第2回定期ミーティング

2024年5月10日(水)

早稲田大学　基幹理工学研究科
電子物理システム学専攻 史研究室
石黒将太郎・野口颯汰

アウトライン

- Stable Diffusionについて
- 本研究の目標
- 提案モデルの紹介
- 今後の計画
- 参考文献

モデル紹介(Stable Diffusion)

- Stable Diffusion[2]は、最も有名な拡散モデルで、ノイズを予測するNNとしてU-Netを採用し、画像とガイドプロンプトを紐づけるためにCLIPを利用

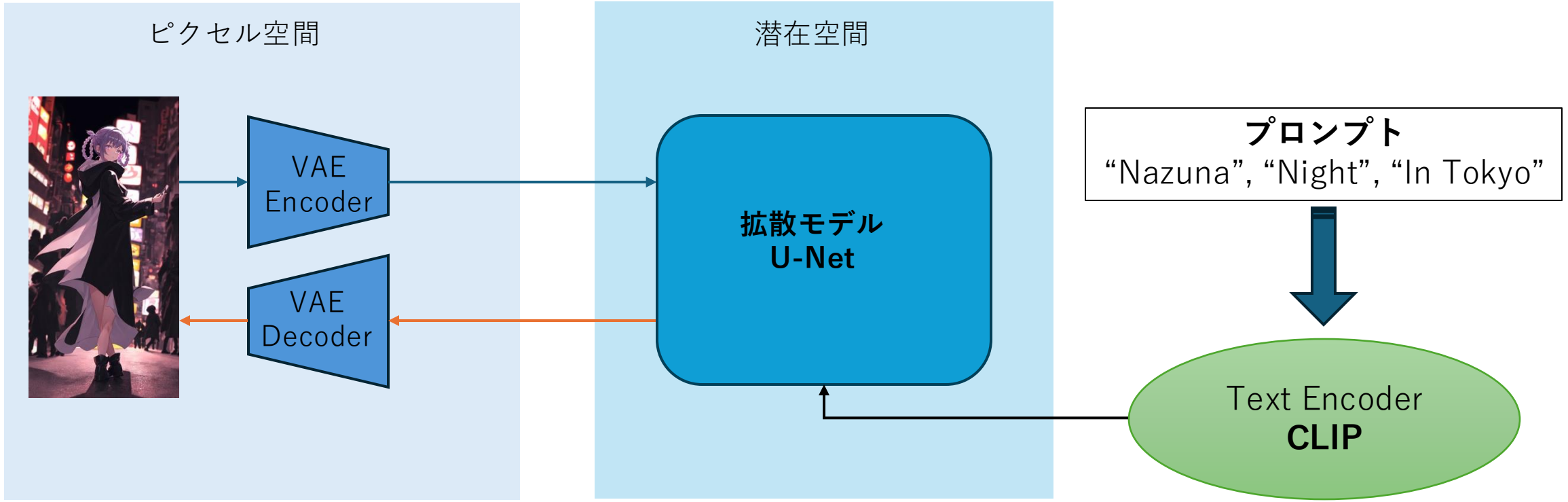

図:Stable Diffusionの模式図.

従来の拡散モデルの問題

- 拡散モデルは、GANに比べて複雑で多様な分布をモデリングでき、学習の不安定さも少ないが、人間が認識できないほどの高周波数特徴をモデル化してしまう
- パラメータ数が多く、RGB高次元空間での勾配計算により計算量が膨大になる
- 1つのモデルに知覚的圧縮と意味的圧縮を任せるのではなく、潜在拡散モデル(LDM)が意味的圧縮を行い、別のオートエンコーダが知覚的圧縮を行えば、勾配計算を全ての画像ピクセルで行う必要がなくなり、計算量を減らせる！

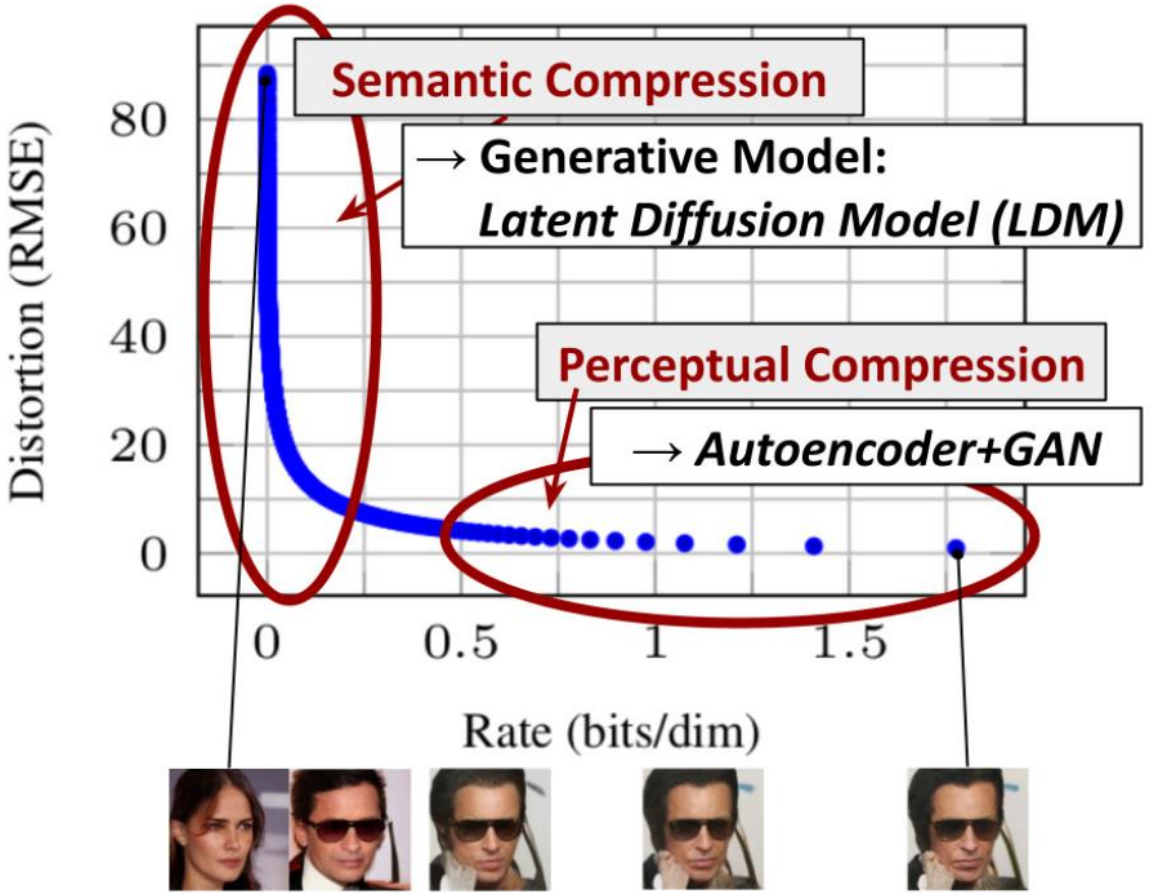

図: 意味的圧縮と知覚的圧縮関係性[1]

LDMのアーキテクチャ

- ピクセル空間にあるオートエンコーダは1回しか訓練する必要がない
- 機械的に3次元をつぶす手法と比べて、空間次元に対するスケーリング特性が向上した

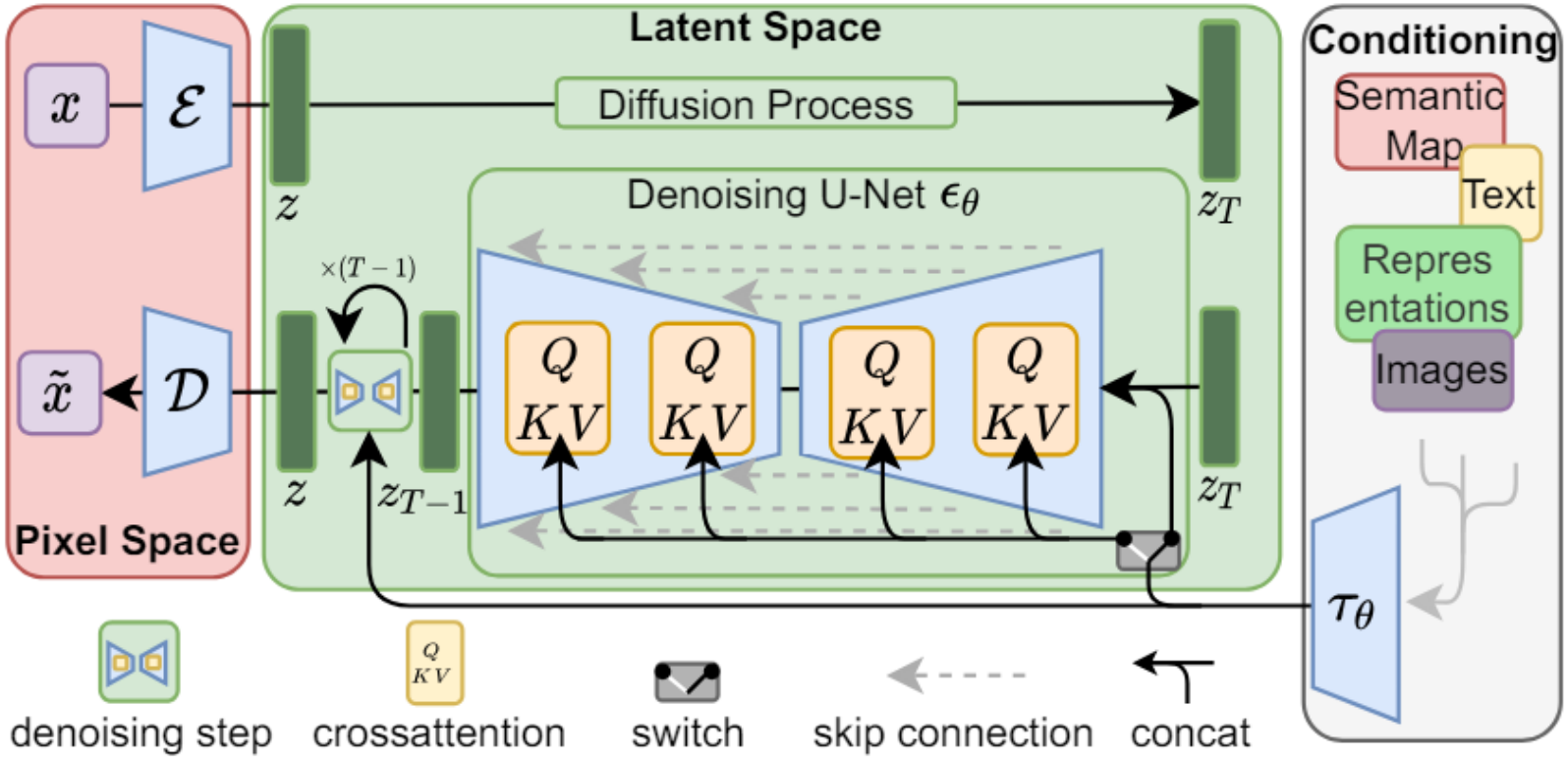

図: Stable Diffusionのアーキテクチャ[1]

オートエンコーダ

- 画像から画像、テキストから画像など、LDMを変更してもオートエンコーダの再学習は不要
- 知覚損失(特徴量比較)とパッチベースの目的関数を組み合わせて学習
- 潜在空間zの高分散化を防ぐため、標準正規分布に対してのKLダイバージェンスペナルティKL-regであると潜在空間を離散的な値に誘導するVQ-regを導入
- 入力画像をdown-sampling因子$f = \frac{H}{h} = \frac{W}{w} = 2^{\mathrm{m}}$でダウンサンプリング

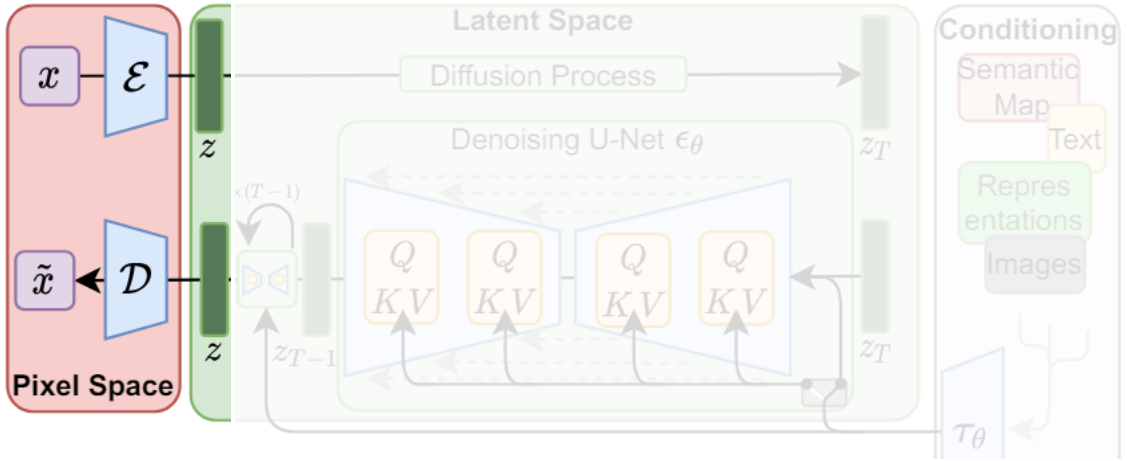

図: Stable Diffusionのアーキテクチャ[1]

Stable Diffusionにおける条件付け

- テキストデータや前処理済み画像データなどを別のエンコーダに入力することで得た特徴ベクトルを、U-Net内のクロスアテンションメカニズムに入れ込むことで条件として機能させることが可能

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \quad \Leftarrow \quad Q = W_Q^{(i)}\cdot\varphi_i(z_t),\ K = W_K^{(i)}\cdot\tau_\theta(y),\ V = W_V^{(i)}\cdot\tau_\theta(y).$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

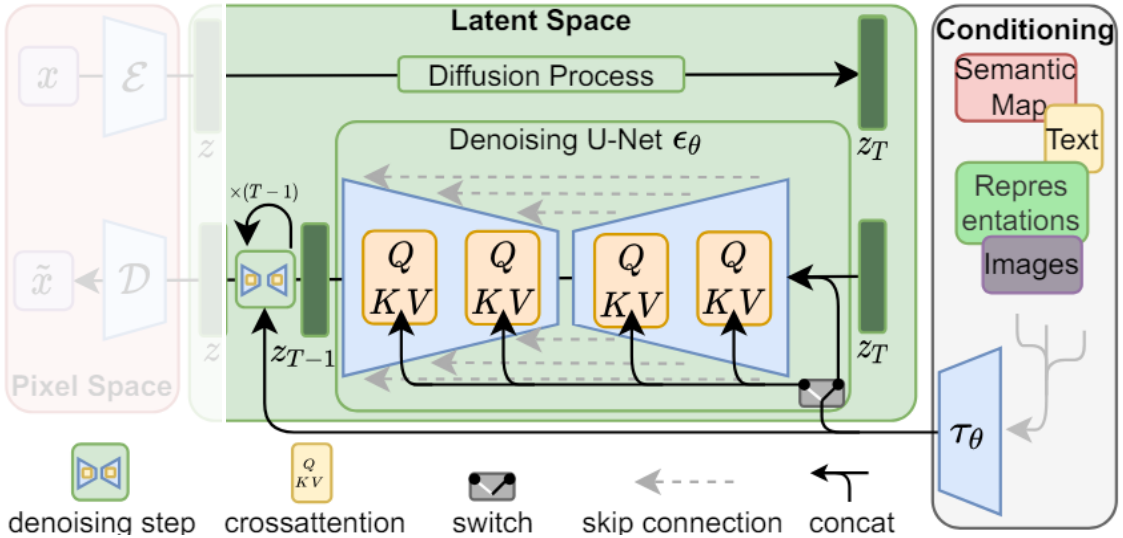

図: Stable Diffusionのアーキテクチャ[1]

LDMの評価(知覚的圧縮f)

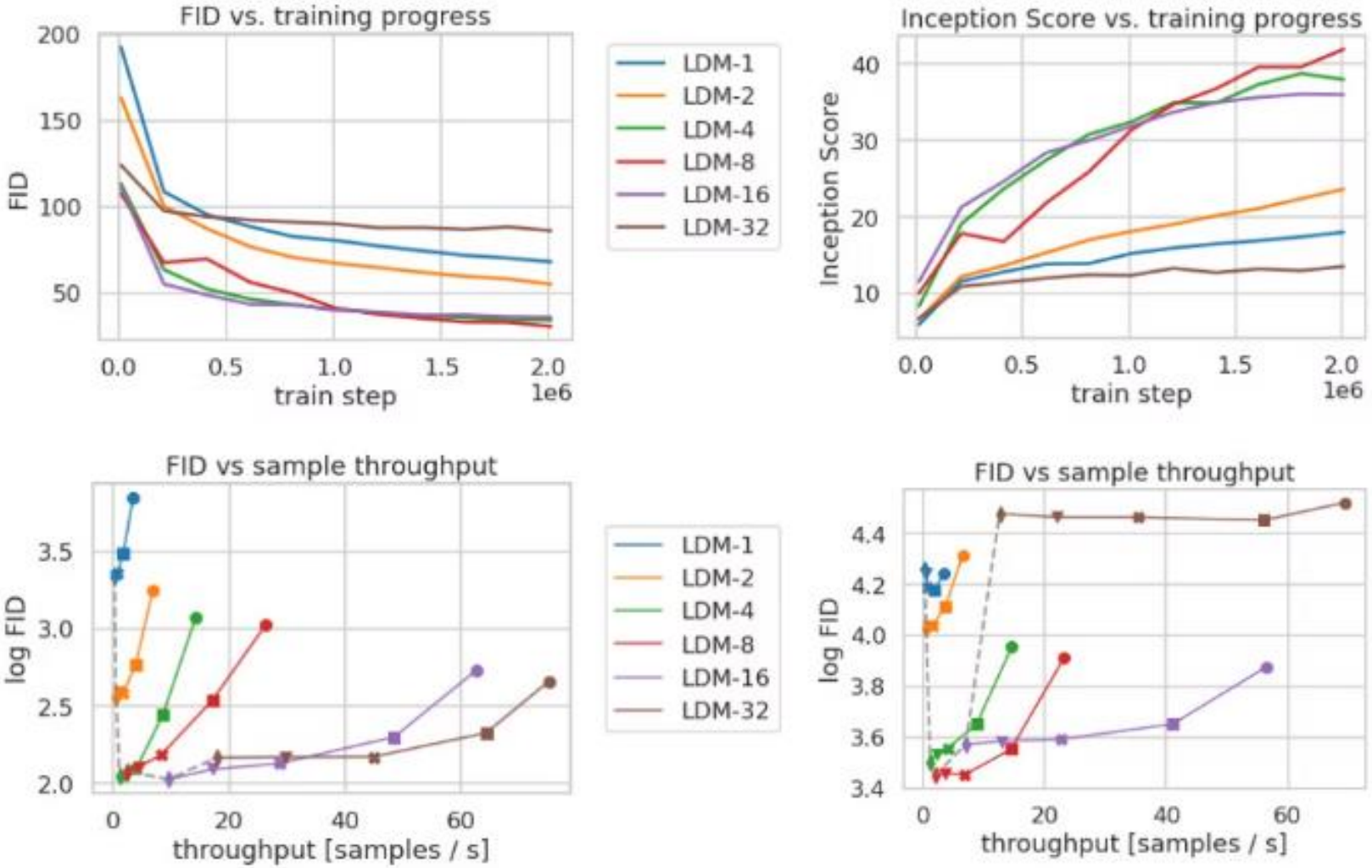

図: fを変化させたときのモデルの性能[1]

LDMの評価(他モデルとの比較)

CelebA-HQ 256 × 256				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T. [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	**3.08**	0.65	0.46
LDM-4 (ours, 500-s†)	**5.11**	0.72	0.49	*LDM-4* (ours, 200-s)	4.98	**0.73**	**0.50**

LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	**Method**	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN [41]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	**0.66**	**0.51**
ProjectedGAN [76]	**1.59**	0.61	0.44	ProjectedGAN [76]	**1.52**	0.61	0.34
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**	*LDM-4* (ours, 200-s)	2.95	**0.66**	0.48

無条件画像生成

テキストガイド付き画像生成

Text-Conditional Image Synthesis				
Method	FID ↓	IS↑	N_{params}	
CogView† [17]	27.10	18.20	4B	self-ranking, rejection rate 0.017
LAFITE† [109]	26.94	26.02	75M	
GLIDE* [59]	12.24	-	6B	277 DDIM steps, c.f.g. [32] $s = 3$
Make-A-Scene* [26]	**11.84**	-	4B	c.f.g for AR models [98] $s = 5$
LDM-KL-8	23.31	$20.03_{\pm 0.33}$	1.45B	250 DDIM steps
*LDM-KL-8-G**	12.63	$\mathbf{30.29}_{\pm \mathbf{0.42}}$	1.45B	250 DDIM steps, c.f.g. [32] $s = 1.5$

図: 他モデルとの比較[1]

研究計画　～アバター生成～　候補①

- 概要
  1．3Dデータを機械的に複数視点の２D画像に変換
  2．2D画像を拡散モデルに入力
  3．ノイズ除去過程で3Dデータからの深度情報・テキスト情報を入力
  4．拡散モデルの出力をCNNで3Dデータに変換

- 背景
  - 拡散モデルは入力と出力の次元は不変
  - StableDiffusionがコード公開されているかつ計算量が少ない

- 問題点
  - ３D点群のデータを扱うエンコーダーを実現可能なのか
  - 骨格変えられないから表情変化が困難
  - CLIPが機能するか（エンコーダーが二つ）

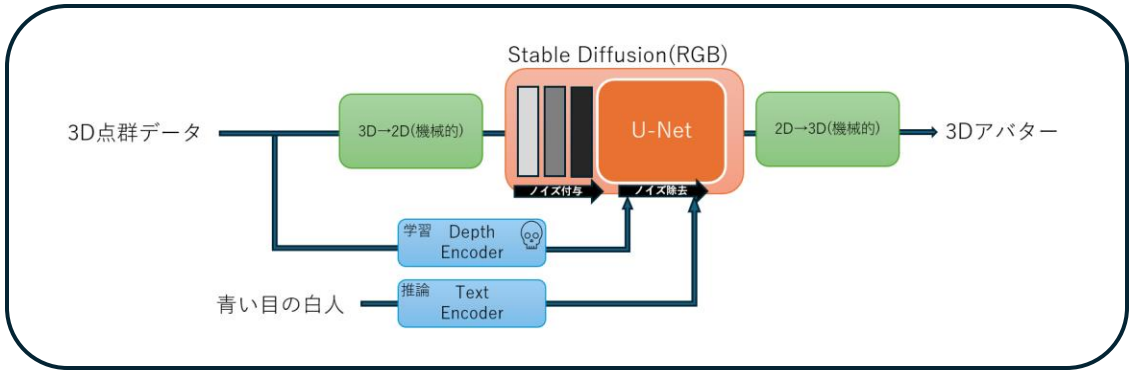

図1. 候補1概要図

研究計画　～アバター生成～　候補②

- 概要
  1．3DデータをCNNを用いて複数視点の潜在空間に変換
  2．潜在空間を拡散モデルに入力
  3．ノイズ除去過程でテキスト情報を入力
  4．拡散モデルの出力をCNNで3Dデータに変換

- 背景
  - 候補①の骨格不変問題
  - 候補①の複数エンコーダーによる条件付けの実現可能性

- 問題点
  - 3D→2Dモデルの実現可能性

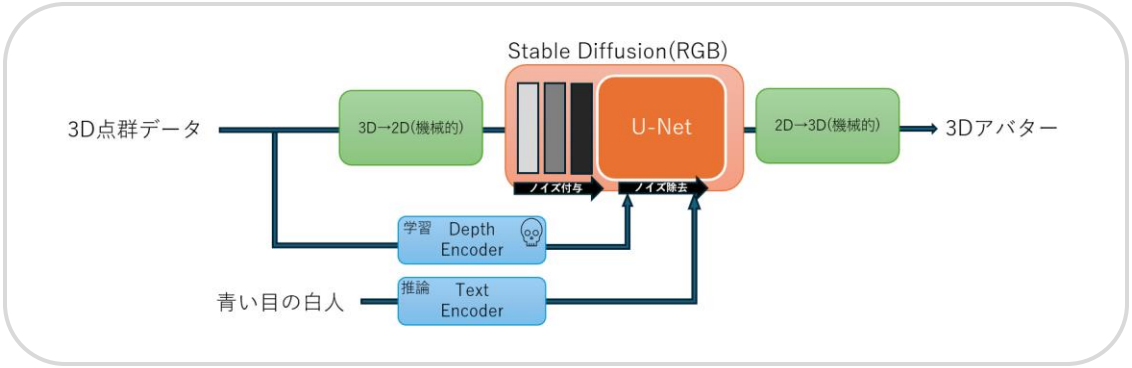

図1. 候補1概要図(再掲)

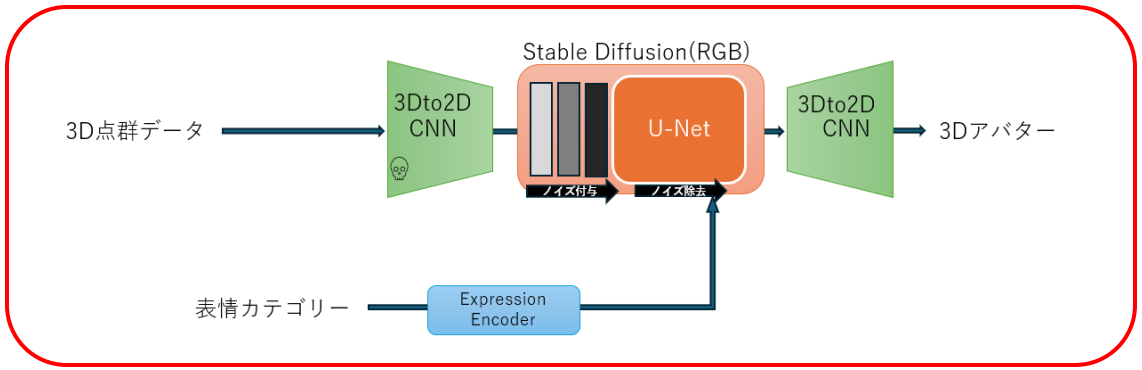

図2. 候補2概要図

研究計画　～アバター生成～　候補③

- 概要
  1．3D点群データをRGBD画像に変換
  2．　RGBD画像を拡散モデルに入力
  3．ノイズ除去過程で表情カテゴリー情報を入力

- 背景
  - 候補②の3D→2Dモデルの実現可能性
  - 候補①、②のテキストエンコーダーの自由度

- 問題点
  - 深度情報の影響度
  - カメラ変数問題
  - すべての候補において入力顔画像のアイデンティティをどう残すか

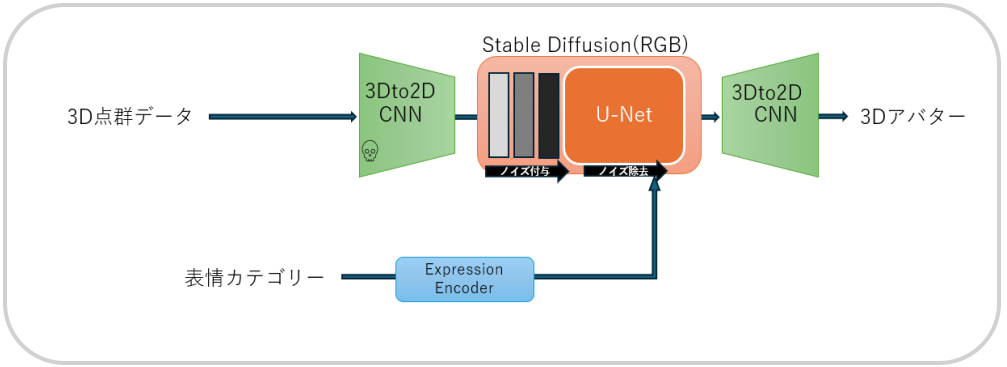

図2. 候補2概要図(再掲)

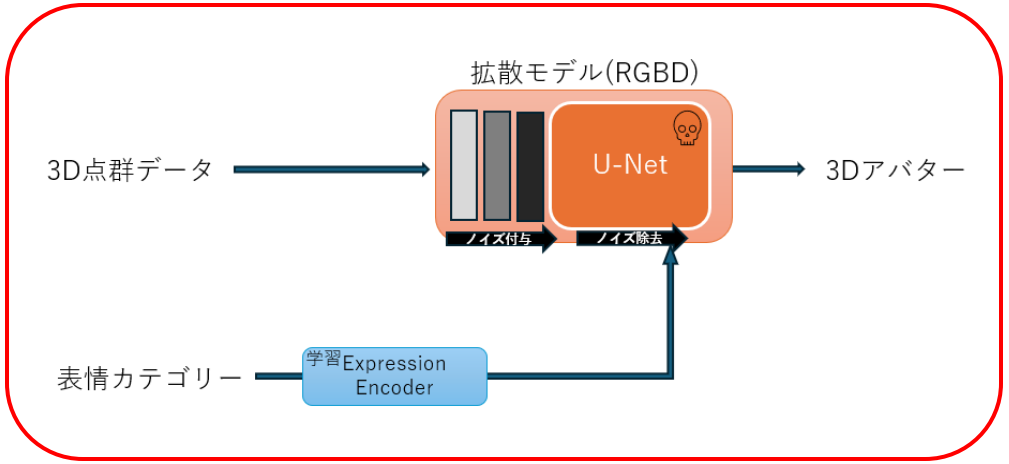

図3. 候補3概要図

今後の研究計画

表１.今後の研究計画

	5月	6月	7月	8月	9月	１０月	１１月	１２月
データセット考察	→							
拡散モデル調査	→	→	→	→	→	→	→	→
実装と検証		→	→	→	→	→	→	→

入力画像問題調査

- ✓ 3DMMを組み込む方法
- ✓ 既存拡散モデルの推論方法の調査

候補③実装・検証

- ✓ 学習時間・推論時間
- ✓ 深度情報の精度
- ✓ 表情の精度

深度情報欠如問題

解決方法：

- ✓ 3D畳み込み
- ✓ GCNを用いた部位ごとに分割して処理
- ✓ 疎なデータセットに対する畳み込みの工夫

参考文献

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer, “*High-Resolution Image Synthesis with Latent Diffusion Models”*, in CVPR 2022, 2022-04-13