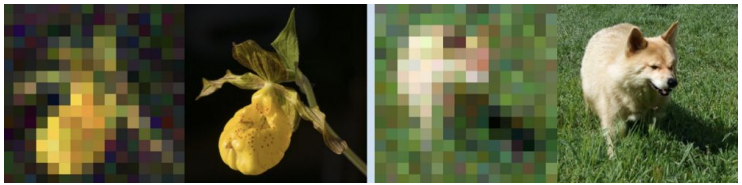
Diffusion-based approaches (2 categories):

1. Conditional diffusion trained on (X, Y) pairs, e.g., SRDiff [Li et al., 2021], Palette [Saharia et al., 2022].
2. Training-free methods based on a diffusion prior, e.g., DDRM [Kawar et al., 2022], DPS [Chung et al., 2023], etc.

See excellent survey [Daras et al., 2024].



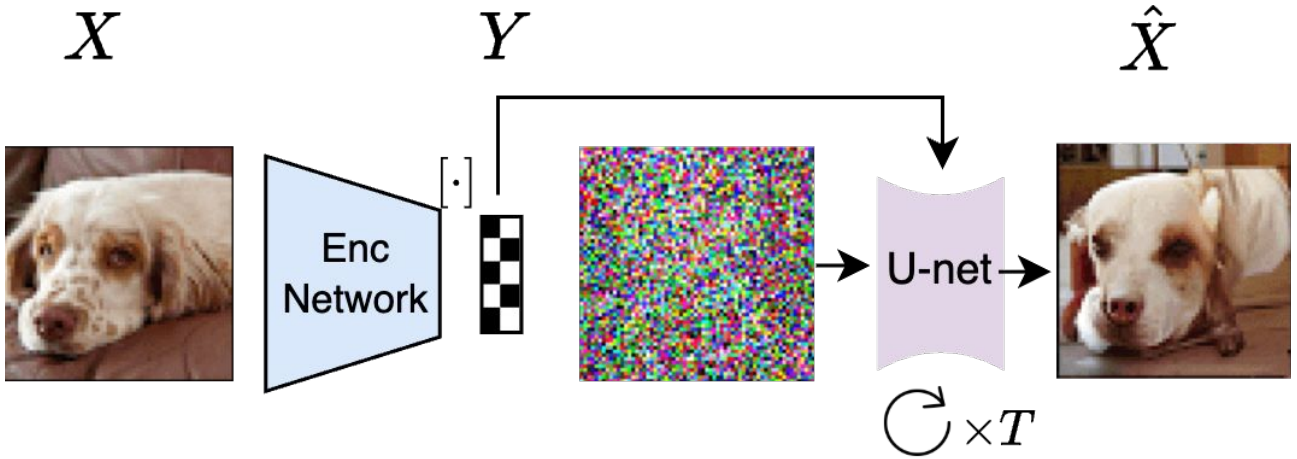
[Saharia et al., 2022]



[Chung et al., 2023]

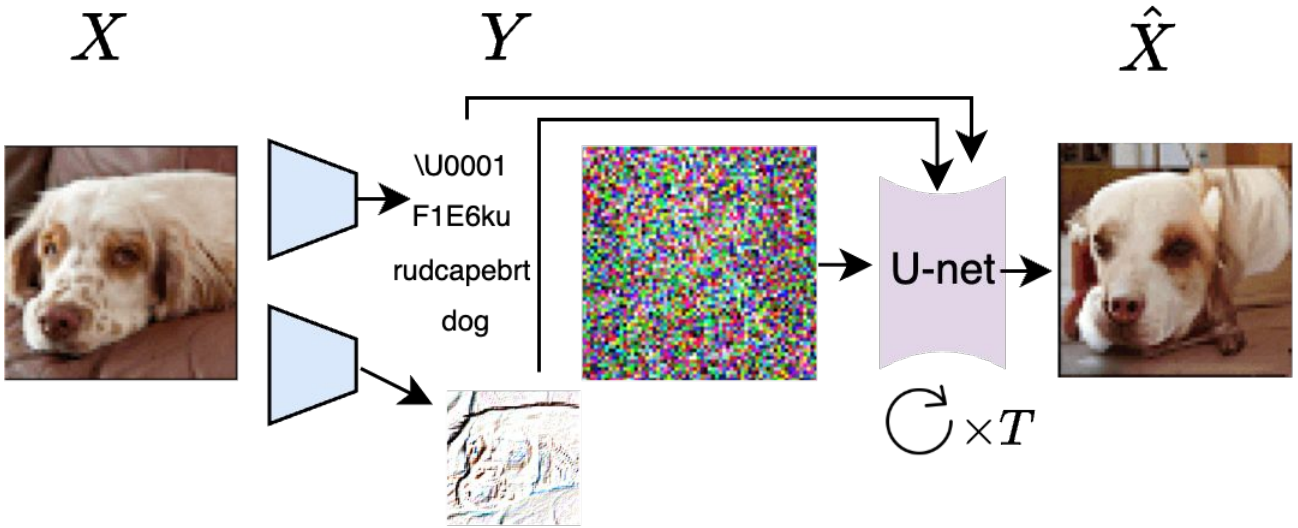
Recent advances in diffusion-based gen. compression

CDC [Yang and Mandt, 2023]



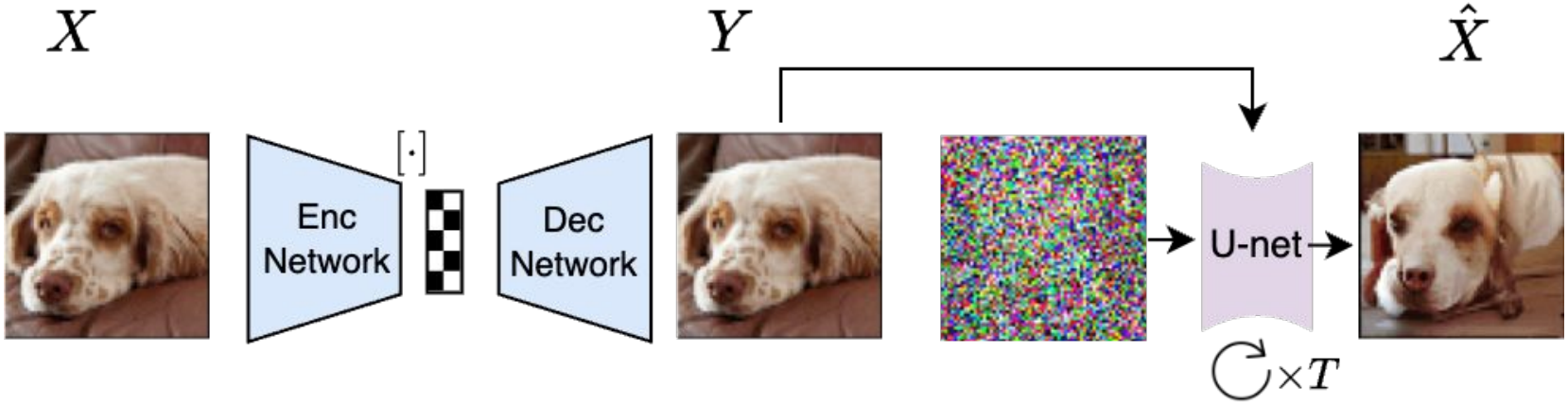
- Y is a **learned quantized embedding**.
- Trained end-to-end on a R-D + realism (conditional diffusion) loss.

Text + sketch [Lei et al., 2023], PerCO [Careil et al., 2023]



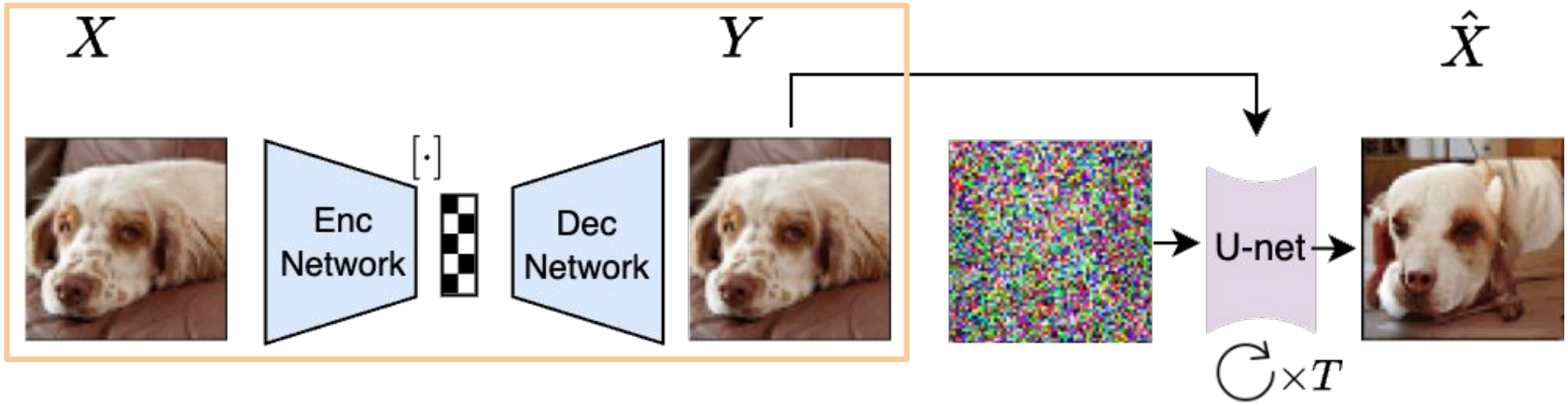
- Y is a **text string** and **(optionally) a color/edge map**.
- Can use pre-trained modules, e.g., image captioning model for encoding and text-to-image diffusion for decoding.

HFD [Hoogeboom et al., 2023], DIRAC [Ghouse et al., 2023]



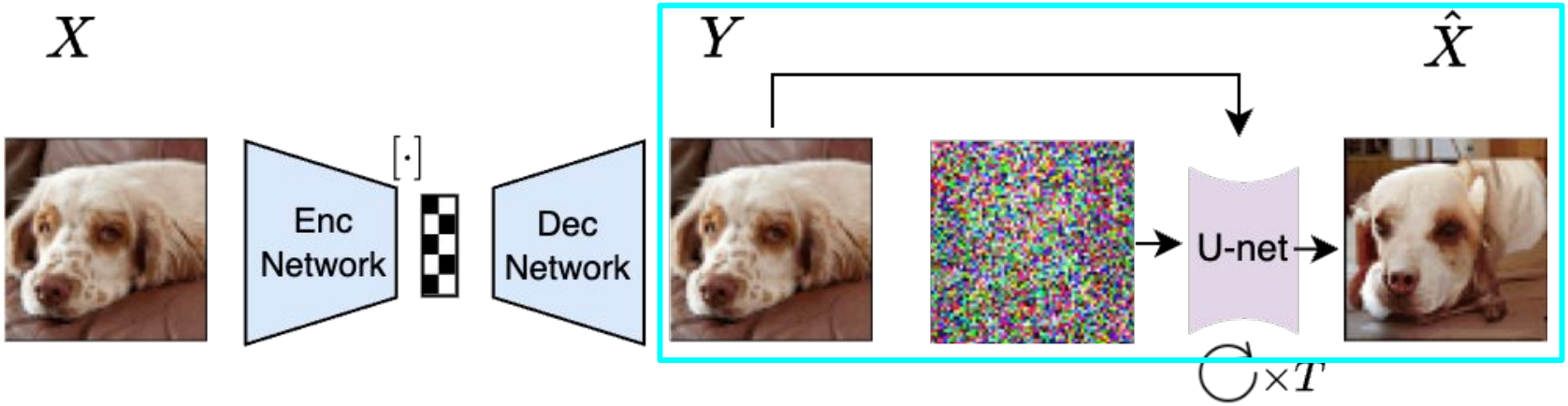
- Y is **the pixel-space reconstruction produced by another codec**.
- Training can be done in two stages:

HFD [Hoogeboom et al., 2023], DIRAC [Ghouse et al., 2023]



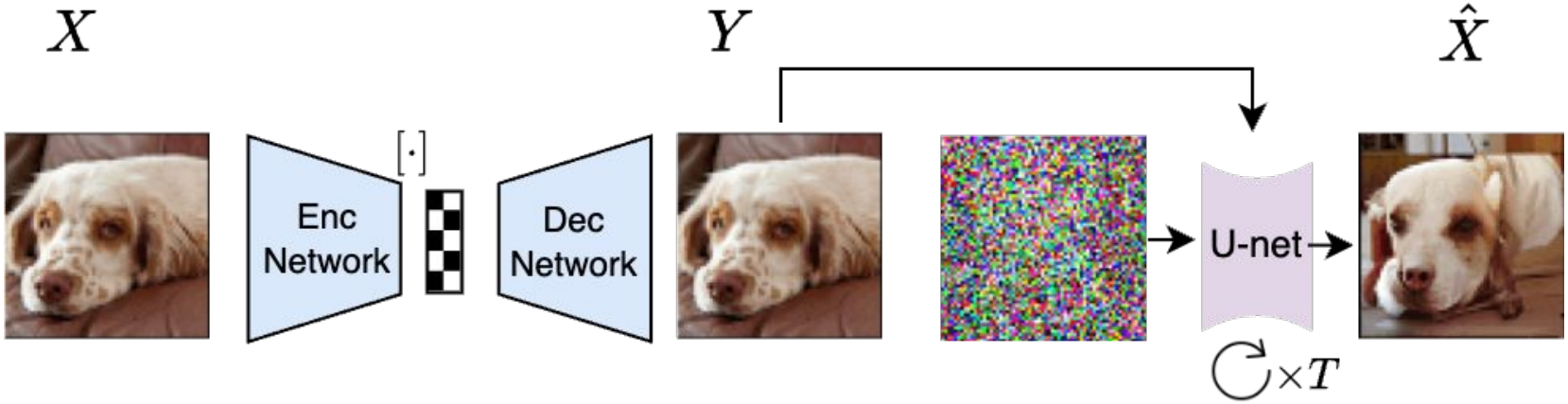
- Y is **the pixel-space reconstruction produced by another codec**.
- Training can be done in two stages:
  1. Train a base compressive autoencoder $X \rightarrow Y$ on R-D loss.

HFD [Hoogeboom et al., 2023], DIRAC [Ghouse et al., 2023]



- Y is **the pixel-space reconstruction produced by another codec**.
- Training can be done in two stages:
  1. Train a base compressive autoencoder $X \to Y$ on R-D loss.
  2. Train a conditional diffusion model $P_{X|Y} \approx P_{X|Y}$; or a flow $P_Y \to P_X$ on the (Y, X) pairs from stage-1 autoencoder.

HFD [Hoogeboom et al., 2023], DIRAC [Ghouse et al., 2023]

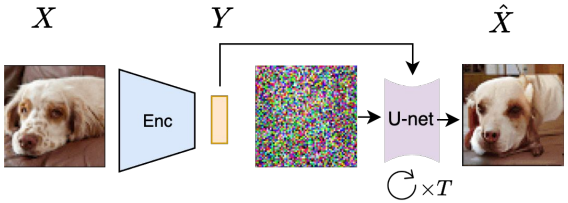


- Lack of end-to-end training doesn't limit the R-D-P performance in theory [Hoogeboom et al., 2023].
- Gives strong rate-distortion & rate-realism performance, outperforming the end-to-end trained baseline [Yand & Mandt, 2023].
- Can be applied to other existing codecs for perfect realism.

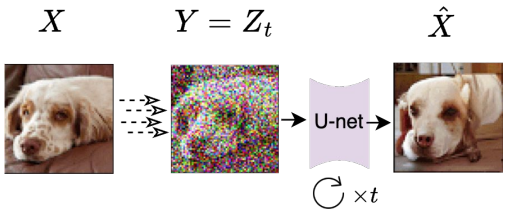
Communicating information with diffusion

Most existing approaches for generative lossy compression:

- The sender picks a representation $Y=y$, makes it available to the receiver via an auxiliary entropy model, and then
- The receiver samples $X|Y=y \sim P_{X|Y=y}$ with a cond. gen. model.



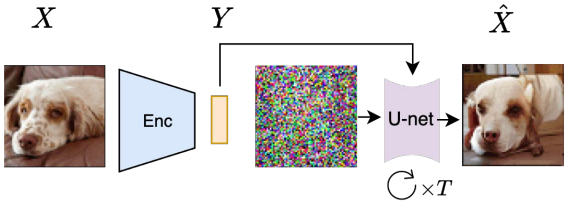
Can we accomplish the above with a single diffusion model?



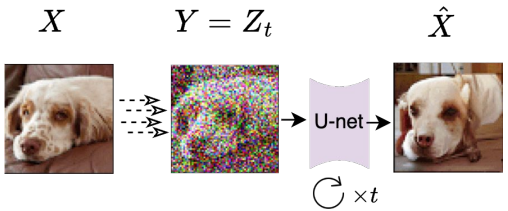
Communicating information with diffusion

Most existing approaches for generative lossy compression:

- The sender picks a representation $Y=y$, makes it available to the receiver via an auxiliary entropy model, and then
- The receiver samples $X|Y=y \sim P_{X|Y=y}$ with a cond. gen. model.



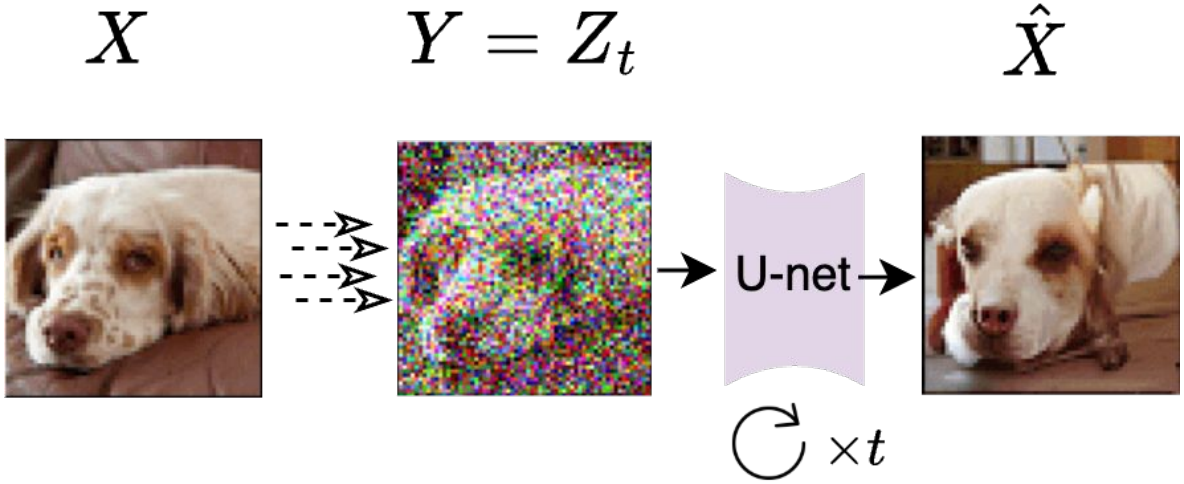
Can we accomplish the above with a single diffusion model?



If so, this can translate to:

- Lower deployment overhead.
- Improved R-D-P performance.
- Other desirable features (e.g., progressive/scalable coding).

Lossy compression with *unconditional* diffusion [Ho et al., 2020; Theis et al., 2022]



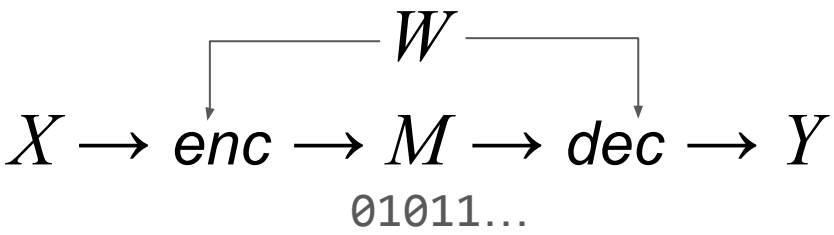
- Y is a **noisy version of X corrupted by Gaussian noise.**

$$Y := Z_t = \alpha_t X + \sigma_t N \quad \text{where} \quad N \sim \mathcal{N}(0, I)$$

- Works on top of a pre-trained variational diffusion model [Kingma et al., 2021]; no additional training required.

Key ingredient: channel simulation

(One-shot) channel simulation/reverse channel coding [Li 2024, section 2.1]:



Setup:

- A source of **common randomness** W, available to both sender and receiver
- Encoder: $(X, W) \to M \in \{0,1\}^*$, decoder: $(X, M) \to Y$

Goals:

- Guarantee $Y|X \sim P_{Y|X}$, for a prescribed channel $P_{Y|X}$.
- Minimize the bit-rate, e.g., $\mathbb{E}[|M|]$
- Minimize the computational complexity.

Key ingredient: channel simulation

Basic result (one-shot CS with unlimited common randomness) [Li and Anantharam, 2021]:

$$I \leq \min \mathbb{E}[|M|] \leq I + \log_2(I+1) + 5$$

where $I := I(X, Y)$

Key ingredient: channel simulation

Basic result (one-shot CS with unlimited common randomness) [Li and Anantharam, 2021]:

$$I \leq \min \mathbb{E}[|M|] \leq I + \log_2(I+1) + 5$$

where $I := I(X,Y)$

If we approximate the marginal distribution of Y by P_Y^θ, then the above rate bound becomes

$$I^\theta \leq \min \mathbb{E}[|M|] \leq I^\theta + \log_2(I^\theta+1) + 5$$

where $\mathbb{E}_{x\sim P_X}[\mathrm{KL}(P_{Y|X=x}\|P_Y^\theta)] := I^\theta \geq I$

Key ingredient: channel simulation

Basic result (one-shot CS with unlimited common randomness) [Li and Anantharam, 2021]:

$$I \leq \min \mathbb{E}[|M|] \leq I + \log_2(I+1) + 5$$

where $I := I(X,Y)$

If we approximate the marginal distribution of Y by P_Y^θ, then the above rate bound becomes

$$I^\theta \leq \min \mathbb{E}[|M|] \leq I^\theta + \log_2(I^\theta+1) + 5$$

where $\mathbb{E}_{x\sim P_X}[\mathrm{KL}(P_{Y|X=x}\|P_Y^\theta)] := I^\theta \geq I$

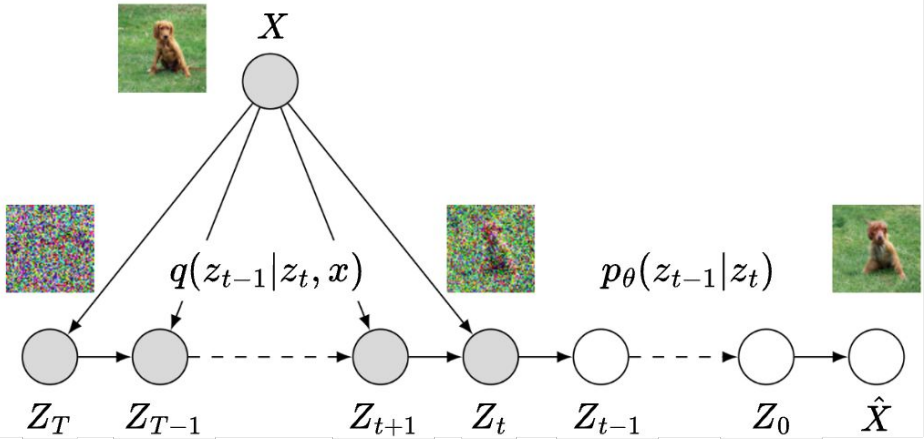
Takeaway: let $Y = RCC(P_{Y|X}, P_Y^\theta, X)$ be the output of a channel simulation / reverse channel coding algorithm, then it holds that

$Y|X \sim P_{Y|X}$ using $\approx \mathbb{E}_{x\sim P_X}[\mathrm{KL}(P_{Y|X=x}\|P_Y^\theta)]$ bits/sample (up to a logarithmic overhead)

Diffusion as a hierarchical latent variable model

A discrete-time diffusion model equivalently optimizes an NELBO:

$$-\log p_\theta(x) \leq \underbrace{\mathrm{KL}(q(z_T|x)p(z_T))}_{:=L_T} + \sum_{t=1}^{T} \underbrace{\mathbb{E}[\mathrm{KL}(q(z_{t-1}|z_t,x)\|p_\theta(z_{t-1}|z_t))]}_{:=L_{t-1}} + \underbrace{\mathbb{E}[-\log p(x|z_0)]}_{:=L_{x|z_0}}$$



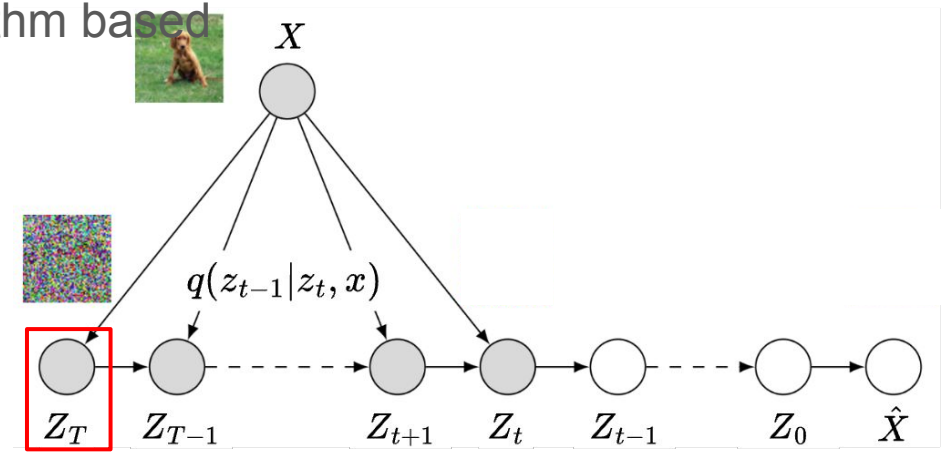
Diffusion as a hierarchical latent variable model

A discrete-time diffusion model equivalently optimizes an NELBO:

$$-\log p_\theta(x) \leq \underbrace{\mathrm{KL}(q(z_T|x)p(z_T))}_{:=L_T} + \sum_{t=1}^{T} \underbrace{\mathbb{E}[\mathrm{KL}(q(z_{t-1}|z_t,x)\|p_\theta(z_{t-1}|z_t))]}_{:=L_{t-1}} + \underbrace{\mathbb{E}[-\log p(x|z_0)]}_{:=L_{x|z_0}}$$

This suggests a *progressive* coding algorithm based on channel simulation [Ho et al., 2020]:

- At time T, simulate $Z_T|X \sim q(z_T|X)$ by running $RCC(q(z_T|X), p(z_T), X)$, costing L_T bits.



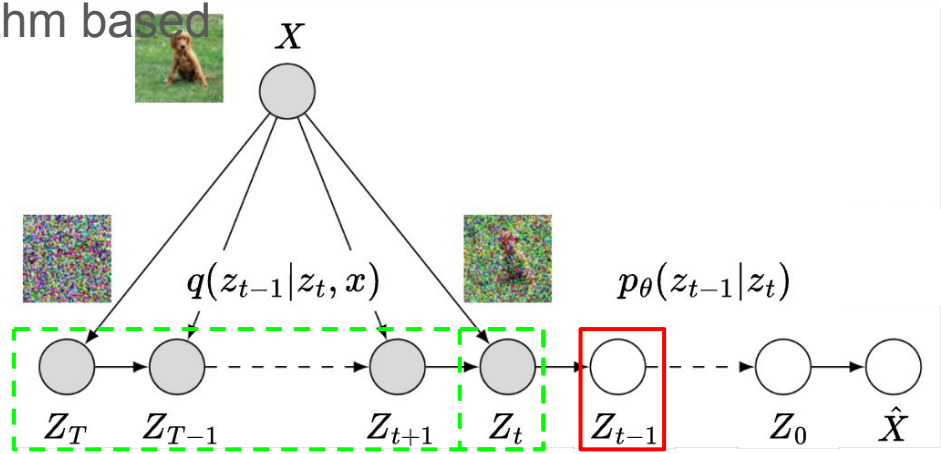
Diffusion as a hierarchical latent variable model

A discrete-time diffusion model equivalently optimizes an NELBO:

$$-\log p_\theta(x) \leq \underbrace{\mathrm{KL}(q(z_T|x)p(z_T))}_{:=L_T} + \sum_{t=1}^{T} \underbrace{\mathbb{E}[\mathrm{KL}(q(z_{t-1}|z_t,x)\|p_\theta(z_{t-1}|z_t))]}_{:=L_{t-1}} + \underbrace{\mathbb{E}[-\log p(x|z_0)]}_{:=L_{x|z_0}}$$

This suggests a *progressive* coding algorithm based on channel simulation [Ho et al., 2020]:

- At time T, simulate $Z_T|X \sim q(z_T|X)$ by running $RCC(q(z_T|X), p(z_T), X)$, costing L_T bits.
- For time t = T-1, T-2, …, 0, simulate $Z_{t-1}|Z_t, X \sim q(z_{t-1}|Z_t, X)$ by running $RCC(q(z_{t-1}|Z_t, X), p(z_{t-1}|Z_t), (Z_t, X))$, costing L_{t-1} bits.



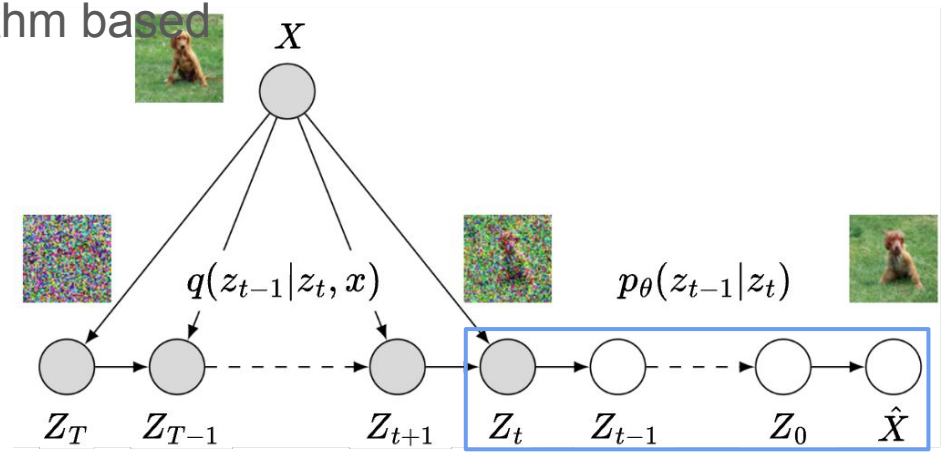
Diffusion as a hierarchical latent variable model

A discrete-time diffusion model equivalently optimizes an NELBO:

$$-\log p_\theta(x) \leq \underbrace{\mathrm{KL}(q(z_T|x)p(z_T))}_{:=L_T} + \sum_{t=1}^{T} \underbrace{\mathbb{E}[\mathrm{KL}(q(z_{t-1}|z_t,x)\|p_\theta(z_{t-1}|z_t))]}_{:=L_{t-1}} + \underbrace{\mathbb{E}[-\log p(x|z_0)]}_{:=L_{x|z_0}}$$

This suggests a *progressive* coding algorithm based on channel simulation [Ho et al., 2020]:

- At time t, the receiver will have simulated $Z_t, Z_{t+1}, ..., Z_T|X \sim q(z_{t:T}|X)$ using $\sum_{s=t-1}^{T} L_s$ bits.
- The receiver can then set $Y = Z_t$, and generate a reconstruction $\hat{X}|Y \sim P_{X|Z_t}$ using the reverse-process model.



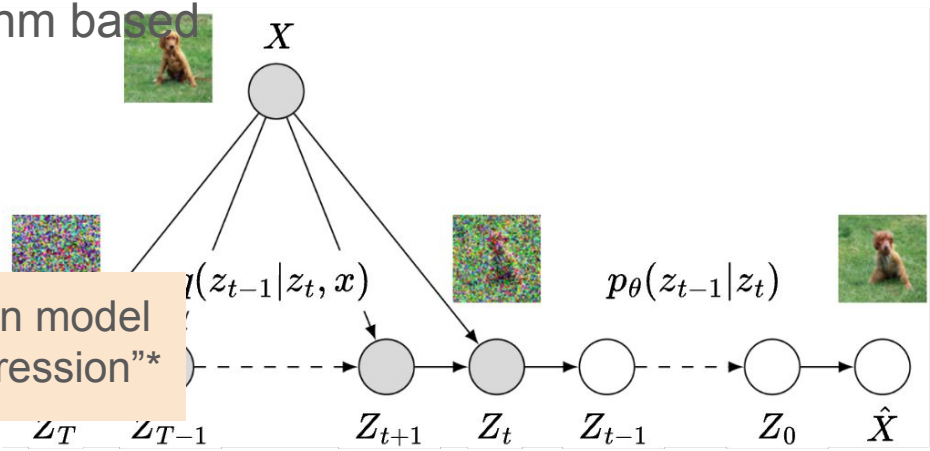
Diffusion as a hierarchical latent variable model

A discrete-time diffusion model equivalently optimizes an NELBO:

$$-\log p_\theta(x) \leq \underbrace{\mathrm{KL}(q(z_T|x)p(z_T))}_{:=L_T} + \sum_{t=1}^{T} \underbrace{\mathbb{E}[\mathrm{KL}(q(z_{t-1}|z_t,x)\|p_\theta(z_{t-1}|z_t))]}_{:=L_{t-1}} + \underbrace{\mathbb{E}[-\log p(x|z_0)]}_{:=L_{x|z_0}}$$

This suggests a *progressive* coding algorithm based on channel simulation [Ho et al., 2020]:

- The expected coding cost of this algorithm is exactly equal to the NELBO; thus

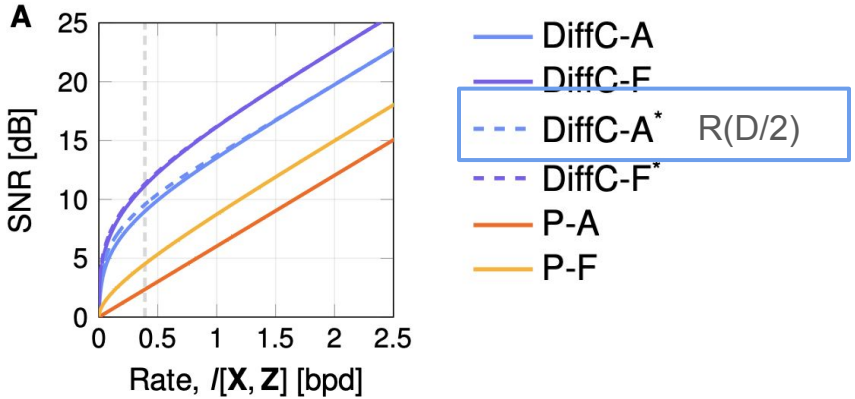


"Variational learning of a discrete-time diffusion model ≈ end-to-end optimizing for progressive compression"*

*Fine print: the connection is exact in the case of lossless compression with bits-back coding (see VDM [Kingma et al., 2021]) ; in lossy compression with channel simulation, the NELBO is only a lower bound on the actual coding cost.

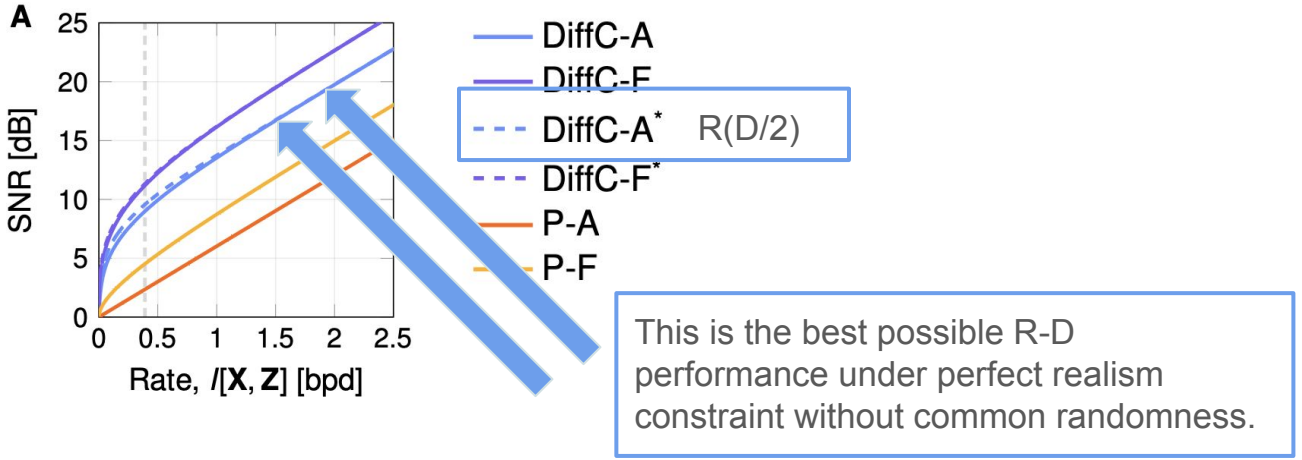
DiffC [Theis et al., 2022]

- First study of the R-D performance of this approach.
- Further optimized the forward (noising) process for R-D performance.
- Results:
  - ➢ DiffC (with optimized noising process + ODE-based reconstruction) does better than R(D/2) on Gaussian sources.



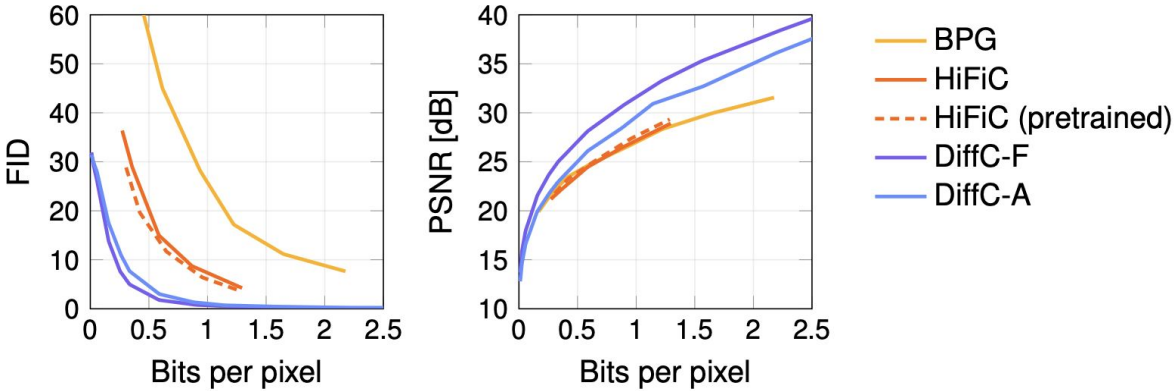
DiffC [Theis et al., 2022]

- First study of the R-D performance of this approach.
- Further optimized the forward (noising) process for R-D performance.
- Results:
  - ➢ DiffC (with optimized noising process + ODE-based reconstruction) does better than R(D/2) on Gaussian sources.



DiffC [Theis et al., 2022]

- Results:
  - ➢ DiffC significantly outperforms BPG and GAN-based HiFiC [Mentzer et al., 2020] on ImageNet 64:

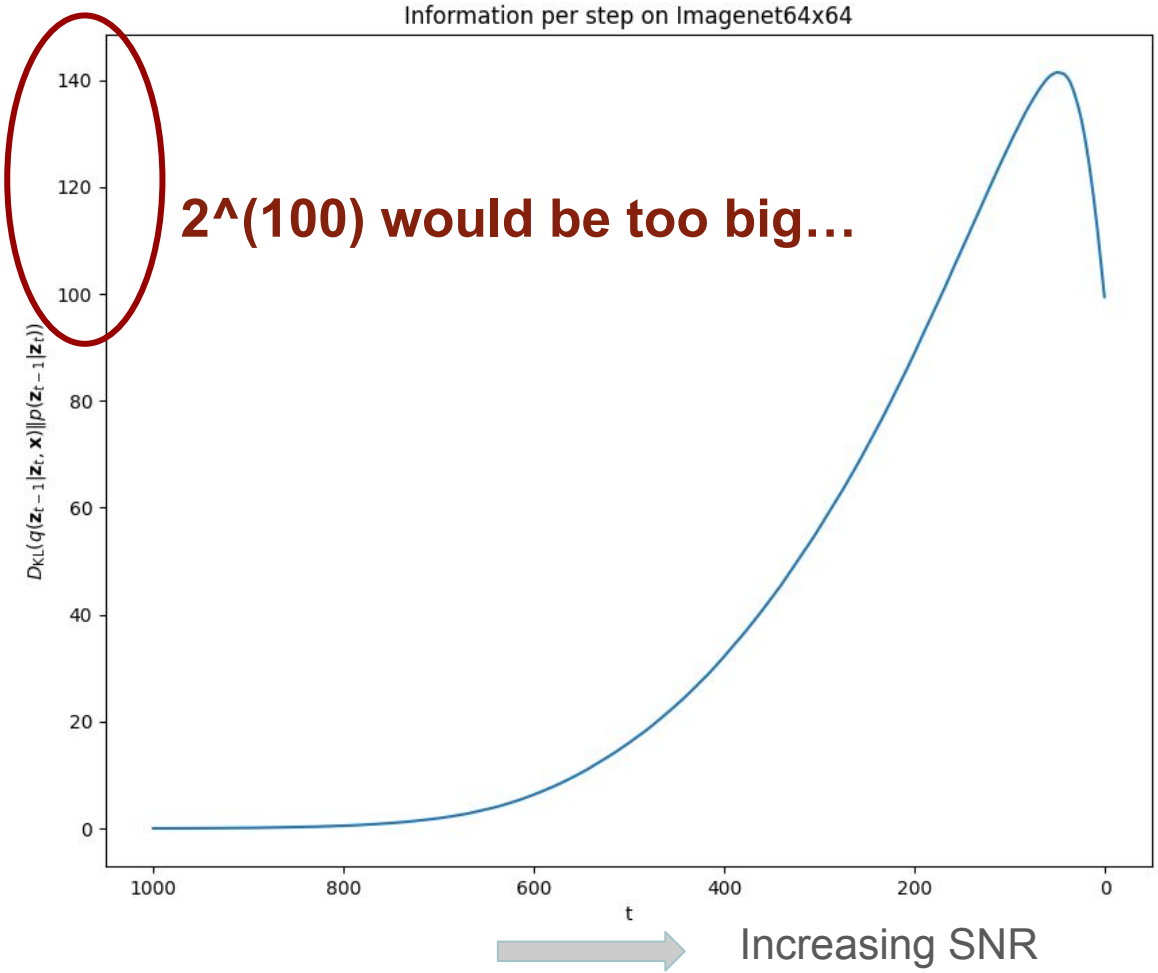


DiffC [Theis et al., 2022]

- *Caveat:* ***results are hypothetical****; naive channel simulation* (*e.g., PFR runtime ~* $\exp$*(bit-rate)* [Theis and Ahmed, 2022]) *would be too expensive.*

DiffC [

- Cave
  PFR
  expe



UQDM [Yang et al., 2025]

- We can bypass the difficulty of simulating high-dim Gaussian channels by simulating uniform-noise channels instead.
- The uniform channel can be simulated efficiently using **Universal (Dithered) Quantization** [Zamir and Feder, 1992], with optimal expected code length and running time.

UQDM [Yang et al., 2025]

Proposed forward process:

We keep the same reverse-time factorization as in Gaussian diffusion, but replace Gaussian distributions with matching uniform distributions:

$$q(z_{0:T}|x) = q(z_T|x) \prod_{t=1}^{T} q(z_{t-1}|z_t, x)$$

$$\approx \mathcal{N}(0, I),$$

keep unchanged

UQDM [Yang et al., 2025]

Proposed forward process:

We keep the same reverse-time factorization as in Gaussian diffusion, but replace Gaussian distributions with matching uniform distributions:

$$q(z_{0:T}|x) = q(z_T|x) \prod_{t=1}^{T} q(z_{t-1}|z_t, x)$$

$$\approx \mathcal{N}(0, I),$$

keep unchanged

Change from Gaussian

$$\mathcal{N}(b(t)z_t + c(t)x, \beta^2(t)I)$$

to uniform:

$$\mathcal{U}([b(t)z_t + c(t)x \pm \frac{\Delta(t)}{2}])$$

where b(t), c(t), and β(t) are specified by the Gaussian noising process, and we choose quantization width $\Delta(t) := \sqrt{12}\beta(t)$ to match the variance of Gaussian.

UQDM [Yang et al., 2025]

Proposed reverse process:

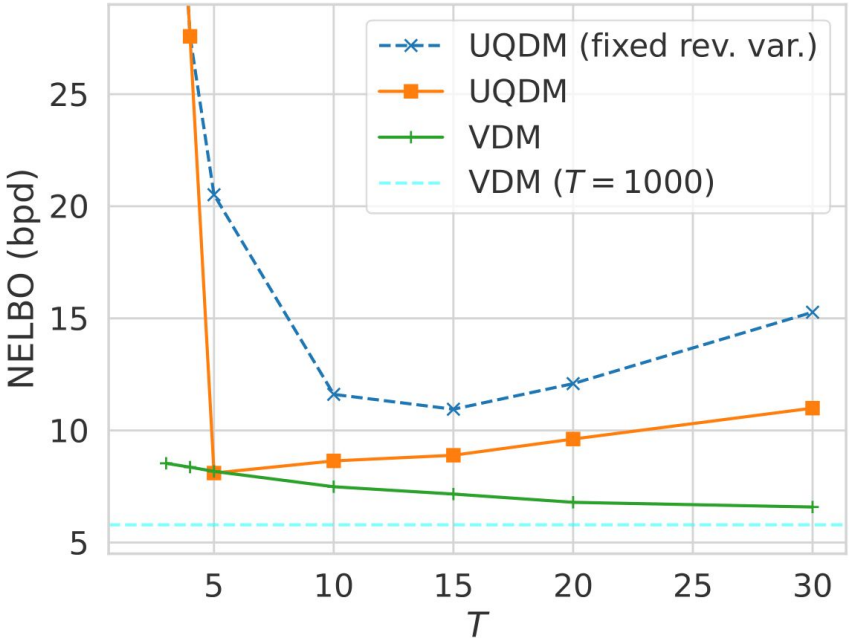
- In Gaussian diffusion, the reverse process model is chosen as

$$p_\theta(z_{t-1}|z_t) = q(z_{t-1}|z_t, x = \hat{x}_\theta(z_t))$$

- We make the same choice, except we additionally broaden it by convolving with $\mathcal{U}([-\frac{\Delta(t)}{2}, \frac{\Delta(t)}{2}])$ for entropy modeling with UQ [Ballé et al., 2018].
- We also make the reverse-process variance learnable.

UQDM [Yang et al., 2025]

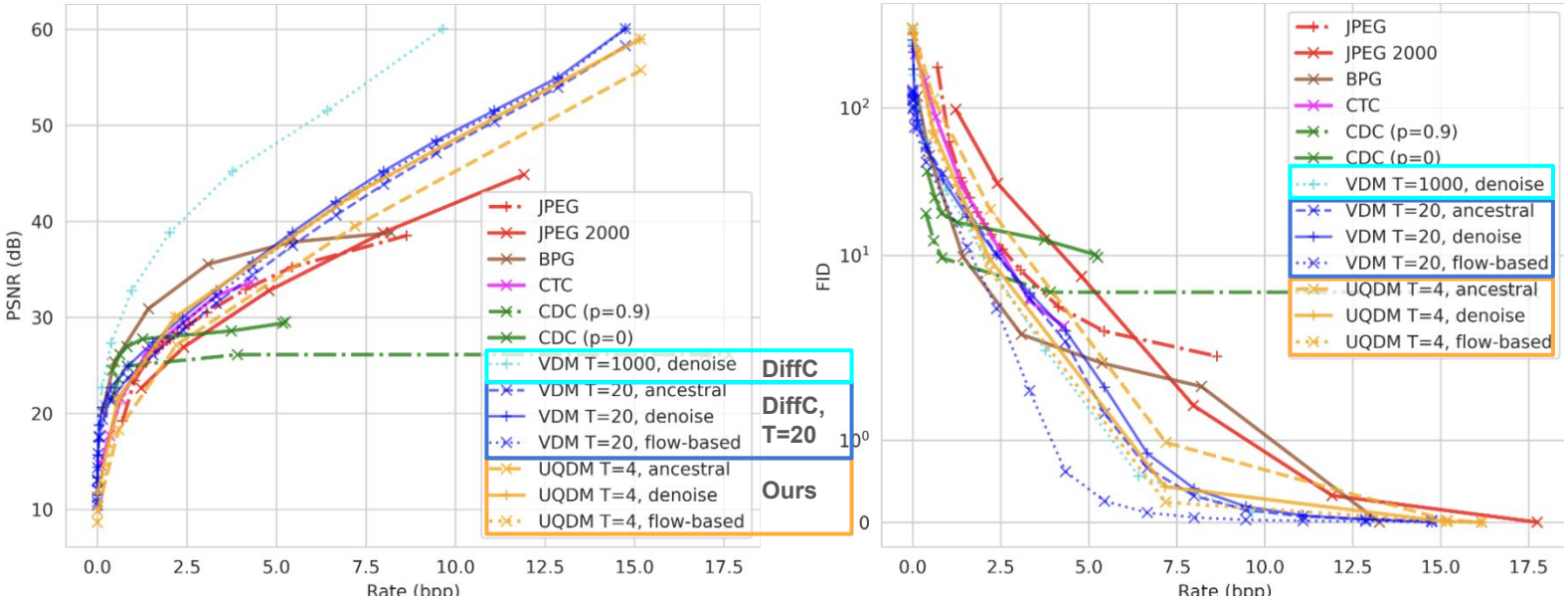
End-to-end trained NELBO v.s. num diffusion steps T:



- Empirically, the NELBO for UQDM appears to diverge with increasing T; $\exists$ optimal $T \leq 10$.
- Finetuning from Gaussian diffusion didn't help.
- Learning reverse-process variance helped a lot.

UQDM [Yang et al., 2025]

Rate-distortion and rate-realism results on ImageNet 64:



Fast encoding/decoding (as fast as evaluating NELBO), ~0.1 sec for ImageNet 64

LCPD [Vonderfecht and Liu, 2025]

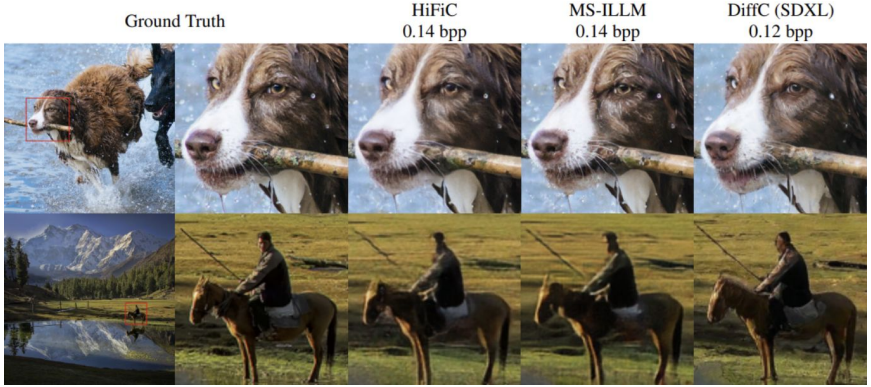
Concurrent work implements DiffC by introducing practical workarounds:

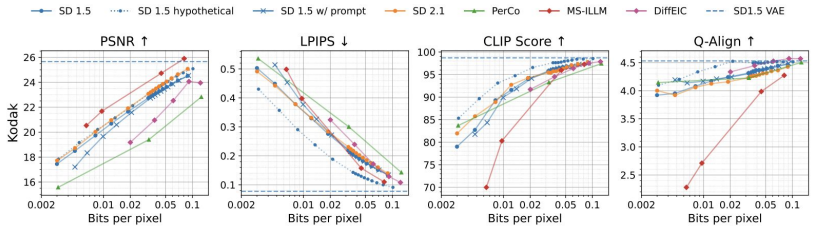
- Skipping CS steps when KL is small:

$RCC(q(z_{999}|Z_{1000}, X), p(z_{999}|Z_{1000}), (Z_{1000}, X))$❌
$RCC(q(z_{900}|Z_{1000}, X), p(z_{900}|Z_{1000}), (Z_{1000}, X))$✅

- "Chunking up" dimensions of Z_t when KL is large [Flamich et al., 2020], and use a GPU implementation of PFR.

Competitive results when applied on top of latent diffusion models.





Discussions – design choices

Type of conditioning (choice of Y):

- Text/language is insufficient for preserving spatial information [Lei et al., 2023].
- Conditioning on a pixel-space reconstruction can work well in theory [Yan et al., 2021], [Hoogeboom et al., 2023], rivaling end-to-end training [Yang & Mandt, 2023].

Pixel-space v.s. latent-space diffusion:

Pixel-space diffusion	**Latent diffusion**
✅Doesn't require training a separate autoencoder; theoretically more optimal.	❌Requires training a separate autoencoder; suboptimal in theory.
✅Yields better control over R-D-P tradeoff.	❌Performance limited by the pre-trained autoencoder.
❌Computationally more expensive.	✅Computationally cheaper
❌Patch-based reconstruction may not be suitable for very low bit-rates.	✅Tends to offer better realism (globally coherent reconstructions) at very low rates.

Discussions – open problems

High computational cost.

- SOTA methods still require on the order of ~10 iterative denoising steps.

Table 2: Encoding and Decoding Speed (in seconds). Kodak

Model	Encoding Speed (in sec.)	Decoding Speed (in sec.)
VTM	16.892 ± 7.574	0.135 ± 0.002
MS-ILLM	0.084 ± 0.021	0.080 ± 0.007
Text-Sketch (PIC)	163.070 ± 0.380	2.725 ± 0.012
Text-Sketch (PICS)	190.231 ± 2.476	19.288 ± 0.251
PerCo - 5 denoising steps	0.080 ± 0.018	0.665 ± 0.009
PerCo - 20 denoising steps	0.080 ± 0.018	2.551 ± 0.018

[Careil et al., 2023]

- Few-step sampling is an active research topic; some solutions: higher-order samplers [Liu 2022, Jolicoeur-Martineau 2021], model distillation [Salimans & Ho 2022], consistency models [Song and Dhariwal, 2023], etc.

Discussions – open problems

Loss of fine details.

HFD [Hoogeboom et al., 2023]

Non-linear transform coding (PO-ELIC), [He et al., 2022]

Original

Discussions – open problems

Difficulty with *evaluation* in the very low bit-rate regime.

- As generation performance improves, distortion metrics such as PSNR and MS-SSIM (even LPIPS) are no longer informative, nor is FID [Careil et al., 2023].
- Many plausible and realistic reconstructions exist – which is the “best”?

Discussions – open problems

Difficulty with *evaluation* in the very low bit-rate regime.

- As generation performance improves, distortion metrics such as PSNR and MS-SSIM (even LPIPS) are no longer informative, nor is FID [Careil et al., 2023].
- Many plausible and realistic reconstructions exist – which is the “best”?
- Some possible solutions:
  - Human assessment [Mentzer et al., 2020].
  - Measure the preservation of semantic / spatial information, e.g., using CLIP / IoU scores [Careil et al., 2023].
  - Alternative distortion/realism metrics, see e.g. Wasserstein distortion [Qiu et al., 2024].

Discussions – open problems

Effectively communicating information with diffusion.

- Random coding / channel simulation can potentially outperform deterministic coding (“compress + refine”) [Theis and Agustsson, 2021], [Theis et al., 2022].
- To achieve this, we can
  - Develop more efficient channel simulation schemes [Vonderfecht and Liu, 2025].
  - Consider non-Gaussian channel [Yang et al., 2025] – would be useful to have I-MMSE relations analogous to the Gaussian case [Guo et al., 2005; Kong et al., 2023].

Thank you! Q & A

References

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., & Johnston, N. (2018). Variational Image Compression with a Scale Hyperprior. *ICLR*.

Blau, Y., & Michaeli, T. (2019). Rethinking lossy compression: The rate-distortion-perception tradeoff. *International Conference on Machine Learning*, 675–685.

Careil, M., Muckley, M. J., Verbeek, J., & Lathuilière, S. (2023). Towards image compression with perfect realism at ultra-low bitrates. *The Twelfth International Conference on Learning Representations*.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., & Ye, J. C. (2023). Diffusion Posterior Sampling for General Noisy Inverse Problems. *International Conference on Learning Representations*. https://openreview.net/forum?id=OnD9zGAGT0k

Daras, G., Chung, H., Lai, C.-H., Mitsufuji, Y., Ye, J. C., Milanfar, P., Dimakis, A. G., & Delbracio, M. (2024). *A Survey on Diffusion Models for Inverse Problems*. https://arxiv.org/abs/2410.00083

References

Flamich, G., Havasi, M., & Hernández-Lobato, J. M. (2020). Compressing Images by Encoding Their Latent Representations with Relative Entropy Coding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 16131–16141).

Ghouse, N. F. K. M., Petersen, J., Wiggers, A. J., Xu, T., & Sautiere, G. (2023). *Neural image compression with a diffusion-based decoder*.

Guo, D., Shamai, S., & Verdú, S. (2005). Mutual Information and Minimum Mean-Square Error in Gaussian Channels. *IEEE Transactions on Information Theory*, *51*(4), 1261–1282.

Gushchin, N., Kholkin, S., Burnaev, E., & Korotin, A. (2024). Light and Optimal Schrödinger Bridge Matching. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 17100–17122). PMLR. https://proceedings.mlr.press/v235/gushchin24a.html

References

He, D., Yang, Z., Yu, H., Xu, T., Luo, J., Chen, Y., Gao, C., Shi, X., Qin, H., & Wang, Y. (2022). Po-elic: Perception-oriented efficient learned image coding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1764–1769.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv Preprint arXiv:2207.12598*.

Hoogeboom, E., Agustsson, E., Mentzer, F., Versari, L., Toderici, G., & Theis, L. (2023). High-fidelity image compression with score-based generative models. *arXiv Preprint arXiv:2305.18231*.

Jolicoeur-Martineau, A., Li, K., Piché-Taillefer, R., Kachman, T., & Mitliagkas, I. (2021). Gotta go fast when generating data with score-based models. *arXiv Preprint arXiv:2105.14080*.

References

Kawar, B., Elad, M., Ermon, S., & Song, J. (2022). Denoising Diffusion Restoration Models. *Advances in Neural Information Processing Systems*.

Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. *Advances in Neural Information Processing Systems*, *34*, 21696–21707.

Kong, X., Brekelmans, R., & Steeg, G. V. (2023). Information-Theoretic Diffusion. *International Conference on Learning Representations*. https://arxiv.org/abs/2302.03792

Lei, E., Uslu, Y. B., Hassani, H., & Bidokhti, S. S. (2023). Text+ sketch: Image compression at ultra low rates. *arXiv Preprint arXiv:2307.01944*.

Li, C. T. (2024). Channel simulation: Theory and applications to lossy compression and differential privacy. *Foundations and Trends® in Communications and Information Theory*, *21*(6), 847–1106.

References

Li, C. T., & Anantharam, V. (2021). A Unified Framework for One-Shot Achievability via the Poisson Matching Lemma. *IEEE Transactions on Information Theory*, *67*(5), 2624–2651. https://doi.org/10.1109/TIT.2021.3058842

Li, H., Yang, Y., Chang, M., Feng, H., Xu, Z., Li, Q., & Chen, Y. (2021). *SRDiff: Single Image Super-Resolution with Diffusion Probabilistic Models*. https://arxiv.org/abs/2104.14951

Liu, L., Ren, Y., Lin, Z., & Zhao, Z. (2022). Pseudo numerical methods for diffusion models on manifolds. *arXiv Preprint arXiv:2202.09778*.

Lui, K. Y. C., Cao, Y., Gazeau, M., & Zhang, K. S. (2017). Implicit manifold learning on generative adversarial networks. *arXiv Preprint arXiv:1710.11260*.

Mentzer, F., Toderici, G. D., Tschannen, M., & Agustsson, E. (2020). High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, *33*, 11913–11924.

References

Qiu, Y., Wagner, A. B., Ballé, J., & Theis, L. (2024). Wasserstein distortion: Unifying fidelity and realism. *2024 58th Annual Conference on Information Sciences and Systems (CISS)*, 1–6.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., & Norouzi, M. (2022). Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Salimans, T., & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. *arXiv Preprint arXiv:2202.00512*.

Song, Y., & Dhariwal, P. (2023). Improved techniques for training consistency models. *arXiv Preprint arXiv:2310.14189*.

Theis, L., & Agustsson, E. (2021). On the advantages of stochastic encoders. *arXiv Preprint arXiv:2102.09270*.

Theis, L., & Ahmed, N. Y. (2022). Algorithms for the communication of samples. *International Conference on Machine Learning*, 21308–21328.

Theis, L., Salimans, T., Hoffman, M. D., & Mentzer, F. (2022). Lossy compression with gaussian diffusion. *arXiv Preprint arXiv:2206.08889*.

References

Vonderfecht, J., & Liu, F. (2025). Lossy Compression with Pretrained Diffusion Models. *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=raUnLe0Z04

Yan, Z., Wen, F., Ying, R., Ma, C., & Liu, P. (2021). On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework. *International Conference on Machine Learning*, 11682–11692.

Yang, R., & Mandt, S. (2023). Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, *36*, 64971–64995.

Yang, Y., Will, J., & Mandt, S. (2025). Progressive Compression with Universally Quantized Diffusion Models. *International Conference on Learning Representations*.

Zamir, R., & Feder, M. (1992). On Universal Quantization by Randomized Uniform/Lattice Quantizers. *IEEE Transactions on Information Theory*, 428–436.

Zhang, G., Qian, J., Chen, J., & Khisti, A. (2021). Universal rate-distortion-perception representations for lossy compression. *Advances in Neural Information Processing Systems*, *34*, 11517–11529.