Text-to-Image モデルにおける生物画像生成の地理的バイアス

吉岡智輝 [1,2] 平尾努 [1,2] 高村大也 [2]
[1] 金沢大学 [2] 産業技術総合研究所
{y2252100242@stu,hirao@se}.kanazawa-u.ac.jp, takamura.hiroya@aist.go.jp

概要

Text-to-Image (T2I) モデルは、テキストプロンプトに基づき高品質な画像を生成できるものの、多様性のあるオブジェクトに対し特定のものを生成しがちであるというバイアスがある。本稿では、哺乳類を対象として、T2I モデルによる生成画像の分布と Global Biodiversity Information Facility (GBIF) の観測分布とを比較することでそのバイアスを明らかにする。具体的には哺乳類の一般名を与えて生成した画像の種の偏りとその実世界の各地域での観測データの偏りとを比較する。この結果、T2I モデルが生成する画像は英語圏に偏っていることがわかった。また、一般名では生成できない哺乳類は、学名や形態的特徴を与えることで生成可能になることもわかった。

1 はじめに

Text-to-Image (T2I) モデルはテキストプロンプトから画像を生成するモデルであり、近年のモデルは高品質な画像を生成することができる。オープンウェイトモデルでは Stable Diffusion [1, 2]、Qwen-Image [3] など、商用モデルでは Gemini 2.5 Flash Image (Nano-Banana)[1] や ChatGPT (DALL-E 3)[2] などがよく知られている。しかしながら、こうしたモデルによる生成画像が常にユーザが意図する文化的・地理的背景と一致するとは限らない。例えば、T2I モデルを科学教育に利用することを考えよう。ラクダという生物を説明するため、T2I モデルにその画像を生成させたところ、コブが 1 つのラクダ、ヒトコブラクダの画像を提示したとする。すると、人々はラクダという生き物はコブが 1 つであると思い込んでしまうのではないだろうか。しかし、フタコブラクダというコブが 2 つのラクダも実在するので、多様性を無視した画像生成は自然を正しく理解するという観点からは大きなリスクとなる。

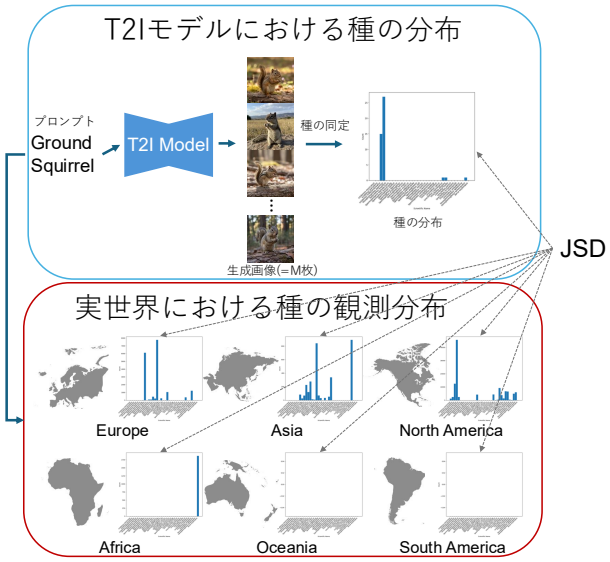

図 1 T2I モデルの種の分布と実世界の種の分布の比較

こうしたバイアスを是正するためには、まずそのバイアスを正確かつ客観的に把握する必要がある。そこで本稿では、客観性に優れる「生物 (哺乳類)」を指標として、モデルのバイアスを定量化する。具体的には、哺乳類の一般名 (Common Name)[3] を入力した際に生成される画像が、分類学的にどの種に該当するかを同定する。そして、モデルによる種の生成頻度分布と、Global Biodiversity Information Facility (GBIF)[4] における現実世界における種の観測頻度分布を比較することで、T2I モデルが内部に持つ一般名に対するデフォルト種と世界の各地域における一般名に対するデフォルト種の違いを定量的に評価する (図 1)。

実験では、生物の遺伝子配列に基づき作成した系統樹である NCBI Taxonomy [4] から哺乳類 5,853

1) https://aistudio.google.com/models/gemini-2-5-flash-image
2) https://cdn.openai.com/papers/dall-e-3.pdf
3) 本稿では複数の種を包括する、分類学上の「属」や「科」などに相当するグループに対する呼称とする。例えば、シカ、ネズミ、クマなどが一般名である。
4) http://www.gbif.org

種の学名[5]を取得し、大規模言語モデル (LLM) を用いて、それらに対する一般名を得た。そして、得られた一般名を T2I モデルに入力して複数画像を生成した後、GPT-4o を用いてそれらの種を同定することで生成頻度分布を作成した。この分布と GBIF の地域ごとの生物の観測分布との距離を JS ダイバージェンスを用いて評価した。Stable Diffusion、Qwen-Image、Nano-Banana を対象に分析を行った結果、以下の知見が得られた。(1) どのモデルも一般名に対して生成される種に顕著な偏りが見られた。(2) その生成分布は、ヨーロッパ・北米・オセアニア地域における観測情報の偏りと似ていた。さらに、一般名のみでは生成が稀な種であっても、プロンプトに学名や形態的特徴を追記することで生成可能であることを確認した。

2 関連研究

T2I モデルのバイアスに関しては、社会、文化という視点での検証が進んでいる。Basu ら [5] は、国名を指定しない場合、生成される画像が米国や欧州のスタイルに偏ることを示した。また、Luccioni ら [6] は人物生成におけるジェンダーと民族性を調査した結果、特定の職業 (CEO や Director など) において、実社会以上に男性・白人に偏っていることを示した。Jha ら [7] は、135 の国籍における視覚的ステレオタイプを分析し、モデルが中立的なプロンプトに対してもステレオタイプな描写に引き寄せられる現象を報告している。Ghosh は [8]、カースト制度という複雑な社会的階層を持つインドにおいて、単に「インド人」というプロンプトをモデルに与えると、上位カーストの特徴をデフォルトとして出力することを明らかにした。さらに、モデルの文化的知識を測るためのデータセット、CCUB [9] や CUBE [10] も整備されている。これらは特定の国やカテゴリ (食事、衣服など) を選定し、生成画像の妥当性を評価するものである。

しかしながら、T2I モデルのバイアスは社会、文化的な概念だけに限らない。Kupferschmidt ら [11] は、河川地形学の観点から Stable Diffusion を評価し、「川」という単純なプロンプトが、学習データに多くを占める北米・欧州の景観が良い山間部の川に偏って生成されることを実証した。

こうしたバイアスが生じる要因の一つとして、学習データセット (LAION-5B [12]) の偏りが指摘されている。LAION は web から収集された画像とそれを説明するテキストの対から成る。Seshadri ら [13] は、プロンプトに曖昧性がある場合、モデルは学習データの分布を忠実に再現する傾向があり、これがバイアスとして観測されることを示した。

これらの先行研究は主に人間社会や物体を対象としており、その解釈に曖昧性を残すものを対象としている。さらに、生成された画像の評価、つまりどこの文化圏に属するかという判断は定性的になりがちであり、客観的な定量化が困難であるという課題がある。

一方で、バイアス是正のため、学習データに含まれない、あるいは頻度の低い文化的概念を生成するためのプロンプトエンジニアリング手法も提案されている。Jeong ら [14] は、Wikipedia 等の外部知識を用いてプロンプトを反復的に洗練する手法を提案し、マイナーな文化語彙に対する画像生成の精度を向上させた。しかしながら、彼らが対象とした文化的概念は、文脈や解釈に依存するため定義が一意に定まるとは限らないという本質的な曖昧さを持つ。

3 アプローチ

本稿では、哺乳類画像の生成を対象として T2I モデルのバイアスを調べる。具体的には、いくつかの種をまとめた概念である一般名を T2I モデルに与え、どの種が何回生成されたかという頻度分布でバイアスを定量化し、それを実際の生物の地域ごとの分布と比較する。生物の「種」は学術界のコンセンサスのもとに定義されたものであり、これまでの研究が扱ってきた文化・社会的背景を持った物体よりも概念としての客観性が高い。さらに生物の分布は、地球環境と進化の過程で決まったものでありこの点においても客観性が高い。

3.1 種と一般名

種として認められた生物には学名が与えられる。つまり、学名を用いればある種を曖昧性なく同定できる。そして、学名に対して英名や和名が与えられる[6]。これとは別に、我々はいくつかの種をまとめた概念である一般名も用いる。例えば、日本にはドブネズミ、ハツカネズミ、カヤネズミなど多数のネ

5) 学名とは、ラテン語で表記される世界共通の生物名であり、属名と種小名から成る。例えば、ハツカネズミの学名は *Mus musculus* であり、*Mus* が属名、*musculus* が種小名である。生物に学名が与えられるということはそれを種として記載する記載論文とタイプ標本が存在することを意味する。

6) 図鑑、動物園や水族館などで用いられる呼称。

ズミの種が生息しているが、多くの人はこれらを細かく区別することなくネズミと呼ぶだろう。こうした一般名は学術的に定義されたものでなく人々の間で歴史を経て定着した名称である。

本稿では、学名を NCBI Taxonomy [4] から得る。これは、NCBI が提供する、生物種とその系統分類を階層構造 (ドメイン、界、門、綱、目、科、属、種) で表したデータベースであり、NCBI の塩基配列データベースに登録されている生物種が登録されている。綱が *Mammalia* (哺乳類) となるすべての種を取得し、未同定種 (*sp.*)、不確定な学名 (*cf.,aff.*) を除く。この結果、哺乳類に対し、5,853 の学名が得られた。次に、得られた学名を LLM に入力し一般名を得る。この際、LLM には 5-shot の例示を与える。なお、適当な一般名がない場合には LLM が「無い」と答えることを許している。プロンプトを付録 A に示す。ある一般名に対して 1 つの種しか割り当てられていない事例、LLM が一般名がないと答えた種を除いた結果、5,159 種に対して 168 の一般名を得た。一般名に割り当てられる種の数の最小、最大、平均、中央値はそれぞれ 2、1,479、30.7、5 であった。種が最も多い一般名は Mouse であった。

3.2 T2I モデルにおける種の分布

T2I モデルに前節で得た一般名を入力し、画像を生成する。これを M 回繰り返すことで 1 つの一般名に対して M 枚の画像を得る。そして、各画像に対して GPT-4o を用いてその学名を同定する。具体的には、GPT-4o に対して画像と共に生成に用いた一般名に対して割り当てられた学名を候補として与え、画像がどの学名に対応するかを選択させる[7)]。なお、該当する学名がない場合には該当なしを選択するように指示した。この処理により、種の頻度分布が得られるので、これを T2I モデルが内部に持っている一般名に対する種の分布とみなす。

3.3 実世界における種の分布

生物の分布は多岐にわたり、ある地域にしか生息しないものや世界各国に生息するものもある。GBIF の Occurrence データベースはある種がいつどこで観測されたかという情報を集めたデータベースである。なお、野生個体だけではなく博物館の標本や動物園で飼育されている個体の観測情報も含む。このデータベースから、北米 (NA)、南米 (SA)、ヨーロッパ (EU)、アジア (AS)、アフリカ (AF)、オセアニア (OC) の各地域において、ある種が何回観測されたかというデータが得られるので、任意の一般名に対して、地域ごとの種の観測頻度分布が得られる。本稿では、これを実世界における種の分布とみなす。

3.4 バイアスの定量化と分布間の距離

3.2 節の手法で得た T2I モデルの種の生成頻度分布に対し、以下の式でバイアスを定量化する：

$$NH(c_k) = (\log |S(c_k)|)^{-1} \sum_{i=1}^{|S(c_k)|} -p_i \log p_i. \quad (1)$$

$NH(c_k)$ は、k 番目の一般名 c_k に対するエントロピーを正規化したものであり、$S(c_k)$ は c_k に対応する種 (=学名) の集合、p_i は i 番目の種の生成確率である。c_k に対して T2I モデルに生成させた総画像数を M、i 番目の種の画像数を g_i としたとき、$p_i = g_i/M$ となる。$NH(c_k)$ が低いと特定の種に偏った生成、高いと多様な生成であることを意味する。

3.3 節の手法により、c_k に対する実世界の各地域での j 番目の種の観測頻度 f_j が得られるので、その観測確率は、$q_j = f_j / \sum_{j=1}^{|S(c_k)|} f_j$ となる。よって、式 (1) の p_i を q_j に置き換えることで同様にエントロピーを計算できる。T2I モデルのエントロピーを $H_{\mathcal{G}}(c_k)$、GBIF から得たエントロピーを $H_{\mathcal{O}}(c_k)$ として、2 つの分布間の距離を以下の JS ダイバージェンスで計算する：

$$D_{\mathrm{JS}}(c_k) = H_{\mathcal{M}}(c_k) - (1/2)H_{\mathcal{G}}(c_k) - (1/2)H_{\mathcal{O}}(c_k). \quad (2)$$

なお、$H_{\mathcal{M}}(c_k)$ は T2I モデルと GBIF による観測分布の平均分布である。これにより、T2I モデルが持つ種の生成頻度分布が実世界におけるどの地域における観測分布に近いかがわかる。

4 結果と議論

4.1 T2I モデルのバイアス

T2I モデルとして Stable Diffusion v1 (2022 年)、XL (2023 年)、Qwen-Image (2025 年)、Nano-Banana (2025 年) を用いて式 (1) で $NH(c_k)$ を計算し、一般名全体で平均を取った結果を表 1 に示す。なお、1 つの一般名に対する生成画像数 (M) は、Stable Diffusion v1、Stable Diffusion XL、および Qwen-Image では 100 枚、

7) 本来、このタスクに対しては BioCLIP [15] や BioCLIP2 [16] を用いて学名を同定すべきであるが、いくつかの事例についてこれらで学名を決定したところ明らかな誤りが多数あったため使用を見合わせた。

表 1　T2I モデルによる種の生成の偏り

Model	SDv1	SDXL	Qwen-Image	Nano-Banana
NH	.308	.241	.180	.207
# of Sp.	567	480	369	368

表 2　T2I モデルと GBIF の間の JS ダイバージェンス

	NA	SA	EU	AS	AF	OC
SDv1	.561	.630	.527	.648	.637	.563
SDXL	.568	.641	.530	.687	.629	.567
Qwen-Image	.553	.652	.494	.683	.620	.547
Nano-Banana	.552	.642	.534	.691	.644	.571

Nano-Banana では 50 枚とした。表より、モデルの発表時期が古いほど生成する種の多様性に富み、新しくなればそれに乏しくなる。Stable Diffusion v1 と Qwen-Image、Nano-Banana では生成される種の数が 200 も違う。

NH が低い代表的な一般名は Bear (クマ)、Camel (ラクダ)、Elephant (ゾウ) などであった。「Bear」については、ハイイログマ (*Ursus arctos horribilis*) の画像しか生成されず、ツキノワグマやアメリカクロクマは生成されなかった。「Camel」はヒトコブラクダ (*Camelus dromedarius*) の画像しか生成されず、フタコブラクダの画像は生成されなかった。「Elephant」に関してもアフリカゾウ (*Loxodonta africana*) の画像ばかりでアジアゾウは生成されなかった。

一方、*NH* が高い一般名は、小型のサルである Tamarin (タマリン) や地上で生活するリスである Ground Squirrel (ジリス) などであり、一般名配下の様々な種が生成された。

4.2　GBIF の分布との比較

前節の結果の通り、T2I モデルは一般名に対して、偏った種の画像を生成することが多い。この偏りが実世界のどの地域における種の分布に近いかを調べるため、T2I モデルの種の生成頻度分布と GBIF による観測頻度分布の距離を式 (2) の JS ダイバージェンスで計算した。その結果を表 2 に示す。

表より、どの T2I モデルもヨーロッパとの距離が近く、次いで北米、オセアニアが近いことがわかる。これらと比較すると、南米、アフリカ、アジアは距離が遠く、その中でもアジアは最も遠い。Stable Diffusion は XL のほうが v1 よりもその傾向が強く、全体では Qwen-Image、Nano-Banana はそれがより顕著である。こうしたバイアスは学習データに起因していると考える。T2I モデルの学習データの多くが

一般名(Camel)　学名(*Camelus bactrianus*)　一般名(Camel)+形態情報

図 2　一般名、学名、一般名+形態的特徴の生成例。左の画像は Qwen-Image に Camel を与えた結果、中央画像はフタコブラクダの学名 *Camelus bactrianus* を与えた結果、右の画像は Camel と Wikipedia から得たフタコブラクダの形態的特徴を与えた結果。

Web 上のデータであることを考えると、生成される哺乳類の画像がヨーロッパ、北米、オセアニアの分布に偏るのは自然だろう。

一方、世界的にではなく地域に限定的に生息する種の一般名に関しては、その地域に対する JS ダイバージェンスが低かった。Sugar Glider (フクロモモンガ) はアジア、Groundhog (ウッドチャック) は北米、Degu (デグー) は南米、Hippopotamus (カバ) はアフリカ、Dunnart (ダンナート) はオセアニアの代表例である。

表 1、2 より、T2I モデルのバイアスは明らかであるが、それらは、生成されることがなかった種について情報を持っていない、つまり、知らないのだろうか。これを検証するため、T2I モデルに学名や種の形態的特徴を与えたところ、多くの種を生成することができた (図 2 は Camel の例)。つまり、T2I モデルは様々な種に対して、テキストと視覚情報を結びつけることができているが、一般名という抽象度の高いプロンプトでは学習データのバイアスの影響を強く受けるのではないかと考える。

5　おわりに

本稿では、T2I モデルのバイアスを、哺乳類の一般名と対応する種がそれぞれ何回生成されるかという観点から定量化した。その結果、多くの場合、T2I モデルは特定の種に偏った生成をすることが明らかとなった。次に、T2I モデルの種の生成分布と世界の各地域における種の観測分布との距離を JS ダイバージェンスで測った結果、ヨーロッパ、北米、オセアニアの観測分布に近いことが明らかとなった。また、プロンプトに学名や種の形態情報を与えることで一般名では生成されることがなかった種でも画像を生成できた。

謝辞

この成果は、NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP25006）の結果得られたものです。

参考文献

[1] 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 (CVPR)**, pp. 10684–10695, June 2022.

[2] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In **The Twelfth International Conference on Learning Representations (ICLR)**, 2024.

[3] Chenfei Wu, et al. Qwen-image technical report, 2025.

[4] Conrad L Schoch, et al. Ncbi taxonomy: a comprehensive update on curation, resources and to ols. **Database**, Vol. 2020, p. baaa062, 08 2020.

[5] Aparna Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. **2023 IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 5113–5124, 2023.

[6] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 56338–56351, 2023.

[7] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 12333–12347, 2024.

[8] Sourojit Ghosh. Interpretations, representations, and stereotypes of caste within text-to-image generators. **Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society**, Vol. 7, No. 1, pp. 490–502, 2024.

[9] Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. Scoft: Self-contrastive fine-tuning for equitable image generation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 10822–10832, June 2024.

[10] Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, **Advances in Neural Information Processing Systems**, Vol. 37, pp. 13716–13747. Curran Associates, Inc., 2024.

[11] C. Kupferschmidt, A.D. Binns, K.L. Kupferschmidt, and G.W. Taylor. Stable rivers: A case study in the application of text-to-image generative models for earth sciences. **Earth Surface Processes and Landforms**, Vol. 49, No. 13, pp. 4213–4232, 2024.

[12] Christoph Schuhmann, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 25278–25294. Curran Associates, Inc., 2022.

[13] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 6367–6384, 2024.

[14] Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 9543–9573, 2025.

[15] Samuel Stevens, et al. Bioclip: A vision foundation model for the tree of life. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 19412–19424, June 2024.

[16] Jianyang Gu, , et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. In **Advances in Neural Information Processing Systems**, Vol. 38, 2025.

A　学名から一般名への変換

学名から一般名を得るために用いたプロンプトは以下の通りである。

Convert the given scientific name into the most common, everyday English name that people would naturally think of when seeing the animal — not a formal taxonomic category.

Rules

1. If a familiar, everyday name exists for the specific species (e.g., Cat, Dog, Tiger), use it.
2. If no such name exists, use the everyday word people would most likely think of upon seeing that animal, even if it represents a broader or informal group (e.g., Mouse, Monkey, Bat, Whale).
3. Avoid obscure or technical names, and do not use phonetic spellings of the scientific name.
4. If there is no suitable everyday name at any level, answer “None”.

Examples

Scientific Name: Felis catus
Common Name: Cat
Scientific Name: Panthera tigris
Common Name: Tiger
Scientific Name: Abeomelomys sevia
Common Name: Mouse
Scientific Name: Pteropus conspicillatus
Common Name: Bat
Scientific Name: Alces alces
Common Name: Moose
Scientific Name: [Extremely obscure species]
Common Name: None
Scientific Name: [species]
Common Name: