「心の中の言葉」はどのように予測できるか？ ―複数のモダリティの特徴に基づく 脳活動デコーディングプロセスの構築―

Muxuan Liu[1]　西田知史[2]　小林一郎[1]
[1] お茶の水女子大学大学院
[2] 情報通信研究機構 未来 ICT 研究所 脳情報通信融合研究センター
{liu.muxuan,koba}@is.ocha.ac.jp　s-nishida@nict.go.jp

概要

本研究では，被験者が画像観察時に想起する内容（想起文）の脳活動表現を解明するための新たな手法を模索し，fMRI 信号と LLaVA による画像と言語の統合表現を用いたデコーディングプロセスを構築した．特に，LLaVA が生成するトークン表現と fMRI 信号の時間軸不一致問題に対して，滑動平均による信号のノイズ低減とリッジ回帰による関連性学習を組み合わせることで，解読手法の整合性を向上させた．さらに，生成方式の違いがモデル性能に与える影響を検討し，想起文の脳活動表現の初歩的なデコーディングプロセスとその可能性を示す結果を得た．

1　はじめに

画像を観察しながら心中で生成される言語内容を解読することは，従来の音声聴取タスクに比べて困難が伴う．音声刺激データは明確な時間軸を持ち，脳信号と時系列的に整合するため，比較的簡単な手法である BOLD 信号の遅延補正やスライディングウィンドウで処理できる．また，音声の時間変動は語音の認識に重要な役割を果たし，例えば，ポッドキャストの再生に基づくデータでは，文を単語やそれ以下の言語構造の最小単位に切り分け，単語速度の予測が可能となる．例えば，Tang ら [1] の研究では，テキストを単語レベルの単位に切り分け，各単語を意味空間の 1 点にマッピングすることで，異なる語彙的概念に対する脳の応答を予測するための意味モデルを構築した．一方，画像観察による文想起タスクには明確な時間軸がなく，低時間分解能の fMRI 信号を用いた時系列モデルの適用が難しい．このため，画像刺激に対する脳活動の時間的特性を捉えるには，異なるアプローチが必要となる．

本稿では，マルチモーダルモデルの特性を活用し，画像刺激を通じて被験者が想起する文（想起文）の解読を試みる．また，fMRI 実験設計では，画像刺激が固定的な時間に提示されることが多いため，fMRI 信号は固定された時間軸を持つデータ（例：形状は (8TR, 93452 ボクセル)）として記録される．一方で，被験者が生成する想起文の長さが不特定であるため，言語モデルで特徴を抽出する際のトークン数も一定ではなく，（異なるトークン数，固定している言語モデルトークンサイズ）のように第一次元のサイズが異なるデータ形式となる．このため，画像観察中の想起文を解読するには，時間軸の曖昧さや言語生成プロセスの不規則性といった課題を解決する必要がある．これらの課題を解決するため，マルチモーダルな統合表現を用いて画像観察時の想起文解読に新たなアプローチを導入することで，解読精度の向上を目指す．

2　関連研究

近年，言語モデル，画像モデル，さらにはマルチモーダルモデルの特徴を活用して脳活動を解読する研究が急速に進展している．特に，脳信号と計算モデルとの対応関係を探るアプローチは，言語や視覚の情報処理に関する理解を深める重要な役割を果たしている．視覚刺激の解読において，生成モデルの活用が注目されている．Lin ら [2] は，fMRI 信号を視覚・言語の埋め込み空間にマッピングし，複雑な画像刺激を再構成する手法を提案し，従来の単純な特徴マッピングを超え，高品質な画像再構成を可能にした．また，Ozcelik と VanRullen [3] は，生成拡散モデルを用いた 2 段階の再構成フレームワークを開発し，自然シーンの詳細な視覚的特徴を再現

することに成功した．また，視覚刺激をもとに言語を生成する研究も進展している．Chen ら [4] は，高次視覚皮質の活動を GPT モデルと結び付け，視覚刺激が誘発する言語的意味情報の解読に成功した．Huang ら [5] は，PT-LDM モデルを開発し，fMRI 信号から短いフレーズや文章を生成することを実現した．Tang ら [1] は，エンコーディングモデルと生成モデルを分離した手法を提案した．GPT-1 で大量のトークンを生成し，これを特徴ベクトルに変換してエンコーディングモデル（リッジ回帰ベース）に入力し，脳信号の予測値を計算する．また，予測値と実際の脳信号を比較し，最も一致するトークンを選択する．これにより，生成されたテキストと脳活動の対応関係を高精度でモデル化した．さらに，マルチモーダルモデルをもとに言語を生成する研究も進展している．Liu ら [6] は，BrainCLIP を提案し，fMRI 信号を視覚および言語の埋め込み空間にマッピングすることで，高次元の意味情報を効果的に捉えた．Du ら [7] は，BraVL と呼ばれる三モーダルフレームワークを提案し，脳・視覚・言語の特徴を統合することで，ゼロショットの脳活動デコードを実現した．これらの研究は，脳活動・視覚・言語の特徴を統合し，脳信号と外部刺激の関係性をより深く解明するための多様なアプローチを示している．本研究では，これらの知見を活用し，視覚刺激に基づく文想起タスクにおける脳信号の解読を目指す．

3　データセットと前処理

3.1　脳活動データの収集

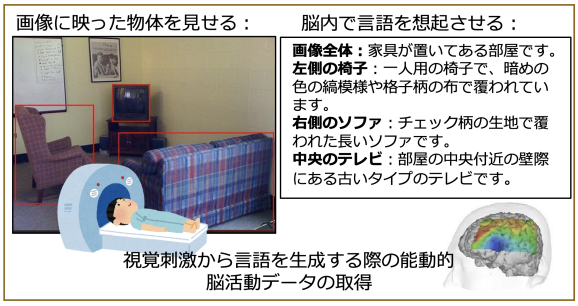


図 1　fMRI 実験全体像

本稿では 1 名の被験者を対象とし，fMRI 装置内で COCO データセット [8] から選出した自然画像を観察する際の大脳皮質データを記録した．被験者から実験前に書面で同意を得た．また，実験プロトコルは情報通信研究機構の倫理審査委員会および安全審査委員会から承認を得た．

実験内容を図 1 に示す．記録された各時間点における大脳皮質データ（ボクセル）のサイズは 93,452 であり，fMRI の撮像は 1TR=1 秒間隔で行われた．実験プロトコルは以下のように実施した：まず各試行の開始時に「Ready...」の表示を提示し，被験者に準備を促した．その後，被験者は画面中央の固視点を 20 秒程度注視した．続いて，8 秒間の画像提示を行い，被験者は画像全体の内容を黙って観察し，脳内で記述文を想起した．次に，同じ画像に対して白枠で示された 3 つの主要な物体について，それぞれ 8 秒ずつ順次観察・想起を行った．観察・想起時は声を出さず，また極力目を動かさないよう指示した．その後，4 秒間の中央固視点の注視を経て，次の試行へ移行した．これを 1 試行とし，16 枚の異なる画像で 1 ブロックを構成した．全 5 ブロックの実験を実施したが，第 2 ブロックと第 5 ブロックでは同一の画像セットを使用し，反応の再現性を確認できる設計とした．合計 560 枚の高解像度画像を提示し，大脳皮質全体をカバーする 3 次元の fMRI データを収集した．fMRI 撮像終了後，被験者にスキャナー内で観察した各画像を再度提示し，fMRI 計測中の無声観察時における被験者の認知内容を想起文テキストとして書き起こした．想起文データは，後の解析において脳活動パターンと言語表現の対応関係を検討する際の基礎データとして使用した．

3.2　画像・想起文データの特徴抽出

被験者が観察した画像と事後に記述した想起内容の言語テキストを Liu ら [9, 10] が開発した LLaVA-1.6-vicuna-7B-hf モデル[1)] に入力し，マルチモーダルな特徴表現を抽出した．このモデルは視覚データと言語データを同時に処理し，統一された特徴表現を提供する．最終的な隠れ層から特徴を抽出する際に，データの前処理として以下の手順を実施した．まず，最初の 5 つのトークンをスキップする処理を行った．これらのトークン（⟨s⟩, _US, ER, :, _）は固定のテンプレート構造を表しており，実際の入力内容には関連しないため削除した．次に，「⟨image⟩」トークンの削除を行った．このトークン（ID: 32000）は画像入力に対応するものであり，隠れ層の特徴抽出において冗長であるため全て削除した．最後に，トークン列「_ASSISTANT:」の削除を実施した．このトークン列は出力テンプレートに含

1）https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf

まれるものであり，特徴抽出に不要であるため取り除いた．

3.3 データの前処理

fMRI データと視覚特徴の時間的整合性を確保するため，以下の前処理を実施した．まず，データは全て平均 0，標準偏差 1 に標準化し，時間系列の整列において，スライディングウィンドウ法（ウィンドウサイズ = 3）を実施した．具体的には，特徴表現の各時点 T_i に対して，中心時点 c_i を計算し，区間 $[c_i - 1, c_i + 1]$ 内の fMRI 応答の平均値を対応させた．境界条件では，利用可能なデータポイントのみを使用して平均値を計算した．例えば，8 時点の fMRI データと 4 時点の特徴表現がある場合，T_0 の時点では t_0 から t_2 までの平均値，T_1 の時点では t_2 から t_4 までの平均値を使用する．

4 実験設計

エンコーディングモデルの学習に使用するために 448 枚の画像を訓練データとして使用し，さらにエンコーディングモデルおよびデコーダーの性能を評価するために 112 枚の画像を検証データとして使用した．予測された応答を実際の fMRI データと比較し，その評価には平均二乗誤差（MSE）および相関係数を用いる．

4.1 エンコーディングモデルの構築

本研究では，Naselaris ら [11] の研究を基に，画像データおよび想起内容から脳活動を予測するエンコーディングモデルを構築した．LLaVA の最終埋め込み層から得られる特徴ベクトルと大脳皮質データを用い，リッジ回帰分析を行い，両者の関係をモデル化した．また，Tang ら [1] の研究を参考に，脳活動のノイズパターンを捉えるノイズモデルを構築した．このモデルは予測残差の共分散行列を計算し，ボクセル間のノイズ相関を考慮し，観測データと予測データの尤度計算を改善する．エンコーディングモデルは主にトークンから fMRI 応答を予測し，ノイズモデルはボクセル間のノイズ相関を考慮して補完的役割を果たす．

4.2 デコーダーの構築

本研究におけるデコーダーは，モデルからテキストを生成する際に複数の生成方式を組み合わせて構築され，LLaVA モデルを用いて生成された完全な日語文を基盤とし，隠れ層の特徴を抽出して fMRI 応答を予測する一連のプロセスから成る．まず，LLaVA モデルを用い，ビームサーチやランダムサンプリングなど異なる生成方式を適用することで複数の候補文を生成する．次に，生成された文を再びモデルに入力し，隠れ層から特徴量を抽出する．この特徴量はエンコーディングモデルを通じて fMRI 応答の予測に用いられ，最後に，予測された応答を実際の fMRI データと比較し，その評価には平均二乗誤差（MSE）および相関係数を用いる．本研究で使用したデータは，被験者が画像を見た後にその内容を思考し，それを文章として表現したものである．このため，データにはいくつかの特性がある．まず，文の生成過程には正確な単語レベルのタイムスタンプが存在しない点が挙げられる．また，fMRI データは被験者の思考全体を反映した神経活動を表しており，逐次的な時系列情報に依存しない特徴がある．デコーダーは，画像や初期プロンプト文：「Describe image in one sentence in Japanese.」を基に複数の候補文を一括生成した後，それぞれの候補文から特徴量を抽出し，エンコーディングモデルを用いて fMRI 応答を予測する仕組みとなっている．さらに，予測された応答と実際の fMRI データを比較し，候補文ごとのスコアを計算する．

4.3 生成方式

付録 A の表 1 で示されているように，デコーダーには以下の生成方式を実装し，それぞれの特徴に基づきプロセスを調整した：Beam Search（複数候補から最良スコアを選択），Random Sampling（確率分布に基づきランダム選択），Temperature Sampling（温度パラメータで分布調整），Top-k Sampling（上位 k 個から選択），および Top-p Sampling（累積確率閾値 p 超の候補から選択）．

4.4 生成方式の評価

各生成方式を試行し，以下の指標を用いて最適な結果を選択する．平均二乗誤差（MSE）と相関係数はノイズモデルを用いない直接評価指標であり，MSE は実際の fMRI データと予測値の差を示し，値が小さいほど良い．また，相関係数は fMRI と予測パターンの類似度を示し，値が大きいほど良い．対数尤度はノイズモデルを考慮した評価指標であり，ボクセル間の共分散構造を考慮して脳活動パターンを包括的に評価する．

①

想起文	白い雪と雪が一目に映っていて山も映っているスキーに乗った人が7、8人写っている
最小MSE	雪に覆われた山に集まった人々がスキーを楽しんでいる。
最高相関係数	グループでスキーを楽しんでいる人たち。
最高尤度	雪に覆われた山の麓で、7人のスキーヤーが集まっている。

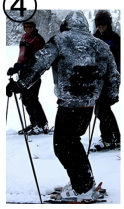
②

想起文	赤いアウターを着た男性、スキーの板に乗って笑っている
最小MSE	雪の上に、赤のスキージャケットを着た人が、スキー杖を持っている。
最高相関係数	人がスキーに乗っている写真。
最高尤度	雪の上に、赤のスキージャケットを着た人が、スキー杖を持っている。

③

想起文	青いアウターを着た女性で赤いヘルメットを被っているゴークルもつけている
最小MSE	雪の上に立つ、赤いヘルメットを着用した、スキー板を持つ人物。
最高相関係数	雪上加霜的人が、準備を終えて、スキーエリートに扮する。
最高尤度	雪の上に立つ、赤いヘルメットを着用した、スキー板を持つ人物。

④



想起文	黒いアウターを着た男性、背中に雪がついている
最小MSE	雪に覆われた3人のスキーヤーが、山の頂上でスキーを楽しんでいる。
最高相関係数	人がスキーを楽しんでいる。
最高尤度	雪の中で、3人のスキーヤーが歩いている。

図 2　生成例

5　実験結果

テストデータで構築したエンコーディングモデルの精度を評価した．付録の図 3 で示されているように，平均 MSE は 0.3445（標準偏差：0.0228）であり，標準偏差が小さいことから，モデル性能のばらつきが少ないことを示している．また，MSE 分布（最小値：0.2908，最大値：0.4320）も狭い範囲に収まり，一貫性が確認された．一方，平均相関係数が 0.0139（標準偏差：0.0300）と低く，細粒度のセマンティック関係を捉えるには限界があると考えられる．これらの結果から，本エンコーディングモデルは一定の安定性を示し，デコーダーにおける fMRI 応答予測の基盤として適していると判断される．

デコーダーの性能評価実験では，448 個の fMRI サンプルに対して，異なる生成方式を試みた．図 2 は生成結果の一例である．

日本語文の生成成功率について，付録 A の表 2 に示されているように，全手法が 90%以上の高い成功率を示したが，Sampling や Top-k などの生成設定で日本語文を出力しようとしても図 2 の 3 番で示されるように他言語が混ざり込むケースがあった．また，生成に成功したサンプルのみを対象に MSE と相関係数を算出したところ，MSE はいずれの手法も 0.463 前後とほぼ同等で，相関係数もおおむね 0.02 前後と大差は見られなかった．

さらに，各生成方式が異なる文生成結果を示すことを観察した．例えば，付録 A の表 3 に示される例では，生成された文の中に「キャンディケース」，「クリーム」，「アパートのリービング」，「寝室」，「和田」，「アメリカ」といった記述が含まれる場合があった．しかし，MSE，相関係数，平均対数尤度の最良値で選ばれた文は，被験者の想起文により近い傾向が見られた．具体的には，付録 A3 の想起文「白い 2-3 人掛けで肘掛けの付いたソファー，上に黄色いクッションが置かれている」に対し，最良生成文はいずれも「ソファー」，「黄色」，「布の描写」といった想起文に関連する語句を含んでいた．一方，「キャンディケース」や「アメリカ」といった語句は選ばれていない．これらの結果は，MSE や相関係数といった評価指標が，生成文のセマンティック適合性を捉える上で有効であることを示唆している．

6　おわりに

本稿では，画像を見て想起された文の脳活動表現を解読するプロセスを試行し，初歩的な有効性を確認した．エンコーディングモデルの精度は限定的ながら，生成文の品質から LLaVA モデルの影響が大きいと考えられる．今回は全文生成後に脳活動データと照合する方式を採用したが，今後は生成各段階で脳活動データを活用し，動的に誘導する手法の検討が必要である．

参考文献

[1] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. **Nature Neuroscience**, Vol. 26, No. 5, pp. 858–866, 2023.

[2] Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind reader: Reconstructing complex images from brain activities. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 29624–29636, 2022.

[3] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. **Scientific Reports**, Vol. 13, No. 1, p. 15666, 2023.

[4] Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. **arXiv preprint arXiv:2309.15729**, 2023.

[5] 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.

[6] Yulong Liu, Yongqiang Ma, Wei Zhou, Guibo Zhu, and Nanning Zheng. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. **arXiv preprint arXiv:2302.12971**, 2023.

[7] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 45, No. 9, pp. 10760–10777, 2023.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context, 2014.

[9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 26296–26306, 2024.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **Advances in neural information processing systems**, Vol. 36, , 2024.

[11] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. **Neuroimage**, Vol. 56, No. 2, pp. 400–410, 2011.

A　付録

表 1　生成方式とその設定

パラメータ/値	Beam Search	Random Sampling	Temperature Sampling	Top-k Sampling	Top-p Sampling
num_beams	4	-	-	-	-
do_sample	False	True	True	True	True
top_k	-	0	-	2, 3, 4	-
temperature	-	0.5, 0.7, 0.9	0.5, 0.7, 0.9	-	-
top_p	-	-	-	-	0.5, 0.7, 0.9

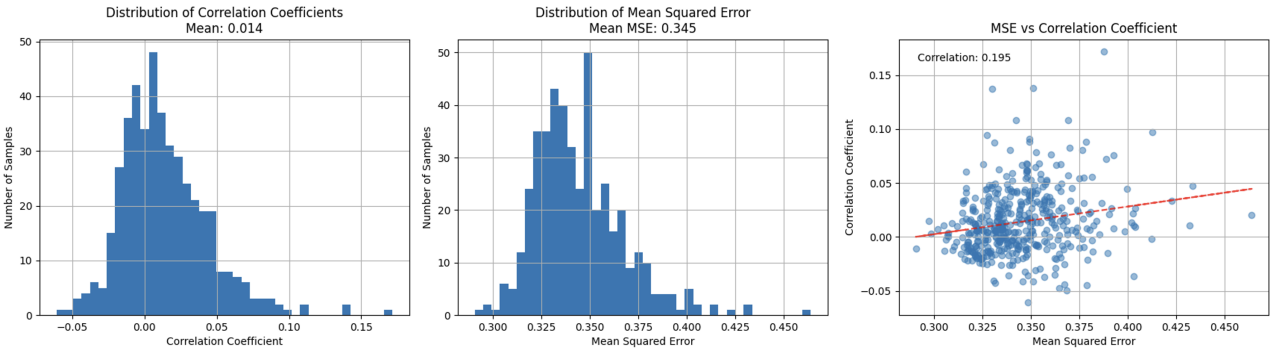


図 3　エンコーディングモデルの精度: MSE と相関係数の分布および比較

表 2　日本語生成成功率統計

手法	成功サンプル数	試行総数	成功率 (%)
Beam Search	1675	1792	93.5
Random Sampling	1322	1344	98.4
Top-k Sampling	1326	1344	98.7
Top-p Sampling	1314	1344	97.8

表 3　「想起文：白い 2-3 人掛けで肘掛けの付いたソファー，上に黄色いクッションが置かれている」に対する完全な生成例

生成方法	生成文	MSE ↓	相関係数↑	平均対数尤度
beam_search_1	白いソファに黄色のカーペットが敷かれ，その上に色彩豊かなカーペットが敷かれています．	0.4001	0.0034	-26.6980
beam_search_2	白いソファに黄色のカーペットが敷かれ，その上に色彩豊かなカーペットが乗っています．	0.4022	0.0024	-26.6604
beam_search_3	白いソファに黄色のカーペットが敷かれ，その上に色彩豊かなカーペットが敷かれている．	0.4021	0.0110	-26.5495
beam_search_4	白いソファに黄色のカーペットが敷かれています．	0.4117	0.0022	-26.6693
random_sampling_0.2	白いソファには，柔らかなカーペットが敷かれています．	0.4140	0.0019	-26.9633
random_sampling_0.5	白いソファに黄色のシーツが敷かれ，カラフルなキャンディケースが敷かれた独特なインテリアが描かれています．	0.4033	-0.0154	-26.7185
random_sampling_0.9	アパートのリビングルームで，クリーム色のソファが見える．	0.4063	-0.0060	-26.5744
top_k_2	そのソファは，白と黄色の毛布で包まれ，複数の色の毛布で裾が飾られています．	0.4051	0.0400	-26.7571
top_k_3	白いソファーの前には，床には独特のパターンがある，豊かな柔らかな雰囲気が漂っている．	0.4017	-0.0018	-26.8489
top_k_4	白いソファーで囲まれた寝室．	0.4139	0.0115	-27.5148
top_p_0.5	白いソファには，黄色のカーペットが敷かれています．	0.4096	0.0150	-26.6981
top_p_0.7	リビングルームの寝室．	0.4093	-0.0425	-27.5075
top_p_0.9	和田のアメリカ屋の部屋．	0.4301	-0.0134	-27.9701