拡散モデルを用いた脳活動条件による視覚体験再構成手法

石﨑文都 [1]　小林一郎 [1]
[1] お茶の水女子大学
{g2020505,koba}@is.ocha.ac.jp

概要

人間の脳活動から視覚体験を再構成する研究は，脳内視覚表象の理解と人工知能技術の発展の両面から注目されている．本研究では，Stable Diffusion の条件付け入力として fMRI 脳活動データを用いる手法を構築し，複数の条件下における画像再構成の精度を検証した．まず，脳活動を Latent Diffusion Model（LDM）の条件付け機構に合致するコンテキスト表現へ変換する Brain-to-Context（B2C）エンコーダを実装し，直接的な信号入力による画像生成の基礎的な性能を評価した．その上で，再構成精度に影響を与える要因を明らかにするため，二つの比較実験を行った．第一に，脳活動データを一括で変換する手法と，主要な脳領域（ROI）ごとに分割して変換する手法を比較し，脳信号の空間的な処理単位が再現性に与える影響を評価した．第二に，アノテーションに基づく CLIP テキスト埋め込みを B2C の損失関数に導入し，テキスト埋め込みとの意味的整合性の有無が再構成結果に及ぼす影響を検討した．実験の結果，出力に不明瞭さが残る場合があるものの，特定の条件下では対象物の概念や色彩といった意味的特徴の復元が確認された．これらの結果は，脳信号の空間的な分割や意味的な制約が視覚再構成の質を決定付ける重要な要因であることを示している．

1　はじめに

人間の脳は外界から得られる視覚情報を高度に処理し，物体認識や情景理解といった複雑な知覚体験を形成している．このような脳内の視覚情報処理を解明することは，神経科学における重要な課題であると同時に，画像認識や生成を中心とした人工知能技術の発展にも大きく貢献すると期待されている．

近年，脳活動データから視覚体験を再構成する試みとして，Latent Diffusion Model（LDM）を活用した研究が注目を集めている [1, 2]．特に，fMRI（functional Magnetic Resonance Imaging）と Stable Diffusion を組み合わせ，脳活動から潜在表現や言語特徴量を推定する手法は，高精細な画像再構成を可能にしている [1]．ただし，これらの手法ではテキスト特徴量や既存の条件付け機構を介して生成を行うため，表現の自由度がその中間媒体に依存する側面があり，脳内表象に含まれる情報を生成過程へ十分に反映する手法の検討が重要となる．

この課題に対し，本研究では Stable Diffusion を基盤とし，脳活動データを LDM の条件として直接利用する新たなアプローチを提案する．本研究の核心は，fMRI から得られた脳活動データを拡散モデルの条件付け機構に適合するコンテキスト表現へ変換する Brain-to-Context（B2C）エンコーダの導入にある．これを用い，テキスト特徴量を補助的に併用する場合と脳活動のみを用いる場合の比較，および脳活動データの処理（脳領域ごとの分割の有無等）による差異を検証する．これにより，脳活動が持つ空間的・意味的情報を直接的に拡散過程へ反映し，視覚体験の再構成精度をどこまで向上させ得るかを探究する．

2　提案手法

図 1 に，本研究で提案する fMRI により取得された脳活動データから視覚体験を再構成する手法の概要を示す．本手法は，事前学習済みの LDM（Stable Diffusion）[1)] を基盤とし，脳活動データを直接条件として画像生成を行うために B2C エンコーダを導入する点に特徴がある．従来の拡散モデルでは，画像生成を制御する条件情報としてテキストキャプションから得られる言語特徴量が用いられてきたが，本研究ではキャプションを用いず，その代替としてヒトの脳活動データを条件情報として利用する．

提案手法は，主に B2C エンコーダと LDM の 2 つの構成要素からなる．B2C エンコーダは，fMRI から得られた脳活動データを，LDM の条件付け機構に適合するコンテキスト表現へと変換する役割を担う．本研究では，Transformer を基盤としたエンコーダを用い

1） https://huggingface.co/CompVis/stable-diffusion-v-1-4-original

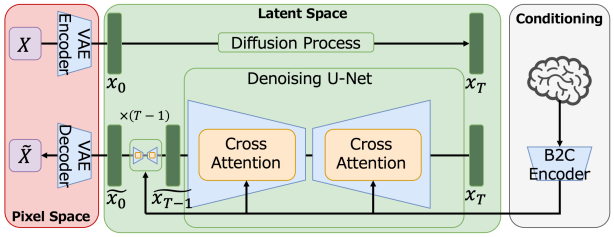

図 1　提案手法の概要図.

ることで，脳活動データに含まれる高次の意味情報を抽出し，拡散過程における条件情報として利用可能な表現を生成する．

事前に大量の画像とキャプションのペアデータを用いて学習された LDM は，高精細な画像生成能力を有している．本研究では，この事前学習済み LDM と B2C エンコーダを組み合わせ，画像と脳活動データのペアを用いて再学習を行うことで，脳活動データを直接拡散過程に反映させる．これにより，視覚体験に対応した内部表象を条件とした画像生成が可能となり，脳活動に基づく視覚体験の高精度な再構成を目指す．

2.1　拡散モデル

拡散モデルは生成モデルの一種であり，ランダムノイズからデータを生成することを目的とした手法である．完全なノイズから直接データを生成することは困難であるが，データを徐々にノイズへと変換する拡散過程と，その逆変換としてのデータ生成過程を定式化することで，この問題に対処している．

拡散モデルでは，元のデータに対して段階的にガウスノイズを付加する拡散過程を定義する．この過程は確率的に設計されており，学習を必要としない．一方，拡散過程の逆である逆拡散過程では，ノイズが付加されたデータから元のデータを復元することを目的とし，この過程におけるノイズ除去を行うネットワークを学習する．

逆拡散過程では，各ステップにおいてノイズ成分を予測し除去することで，ノイズの多い状態から徐々にデータ構造を回復させる．この枠組みにより，拡散モデルは画像や音声といった高次元かつ複雑なデータに対しても安定した学習が可能となり，高品質な生成結果を実現している．

Latent Diffusion Model　LDM [3] は，拡散モデルを元画像空間ではなく潜在空間上で実行するよう拡張した生成モデルである．一般的な拡散モデルでは，高次元の画像空間において拡散過程と逆拡散過程を繰り返す必要があるため，計算コストが大きくなるという課題がある．LDM では，この問題に対処するため，Variational Autoencoder（VAE） [4] を用いて画像を低次元の潜在空間へ圧縮し，その潜在表現に対して拡散処理を行う．

VAE による潜在空間への写像では，人間の知覚にとって重要な視覚的特徴を保持しつつ，冗長な情報を除去する知覚的圧縮が行われる．これにより，生成品質を維持したまま計算量を大幅に削減することが可能となる．LDM の逆拡散過程においてノイズ除去を行うネットワークには U-Net [5] が用いられ，潜在表現上で段階的にノイズを除去することで画像生成が行われる．

さらに，LDM では意味的圧縮の概念が取り入れられており，具体的なピクセル値の再現ではなく，抽象的な意味表現を保持した潜在表現を用いて生成を行う点に特徴がある．条件付き生成においては，与えられた条件情報が潜在空間の意味的特徴と注意機構を介して関連付けられ，生成される画像が条件情報と意味的に整合するよう設計されている．このような性質により，LDM は高品質かつ柔軟な条件付き画像生成を可能にしている．

Stable Diffusion　Stable Diffusion は，LDM を基盤として開発された拡散モデルであり，テキスト条件に基づく高精細な画像生成を可能にする．Stable Diffusion では，入力されたテキストから言語特徴量を抽出するために CLIP エンコーダ [6] が用いられ，得られた特徴量は U-Net 内の注意機構を介して潜在表現に反映される．この構造により，生成される画像は与えられたテキスト条件と意味的に整合するよう制御される．

本研究では，Stable Diffusion の学習済みモデルを基盤とし，従来用いられていたテキスト条件の代替として，B2C エンコーダによって変換された脳活動データを条件情報として用いる（図 1 中，Conditioning 参照）．具体的には，B2C エンコーダが出力するコンテキスト表現を，Stable Diffusion における注意機構への入力として与えることで，脳活動データを拡散過程に直接反映させる．

さらに，画像と脳活動データのペアを用いたファインチューニングをすることで，脳活動と生成画像との意味的一貫性を高める．これにより，Stable Diffusion が本来有している高精細な画像生成能力を保持したまま，脳活動データに基づく視覚体験の再構成を実現することを目指す．

2.2 Brain-to-Context エンコーダ

B2C エンコーダは，fMRI により取得された脳活動データを，LDM の条件付け機構に適合するコンテキスト表現へ変換するモジュールである．本研究では，Stable Diffusion の注意機構に入力可能な柔軟な条件表現を学習するため，Transformer エンコーダを基盤とした構造を採用する．構造設計にあたっては，Photonic Band における Material-to-Context エンコーダ [7] を参考とした．

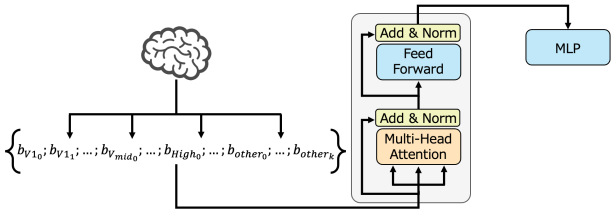

図 2　Brain-to-Context エンコーダとパッチ処理のモデル図.

B2C エンコーダでは，前処理された脳活動データを ROI ごとのパッチ単位へ分割するか直接入力するかのいずれかを経て各バッチを Transformer エンコーダブロックへ入力する．内部機構により，異なる脳領域間の相互依存関係や，脳活動に内在する時空間的特徴を統合的に捉えた表現が生成される．これらによって，単一ボクセルや局所的な活動に依存しない，高次の意味情報を含む特徴表現を獲得することが可能となる．

Transformer エンコーダの出力は，最終段の全結合層を介して次元変換され，Stable Diffusion の注意機構に直接入力可能なコンテキスト表現として整形される．このようにして得られたコンテキスト表現を拡散過程の条件情報として用いることで，脳活動データが持つ高次の意味情報を画像生成過程に直接反映させ，視覚体験に対応した画像再構成精度の向上を目指す．

3 実験

3.1 データセット

本研究では，大規模 fMRI データセットである Natural Scenes Dataset(NSD) [8] を使用した．NSD は，8 人の被験者が MS-COCO データセット [9] から選定された数千枚の自然画像を視聴した際の脳活動を記録したものである．本解析では，全画像を視聴した 4 名の被験者のうち subj01 のデータを対象とした．

脳活動データとしては，スライス時間補正や頭部運動補正，空間歪み補正が施された前処理済みの機能的画像（1.8mm 等方性ボクセル）を用いた．各試行における脳活動の定量化には，一般線形モデル（GLM）に基づき算出された単一試行ごとの β 値（betas_fithrf_GLMdenoise_RR）を使用している．この値は，GLMdenoise およびリッジ回帰を組み合わせることで，ノイズが低減された高精度な神経活動の応答強度を示す指標である．

解析対象とするボクセルの抽出には，NSD が提供する NSDGeneral ROI マスクを適用した．このマスクは，FreeSurfer[2] による構造画像解析に基づいて定義された機能的視覚領域を網羅し，かつ刺激に対して高い反応性を示すボクセルを選定したものである．本研究では，subj01 専用のこのマスクを通じて抽出された 15,724 ボクセルを，画像入力に対する脳活動表現としてデータセットに使用した．

3.2 Brain-to-Context エンコーダの構築

本研究では，B2C エンコーダとして Transformer エンコーダを採用した．入力には，NSDGeneral ROI に基づく前処理済み脳活動データを用い，入力形式として全ボクセルを直接入力する手法と解剖学的な脳領域ごとにパッチ化し，各パッチをトークン列として順次入力する手法の 2 種類を検討した．

エンコーダは，脳領域ごとの特性を保持するために設計された ROI ベースの投影層と，6 層の Transformer ブロックにより構成される．まず，入力された脳活動データは，解剖学的知見に基づく 4 つのグループ（V1, V2〜4, Higher Visual, Other）に分割される．各グループは独立した線形層によって投影され，合計 77 個の潜在トークン（各 768 次元）へと変換される．ここで，パッチ化を行わない場合には，脳活動データの分割をせずに潜在トークンへ線形射影される．次に，トークン間の順序および領域情報を付加するため，Transformer ブロックへの入力直前に Positional Encoding が加算される．各 Transformer ブロック内では，Multi-head Attention を通じて脳領域間の動的な相互作用や空間的特徴の統合が行われ，高次の意味表現が抽出される．本エンコーダは学習の安定化のために Pre-Norm 構造を採用しており，最終層の出力は Layer Normalization を経て，LDM の Cross Attention における Key および Value として供給される．

学習においては，U-Net 側の拡散プロセスに基づく損失に加え，B2C エンコーダの出力コンテキストと真の画像に対応する CLIP 埋め込みとの整合性を高めるための MSE およびコサイン類似度損失を組み合わ

2）https://surfer.nmr.mgh.harvard.edu/

せた多重損失を用いる場合を検証した．この最適化により，B2C エンコーダは脳活動データから Stable Diffusion の条件付けに最適な潜在表現への変換を学習し，高次の意味情報を画像生成過程に反映させることを目的とした．

3.3 LDM の学習

事前学習済モデル　本研究では，StabilityAI 社がリリースした LDM である Stable Diffusion バージョン 1.4[3)] を事前学習済み拡散モデルとして使用した．Stable Diffusion は，CLIP ViT-L/14 のテキストエンコーダーを用い，約 58 億のカラー画像とテキストのペアからなる LAION-5B データセット [10] のサブセット（laion-2B-en や laion-aesthetics など）で学習され，高精細なテキスト条件付き画像生成（text-to-image）を可能にする．

ファインチューニング　本研究では，Stable Diffusion の事前学習済み拡散モデルをベースとし，VAE のパラメータを固定した状態で U-Net のファインチューニングを行った．モデルへの条件付けに関しては，通常のテキストプロンプトによる入力を排除し，代わりに B2C エンコーダから出力された潜在コンテキストを入力として用いた．これにより，画像生成プロセスにおいて外部のテキスト情報を直接的な入力条件として介在させず，脳活動由来の情報のみによる条件付けを実現している．また，既存の知識を保持しつつ効率的に適応させるため，U-Net 内の Cross Attention 層における Key および Value のパラメータのみを学習対象とした．学習時の損失関数には，従来のノイズ予測に基づく平均二乗誤差に加え，拡散過程におけるノイズと元の信号（x_0）の速度成分をターゲットとする V-prediction アルゴリズムに基づく予測誤差と，予測された元画像（$\tilde{x}_0$）と真の画像（x_0）間の MSE $\tilde{x}_0$ とコサイン類似度および構造類似度（SSIM）による予測誤差の損失関数を用いることで，脳活動から得られるコンテキストに基づいた，より構造的整合性の高い画像生成を最適化した．

3.4 実験結果・考察

図 3 に脳活動からの再構成結果を示す．脳活動からの視覚像再構成において，脳領域ごとのパッチ化処理と CLIP 埋め込みによる意味的制約が，出力結果の質に与える影響を比較したものである．

CLIP 損失を導入せずパッチ化を行わない場合（左から 2 列目），テストデータの半数以上は真っ黒な画像が出力された．脳領域をパッチ単位で扱った場合（左から 3 列目）には，色味や大まかなコントラストは出始めているが，形状は崩壊しており，画像全体が未分化で波状のパターンに支配されやすい．また，CLIP 損失のみ用いること（左から 4 列目）でも，模様のようなものが見えるが，物体としては成立していない．CLIP 損失とパッチ化を併用した条件（左から 5 列目）において最も高い具象性が得られ，「人物」や「屋内」といった具体的なオブジェクトの概念が生成画像に付与された．さらに，Pairwise Identification は 0.7412 に達し，脳活動から抽出された意味情報が CLIP の埋め込み空間において高い精度で予測されていることを裏付けている（詳細は付録表 1 参照）．一方で，詳細な背景やオブジェクトの誤認は，CLIP の持つ強力な意味生成能力が，脳活動由来の空間情報を上書きしてしまったと考えられる．

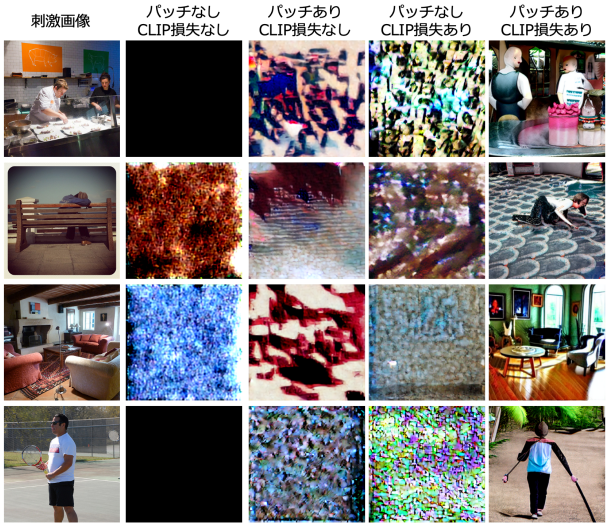

図 3　異なる学習条件による脳活動からの視覚像再構成結果の比較．（その他の出力例については付録 A.1 を参照）

4　まとめ

本研究では，ヒトの脳活動データを条件付け入力として Stable Diffusion に適用し，視覚体験に基づく画像再構成の精度を検証した．従来のテキスト入力に代わり，脳活動データを直接潜在拡散モデルの条件表現へ変換する B2C エンコーダを構築し，CLIP 損失の導入およびパッチ化の有無が再構成品質に与える影響を比較検討した．実験の結果，特定の条件下で視覚的再構成品質に劇的な改善が見られ，本手法が脳活動から画像生成をガイドするための構造的制約や意味的文脈を効果的に抽出できていることを示した．

3）　https://huggingface.co/CompVis/stable-diffusion-v-1-4-original

謝辞

本研究は，科研費・海外連携研究（24KK0189）の支援を受けた．ここに深謝する．

参考文献

[1] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 14453–14463, 2023.

[2] Kotaro Yamashiro, Nobuyoshi Matsumoto, and Yuji Ikegaya. Diffusion model-based image generation from rat brain activity. **PLoS One**, Vol. 19, No. 9, pp. e0309709–, 9 2024.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

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

[6] 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, 2021.

[7] Valentin Delchevalerie, Nicolas Roy, Arnaud Bougaham, Alexandre Mayer, Benoît Frénay, and Michaël Lobet. Towards photonic band diagram generation with transformer-latent diffusion models, 2025.

[8] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. **Nature neuroscience**, Vol. 25, No. 1, pp. 116–126, January 2022.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[10] 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, 2022.

A　追加の実験結果

本節では，本文の議論を補完するため，多様なサンプルに対する定性的な比較，および追加の評価指標を用いた定量評価結果を提示する．

A.1　追加の定性比較

本文で示した結果を補足するため，異なる刺激画像に対する再構成結果の追加分を図 4 に示す．

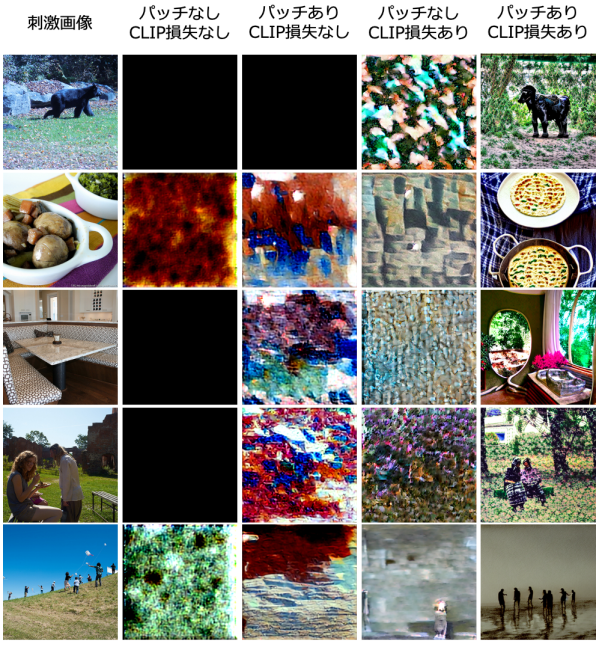

図 4　異なる学習条件による脳活動からの視覚像再構成結果の比較．

A.2　評価指標

本研究では，脳活動からの画像再構成を以下の 3 つの階層で評価した．

1. 画像レベル:　生成画像と正解画像の視覚的一致度を測定する．画素値の直接的な差を測る MSE，輝度パターンの相関を示す PCC（Pearson Correlation Coefficient），および人間の視覚特性に近い SSIM（Structural Similarity Index; 輝度・コントラスト・構造）を用いる．

2. 潜在空間レベル:　Stable Diffusion の VAE 潜在空間における再現性を評価する．潜在変数 z の各次元において MSE および PCC を算出し，VAE による可視化前の生成の種の予測精度を確認する．

3. 意味レベル:　B2C エンコーダが抽出した意味情報の精度を評価する．予測ベクトルと正解の余弦類似度を測る Cosine 類似度，および 2 ペアの脳活動とベクトルから正しい組み合わせを識別する Pairwise Identification により，個別の画像の意味的識別能力を測る．

A.3　定量比較

表 1 は，パッチ化（ROI ごとの分割）および CLIP 損失の有無による再構成性能を，画素・潜在空間・コンテキストの 3 つの観点から比較したものである．

まず，CLIP 損失を導入することで，意味レベルの指標（Cosine 類似度および Pairwise Identification）が劇的に向上した．特に，パッチ化と CLIP 損失を併用した条件において，Cosine 類似度は最高値である 0.5650，Pairwise Identification は 0.7412 に達し，脳活動から抽出された意味情報が CLIP の埋め込み空間において高い精度で正しく予測されていることが示された．

また，潜在空間においても，パッチ化と CLIP 損失の両方を用いた条件で MSE が最小（1.6619），PCC が最大（0.0112）となり，最終的な画像生成の基盤となるベクトル表現がより頑健に推定できていることが確認された．一方で，画像レベルの SSIM や MSE は条件間で複雑な挙動を示しているが，これは画素単位の単純な一致よりも，高次な意味情報の保持が再構成画像の質に大きく寄与しているためと考えられる．

表 1　異なる条件における画素（Pixel），潜在空間（Latent），およびコンテキスト（Context）の評価指標比較

条件		Pixel			Latent		Context	
パッチ	CLIP 損失	MSE ↓	PCC ↑	SSIM ↑	MSE ↓	PCC ↑	Cosine ↑	Pairwise ↑
		0.6725	0.0133	**0.0164**	3.0996	0.0095	0.0078	0.4984
✓		0.6096	0.0049	0.0105	2.7721	0.0063	0.01423	0.5214
	✓	**0.4825**	0.0221	0.0062	2.1907	0.0083	0.5055	0.6897
✓	✓	0.5608	**0.0466**	0.0103	**1.6619**	**0.0112**	**0.5650**	**0.7412**

注：↑は値が高いほど，↓は値が低いほど性能が良いことを示す．数値は小数点第 4 位までを表示．