Diffusion Modelのtext-to-Imageへの応用（GLIDE）

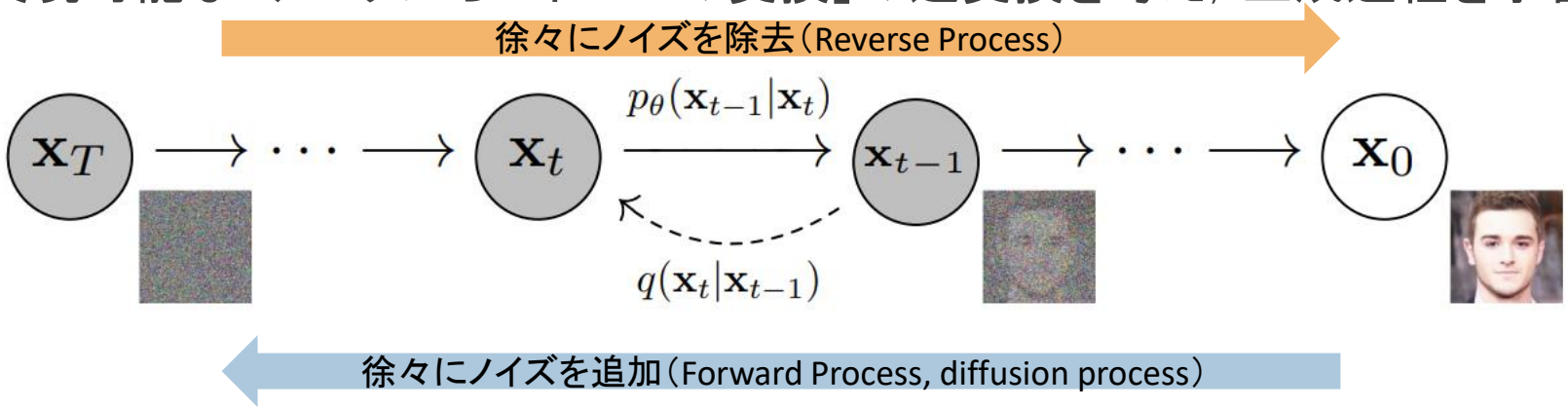
小林研修士1年 海老澤優

GLIDE

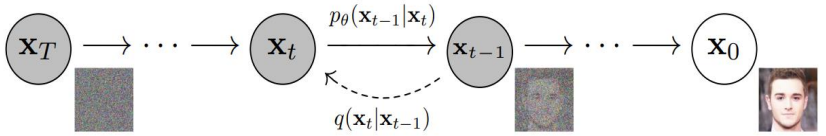
- GLIDE(Guided Language to Image Diffusion for Generation and Editing)
- 2022年にOpenAIによって発表.
- Diffusion modelを利用してテキスト情報から画像を生成するモデル
- OpenAIが以前に提案したDALL-Eよりも高性能なモデル
  画像の質，キャプションと画像の関連性

DDPMのおさらい

- 拡散過程に基づく生成モデル
- 一般にノイズを複雑なデータに変換するのは難しい
  →容易に実現可能な「データからノイズへの変換」の逆変換を考え，生成過程を学習する.



- 物理現象のアナロジーで時刻の概念を導入（時刻0がデータ，時刻Tが完全なノイズ）
- 複雑な分布x_0を簡単な分布x_Tに変換するマルコフ連鎖$q(x_t|x_{t-1})$を定義.
- この逆変換を与える過程$p_\theta(x_{t-1}|x_t)$を学習し，簡単な分布x_Tを用いて複雑なデータ分布のサンプリングを実現した.



Forward Process（Diffusion Process）

- 各時刻において，前時刻のデータを減衰させ，ノイズを付加する．つまり，

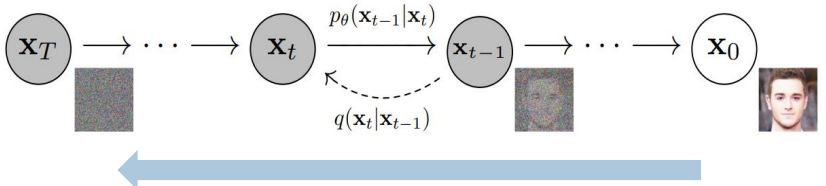
$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon_{t-1} \quad \text{ただし，} \varepsilon_{t-1} \sim N(\mathbf{0}, \mathbf{I}), 0 < \beta_t < 1$$

- β_tは各時刻で付加するノイズの強度を表すハイパーパラメータ．
時間が経つにつれノイズが大きくなるようにしたいので$\beta_1 < \beta_2 < \cdots < \beta_T$を仮定

- 今後のために条件付き確率の形で書くと，

$$q(x_t|x_{t-1}) \coloneqq \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right)$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \text{ ただし，} (x_{1:T} = \{x_t\}_{i=1}^{T})$$



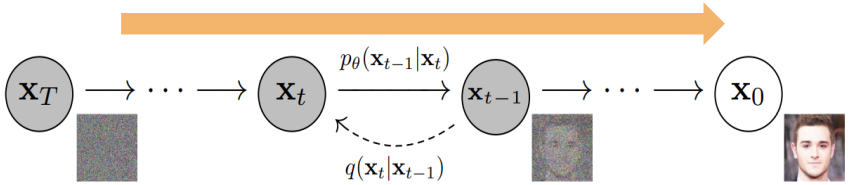
Forward Process（Diffusion Process）

- このプロセスが優れている点は，任意の時刻tにおけるx_tを閉形式（加減乗除と初等関数の合成関数による解の表し方）で表せる点．
- データx_0が与えられたときのx_tは，$\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$とすると，

$$
\begin{aligned}
x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon_{t-2}\right) + \sqrt{1-\alpha_t}\varepsilon_{t-1} \\
&= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\varepsilon}_{t-2} \quad (*) \\
&= \cdots \\
&= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon
\end{aligned}
$$

時刻tのデータ=データx_0とノイズの重み付き和

- (*)は$X \sim N(\mathbf{0}, \sigma_1^2\mathbf{I}), Y \sim N(\mathbf{0}, \sigma_2^2\mathbf{I})$で互いに独立のとき，$X + Y \sim N(\mathbf{0}, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$となることを利用．



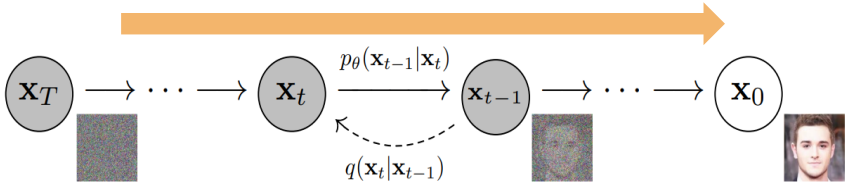
Reverse Process（データの生成）

- データ生成の過程はForward Processの逆変換として定義できる.
- 微小時刻間の変化が十分小さいcontinuous diffusionにおいて，その逆過程も同様の関数系で表せることが知られている.（Kolmogorov equation）
- よって，β_tを十分小さくとることで，逆方向の条件付き確率$q(x_{t-1}|x_t)$もガウス分布になる.
  →平均と分散をモデルで推定すればよい.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

- ここでは$\Sigma_\theta(x_t, t) = \sigma_t^2 \mathbf{I}$と単純化して議論する.（$\sigma_t^2$はハイパーパラメータ）
  よって，$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z$ $(z \sim N(0, \mathbf{I}))$で$x_t$から$x_{t-1}$をサンプルできる.



Reverse Process（データの生成）

- データ生成の過程はForward Processの逆変換として定義できる.
- 微小時刻間の変化が十分小さいcontinuous diffusionにおいて，その逆過程も同様の関数系で表せることが知られている.（Kolmogorov equation）
- よって，β_tを十分小さくとることで，逆方向の条件付き確率$q(x_{t-1}|x_t)$もガウス分布になる.
  →平均と分散をモデルで推定すればよい.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

- ここでは$\Sigma_\theta(x_t, t) = \sigma_t^2 \mathbf{I}$と単純化して議論する.（$\sigma_t^2$はハイパーパラメータ）
  よって，$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z$ $(z \sim N(0, \mathbf{I}))$で$x_t$から$x_{t-1}$をサンプルできる.

Diffusion modelの損失関数

- 損失関数は

$$L_{VLB} = E_q\left[\underbrace{D_{KL}(q(x_T|x_0)||p_\theta(x_T))}_{L_T} + \sum_{t\geq 2}\underbrace{D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0|x_1)}_{L_0}\right]$$

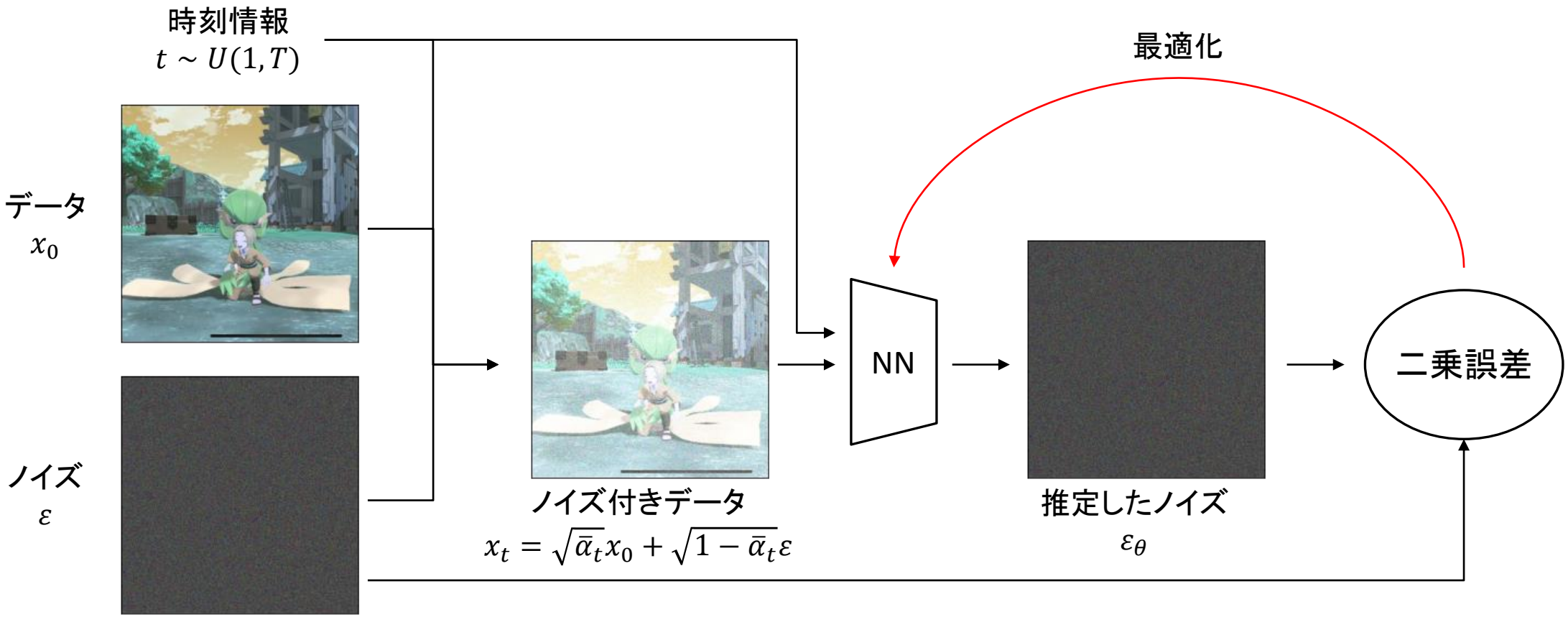
- DDPMでは最終的にこれを簡略化した損失関数として以下を用いた.

$$L_{\text{simple}}(\theta) = E_{t,x_0,\varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t\right)\right\|^2\right]$$

学習の概要

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged



Σ_θの学習

- DDPMではΣ_θを固定していたが，ステップ数が少ない場合は分散も学習させた方が精度がよくなることがわかっている（Improved Denoising Diffusion Probabilistic Models.2021）
- Σ_θを直接学習するのではなく，

$$\Sigma_\theta(x_t, t) = \exp\left(v \log \beta_t \mathbf{I} + (1 - v) \log \tilde{\beta}_t \mathbf{I}\right)$$

のvを学習するようにする．

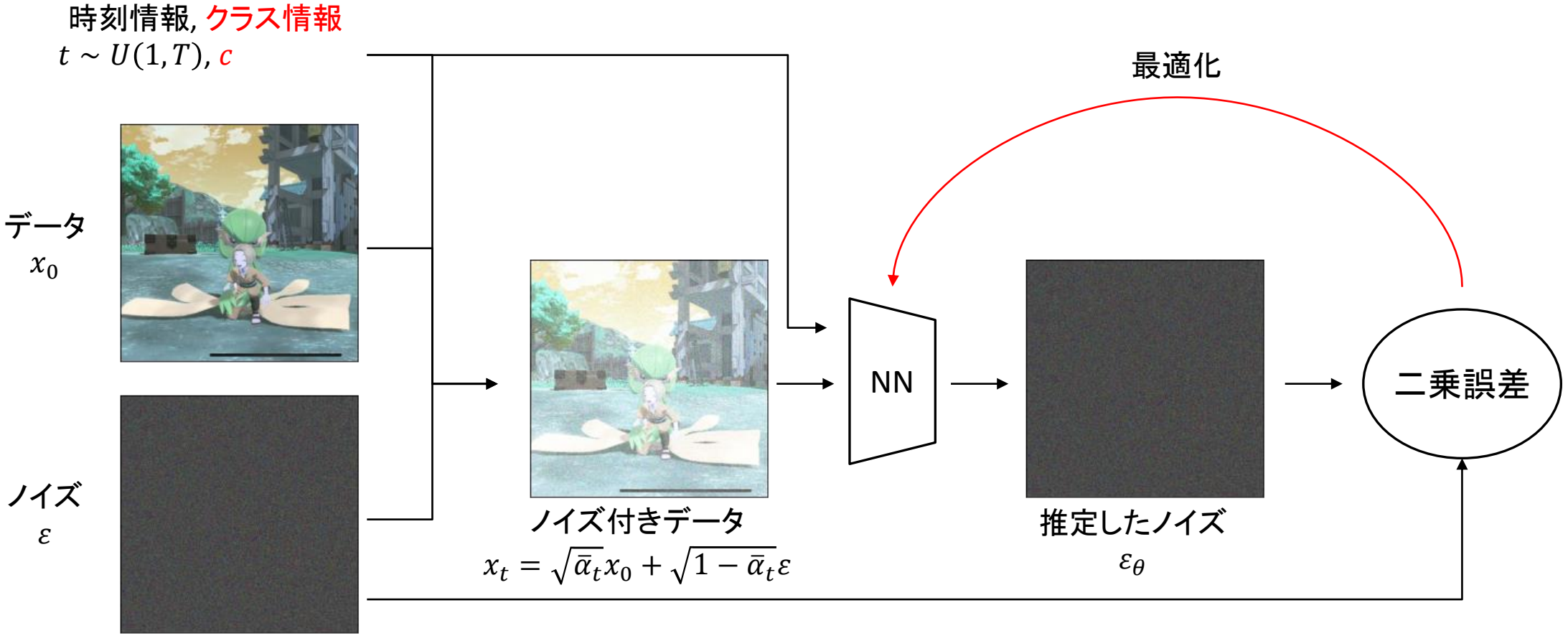
$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t \quad (q(x_{t-1}|x_t, x_0)\text{の分散項})$$

- 分散の情報も用いるので，目的関数を

$$L_{\text{simple}} + \lambda L_{\text{VLB}}$$

に変更．λはハイパーパラメータ．

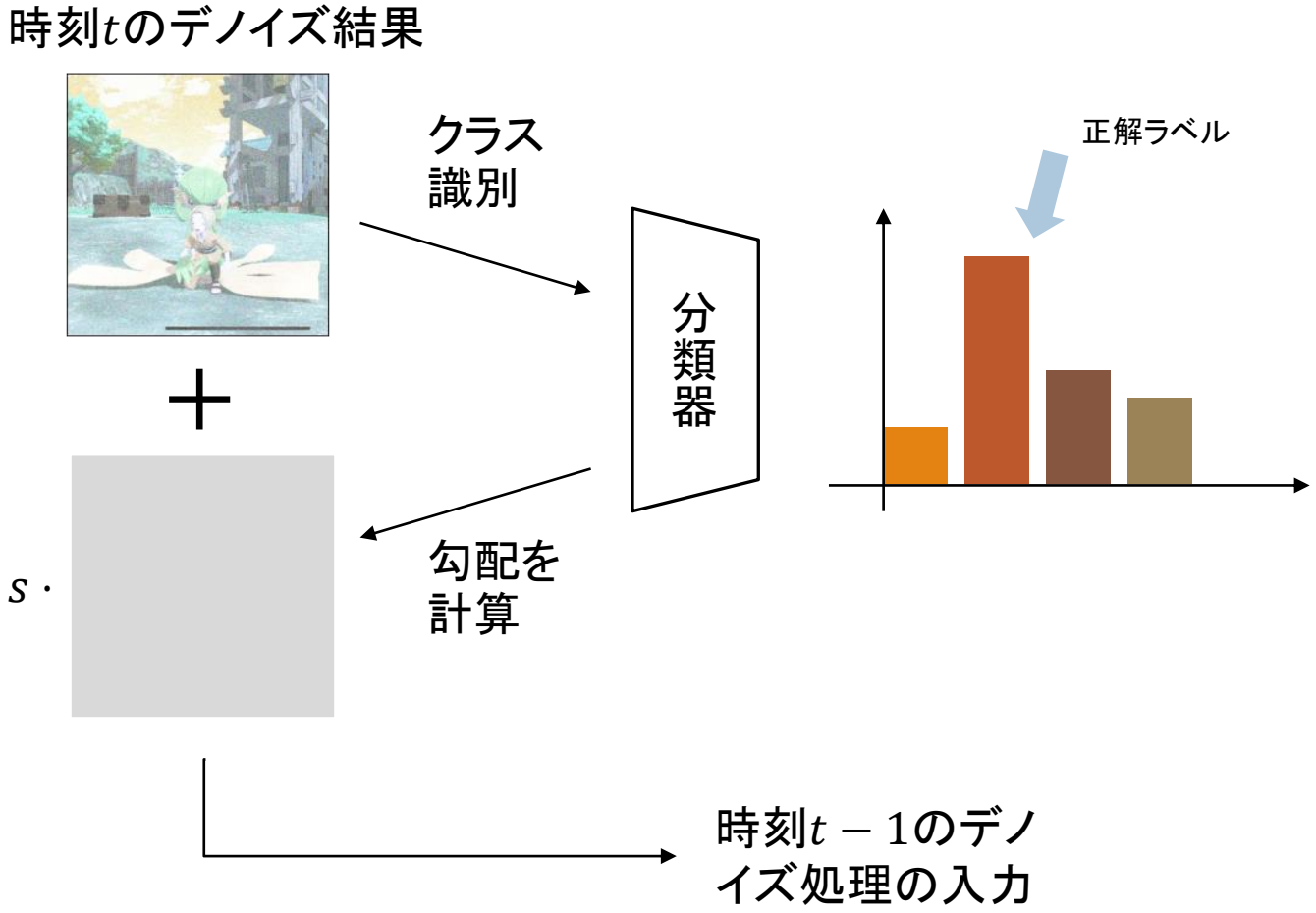
Class Conditional DDPM



Classifier Guidance

- 識別器を用いて各時刻のでノイズ結果をクラス情報が大きく反映させるようにずらし，本物画像らしくなるように誘導する.
- 別途学習した分類モデル$p_\phi(y|x_t)$の勾配$\nabla_{x_t}\log p_\phi(y|x_t)$を利用.
- 通常のdiffusion modelでは$p_\theta(x_{t-1}|x_t)=\mathcal{N}(x_{t-1};\mu_\theta(x_t|y),\Sigma_\theta(x_t|y))$に従ってデータを生成するが，Classifier Guidanceでは
$$\hat{\mu}_\theta(x_t|y)=\mu_\theta(x_t|y)+s\cdot\Sigma_\theta(x_t|y)\nabla_{x_t}\log p_\phi(y|x_t)$$
と更新することでラベルにあった画像が生成できるように誘導する.
- sはスケールパラメータといい，ノイズ結果をずらす強さを指定するパラメータ.
- 綺麗な画像だけでなく，ノイズの混ざった画像も分類できるような分類器を別途用意しなければならないのが欠点.

Classifier Guidance

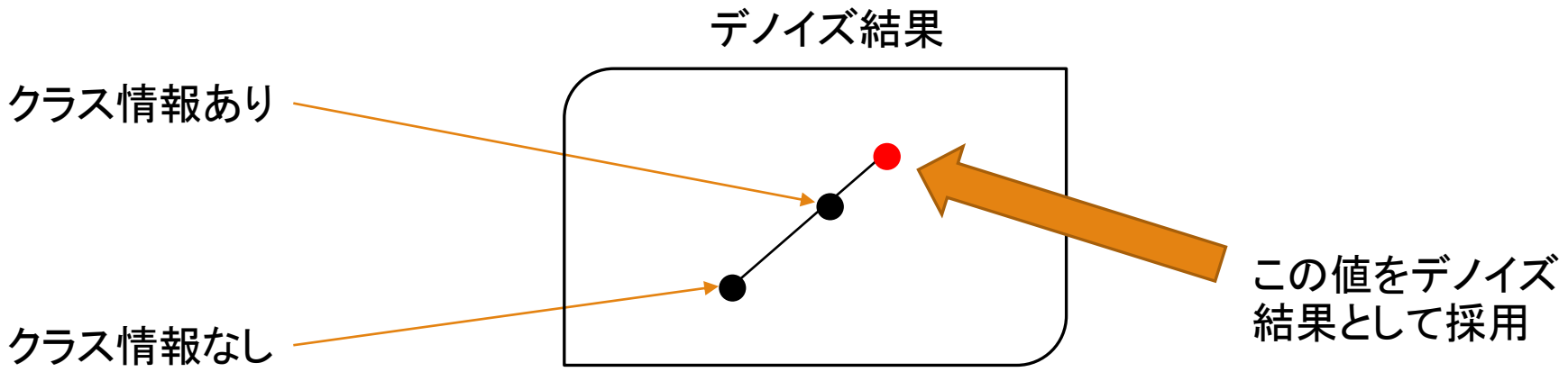


Classifier-free guidance

- Class Conditional DDPMのみを用いて，同じタスクを実行.
- Class情報なしでもDDPMを学習し，その差を利用.（一定確率でClass情報をnullにして学習）
- 時刻tのデノイズ結果を

$$\hat{\varepsilon}_\theta(x_t|y) = \varepsilon_\theta(x_t|\emptyset) + s \cdot \left(\varepsilon_\theta(x_t|y) - \varepsilon_\theta(x_t|\emptyset)\right)$$

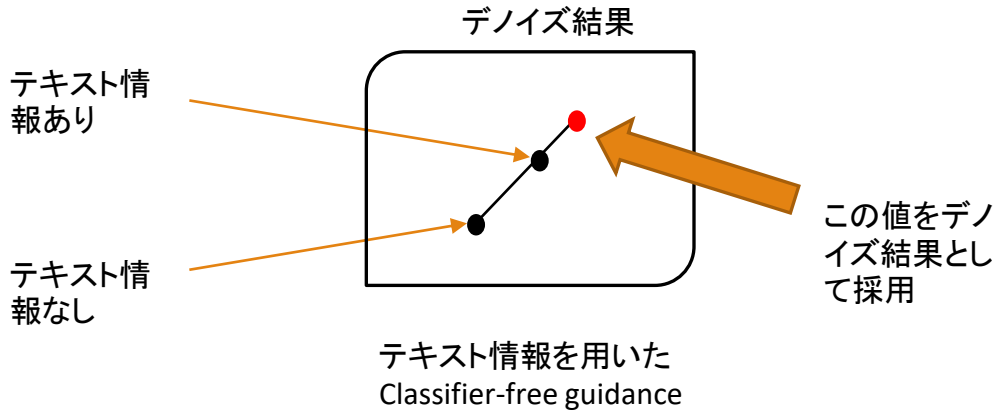
とし，$\varepsilon_\theta(x_t|y)$の方向に誘導する.



GLIDE Text-to-Imageへ

- クラス情報の代わりにテキスト情報を用いる.
  →基本的な考え方はConditional DDPMと変わらない

- Conditional DDPMでクラス情報が用いられていた場面.
  - 学習時，時刻情報と同時にモデルに入力
    →クラス情報をテキスト情報に置き換えれば学習可能
  - 生成時，Classifier guidanceで利用.
    →工夫が必要.（ほぼ同じだが）
    ① テキスト情報を用いたClassifier-free guidance
    ② CLIP guidance (CLIPの類似度を分類器の代わりに利用)

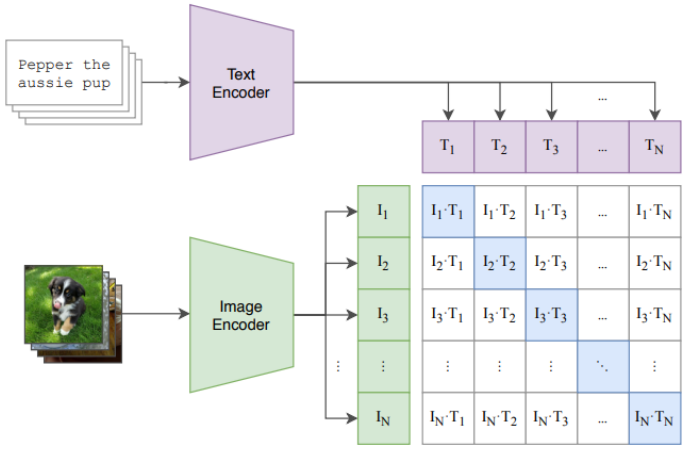


CLIP guidance

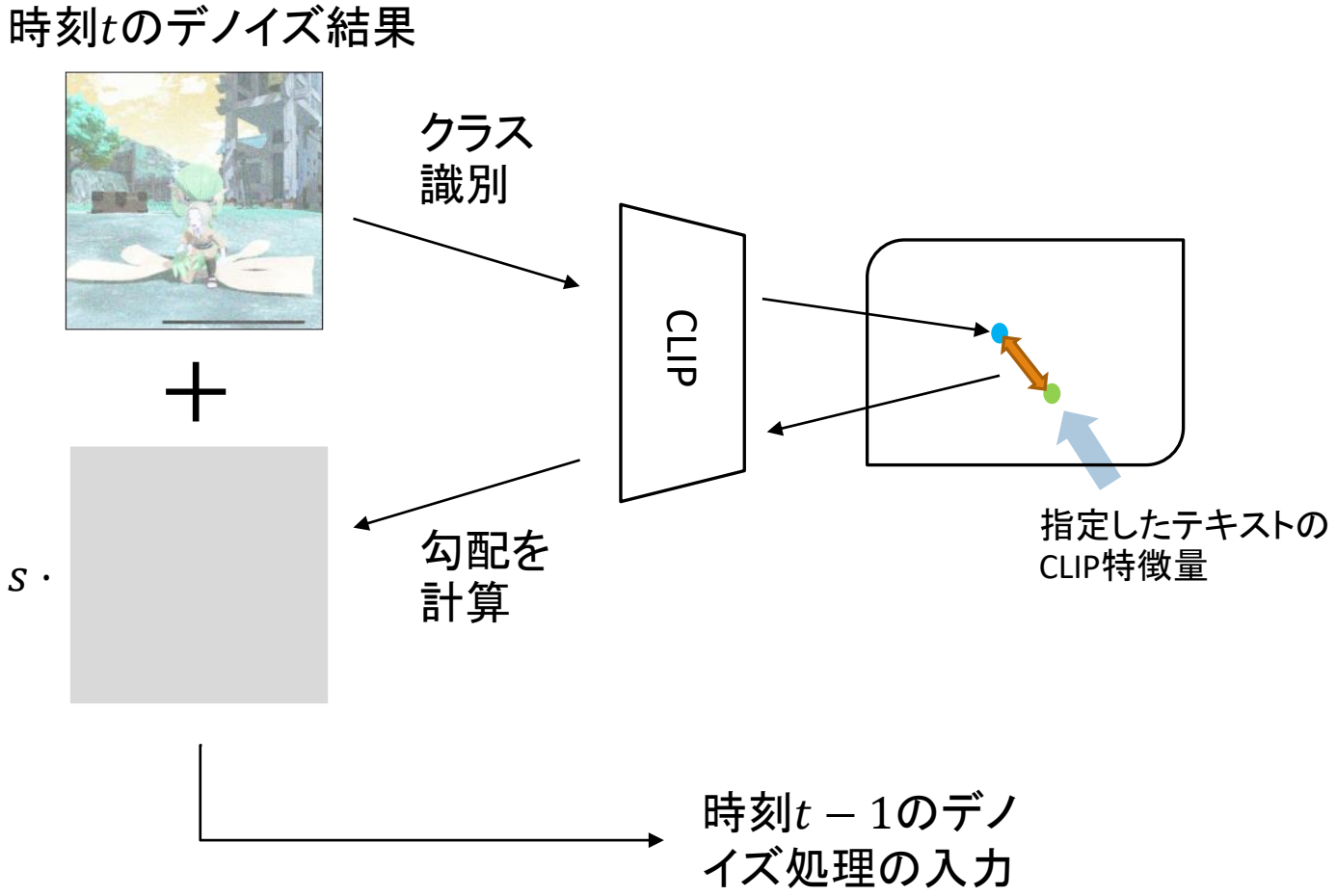
- 通常の分類モデルは任意のキャプションには対応していないので，その問題が解決できるようCLIPを利用
- CLIPは画像とテキストを直接比較できるように特徴量を変換する.
- 画像x_tとText cをそれぞれ同次元の空間$f(x_t), g(c)$に埋め込み，もともとペアだった場所のコサイン類似度（内積）が大きくなるように学習する.
- データ生成の際，平均を

$$\hat{\mu}_\theta(x_t|y) = \mu_\theta(x_t|y) + s \cdot \Sigma_\theta(x_t|y)\nabla_{x_t}\big(f(x_t), g(c)\big)$$

で更新する.



CLIP guidance



GLIDEの実験

- Diffusion Model
  DDPMを改良したADM(Ablated Diffusion Model)を利用
- テキスト情報はTransformerでADMのインプットになるよう変換
- データセット
  インターネットから収集した2億5千万のテキスト-画像ペア
- サンプリング
  Classifier-free guidanceとCLIP guidanceを別々に利用
- 学習データのうち20%について、キャプションをϕに置き換え，無条件で画像を生成することができるように、ファインチューニングを行う

性能

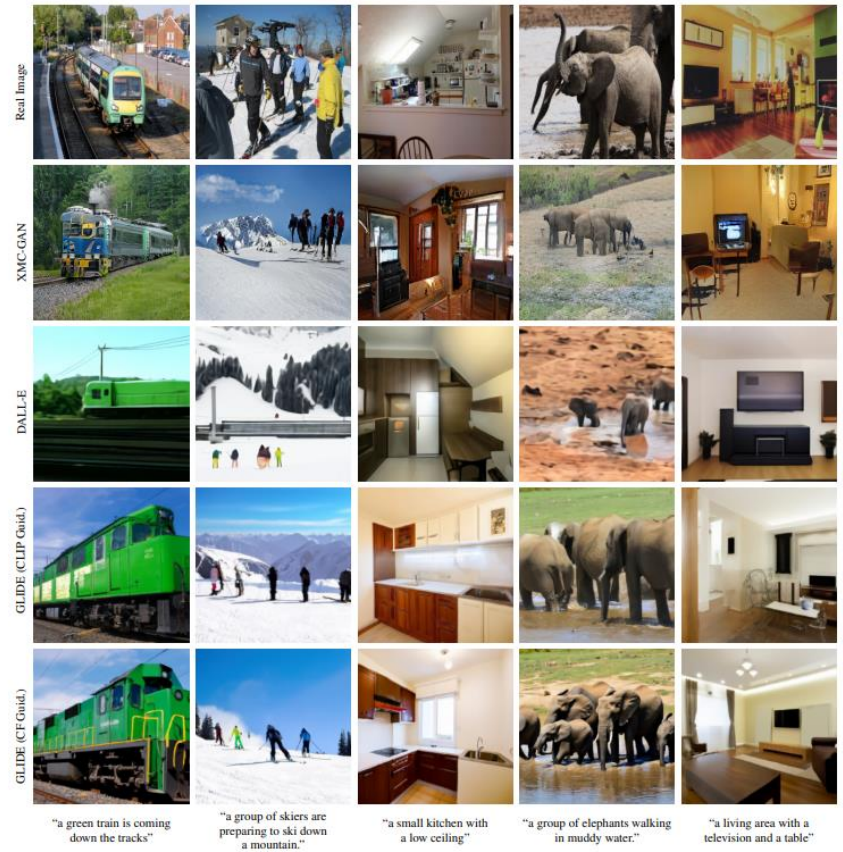


Figure 5. Random image samples on MS-COCO prompts. For XMC-GAN, we take samples from Zhang et al. (2021). For DALL-E, we generate samples at temperature 0.85 and select the best of 256 using CLIP reranking. For GLIDE, we use CLIP guidance with scale 2.0 and classifier-free guidance with scale 3.0. We do not perform any CLIP reranking or cherry-picking for GLIDE.

Guidance	Photorealism	Caption
Unguided	-88.6	-106.2
CLIP guidance	-73.2	29.3
Classifier-free guidance	**82.7**	**110.9**

- GLIDEでは繊細で精度の高い画像が生成されている.
- 定量的にはClassifier-free guidanceの方が優秀

画像の一部を変換

- Fine-tuningすることで画像を局所的に変更することも可能.
- 入力を元画像と変更したい領域のマスク画像になるようにfine-tuneした.

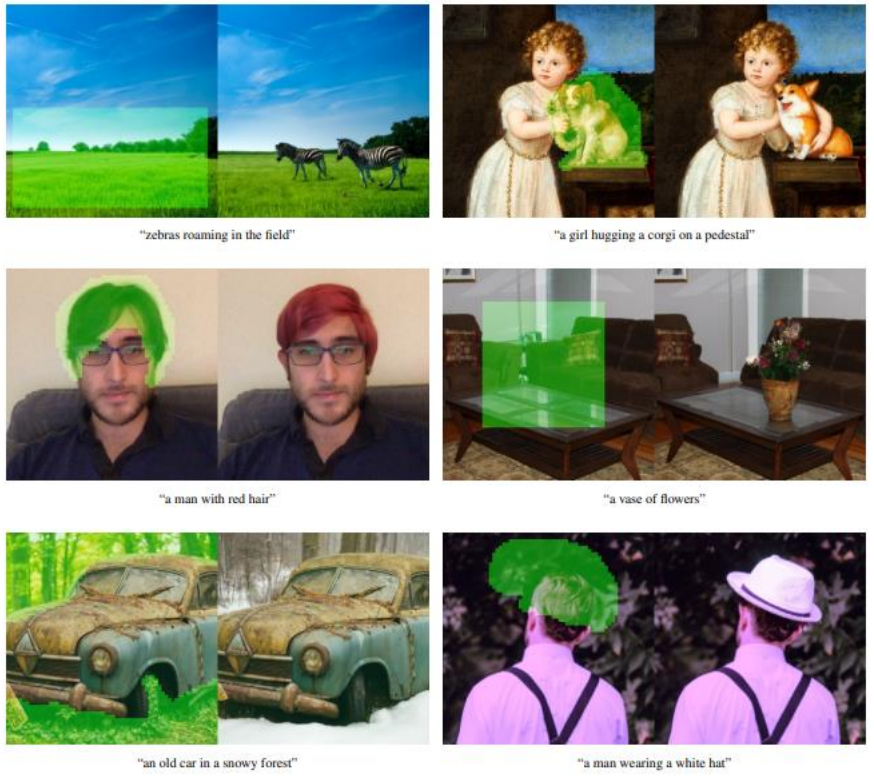


Figure 2. Text-conditional image inpainting examples from GLIDE. The green region is erased, and the model fills it in conditioned on the given prompt. Our model is able to match the style and lighting of the surrounding context to produce a realistic completion.

参考文献

- Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." *arXiv preprint arXiv:2112.10741* (2021).
- Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *Advances in Neural Information Processing Systems* 34 (2021): 8780-8794.
- Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." *arXiv preprint arXiv:2207.12598* (2022).

GLIDE(filterd)の実装

100% 100/100 [00:53<00:00, 1.84it/s]

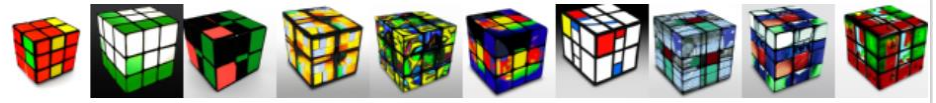
a hedgehog using a calculator

100% 100/100 [00:53<00:00, 1.85it/s]

an illustration of Pokemon

100% 100/100 [00:53<00:00, 1.85it/s]

a photo of Rubik's Cube



100% 100/100 [00:53<00:00, 1.83it/s]

a photo of Shibuya

100% 100/100 [01:07<00:00, 1.68it/s]

a blonde hair woman

100% 100/100 [01:01<00:00, 1.70it/s]

an illustration of apple