画像キャプションを介した脳活動からの視覚体験再構成

高木優 [1,2*] 西本伸志 [1,2]
[1] 大阪大学 大学院生命機能研究科 [2] 情報通信研究機構

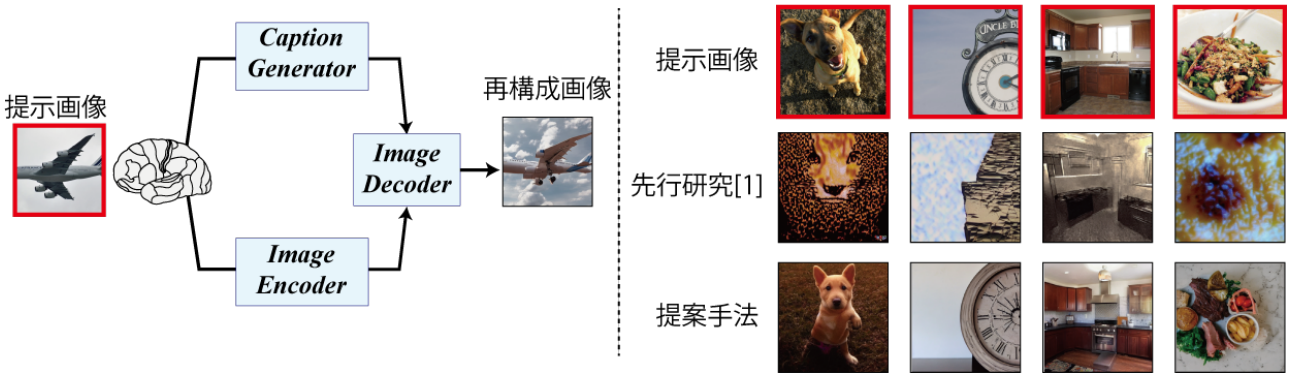

図 1: （左）脳活動からのキャプションを介した視覚再構成の概要図.（右）先行研究 [1] との比較

概要

ヒト脳活動から視覚体験を映像化する（視覚再構成）精度が，大規模自然画像データで学習した深層学習モデルを用いることで飛躍的に向上している．近年我々が提案した LDM（Latent Diffusion Model）を用いた手法 [1] は，LDM の潜在表現と脳活動との間に線形モデルを構築するだけで高解像度の視覚再構成が可能なことを示した一方で，生成画像がぼやける問題があった．本研究では，脳活動から生成された画像キャプションを介することで，高精細な再構成が可能となることを示す．また，キャプション生成モデルの内部表現と脳との対応関係を探ることで，モデルの内部で視覚情報が意味情報へと動的に変換されていく過程を定量的に示す．

1 はじめに

ヒト脳活動から視覚体験を映像化する（視覚再構成）精度が，大規模自然画像データで学習した深層学習モデルを用いることで飛躍的に向上している [2, 3, 4, 5, 6, 7, 8]．特に近年我々が提案した fMRI（functional Magnetic Resonance Imaging）により取得された脳活動データと LDM[9]（Stable Diffusion）を用いた手法 [1] は，LDM で用いる画像・意味を表現する潜在表現と脳活動との間にシンプルな線形モデルを構築するだけで，高解像度かつ意味的に妥当な画像を出力できる．また，深層学習モデルと脳活動との対応関係を探ることで，ブラックボックスである深層学習モデルを生物学的な観点から理解する研究も進んでいる [10, 11]．

LDM は，入力文に対応した高精細な画像を出力できる．それら高精細な画像と比較して，脳活動から生成された画像はぼやけてしまう問題があった [1]．ここで興味深いことに，脳活動から生成された画像はぼやけてはいるものの，元画像の意味情報はよく表現している．このことは，離散的なテキスト入力と対応する自然画像を用いて訓練された LDM が，意味潜在表現空間の全域で自然画像と対応しているわけではない，もしくは部分的に滑らかではなく，脳活動からの推定に伴うノイズに対して脆弱なことを示唆する．

我々は上記の観察を踏まえて，脳活動から画像の意味潜在表現を予測する問題を，脳活動からの画像キャプション生成問題へと置き換える．これによって，脳活動から連続値で推定されノイズを含む潜在表現を，ノイズの少ない離散的なテキストへと変換する．そうして予測されたテキストを LDM の入力に用いることで，高精細かつ意味的に妥当な画像を出力できるという仮説を立てた．

本研究ではこの仮説を検証するために，大規模データセットによって訓練されたキャプション生成モデル [12] を利用する．具体的には，視覚刺激提示中に fMRI から取得されたヒト脳活動を用いて提示

*連絡先：takagi.yuu.fbs@osaka-u.ac.jp

画像のキャプションを生成し，そのキャプションを介した視覚再構成を行った（図 1）．提案手法は，先行研究に比べて高精細かつ意味的に妥当な画像を出力することができた．また，キャプション生成モデルと脳活動との対応関係を探ることで，モデル内で視覚表現が意味表現へと変換される過程を定量的に示した．

2 関連研究

2.1 脳活動デコーディング

fMRI データを用いた視覚再構成（デコーディング）は，ノイズが多くサンプル数も少ないため一般に難しかった．だが近年，大規模自然画像で学習した深層学習モデルを用いることで高精度な再構成が可能となった [2, 3, 4, 5, 6, 7, 8]．これらの先行研究では，fMRI データまたは fMRI 実験で使われた刺激を用いた複雑な深層学習モデルの訓練・ファインチューニングを行う必要があった．

我々が提案した LDM を用いた手法 [1] は，LDM の潜在表現と脳活動との線形予測モデルを構築するだけで，高解像度かつ意味的に信頼度が高い視覚再構成ができることを示した．この手法は簡便かつ高精度な再構成を可能としたが，それでもなお，LDM にテキストを直接入力した場合に比べて画像がぼやける問題があった．

本研究で我々は，脳活動から生成されたキャプションを介することによって上記の問題を解決する．脳活動からのキャプション生成について，限定的なデータで試みた研究は少数存在するが [13, 14]，視覚再構成に利用した研究は存在しない．今回我々は，大規模データで訓練されたキャプション生成モデル [12] を用いて，簡便な手法で脳活動からのキャプション生成が可能なことを示す．加えて，キャプションを介することにより視覚再構成の精度が向上することを示す．

2.2 脳活動エンコーディング

深層学習モデルを生物学的観点から解釈するために先行研究は，深層学習モデルの異なる層から特徴量を抽出し，それぞれの特徴量から脳活動を予測するモデル（エンコーディング）を構築してきた．例えば，畳み込みニューラルネットワーク（CNN）の階層表現と，ヒトの視覚野の階層表現との対応などが示されてきた [10, 15, 16, 17, 11, 18]．

CNN などの視覚系モデルと比べ，言語など高次機能に関わるモデルと脳との関連を探る研究は少ない [19]．特に，キャプション生成モデルと脳との関係を調べた研究は存在しない．本研究では，キャプション生成モデルの各構成要素と対応する脳活動を探ることで，画像が言語情報へと変換されていく内部過程を生物学的観点から理解する．

3 提案手法

3.1 データセット

本研究では，Natural Scenes Dataset（NSD）を用いた [20]．[1]NSD は，7 テスラの fMRI 内で各被験者が MS-COCO から選定された 10,000 枚の画像を 3 回繰り返し見た際の脳活動データを提供している．本研究では，全画像を視聴した 4 被験者（subj01，subj02，subj05，および subj07）のデータのうち，公開されている 27,750 試行をデータとして用いた．このうち 2,770 試行（982 枚の画像に対応）は全被験者が同一の画像を視聴しており，これらの試行をテストデータに，残りの試行（24,980 試行）を訓練データに用いた．脳画像データとして，NSD が提供している前処理済の脳画像を用いた．関心領域（ROI）には NSD が提供する視覚野を用いた．詳細は A.1 を参照．

3.2 脳活動を用いたキャプション生成

本研究では，キャプション生成モデルである BLIP[12] を用いて，脳活動からのキャプション生成を行う．BLIP は，入力画像 $\mathbf{X}$ から Vision Transformer(ViT) 特徴量 $\mathbf{z}_\mathbf{v}$ を抽出し，そこから BERT[21] を利用した言語生成を行う (図 2 上段)．詳細は A.2 を参照．本研究では，脳活動から予測された $\mathbf{z}_\mathbf{v}$ を用いてキャプションを生成する．予測モデルの重みは L2 正則化線形回帰を用いて訓練データから推定し，その後テストデータに適用した．生成されたキャプションに関して，テキストベースの評価（BLEU-4）[22]，テキストと画像の類似度による評価（CLIP[23]），人手評価を行った．人手評価では 6 人の評価者に対して，ランダムに抽出した他の画像のキャプションと比べてどちらが提示画像をよく表しているかを回答させた (N=600 枚)．

1 http://naturalscenesdataset.org/

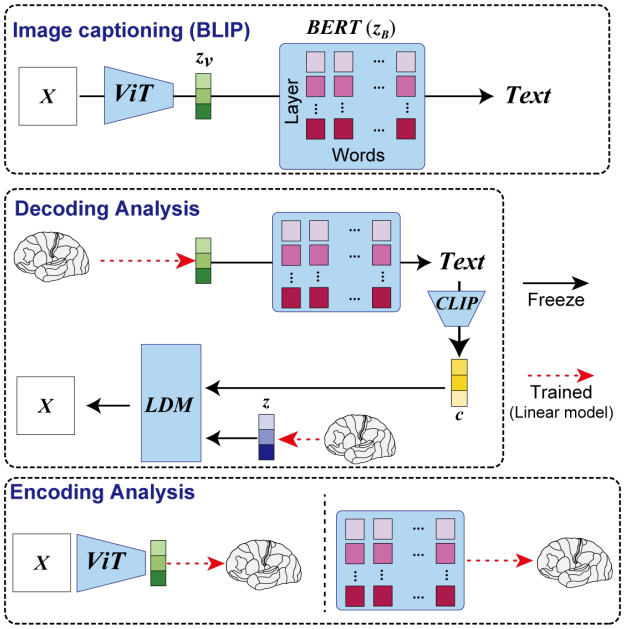

図 2: 提案手法の概要図.(上段) 本研究で用いるキャプション生成モデル BLIP.（中段）脳活動から生成されたキャプションを利用した視覚再構成の概要.（下段）BLIP の内部表現と脳活動との対応関係を探るためのエンコーディング解析の概要.

3.3 脳活動を用いた視覚再構成

本研究は，我々が近年提案した LDM（Stable Diffusion）を用いた視覚再構成手法を用いた [1](図 2 中段)．先行研究ではまず，入力画像 **X** をオートエンコーダのエンコーダに通した出力である潜在表現 **z** を脳活動から線形モデルで予測した．次に，画像に付随するテキストアノテーションを CLIP エンコーダに通し，その出力である潜在表現 **c** を脳活動から予測した．予測した **z** 及び **c** を LDM の逆拡散過程に通し再構成を行った．詳細は A.3 及び A.4 を参照．本研究では，**c** を脳活動から直接推定するのではなく，脳活動から生成したキャプションを介することで推定する点が先行研究と異なる．再構成画像に関して，画像ベースの評価（CLIP）と，人手評価を行った．画像ベースの評価については，先行研究による再構成画像と提案手法による再構成画像について，提示画像との CLIP 類似度を比較した．人手評価では，6 人の評価者に対して，先行研究による再構成画像と提案手法による再構成画像のどちらが提示画像とよく似ているかを尋ねた (N=400 枚).

3.4 脳活動エンコーディングモデル

次に，BLIP の内部表現について定量的に解釈するために，BLIP の各潜在表現と脳活動との対応関係を探った．そのために，BLIP で用いている ViT 特徴量 $\mathbf{z}_v$ と BERT 特徴量 $\mathbf{z}_B$ を用いて，それぞれから脳活動を予測するモデルを構築した (図 2 下段)．また，BERT の中で視覚的な情報が言語的な情報へと変換される過程を見るために，$\mathbf{z}_v$ と $\mathbf{z}_B$ の両方を組み込んだ予測モデルも構築し，予測に対する $\mathbf{z}_B$ の寄与度を検証した．全ての予測モデルは BERT の各層ごとに構築し，$\mathbf{z}_B$ の予測力が層を経るごとにどう変化していくのかを検証した．

予測モデルの重みは L2 正則化線形回帰を用いて訓練データから推定し，その後テストデータに適用した．評価には，予測 fMRI 信号と実際の fMRI 信号のピアソン相関係数を使用した．統計的有意性は，2 つの独立な乱数間の相関を比較することによって計算した．統計的閾値は $P < 0.05$ とし，FDR 法により多重比較補正を行った．

図 3: 脳活動から生成したキャプションの例．各行は異なる被験者の脳活動から生成されたキャプション.

4 結果・議論

4.1 脳活動からのキャプション生成

図 3 に，全 4 名の被験者それぞれの脳活動を用いて生成したキャプションの例を示す．被験者を通じて，提案手法はヒトの視覚体験をよく表現するキャプションを生成できていることがわかる．次に定量評価の結果を表 1 に示す．テキストベース指標，画像ベース指標，人手評価のいずれも，ランダムに選択された他のキャプションに比べて提案手法で生成されたキャプションは画像をよく説明することが示された.

表 1: キャプションの定量評価. 括弧内はチャンスレベル.

BLEU-4	11.06 (2.4 ± 1.7)
CLIP (cosine sim.)	0.57 (0.42 ± 0.02)
Human (%correct)	90.1 ± 3.5% (50%)

4.2 脳活動からの視覚再構成

図 4 に，キャプションを介して視覚再構成を行った場合と，介さずに視覚再構成を行なった場合の生成画像の例を示す．先行研究に比べ，キャプションを介した視覚再構成がより精細な画像を生成できていることがわかる．定量評価の結果，画像ベース指標の場合は 75.6%，人手評価の場合は 64.6% の割合でキャプション生成を介した再構成画像が既存手法による再構成画像よりも提示画像と類似していた．以上の結果より，キャプション生成を介した視覚再構成によって精度が高まることが示された．

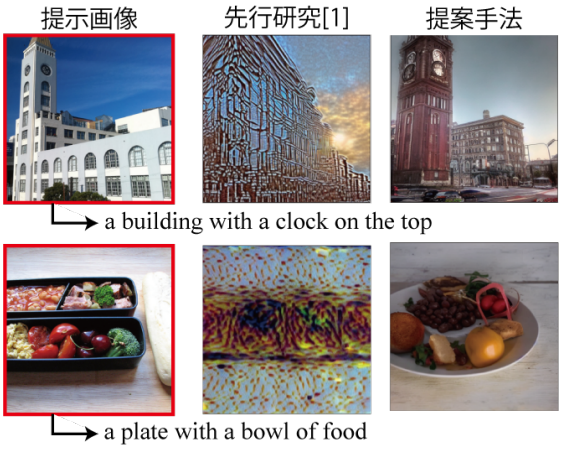

図 4: 脳活動から生成したキャプション（各画像の下に提示）を介した視覚再構成の例．図は単一の被験者の例．その他の被験者の結果は図 B.1 を参照．

4.3 エンコーディングモデル

図 5 に，単一被験者の脳活動を BLIP の潜在表現から予測した結果を示す．どちらの特徴量も視覚野（後頭部）をよく説明するが，ViT 特徴量 $\mathbf{z}_v$ はより後部の低次視覚野をよく説明し，BERT 特徴量 $\mathbf{z}_B$ はより前部にある高次視覚野をよく説明した．また，BERT 内は高次層おど予測力が高く，キャプション生成中期で最も説明力が高かった（図 6）．

次に，BERT の階層表現と脳部位の対応を検証するため，$\mathbf{z}_v$ と $\mathbf{z}_B$ を同時にモデルに組み込んだ場合の $\mathbf{z}_B$ の説明力の変化を探った．図 7 は，脳の各ボクセルを最も高い精度で予測した BERT の層番号を図示したものである．BERT 層が上がるに連れて，対応する脳領域も低次から高次へとシフトしており，意味的な情報が獲得されていくことがわかる．

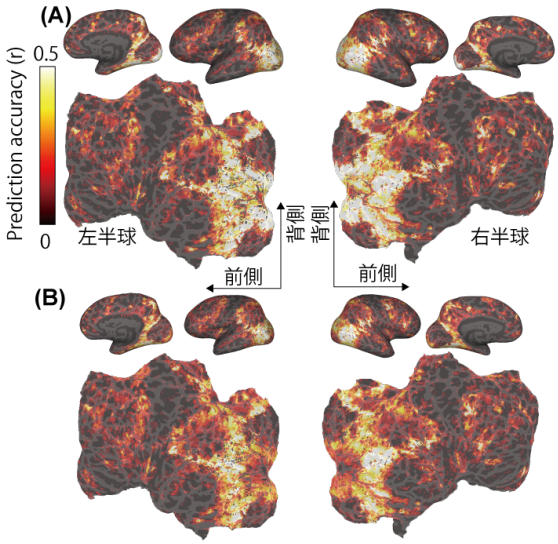

図 5: BLIP 特徴量を用いて脳活動を予測した結果．後頭初期視覚野を含む視覚野全般を予測する ViT (A) に比べ，BERT (B) は高次視覚野を中心に予測する．図は左右大脳皮質（上）およびその平面図（下）を表し，有意な予測を示したボクセルのみ色を付与．

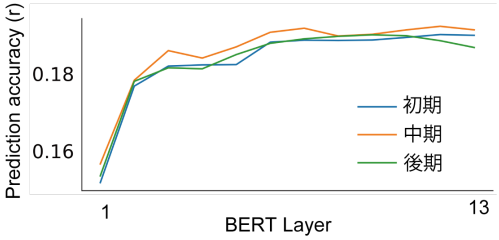

図 6: BERT 各層の表現を用いた視覚野活動予測精度の全ボクセル平均値を生成初期・中期・後期に分けて図示．

5 結論

本研究では，ヒト脳活動から視覚体験を映像化する手法について，脳活動から得られた意味情報をそのまま特徴ベクトルとして扱うのではなく一旦キャプションに変換することで画像生成精度が高められることを示した．また，急速に発展している画像・言語生成モデルの内部表現に生物学的観点から定量的解釈を提供した．

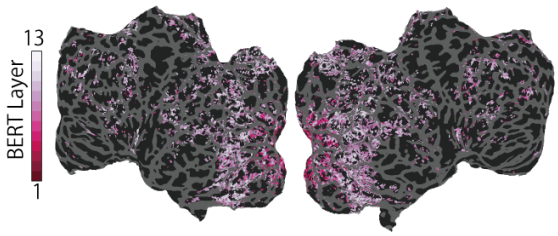

図 7: BERT 各層表現の予測精度．初期視覚野から高次視覚野へと予測力が高い領域がシフトしている．有意な予測を示したボクセルのみ色を付与．

謝辞

本研究はJSPS科研費19H05725，JP18H05522，JST CREST JPMJCR18A5，およびERATO JPMJER1801の助成を受けた．

参考文献

[1] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. **bioRxiv**, 2022.

[2] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. **NeuroImage**, Vol. 181, pp. 775–785, 2018.

[3] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. **PLoS Computational Biology**, Vol. 15, , 2019.

[4] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. **Frontiers in Computational Neuroscience**, p. 21, 2019.

[5] Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. **NeuroImage**, Vol. 254, , 7 2022.

[6] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. **Advances in Neural Information Processing Systems**, Vol. 32, , 2019.

[7] Lynn Le, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity. **BioRxiv**, 2021.

[8] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. **Advances in Neural Information Processing Systems**, 9 2022.

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 10684–10695, 2022.

[10] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. **Proceedings of the national academy of sciences**, Vol. 111, No. 23, pp. 8619–8624, 2014.

[11] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. **Journal of Neuroscience**, Vol. 35, No. 27, pp. 10005–10014, 2015.

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. **arXiv preprint arXiv:2201.12086**, 2022.

[13] Wei Huang, Hongmei Yan, Kaiwen Cheng, Chong Wang, Jiyi Li, Yuting Wang, Chen Li, Chaorong Li, Yunhan Li, Zhentao Zuo, et al. A neural decoding algorithm that generates language from visual activity evoked by natural images. **Neural Networks**, Vol. 144, pp. 90–100, 2021.

[14] Eri Matsuo, Ichiro Kobayashi, Shinji Nishimoto, Satoshi Nishida, and Hideki Asoh. Generating natural language descriptions for semantic representations of human brain activity. In **Proceedings of the ACL 2016 student research workshop**, pp. 22–29, 2016.

[15] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. **Nature communications**, Vol. 8, No. 1, pp. 1–15, 2017.

[16] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. **Cerebral cortex**, Vol. 28, No. 12, pp. 4136–4160, 2018.

[17] Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. **Proceedings of the National Academy of Sciences**, Vol. 116, No. 43, pp. 21854–21863, 2019.

[18] Iris IA Groen, Michelle R Greene, Christopher Baldassano, Li Fei-Fei, Diane M Beck, and Chris I Baker. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. **Elife**, Vol. 7, , 2018.

[19] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. **Nature neuroscience**, Vol. 25, No. 3, pp. 369–380, 2022.

[20] 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, pp. 116–126, 1 2022.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International Conference on Machine Learning**, pp. 8748–8763. PMLR, 2021.

A　手法

A.1　データセット

本研究では，Natural Scenes Dataset（NSD）が公開している前処理済み脳機能画像（解像度=1.8mm）を用いた．前処理には，スライス時間の差を補正する時間的リサンプリングと，頭部の動きと空間的歪みを補正する空間的補間が含まれている．NSD は，GLM で推定した 3 種類の単一試行ごとのベータ値を提供している．本研究では，*betasfithrfGLMdenoiseRR* を使用した．加えて NSD は，各被験者について複数の関心領域（ROI）を提供している．本研究では， *streams* アトラスに含まれる視覚野のうち，$\mathbf{z}_v$ および $\mathbf{c}$ の推定には視覚野全体を，$\mathbf{z}$ の推定には初期視覚野を ROI として用いた．テストデータでは各画像について 3 試行の平均値を用いた．訓練データでは，平均化せずに 3 試行をそのまま用いた．

A.2　キャプション生成モデル

本研究では，キャプション生成モデルとして BLIP[12] を用いた．fMRI データが刺激として MS-COCO を用いていることから，モデルは公式からリリースされている LAION-5B で訓練された Base モデル[2]を用いており，MS-COCO でファインチューニングされたキャプショニングモデルは用いていない． 各種パラメータについて，繰り返しに対するペナルティを 1 .5 に設定し，その他はデフォルトの数値を用いた．

A.3　LDM (Latent Diffusion Model)

本研究では，StabilityAI 社がリリースした LDM である Stable Diffusion のバージョン 1.4 を用いた．Stable Diffusion は，テキストで条件付けた高精細な画像生成（text-to-image）を可能にする．LAION-5B のサブセットを用いて学習され，CLIP ViT-L/14 のテキストエンコーダーが用いられている．コード及びパラメータは著者らが提供しているコード及びデフォルトパラメータを使用した．[3]

A.4　関連研究（Takagi and Nishimoto 2022）手法詳細

本研究では，我々が先行研究で行った LDM を用いた視覚再構成手法を用いた [1]．以下，再構成手法の概要を述べる．

まず，初期視覚野の活動を用いて，提示画像 $\mathbf{X}$ をオートエンコーダーのエンコーダーに通した潜在表現 $\mathbf{z}$ を予測した．次に，$\mathbf{z}$ をデコーダーに通し，320×320 の粗い画像 $\mathbf{X}_\mathbf{z}$ を生成し，さらに $\mathbf{X}_\mathbf{z}$ を 512×512 にリサイズした．次に，リサイズした $\mathbf{X}_\mathbf{z}$ をエンコーダに通し，さらに拡散過程を通すことで，ノイズを付加した潜在表現 $\mathbf{z}_\mathbf{T}$ を作成した．同時に，脳活動から生成したキャプションまたは視覚野全体から，CLIP エンコーダーの潜在表現 $\mathbf{c}$ を推定した．最後に，$\mathbf{z}_\mathbf{T}$ と $\mathbf{c}$ を用いて 512×512 の再構成画像を生成した．モデル構築には L2 正則化線形回帰を用い，すべてのモデルは被験者ごとに構築された．重みは訓練データから推定し，正則化パラメータは 5 分割交差検定を用いて訓練データ内で探索した．

B　全被験者の視覚再構成結果

図 B.1: 全被験者の結果. 上段が提案手法，下段が先行研究による再構成 [1]

2　https://github.com/salesforce/BLIP
3　https://github.com/CompVis/stable-diffusion/blob/main/scripts/