広告画像ランキングによる視覚言語モデルの評価

大竹 啓永 [1]　張 培楠 [2]　坂井 優介 [1]　三田 雅人 [2]　大内 啓樹 [1,2]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 株式会社サイバーエージェント
otake.hiroto.od2@naist.ac.jp
{sakai.yusuke.sr9, hiroki.ouchi, taro}@is.naist.jp
{zhang_peinan, mita_masato, ouchi_hiroki}@cyberagent.co.jp

概要

視覚言語モデルは高い画像認識能力を有しており、広告画像理解への応用に注目が集まっている。しかし、既存研究は広告画像単体を対象にしたものに限られている。そのため広告される対象である商材との関係を捉えた広告画像への視覚言語モデルの理解能力は明らかにすることで、自動生成された広告の推薦システム開発の支援に繋がる。そこで本研究では、モデルと人間との解釈の相関を測る評価用フレームワークを広告分野へ応用することでこの能力を明らかにする。実験結果より、視覚言語モデルは広告画像のランキング生成において関係性を十分に考慮できない場合が多かったが、一部のモデルで購買意欲を刺激するという観点を与えた場合に商材と広告画像の関係性を考慮し、ランキングに変化をもたらす可能性が示唆された。

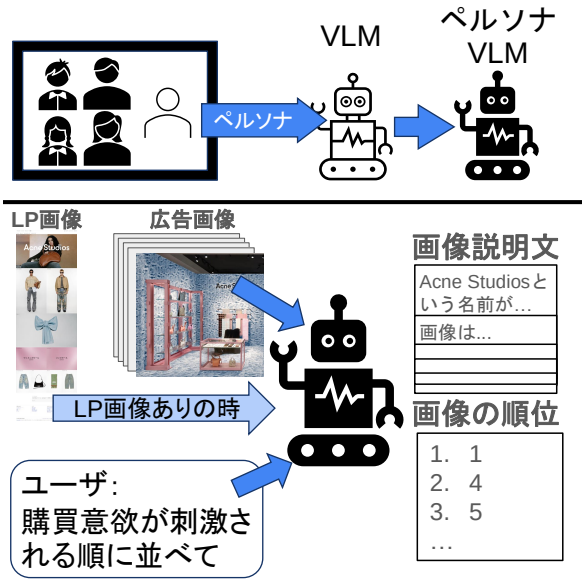

図 1: フレームワークの概要

1　はじめに

広告は人間の意思決定に対して影響を与えるものである。広告される対象である商材は地域サービスから日用雑貨まで多種多様である。広告によって、サービス利用者や消費者は行動を変化させられる。また、求人募集への申し込みなど具体的な行動がランディングページ (Landing Page, LP) と呼ばれるウェブサイト訪問者が最初に訪れるページで提供され、インターネット広告ではより直接的にユーザの行動を変化させている。この結果、サービス利用や定期購入、高額な買い物などと直接結びつき、広告業界は市場規模を拡大している [1]。

急速な発展に伴い、インターネット広告制作の自動化の需要が高まったことや事前学習済み言語モデルによる自然言語生成の技術革新 [2, 3, 4] により、自然言語処理を用いた広告生成・理解に関する研究が活発化している [5, 6, 7]。さらに、広告制作の自動化に伴い広告の評価手法やベンチマークが提案されてきた [8, 9]。

視覚情報はテキスト情報と比べ、より多くの情報を一度に伝達可能なため、短い時間で商材の情報を届けることができ、広告画像や広告動画はインターネット広告において重要である。そのため、機械学習による広告画像の生成 [10, 11]・理解 [12, 13, 14] についての研究が注目されている。さらに、視覚言語モデル (Vision Language Model, VLM) [15, 16, 17] は高い画像認識能力を持っており、VLM の広告画像への理解能力にも注目が集まってきている [18, 19]。しかし、既存研究は広告画像単体を対象にしたものに限られており、商材と広告画像の関係を捉えた評価については研究の余地が大いに残されている。また、LP などの商材情報に基づいて広告が制作されるため、広告画像の品質を測る上で両者の関係性を捉えることは重要である。さらに、VLM が両者の関係を捉えられることにより、自動生成された広告

への推薦システム開発の支援につながる。

本研究では、VLM における、広告画像と商材情報が載っている LP 画像との関係の理解度を調査する。調査手法として、Hayashi ら [20] が提案した Image Review Rank（IRR) というフレームワークを導入し、LP 画像の有無や異なる 2 種類の観点、人物像の設定を組み合わせた複数の条件を与えた VLM に広告画像のランキングを生成させ、条件ごとのランキングと人間のランキングとの相関を計算することで VLM を評価する。

実験の結果、多くの設定で VLM は LP 画像との関係を考慮した広告画像のランキングを生成できなかった一方で、購買意欲を刺激する観点を与えた場合に実験で使用した VLM のうちの 1 つが広告画像のランキングを変化したと示唆する結果が得られた。本研究を通じて、VLM の広告画像と商材との関係に対する VLM の理解能力を測定するためには、より適切な問題設定が必要であると示された。

2 関連研究

広告画像の理解 広告の自動評価を行うために、広告画像・広告動画のデータセットを作成し、広告が伝えるメッセージを予測するタスクが提案され、学習により理解能力が向上することが示された [12, 13]。また、広告画像中の物体が暗示していることを学習することでそれまでのベースラインと比べ広告画像の理解能力が向上することも示された [14]。さらに、既存の広告理解タスク [12] を用いて、高い画像理解能力を持つ VLM に対する広告画像の理解能力が調査され、高い理解能力が示された [18]。一方で新たに提案した評価タスクにより VLM が複雑な広告画像を理解することは困難であり従来の広告理解タスクの問題点が指摘されている [19]。しかし、既存研究は広告画像単体を対象とするものに限られている。ゆえに本研究では商材と広告画像との関係の理解度に注目し、広告画像を評価する。

IRR 人間の批評文への評価とモデルの出力との相関を測ることで VLM の性能を評価する新たなフレームワークを提案し、VLM に対する既存の評価手法の限界を指摘している [20]。本研究ではこの研究の流れを汲み、新たな VLM の評価手法を模索する。

3 調査手法

図 1 は本研究で用いたフレームワークの流れを示している。IRR では、一枚の画像に対して、複数の観点に基づく批評文を 5 つ生成させ、5 つの批評文に対してランキングを行なっている。本研究では IRR に変更を加え、VLM と人手による広告画像のランキングと、配信実績に基づくランキングの関係を測るためのフレームワークを構築する。

フレームワークの詳細 VLM に対し、年齢や性別が異なる 5 つの人物像、2 種類の評価観点、LP 画像の有無という 3 つの設定を組み合わせた計 20 通りの条件を提示し、広告画像の説明文生成およびランキングを実施する。具体的な人物像の設定に関しては、20-34 歳の女性 (**F1**)、35-49 歳の女性 (**F2**)、20-34 歳の男性 (**M1**)、35-49 歳の男性 (**M2**) とペルソナを与えない (**NO**) である。また、評価観点は、購買意欲を刺激するか、興味を惹かれるかの 2 種類である。

VLM に条件を与えた後、説明文の生成を行う。具体的には、1 つの LP 画像をもとに作成された 5 つの広告画像を対象とし、条件が与えられた VLM にこれらの広告画像に対する説明文を生成させる。その後、生成された説明文をもとに、5 つの広告画像を購買意欲を刺激される順、または興味を惹かれる順にランキングを生成させる。

4 実験設定

VLM に与える条件の違いによる評価の変化を調査し、VLM と人間による評価の違いを比較する実験を行なった。具体的には、3 節で構築したフレームワークを用いて、複数の VLM に対して各条件ごとの広告画像のランキングを生成させた。その後、VLM によるランキング、人手によるランキング、および配信実績に基づくランキングとの相関を測定した。

4.1 モデル

本研究では、GPT-4o-mini[15]、Gemini-1.5-flash[16]、Pixtral Large[17] の 3 つの VLM を選択し使用した。

VLM への指示 本実験では、広告画像を対象とした以下の 2 つのタスクを設定し、それぞれに対応する指示を作成した。全ての指示文は英語で作成し、実験で一貫性を保つようにした。ただし、今回の実験ではタスクごとに出力をさせるのではなく

VLM には 1 度で 2 つのタスクに返答させた。

1 つ目のタスクは、広告画像の説明生成である。このタスクでは、「広告画像間の違いがわかるように (**興味プロンプト**)」と、「購買意欲に基づいて (**購買意欲プロンプト**)」の 2 種類の文言からなるプロンプトを用意した。

2 つ目のタスクは、広告画像のランキング生成である。このタスクでは、特定の性別や年齢（例:「35 から 49 才の男性にとって」）を反映したペルソナに基づくランキングを生成するよう指示した。広告画像がペルソナ（特定のユーザー層）に基づくランキングがされるように「35 から 49 才の男性にとって」というような特定の性別・年齢の人物像を与える文言を指示に加えた。また、LP 画像がある場合には、LP 画像の内容に矛盾しない広告画像が上位にランク付けされるよう、「LP 画像に広告画像が矛盾していないかに注意してください」という文言を追加し、内容の整合性を確保した。さらに、**購買意欲プロンプト**では「購買意欲を刺激するという観点から」、**興味プロンプト**では「興味を惹かれるという観点から」という指示を加え、それぞれの観点に基づいてランキングを生成するように設計した (付録 A 参照)。

4.2 データの前処理

データ収集 本研究では、配信実績値を確認できる 43,412 件の LP 画像と、それぞれの LP 画像に対応する 217,060 件の広告画像を収集した。

フィルタリング 各 LP 画像に対応する 5 枚の広告画像について、配信実績値を用いて分散を算出し、すべての LP 画像に対して広告画像の分散を評価した。その結果、分散に有意な差が見られない組を除外した。さらに、VLM が処理できないサイズの LP 画像も対象から除外した。また、広告画像 5 枚の中で類似度が高いものや明確な違いが認められないものについても除外を行った。加えて、LP 画像が正しく読み込めていない場合や、アンケートやアプリストアに関連する LP 画像も分析対象から除外した。

これらのフィルタリングを経て、最終的に人手による評価が可能だった 22 件の LP 画像と 110 件の広告画像を本実験で使用した。また、Pixtral Large[17] が入力可能な画像サイズに制限があったため、この制限を超える画像に対してはアスペクト比を維持したまま画像を縮小する処理を実施した。

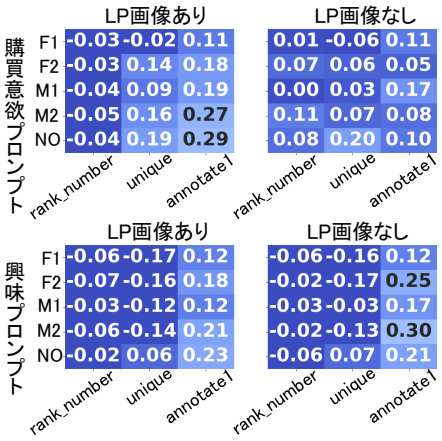

図 2: Gemini-1.5-flash での相関結果

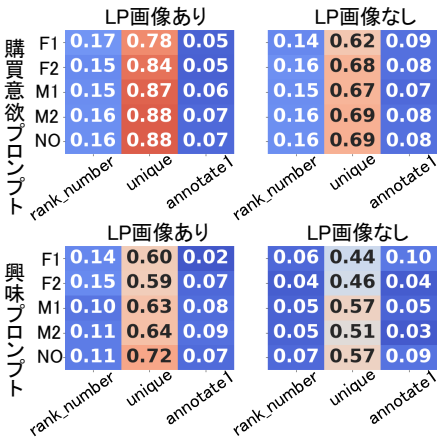

図 3: GPT-4o-mini での相関の結果

4.3 人手によるランキング

本実験における人手によるランキングでは、広告画像を購買意欲が高まる順に並べるように指示を与えた。この評価は以下の 2 段階で実施された。最初の段階では、評価者に LP 画像を提示せず、広告画像のみをもとにランキングを行った。次に第二段階では、LP 画像を提示し、広告画像の内容が LP 画像と整合しているかを考慮したうえで再度ランキングを実施した。この手順により、LP 画像の有無が広告画像のランキングに与える影響を評価することを目的とした。

5 結果

VLM (**F1**、**F2**、**M1**、**M2**、**NO**)・人手 (**annotate1**)・配信実績値 (**rank_number**) によるランキングと画像の入れた順番 (**unique**) に対して、ケンドールの順位相関を適用した。図 2 に示すように Gemini-1.5-flash は多くの条件で **annotate1** との相関が最も高い結果を示したが、この傾向は一貫していないことがわか

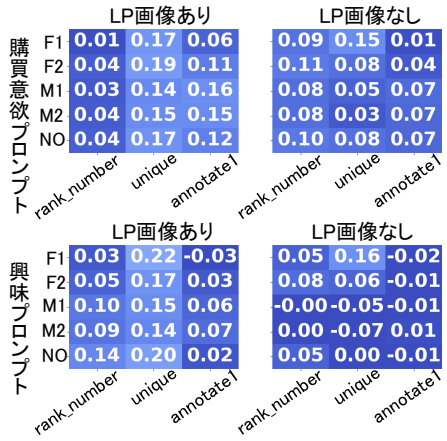

図 4: Pixtral での相関結果

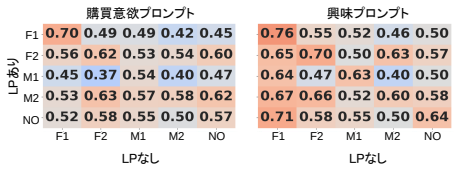

図 5: Gemini-1.5-flash での LP の有無間の相関結果

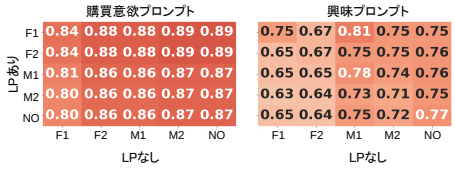

図 6: GPT4o-mini での LP の有無間の相関結果

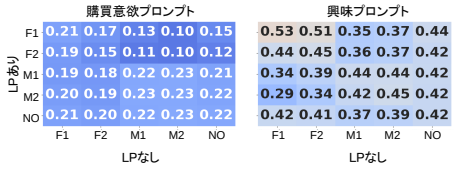

図 7: pixtral での LP の有無間の相関結果

る。一方、図 3 では、GPT-4o-mini が **unique** との相関が、**annotate1** 列と **rank_number** 列とに比べ、顕著に高い値を示し、この傾向は一貫している。この結果は、複数の画像を一度に入れることにより、個々の画像の比較が困難になったことに起因すると考えられる。

また、図 4 では、LP 画像あり・なしの条件を比較すると、どちらの観点でも LP 画像ありの場合の方が **unique** との相関がやや高くなっていることが確認された。さらに、図 8 を見ると、**興味プロンプト**よりも**購買意欲プロンプト**の場合において、LP 画像あり・なし間の相関が低いことがわかる。同時に、図 7 では、LP 画像ありの場合の方が、LP 画像なしと比較して両プロンプト間の相関が高いことが示されている。これらのことから、**興味プロンプト**では LP 画像の有無による評価の変化がほとんど

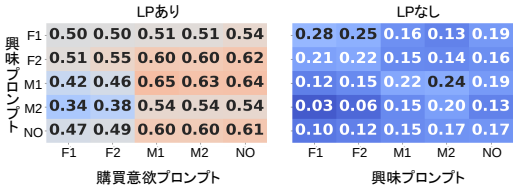

図 8: pixtral での観点間の相関結果

見られない一方、**購買意欲プロンプト**では LP 画像の有無によって Pixtral Large が広告画像のランキングを変化させたことが示唆される。

一方、図 5 および図 6 では、Geimini-1.5-flash や GPT-4o-mini について、Pixtral Large とは異なり、LP 画像の有無間の相関において観点の違いによる大きな差は見られなかった。これにより、LP 画像の有無によるランキングの変化は Pixtral Large 特有の現象であることが示唆される。

全体として、大半の条件で VLM は LP 画像の有無による広告画像に対するランキングの変化を適切に反映できなかった。しかし、購買意欲という特定の観点を与えた場合、1 つの VLM においては LP 画像の有無に応じた広告画像の評価を変化させる可能性が示唆された。このことから、特定の条件下における LP 画像の有無による広告画像のランキングの変化が VLM の一般的な特性であるかを明らかにするため、さらなる調査が必要である。加えて、本研究では複数の画像を一度に入力すること、個々の画像の比較が難しくなる問題が発見された。この課題を解決するためには、広告画像一枚ごとに説明文を生成させた上で、その説明文に基づいてランキングを行う手法が求められる。さらに、生成された説明文に対する分析として、広告画像に対する説明の十分性の評価や条件ごとの説明文の比較を行うことも重要である。これらの課題は、今後の研究で取り組むべき課題といえる。

6 おわりに

本稿では、VLM に複数の設定を与えた上で広告画像の説明文・ランキングを生成させた。実験結果から既存の VLM が複数の画像の関係を捉え、画像の評価を行うことは難しいとわかった。しかし、特定の VLM・条件下では振る舞いの変化が示唆された。また、構築したフレームワークの課題が見つかった。さらに、生成された説明文への分析や VLM への調査が必要とわかった。これらを今後の改善に繋げていきたい。

参考文献

[1] 総務省. 情報通信白書. 日経印刷 and 全国官報販売協同組合 (発売), 2024.

[2] A Vaswani. Attention is all you need. **Advances in Neural Information Processing Systems**, 2017.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[5] Hidetaka Kamigaito, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. An empirical study of generating texts for search engine advertising. In Young-bum Kim, Yunyao Li, and Owen Rambow, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers**, pp. 255–262, Online, June 2021. Association for Computational Linguistics.

[6] Konstantin Golobokov, Junyi Chai, Victor Ye Dong, Mandy Gu, Bingyu Chi, Jie Cao, Yulan Yan, and Yi Liu. Deepgen: Diverse search ad generation and real-time customization, 2022.

[7] Go Inoue, Akihiko Kato, Masato Mita, Ukyo Honda, and Peinan Zhang. CAMERA^3: An evaluation dataset for controllable ad text generation in Japanese. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 2702–2707, Torino, Italia, May 2024. ELRA and ICCL.

[8] Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. Learning to create better ads: Generation and ranking approaches for ad creative refinement. In **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**, CIKM '20, p. 2653–2660, New York, NY, USA, 2020. Association for Computing Machinery.

[9] Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. Adtec: A unified benchmark for evaluating text quality in search engine advertising, 2024.

[10] Zhenbang Du, Wei Feng, Haohan Wang, Yaoyu Li, Jingsen Wang, Jian Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junsheng Jin, Junjie Shen, Zhangang Lin, and Jingping Shao. Towards reliable advertising image generation using human feedback. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, **Computer Vision – ECCV 2024**, pp. 399–415, Cham, 2025. Springer Nature Switzerland.

[11] Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, and C. L. Philip Chen. Desigen: A pipeline for controllable design template generation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 12721–12732, June 2024.

[12] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 1100–1110, 2017.

[13] Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. Interpreting the rhetoric of visual advertisements. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 43, No. 4, pp. 1308–1323, 2021.

[14] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In **Proceedings of the European Conference on Computer Vision (ECCV)**, September 2018.

[15] OpenAI, :, Aaron Hurst, et al. Gpt-4o system card, 2024.

[16] Gemini Team, Petko Georgiev, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[17] Mistral AI team. Pixtral large: Frontier-class multimodal model. **Mistral AI News**, November 2024. Accessed: 2025-01-07.

[18] Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. KAFA: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)**, pp. 772–785, Toronto, Canada, July 2023. Association for Computational Linguistics.

[19] Anna Bavaresco, Alberto Testoni, and Raquel Fernández. Don't buy it! reassessing the ad understanding abilities of contrastive multimodal models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 870–879, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[20] Kazuki Hayashi, Kazuma Onishi, Toma Suzuki, Yusuke Ide, Seiji Gobara, Shigeki Saito, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Irr: Image review ranking framework for evaluating vision-language models, 2024.

A　VLM に与えたプロンプトの例

VLM に対して、異なる観点に基づく返答を得るために観点別の 2 種類のプロンプトを設計した。以下に**興味プロンプト、購買意欲プロンプト**の具体例を示す。2 種類のプロンプト間で異なる点は太字となっている。

興味プロンプトの場合の例

Describe the five Images in detail so that we can understand the difference each.
The output must strictly follow the JSON format. Example output JSON format:{"Image1": <_>,"Image2": <_>,"Image3": <_>,"Image4": <_>,"Image5": <_>}
Constraints:
- The output must be in JSON format.
- The JSON object should contain exactly five keys: "Image1," "Image2," "Image3," "Image4," and "Image5."
- Each key's value must be a specific and clear description of the corresponding image.
- Descriptions should provide details such as the main subject, colors, setting, and any notable features.

Next task:
Please rank by which advertisemental images are interesting for 35-49 old man.
Note that the ad image does not contradict this LP image.The output must strictly follow the JSON format. Example output format:{"Image1": 1,"Image2": 4,"Image3": 3,"Image4": 5,"Image5": 2}
Constraints:
- The output must be in JSON format.
- The JSON object should contain exactly five keys: "Image1," "Image2," "Image3," "Image4," and "Image5."
- Each key's value must be a numerical ranking between 1 and 5.
- The rankings should be unique, with no two images sharing the same rank.
- A rank of 1 represents the highest rank, and a rank of 5 represents the lowest rank.

購買意欲プロンプトの場合の例

Describe in detail the five advertising images based on the perspective that they stimulate the desire to buy.
The output must strictly follow the JSON format. Example output JSON format:{"Image1": <_>,"Image2": <_>,"Image3": <_>,"Image4": <_>,"Image5": <_>}
Constraints:
- The output must be in JSON format.
- The JSON object should contain exactly five keys: "Image1," "Image2," "Image3," "Image4," and "Image5."
- Each key's value must be a specific and clear description of the corresponding image.
- Descriptions should provide details such as the main subject, colors, setting, and any notable features.

Next task:
Please rank which ad images stimulate your purchasing decisions for 35-49 old man.
Note that the ad image does not contradict this LP image.The output must strictly follow the JSON format. Example output format:{"Image1": 1,"Image2": 4,"Image3": 3,"Image4": 5,"Image5": 2}
Constraints:
- The output must be in JSON format.
- The JSON object should contain exactly five keys: "Image1," "Image2," "Image3," "Image4," and "Image5."
- Each key's value must be a numerical ranking between 1 and 5.
- The rankings should be unique, with no two images sharing the same rank.
- A rank of 1 represents the highest rank, and a rank of 5 represents the lowest rank.