クラス名に依存しない視覚的記述子探索による CLIP の OOP 適応

川越 壮[1,a)]　内海 ゆづ子[1,b)]　岩村 雅一[1,c)]

概要

CLIP は事前学習で見たことのない概念 (Out-of-Pre-training, OOP) では画像とクラス名を対応づけた Zero-shot 分類に失敗する．本稿では，クラス名も勾配学習も用いず，記述子プールから少数のサポート画像を手がかりに視覚的な記述子を貪欲探索で選択，加算することで，OOP クラスの画像埋め込みと対応するテキスト埋め込みを構築する手法を提案する．提案手法は Zero-shot 分類精度を 12.9%から 54.2%へ改善し，さらに 1,2-shot では既存手法である FSNL を上回る．

1. はじめに

画像と言語を共通の特徴空間で結びつける CLIP [9] は，クラス名を含むプロンプトの埋め込みを利用することで追加学習なしの Zero-shot 分類を可能にした．しかし，事前学習で獲得されていない概念は，クラス名が視覚特徴と対応しておらず，Zero-shot では識別できない．LAION-Beyond [3] は，CLIP の事前学習に含まれる概念を In-Pre-training (IP)，含まれない概念を Out-of-Pre-training (OOP) として明示的に定義し，この問題を顕在化させた．

OOP クラスでは，画像特徴はクラスごとにまとまり区別できるにもかかわらず，クラス名プロンプトの埋め込みが事前学習で視覚特徴と対応していない．実際，クラス名プロンプトを用いる Zero-shot 分類は IP クラスで 31.3%を保つ一方，OOP クラスでは 12.9%まで低下する．失敗の原因は画像側ではなくテキスト側，すなわちクラス名が特徴空間で視覚概念に対応づいていない点にあり，OOP ではクラス名がボトルネックになる．そこで，Huang [3] らは Few-shot Name Learning (FSNL) を提案し，IP のテキスト分類器を凍結したまま OOP のクラス名埋め込みのみを Few-shot で学習して画像との不整合を直接埋め，OOP で高い精度を示した．しかし得られる表現は勾配学習による連続的な埋め込みであり，学習にコストがかかる．

そこで，FSNL のようにクラス名埋め込みを学習し直すのではなく，クラス名を使わずに視覚的な特徴を表す記述子でクラスを表現することを考える．縞模様や羽，細長い形状といった汎用的な視覚句は，事前学習で画像と対応しており，OOP クラスの識別に活用できる．記述子でクラスを表現する試みは，概念や属性を介して分類する概念ボトルネックモデルや，記述子ベースの Zero-shot 分類として研究されてきた．しかしこれらの多くは，記述子の生成や選択，あるいは概念からクラスへの対応づけの段階でクラス名を用いる．OOP ではクラス名が画像と対応していないため，クラス名に依存せずに記述子を構成する必要がある．

本研究はクラス名も勾配学習も用いず，少数のサポート画像（適応に使う少数のラベル付き画像）から得た画像埋め込み平均（画像プロトタイプ）を手がかりに，クラス名を含まない汎用記述子を貪欲探索で選び，各記述子の埋め込みを線形加算した埋め込みで OOP クラスの分類器を構成する記述子探索を提案する．選ばれるのが自然言語の記述子列であるため解釈可能である．

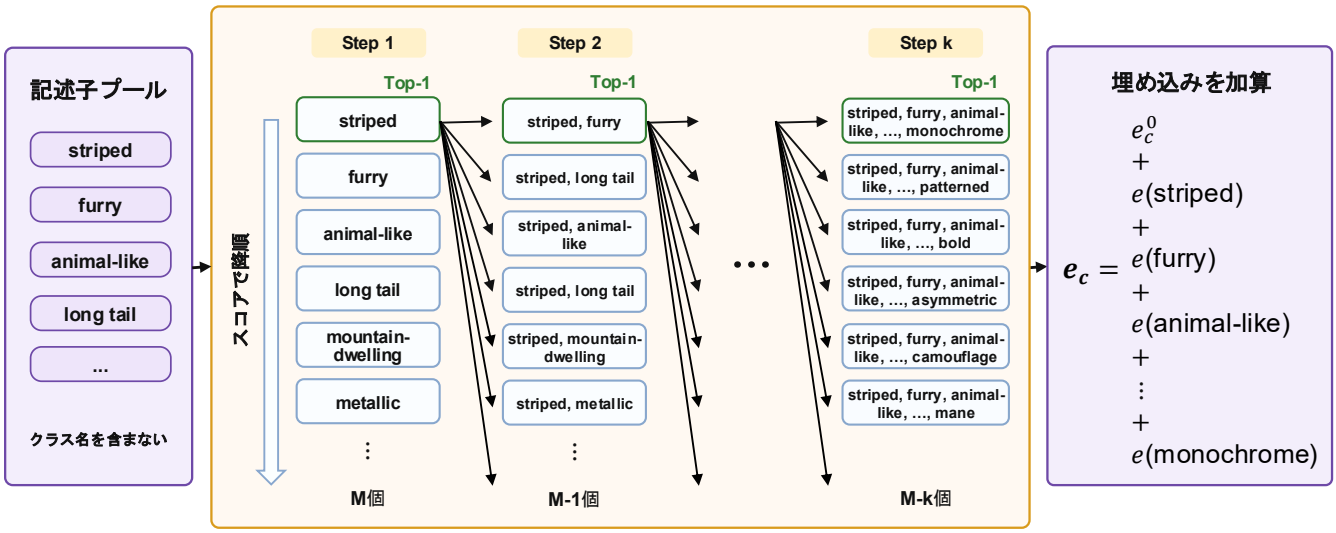
本研究の貢献は次の 3 点である．(1) 提案手法は LAION-Beyond の OOP 674-way において，4-shot でクラス名プロンプトを用いる Zero-shot の分類精度を 12.9%から 54.2%へ大きく改善する．(2) 1,2-shot では勾配学習を行う FSNL を上回る．(3) 記述子探索は勾配学習を要さず，FSNL の学習時間に対して 1-shot で約 1/74，16-shot で約 1/627 の計算時間で済み，かつ OOP クラスを人間が読める自然言語記述子として表現できる．

2. 関連研究

2.1 記述子に基づく Zero-shot 分類

CLIP のクラス名プロンプトを大規模言語モデルの生成するクラス説明文で補強する方向が広く研究されている．Menon と Vondrick [6] は，カテゴリ名の代わりに視覚的特徴の有無で分類する classification by description を提案した．CuPL [8] は言語モデルにクラスごとの記述文を生成させ，CLIP-A-self [5] は生成した多数の説明文を

[1] 大阪公立大学
[a)] sp25249p@st.omu.ac.jp
[b)] yuzuko@omu.ac.jp
[c)] masa.i@omu.ac.jp



(a) (b) (c)

図 1 提案手法の概要．少数のサポート画像から画像プロトタイプ $\boldsymbol{v}_{\text{proto},c}$ を作り，これをターゲットとして記述子を選ぶ．図は左から，(a) クラス名を含まない記述子プール，(b) 記述子探索，(c) 埋め込みの加算を表す．

Self-Attention で集約する．FuDD [1] は，紛らわしいクラスを区別する差分的な記述を画像ごとに生成する．一方 WaffleCLIP [10] は，ランダムな単語列でも同程度の精度向上が得られることを示し，意味のある記述子の必要性に疑問を投げかけた．これらはいずれもクラス名プロンプトに記述子を付加する枠組みであり，クラス名そのものを用いる点で本研究と異なる．これに対し Ma ら [4] は，識別的な記述子を数個選ぶだけで，記述子をクラス名に付加する従来手法に匹敵する精度が得られることを示し，クラス名に依存しない記述子表現の可能性を示唆した．ただし，その評価は OOP では行われておらず，また記述子の選択にあたって混同しやすいクラスをクラス名の埋め込みから特定する段階を含むため，クラス名が信頼できない OOP では精度が低下すると予想される．本研究はこの方向をさらに進め，クラス名を一切用いず，画像から得たクラス方向のみを手がかりに記述子を合成して OOP クラスを表現する．

2.2 概念ボトルネックと解釈可能な分類

人間が読める概念の集合を介して分類を行う概念ボトルネックモデルも，解釈可能性の観点で本研究と関連する．LaBo [12] は言語モデルが生成した概念文から識別的なものを選び，Label-Free CBM [7] は任意のネットワークを概念ボトルネックに変換する．いずれも概念を介した解釈性をもつが，概念とクラスを対応づける分類器の学習を要する．Zero-shot Concept Bottleneck Models (ZCBM) [11] は，大規模な概念バンクからのクロスモーダル検索によって学習なしで概念を選ぶ点で，本研究の汎用記述子プールからの離散選択に近い．本研究はこの ZCBM が構築した大規模な記述子プールを探索対象として用いる．ただしこれらはいずれも，概念からクラスを予測する段階でクラス名を用いる点で本研究と異なる．

2.3 OOP 一般化と Few-shot 適応

Huang ら [3] は，事前学習コーパスに含まれない概念への一般化を OOP 問題として定式化し，9 ドメインからなるベンチマーク LAION-Beyond を構築した．この論文で提案された FSNL は，CLIP を凍結したままクラス名埋め込みのみを Few-shot で学習する name-tuning であり，OOP で高い精度を示す．Few-shot 適応の一般的な手法としては，連続的なコンテキストベクトルを学習する CoOp [14] や CoCoOp [13]，特徴量にアダプターを挿入する CLIP-Adapter [2] がある．これらは OOP の精度を Zero-shot から向上させるものの，クラス名プロンプトの埋め込み自体は凍結したままその周囲を最適化するに過ぎず，視覚特徴と特徴空間で対応しないという根本問題は解消されない．いずれも勾配学習による連続表現の最適化であり，本研究は離散的な記述子を学習なしで探索する点で対照的である．

3. 提案手法

提案手法の全体像を図 1 に示す．OOP クラス c に対し，その K 枚のサポート画像から CLIP 画像エンコーダーで特徴量を求め，その平均を画像プロトタイプ $\boldsymbol{v}_{\text{proto},c}$ とする．これを探索のターゲット $\boldsymbol{t}_c$ とし，クラス名を含まない記述子の埋め込みの和で $\boldsymbol{t}_c$ を近似する．提案手法は，(a) 記述子プール，(b) 記述子探索，(c) 埋め込みの加算からなる．

(a) 記述子プール（図 1(a)）． 記述子は，クラス名を含まない汎用的な視覚句からなる大規模な記述子プールから

表 1 各手法の性質と shot 数ごとの OOP 674-way top-1 精度 (%)．CLIP ZS はサポート画像を用いない 0-shot 基準値．**太字**は提案手法と FSNL のうち高い方．

手法	勾配学習	クラス名	解釈可能	OOP top-1 (%)					
				0-shot	1-shot	2-shot	4-shot	8-shot	16-shot
CLIP Zero-shot	不要	要	○	12.9	-	-	-	-	-
提案手法 (Ours)	不要	不要	○	-	**35.0**	**45.4**	54.2	59.7	62.2
FSNL [3]	要	不要	×	-	31.4	44.3	**56.3**	**65.0**	**71.4**

選ぶ．各記述子 d は “a photo of a d” という短い文として CLIP テキストエンコーダーで埋め込み，その埋め込みを $\boldsymbol{e}(d)$ とする．これらはあらかじめ計算しておく．

(b) 記述子探索（図 1(b)）．初期ベクトル $\boldsymbol{e}_c^0$（“a photo of something” の埋め込み）から始め，追加する記述子は貪欲探索で選ぶ．まずプール全体から $\boldsymbol{t}_c$ へのコサイン類似度が高い上位 M 句に候補を絞り込む（prefilter）．ステップ i $(i \geq 1)$ では，前ステップのクラス埋め込み $\boldsymbol{e}_c^{i-1}$ に候補記述子 d の埋め込み $\boldsymbol{e}(d)$ を加えた $\boldsymbol{e}_c^i = \boldsymbol{e}_c^{i-1} + \boldsymbol{e}(d)$ のスコアを計算し，スコアが最も高くなる記述子を選ぶ．クラス埋め込み $\boldsymbol{e}_c^i$ のスコアを次式で定義する．

$$\mathrm{score}(\boldsymbol{e}_c^i) = \cos(\boldsymbol{e}_c^i, \boldsymbol{t}_c) - \lambda \max_{c' \neq c} \cos(\boldsymbol{e}_c^i, \boldsymbol{t}_{c'}) \qquad (1)$$

第 1 項はターゲットへの近さ，第 2 項は最も紛らわしい他クラスのターゲットから離す識別的な罰則であり，λ でその強さを調整する．これにより，ターゲットに似せるだけでなく，紛らわしい隣のクラスと見分けがつくように記述子が選ばれる．スコアが改善しなくなったら停止する．

(c) 埋め込みの加算（図 1(c)）．最終的なクラス c の埋め込み $\boldsymbol{e}_c$ は，探索停止時の累積埋め込み $\boldsymbol{e}_c^i$（初期ベクトル $\boldsymbol{e}_c^0$ に選んだ記述子の埋め込み $\boldsymbol{e}(d)$ をすべて加えたもの）を正規化して単位ベクトルにしたものとする．最終的に各クラスを $\boldsymbol{e}_c$ で表現し，全クラスとのコサイン類似度で分類する．

4. 実験

4.1 実験設定

全ての手法に共通のバックボーンとして OpenCLIP ViT-B/32 (LAION-400M) を用い，評価は LAION-Beyond の 674 OOP クラスで行う．記述子プールは ZCBM [11] の約 914 万句，prefilter 幅 $M = 20{,}000$，識別罰則 $\lambda = 0.1$ とした．識別罰則 λ は精度をわずかに左右し，予備実験では $\lambda = 0.1$ 付近で最大となった．サポート画像枚数（shot 数）を 1, 2, 4, 8, 16 と変えて評価する．

4.2 OOP の結果

表 1 に，各手法の性質と shot 数ごとの OOP 674-way top-1 精度を示す．提案手法は，クラス名にも勾配学習にも依存せず，4-shot でクラス名プロンプトを用いる Zero-shot の分類精度を 12.9%から 54.2%へ大きく改善する．さらに，

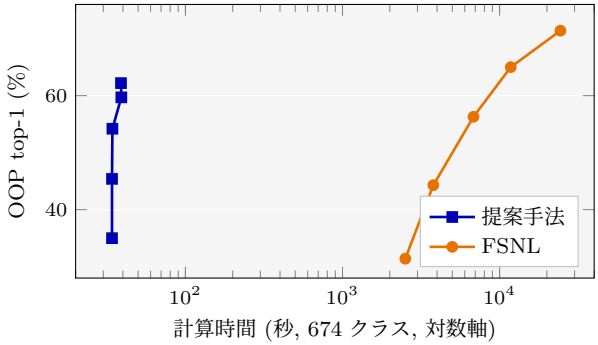


図 2 計算時間と精度（OOP 674-way, 1/2/4/8/16-shot）．提案手法（青）は shot によらず約 34〜39 秒で完了し，FSNL の学習（橙, 2510〜24394 秒）の数十〜数百分の一の時間で動作する．低 shot ほど提案手法が左上（高速・高精度）に位置する．

1-shot（35.0%対 31.4%）と 2-shot（45.4%対 44.3%）では FSNL を上回る．記述子の加算という制約が，サポート画像が少ないときに暗黙の正則化として働き，過学習を抑えるためと考えられる．一方，4-shot 以降はクラス名埋め込みを直接学習する FSNL が上回り，その差は shot 数とともに広がる．ただし FSNL が勾配学習を要し表現も解釈できないのに対し，提案手法は学習なしで解釈可能な記述子表現を与える．

4.3 精度と計算時間

提案手法の探索は勾配学習を行わないため，FSNL の学習に比べて計算コストが極めて小さい．図 2 に，674 クラス分の適応に要する計算時間（対数軸）と精度の関係を示す．提案手法は shot 数によらず約 34–39 秒で完了するのに対し，FSNL の学習は 1-shot で約 42 分（2510 秒），16-shot で約 407 分（24394 秒）を要する．すなわち提案手法は FSNL の学習に対し 1-shot で約 74 倍，16-shot で約 627 倍高速である．とくに低 shot では，提案手法は FSNL より高速かつ高精度であり，計算コストと精度の両面で優位に立つ．

4.4 記述子の解釈性

提案手法の出力は連続ベクトルではなく自然言語の記述子列であり，各クラスをどの視覚特徴で表現したかを読み取れる．選ばれた記述子の視覚的妥当性を大規模言語モデルに 0〜2 の 3 段階で採点させたところ，提案手法が選ん

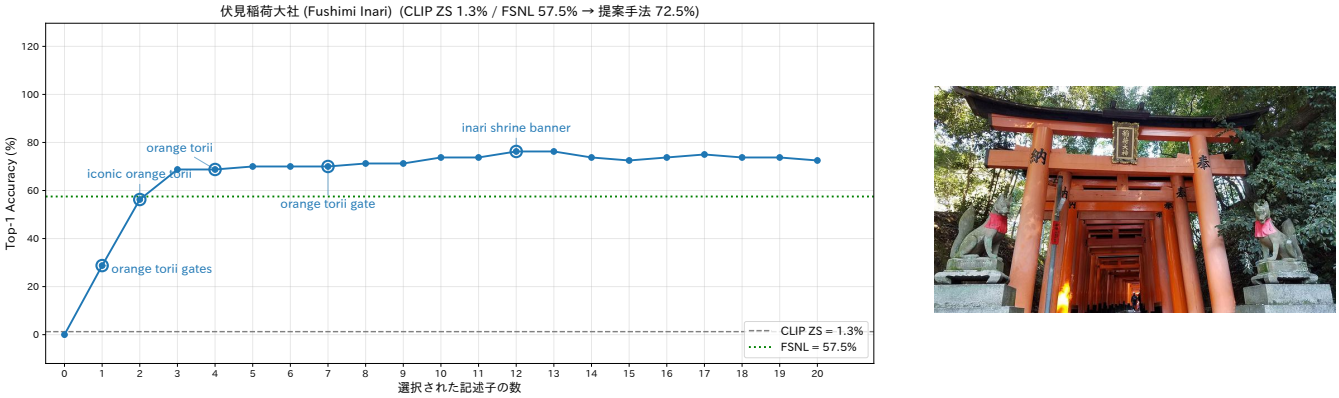


図 3 「伏見稲荷大社」（OOP の Landmark クラス）における記述子の追加（横軸=記述子数）に伴う top-1 精度の推移（左）と，そのクラスの画像例（右）．クラス名プロンプトを用いる Zero-shot 分類が 1.3%しか当てられないのに対し，“iconic orange torii” などの鳥居をとらえる記述子が選ばれて精度が伸び，提案手法（75.0%）は FSNL（57.5%, 破線）をも上回る．各点のラベルはそのステップで追加された記述子．

表 2 提案手法が選んだ記述子の例（9.14M プール, 4-shot）．**太字**は LLM 評価が 2（明確な視覚記述子）．建築やランドマークでは妥当な視覚語が選ばれる一方，近縁種の多い生物や架空の固有名では別種名・無関係語に流れる．

OOP クラス（ドメイン）	選ばれた記述子（抜粋）
伏見稲荷大社 (Landmark)	**iconic orange torii**, **orange torii gate**, fushimi inari taisha, vivari shrine banner …
Familok (Architecture)	**a large red brick building**, **red brick prudential**, fortified residence …
Western Box Elder Bug (Insects)	mangrove shield bugs, bivoltine orange, red flat bark beetle …
Quaquaval (Pokemon)	mandarin ducks, blue orange trim, dancing mudkip, zazu exclusive …

だ記述子の平均は 0.61 で，同じプールからのランダム抽出の 0.20 を大きく上回った．記述子は一定の視覚的意味をもって選ばれている．

図 3 に，成功例について記述子を 1 つずつ追加したときの精度の推移を示す．建築やランドマーク，装身具のように視覚的な特徴が際立つ人工物では，それを的確にとらえる記述子が選ばれて精度が伸びる．たとえば伏見稲荷大社（OOP の Landmark クラス）では，クラス名プロンプトを用いる Zero-shot 分類が 1.3%しか当てられないのに対し，“iconic orange torii” や “orange torii gate” といった鳥居をとらえる記述子が選ばれ，提案手法は 75.0%と FSNL の 57.5%をも上回る．一方，近縁種が多い細粒度の生物では，記述子が近縁種や無関係語に流れて精度が伸びにくい．表 2 に，こうした成功例（建築・ランドマーク）と失敗例（生物・架空の固有名）で選ばれた記述子の例を示す．解釈性の高さ（妥当な視覚記述子が選ばれること）と精度の高さは同じ要因から生じており，記述子の可読性が「なぜ当たる／外すか」の説明になっている．

また，提案手法は 1 クラスを平均約 26 句の記述子で表現する．この選択数に対して prefilter 幅 $M = 20{,}000$ は十分大きく，選ばれる記述子の多くは画像プロトタイプへのコサイン類似度では上位に来ないため，M を選択数程度に絞ると有効な記述子を取りこぼす．

5. まとめ

本研究では，クラス名を用いず，勾配学習も行わずに，ZCBM の約 914 万句の記述子プールから貪欲探索で選んだ視覚記述子の加算だけで OOP クラスを表現する記述子探索を検討した．LAION-Beyond の 674 OOP クラスにおいて，提案手法は 4-shot でクラス名プロンプトを用いる Zero-shot の分類精度を 12.9%から 54.2%へ大きく改善し，1,2-shot では FSNL を上回った．4-shot 以上では FSNL に及ばないものの，提案手法は学習を要さず FSNL の最大数百分の一の計算時間で動作し，OOP クラスを解釈可能な自然言語記述子として表現できる．

今後は，記述子プールのノイズ除去によって解釈性と精度をさらに高めるとともに，FSNL のような高精度な学習手法と提案手法を補完的に組み合わせる方向を検討する．

謝辞

本研究は，JSPS 科研費 24K03020 の助成を受けて実施された．

参考文献

[1] Esfandiarpoor, R. and Bach, S. H.: Follow-Up Differential Descriptions: Language Models Resolve Ambiguities for Image Classification, *Proceedings of the International Conference on Learning Representations* (2024).

[2] Gao, P. and et al., S. G.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters, *International*

Journal of Computer Vision, Vol. 132, No. 2, pp. 581–595 (2023).

[3] Huang, X., Chen, Z., Fan, X., Wang, K., Zhou, Y., Guan, Q. and Lin, L.: Reproducible Vision-Language Models Meet Concepts Out of Pre-Training, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025).

[4] Ma, P., Rietdorf, L., Kotovenko, D., Hu, V. T. and Ommer, B.: Does VLM Classification Benefit from LLM Description Semantics?, *Proceedings of the AAAI Conference on Artificial Intelligence* (2025).

[5] Maniparambil, M. and et al., C. V.: Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts, *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 262–271 (2023).

[6] Menon, S. and Vondrick, C.: Visual Classification via Description from Large Language Models, *Proceedings of the International Conference on Learning Representations* (2023).

[7] Oikarinen, T., Das, S., Nguyen, L. M. and Weng, T.-W.: Label-Free Concept Bottleneck Models, *Proceedings of the International Conference on Learning Representations* (2023).

[8] Pratt, S. and et al., I. C.: What Does a Platypus Look Like? Generating Customized Prompts for Zero-Shot Image Classification, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023).

[9] Radford, A. and et al., J. W. K.: Learning Transferable Visual Models From Natural Language Supervision, *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763 (2021).

[10] Roth, K. and et al., J. M. K.: Waffling Around for Performance: Visual Classification with Random Words and Broad Concepts, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023).

[11] Yamaguchi, S. and et al., K. N.: Zero-shot Concept Bottleneck Models, *Proceedings of the International Conference on Learning Representations* (2025).

[12] Yang, Y. and et al., A. P.: Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).

[13] Zhou, K., Yang, J., Loy, C. C. and Liu, Z.: Conditional Prompt Learning for Vision-Language Models, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16795–16804 (2022).

[14] Zhou, K., Yang, J., Loy, C. C. and Liu, Z.: Learning to Prompt for Vision-Language Models, *International Journal of Computer Vision*, Vol. 130, No. 9, pp. 2337–2348 (2022).