命題論理に基づく ガイダンスによる ユーザの意図に忠実な 画像生成拡散モデル

松原 崇(北海道大学)

第63回 NLPコロキウム

書誌情報：Kota Sueyoshi, Takashi Matsubara,
"Predicated Diffusion: Predicate Logic-Based Attention Guidance for Text-to-Image Diffusion Models,"
CVPR, 2024. (highlight)

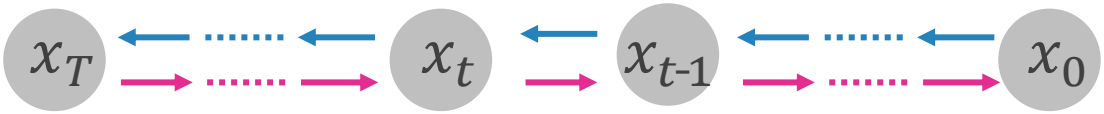
背景

- **拡散モデル**
  - 高品質で多様性のある画像生成が可能
  - 高い拡張性(画像，動画，音声)

拡散過程(推論過程)

徐々にノイズを加えてランダムノイズにしていく

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$$



ノイズと実データ間の関係を学習

逆拡散過程(生成過程)

ランダムノイズを初期状態として徐々にノイズを減らす

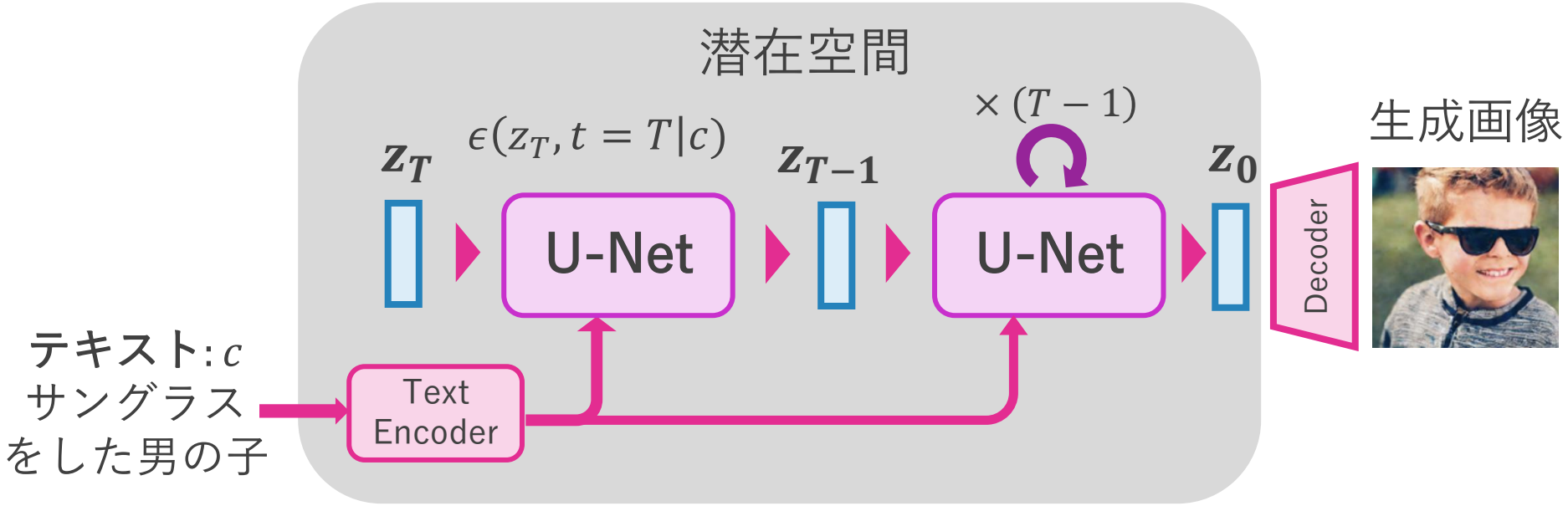
$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$$

の部分をU-netで学習

背景

- **研究対象：Text-to-image model**
  - 例：Stable Diffusion, DALL-E 3,　Midjourney
  - 低次元の潜在空間で拡散モデルを学習することで計算コスト低減
  - タスクに特化した入力エンコーダーを導入(今回はテキスト)

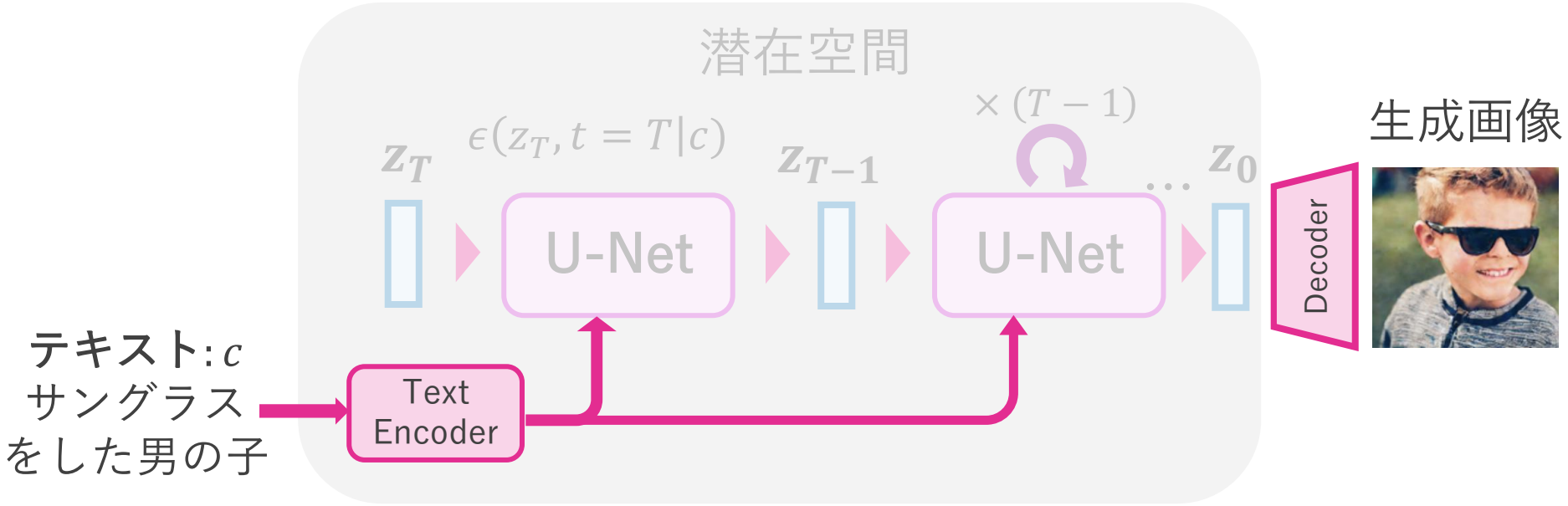
[Rombach+, CVPR2022]



生成時のプロセス

背景

- 研究対象：Stable Diffusion
  - テキストによる条件づけ Text Encoder
  - テキストをCLIPという埋め込みを用いてベクトルに変換
  - 取得したベクトルをU-Net内に組み込む(Cross-Attention)



生成時のプロセス

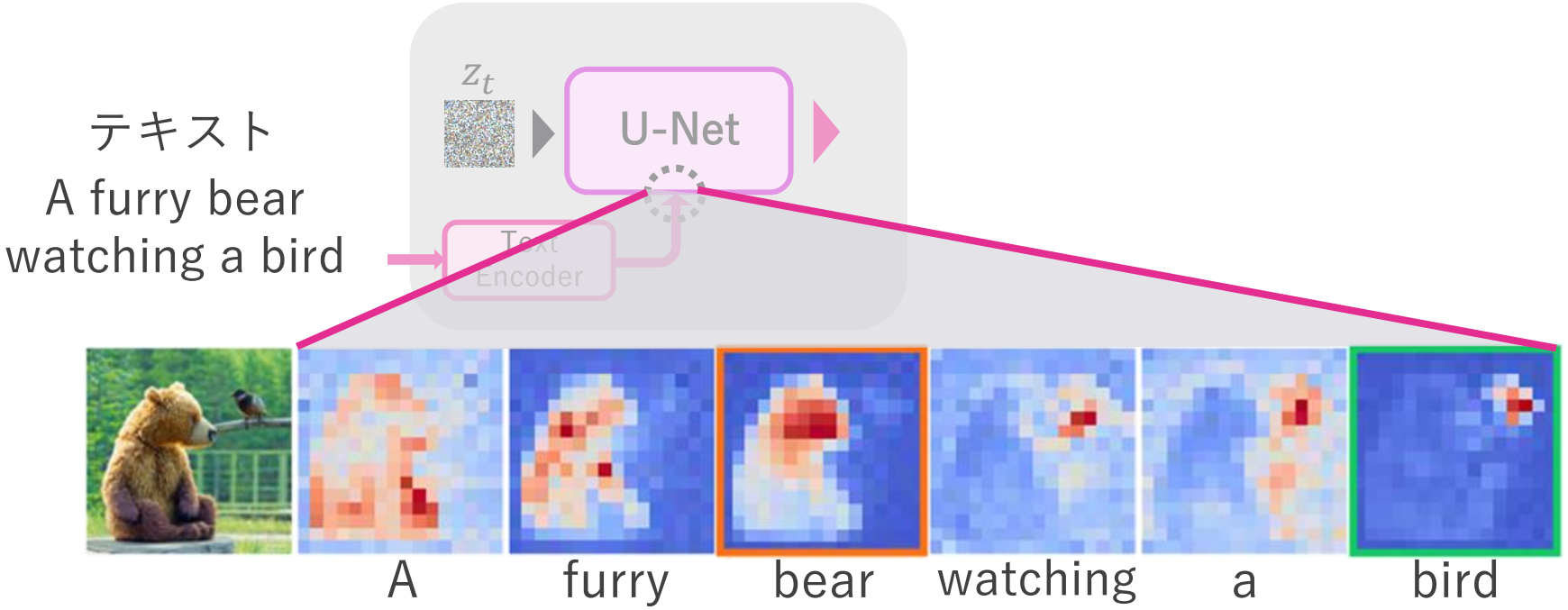
問題点

- Text-to-image modelの生成で失敗する例

	A yellow car and a blue bird	A bird and a cat	A green balloon and a purple clock	A boy grasping a soccerball
Stable Diffusion				
	物体の消失	物体の混合	属性の漏れ	所有の失敗

問題点

- U-Net内のAttention層で得られるAttention map
  - Attention mapの強度が高いピクセルは対応する物体や概念を示唆



  - Attention mapが適切に反応するように，画像を補正すれば，よりテキストに忠実な生成ができる可能性あり
  - 適当な損失関数$\mathcal{L}$を用意して更新時に使う(ガイダンス)

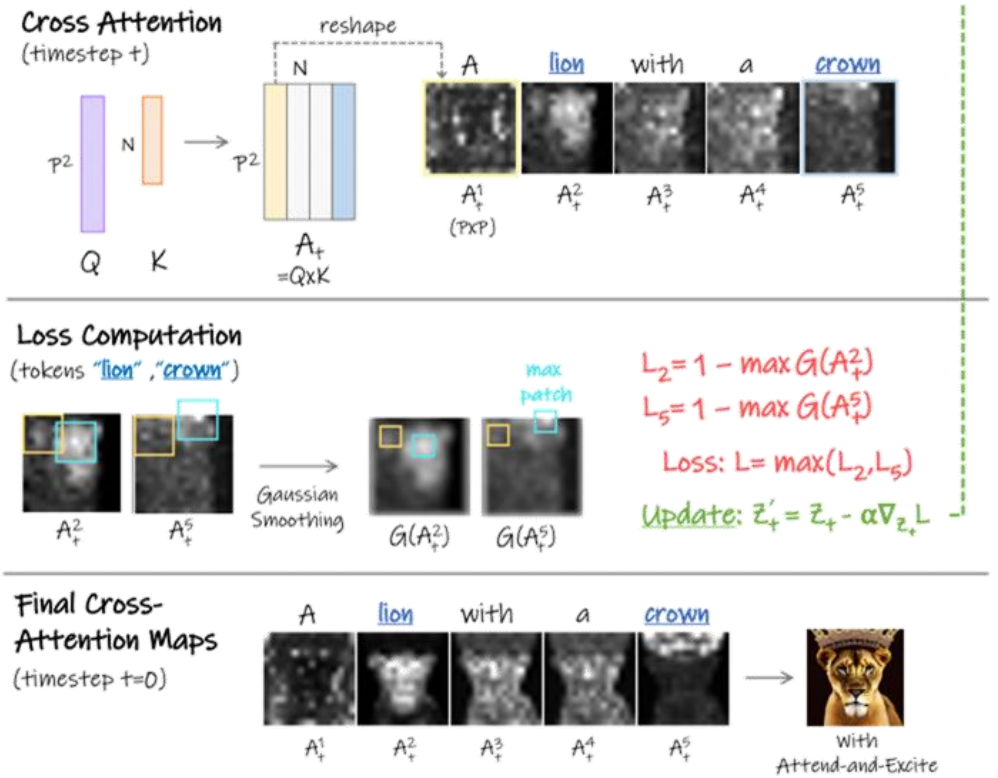
$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z \boxed{-\eta\nabla\mathcal{L}}$$

関連研究①：Attend and Excite

- 物体混合と物体消滅
  - 物体混合/消滅している画像ではAttentionが反応してない
  - 物体のAttentionを活性化させるような損失を導入
    - $\mathcal{L}_1 = 1 - \max_i A_1$

      $\mathcal{L}_2 = 1 - \max_i A_2$

      $\mathcal{L} = \max(\mathcal{L}_1, \mathcal{L}_2)$



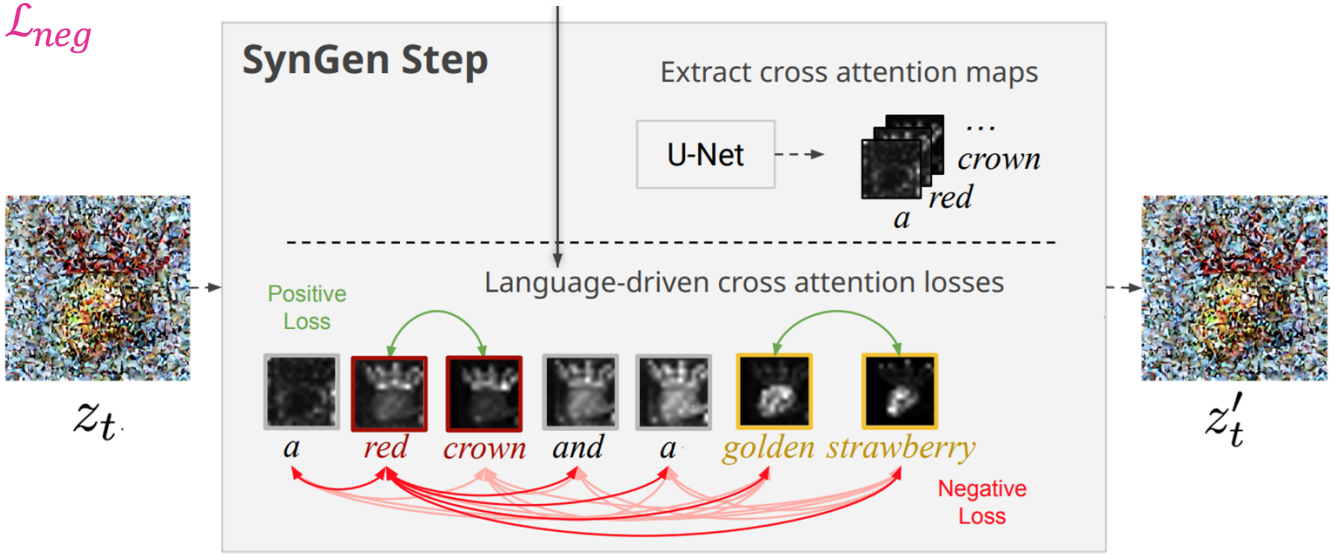
Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models (SIGGRAPH 2023)
https://arxiv.org/abs/2301.13826

関連研究②：SynGen

- 属性の漏れ
  - 形容詞と名詞の対応を修正するように損失を定義
  - 形容詞と名詞のAttentionの分布を一致させる
    - $\mathcal{L}_{pos} = \sum_{i=1}^{k} \sum_{(m,n)\in P(S_i)} dist(A_m, A_n),\ dist(A_i, A_j) = \frac{1}{2} D_{KL}(A_i || A_j) + \frac{1}{2} D_{KL}(A_j || A_i)$

$$\mathcal{L}_{neg} = -\sum_{i=1}^{k} \frac{1}{U(S_i)} \sum_{(m,n)\in P(S_i)} \sum_{u\in U(S_i)} \frac{1}{2} \left(dist(A_m, A_u) + dist(A_u, A_n) \right)$$

$$\mathcal{L} = \mathcal{L}_{pos} + \mathcal{L}_{neg}$$



Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment (NeurIPS2023)
https://arxiv.org/abs/2306.08877

研究背景:Attention guidance

物体の消失

物体の混合

Attend-and-Excite [Chefer+, SIGGRAPH2023]

- 物体のAttentionを活性化

属性の漏れ

SynGen [Rassin+, NeurIPS2023]

- 形容詞と名詞の対応を修正

課題別のアプローチは存在するが
統一的な解決はされていない

述語論理を用いた包括的な手法を提案

提案手法Predicated Diffusion

- **Attention mapを論理命題とみなす**
  - 単語Pに対し「xはPである」という命題$P(x)$を
    単語Pに対応するAttention mapの
    i番ピクセルの強度 $A_P[i]$に対応付ける
  - ピクセルの強度は連続値なのでファジィ論理(product fuzzy)を用いる

命題	Attention Map
true	1
false	0
$P(x)$	$A_P[i]$
$\neg P(x)$	$1 - A_P[i]$
$P(x) \wedge Q(x)$	$A_P[i] \times A_Q[i]$
$P(x) \rightarrow Q(x)$	$1 - A_P[i] \times (1 - A_Q[i])$
$\forall x.\, P(x)$	$\prod_i A_P[i]$
$\exists x.\, P(x)$	$1 - \prod_i (1 - A_P[i])$

提案手法Predicated Diffusion

存在　Prompt: There is a dog

- プロンプトを$\exists x.Dog(x) = \neg\big(\forall x.\neg Dog(x)\big)$と解釈
- ファジィ論理で表すと$1-\prod_i\big(1-A_{Dog}[i]\big)$
- 負の対数をとり，損失とする

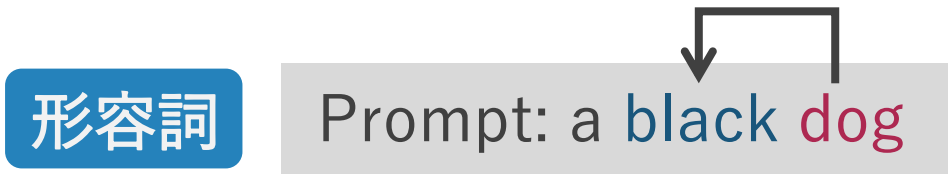
  $\mathcal{L}[\exists x.Dog(x)] = -\log\big(1-\prod_i\big(1-A_{Dog}[i]\big)\big)$
  - ベルヌーイ交差エントロピー損失のイメージ

提案手法Predicated Diffusion

共存　Prompt: a dog and a cat

- 両方の物体が存在することを確実にしたい
- $\mathcal{L}\left[\left(\exists x.\, Dog(x)\right)\wedge\left(\exists x.\, Cat(x)\right)\right]$
  $= \mathcal{L}[\exists x.\, Dog(x)] + \mathcal{L}[\exists x.\, Cat(x)]$
  $= -\log\left(1-\prod_i\left(1-A_{Dog}[i]\right)\right) - \log(1-\prod_i(1-A_{Cat}[i]))$

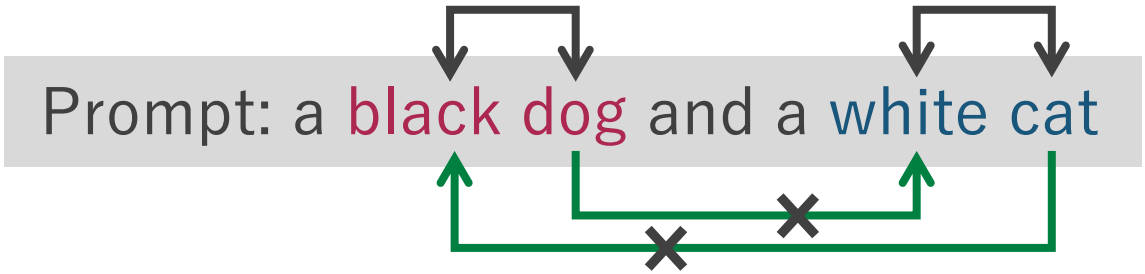
提案手法Predicated Diffusion



- 「黒い犬」を「犬ならば黒い」と解釈
- $Dog(x) \longrightarrow Black(x) = \neg\big(Dog(x) \wedge \neg Black(x)\big)$
- ファジィ論理では$1 - A_{Dog}[i] \times (1 - A_{Black}[i])$
- $\mathcal{L}[Dog(x) \longrightarrow Black(x)] = -\log\Big(1 - A_{Dog}[i] \times (1 - A_{Black}[i])\Big)$

提案手法Predicated Diffusion

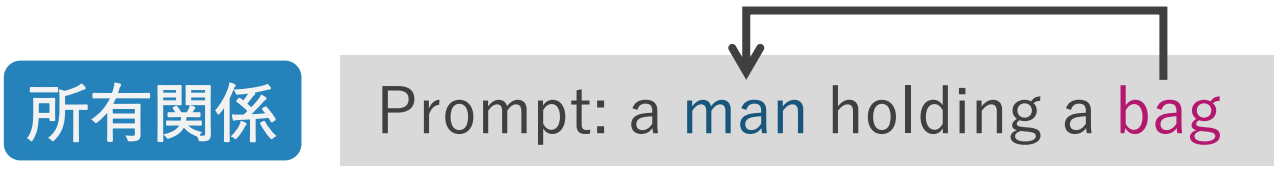
一対一対応



- 名詞と形容詞の対応関係を確実にしたい
- 「犬ならば黒い」だけでなく，「黒いものは犬」「犬は白くない」
- $\mathcal{L}_{\text{one-to-one}}$

$$= \mathcal{L}[\forall x. Dog(x) \leftrightarrow Black(x)] + \mathcal{L}[\forall x. Cat(x) \leftrightarrow White(x)]$$
$$+\alpha\mathcal{L}[\forall x. Dog(x) \longrightarrow \neg White(x)]$$
$$+\alpha\mathcal{L}[\forall x. Cat(x) \longrightarrow \neg Black(x)]$$

提案手法Predicated Diffusion



- 「人がバッグを持っている」＝「バッグが人の一部を成す」
- $\mathcal{L}[\forall x.\, Bag(x) \longrightarrow Man(x)]$

$$= -\sum_i \log\left(1 - A_{Bag}[i] \times (1 - A_{Man}[i])\right)$$

実験結果

- さきほどの例

	A yellow car and a blue bird	A bird and a cat	A green balloon and a purple clock	A boy grasping a soccerball
Stable Diffusion				
Predicated Diffusion				
	物体の消失	物体の混合	属性の漏れ	所有の失敗

実験

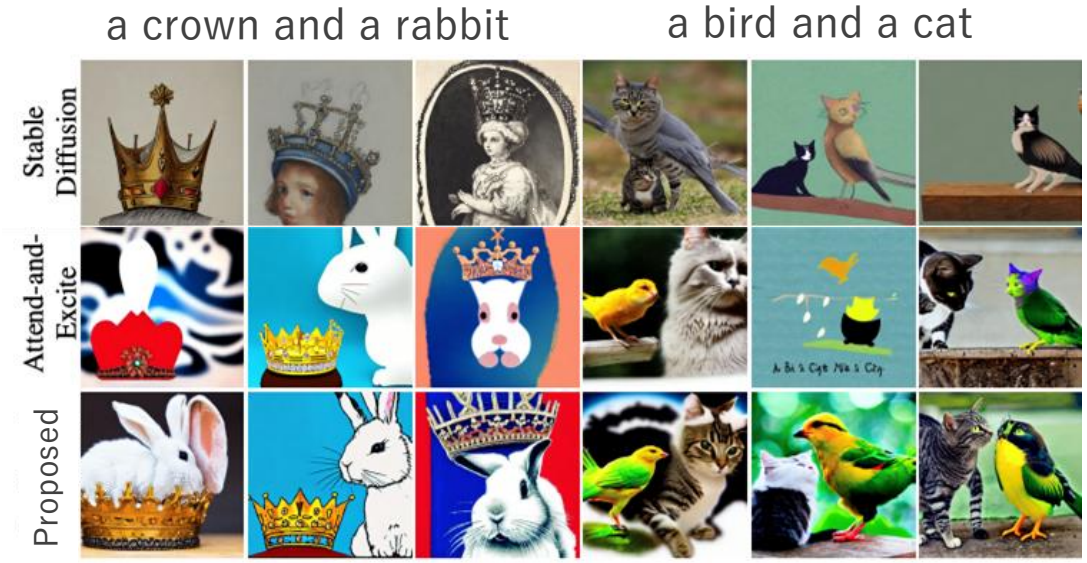
- **ユーザー評価**
  - 物体が生成されているかどう(0個,1個,2個,混合)を回答
  - 手法間で一番テキストに忠実なものを選択
- **自動評価**
  - CLIP埋め込みによる画像-テキスト間のcos類似度
  - BLIPによるキャプションとテキスト間のcos類似度
  - CLIP-IQA（生成画像が自然かどうか）

実験	枚数（ユーザ）	枚数（自動）	プロンプトの型
(i) 共存	400	10k	a [Object A] and a [Object B]
(ii) 一対一対応	400	10k	a [Adjective A] [Object A] and a [Adjective B] [Object B]
(iii) 所有関係	200	10k	a [Subject A] is [Verb C]-ing a [Object B]

実験結果 (i)共存

- 物体混合・消滅などの問題が改善されている

	ユーザー評価		自動評価	
Models	物体の消失や混合 ↓	忠実度 ↑	**CLIP / BLIP 類似度** ↑	画像品質 ↑
Stable Diffusion	66.0	11.0	0.326 / 0.767	0.761
Composable Diffusion	82.3	2.5	0.317 / 0.739	0.764
Structure Diffusion	64.5	12.0	0.325 / 0.763	0.763
Attend-and-Excite	36.3	29.5	0.342 / 0.814	0.766
Proposed	**28.5**	**30.3**	**0.348 / 0.825**	**0.775**



実験結果 (i)共存

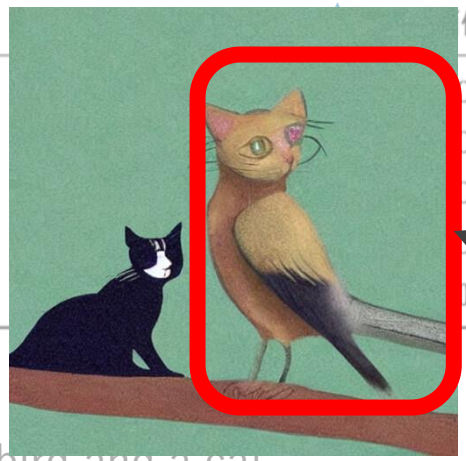
- 物体混合・消滅などの問題が改善されている

a crown and a rabbit

a bird and a cat

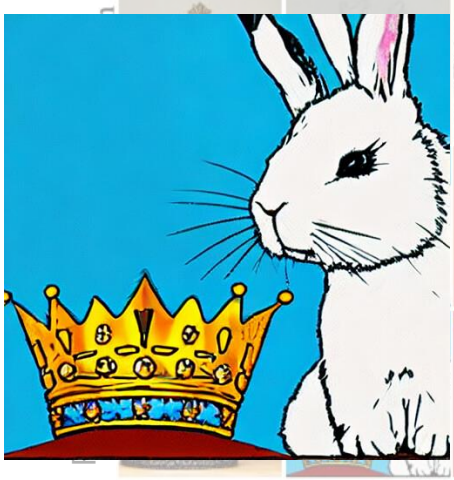
Stable Diffusion

物体の消滅



物体の混合

Proposed



実験結果 (ii)一対一対応

- 属性のもれ，物体消滅などの問題が改善されている

Models	ユーザー評価 物体の消失と混合 ↓	ユーザー評価 属性の漏れ ↓	ユーザー評価 忠実度 ↑	自動評価 CLIP / BLIP 類似度 ↑	自動評価 画像品質 ↑
Stable Diffusion	73.5	88.5	6.0	0.345 / 0.744	0.756
Composable Diffusion	83.5	88.5	3.8	0.348 / 0.729	0.757
Structure Diffusion	69.5	86.5	5.8	0.346 / 0.741	0.760
Attend-and-Excite	35.8	64.5	19.3	0.367 / 0.792	0.761
SynGen	29.3	40.3	36.8	0.367 / 0.801	0.750
Proposed	**16.5**	**33.0**	**44.8**	**0.379 / 0.811**	**0.769**

実験結果 (ii)一対一対応

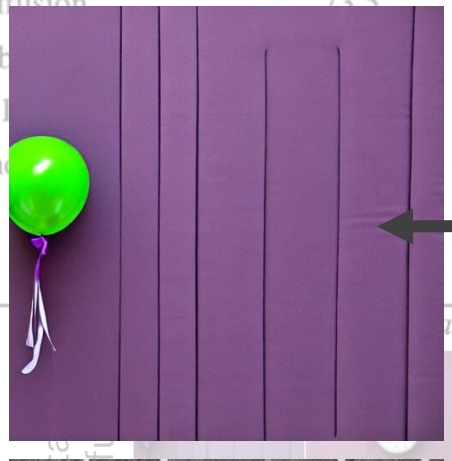
- 属性のもれ，物体消滅などの問題が改善されている

a green balloon
and
a purple clock

a blue turtle
and
a white bear

Stable Diffusion

Proposed



物体の消滅
＋
属性の漏れ

物体の消滅

実験結果 (iii)所有関係

- 所有関係の問題が改善されている

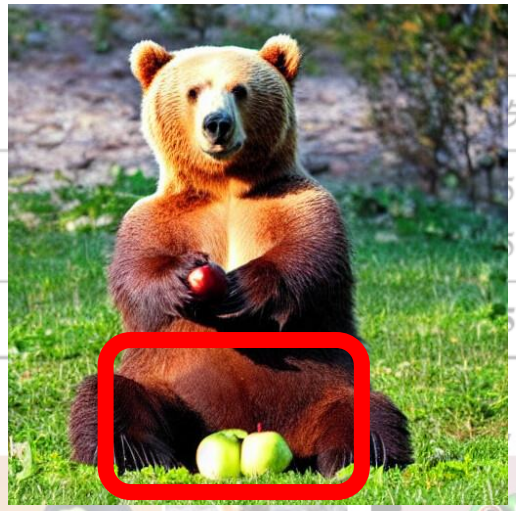
	ユーザー評価			自動評価	
Models	物体の消失と混合↓	所有の失敗↓	忠実度↑	**CLIP / BLIP 類似度↑**	画像品質↑
Stable Diffusion	36.0	52.5	33.5	0.320 / 0.811	0.762
Attend-and-Excite	17.0	51.5	27.5	0.334 / 0.843	0.760
Proposed	**7.0**	**29.5**	**52.0**	**0.345 / 0.855**	**0.765**

実験結果 (iii)所有関係

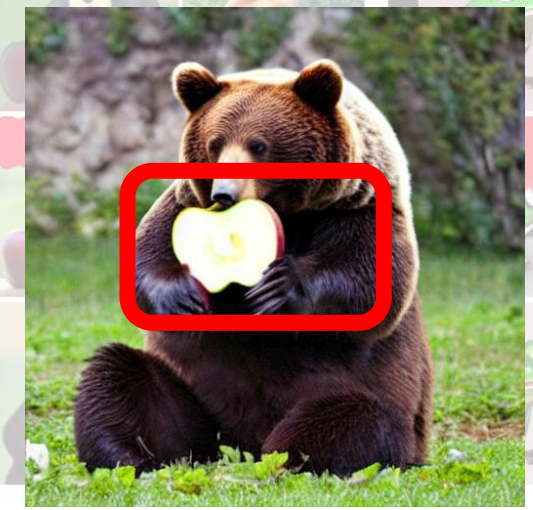
- 所有関係の問題が改善されている

a bear having an apple

Stable Diffusion



Proposed



- Stable Diffusionでは熊がりんごを持っていない(所有の失敗)
- 提案手法では，りんごを持つ熊が生成されている

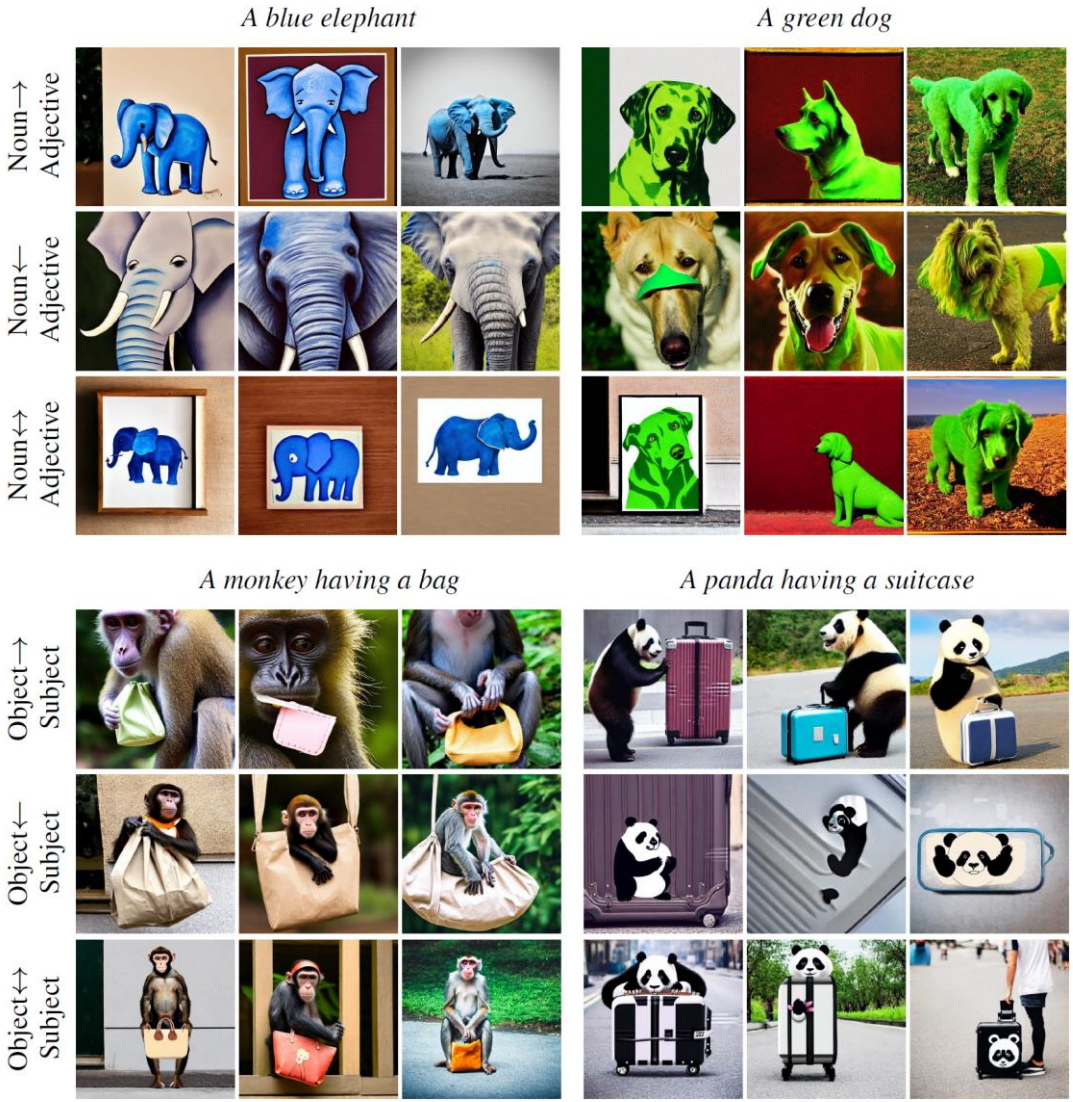
実験結果 (iv)さらに複雑なプロンプト

- 3物体以上にも対応
  - 15個以上のガイダンスを導入しても安定して画像生成可能

  - ORも可能: $x.Bird(x) \rightarrow (Green(x) \vee Grey(x))$
  - 否定も可能だが，negative promptで十分

実験結果 (v)含意の検証

- 形容詞や所有の含意の方向が正しいことがわかる
  - 名詞→形容詞
    染まるが漏れる
  - 名詞←形容詞
    部分的に染まる
  - 名詞↔形容詞
    限定して染まる
  - 所有物→所有者
    所有される
  - 所有物←所有者
    逆転する
  - 所有物↔所有者
    安定しない

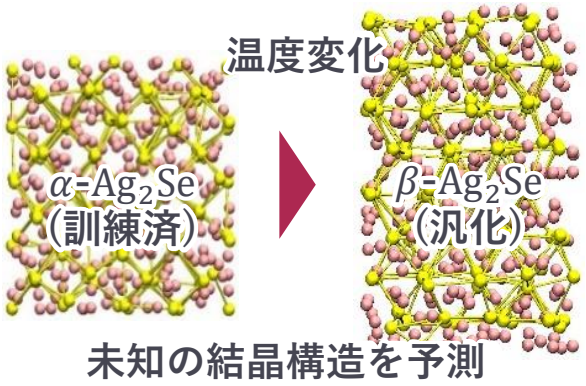


まとめ

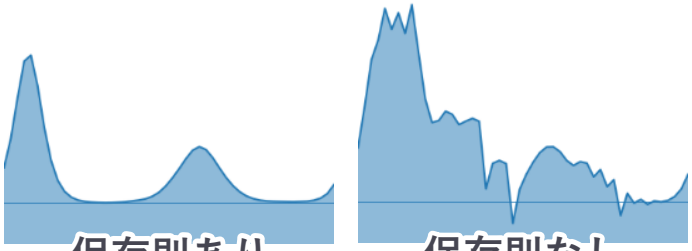
- **プロンプトの意図に忠実な画像生成のためのガイダンス**
  - プロンプトの意図を述語論理で表現
  - ファジィ論理に基づくガイダンスを導入
- **自動設計の可能性**
  - 今回の実験は命題論理を手動設計
  - 構文解析でほぼ自動設計可能
    - 名詞なら共存，修飾語なら含意，所有を意味する動詞なら含意
    - “with”が共存か所有か区別できない
- **想定される拡張**
  - SDXLはアテンションが物体に反応していないので要改善
  - 二項述語，位置関係，時間関係(時相論理)
  - 公平性への応用(人種や性別の組み合わせに堅牢な生成)
- **その他**
  - 論文に掲載した画像は約500枚．学生と2人でやる仕事量じゃない

科学技術機械学習

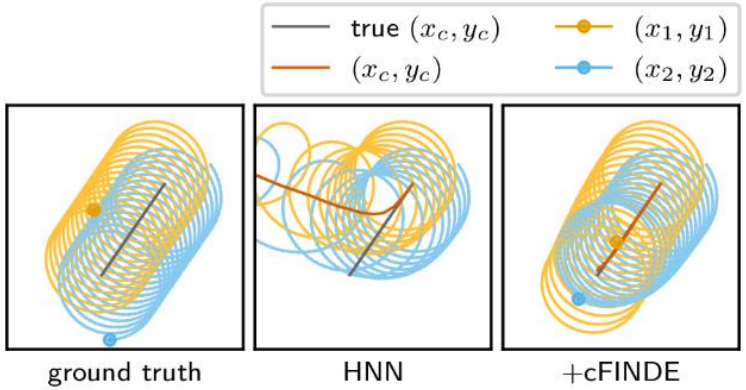
- 科学技術機械学習 (scientific machine learning; SciML)
  - 科学技術計算×機械学習
  - データから力学系をモデル化＆物理法則を同定
  - 計算機シミュレーションの高速化＆物理法則への忠実化
  - ロボットなどの制御への応用も



未知の結晶構造を予測



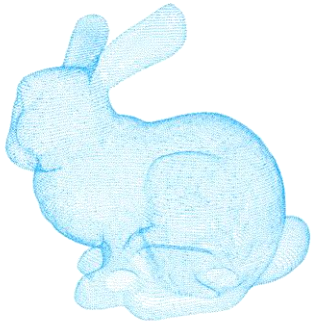
保存則に忠実なモデル化



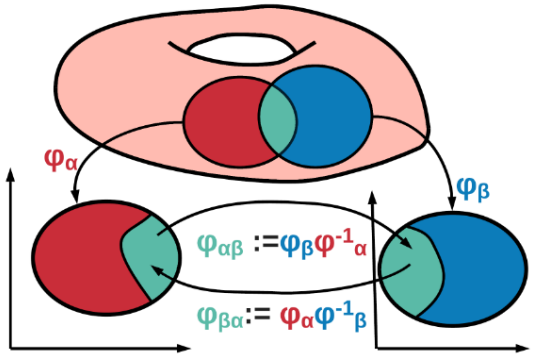
データから物理法則を同定

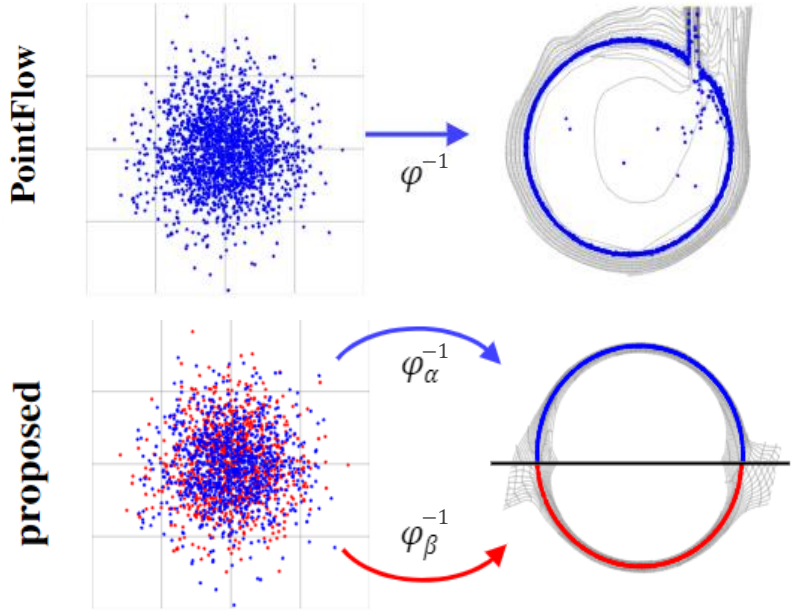
その他の研究：位相構造を利用した点群生成

- **点群は物体の表面を表現する方法**
  - ボクセルより高解像度，メッシュより処理が容易
  - 点群の生成は補間やノイズ除去に応用可能



- **表面なので位相構造を持つ**
  - 単一の写像では表現しきれない
  - 可微分多様体に倣って
    複数の写像で表現

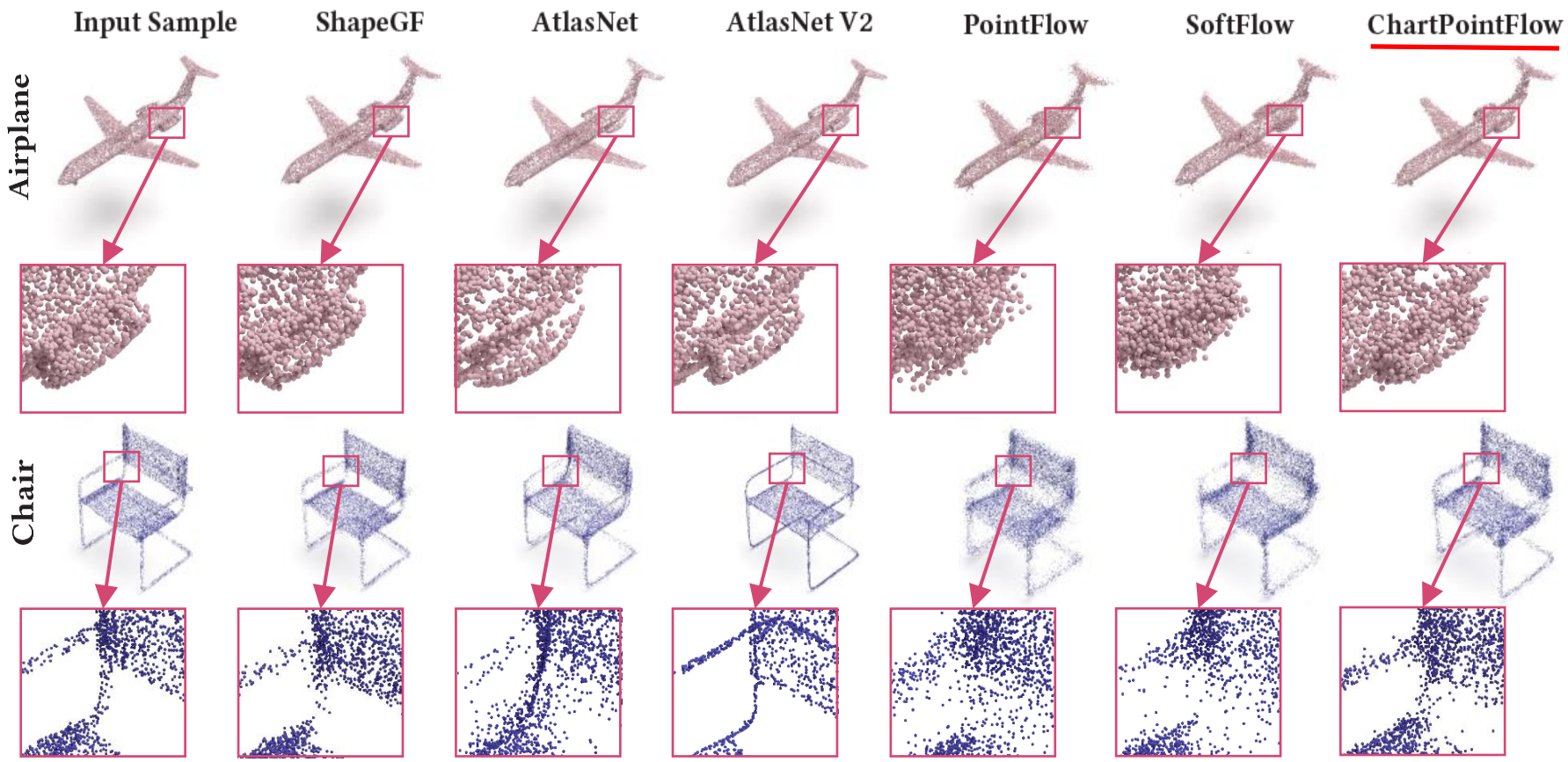




PointFlow: 3D Point Cloud Generation With Continuous Normalizing Flows(ICCV2019), arXiv:1906.12320
ChartPointFlow for Topology-Aware 3D Point Cloud Generation (ACMMM2021) arXiv:2012.02346

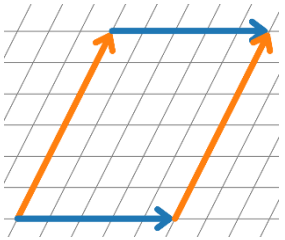
その他の研究：位相構造を利用した点群生成

- 小さな突起物や穴に強い点群生成を実現

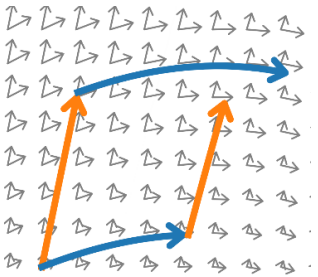


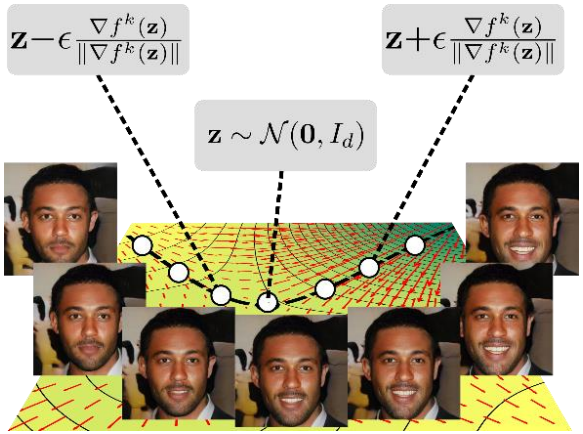
その他の研究：潜在変数の構造と画像編集

- 生成モデル(GAN)の潜在変数を操作する
  - Word2vecみたいに線形な操作を想定する
    - 線形な深層学習には厳しすぎる過程
    - 実データは歪んでいるので，きれいに並ぶことは考えにくい



  - ベクトル場を積分する
    - 非線形だが，ベクトル空間の構造を無視している
      - 非可換である，つまり$z + a + b \neq z + b + a$.

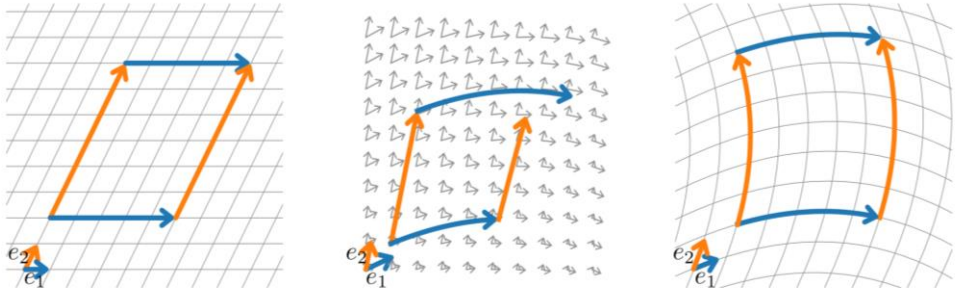


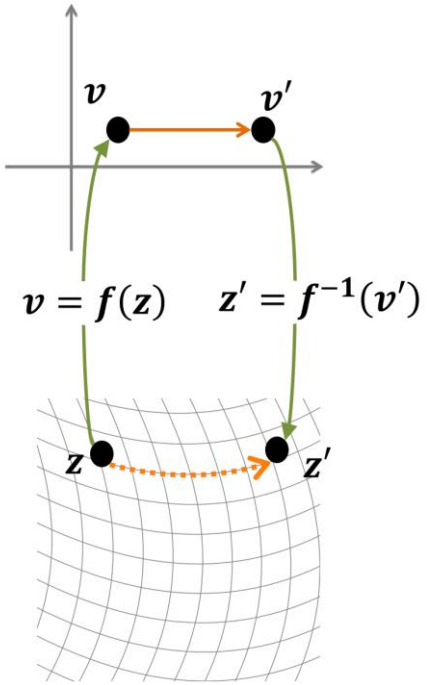


Unsupervised Discovery of Interpretable Directions in the GAN Latent Space (ICML 2020), arXiv:2002.03754
Finding Non-Linear RBF Paths in GAN Latent Space (ICCV 2021), arXiv:2109.13357

その他の研究：潜在変数の構造と画像編集

- 可換なベクトル場＝曲線座標系
  - 可換なベクトル場のフローを用いて編集を定義
  - ベクトル空間の構造を保ちながら非線形化

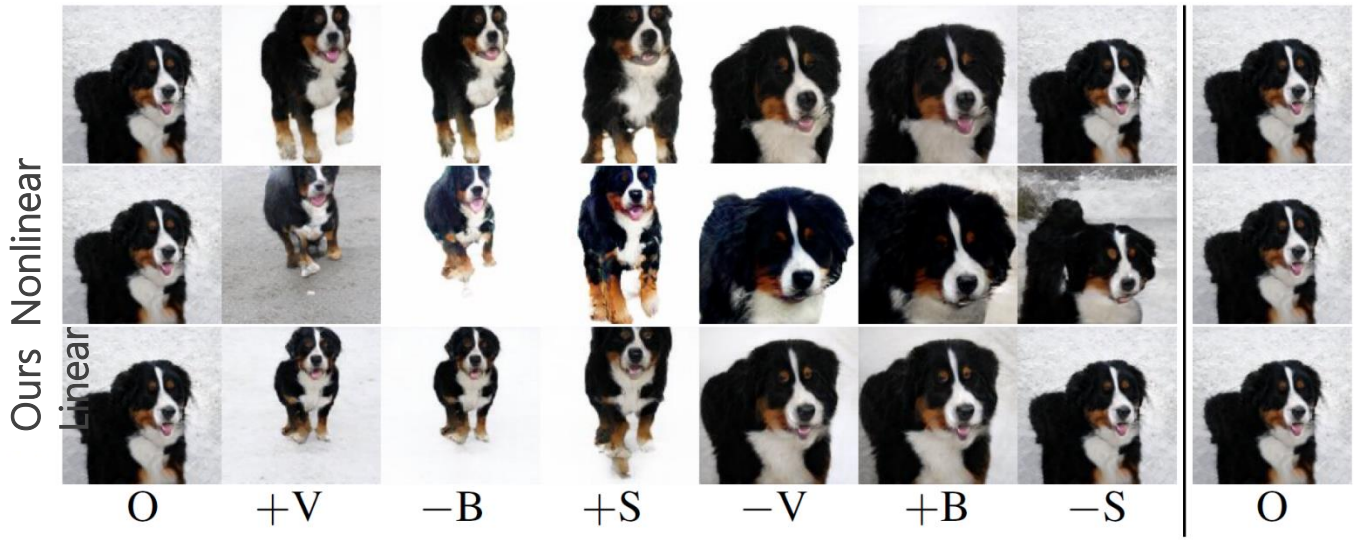
	Linear arithmetic	Vector fields/Local basis	DeCurvEd
Global coordinate	oblique	(only local)	curvilinear
No retraining	✓	✓	✓
Nonlinear edit	✗	✓	✓
Commutative edit	✓	✗	✓
Conceptual diagram			



Deep Curvilinear Editing: Commutative and Nonlinear Image Manipulation for Pretrained Deep Generative Model (CVPR2023), arXiv:2211.14573

その他の研究：潜在変数の構造と画像編集

- 可換性と編集精度を両立
  - disentanglementも促進



	精度	可換性
Ours	✓	✓
Nonlinear	✗	✗
Linear	✓	✓

AnimeFaces, hair color.

ProgGAN, smile.