画像生成における倫理の問題

～著作権，プライバシー，偏見～

東京工業大学工学院
准教授　川上　玲
reikawa@sc.e.titech.ac.jp

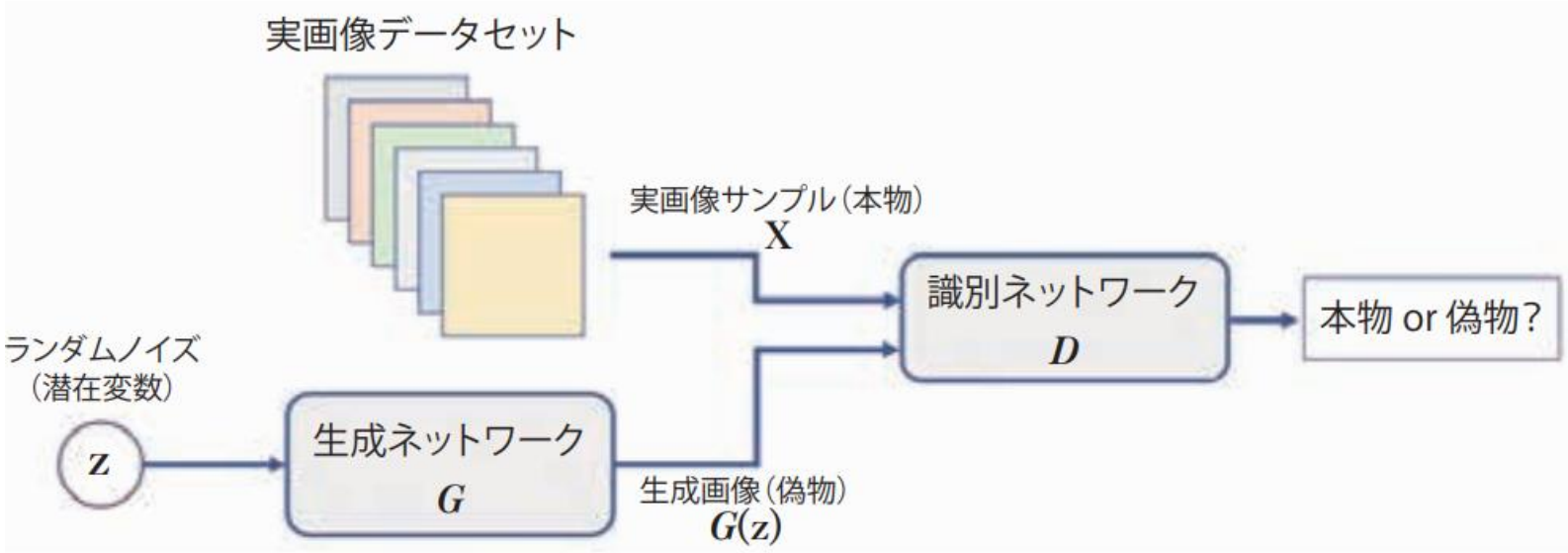
令和6年3月14日（木）
電気学会全国大会

画像生成AIの技術的進展

Generative Adversarial Network (GAN)

- 生成ネットワークGと識別ネットワークDを交互に学習
  - G：ガウスノイズから画像を生成
  - D：生成された画像と実画像を識別

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$



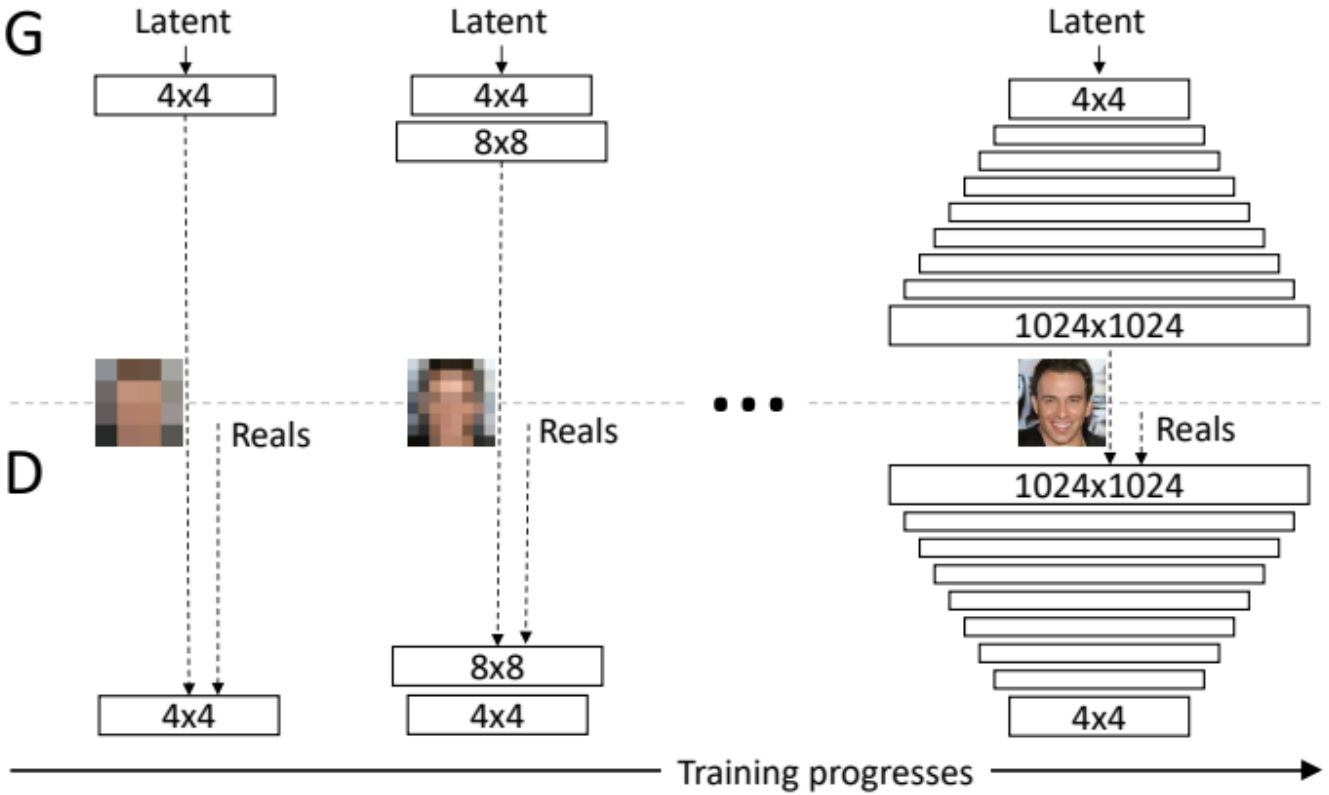
GAN
(2014)

[I. Goodfellow et al. Generative Adversarial Networks, 2014]

[中山 英樹, 5分で分かる!? 有名論文ナナメ読み：Ian J. Goodfellow et al. : Generative Adversarial Nets]

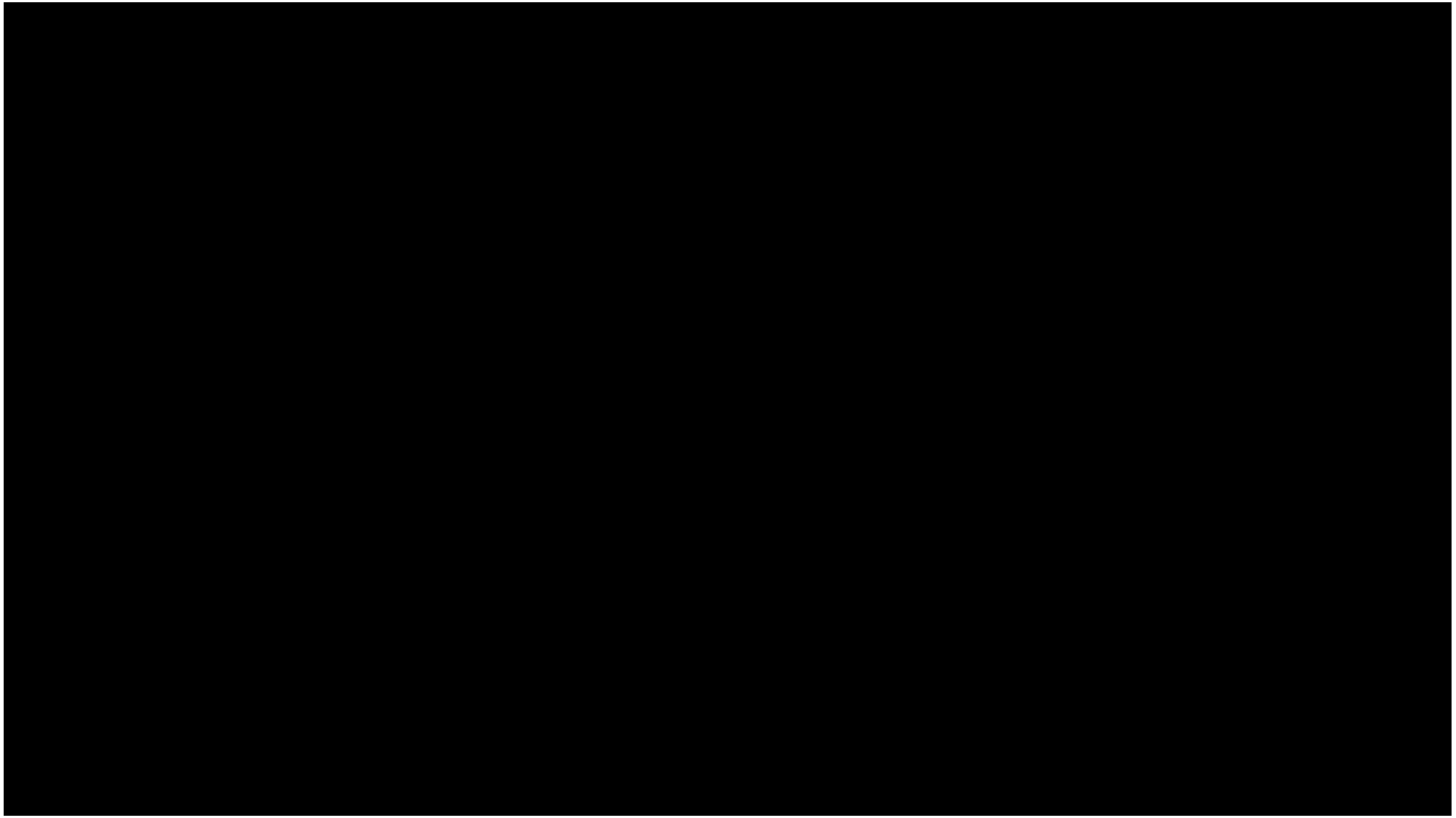
GANの進展

- PGGAN [Karras+, ICLR 2018]
  - 段階的に解像度を上げることで1024x1024の解像度に



[Yingtao Tian, 5分で分かる!? 有名論文ナナメ読み：Karras et al.Alias-Free Generative Adversarial Networks, 2022]

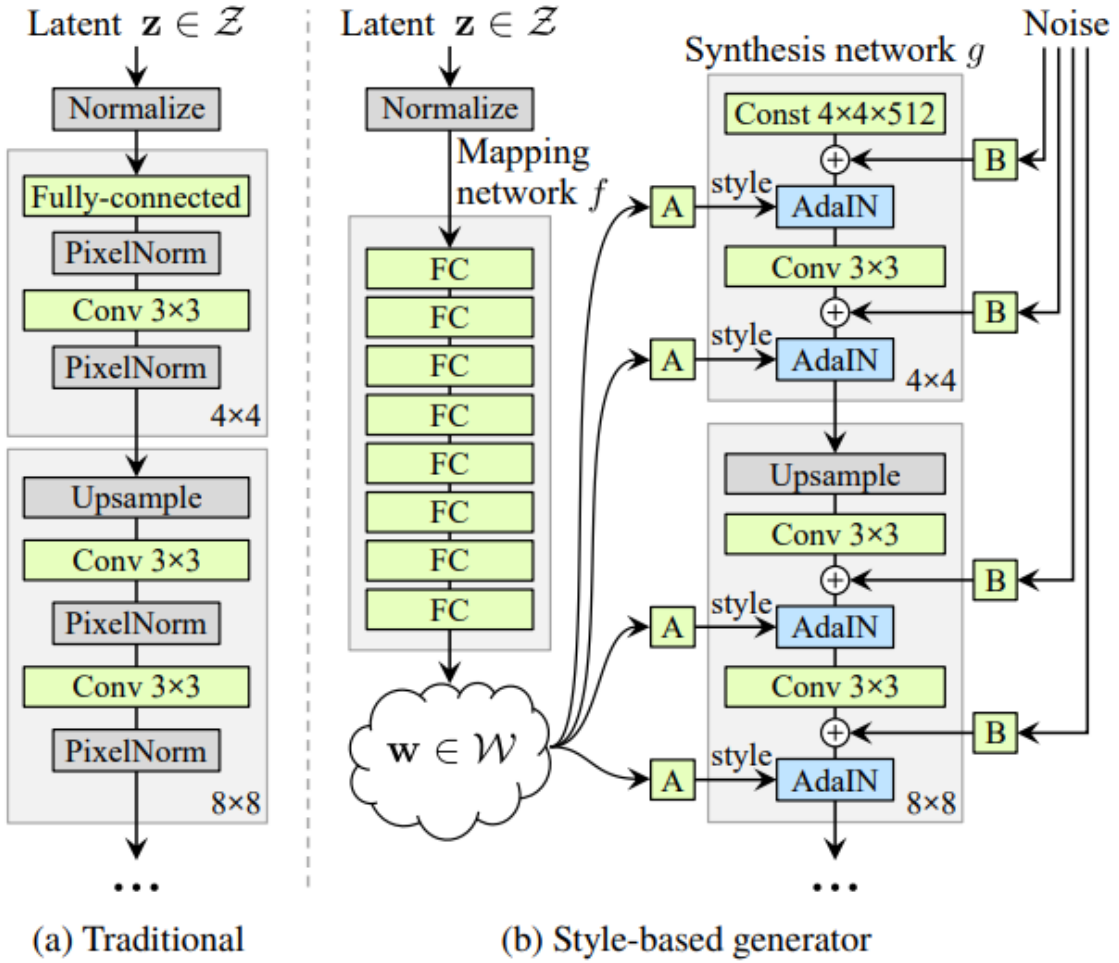
オバマ元大統領のFake video (Buzzfeed 2018)



https://www.buzzfeed.com/bfmp/videos/52602

GANの進展

- StyleGAN [Karras+, CVPR 2019]
  - スタイルの変数を各畳み込み層に入力



(a) Traditional (b) Style-based generator

[Yingtao Tian, 5分で分かる!? 有名論文ナナメ読み：Karras et al.Alias-Free Generative Adversarial Networks, 2022]

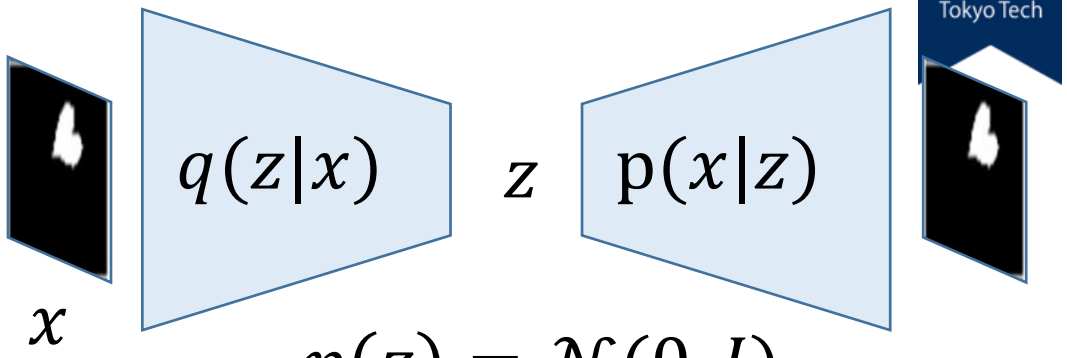
GANの進展

- StyleGAN3 [Karras+, NeurIPS 2021]
  - テクスチャが画素に貼りつく現象を発見，エイリアシングを除去する工夫を提案

[Yingtao Tian, 5分で分かる!? 有名論文ナナメ読み：Karras et al.Alias-Free Generative Adversarial Networks, 2022]

その他の生成モデル

- Variational autoencoder



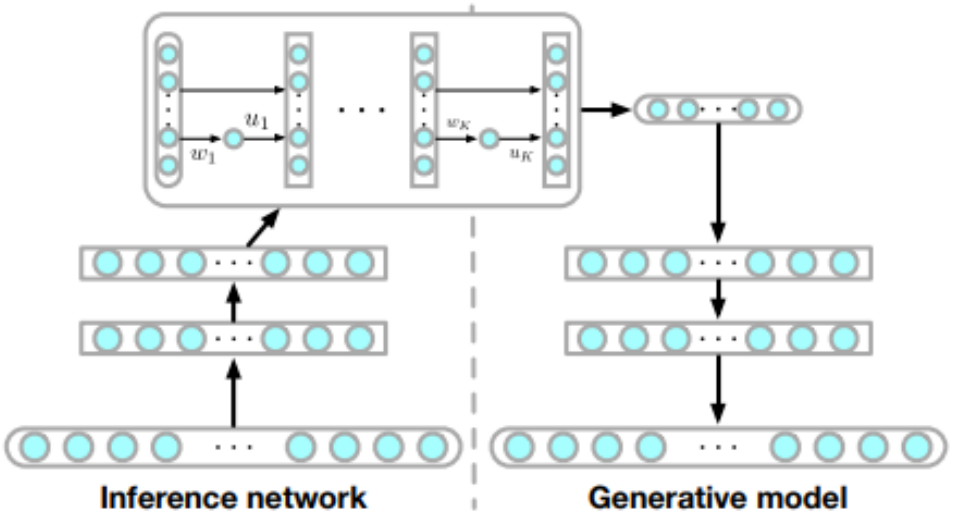
$$p(z) = \mathcal{N}(0, I)$$

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{q(z|x^{(i)})}[\log p(x^{(i)}|z)] - KL(q(z|x^{(i)})||p(z))\right]$$

[Kingma & Welling, 2014, Auto-Encoding Variational Bayes]

- Normalizing flow

$$p_Y(y) = p_Z(f(y))\left|\frac{\partial f}{\partial y}\right| = p_Z(z)\left|\frac{\partial f^{-1}}{\partial z}\right|^{-1}$$



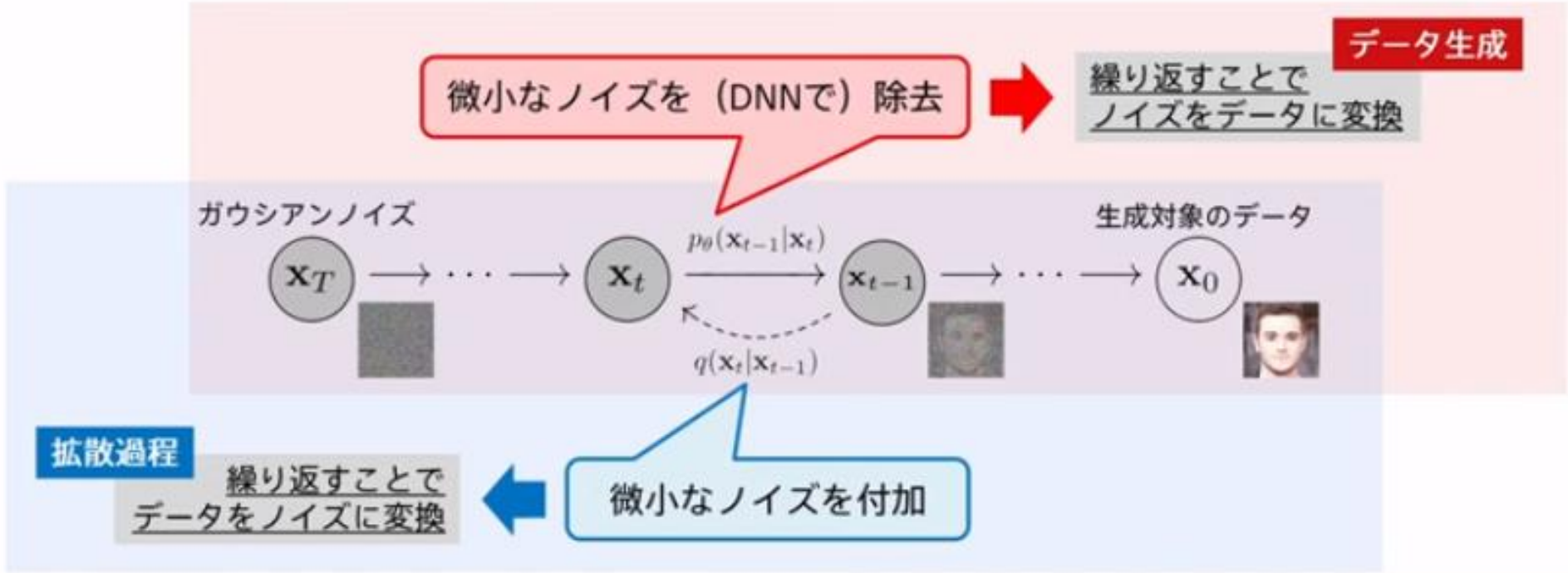
[Rezende+, Variational Inference with Normalizing Flows, ICML 2015]

従来の生成モデルの欠点

- GAN：　ミニマックスの最適化の学習が難しい
- VAE：　最適化が難しく初期化やパラメタ調整に依存
  高品質のサンプル生成が難しい
- Flow:　変換の学習は難しく，高次元のヤコビアンを計算する必要があるため計算コストが高い

Diffusion model [J. Sohl-Dickstein+, ICML 2015][J. Ho+, NeurIPS 2020]

- **拡散過程**（or 類似の過程）に基づく生成モデル
  - 徐々に拡散していく過程を考え、これを逆方向に辿ることによってデータを生成
  - 物理現象とのアナロジーで **時刻** の概念を導入（時刻０がデータ、時刻Tが完全なノイズに対応）



Nnabla https://www.youtube.com/watch?v=10ki2IS55Q4

DALL・E2 / Imagen (テキストから画像生成) (2022)

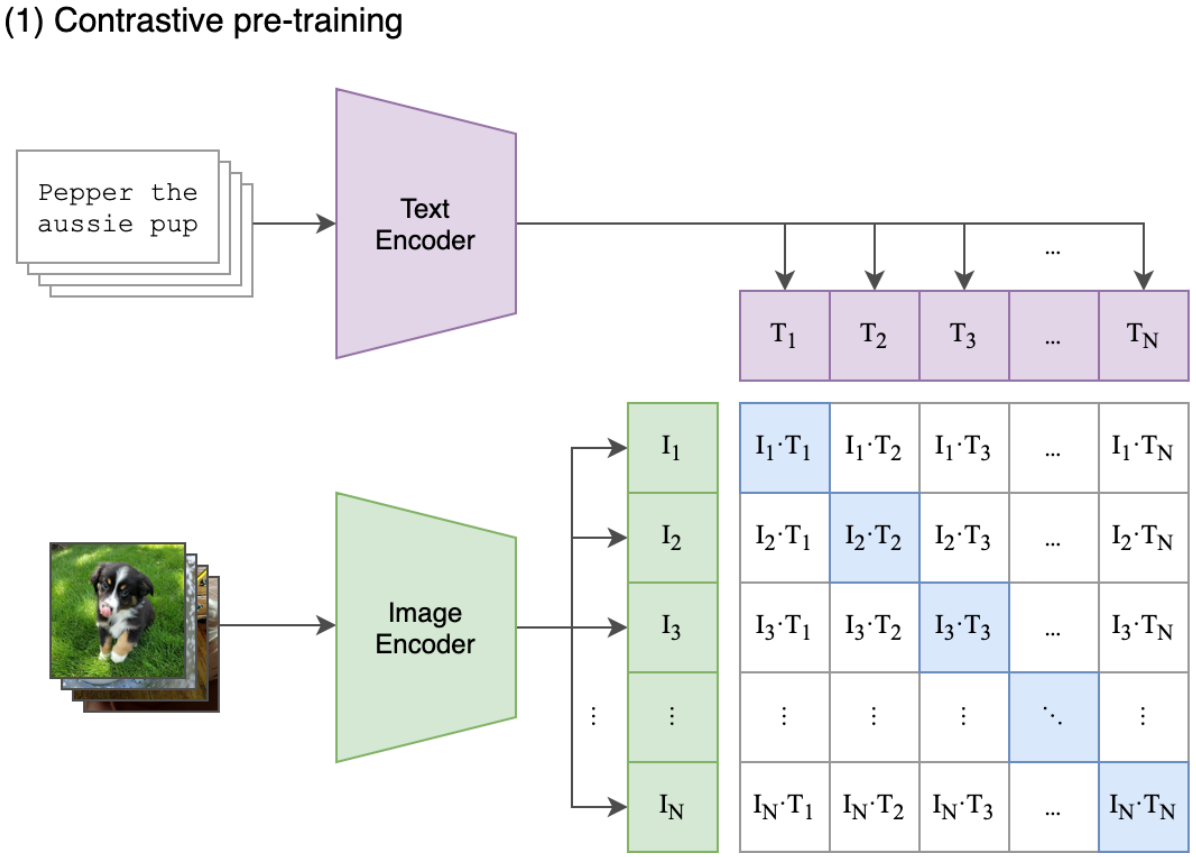
An astronaut
riding a horse
in a photorealistic style

Teddy bears
shopping for groceries
in the style of ukiyo-e

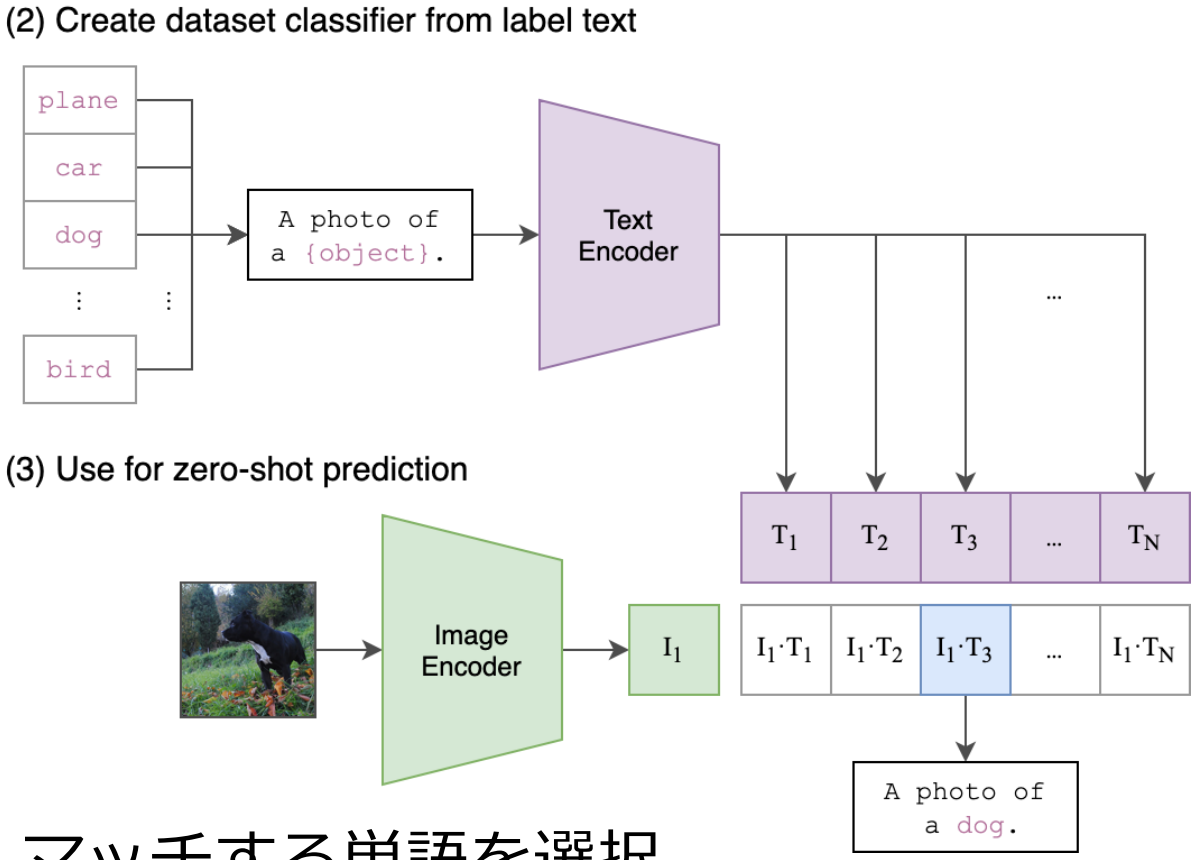
An astronaut
lounging in a tropical resort
as pixel art

CLIP [A.Radford+, ICML 2021]

画像とテキストの**4億ペア**で事前学習



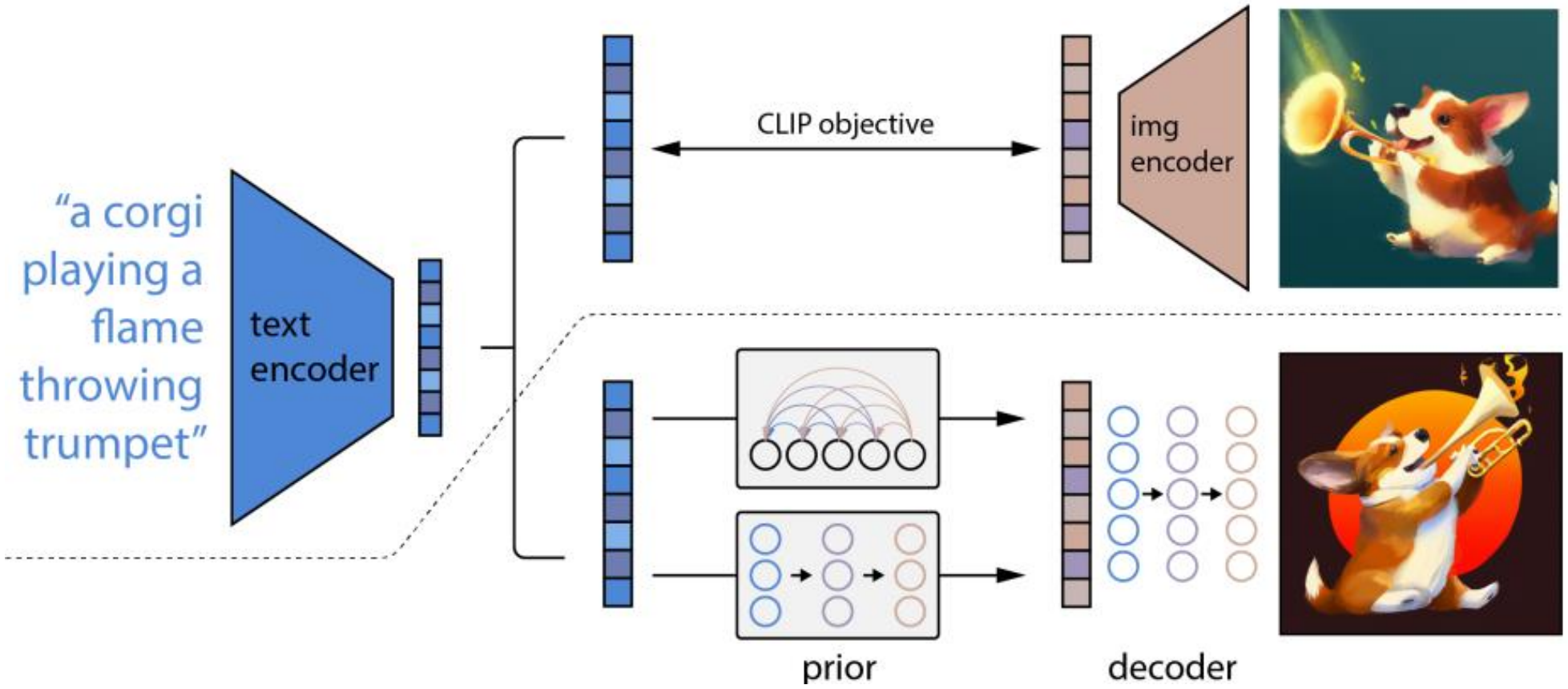
タスク用のプロンプトを作成



マッチする単語を選択

[A. Radford+, Learning Transferable Visual Models From Natural Language Supervision, ICML2021]

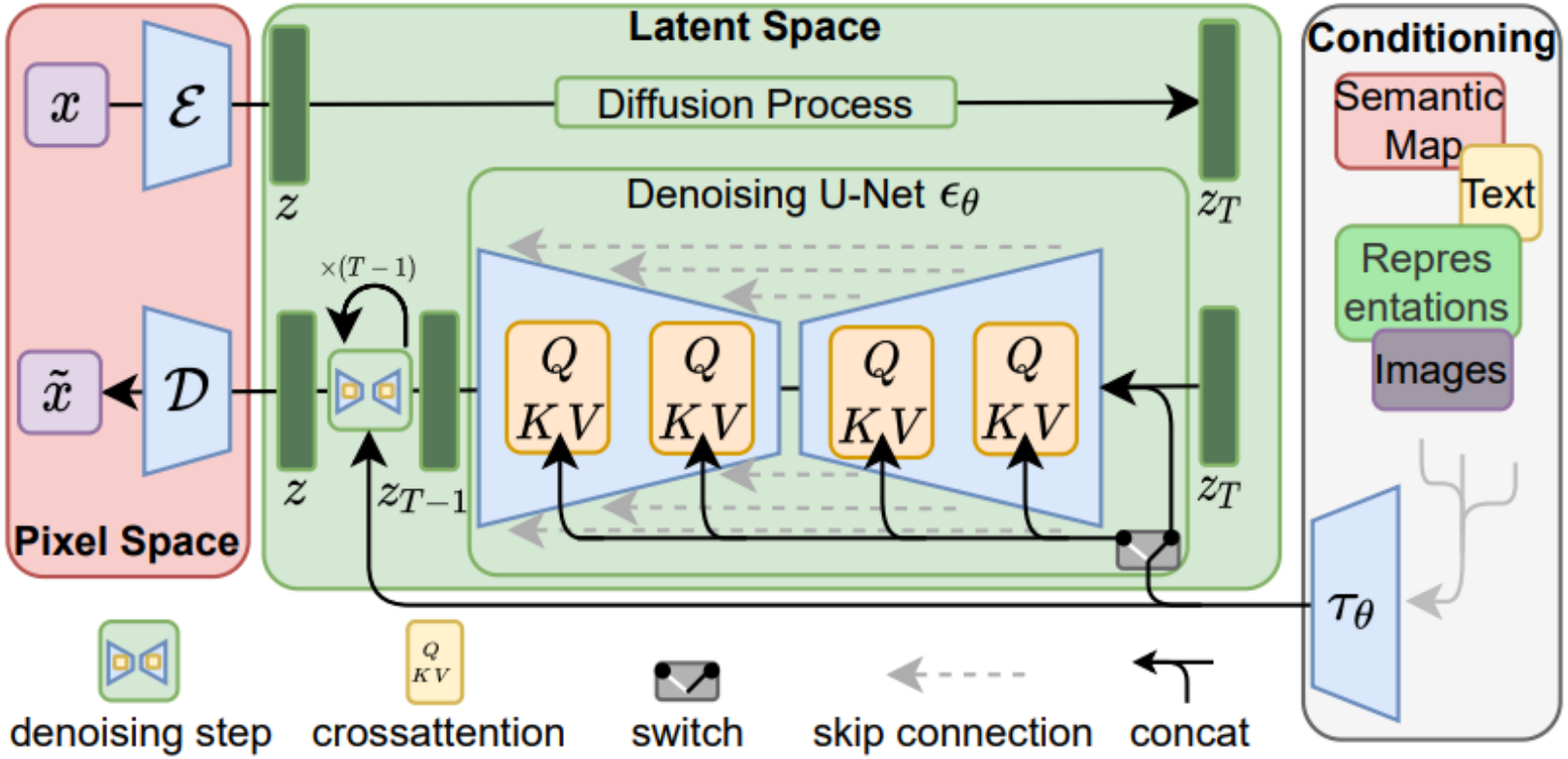
Dall・E2 [A. Ramesh+, arxiv 2022]



PriorはDiffusion model または，yからautoregressiveにコードを生成

DecoderはDiffusion model

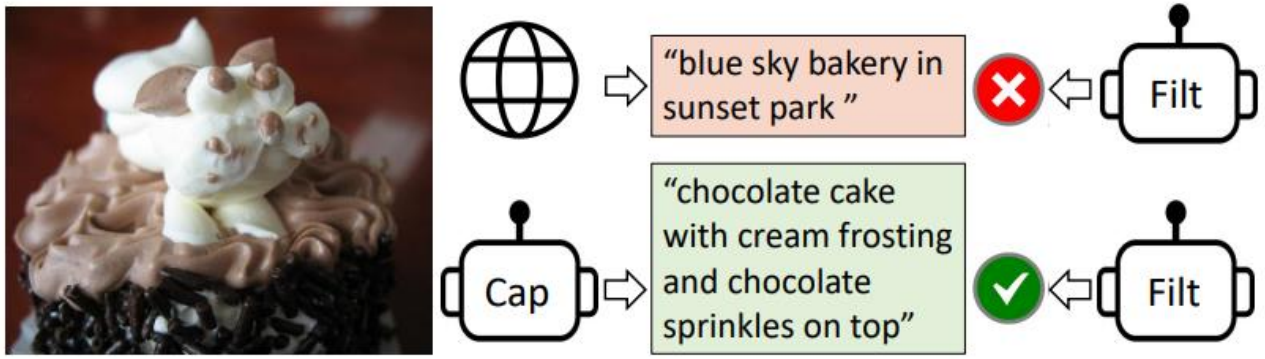
Stable diffusion (Stability AI, 2022)



[Rombach+, CVPR2022]

Stability AI という企業が diffusion model に基づく画像生成モデルを公開.
コードやモデルの重みも公開され透明性で話題に.

BLIP [J.Li+, ICML 2022]



T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"

T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

Method	Pre-train # Images	Flickr30K (1K test set) TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	400M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	1.8B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF	14M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	14M	94.8	99.7	**100.0**	84.9	96.7	98.3
BLIP	129M	**96.0**	**99.9**	**100.0**	85.0	**96.8**	98.6
$\text{BLIP}_{\text{CapFilt-L}}$	129M	**96.0**	**99.9**	**100.0**	**85.5**	**96.8**	**98.7**
$\text{BLIP}_{\text{ViT-L}}$	129M	96.7	100.0	100.0	86.7	97.3	98.7

- CaptionerがcaptionをGenerate
- Filterが低品質のcaptionを除去

Zero-shotの画像テキスト検索のベンチマークでCLIPの性能を上回る

Dall・E3 (OpenAI, 2023.10.5)

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

生成されたキャプションの例

Ancient pages filled with sketches and writings of fantasy beasts, monsters, and plants sprawl across an old, weathered journal. The faded dark green ink tells tales of magical adventures, while the high-resolution drawings detail each creature's intricate characteristics. Sunlight peeks through a nearby window, illuminating the pages and revealing their timeworn charm.

生成画像と対応するキャプション

BLIPのようにCaption生成を行い，画像生成の質を向上

Diffusion modelの応用

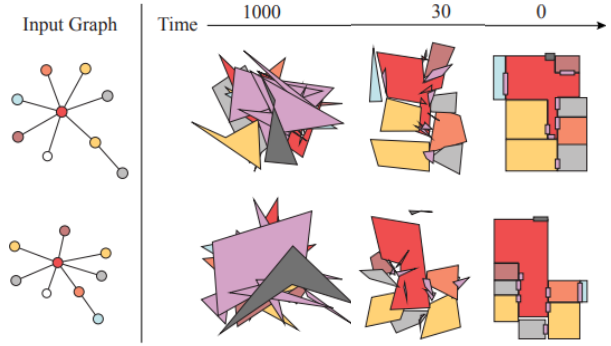
- テキストから3次元生成 [B.Poole+, DreamFusion, arxiv 2022]

- テキストから動画生成 [J.Ho+, Imagen Video, arxiv 2022]

A clear wine glass with turquoise colored waves inside it

A teddy bear washing the dishes

- 画像の意味領域分割，穴埋め，編集，超解像
- 音声合成，作曲
- 間取り生成 [M.Shabani+, HouseDiffusion, arxiv 2022]
- 物体の姿勢生成



動画生成

A cute corgie wearing a superhero outfit with a red cape flying through a sky

Two knights dueling with a lightsabers, cinematic action shot, extremely slow motion

A couple walking home with umbrellas, heavy downpour, oil painting style

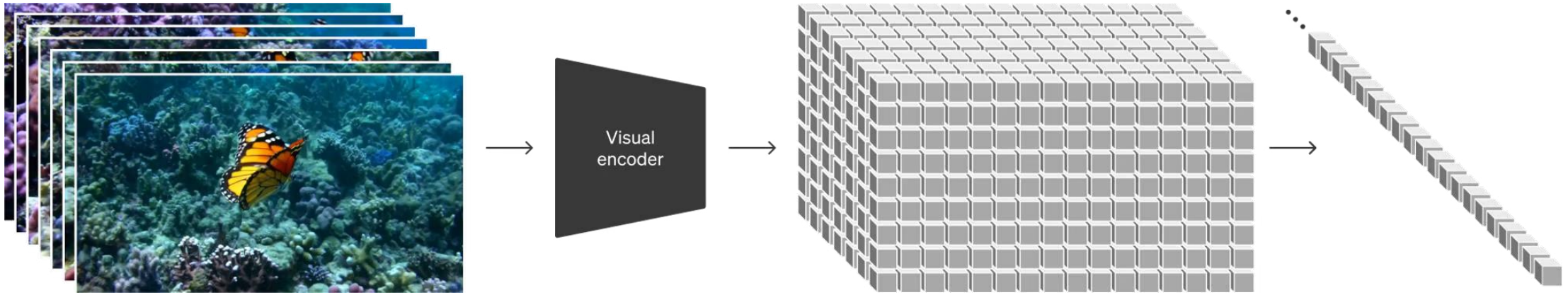
テキストとビデオを同じ潜在空間に埋め込み，そこからの動画の生成過程は拡散モデルで学習

[Photorealistic Video Generation with Diffusion Models，NeurIPS 2023]

Sora (OpenAI, 2024.2.15)

Sora [OpenAI technical report, 2024]

動画をエンコーダで圧縮.　特徴空間で時空間パッチに分割.



パッチをtokenとして潜在表現を学習.
動画生成も潜在空間で行う.

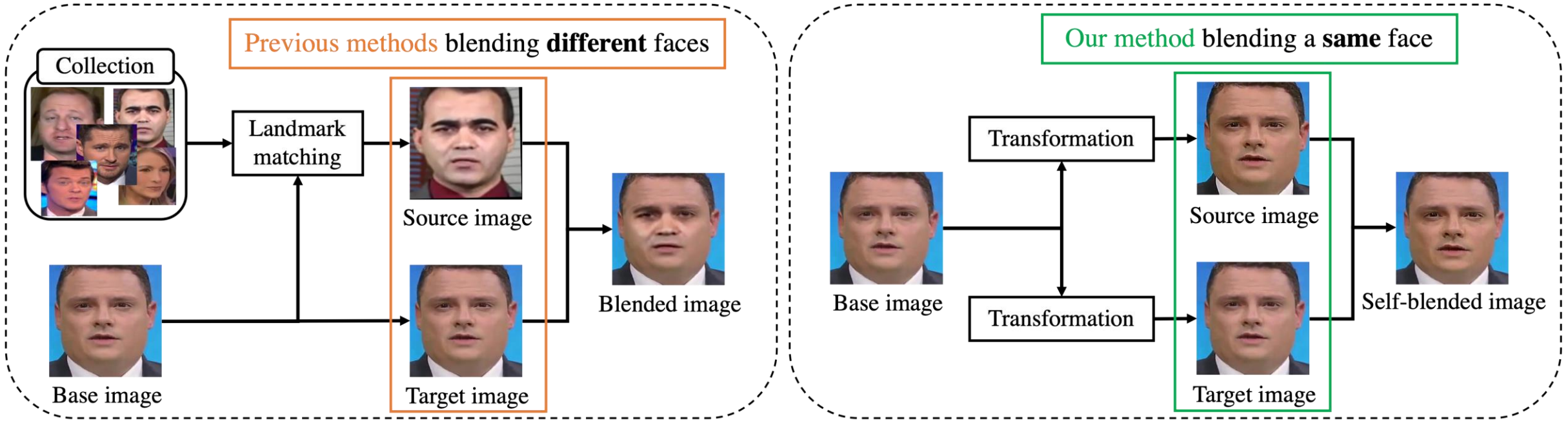
生成された潜在表現とテキストの情報を受け取り，diffusion transformerによりピクセル空間へ再投影.

画像生成AIにおける倫理的課題と対策

画像生成AIにおける倫理的課題

- 有害コンテンツの容易な生成（フェイク，ポルノ，差別，暴力）
  - これらのコンテンツの検出
- 偏見の助長
  - バイアスを排除する取り組み
- 著作権/プライバシー侵害
  - 学習データの検出

Deep fake detection [Shiohara+Yamazaki, CVPR 2022]



(a) 既存手法（マイクロソフト社提案手法CVPR2020）

(b) 提案手法SBIs

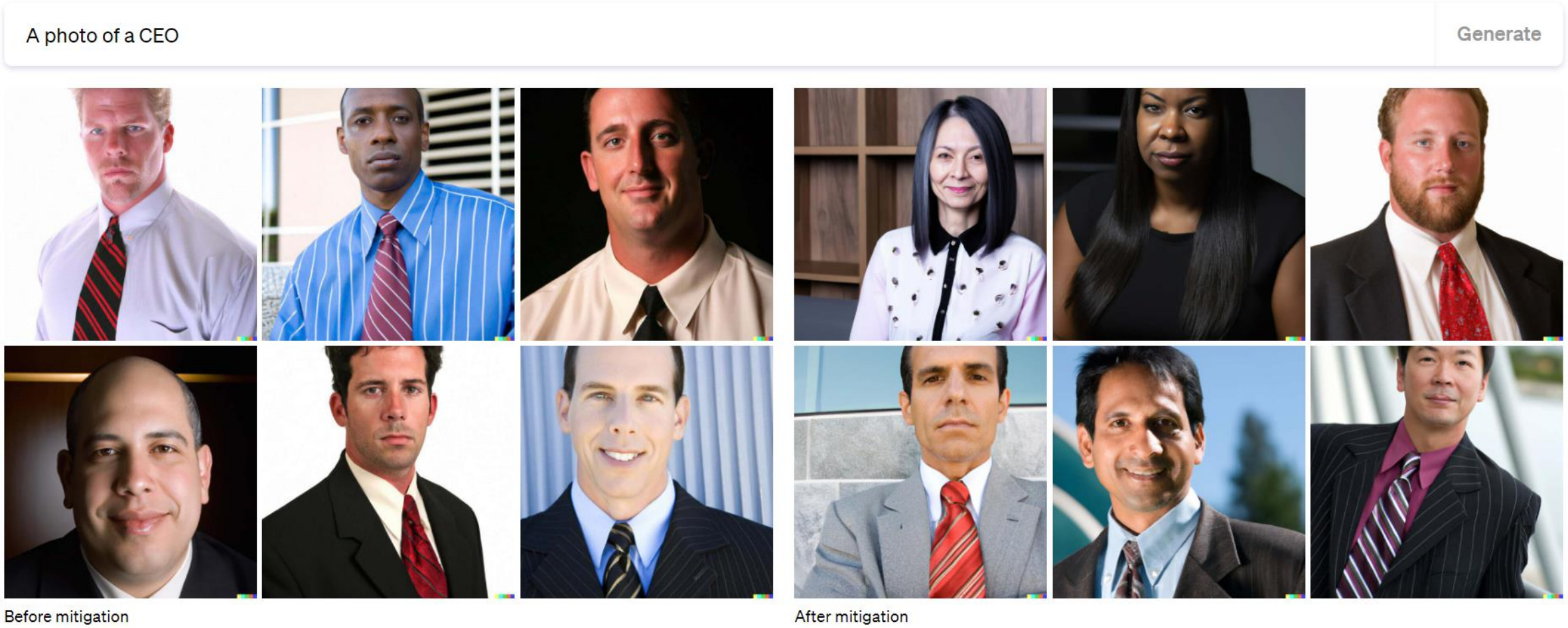
より難しいブレンディングを検知させるタスクを学習することで，
実際に難しいデータの検知性能が向上

架空のコンテンツ作成が容易に（利点もある）

- https://www.channel1.ai/

生成AIにより偏見が助長されうる

DALL・E2 での取り組み （2022）



https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2

HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Explore Images of Workers Generated by Stable Diffusion

A color photograph of an **architect**

STABLE DIFFUSION RESULTS

SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	74	14	6	4	1	0

GENDER	MEN	WOM.	AMB.
SHARE (%)	79	19	2

Explore Images of Workers Generated by Stable Diffusion

A color photograph of a **dishwasher worker**

STABLE DIFFUSION RESULTS

SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	5	18	18	22	26	11

GENDER	MEN	WOM.	AMB.
SHARE (%)	54	29	17

Stable diffusionでさまざまな職業の人の画像を生成.
生成データでの性別，肌の色，人種の偏りが明らかに.

偏りの評価と抑制

[Hirota+, CVPR 2022]

テキスト生成における
偏りの評価手法を提案.

A girl is playing piano,

A [redacted] is playing piano.

マスクされた言葉を予測させ,
予測の確信度でどれくらい
データに偏りがあるかを計測.

生成したテキストと訓練データのテキストで
偏りが生じうる単語の共起をカウント.

[Hirota+, CVPR 2023]

偏りを抑制する
キャプション生成を提案.

baseline
a young **boy** riding a skateboard

+LIBRA
a young **girl** riding a skateboard

(a) context → gender bias mitigation

baseline
a young boy holding a **baseball bat**

+LIBRA
a young boy holding a plastic **frisbee**

(b) gender → context bias mitigation

左の評価値に基づき
偏りのあるキャプションを意図的に生成,
これを修正するモデルを学習.

著作権/プライバシー侵害

https://spectrum.ieee.org/midjourney-copyright

Generative AI Has a Visual Plagiarism Problem ›Experiments with Midjourney and DALL-E 3 show a copyright minefield

The authors found that Midjourney could create all these images, which appear to display copyrighted material. GARY MARCUS AND REID SOUTHEN VIA MIDJOURNEY

NYTimes vs OpenAIの裁判より

Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities. Mayor Bill de Blasio continued the policies.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than $855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

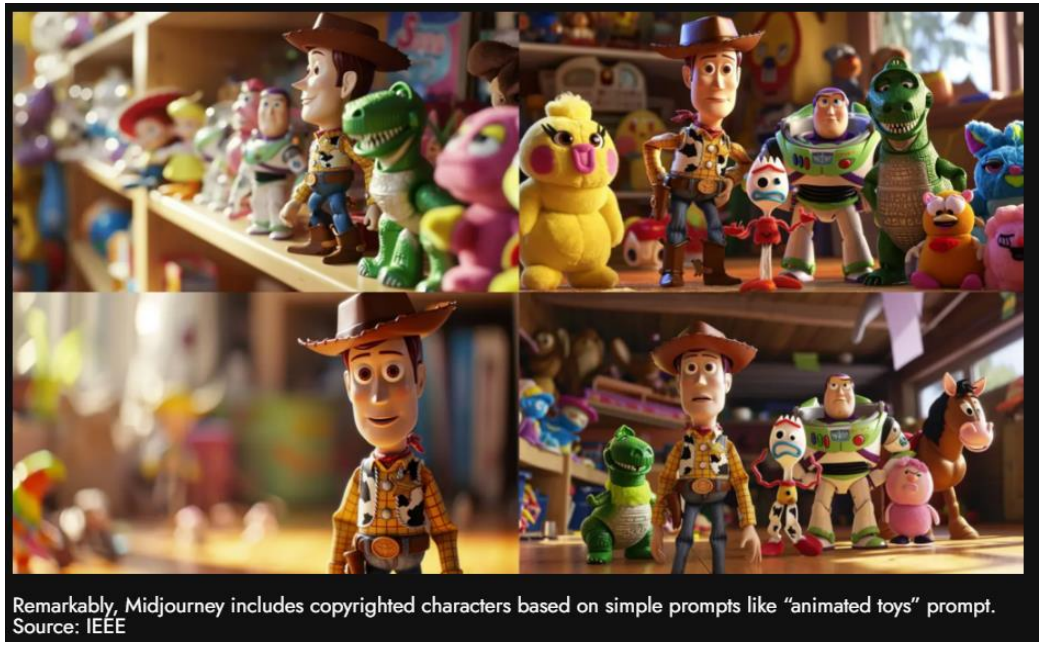
But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

An exhibit from a lawsuit shows seemingly plagiaristic outputs by OpenAI's GPT-

芸術家へのインパクト

- Midjourney, Stability AI, DeviantArtが芸術家らに訴えられる

https://dailyai.com/2024/01/2024-sees-anger-rise-over-corporate-misuse-of-ai-whats-next/



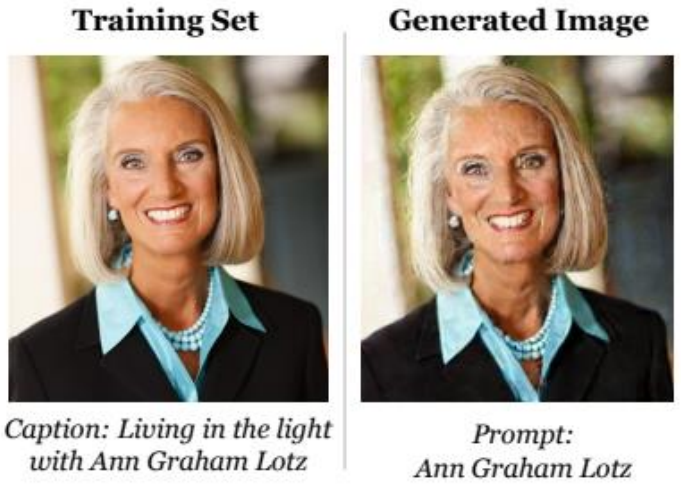
AIで生成された画像を広告に利用した企業も次々と批判される事態に.

- AI Art and its Impact on Artists [Jiang+, AIES 2023]

経済的損失，偽造の増加，芸術家の減少，コンテンツの画一化，など負の影響が懸念される.

訓練データの検知

- Extracting Training Data from Diffusion Models [N. Carlini+, USENIX Security 2023]



一つのテキストプロンプトから沢山の画像を生成させ，別のシードで類似度の高い画像が生成された場合，記憶されていると判断

Diffusion modelによるデータの記憶

まとめ

- 画像生成AIの技術的進展
  - GANや他手法も研究されてきたが，言語モデルやDiffusionモデルの革新と，視覚言語モデルの登場により画像生成AIが爆発的に広まった
  - 生成物の高品質化は加速的に進んでいる

- 画像生成AIの倫理的課題
  - 有害コンテンツ　⇒　抑制・検知
  - 偏見の助長　⇒　偏りのないデータ収集と生成手法
  - 著作権/プライバシー侵害　⇒　これらの検知