指示文からの画像生成における配置指定能力の調査

木村 昴 [1] 工藤 慧音 [1] 赤間 怜奈 [1,2] 鈴木 潤 [1,2]
[1] 東北大学　[2] 理化学研究所
{subaru.kimura.s4, keito.kudo.q4}@dc.tohoku.ac.jp
{akama, jun.suzuki}@tohoku.ac.jp

概要

指示文から品質の高い画像を生成する画像生成モデルが大きな関心を集めている．現在，生成画像の品質や忠実さの向上に焦点を当てた研究が盛んに行われている．一方で，生成画像内の各物体の位置や関係性といった細かい指定を実現する方法については，未だほとんど取り組まれていない．本研究では，一般に広く使われている拡散モデルに基づく指示文からの画像生成方法の配置指定能力を調査する．単純な位置指定表現を加えた指示文により物体の配置ができるかを実験により検証する． 実験結果より，単純な位置指定表現では意図通りに作用しないことを報告する．

1　はじめに

現在，拡散モデル [1] が画像生成モデルに取り入れられたことにより，**指示文 (システムに入力されるプロンプト)** の内容を忠実に画像へ描写できる画像生成技術が向上している．例えば，生成したい画像に対する物体の指定や全体的な画風などの指定に対する生成画像の品質は以前に比べて格段に向上した．しかし，生成内容の配置を自由に指示文から指定することは容易ではない．図 1 の図下の例では，猫を画像内の左下に描写することを意図した指示文を入力としたときの実際の画像生成例を示している．生成画像を眺めると意図通りに猫を配置できていないことが分かる．一方で，本研究が目指す生成画像の理想は，図 1 の図上のように指示文通りに配置をすることである．

生成画像内容の配置を指定する研究は Generative Adversarial Networks (GANs) [2] によって画像生成が可能になってから徐々に研究され始めた [3]．その後は，物体領域やセグメンテーションを用いた配置図をモデルの入力に加えることで配置を指定する研究 [4, 5, 6, 7, 8] が報告されている．また，画像内の

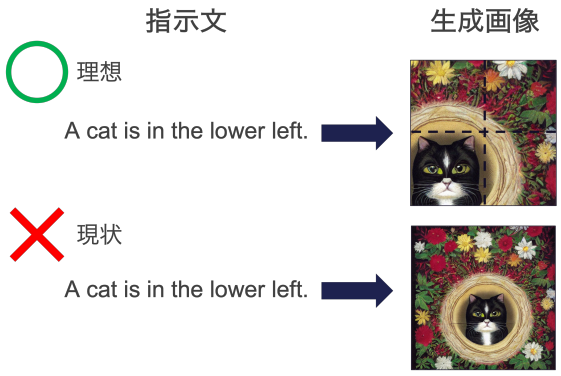

図 1　現状の Stable Diffusion における位置指定付き指示文での実際の生成例 (図下) と本研究で目指す指示文のみでの配置指定の理想像 (図上).

物体同士の位置および意味的な関係を表現したグラフを用いて多くの物体を含む複雑な文から忠実な画像を生成する研究 [9] などが実施されている．これらの関連研究の内，モデル構造に拡散モデルを用いているのは最近の研究のみで [4]，その他は拡散モデル以外のモデル構造を用いている [3, 5, 6, 7, 8, 9]．拡散モデルを用いた画像生成において指示文のみで配置指定ができるかは明確に明らかにされていない．また，画像生成モデルが有用であるには，生成内容を高度に制御できることが望ましい．その上で，配置図などを作成せずとも指示文のみで配置指定を制御できることは有用であると考える．

そこで本研究では，一般に使われている Stable Diffusion [10] を対象に，「指示文のみでの配置指定」をすることが可能かを実験により確かめ，指示文からの画像生成モデルの配置指定能力を調査する．また，現在 Imagic [11] や Texutual Inversion [12], DreamBooth [13] などの小規模の微調整で生成内容を個別の目的に最適化する方法が注目されている．これらの手法 [11, 12, 13] と微調整に使うデータサイズや工夫は異なるが，単純な微調整の有用性を調べるために画像内の物体領域の位置情報を加えた説明文を用意し，その説明文と画像のペアを用いて微調

図 2 実験の流れ：Visual Genome データセットから説明文を作成し，その説明文を用いて Stable Diffusion で画像生成または説明文画像ペアで微調整してから画像生成する．

整した上で画像生成した場合の結果についても調査する．

2 調査方法

今回調査対象とした画像生成モデルは Stable Diffusion である．次にモデルの入出力を述べる．図 2 のように，画像内の一つの物体に注目し，その物体の位置から説明文を作成する．ここでの説明文とは画像中心を原点とした 4 象限のどこに物体が属しているかを明示する位置指定文とその物体について記述した文の 2 文からなる文である．これを指示文として対象モデルに入力し生成された画像における注目物体の生成位置を評価し，対象モデルが指示文で配置指定が可能かを調査する．また，作成した説明文と作成元画像のペアのデータセットを用いて，Stable Diffusion モデルを微調整した複数のモデルを用意し，それぞれのモデルで指示文のみでの配置指定が可能かを調査する．調査する観点をまとめると以下の 2 点になる．

1. そのままの Stable Diffsuion モデルで指示文による注目物体の配置指定が可能か．
2. 説明文画像ペアのデータセットを用いて Stable Diffusion モデルを微調整したモデルにおいて指示文による注目物体の配置指定が可能か．

3 実験

2 章で述べた内容を実験的に調べる．

3.1 調査に用いるデータセット

今回作成するデータセットの元として，計 108,077 枚の画像と詳細な各画像内の物体の記述を持つ Visual Genome [14] を用いる．画像内の物体一つを選び出し説明文を Visual Genome の各画像から作成した計 94,079 枚の画像と説明文からなるデータセットを構築した．位置指定の種類は画像中心を原点として 4 象限に分けた時，どこに物体が属しているかで 4 種類ある．これらの位置と位置指定文の例の対応を表 1 に示す．

表 1 画像中心を原点として 4 象限に分けた位置と位置指定文の例の対応表．

位置	位置指定文の例
左上	Sign is in the upper left.
左下	Sign is in the lower left.
右上	Sign is in the upper right.
右下	Sign is in the lower right.

表 2 作成データセット内の選択物体領域の位置内訳．

	右下	右上	左下	左上
位置内訳	29 %	26 %	24 %	21 %

次にデータセット作成の手順を示す．

1. 画像の幅，高さが異なる画像を画像の左上頂点から幅と高さの内小さい方の長さで 1:1 に切り出す．
2. 画像中心を原点として 4 象限に分けた位置のどれかに物体が属している物体を一つ選び出す．選び出せない場合や物体についての記述がない画像については除外．
3. 選んだ物体の物体領域を拡大縮小し，学習・生成画像サイズである 512 に合わせる．
4. サイズ調整した物体領域の中心を算出．
5. 算出した中心が画像中心を原点として 4 象限に分けた位置のどれに位置するかで説明文を作成．
6. 作成した説明文画像ペアのデータセットを訓練，検証，評価セットに分割．このときの分割比は，Frolov らの研究 [8] にならい，訓練：検証：評価を 14:1:1 の比率とする．

作成したデータセット内の選択物体領域の位置内訳を表 2 に示す．また，データセット分割時は表 2 の位置内訳が保たれるように分割した．

3.2 調査に用いたモデル

一般に，指示文からの画像生成モデルは指示文を解釈するテキストエンコーダとその出力を用いて画像を生成する生成モデルで構成される．Stable Diffusion はテキストエンコーダとして OpenAI が発表した言語画像事前学習の Contrastive Language Image Pretaraining で事前学習されたモデルである CLIP [15] のテキストエンコーダを使用している．また，生成モデルとして潜在拡散モデル [10] を使用している．潜在拡散モデルは画像から潜在空間への符号化と潜在空間から画像への復号化をする

Variational Auto-Encoder (VAE) [16] と潜在空間で拡散モデルとして機能する Unet [17] から構成されている．また，Stable Diffusion は主に LAION-5B [18] の英語データセットで学習されている．

今回は，元の Stable Diffusion だけでなく，作成データセットの訓練データに含まれる説明文画像ペアで Stable Diffusion モデルを微調整したものを用意する．このとき，潜在拡散モデルの Unet だけを微調整する場合と，Unet に加えてテキストエンコーダも微調整する場合の 2 通りのモデルを用意した．

3.3 評価指標

生成画像を評価する指標として指示文で指定した位置に物体が配置されているかを評価する指標と画像の質の指標を用意する．

位置評価 指示文で指定した位置に物体が配置されているかを評価する指標として独自に評価指標を 3 種類用意した．Frolov ら [8] は，配置指定ができているかを評価するために指定物体領域を切り出し，その領域の説明文との CLIPscore [19] を計算している．CLIPscore とは CLIP を用いて説明文と画像の一致度を評価する指標である．しかし，この方法では切り出す領域以外に物体が配置されたときに評価できない．ゆえに，本研究では物体検出器を用いた独自の評価指標群を利用する．物体検出器には検出クラス数の制限がなく，文を入力として検出できる物体検出器 [20] を用いる．物体検出において検出物体が意図した物体かどうかを 0 から 1 で示す確信度の閾値は 0.01 とした．我々の位置評価では指示文に含まれる物体名で画像内の同一の物体を検出し，確信度が閾値を超えて検出された物体の物体領域全てについて正解画像の正解物体領域との IoU を求め，最大値をとる物体領域を検出物体の代表物体領域として採用する．このときの IoU を一つ目の評価指標とする．IoU とは Intersection over Union の略で (1) 式で計算される物体領域同士の重なり具合を評価する指標である．(1) 式の A, B は物体領域を表す．

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

二つ目の評価指標は，正解物体領域の中心と代表物体領域の中心との距離 D (単位は pixel) である．そして三つ目の評価指標は，正解物体領域の 4 象限での存在位置を正解データとして，代表物体領域の中心が同じ象限に生成されている場合に正解，そうでない場合に不正解としたときの正解率を採用する．

画像の質の評価 一般に使われている Inception Score (IS) [21] を用いる．IS は，画像の識別しやすさと多様性を評価する指標であり，高いほど良い．

4 実験結果および考察

4.1 結果

3.2 節で用意した 3 つのモデルに対して評価データの説明文を指示文とし 512×512 の画像を 5880 枚生成した．生成条件として，位置指定文を説明文から除くどうか．加えて，位置指定文がある場合に言語誘導係数 [22] を変える場合を足した 3 通りの場合について 3 つの各モデルで画像を生成した．言語誘導係数とは画像生成時の設定値の一つで指示文の内容をどれだけ画像内容に強く反映させるかの設定値である．この値を高くしたときの配置指定能力についても調査する．各モデルの各生成条件ごとに，生成した各画像の評価値の平均値を表 3 に示す．

位置評価の注釈 本研究の位置評価の有効性を考える．正解物体領域の選択元である評価データの画像 (正解画像) に対して考案した位置評価指標を用いて評価値の計算を行なった．これをオラクルの評価値とする．オラクルの評価値は検出物体領域が多少違えど同じ正解の物体領域を検出し，各値がベストに近い値を出すと考えられる．オラクルの評価値は表 3 一行目の評価データ画像の IoU，D，正解率の各値を見ると各指標のベストに近いため，我々の位置評価指標は有効であると判断した．この値を基準として，モデルの生成画像に対する評価値がどの程度になるかを測定し，評価を行なった．

しかし，今回は微調整モデルの生成画像の質が良くなかったため物体検出器の確信度の閾値を経験的に 0.01 と低くした．ゆえに検出対象以外の物体領域を誤検出しうるという欠点があることを注釈する．

4.2 考察

Stable Diffusion 表 3 より，Stable Diffusion の位置指定あり・言語誘導係数 7.5 での結果を見ると，位置指定ありで生成した時と位置指定なしの時を比較すると，IoU で見ると 2 倍ほどの差があり，正解との物体領域中心同士の距離 D が 136 pixel ほどと最も低いため数値上はある程度位置指定の効果があるとも考えられる．しかし，このときの Stable Diffusion の生成画像を眺めると同一物体を複数生成する場合が散見されたため，配置指定できたという

表 3　各モデルの各生成条件ごとに，生成した各画像の評価値の平均値. 微調整の有無・位置指定の有無・言語誘導係数を変えた計 9 つの組み合わせで，5,880 枚の評価データ画像から作成した説明文で生成した画像の評価結果．IoU_{all} は生成画像数，IoU，正解との物体領域中心同士の距離 D，代表物体領域中心が正解位置にあるかの正解率は検出画像数で平均を取った値である．位置指定の列の「-」は位置指定なしの説明文を，「✓」は位置指定ありの説明文を使うことを意味する．

	位置指定	言語誘導係数	検出画像数	IoU_{all}	IoU	D (pixel)	正解率	IS
評価データ画像	/	/	5,710	0.6336	0.6525	12.3	98.49	33.19
Stable Diffusion	-	7.5	4,071	0.0364	0.0526	160.1	39.16	**28.60**
	✓	7.5	3,620	0.0588	0.0957	**135.6**	52.30	26.49
	✓	15	3,690	0.0551	0.0879	138.0	50.90	25.39
→ 微調整 (Unet のみ)	-	7.5	3,906	0.0530	0.0797	161.9	49.04	21.61
	✓	7.5	4,114	0.0700	**0.1000**	152.4	56.39	19.93
	✓	15	**4,616**	**0.0773**	0.0985	146.7	**57.49**	25.62
→ 微調整 (Unet +テキストエンコーダ)	-	7.5	3,991	0.0517	0.0763	158.5	48.29	21.95
	✓	7.5	4,109	0.0634	0.0909	152.7	53.27	19.90
	✓	15	4,540	0.0691	0.0895	150.4	55.08	22.40

より，画像内に物体を大量に生成することでその中で正解物体領域と近いものが選ばれた可能性がある．また，正解率は 5 割でオラクルの評価値と比べると配置指定が正確にできているとは言えない．

微調整の結果　図 3 の生成画像の代表例を眺めると，元の Stable Diffusion で生成した画像は，図 3 の 1 行目の画像のように指示文の注目物体のみを大きなサイズで生成し，他のものを生成しない傾向がある．一方で，微調整したモデルで生成した画像は，図 3 の 2，3 行目の画像のように指示文には存在しない物体も生成する傾向があるため，検出画像数やその他の位置評価の値に差が出た可能性がある．また，テキストエンコーダも含めて微調整しても評価値や生成内容に大きな差は見られなかった．

言語誘導係数　Stable Diffusion モデルで言語誘導係数を 15 に上げても表 3 の評価結果や目視で確認した配置指定能力にも大きな差は見られなかった．一方で，目視で確認すると微調整したモデルの生成内容では指示文の注目物体を明確に生成する頻度が増えた．これに関連して表 3 の微調整したモデルでの検出画像数の増加や IS の向上が生じたと考える．

今後の展望　表 3 の結果より，微調整したモデルでは生成画像の IS が減少した．この原因として，微調整に使用した説明文画像ペアの内容の一致が十分でないことが学習の際に悪影響した可能性がある．そして，説明文画像ペアの内容を一致させるためには，説明文の作成時に画像内の物体の選び方を工夫する余地がある．例えば，今回のように画像内の物体を一つ選ぶのみでなく全体的に画像内容と一致するように複数選び出して説明文を作ることや，画像内の最大の物体を選び説明文を作ることが考えられる．また，配置指定として経験的に on や under での配置指定はできる傾向があったため，他の位置指定によっては配置指定可能な余地がある．

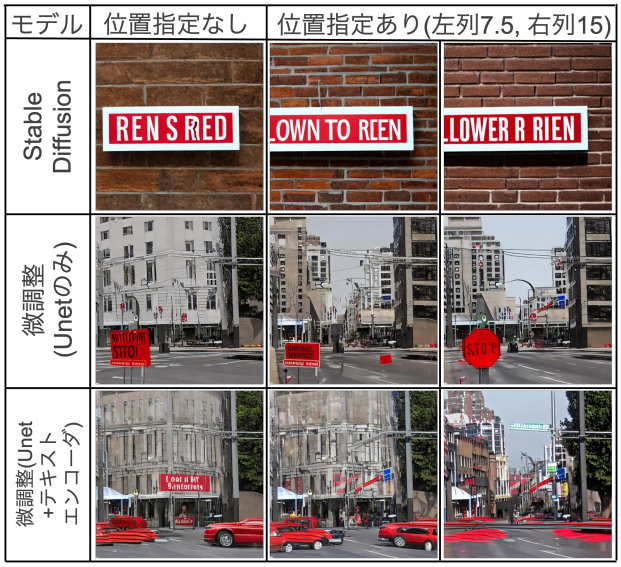

図 3　生成画像の代表例．指示文は位置指定ありの時，「Sign is in the lower left. Sign is red.」．位置指定なしの時，「Sign is red.」．右 2 列は言語誘導係数が各列で異なる．

5　おわりに

本稿では，一般に使われている Stable Diffusion の生成内容の配置指定能力を調査した．実験の結果から，今回調査した位置指定文では，指示文による配置指定が安直にはできないことを物体検出を用いた独自の位置評価で定量的に明らかにした．

今後は，データセットや指示文の改良，モデル構造自体を変える必要性について調査していきたい．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 (fundamental research) の助成を受けて実施されたものである．本研究を進めるにあたり，有益な助言を頂きました東北大学岡谷研究室の菅沼助教へ，記して感謝いたします．

参考文献

[1] Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. **CoRR**, Vol. abs/2209.00796, , 2022.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In **NeurIPS**, 2014.

[3] Scott E. Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw, 2016.

[4] Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John P. Collomosse, Jason Kuen, and Vishal M. Patel. Scenecomposer: Any-level semantic image synthesis. **CoRR**, Vol. abs/2211.11742, , 2022.

[5] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. 2019.

[6] Nan Liu, Shuang Li, Yilun Du, Joshua B. Tenenbaum, and Antonio Torralba. Learning to compose visual relations. In **NeurIPS**, 2021.

[7] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R. Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In **AAAI**.

[8] Stanislav Frolov, Prateek Bansal, Jörn Hees, and Andreas R. Dengel. Dt2i: Dense text-to-image generation from region descriptions. In **ICANN**, 2022.

[9] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In **CVPR**, 2018.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **CVPR**, 2022.

[11] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. **CoRR**, Vol. abs/2210.09276, , 2022.

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. **CoRR**, Vol. abs/2208.01618, , 2022.

[13] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. **CoRR**, Vol. abs/2208.12242, , 2022.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **International Journal of Computer Vision**, Vol. 123, pp. 32–73, 2016.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **ICML**, 2021.

[16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In **ICLR**, 2014.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In **MICCAI**, 2015.

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. **CoRR**, Vol. abs/2210.08402, , 2022.

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Joseph Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In **EMNLP**, 2021.

[20] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. **CoRR**, Vol. abs/2205.06230, , 2022.

[21] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[22] Jonathan Ho. Classifier-free diffusion guidance. **CoRR**, Vol. abs/2207.12598, , 2022.

表 4　作成データセットの詳細情報：各画像で選ばれた物体名の上位 10 位までの物体名とその個数.

上位 10 位の物体名と個数	「Sign」：1797，「Shirt」：1680，「Hair」：1659，「Light」：1287，「Hat」：1158，「Helmet」：1142，「Window」：1049，「Tree」：963，「Eye」：947，「Sky」：917

表 5　微調整時の学習設定.

設定値名	設定値
使用 GPU	NVIDIA RTX A6000 48 GB
使用 GPU 数	4
学習ステップ数	20000
学習バッチサイズ	3
解像度	512
勾配正規化最大値	1
学習率	0.00001
学習率スケジューラ	一定

表 6　画像生成時の設定値.

設定値名	設定値
推論ステップ数	50
生成画像サイズ	512 × 512
言語誘導係数	7.5 または 15
ノイズスケジューラ	PNDM スケジューラ
指示文あたりの生成画像数	1
乱数シード	42

参考情報

実験の実装をするにあたっては，Hugging Face の diffusers ライブラリ[1)]を用いた．また，Stable Diffusion のモデルについても Hugging Face 上で公開されているモデルを使用した[2)]．

作成データセットの詳細情報　表 4 に作成データセットの詳細情報として，作成データセット内の各画像で選ばれた物体名の上位 10 位までの物体名とその個数を示す.

微調整時の学習設定値　表 5 に微調整時の学習設定値を示す.

画像生成時の設定値　表 6 に画像生成時の設定値を示す.

評価指標一覧　以下に評価指標をまとめる.

- 検出画像数：画像の精細さや目的の物体を生成できているかが関係すると考えられる.
- $\mathrm{IoU}_{\mathrm{all}}$：正解物体領域と代表物体領域の重なり具合の指標．最大値 1．生成画像数を分母に平均を取った値.
- IoU：正解物体領域と代表物体領域の重なり具合の指標．最大値 1．検出画像数を分母に平均を取った値.
- D：正解物体領域中心と代表物体領域中心の距離．検出画像数を分母に平均を取った値．低いほど良い.
- Acc：検出代表物体領域の中心座標を四象限分類したときの正解率．検出画像数を分母に平均を取った値.
- IS：生成画像の識別しやすさと多様性を評価する指標．生成画像数を分母に平均を取った値．高いほど良い.

1）　https://github.com/huggingface/diffusers
2）　https://huggingface.co/runwayml/stable-diffusion-v1-5