画像言語モデルにおけるハルシネーションの発生と Visual Attention 精度の関係の調査

富田 雅代 [1] 林 克彦 [2] 金子 知適 [2]
[1] 東京大学教養学部 [2] 東京大学総合文化研究科
mtomita143@gmail.com katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp
kaneko@graco.c.u-tokyo.ac.jp

概要

画像言語モデルは、与えられた画像と矛盾する出力を生成することが報告されている。この現象はハルシネーションと呼ばれ、画像言語モデルの信用を損ない、普及を阻害する要因となっている。しかし、ハルシネーションの発生と、画像言語モデルがテキスト生成時に画像の適切な領域へ注目しているかどうかは関連すると予想されるが、データでは示されていない。本研究では、ハルシネーションの有無に関する評価指標である POPE を用い、画像内に実際に写っている物体と写っていない物体を対象として、ハルシネーションの発生と画像内の注目領域との関係を検証する。

1 序論

CLIP [1] や LLaVA [2] に代表される画像言語モデル（VLM; Vision-language Model）は急速に発展し、注目を集めるようになった。画像言語モデルとは、画像エンコーダとテキストエンコーダを備え、画像情報とテキスト情報を同時に処理できるモデルである。これらのエンコーダによって、画像質問応答 [3] や画像説明文生成 [4]、画像検索 [5] などのマルチモーダルタスクが実現され、自動運転 [6] や医療診断 [7] といった他分野へも応用が広がっている。しかし、応用範囲の拡大とともに、ハルシネーションの問題がいっそう顕在化してきた。

ハルシネーションとは、入力画像と出力テキストの間に生じる矛盾である [8]。ハルシネーションの発生の有無を評価する手法として POPE [9] が提案され、広く用いられている。POPE とは、画像に含まれる、または含まれない物体について質問し、回答が正しいかを検証する手法である [9]。

Prompting はハルシネーションを低減する手段の一つとして提案された [10]。とりわけ、画像に赤い丸 [11] や矢印 [12] などの印を付して対象物を強調する Visual Prompting は、ハルシネーションの抑制に寄与することが知られている [13]。なかでも、Attention Prompting on Image（API）と呼ばれる Visual Prompting が提案され、画像に存在する物体に対する POPE において、正答率の改善を示した [14]。API Prompting は、画像言語モデル内部の Visual Attention を画像に重ね合わせる手法であり、従来の手作業による Visual Attention とは一線を画している。しかし、先行研究 [14] では、実験は画像内に存在する物体のみを対象としており、存在しない物体を誤って「存在する」と判断する、ハルシネーションの重要な側面は検証されていない。また Visual Attention に依存する手法であるが、Visual Attention の精度に関係あると予想される Visual Attention を抽出する際の画像言語モデルの出力と Visual Attention の精度との関係が十分に検証されていない。

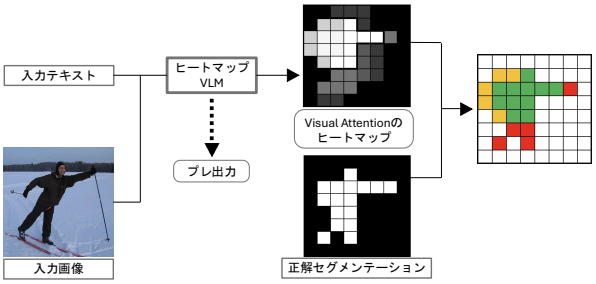

図 1 Visual Attention の精度: 入力画像とテキストから画像言語モデルによって生成された Visual Attention のヒートマップは、正解セグメンテーションと比較され、MSE、Precision、Recall、IoU が計算されて精度が評価された。

本研究の主な貢献は以下である。

1. 画像内に含まれる物体と含まれない物体において、API の性能を評価する。
2. VLM の出力の正しさと Visual Attention の精度に関係があるか検証する。

2 方法

2.1 API Prompting

API Prompting は画像言語モデルから取得される Visual Attention のヒートマップを用いて、画像の重要な物体を強調する Visual Prompting の手法の一つである [14]。画像言語モデルから各画像パッチがどの程度出力に寄与するかを表す Attribution Map Ψ を取得し、畳み込みし、画像サイズにリサイズして、Visual Attention のヒートマップを得る。取得したヒートマップを元の画像に重ねる。先行研究 [14] に倣い、Vision Transformer [15] ベースの CLIP と LLaVA の 2 種類を用い、以下それぞれ CLIP API Prompting、LLaVA API Prompting と呼ぶ。

CLIP の Visual Attribution Map CLIP はテキスト表現 $\hat{T}$ と画像表現 $\hat{I}$ の類似度を計算するモデルであるが、Attribution Map Ψ はその類似度 $sim(\hat{I},\hat{T})$ を分解することによって求められる。

CLIP の画像表現 $\hat{I}$ は最初の層のクラストークン、各層の Multihead Self-Attention（MSA）と Multi-Layer Perceptron （MLP）の出力値を線形結合したものとして表現できるが、先行研究 [16] より MSA の後半の層が画像表現に大きな影響を与えるため MSA の項以外無視してよく、類似度は

$$\mathrm{sim}(\hat{I},\hat{T}) \approx \mathrm{sim}\left(\sum_{l=L'}^{L} \mathscr{L}(\mathrm{MSA}^l([Z^{l-1}]))_{\mathrm{cls}}, \hat{T}\right).$$

と表せる。ただし Z^l は l 番目の Transformer 層への入力トークンを示し、L' は Attirubution 取得を開始する層の id を表す。$\mathscr{L}$ は、全結合層と正規化処理を含む線形変換を示す。

これは Attribution Map Ψ^{cls} と捉えて良いが、しばしば単色領域など重要な情報を持たない領域と $\hat{T}$ の相関が大きいため [17]、無視すべき領域と識別するため、類似度スコアの反転スコアとして計算される補完的な Attribution Map Ψ^{comp} を考える。

$$\Psi^{\mathrm{comp}}_{i,j} \triangleq 1-\mathrm{sim}(\mathscr{L}(Z^L_t),\hat{T}), \quad \text{where} \quad t = 1+j+P\cdot(i-1).$$

ただし、$\Psi^{\mathrm{comp}}_{i,j}$ は Ψ^{comp} の成分であり、Z^L_t は最後の Transformer 層からの t 番目の出力トークンを指し、P は画像の一辺あたりの分割数を表す。

この二つの Attribution を合わせて、次のように CLIP の Attribution Map Ψ が定義される。

$$\Psi = \Psi^{\mathrm{cls}} + \Psi^{\mathrm{comp}} - \Psi^{\mathrm{comp}} \odot \Psi^{\mathrm{cls}}.$$

LLaVA の Visual Attribution Map LLaVA は入力画像の理解に MSA を用いる画像言語モデルである。出力テキストの各トークンと画像トークン間の Visual Attention の重みを用いて、対応する画像のパッチへの出力の寄与度 Attribution Map Ψ を決定する。Attribution Map Ψ は、生成されたテキスト全体およびすべての Attention Head を平均化して算出され、以下のように定義できる。

$$\Psi_{i,j} \triangleq \frac{1}{MH}\sum_{m=1}^{M}\sum_{h=1}^{H} A^{(\bar{L},h)}_{m,t}, \quad \text{where} \quad t = j + P\cdot(i-1).$$

ただし、$\Psi_{i,j}$ は Ψ の成分であり、M は出力トークンの長さ、H は Attention Head の数、P は画像の一辺あたりの分割数を表し、$A^{(\bar{L},h)}$ は $\bar{L}$ 番目の層、h 番目の Attention Head での出力テキストと画像トークン間の Cross Attention を表す。

2.2 Visual Attention の精度評価

画像に存在するオブジェクトについて、対象物体のセグメンテーションデータと Visual Attention のヒートマップとの Precision、Recall、Intersection over Union （IoU）、平均二乗誤差 （MSE）を計算する。

セグメンテーションデータは 0 と 1 で構成される二値配列であり、Visual Attention のヒートマップは 0 から 255 の整数値をとるため、ヒートマップを 255 で割って平均二乗誤差を計算する。

Precision、Recall、IoU については、Visual Attention のヒートマップの平均値を閾値として、それ以上であれば 1、未満であれば 0 として二値配列に変換し、True Positive（図 1 の緑の部分、ヒートマップで 1 かつ正解セグメンテーションで 1）、False Positive （図 1 の黄色の部分、ヒートマップで 1 かつ正解セグメンテーションで 0）と False Negative （図 1 の赤の部分、ヒートマップで 0 かつ正解セグメンテーションで 1）を求め、Precision、Recall、IoU を計算する。

2.3 プレ出力の取得

ヒートマップ生成時の出力を同一の画像言語モデル （ヒートマップ VLM）から取得し、プレ出力として評価した。

CLIP においては、[‘photo of <object>s’, ‘photo of no <object>s’] という分類タスクに対する確率を算出する。‘photo of <object>s’ の確率が 0.5 以上であれば、出力を「Yes」と分類する。

LLaVA は Visual Attention のヒートマップを作成する際の出力をプレ出力と扱う。

VLM	Prompting の有無	ヒートマップ VLM	プレ出力	出力の評価 [%] Acc.	Pre.	Rec.	TNR	F1
LLaVA	なし	-	-	86.23	84.21	89.19	83.27	86.63
	あり	CLIP	-	**86.52**	**84.78**	89.02	**84.02**	**86.85**
			正答 （65%）	▲89.40	▲86.17	▲93.87	▲84.93	▲89.85
			誤答 （35%）	▼82.54	▼81.42	▼84.32	▼80.76	▼82.85
	あり	LLaVA	-	86.11	**84.72**	88.12	**84.10**	86.39
			正答 （92%）	▲89.63	▲88.24	▲91.45	▲87.81	▲89.82
			誤答 （8%）	▼43.06	▼43.61	▼47.44	▼38.67	▼45.45

表 1 MSCOCO データセットにおいて POPE を適用し、Prompting なし、CLIP API Prompting、LLaVA API Prompting 画像での LLaVA の出力の、Accuracy、Precision、Recall、True Negative Rate （TNR）、F1 スコアを示す。CLIP および LLaVA API Prompting では、データ全体に加えて、Visual Attention を生成する際の出力であるプレ出力が正答、誤答に限定した結果も示す。Prompting なしと比べて改善した指標は太字で表され、各ヒートマップ VLM において、全体の結果と正答、誤答の場合の結果を比べた差を上三角または下三角を用いて記した。

プレ出力 \ ヒートマップ VLM	Visual Attention の精度 [%] CLIP Pre.	Rec.	IoU	MSE	LLaVA Pre.	Rec.	IoU	MSE
正答	**15.88**	**83.23**	**13.64**	**27.06**	**10.55**	**65.56**	**8.03**	13.51
誤答	9.63	83.13	8.30	31.28	4.57	61.16	3.30	**11.68**

表 2 画像に含まれる物体の POPE において、Visual Attention のヒートマップと正解のセグメンテーションデータを比較した結果を Visual Attention を生成する際に CLIP、LLaVA が正答、誤答を返した場合ごとに示す。Precision、Recall、IoU は、ヒートマップの平均値を閾値として算出した。各カテゴリーの最良結果を太字で表し、5%以上の差を緑色で示している。

3 実験

実験設定 MSCOCO データセット [18] から 3,860 枚の画像を用い、CLIP と LLaVA から Visual Attention のヒートマップを作成し、元の画像に黒いマスクとして混ぜ、加工済み画像と、POPE の質問を LLaVA に入力して出力を評価した。この時、Visual Attention の精度も評価した。

画像に含まれる物体は MSCOCO ラベルから取得し、残りの 80 ラベルからランダムに 3 つの含まれない物体を選び、含まれる物体のセグメンテーションデータを正解セグメンテーションデータとして取得した。API Prompting の実験方法は先行研究 [14] に概ね準拠したが、先行研究論文中ではアルファチャネルを用いて混ぜるとしているが、本実験では黒いマスクとして混ぜた。CLIP API Prompting には事前学習済みの CLIP ViT-L/14@336px を、LLaVA API Prompting には事前学習済みの llava-v1.5-7b を使用し、それぞれ、22 層以降の層、20 層から Visual Attention のヒートマップを作成した。その時、kernel サイズを 3 として畳み込み、LANCZOS 法を用いてリサイズした。POPE の質問への回答は、事前学習済みの llava-hf/llava-1.5-7b-hf (vLLM バージョン) から取得し、温度を `0.8`、top_p を `0.9` に設定した。「Answer Yes, No, or Not Sure」を文末に付加し、回答を先頭トークンで分類した。回答が「Yes」「No」「Not Sure」以外で始まる場合は除外した。

実験結果 表 1 は、POPE における Prompting なしと API Prompting ありの結果の比較を示す。Prompting ありでは、ヒートマップ VLM ごとに、プレ出力が正答、誤答である場合に限定した結果も示した。Prompting ありでは、なしと比べ、Precision と TNR が改善した。CLIP では 21,651 件中 14,015 件がプレ出力が正答で、LLaVA では 20,016 件が正答であった。プレ出力が正答である時には、POPE の各指標が Prompting ありの正答、誤答を分けない場合と比べ改善された。誤答時では、ヒートマップ VLM が LLaVA である場合、特に悪化が顕著であった。

表 2 は、画像に含まれる物体の POPE ついて、プレ出力が正答、誤答である場合に分けてに Visual Prompting の精度を比較した結果である。CLIP では 10,071 件中 4,956 件プレ出力が正答で、LLaVA では 9,310 件が正答であった。LLaVA の Visual Attention の MSE を除き、プレ出力が正答である時、どの指標も改善された。特に Precision、IoU が正答の場合と誤答の場合で差が大きかった。

4 考察

API Prompting の効果 表 1 から、Visual Attention を用いて画像の重要箇所を強調することで、画像に写っていない物体を正しく「写っていない」と判断する能力が示され、API Prompting ハルシネーションを減少させる効果があることが示された。

また、Visual Attention を生成するときの画像言語モデルの出力（プレ出力）の正答、誤答の場合を比べると、プレ出力が正答である場合は POPE の質問への回答の正答率が高かった。これは特に、ヒートマップを作成する VLM と最終的に POPE に回答する VLM が同一であり、プレ出力と最終的な回答で同じ問題を同様に間違えやすい LLaVA + LLaVA API Prompting でより顕著であったが、異なる VLM を用いる、LLaVA + CLIP API Prompting の場合であっても、同様の傾向が確認された。表 2 の結果と合わせて、プレ出力が正答であり、Visual Attention の精度が高い時、プレ出力が誤答であり、Visual Attention の精度が低くい時と比べ、POPE の質問へより正しく回答できることを示している。プレ出力が誤答である時、Visual Attention の精度が低く、POPE へ正しく回答するようになるとは限らないため、API Prompting の使用には注意が必要である。

Visual Attention の精度 表 2 から分かるように、画像言語モデルが画像に含まれる物体を誤って「含まれない」と判断した場合、Visual Attention が生成される際にその精度が低下し、特に Precision や IoU に顕著な差が見られた。これは、画像言語モデルの出力が誤っている場合、対象物体を見落としている可能性が高いことを示唆している。

Precision はどれだけ画像言語モデルが正しい対象物に集中し、誤認識を最小限に抑えられるかを測定する指標である。低い Precision は、画像言語モデルが関連する領域を正しく識別せず、関連しない領域に焦点を合わせていることを示す。IoU は、Visual Attention がどれだけ正解セグメンテーションと一致しているかを測定している。IoU が低い場合、正解セグメンテーションと違う場所を見ていることが示唆される。以上より、ハルシネーションが起きているとき、画像言語モデルが対象物体とは違う箇所に注意を向けている傾向があることが示唆される。

Precision や IoU が VLM の出力の正しさと関連がある一方、Recall はあまり関係がないと示唆された。Recall は物体のセグメンテーション領域のうち、どの程度注意が向けられたかを表している。物体によってどの程度見ればその物体か判断できる領域は大きく異なると推定でき、あまりハルシネーションに寄与しなかったと考えられる。

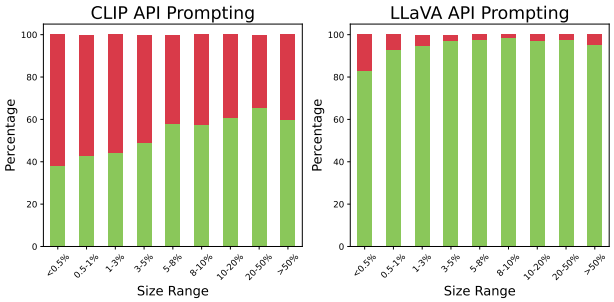

図 2 対象物体のサイズと Visual Attention 生成時の画像言語モデルの出力の関係 緑は正答した割合を示し、赤は誤答した割合を示している。

対象物体のサイズは図 2 に示されるようにハルシネーションの起こりやすさにも影響するとともに、Precision、Recall にも大きく影響する。対象物体が小さい時、Precision は低くなりやすく、反対に Recall は高くなりやすい。そのため、Precision とハルシネーションの関係は過大評価されている可能性があることに留意する必要がある。

課題 本研究の制約として、対象とした画像言語モデルを LLaVA のみに限定しているため、他の画像言語モデルに対して結果の一般化が困難である点が挙げられる。また、画像に含まれない物体に対する Visual Attention の精度を評価していないため、この側面についてはさらなる分析が求められる。

5 結論

本研究では、MSCOCO データセット上で、画像に写る物体と写らない物体を対象とした POPE に対して、API Prompting での性能向上の有無と CLIP と LLaVA を用いて生成された Visual Attention を評価した。主な結論は次の 2 点である。

1. API Prompting は画像に写っていない物体にも正しく「ない」と判定できる能力を示す。
2. 画像言語モデルが画像に写っている物体を見落とす時、画像言語モデルが対象物体とは違う箇所に注意を向けている傾向がある。

以上の結果から、画像の重要領域を強調することはハルシネーションに効果があり、画像言語モデルの出力と Visual Attention の精度は関係があることが示唆される。今後は、Visual Attention の領域の精度をいかに高めるかが重要な課題となるだろう。

参考文献

[1] 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. In **Proc. of ICML**, Vol. 139, pp. 8748–8763, 2021.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **Proc. of NeurIPS**, 2023.

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proc. of IEEE/CVF**, pp. 26286–26296, 2024.

[4] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. In **Proc. of NeurIPS**, 2023.

[5] Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. Enhancing interactive image retrieval with query rewriting using large language models and vision language models. In **Proc. of ICMR**, pp. 978–987, 2024.

[6] Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C. Knoll. Vision language models in autonomous driving and intelligent transportation systems. **CoRR**, Vol. abs/2310.14414, , 2023.

[7] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. **CoRR**, Vol. abs/2403.02469, , 2024.

[8] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, and Others. A survey on hallucination in large vision-language models. **CoRR**, Vol. abs/2402.00253, , 2024.

[9] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In **Proc. of EMNLP**, pp. 292–305, 2023.

[10] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip H. S. Torr. A systematic survey of prompt engineering on vision-language foundation models. **CoRR**, Vol. abs/2307.12980, , 2023.

[11] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for vlms. In **Proc. of IEEE/CVF**, pp. 11953–11963, 2023.

[12] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In **Proc. of IEEE/CVF**, pp. 12914–12923, 2024.

[13] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Ruiyi Zhang, Subrata Mitra, Dimitris N. Metaxas, Lina Yao, Jingbo Shang, and Julian J. McAuley. Visual prompting in multimodal large language models: A survey. **CoRR**, Vol. abs/2409.15310, , 2024.

[14] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In **Proc. of ECCV**, Vol. 15088, pp. 251–268, 2024.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **Proc. of ICLR**, 2021.

[16] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In **Proc. of ICLR**, 2024.

[17] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In **Proc. of ICLR**, 2024.

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In **Proc. of ECCV**, Vol. 8693, pp. 740–755, 2014.

A　参考画像

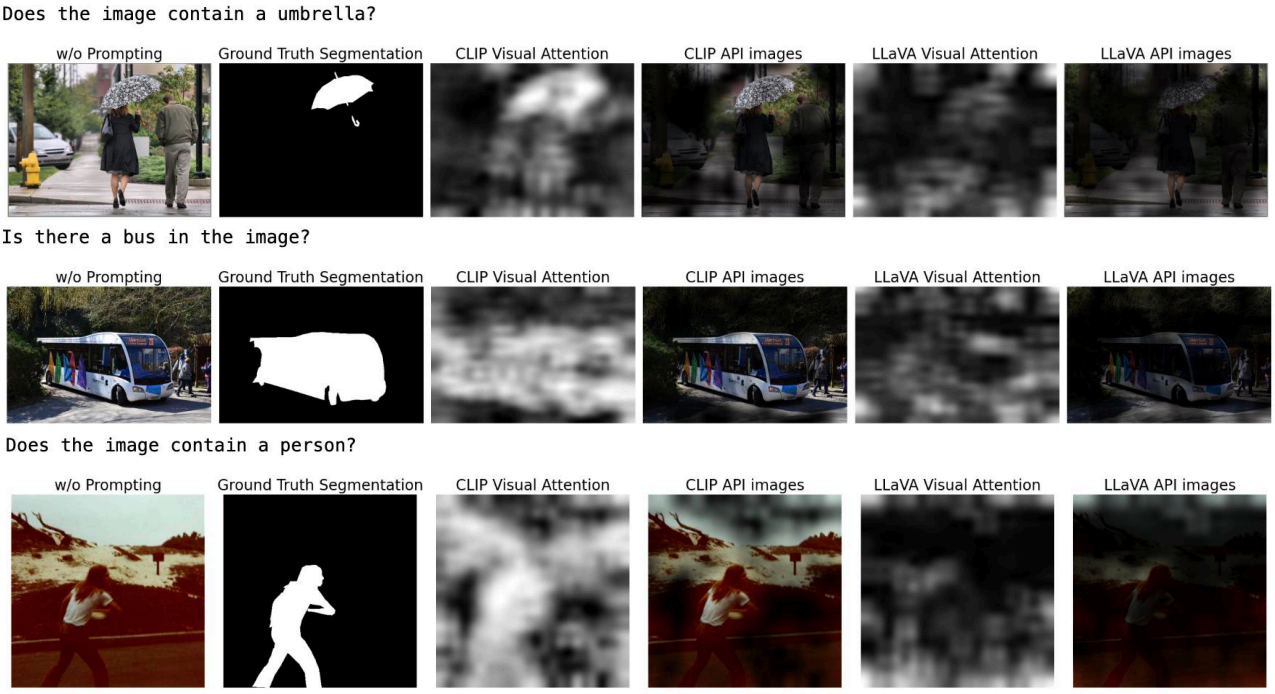

図 3　画像に含まれる物体について、POPE の質問文と左から、元の画像、正解セグメンターションデータ、CLIP の Visual Attention のヒートマップ、CLIP のヒートマップにより加工した画像、LLaVA の Visual Attention のヒートマップ、LLaVA のヒートマップにより加工した画像を並べて表示する。

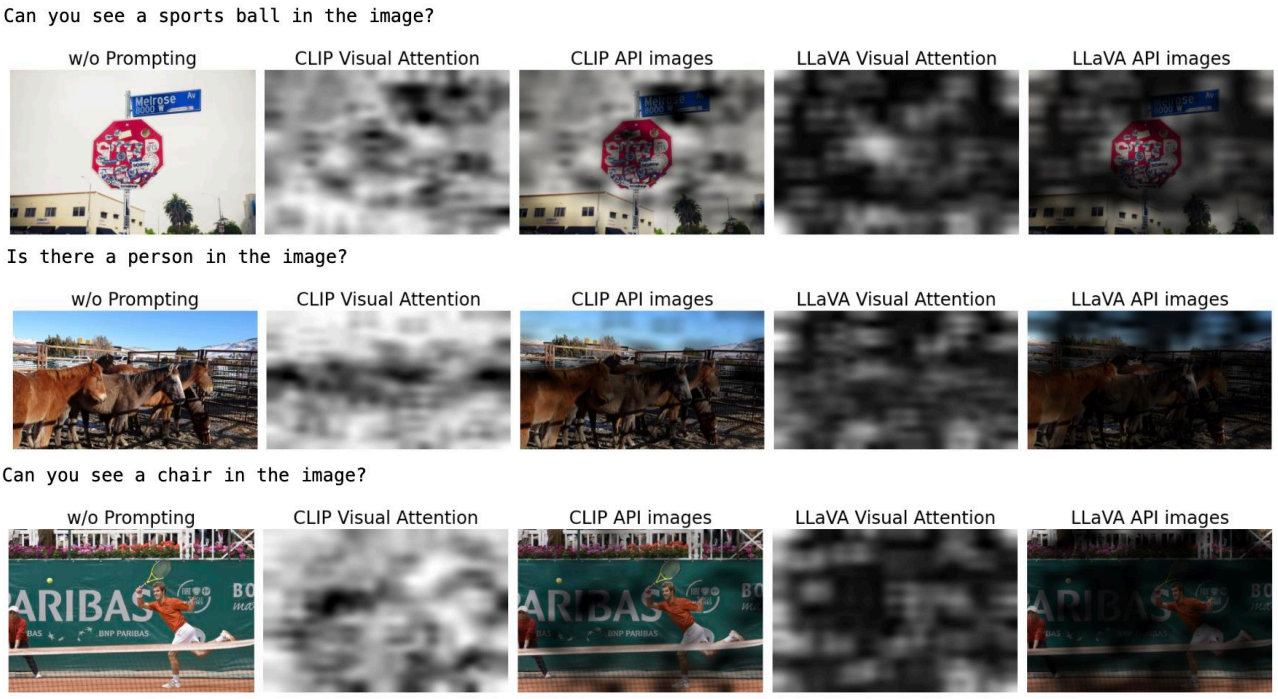

図 4　画像に含まれない物体について、POPE の質問文と左から、元の画像、CLIP の Visual Attention のヒートマップ、CLIP のヒートマップにより加工した画像、LLaVA の Visual Attention のヒートマップ、LLaVA のヒートマップにより加工した画像を並べて表示する。