バナー広告における画像と広告コピーの評価ベンチマーク構築

遠藤 洸亮 [1]　脇本 宏平 [2]　宮西 洋輔 [2]　岡崎 直観 [1]
[1] 東京科学大学情報理工学院　[2] 株式会社サイバーエージェント
{kosuke.endo@nlp.,okazaki@}comp.isct.ac.jp
{wakimoto_kohei,miyanishi_yosuke}@cyberagent.co.jp

概要

インターネット広告市場の拡大に伴い，バナー広告の需要が急増している．バナー広告は視覚的に訴求する画像と，商品の価値や特徴を伝える広告コピーで構成され，その組み合わせの「相応しさ」は広告効果に影響するとされる．しかし，この相応しさを評価する研究はこれまで行われていない．本研究では，相応しさを評価するため，「コピーがバナー画像の内容を再現しているか」および「バナー画像がコピーの内容を再現しているか」に基づくバナー画像とコピーの表現再現性評価タスクを提案し，ベンチマークを構築する．また，表現再現性とバナー広告としての相応しさの関係を調査する．そして，既存手法による表現再現性の評価実験を行い，評価の課題を明らかにする．

1　はじめに

近年拡大しているインターネット広告市場 [1] では，画像と広告コピー（以下コピー）を組み合わせ，視覚的に情報を伝えるバナー広告（図 1）が注目されている．バナー広告は配信対象者の性別や興味関心などを考慮し，条件に応じて自動的に個別配信されている．これに伴い，広告制作会社は各条件で高い効果を発揮するバナー広告を大量に制作する必要がある．この高い需要に応えるため，画像生成モデル [2, 3] と大規模言語モデル [4, 5, 6] から生成した画像とコピー [7] を組み合わせたり，同じ商材を対象にした異なるバナー広告から，画像とコピーを組み合わせるなどして，バナー広告の自動作成が検討されている．

しかし，バナー画像とコピーを単純に組み合わせる手法では，画像とコピーの内容が一貫せず，広告効果の低いバナー広告が生成される可能性がある．例えば，図 1 のバナー画像に対して「40 代から始める自分磨きって？」というコピーは年齢に齟齬があり，組み合わせたときの広告効果が低くなるだけでなく，閲覧者の印象を著しく損なう可能性がある．広告分野では広告効果を高めるために数多くの研究が進められてきたが，バナー画像とコピーの**組み合わせの相応しさ**に関する研究は行われていない．

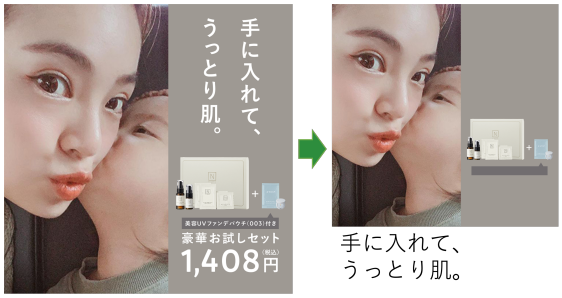

図 1: バナー広告と構成するバナー画像とコピー

例えば，張ら [8] は広告文容認タスクを提案し，文章がコピーであるかどうか，言語モデルが理解できるかを検証した．しかし，画像とコピーの組み合わせがバナー広告として容認できるかには取り組んでいない．Hussain ら [9] はバナー広告理解タスクを提案し，視覚言語モデルがバナー広告の訴求内容を理解できるかを調べたが，画像とコピーの組み合わせの相応しさは研究していない．さらに，バナー画像とコピーの組み合わせの相応しさの評価は画像キャプション生成タスク [10] におけるキャプションの評価とは異なる．画像キャプション生成の目的は画像を忠実に言葉で再現することであり，その評価は画像をどれだけ正確かつ詳細に伝えているかに基づく．一方で，コピーの目的は閲覧者に購買行動やブランド認知などの行動変容を促すことである．そのため，コピーは画像に忠実である必要はなく，画像が持つ訴求内容をより幅広い形で再現し，閲覧者の興味を喚起することが求められる．

そこで本研究では，バナー画像とコピーの組み合わせの相応しさを定量的に評価するため，**バナー広告における表現再現性評価タスク**を提案し，そのた

めのベンチマークを構築した．表現再現性とは，**コピーがバナー画像の重要箇所を再現しているか**，および**バナー画像がコピーの重要箇所を再現しているか**を示す．調査を通じて，表現を再現し合うバナー画像とコピーの組み合わせに対して，人間はバナー広告として相応しいと判断する傾向があることが判明した．また，提案ベンチマークに対して既存手法で実験したところ，既存手法は表現再現性評価タスクで高い精度を達成できないことが示された．そして，zero-shot もしくは few-shot の汎用的な大規模視覚言語モデル [11] よりも，このタスクにファインチューニングした比較的小規模なマルチモーダル基盤モデル（CLIP [12]）の方が性能が良いことが分かった．このことから，バナー画像とコピーの相応しい組み合わせを理解するには，広告ドメインに特化したデータが必要であることが明らかになった．

2　バナー画像とコピーの表現再現性評価ベンチマーク

本研究では，バナー広告における画像とコピーの組み合わせの相応しさを定量的に評価するため，**バナー画像とコピーの表現再現性評価タスク**を提案し，そのためのベンチマークを構築した．バナー画像とコピーの表現再現性は以下の二つの要素から構成される：(1) コピーがバナー画像の重要箇所を再現しているかを示す**コピーにおけるバナー画像の表現再現性**と，(2) バナー画像がコピーの重要箇所を再現しているかを示す**バナー画像におけるコピーの表現再現性**である．さらに表現再現性の有効性を検証するため，バナー画像とコピーの表現再現性とバナー広告の相応しさの関係を調査した．

バナー画像とコピーの表現再現性を定義する前に，**バナー画像の表現要素**と**コピーの視覚表現要素**を定める．**バナー画像の表現要素**とはコピーで表現した方が組み合わせたバナー広告として相応しいと考えられるバナー画像の重要箇所である．具体的にはバナー画像に含まれる物体や図形から，商品，物体を保持する手や皿，背景にある物体を除いたものが該当する．なお，商品画像を表現要素から除外したのは，広告主固有の特徴を排除した汎用的なデータセット構築を目指すためである．

コピーの視覚表現要素とはバナー画像で表現した方が組み合わせたバナー広告として相応しいと考えられるコピー中の重要箇所である．具体例と例外を表 1 に示す．なお，金融サービスに関係する言葉が例外となるのは，バナー広告ではお金を画像で描写することを避けるからである．また，広告商品の使用動作が例外となるのは，商材から使用動作が容易に想起され，画像描写が不要であるからである．

表 1: コピーの視覚表現要素と例外項目

コピーの視覚表現要素	**例**
固有名詞を除く，視認できる物体・状態	体重計，(皮膚の) 赤み，男，20 代
潤い，輝き，香りに関わる言葉	ぷるぷる，つや，香り
画像で表現されるべき代名詞や人	これ，私
視覚表現要素としない例外項目	**例**
否定の意味を含む表現	○○要らず
金融サービスに関係する言葉	返金，100 万円
抽象的で画像で表現しにくい言葉	サポート，体調
時間に関わる言葉	夏，朝
感情と関わる言葉	好き，疲れ
広告表示者や広告主を表す言葉	あなた，当社
広告を対象とする言葉	ご案内，クリック
広告商品の使用動作	飲む，つける

前述のバナー画像とコピーの表現要素を用いて，**コピーにおけるバナー画像の表現再現性評価タスク**と**バナー画像におけるコピーの表現再現性評価タスク**を定義する．各タスクはバナー画像やコピーに表現要素が存在するか，一方の素材が他方のすべての表現要素を再現しているかに基づき，バナー画像とコピーの組み合わせを以下の各タスクにおける三つのクラスのいずれかに分類するタスクである．

コピーにおけるバナー画像の表現再現性評価タスク

- コピーがバナー画像のすべての表現要素を再現している
- コピーがバナー画像のいずれかの表現要素を再現していない
- バナー画像に表現要素が存在しない

バナー画像におけるコピーの表現再現性評価タスク

- バナー画像がコピーのすべての視覚表現要素を再現している
- バナー画像がコピーのいずれかの視覚表現要素を再現していない
- コピーに視覚表現要素が存在しない

2.1　ベンチマークの構築

本研究では，バナー画像とコピーの組に対して表現再現性評価タスクのアノテーションを行い，ベン

表 2: 人手による表現再現性評価ラベル分布

コピーにおけるバナー画像の表現再現性評価タスク				
ラベル	**学習**	**検証**	**評価**	**合計**
コピーがバナー画像のすべての表現要素を再現している	94	136	114	**344**
コピーがバナー画像のいずれかの表現要素を再現していない	182	154	344	**680**
バナー画像に表現要素が存在しない	252	130	464	**846**
各データセットの総数	528	420	922	**1,870**

バナー画像におけるコピーの表現再現性評価タスク				
ラベル	**学習**	**検証**	**評価**	**合計**
バナー画像がコピーのすべての視覚表現要素を再現している	75	127	109	**311**
バナー画像がコピーのいずれかの視覚表現要素を再現していない	218	172	527	**917**
コピーに視覚表現要素が存在しない	219	97	217	**533**
各データセットの総数	512	396	853	**1,761**

チマークを構築し，既存手法による評価性能を示す．バナー画像とコピーの組に対し，バナー広告としての**組み合わせの相応しさ**に評価の焦点を当てるため，一定の広告効果のあったバナー広告の画像と，同じ商品で異なるバナー広告のコピーを組み合わせ，擬似バナー広告ペアデータを作成した．なお，一般的なバナー広告の作成状況を模擬するため，図 1 のようにテキストを除去したバナー画像とコピーを組み合わせてペアデータとした．そして，配信済みバナー広告の画像とコピーのペアデータと作成した擬似ペアデータに対して，表現再現性評価タスクのアノテーションを行い，ベンチマークを構築した．日本語母語話者であるアノテータ 3 人がガイドラインの説明を受けた上でアノテーションを実行した．正解ラベルは多数決により定め，一意に定まらないデータはベンチマークから除外した[1)]．また商材，コピー，バナー画像が混合しないよう学習，検証，評価データセットに分割した．その結果，表現再現性評価ベンチマークの人手評価のラベル分布は表 2 となった．アノテータ間の一致度合いを示す Fleiss' Kappa [14] はコピーにおけるバナー画像の表現再現性評価タスクで 0.54, バナー画像におけるコピーの表現再現性評価タスクで 0.43 であり，人間の間で一致を得るのが難しいタスクであった．

2.2 表現再現性とバナー広告としての相応しさの関係

バナー広告作成における表現再現性の有効性を検証するため，同一素材に対して表現再現性の異なる素材ペアを比較し，バナー広告として相応しいと人間に判断される表現再現性ラベルの割合を調査した．同一の画像が再現しているコピーと再現していないコピーの比較調査では，73%の割合で再現しているコピーの方が相応しいと判断された．また，同一のコピーが再現している画像と再現していない画像の比較調査では，80%の割合で再現している画像の方が相応しいと判断された．なお，詳細は付録の表 4 に示す．この結果から，バナー画像とコピーで互いに表現を再現し合う組み合わせがバナー広告として相応しいと判断される傾向があり，バナー広告における表現再現性の有効性が示された．調査はベンチマーク作成者とは異なる 3 人が担当した．

3 実験設定

2.2 節で，互いの表現を再現するバナー画像とコピーの組み合わせがバナー広告として相応しいと判断される傾向があると分かった．そこで，人手によらずとも，機械学習モデルが画像とコピーの相応しい組み合わせを判断できるか検証するため，既存の機械学習モデルによる表現再現性の評価性能を実験で確認した．

既存モデルとして，画像およびテキストの入力が可能な大規模視覚言語モデル (VLM) と比較的小規模な視覚と言語のマルチモーダル基盤モデルである CLIP [12] に，二層のパーセプトロンを結合した CLIP-MLP を用いた．実験は構築したベンチマークの評価データセット上で行った．VLM には GPT-4o-2024-11-20 [15]，Llama-3-EvoVLM-JP-v2 [16, 17]，llava-calm2-siglip [18]，および VILA-jp [19] を使用した[2)]．CLIP には clip-japanese-base [20] を使用した．なお，実験の詳細は付録 B に記載した．

大規模言語モデルは解答例を入力文に含めることで，性能が向上するという文脈内学習 (ICL) の効果が報告されている [5, 21]．そこで，VLM の実験では，各正解ラベルの数が等しくなる

1) 作業には FAST: Fast Annotation tool for SmarT devices[13] https://www.fast-annotation-tool.app/ を用いた．

2) 各モデルを公開する Hugging Face 上に記載された Chat Template を適用した．

表 3: ベンチマークにおける性能．下線は同一モデル内の最高スコア，**太字**は全モデル中の最高スコア．

Model	# of shots	コピーにおけるバナー画像の表現再現性評価タスク		バナー画像におけるコピーの表現再現性評価タスク	
		Macro F1 ↑	指示追従率 ↑	Macro F1 ↑	指示追従率 ↑
CLIP-MLP	Fine-tuned	**0.646**	**1.000**	**0.546**	**1.000**
GPT-4o	0	0.543	**1.000**	0.409	**1.000**
	6	0.323	**1.000**	0.498	**1.000**
	12	0.312	0.996	0.416	**1.000**
EvoVLM-JP-v2	0	0.307	**1.000**	0.355	**1.000**
	6	0.221	0.980	0.139	0.943
	12	0.008	0.005	0.000	0.000
calm2-siglip	0	0.254	**1.000**	0.319	**1.000**
	6	0.241	0.980	0.137	0.984
	12	0.231	0.975	0.144	0.978
VILA-jp	0	0.248	**1.000**	0.165	**1.000**
	6	0.000	0.000	0.000	0.000
	12	0.000	0.000	0.000	0.000

ように解答例を入力に追加した実験も行った．VLM の自動評価のため，選択肢に数値を割り当て，次のフォーマット指定文を入力文に付加した．

> [注意]
> 数字を答えてからその理由を答えてください。

分類性能の評価指標として正解候補の 3 クラスの F1 スコアから計算した **Macro-F1 スコア**を用いた．また，VLM が指示に従わず，選択肢を出力しないことがあった．そこで，全体の評価データセットのうち選択肢を出力した割合を**指示追従率**とし，指示追従性能の評価指標に用いた．

4 実験結果と考察

3 節の実験結果を表 3 に示す．コピーにおけるバナー画像の表現再現性評価タスク，バナー画像におけるコピーの表現再現性評価タスク共に，CLIP-MLP が最も高い F1 スコアを記録した．ファインチューニングにより，画像とコピーの表現の関係を VLM 以上に理解できるようになったと考えられる．このことは，表現再現性評価が画像と言語の汎用的なタスクではなく，広告ドメインに特化したタスクであることを示している．最も F1 スコアの高い CLIP-MLP であっても，コピーにおけるバナー画像の表現再現性評価タスクとバナー画像におけるコピーの表現再現性評価タスクにおける F1 スコアは，それぞれ，0.646 と 0.546 であり，中程度のスコアであった．このことは，広告ドメインにおける表現再現性評価タスクが機械学習モデルにとって難しいことを示唆する．

4.1 文脈内学習の実験結果と考察

ICL によって GPT-4o はバナー画像におけるコピーの表現再現性評価タスクの性能が向上した．特に，6-shot 時にモデル内最高の F1 スコアを記録した．一方で，コピーにおけるバナー画像の表現再現性評価タスクでは分類性能の向上が見られなかった．この結果から，画像はコピーより few-shot 事例を利用した表現要素の理解が難しく，ICL では画像の活用が困難であったことが示唆される．EvoVLM-JP-v2, calm2-siglip, VILA-jp では，ICL によって F1 スコア及び，指示追従率が低下した．これらのことから，ICL によって性能が改善されるかは，モデルとタスクに依存することが分かった．

5 おわりに

本研究では，バナー広告におけるバナー画像とコピーの組み合わせを，表現再現性に着目して定量的に評価する方法を提案した．提案手法によってバナー広告としての相応しさを判断できる傾向が確認された．また，機械学習モデルがバナー画像とコピーの相応しい組み合わせを理解するには，広告ドメインに特化したデータが必要であること，並びに現時点の機械学習モデルでは，提案手法を用いた評価を人間と同様に行うことは困難であることが分かった．具体的には，大規模視覚言語モデルを使用しても，CLIP を用いたモデルの性能を超えることはできず，CLIP 自体の性能も十分に高いとは言えなかった．今後は，機械学習モデルが正確に評価できなかったケースを分析し，機械学習モデルの評価性能を向上させるための研究を進める予定である．

謝辞

本研究にあたってご助言いただいた，株式会社サイバーエージェントの皆様に感謝いたします．

参考文献

[1] 株式会社 CARTA COMMUNICATIONS 株式会社 D2C 株式会社電通株式会社電通デジタル. 「2021 年日本の広告費 インターネット広告媒体費 詳細分析」 - news（ニュース） - 電通ウェブサイト. https://www.dentsu.co.jp/news/release/2022/0309-010503.html, 2022. 閲覧 2025 年 1 月 9 日.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752, 2022.

[3] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image diffusion models in generative AI: A survey. arXiv:2303.07909, 2024.

[4] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. arXiv:2203.02155, 2022.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. arXiv:2005.14165, 2020.

[6] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. arXiv:2303.18223, 2024.

[7] 近藤昌也, 丹治直人, 狩野芳伸. 効果の高い広告文生成のための LLM の instruction tuning と関連する広告属性の分析. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 2G5GS601–2G5GS601, 2024.

[8] 張培楠, 坂井優介, 三田雅人, 大内啓樹, 渡辺太郎. AdGLUE: 広告言語理解ベンチマーク. 言語処理学会第 29 回年次大会 発表論文集, pp. 1226–1231, 2023.

[9] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. arXiv:1707.03067, 2017.

[10] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. arXiv:2107.06912, 2021.

[11] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. arXiv:2404.07214, 2024.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. arXiv:2103.00020, 2021.

[13] Shunyo Kawamoto, Yu Sawai, Peinan Zhang, and Kohei Wakimoto. FAST: Fast annotation tool for smart devices, 2021. Open source software available from https://github.com/CyberAgent/fast-annotation-tool.

[14] Joseph Fleiss. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, Vol. 76, pp. 378–, 11 1971.

[15] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 閲覧 2025 年 1 月 9 日.

[16] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. arXiv:2403.13187, 2024.

[17] Inoue Yuichi, Akiba Takuya, and Makoto Shing. Llama-3-evovlm-jp-v2. https://huggingface.co/SakanaAI/Llama-3-EvoVLM-JP-v2.

[18] サイバーエージェント株式会社. 独自の日本語 llm「cyberagentlm2」に視覚を付与した vlm（大規模視覚言語モデル）を一般公開 ―商用利用可能な画像チャットモデルを提供―. https://www.cyberagent.co.jp/news/detail/id=30344, 2024. 閲覧 2025 年 1 月 9 日.

[19] Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. Constructing multimodal datasets from scratch for rapid development of a japanese visual language model. arXiv:2410.22736, 2024.

[20] Peifei Zhu Shuhei Nishimura Shuhei Yokoo, Shuntaro Okada and Naoki Takayama. CLIP japanese base. https://huggingface.co/line-corporation/clip-japanese-base.

[21] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. arXiv:2301.00234, 2024.

表 4: 同一素材に対して再現性のある素材と，再現性のない素材（または表現要素がない素材）を比較した際，バナー広告としてより相応しいと判断された素材の表現再現性ラベルの割合と調査件数．アノテータが判断に迷う場合があったため，割合の合計が 100%にならない．

ラベル	再現している素材と 再現していない素材の比較	再現している素材と 表現要素がない素材の比較
画像が再現しているコピー	73%	85%
画像が再現していないコピー	22%	-
表現要素がないコピー	-	15%
調査件数	104 件	60 件
コピーが再現している画像	80%	94%
コピーが再現していない画像	19%	-
表現要素がない画像	-	6%
調査件数	108 件	51 件

表 5: GPT-4o 及び calm2-siglip における，より多い Shot 設定の性能

Model	# of shots	コピーにおけるバナー画像の 表現再現性評価タスク		バナー画像におけるコピーの 表現再現性評価タスク	
		Macro F1 ↑	指示追従率 ↑	Macro F1 ↑	指示追従率 ↑
GPT-4o	24-shot	0.358	0.992	0.467	1.000
	36-shot	0.317	0.997	0.376	1.000
	72-shot	0.285	1.000	0.432	1.000
calm2-siglip	24-shot	0.238	0.973	0.172	0.302
	36-shot	0.238	0.948	0.105	0.185

表 6: ハイパーパラメータ一覧

GPT-4o

Parameter	Value
temperature	0.0
top_p	1.0
frequency_penalty	0.0
presence_penalty	0.0
seed	0
max_tokens	100
detail	low

EvoVLM-JP, calm2-siglip, VILA-jp

Parameter	Value
top_p	1.0
max_new_tokens	64
num_beams	1
do_sample	False
no_repeat_ngram_size	3

CLIP-MLP

Parameter	Value
学習エポック数	30
最適化アルゴリズム	Adam
活性化関数	ReLU
学習率	5×10^{-5}
損失関数	CrossEntropyLoss
損失関数のクラス重み	学習セット内の クラス出現割合の逆数

A few-shot 追加実験

本論における評価実験結果（表 3）によると GPT-4o と calm2-siglip は，shot 数の増加による指示追従率の低下が他のモデルと比べて小さかった．そこで，shot 数をさらに増加させた時に，性能向上が見られるか調べるため，shot 数を増やして評価実験を行った．表 5 に結果を示す．shot 数を増加させても，性能向上は見られなかった．GPT-4o はこの範囲の shot 数で，指示追従率の低下が見られず，複数枚の画像入力に対して，他の VLM よりも指示に従えることが分かった．一方で，calm2-siglip は shot 数を増加させた場合，コピーにおけるバナー画像の表現再現性評価タスクでは，指示追従率の低下が小さかったものの，バナー画像におけるコピーの表現再現性評価タスクでは，大幅に指示追従率が低下した．

B 既存手法の実験設定

CLIP-MLP の説明をする．CLIP でバナー画像とコピーから画像特徴量とテキスト特徴量を抽出，結合し，1024 次元の特徴量を作成する．この特徴量を二層の全結合層で 512 次元を経て 3 次元に出力し，分類を行った．学習時は学習データと検証データを用いて，追加した 2 層の全結合層をファインチューニングした．なお，検証データにおける Macro-F1 スコアが最良のモデルを記録した．

表 6 に大規模視覚言語モデルにおける推論ハイパーパラメータと CLIP-MLP の学習ハイパーパラメータを示す．