マルチモーダル大規模言語モデルにおけるゼロと無の内部表象

江部 正周 [1] 青山 敦 [2]
[1] 慶應義塾大学 SFC 研究所 [2] 慶應義塾大学 環境情報学部
{ebe,aaoyama}@sfc.keio.ac.jp

概要

ゼロは人類史で自然数とは異なる過程で導入された数であり, 脳でも自然数とは異なる表象であることが知られている. マルチモーダル大規模言語モデル（MLLM）を対象とした研究では, 画像でのカウント課題が行われているものの, ゼロや無の内部表象は十分に明らかにされていない. 本研究では MLLM を対象として, 自然画像とそれを加工した数量情報のない画像でカウント課題を行わせ, 0–6 の内部表象を解析した. その結果, 1–6 までの内部表象は順序を保って並ぶが, 0 は異なる領域として表象された. さらに, MLLM の深い隠れ層ほど異なる画像条件の 0 の表象が有意に近くなることがわかった. これは無からゼロの概念が表象される過程と解釈できる.

1 はじめに

ゼロという数は, 人類史において自然数（1,2,3,...）とは異なる過程で導入された数の概念であり, 目に見える対象を数え上げる手続きと結びついた自然数とは対照的に, 「対象の不在」を数量として扱う抽象的操作を必要とする. 空集合は要素を持たないため, 数え上げの手続きのみからは直接評価できず, ゼロを数値的概念として理解するには, 数える経験から切り離された概念操作が求められる [1].

神経科学でも, ゼロは自然数とは異なる表象を持つことが報告されている. 人間の頭頂皮質には数量に対応する地図状の構造が存在し [2], 霊長類の頭頂葉ではゼロに選択的な神経細胞の活動が報告されている [3]. また, 霊長類の腹側頭頂間野（VIP）では視覚的な「有/無」を処理し, 前頭前皮質（PFC）では数量としての 0 を処理するという階層性が報告されている [4]. 加えて, 人工ニューラルネットワーク（ANN）においても, 明示的な数の学習なしに 0 に相当する表象が自発的に出現することが報告されている [5]. さらに, 神経を模した ANN での学習により頭頂間溝（IPS）に対応する領域で, 絶対・相対的数量表象が再編成されることも報告されている [6, 7].

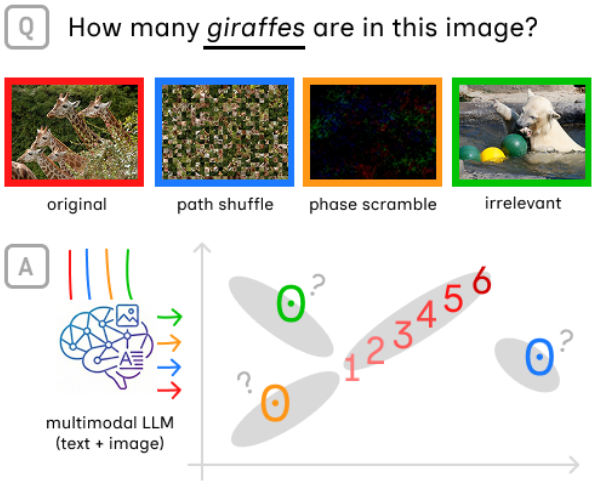


図 1 本研究の概要図

一方, MLLM の数量理解に関する研究では, 自然画像中の数量を問うカウント課題が広く用いられてきた [8, 9]. また, 数表象の内部構造に関して,MLLM が持つ数の表現空間構造や領域を調べる研究 [10], 数字トークン埋め込みのクラスタリング等により 0 が単独クラスタになりやすいことを報告する研究 [11], 数値属性方向の読み出し・介入可能性を示す研究 [12, 13, 14, 15] などが行われてきた. しかし, これらは 0 を含む数の扱いを観察するに留まり, MLLM において 0 および無に関わる内部表象がどのように構成され, 隠れ層の処理階層毎にどのように変化するかを明確に検討していない.

本研究は, MLLM の自然画像のカウント課題において, 0 と 1–6 の自然数の表象の関係と, 無から 0 の概念がどのように表象されるかを明らかにすることを目的とする. その結果, 1–6 の自然数の表象は数直線的構造を形成するものの 0 はそれらと独立した関係を示すこと, さらに, 0 や無を誘発する無情報の画像条件間では, 各 0 表象の距離は隠れ層が深くなるにつれて収束, すなわち, 無とゼロの意味が近くなることが確認された. 図 1 に本研究の概要図を示した.

2　実験設定

2.1　データセット

実験に用いる画像は Microsoft COCO データセット [16] に基づいて作成した. 画像に含まれる最頻カテゴリのインスタンス数が 1–6 である画像を各 100 件, 合計 600 件抽出した. インスタンスとは COCO の物体カテゴリ（例：person, giraffe 等）に対して付与された注釈を指す. また, 対象が小さすぎてカウントが困難になることを避けるため, バウンディングボックス面積が画像全体の 2%未満のインスタンスは抽出対象から除外した.

また, 0 を回答させる条件として, (i) original の自然画像に対し「画像に含まれないカテゴリ」を問うことで 0 を正答とする条件（original 条件 0）, および (ii) 画像情報を破壊した加工画像に対して 0 を正答とする条件（patch shuffle 条件と phase scramble 条件）を用意した. 画像入力条件のための加工画像は, 各画像に対して 16×16 画素パッチ単位で並べ替える patch shuffle 処理, およびフーリエ振幅を保持し位相のみをランダム化する phase scramble 処理を行った. これらの画像加工には NumPy[17] と Pillow[18] を用いた（画像の例については図 1 と付録 A.1 を参照）.

2.2　モデル

実験対象とする MLLM は, Qwen2.5-VL-7B-Instruct[19, 20, 21] を用いた. 推論は Hugging Face Transformers[22] で行い, 推論時の隠れ層の状態を取得した. 計算には NVIDIA L4 GPU を利用した.

2.3　プロンプトと出力制約

本研究では内部表象を安定させる目的で, 出力の自由度を低減するため, 1 トークンのみの出力を強制した. すなわち, カウント課題では次の指示を用い, 0–9 の 1 桁数字のみを出力させた：

> Answer the question using ONLY the image.
> Output EXACTLY one digit (0-9).
> Rules: Output only a single digit. No words, no punctuation, no explanation.
> If the count is 10 or more, output 9.
> Question: How many {instance} are in this image?

これらのテキストと, 条件ごとの画像を MLLM に入力した. なお, 上記プロンプトの {instance} に people や giraffes 等の名詞が入る.

3　内部表象の抽出

3.1　対象の隠れ層と入力系ごとの表象

対象 MLLM の隠れ層は 28 層であるため, 内部表象は層別に抽出した. 抽出対象の隠れ層は, 0%・25%・50%・75%・最終直前層に対応する {0,7,14,21,27} 層とした. また, テキストと画像のマルチモーダル入力における表象を切り分けるため, 各層で以下の 3 種類の表象を保存した：

- **視覚系表象**：入力系列中の画像トークン位置の表現平均.
- **言語系表象**：入力系列における最終有効トークン位置の表現. なお, これは最終出力トークンではなく, その直前の表象である.
- **全平均表象**：全入力トークン（画像＋テキスト）表現の平均.

4　分析

4.1　カウント課題の正答率

COCO データセットの注釈のインスタンス数を正解として, MLLM の出力との一致によりカウント課題の正答率を算出した. また, チャンスレベルとの比較により有意性を評価した.

4.2　表現類似性分析

表現類似性分析（representational similarity analysis, RSA）[23] は, 異なる入力に対する内部表象がどれだけ似ているか（あるいは異なるか）を非類似度行列（representational dissimilarity matrix, RDM）で表し, 理論的に予測される構造との一致度を相関係数で評価する手法である. この解析手法は, 神経科学だけではなく, LLM/MLLM の内部表象解析にも用いられている [24, 25].

1–6 の数直線性の評価　1–6 の数直線性の評価には, 1–6 のみを対象とし, RDM と理論モデル

$$\mathrm{RDM}^{\mathrm{model}}_{1\text{–}6}(i,j) = |i-j| \quad (i,j \in \{1,\ldots,6\})$$

との Spearman 順位相関 $\rho_{1\text{–}6}$ を算出して評価した.

0 の配置モデル比較　ゼロの表象の独立性を評価するため, 0 を含む 0–6 の内部表象配置について, 次の 2 モデルを比較した.

- **ゼロ数直線モデル**：0 も数直線上に配置される

理論モデル

$$\mathrm{RDM}_{\mathrm{ord}}(i,j) = |i-j|$$

- **ゼロ独立モデル**：非 0 同士は数直線上だが, 0 は非 0 から独立して配置される理論モデル

$$\mathrm{RDM}_{\mathrm{ind}}(i,j) = \begin{cases} 0 & i = j \\ |i-j| & i \neq 0 \wedge j \neq 0 \\ c & (i=0 \vee j=0) \wedge i \neq j \end{cases}$$

4.3 異なる 0 表象の収束性の分析

0 の回答を誘発する条件（original / patch shuffle / phase scramble 条件の 0）で, 0 表象が隠れ層の深さ方向に original 条件の 0 表象へ近づくかを定量化するため, original 条件の 0 表象の重心をアンカーとして patch shuffle と phase scramble の表象の重心との距離を求め, 入力種別（視覚系/言語系/全平均）ごとにトレンドを求め, 置換検定 [26, 27] を行った.

5 結果

5.1 カウント課題の正答率

表 1 にカウント課題の正答率を示す. 1–6 は数が大きくなるほど正答率が低下する傾向が見られたが, いずれもチャンスレベルと比較して有意に正答した. 0 条件では条件間で正答率に差があるものの, 0 出力が一定程度得られていることが確認された.

表 1 カウント課題の正答率

画像条件	正答率	回答の誤差	p 値
original 0	0.78	0.87	$p < .001$
original 1	0.87	0.39	$p < .001$
original 2	0.78	0.48	$p < .001$
original 3	0.37	0.86	$p < .001$
original 4	0.39	1.17	$p < .001$
original 5	0.34	1.44	$p < .001$
original 6	0.18	1.59	$p < .05$
patch shuffle 0	0.51	4.36	$p < .001$
phase scramble 0	0.98	0.55	$p < .001$

5.2 1–6 の数直線性と 0 の配置

RSA 解析により, 1–6 のみの数直線構造を評価したところ, 視覚系表象では全層で強い数直線性が得られ（ρ = .83–.97, all $p < .001$）, 全平均表象でも同様に高い相関が得られた（ρ = .78–.93, all $p < .001$）. 一方, 言語系表象では層依存が大きく, 0 層と 7 層では有意でない一方（ρ = .40, p = .141; ρ = .13, p = .647）, 中・後段層で有意となった（14 層：ρ = .55, $p < .05$; 21 層：ρ = .81, $p < .001$）. ただし, これは 0 を含まない 1–6 の表象の構造であることに注意されたい. 図 2 に言語系表象の後段層（21 層）における RDM を示す.

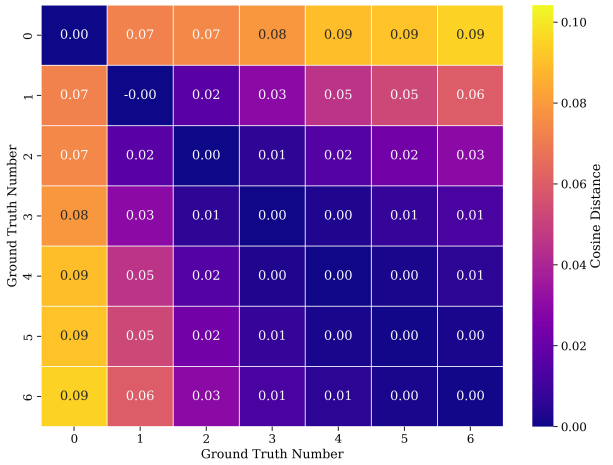


図 2 言語系表象の 21 層における RDM

また, 言語系表象の 21 層における内部表象を主成分分析によって 2 次元に圧縮して可視化した（その他の隠れ層と表現形式の主成分分析は, 付録 A.2 を参照）.

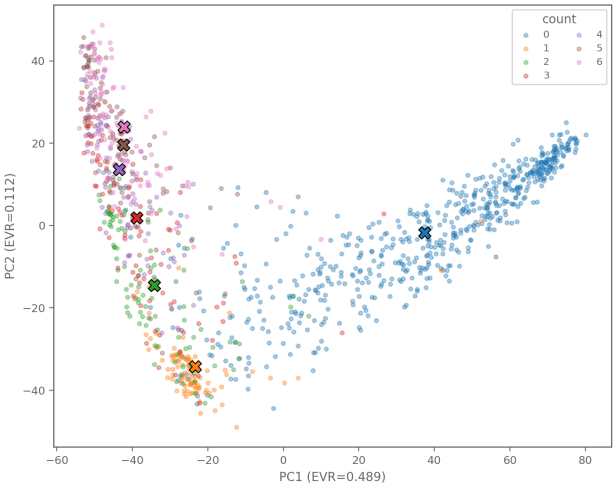


図 3 言語系表象の 21 層における主成分分析

0 を含む 0–6 の表象の配置について, ゼロ数直線モデルとゼロ独立モデルの実験データへの当てはまりを比較した. 視覚系表象および全平均表象ではゼロ数直線モデルが高い当てはまりを示す傾向はあるものの, 有意差は確認されなかった（視覚系表象：Steiger p = .14–.21; 全平均表象：Steiger p = .12–.83）. 対照的に, 言語系表象では後段層（21 層）においてゼロ独立モデルがゼロ数直線モデルを有意に上回

り, 0 が非 0 から独立した配置が支持された（21 層：$\rho_{\text{ind}} = .92$ vs. $\rho_{\text{ord}} = .71$, $Z = -2.79$, $p < .01$）. ここで重要なのは, 0 の独立が全ての表象で一様に観察されるわけではない点である. 言語系表象の後段層（21 層）ではゼロ独立モデルが有意に当てはまり, 0 が「数直線上の端点」ではなく「独立した別カテゴリ」となる結果となった.

5.3 異なる 0 表象の収束性

original 条件 0 表象の重心を基準とした距離 $d_{i,l}$ を算出し, 画像入力条件（patch shuffle / phase scramble の 0 表象）ごとに隠れ層方向の収束トレンドを評価した. その結果, いずれの条件でも傾き T は負となり, 層が深くなるほど original 条件の 0 表象との重心距離が有意に減少する傾向が見られた（patch shuffle 条件：$T = -1.19 \times 10^{-3}$, median slope $= -1.35 \times 10^{-3}$, $p < .001$; phase scramble 条件：$T = -2.29 \times 10^{-3}$, median slope $= -2.33 \times 10^{-3}$, $p < .001$）.

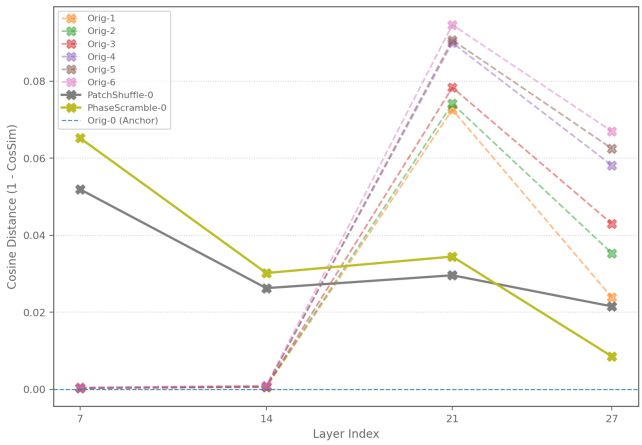


図 4　言語系表象の original 条件 0 を基準とした重心距離

6 考察

本研究の実験設定では, MLLM の自然画像のカウント課題の 1–6 の自然数の表象は, 数直線的構造を形成していた. 先行研究では, 数量の内部表象が直線的構造 [13] や桁ごとの循環構造 [15] であることが報告されており, MLLM の本課題においては数直線的構造となった. これは, 先行研究は画像ではなくテキストのみのモデルが多く, 入力モダリティによる内部表象の違いを反映していると考えられる. また, 特に言語系の表象において, 0 の表象はその自然数の数直線からは連続しているものの, 独立した表象となることがわかった. これは 1–6 との整数としての連続値であることと, 無や空集合を示す独立したカテゴリの性質とを同時に満たす内部表象の配置であると解釈できる. さらに, 0 を誘発する無情報の画像では, 0 の表象が異なる条件で隠れ層毎に収束, すなわち, 無とゼロの意味が近くなることが確かめられた. 自然画像中に数え上げを指示された対象がない 0 や無と, 画像に情報がない 0 や無では, 同じ 0 であるものの質的には異なるものであると考えられる. それらの 0 の表象が隠れ層が深くなるごとに近くに配置され, 似た 0 の意味を持つようになることは, 無からゼロの概念が表象される過程と解釈できる.

これらの結果は ANN で数の概念が層が深くなる過程で表象される例 [6] とも整合しており, 自然数と 0 の概念は表象されるタイミングが異なるものの, 層が深くなるごとに表象されるという意味では似た過程を経ると言える. また, このような階層的変換は, 霊長類で VIP が視覚的有無を処理し PFC が数量としての 0 を処理するという階層性 [4] と, 処理階層に伴う表象の抽象化という観点で対応しており, 0 や無の表象の共通する性質を示唆するものである.

7 研究の限界と今後の展望

第一に, 本研究は単一モデル（Qwen2.5-VL-7B-Instruct）のみを対象としているため, 一般の MLLM に普遍的に成り立つ結論ではない可能性がある. 今後は, 他モデルを比較し, 表象構造がどの程度一般化するかを検討する必要がある.

第二に, 0 を誘発する画像としての patch shuffle や phase scramble は視覚情報を大きく破壊するため,「0（空集合）」という概念表象だけでなく,「視覚的に判断困難」に起因する方略が混入する可能性がある. 今後は,「0（空集合）」を明確に分離する出力設計等により, 0 表象の解釈を厳密化する必要がある.

8 結論

本研究は, MLLM の自然画像のカウント課題において, 0 と 1–6 の自然数の表象の関係と, 無から 0 の概念がどのように表象されるかを実験的に検討した. その結果,（i）1–6 の自然数の表象は数直線的構造を形成し, 特に言語系の表象においては, 0 の表象は自然数の数直線からは連続しているものの, 独立した表象となることがわかった. さらに,（ii）0 を誘発する無情報の画像条件間では, 異なる画像条件でも 0 の表象が隠れ層毎に収束, すなわち, 無とゼロの意味が近くなることが確かめられた. これは無からゼロの概念が表象される過程と解釈できる.

参考文献

[1] Andreas Nieder. Representing something out of nothing: The dawning of zero. **Trends in Cognitive Sciences**, Vol. 20, No. 11, pp. 830–842, 2016.

[2] Ben M Harvey, Barrie P Klein, Natalia Petridou, and Serge O Dumoulin. Topographic representation of numerosity in the human parietal cortex. **Science**, Vol. 341, No. 6150, pp. 1123–1126, 2013.

[3] Sumito Okuyama, Toshinobu Kuki, and Hajime Mushiake. Representation of the numerosity 'zero' in the parietal cortex of the monkey. **Scientific Reports**, Vol. 5, No. 1, p. 10059, 2015.

[4] Araceli Ramirez-Cardenas, Maria Moskaleva, and Andreas Nieder. Neuronal representation of numerosity zero in the primate parieto-frontal number network. **Current Biology**, Vol. 26, No. 10, pp. 1285–1294, 2016.

[5] Khaled Nasr and Andreas Nieder. Spontaneous representation of numerosity zero in a deep neural network for visual object recognition. **Iscience**, Vol. 24, No. 11, 2021.

[6] Percy K Mistry, Anthony Strock, Ruizhe Liu, Griffin Young, and Vinod Menon. Learning-induced reorganization of number neurons and emergence of numerical representations in a biologically inspired neural network. **Nature Communications**, Vol. 14, No. 1, p. 3843, 2023.

[7] Bhavesh K Verma and Rakesh Sengupta. Emergence of behavioral phenomena and adaptation effects in human numerosity decoder using recurrent neural networks. **Scientific Reports**, Vol. 13, No. 1, p. 19571, 2023.

[8] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 33, pp. 8076–8084, 2019.

[9] Muhammad Fetrat Qharabagh, Mohammadreza Ghofrani, and Kimon Fountoulakis. Lvlm-count: Enhancing the counting ability of large vision-language models. **arXiv preprint arXiv:2412.00686**, 2024.

[10] Ivana Kajic and Aida Nematzadeh. Probing representations of numbers in vision and language models. In **SVRHM 2022 Workshop@ NeurIPS**, 2022.

[11] Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Understanding counting mechanisms in large language and vision-language models. **arXiv preprint arXiv:2511.17699**, 2025.

[12] Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. **arXiv preprint arXiv:2403.10381**, 2024.

[13] Ulme Wennberg and Gustav Eje Henter. Exploring internal numeracy in language models: A case study on albert. **arXiv preprint arXiv:2404.16574**, 2024.

[14] Fangwei Zhu, Damai Dai, and Zhifang Sui. Language models encode the value of numbers linearly. In **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 693–709, 2025.

[15] Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 385–395, 2025.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In **European conference on computer vision**, pp. 740–755. Springer, 2014.

[17] Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, and ... Array programming with NumPy. **Nature**, Vol. 585, No. 7825, pp. 357–362, September 2020.

[18] Alex Clark. Pillow (pil fork) documentation, 2015.

[19] Qwen Team. Qwen2.5-vl, January 2025.

[20] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. **arXiv preprint arXiv:2409.12191**, 2024.

[21] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. **arXiv preprint arXiv:2308.12966**, 2023.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2019.

[23] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. **Frontiers in systems neuroscience**, Vol. 2, p. 249, 2008.

[24] Max Klabunde, Mehdi Ben Amor, Michael Granitzer, and Florian Lemmerich. Towards measuring representational similarity of large language models. **arXiv preprint arXiv:2312.02730**, 2023.

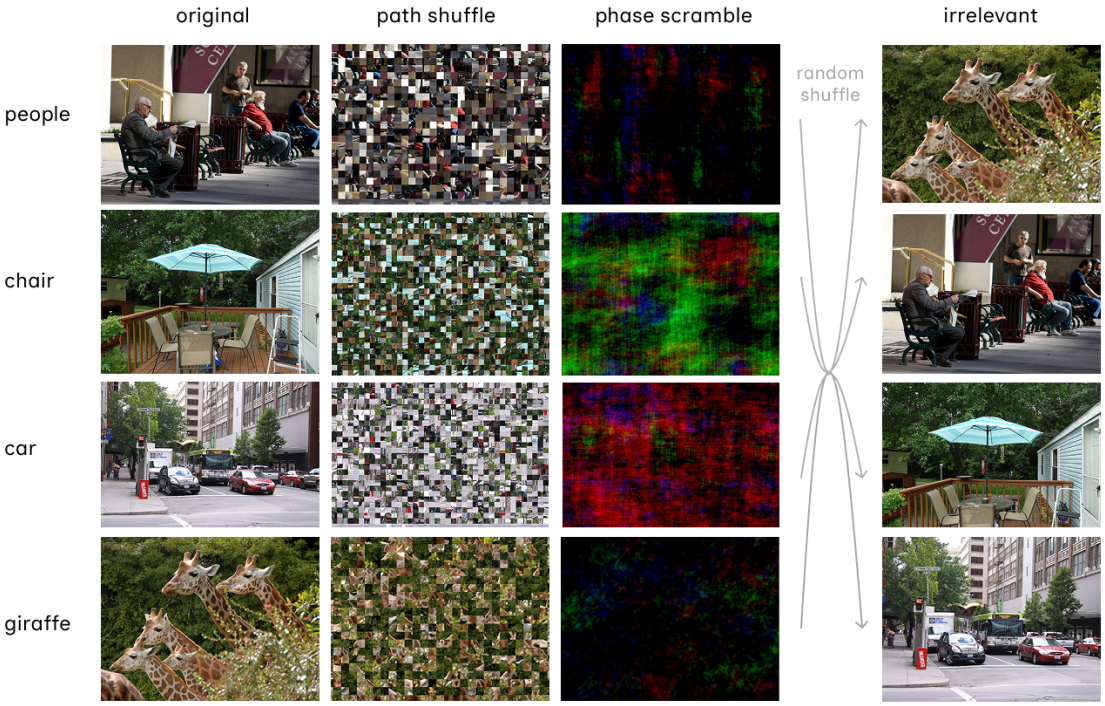
[25] Ningyu Xu, Qi Zhang, Chao Du, Qiang Luo, Xipeng Qiu, Xuanjing Huang, and Menghan Zhang. Revealing emergent human-like conceptual representations from language prediction. **Proceedings of the National Academy of Sciences**, Vol. 122, No. 44, p. e2512514122, 2025.

[26] Andrew P Holmes, RC Blair, JDG Watson, and I Ford. Nonparametric analysis of statistic images from functional mapping experiments. **Journal of Cerebral Blood Flow & Metabolism**, Vol. 16, No. 1, pp. 7–22, 1996.

[27] Thomas E Nichols and Andrew P Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. **Human brain mapping**, Vol. 15, No. 1, pp. 1–25, 2002.

A　付録

A.1　COCO2017 から取得した画像入力と加工のサンプル



A.2　画像入力条件ごとの隠れ層の内部表象の主成分分析