jh250029

マルチモーダル基盤モデルのための選択的忘却に関する研究

入江　豪（東京理科大学）

概要　本研究では、マルチモーダル基盤モデルにおける不要な知識の忘却を目的として、選択的忘却技術の創出および評価基盤の整備に取り組んだ。深いブラックボックスモデルに対しては、モデル内部に一切アクセスできない条件下でも適用可能な離散プロンプト最適化に基づく忘却法を提案し、その有効性を実証した。合わせて忘却対象知識の多様化を目的として、ドメイン単位での忘却や破滅的忘却を抑制するバイアス校正手法を開発し、知識制御の精緻化を実現した。さらに、拡散画像生成モデルに対しては、プロンプト操作のみによる軽量な概念抑制手法と、忘却済み概念を復元する攻撃法を提案し、安全性評価の重要性を明らかにした。加えて、マルチモーダル大規模言語モデルに対する忘却性能を評価するベンチマークを構築し、既存手法の限界を体系的に分析した。以上により、選択的忘却の適用範囲を拡張するとともに、その有効性と課題を明確化した。

1　共同研究に関する情報

(1) 共同利用・共同研究を実施している拠点名

東京大学 情報基盤センター

mdx I

(2) 課題分野

データ科学・データ利活用課題分野

(3) 参加研究者一覧と役割分担

入江　豪（東京理科大学）　選択的忘却技術の設計・開発、JHPCN 計算基盤を用いた評価実験を主導。研究代表者として、課題全体の統括を担当。

相澤　清晴（東京大学）　マルチモーダル大規模言語モデル（MLLM）に対するベンチマーク設計を主導。東大情報基盤センター連携のもと、JHPCN 計算基盤の活用・大規模実験環境の整備を担当。

2　研究の目的と意義

CLIP や ChatGPT、Stable Diffusion などに代表される**マルチモーダル基盤モデル**は、言語、画像、音声などモーダルの異なる複数の情報を統合的に学習することにより、従来の画像認識や音声認識といった特化型 AI モデルとは一線を画す広範な問題解決能力を獲得した汎用 AI モデルである。その応用範囲は医療、教育、製造など多岐にわたり、現代 AI の中核的技術として、ビッグテックをはじめ多くの研究機関で活発に研究開発・社会実装が進められている。

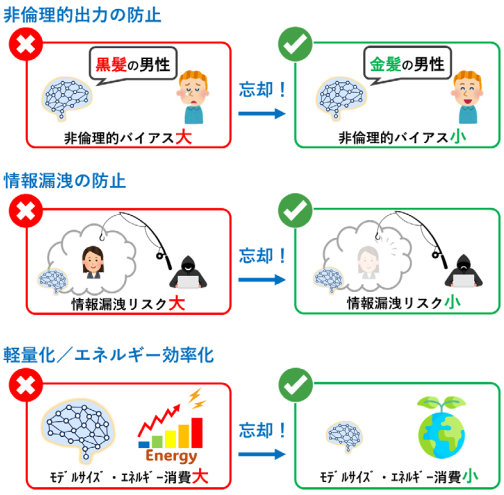

図 1. マルチモーダル基盤モデルのリスクと選択的忘却によるアプローチ

一方、マルチモーダル基盤モデルは膨大なデータを用いた学習を経て作られている。この結果、前述の高い汎用性とタスク解決能力を獲得している反面、次のような重大な社会的リスクを顕在化させて

おり、その実用化・社会展開を阻む深刻な要因となっている（図 1）。

- **非倫理的バイアス**：学習データに潜む望ましくないバイアスにより、モデルが倫理的に不適切な結果を出力してしまう問題が指摘されている。例えば CLIP は「金髪」の男性の髪色を「黒髪」と誤認しやすいことが知られている。このようなバイアスは利用者への不利益や不適切な倫理観を助長しうる。
- **情報漏洩**：学習データをモデル自体から復元するモデル反転攻撃と呼ばれる攻撃法があり、セキュリティリスクとなっている。多様かつ膨大なデータで学習されたマルチモーダル基盤モデルは、従来型モデル以上に甚大な情報漏洩リスクを抱える。
- **運用コスト**：マルチモーダル基盤モデルは多種多様な物体を認識できるが、あらゆる物体の認識が必要な実応用はそう多くない（例：自動運転では「車」や「歩行者」の認識が重要であり、「料理」は認識不要であろう）。不要な知識はモデルの過剰な肥大化を招き、計算資源・エネルギー浪費の要因となる。

本研究では、こうした課題の根本原因を**「モデルが不要な知識を過剰に保持していること」**に見出し、**モデルが持つ知識のうち、不要なもののみを消去可能にする『選択的忘却技術』を創出**する。特に、深いブラックボックスモデル（テーマ①）、マルチモーダル画像生成モデル（テーマ②）、マルチモーダル大規模言語モデル（テーマ③）に対する選択的忘却法の実現に向けた研究を推進する。

本研究を通じて AI 技術の社会的信頼性と効率性を向上させ、安全で持続可能な AI 運用基盤の構築に貢献する。具体的には、以下の側面からの貢献が挙げられる。

- **「忘れられる権利」への対応**：AI サービスが特定情報の削除を求められた場合、該当データを削除しモデルを再トレーニングする必要がある。しかし、大規模モデルの再トレーニングは膨大なエネルギーを要し、環境負荷も大きい。選択的忘却は、モデルの再学習を不要とする効率的な解決策を提供できる。
- **プライバシー保護機能の強化**：不要な情報をモデルから選択的に削除することで、個人情報漏洩リスクを低減する。AI 技術のセキュリティリスクを低減し、受容性を高める効果が期待される。
- **運用効率の向上**：モデルに必要最小限の情報のみを保持させることでモデルの軽量化を可能とし、計算コスト削減・エネルギー効率向上を実現する。これにより、AI システムの持続可能性が向上する。
- **柔軟な適応性の実現**：新たなタスクや環境に迅速かつ柔軟に適応可能なモデルの構築が可能となる。これにより、AI の適用範囲が拡大し、より多様な産業分野における適用性が高まることが期待できる。

3　当拠点の公募型共同研究として実施した意義

本研究を当拠点の公募型共同研究として実施した意義として、学際性、および、計算資源の観点が挙げられる。

学際性　マルチモーダル基盤モデルは **AI 分野、及び、その応用領域全体に波及しうる学際課題**である。本課題の遂行には、入江が取り組む選択的忘却技術に関する知見に加え、相澤が持つマルチモーダル生成モデルの学習に関する見識の双方が不可欠となる。さらに本研究はマルチモーダル大規模言語モデルの選択的忘却のための評価ベンチマーク設計をその計画に入れており、相澤のマルチモーダル大規模言語モデルの評価ベンチマーク設計に関する知見も必須となる。

計算資源　課題の特性上、大規模マルチモーダル言語モデルを研究対象とする。これらは数十〜数百億規模のパラメータ数を持ち、その学習には当拠点の計算資源が不可欠となる。

上記観点から、当研究を効率的・効果的に研究を推進することができる。

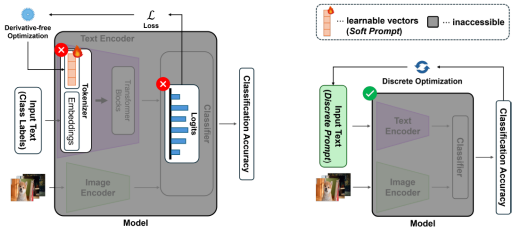

図 2. 従来の緩いブラックボックス忘却法（左図）と提案する厳密なブラックボックス忘却法（右図）。従来法はトークナイザにアクセスできることを仮定していたのに対して、本手法はモデルの入出力以外いかなる情報も観測できない場合を仮定する。

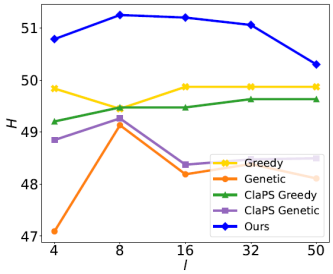

図 3. 離散プロンプト最適化によるブラックボックス忘却法の比較実験結果。提案法はプロンプトの長さ（横軸）によらず一貫して良好な忘却性能（縦軸）を示している。

4 前年度までに得られた研究成果の概要

該当なし

5 今年度の研究成果の詳細

本研究では**「①深いブラックボックスモデルに対する選択的忘却の実現」**、**「②マルチモーダル画像生成モデルに対する選択的忘却の実現」**、**「③マルチモーダル大規模言語モデルにおける選択的忘却ベンチマーキングの整備」**の 3 つを主要課題として設定し、これらを並列に遂行してきた。

5.1 深いブラックボックスモデルに対する選択的忘却の実現

従来の選択的忘却手法の多くは、忘却対象の概念を認識できなくなるように忘却しつつ、それ以外の概念の認識精度を維持するようにモデルのパラメータをチューニングすることによって忘却を実現する。一方で、事前学習済みの視覚言語モデルは、倫理的・商業的理由などからモデルの内部情報が非公開、すなわちブラックボックスモデルであるものも少なくない。このようなモデルは、パラメータや勾配などに一切アクセスできない。このような条件下では、既存の忘却手法の多くが前提としている内部情報へのアクセスが不可能であるため、そのまま適用することができないという問題がある。そこで本研究ではこのようなブラックボックスモデルに対しても適用可能なブラックボックス忘却法の創出に取り組んだ。さらに、現状の選択的忘却技術により忘却可能な知識の種類は極めて限定されているという課題を鑑み、忘却対象知識の拡大にも取り組んだ。本節ではこれら一連の成果について述べる。

A) 離散プロンプト最適化によるブラックボックス忘却

ブラックボックスモデルに対する選択的忘却法はこれまでほとんど検討されてこなかったが、ごく最近モデルパラメータやその勾配に一切触れることなく忘却を実現するブラックボックス忘却法が提案された。しかしながら、この方法は依然としてトークン埋め込み（トークナイザ）へのアクセスを前提とする“緩い”ブラックボックス条件を仮定している（図 2 左）。現実のブラックボックスモデルは API などを通じた入出力のみにしかアクセスできず、トークン埋め込みにさえアクセスできないものも多いため、このような“厳密な”ブラックボックス条件における選択的忘却は未解決の課題であった。

本研究ではこの課題に対し、モデル内部のいかなる情報にもアクセスすることなく、入力として与えるプロンプトの接尾辞のみを最適化することで選択的忘却を実現する新たな手法を提案した（図 2 右）。具体的には、忘却対象クラスの認識性能を低下させつつ、維持対象クラスの認識性能を維持するような接尾辞列（単語列）を離散的に探索する問題として定式化し、分類精度のみを手がかりに最適化を行う。離散最適化法として標準的な発見的探索法である貪欲法や遺伝的アルゴリズムでは、単語間の共起構造や複数の有望解の存在を十分に捉えられず、探索が不

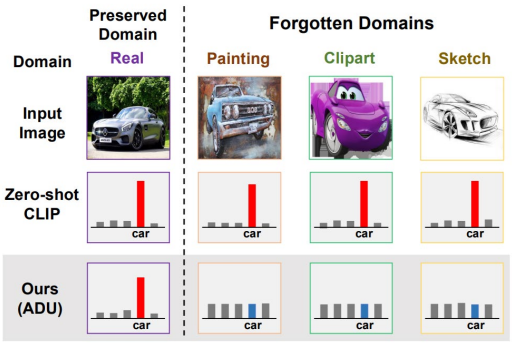

図 4. ドメイン忘却の概要図。従来クラス単位（例：車）での忘却に制限されていたところ、本研究ではドメイン単位（例：イラストの車）での忘却を実現した。

安定になりやすいという問題がある。そこで本研究では、単語選択を多項分布混合で表現し、その事前分布として Dirichlet 分布を導入する確率的多点探索法を考案した。さらに、エリート解に含まれる単語の出現頻度を指数移動平均で反映することで、複数の有望な単語共起構造を安定に保持しながら探索を進める方式を導入した。

複数種類のベンチマークを用いた評価の結果、提案法は、厳密ブラックボックス条件下で比較対象となる貪欲法、遺伝的アルゴリズム、および既存の離散プロンプト探索法を一貫して上回る選択的忘却性能を示した（図 3）。比較的多様かつ大規模な画像を含む ImageNet-1k を用いた実験でも優位性が確認できており、大規模データに対しても適用可能である。以上の結果は、厳密ブラックボックス条件下であっても、離散プロンプトの確率的最適化を通じて実用的選択的忘却を実現可能であることを示すものであり、外部公開 API 型の基盤モデルに対する柔軟な知識制御の可能性を拓く重要な成果である。

本成果は研究会にて発表しているほか、現在国際誌へ投稿中である。

B) ドメインの忘却

CLIP に代表される事前学習済み視覚言語モデルは、高い汎化性能を有しており、ドメインの種類（例：実物かイラストか）によらず、様々な物体を追加学習することなく認識できる（例：実物の車もイラストの車

図 5. ドメインごとの特徴分布を可視化した結果。色はドメイン（実画像、絵画、クリップアート、スケッチ）を表す。左図は元の事前学習済み視覚言語モデルの特徴分布である。異なるドメインのデータが複雑に重なり合っているため、ドメインを個別に忘却することが難しい。右図は我々のドメイン間の特徴分布を分離するための損失関数を適用した後の特徴分布である。異なるドメインを互いに分離し、個別に制御することができるようになる。

も同じ「車」と認識できる）。一方で実応用においては、特定のドメインに属する対象のみを選択的に認識させないことが求められる場合がある。例えば自動運転では、実環境中の物体は正確に認識する必要がある一方、広告やポスターなどに描かれた物体を誤認識することは安全性の観点から望ましくない。このような背景のもと、本研究では、同一クラス内であっても特定ドメインに属する知識のみを選択的に忘却可能にするドメイン単位の選択的忘却を新たに定式化し（図 4）、その実現に取り組んだ。

本課題の難しさは、事前学習済み視覚言語モデルが持つ強いドメイン汎化能力に起因する。すなわち、異なるドメインの特徴分布が潜在空間上で密接に結びついているため、従来の忘却手法ではドメインごとの精密な制御が困難である（図 5 左）。この問題に対し、本研究では、ドメイン間の特徴分布を分離するための損失関数を導入するとともに、入力画像ごとの特徴に応じてプロンプトを動的に調整する機構を組み合わせることで、ドメイン依存の表現を制御する手法を提案した（図 5 右）。これにより、保持すべきドメインの認識性能を維持しつつ、忘却対象ドメインに対してのみ認識できないよう忘却することを可能にした。

複数のマルチドメインデータセットおよび視覚言語モデルを用いた評価の結果、提案手法はベースライン手法と比較して、保持性能と忘却性能の両立に

おいて一貫して優れた結果を示した。これにより、従来困難であったドメイン単位での選択的忘却が実現可能であることを示すとともに、マルチモーダル基盤モデルにおける知識制御の新たな設計指針を提示した。

本成果は、機械学習分野のトップ国際会議 NeurIPS 2025 に Spotlight 論文として採択されたほか、画像認識分野国内最大規模のシンポジウムである画像の認識理解シンポジウム (MIRU 2025)においても口頭発表論文として採択され、国内外で高く評価されている。

C) 破滅的忘却を起こさないバイアス校正

事前学習済みの視覚言語モデルは高い汎用性を有する一方で、集団属性（性別、職業、人種等）との偽相関に起因する望ましくないバイアスに基づいて予測を行うことがあり、公平性や信頼性の観点から重要な課題となっている。この影響を低減すべく、集団属性によるバイアスの影響を抑制する学習法が多数研究されてきた。しかしながら既存研究の多くは、バイアスの低減効果については検証してきている一方で、それによる負の影響、すなわち視覚言語モデルが本来有している汎化性能の低下については、十分に評価してこなかった。

本研究ではまず、プロービング、アダプタ、部分空間射影という 3 つの代表的なアプローチに基づく 10 種類の既存手法について、そのバイアス低減効果と汎化性能への影響を体系的に評価した。11 種類・2,000 クラスを超える複数のデータセットを用いた網羅的な評価の結果、多くの既存手法がバイアス低減効果と引き換えに汎化性能を著しく低下させる、いわゆる破滅的忘却を引き起こすことを確認した（図 6）。この結果は、既存手法の多くが特定の集団属性に対する公平性を改善する一方で、モデルが保持していた広範な知識表現を同時に損なうという本質的なトレードオフを内在していることを示唆している。特に、プロービングやアダプタに基づく手法では、モデル内部の表現全体を変更するため、バイアスに関連する特徴のみならず、それ以外の有用な特徴まで破壊してしまう傾向が観察された。一方で、特徴空間における特定の方向成分を除去する部分空間射影に基づく手法は、他の手法と比較して汎化性能の低下を相対的に抑制できることが確認され、バイアス成分とそれ以外の知識成分が幾何学的に分離可能である可能性が示唆された。

この知見に基づき、本研究では部分空間射影に基づく既存手法を基盤として、バイアス低減と汎化性

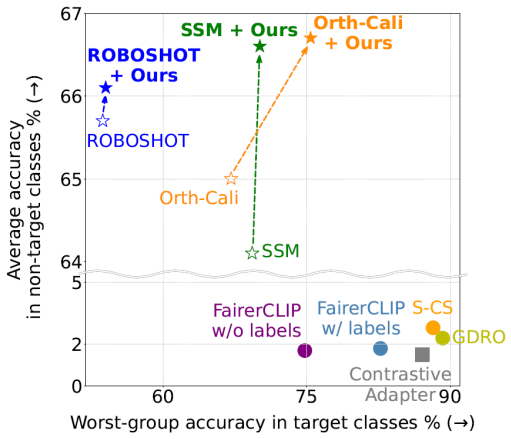

図 6. 既存のバイアス低減手法を評価した結果。各点は既存のバイアス低減手法の性能をバイアス低減効果（横軸）と汎化性能（縦軸）で表している。プロービングに基づく方法（●）、アダプタに基づく方法（■）は高いバイアス低減効果を有する代わりに、著しい汎化性能低下を起こしている。一方、部分空間射影に基づく方法（☆）は比較的その影響を受けにくいことがわかる。提案法は部分空間射影を校正することで、バイアス低減効果を維持しつつ、さらにその汎化性能低下を抑制できる。

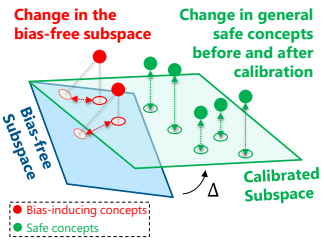

図 7. 提案する射影校正法の概要図。元の射影を極力変化させないようにしつつ、一般語彙に対応する特徴分布を可能な限り変化させないように射影を校正する。

能維持を両立する新たな校正手法を提案した。具体的には、従来の射影操作がバイアスに関連する集団属性のみならず、一般的な概念に関する特徴にまで影響を及ぼしてしまう点に着目し、元のバイアス低減効果を維持しつつ、一般語彙に対応する特徴分布を可能な限り保持するように射影行列を補正する新たな定式化を導入した（図 7）。本定式化は閉形式解を持ち、数秒で求解できる利点を有する。

複数のベンチマークを用いた評価の結果、提案手法は既存の部分空間射影に基づくバイアス低減手法に適用することで、そのバイアス低減効果を維持したまま、汎化性能の低下を抑制できることを確認した（図 6）。この結果は、本手法が、従来両立困難であった公平性と汎化性能のバランスを一貫して向上させることができる汎用手法であることを示唆している。同時に、視覚言語モデルにおけるバイアス校正が単なる性能改善の問題にとどまらず、知識の選択的保持と忘却の観点から設計されるべきであることを示しており、破滅的忘却を回避しつつ信頼性の高いモデルを構築するための重要な指針を与えるものである。

本成果は、MIRU 2025 において口頭発表論文として採択されたほか、研究会発表にも結び付いている。

D) 特徴変換に基づく選択的忘却法

事前学習済みのモデルを仮定するのではなく、新たにモデルを学習する段階から介入できる場合には、あらかじめ学習データに特殊な加工を施すことで、記憶の維持と忘却を制御しやすくするアプローチを採ることができる。従来は入力データ（例えば画像）に対して特殊なパターンを加算する方法が試みられてきたが、記憶の制御は実際には入力そのものではなく特徴空間上の分布に依存する。このため、入力空間での操作は必ずしも直接的な制御を実現できているとは言い難い。本研究ではこの点に着目し、特徴表現そのものに変換を加える新たな枠組みを提案した。

具体的には、従来は入力画像に加算されていたクラス固有のパターン（コード）を中間特徴へ直接加算することで、モデル内部の表現を直接制御する手法を構築した。一方で、中間特徴は画素値とは異なりスケールが不定であるため、単純な加算では安定した情報の埋め込みが困難であるという課題がある。この問題に対し、特徴のスケール整合とコード埋め込みを同時に行う正規化層を導入し（図 8）、特徴分布のばらつきを抑えつつ、忘却に必要な情報を適切に付加できるようにした。

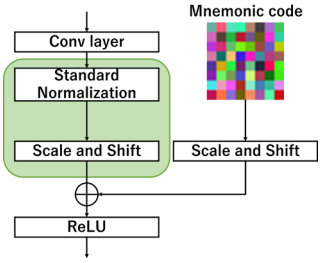

図 8. 提案する正規化層。中間特徴に特殊なパターン（mnemonic code）を加算するが、その際に双方を標準正規化することにより両者のスケールを合わせる。

複数のベンチマークを用いた評価の結果、提案法が従来法と比較して忘却性能・安定性の両面において優れた性能を持つことを確認した。

本成果は論文誌へ採録された。

5.2 マルチモーダル画像生成モデルに対する選択的忘却の実現

拡散生成モデルに代表されるマルチモーダル画像生成モデルは、テキストから高品質な画像を生成できる一方で、不適切な視覚的概念を含む画像や著作権上問題となり得る画像を生成してしまう危険性を内在している。このため、特定の視覚概念のみを選択的に生成しないようモデルを補正する選択的忘却技術の確立は、画像生成 AI の安全・公正な運用に向けた重要課題であるといえる。

とりわけ実運用上は、多くの基盤モデルや視覚言語モデル同様、対象となる拡散モデルが外部サービスとして提供され、モデル本体の内部構造やパラメータに直接アクセスできない場合も少なくない。そのため、モデル本体の再学習や内部パラメータ更新を前提としない軽量な忘却法が必要となる。さらに、仮に忘却法を適用できたとしても、悪意ある攻撃によって忘却済みの概念を復元できてしまっては、そ

図 9. 拡散画像生成モデルに対する忘却法の比較。我々の方法は指定された概念を含まず、かつ、より自然な画像を生成できる。

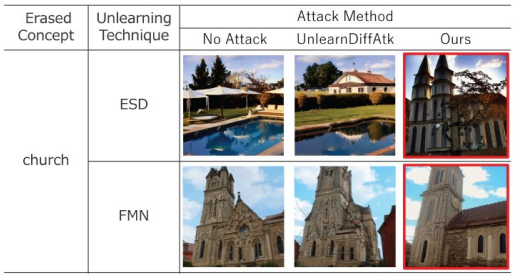

図 10. 忘却概念復元攻撃法の比較。提案法はより確実に忘却した概念を復元できる。

の安全性が脅かされることになりえる。以上の観点から本研究では、拡散モデルにアクセスしない選択的忘却法の検討に加え、拡散画像生成モデルに対する忘却概念復元攻撃法の検討も行った。

以下、これらの成果について順に概説する。

A) 拡散モデルにアクセスしない選択的忘却法

拡散モデル本体にアクセスすることなく、入力プロンプト側の操作のみで特定概念の生成を抑制する選択的忘却法を提案した。

具体的には、学習可能な Escape Token を元のテキストプロンプトの先頭に導入し、そのトークンを適切に最適化することで、対象概念を含む画像が生成されにくい方向へ生成過程を誘導する枠組みを構築した。本手法は、拡散モデル本体やその内部表現を直接更新する既存法と異なり、CLIP テキストエンコーダのみを用いて学習可能である点に特徴がある。さらに、対象概念の影響を明示的に差し引いて背景側の特徴を抽出する Target Concept Subtraction と、文全体ではなく各トークンレベルで背景特徴との整合を課す Dense Token Alignment Loss を導入することで、文脈補完効果によって対象概念が残存してしまう問題を抑制した。

実験では CIFAR-10 の各クラス概念を対象として評価を行った結果、拡散モデルへ直接アクセスできない条件下でありながら、既存法に匹敵する忘却性能を達成するとともに、計算資源とメモリ消費を大幅に削減できることを示した（図 9）。既存法と比較して必要メモリおよび学習対象パラメータ数を削減しつつ、有効な概念抑制を実現できることを確認しており、実利用上の有効性が高い手法である。

本成果は、MIRU 2025 において口頭発表論文として採択された。

B) 拡散画像生成モデルに対する忘却概念復元攻撃法

既存の忘却概念復元攻撃法は、被攻撃モデルの内部構造やパラメータ、さらには忘却対象概念の実例画像へのアクセスを仮定するなど、強いホワイトボックス条件に依存していた。これに対し本研究では、より現実的なブラックボックス条件のもとで、忘却対象概念の近傍概念を線形結合することにより、その概念を再構成して生成に用いる攻撃法を提案した。

具体的には、公開語彙と外部 CLIP を用いて忘却概念に近い概念群を選定し、それらの線形結合係数を最適化することで、忘却済みモデルからも忘却概念に対応する画像を生成させることを目指した。その結果、既存の忘却手法を適用した Stable Diffusion に対して、忘却した概念を一定割合で復元できることを確認し、一部条件では既存のホワイトボックス攻撃法に匹敵、あるいはそれを上回る性能も観測された（図 10）。

これらの結果は、現在の忘却技術が単に「通常のプロンプトでは生成されにくい」状態を実現するだけでは不十分であり、周辺概念を介した概念復元攻撃に対する頑健性まで含めて評価される必要があることを示している。

本成果も MIRU 2025 において口頭発表論文として採択された。

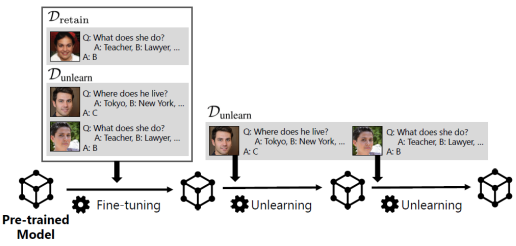

図 11. PULSE の概念図

5.3 マルチモーダル大規模言語モデルにおける選択的忘却ベンチマーキングの整備

マルチモーダル大規模言語モデル（MLLM）は、画像とテキストを統合的に処理し、高度な質問応答や推論を実現できる一方で、学習過程で獲得した個人情報や著作権上の問題を含む知識を保持し続ける可能性がある。このため、不要な知識を後から選択的に忘却させる技術の重要性が高まっている。しかし、MLLM に対する既存の忘却研究は、主として手法提案に焦点が置かれており、それらをどのような条件下で評価すべきかについては十分に整理されてこなかった。特に、既存の評価設定の多くは、直前の fine-tuning により追加された知識を一度だけ忘却する状況を想定しており、実運用上より重要となる「事前学習時に獲得した知識を忘却できるか」あるいは「複数回の忘却要求に継続的に耐えられるか」といった観点を十分に扱えていなかった。

この課題に対し本研究では、MLLM における忘却性能をより実運用に近い条件で評価するための新たな評価プロトコル PULSE を構築した（図 11）。PULSE は、従来の fine-tuning 後に一度だけ忘却処理をする場合の評価に加え、二つの新しい観点を導入している。一点目は事前学習知識の忘却であり、モデルが事前学習段階で獲得した知識を忘却対象とする設定である。二点目は逐次忘却であり、複数の忘却要求が逐次的に与えられる現実的な状況を想定し、連続的な忘却操作に対して性能がどのように変化するかを評価する設定である。これにより、MLLM の忘却性能を単発の成功率だけでなく、知識獲得段階の違いや運用継続時の安定性まで含めて評価できる枠組みを整備した。

実験では、代表的なオープンソース MLLM である

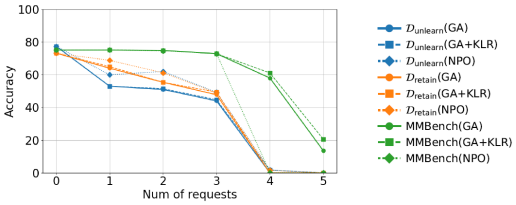

図 12. 逐次忘却の結果。忘却回数の増加に伴い、忘却対象知識の忘却（青）は進むものの、維持すべき知識（オレンジ・緑）まで急速に失われていくことが確認された。

LLaVA-v1.5-13B を対象とし、既存の代表的忘却手法を PULSE 上で比較した。結果、従来想定されてきた fine-tuning 後の知識の単一回数忘却では一定程度の忘却は可能である一方、事前学習において獲得した知識の忘却では、忘却自体は進むものの、維持すべき一般的な知識に対する推論能力が大きく低下し、モデルの汎化能力が損なわれることが分かった。また、逐次忘却では、忘却対象に対する性能低下とともに、維持すべき知識に対する推論能力も急速に失われ、数回の忘却操作で実用上無視できない性能劣化が生じることが確認された（図 12）。さらに、画像入力を伴うマルチモーダル課題よりも、テキストのみの課題の方が忘却されにくい傾向も観察されており、既存手法が真に知識を忘却しているのではなく、画像と知識の対応付けを部分的に破壊しているだけである可能性も示唆された。

以上の結果は、MLLM における選択的忘却技術の有効性を評価するうえで、単一の fine-tuning 忘却シナリオのみでは不十分であり、事前学習知識への忘却可能性と逐次忘却に対する持続性を含めた評価が不可欠であることを示している。PULSE は、こうした現実的要請を反映した評価基盤として、既存手法の限界を可視化するとともに、今後の MLLM 忘却法が目指すべき方向性、すなわち高い忘却性能と汎化性能維持、さらに長期的安定性を両立する設計の必要性を明確にした。

本成果は、MLLM における忘却評価の標準化に向けた基盤的研究として、MIRU 2025 において口頭発表論文として採択されたほか、NeurIPS 2025 Workshop

に採択されている。

6 進捗状況の自己評価と今後の展望

全体として当初計画時の目標通りの成果を創出することができた。学術成果としては学術論文誌 1 件、国際会議 2 件、国内会議発表 6 件の発表に加え、2 件の論文を国際誌へ投稿中である。

本研究では、選択的忘却を中核原理として、(1) 深いブラックボックスモデル、(2) マルチモーダル画像生成モデル、(3) マルチモーダル大規模言語モデル（MLLM）という異なる設定に対し、忘却手法の開発および評価基盤の整備を進めた。結果、モデル内部にアクセスできない厳密ブラックボックス条件における忘却の実現、生成モデルに対する軽量な概念抑制手法およびその安全性評価手法の提案、さらに MLLM における実運用を想定した忘却ベンチマークの構築に至り、選択的忘却技術の適用範囲を大きく拡張することができた。

特に重要な成果として、(i) モデル内部情報に依存しない忘却実現手法の確立、(ii) 忘却後のモデルに対する攻撃可能性の実証、(iii) 従来考慮されてこなかった評価軸（事前学習知識の忘却・逐次忘却）の導入が挙げられる。これらの成果により、選択的忘却は単なる知識削除手法にとどまらず、モデルの運用条件や攻撃環境を含めて設計・評価されるべき技術であることが明確となった。一方で、現時点の技術にはいくつかの限界も明らかとなっている。すなわち、(a) 忘却は可能であるがモデルの規模自体は削減されないこと、(b) 忘却した知識が周辺概念やプロンプト操作を通じて再生成されうること、(c) MLLM に対する忘却性能の評価プロトコルは依然限定的であることである。これらの課題は、選択的忘却を基盤モデルの実運用に適用する上で本質的な制約となるものであり、本研究を通じて体系的に認識された重要な問題である。このような認識に基づき、今後は単なる忘却性能の向上にとどまらず、基盤モデルの運用性および安全性を同時に向上させる方向へと研究を発展させる必要がある。

具体的には、次年度以降は以下の三つの方向性に基づき研究を展開する予定である。第一に、不要な知識の削減を通じてモデルの規模そのものを縮小する忘却駆動型のモデル圧縮技術の探求である。これは、選択的忘却を単なる知識編集ではなく、モデル構造の最適化へと拡張するものであり、基盤モデルの運用コスト低減に直結する。第二に、忘却後モデルに対する知識復元攻撃に対する防御技術の確立である。本研究で明らかとなった攻撃可能性を踏まえ、レッドチーミングに基づく体系的評価と防御機構の設計を行い、忘却技術の安全性を向上させる。第三に、言語を跨いだ忘却の一貫性を評価する多言語ベンチマークの構築である。これは、国際展開を前提とした基盤モデルにおいて不可避となる言語依存的リークの問題に対処するものである。

以上のように、本年度の研究により選択的忘却の基礎技術と評価基盤は一定の成熟に達した一方で、その運用性および安全性に関する新たな課題も明確化された。今後はこれらの課題に対し、忘却原理を基盤モデルの設計・運用に統合することで、持続可能なマルチモーダル基盤モデルの実現に向けた研究を推進する。