MiniGPT-v2: 大規模言語モデルを統一的に視覚言語マルチタスク学習のためのインターフェース

Jun Chen1,2 Deyao Zhu1 Xiaoqian Shen1 Xiang Li1 Zechun Liu2 Pengchuan Zhang2
ラグーラマン•クリシュナムーティ2ヴィカス•チャンドラ2ユンヤン•シオン2 † モハメド•エルホセイニー1 †
1キング•アブドラ科学技術大学（KAUST）
2Meta AIリサーチ

抽象的な

大規模言語モデルは、様々な言語関連アプリケーションの汎用インターフェースとして優れた能力を発揮してきました。これを踏まえ、私たちは画像記述、視覚的質問応答、視覚的グラウンディングなど、多くの視覚言語タスクを実行するための統一インターフェースの構築を目指しています。
課題は、単一のモデルを用いて、シンプルなマルチモーダル指示で多様な視覚言語タスクを効果的に実行することです。この目標達成に向けて、様々な視覚言語タスクをより適切に処理するための統合インターフェースとして扱うことができるモデル、MiniGPT-v2を紹介します。モデルのトレーニング時に、異なるタスクに固有の識別子を使用することを提案します。これらの識別子により、モデルは各タスク指示をより容易に区別できるようになり、各タスクのモデル学習効率も向上します。3段階のトレーニング後、実験結果から、MiniGPT-v2は、他の視覚言語ジェネラリストモデルと比較して、多くの視覚的質問応答および視覚的グラウンディングベンチマークで優れたパフォーマンスを発揮することが示されました。モデルとコードはhttps://minigpt-v2.github.io/で入手できます。

1 はじめに

マルチモーダル大規模言語モデル（LLM)は、視覚AIアシスタント、画像キャプション作成、視覚的質問応答（VQA）、参照表現理解（REC)など、視覚言語コミュニティにおける豊富なアプリケーションを備えた刺激的な研究トピックとして浮上しています。 マルチモーダル大規模言語モデルの主な特徴は、LLMから高度な機能（論理的推論、常識、強力な言語表現など)を継承できることです[32, 49, 50, 8]。 適切な視覚言語指示で調整されたマルチモーダルLLM、特に視覚言語モデルは、詳細な画像記述の生成、コードの生成、画像内の視覚オブジェクトのローカライズ、さらには複雑な視覚的質問により適切に答えるためのマルチモーダル推論の実行など、強力な機能を発揮します[59, 26, 55, 53, 7, 10, 58, 6, 60]。この LLM の進化により、個人とのコミュニケーションにおける視覚と言語の入力の相互作用が可能になり、ビジュアル チャットボットの構築に非常に効果的であることが証明されています。

しかし、複数の視覚言語タスクを効果的に遂行し、それらに対応するマルチモーダル指示を策定することを学習することは、異なるタスクに内在する複雑さのために、大きな課題となります。例えば、「人の位置を教えてください」というユーザー入力があった場合、具体的なタスクに応じて解釈や応答の方法は多岐にわたります。指示表現理解タスクの文脈では、人物の位置を示す境界ボックスを1つ示すだけで回答できます。

視覚的な質問応答タスクでは、モデルは人間の自然言語を用いて空間的な位置を説明するかもしれない。人物検出タスクでは、モデルはすべての空間的な位置を特定するかもしれない。

Meta AIでのインターンシップ中に部分的に行われた作業 †最後の著者と同じ

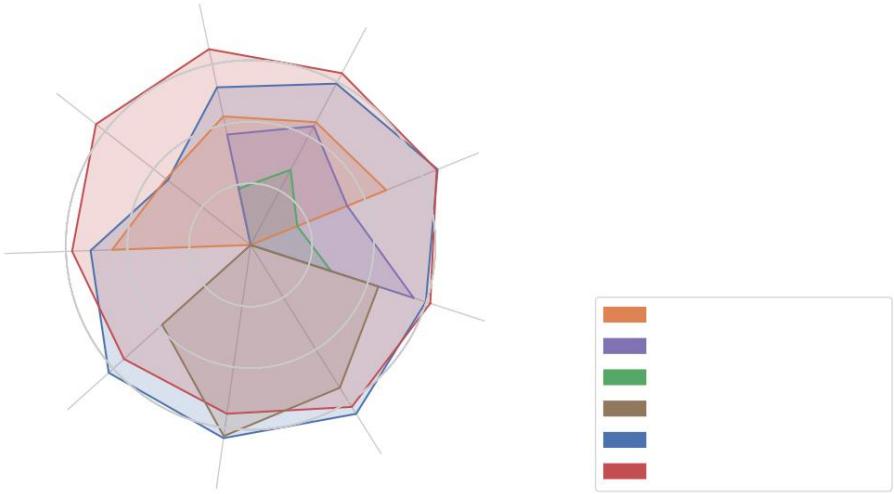


図 1: 当社の MiniGPT-v2 は、他のジェネラリスト モデルと比較して、幅広い視覚言語タスクで最先端のパフォーマンスを実現します。

与えられた画像内の各人物。この問題を軽減し、統一的なアプローチに向けて、マルチモーダルな指示の曖昧性を低減するタスク指向指示学習スキームと、視覚言語モデルであるMiniGPT-v2を提案します。具体的には、各タスクに固有のタスク識別子トークンを提供します。例えば、視覚的質問応答タスクからのすべてのデータサンプルを学習するために、[vqa]という識別子トークンを提供します。モデルの学習段階では、合計6つの異なるタスク識別子を提供します。

私たちのモデルであるMiniGPT-v2は、シンプルなアーキテクチャ設計を採用しています。ViTビジョンエンコーダ[12]から視覚トークンを直接取り込み、大規模言語モデル[50]の特徴空間に投影します。
視覚認識を向上させるため、学習には高解像度画像（448×448)を使用します。ただし、これにより画像トークンの数が増加します。モデルの学習効率を高めるため、隣接する画像トークン4つを1つのトークンに連結することで、トークンの総数を75%削減しています。さらに、弱ラベル付きで細粒度な画像テキストデータセットとマルチモーダルな教育データセットを組み合わせ、各段階で学習の焦点を変えることで、3段階の学習戦略を採用し、モデルを効果的に学習します。

モデルの性能を評価するために、 （詳細な)画像/グラウンデッドキャプション、ビジョン質問応答、ビジュアルグラウンディングなど、多様な視覚言語タスクで広範な実験を実施しました。結果は、MiniGPT-v2が、 MiniGPT-4 [59]、 InstructBLIP [10]、 LLaVA [26] 、Shikra [7]などの以前の視覚言語ジェネラリストモデルと比較して、さまざまなベンチマークでSOTAまたは同等の性能を達成できることを示しています。たとえば、MiniGPT-v2はVSRベンチマーク[25]でMiniGPT-4を21.3%、InstructBLIPを11.3%、LLaVAを11.7%上回り、RefCOCO、RefCOCO+、RefCOCOgのほとんどの検証で、以前に確立された強力なベースラインであるShikraよりも優れた性能を発揮します。私たちのモデルは、図1に示すように、視覚言語ジェネラリスト モデルの中でこれらのベンチマークにおいて新たな最先端の結果を確立しました。

2 関連研究

視覚的アライメントのための高度な大規模言語モデルとマルチモーダル LLM に関する関連研究を簡単にレビューします。

高度な大規模言語モデル（LLM）。GPT -2 [38]や BERT [11]などの初期段階のモデルは、ウェブ規模のテキストデータセットで学習された基礎モデルであり、NLP分野における画期的な進歩でした。
基礎モデルの成功に続いて、GPT-3 [4]、メガトロンチューリングNLG [46]、 PaLM [9]、Gopher [39]など、より高い容量とより多くのトレーニングデータを備えたLLMが開発されました。

Chinchilla [16]、 OPT [57]、 BLOOM [41]などが挙げられます。最近では、LLMを人間による指示やフィードバックと効果的に連携するように改良することに注力しています。この方向への取り組みの代表的なものとしては、InstructGPT [34]とChatGPT [32] が挙げられます。これらは、多様な言語の質問に答えたり、人間と会話したり、文章の洗練やコーディング支援といった複雑なタスクを学習したりするなど、優れた能力を示しています。

LLMのこうした進歩と並行して、LLaMA [49]言語モデルの台頭が見られます。ChatGPTと同様の人間の指示追従能力を実現するために、いくつかの研究では、高品質な指示データセットを追加してLLaMAモデルを微調整しようと試みています[1]。これらのモデルの例としては、 Alpaca [47]、 Vicuna [8]、MPT [48]などが挙げられます。Falcon [35]やLLaMA-2 [50]など、人間のフィードバックデータから学習したオープンソースの言語モデルも、 NLPコミュニティに導入され、優れた性能を示しています。

LLMによる視覚的アライメント。LLMの優れた一般化能力を利用して、興味深い研究では、視覚入力をLLMにアライメントすることで、LLMをマルチモーダル領域に拡張しています。VisualGPT [5]やFrozen [51]などの初期の研究では、事前学習済みの言語モデルを使用して、画像キャプションと視覚的な質問応答における視覚言語モデルを改善しました。この初期の探求は、 Flamingo [2]やBLIP-2 [22]などのその後の視覚言語研究への道を開いた。最近では、GPT-4がリリースされ、手書きのテキスト指示に基づいてウェブサイトのコードを生成するなど、多くの高度なマルチモーダル機能を実証しています。これらの実証された機能は、適切な指示チューニングを使用して画像入力を大規模言語モデルVicuna [8]にアライメントするMiniGPT-4 [59]やLLaVA [26]などの他の視覚言語LLMに影響を与えました。これらの視覚言語モデルは、アライメント後にも多くの高度なマルチモーダル機能を示しています。

Vision-LLM [53]、 Kosmos-2 [36]、 Shikra [7]、そして我々の同時研究であるQwen-VL [3]などの最近の研究でも、マルチモデルLLMモデルが言語モデルを通して境界ボックスのテキスト形式を生成することで視覚的なグラウンディングを実行できることが実証されている。

3 方法

まず、私たちの視覚言語モデルである MiniGPT-v2 を紹介し、次にトレーニング用のタスク識別子を使用したマルチタスク命令テンプレートの基本的な考え方について説明し、最後にタスク識別子の考え方を適応させてタスク指向の命令チューニングを実現します。

3.1 モデルアーキテクチャ

提案するモデルアーキテクチャMiniGPT-v2を図2に示す。これは、視覚バックボーン、線形投影層、大規模言語モデルの3つのコンポーネントから構成される。各コンポーネントを以下のように説明する。視覚バックボーン。MiniGPT -v2は、視覚バックボーンとしてEVA [12]を採用している。

モデルの学習中は、視覚バックボーンを固定する。モデルは448×448の画像解像度で学習し、より高い画像解像度に合わせて位置エンコーディングを補間する。

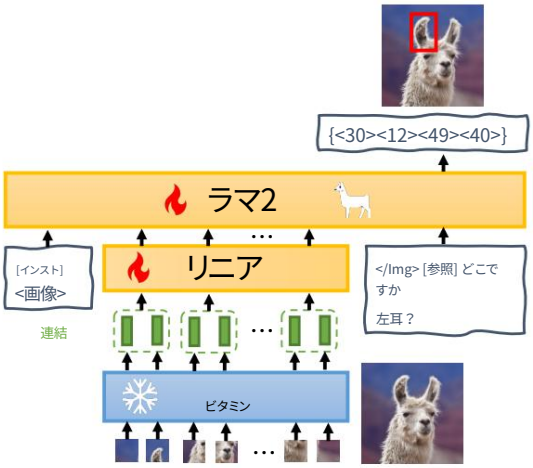


図2： MiniGPT-v2のアーキテクチャ。このモデルはViT視覚バックボーンを採用し、これはすべての学習フェーズで固定されたままです。ViTバックボーンから隣接する4つの視覚出力トークンを連結し、線形射影層を介してLLaMA-2言語モデル空間に射影します。

線形投影層。我々は、凍結された視覚バックボーンからすべての視覚トークンを言語モデル空間に投影することを目的としています。しかし、448x448などの高解像度画像の場合、すべての画像トークンを投影すると、非常に長いシーケンス入力（例えば1024トークン）となり、学習と推論の効率が大幅に低下します。そこで、埋め込み空間で隣接する4つの視覚トークンを連結し、大規模言語モデルの同じ特徴空間に単一の埋め込みに投影することで、視覚入力トークンの数を4分の1に削減します。この操作により、MiniGPT-v2は学習および推論段階で高解像度画像をより効率的に処理できます。

大規模言語モデル。MiniGPT -v2は、オープンソースのLLaMA2-chat (7B) [50]を言語モデルのバックボーンとして採用しています。本研究では、言語モデルを様々な視覚言語入力のための統一インターフェースとして扱います。LLaMA-2言語トークンを直接利用して、様々な視覚言語タスクを実行します。空間位置の生成を必要とする視覚グラウンディングタスクでは、空間位置を示す境界ボックスのテキスト表現を言語モデルに直接生成させます。

3.2 マルチタスク指示テンプレート

視覚的な質問応答、画像キャプション、参照表現、グラウンデッド画像キャプション、地域識別などの複数の異なるタスク用に単一の統合モデルをトレーニングする場合、マルチモーダル モデルは、視覚トークンを言語モデルに合わせるだけでは各タスクを区別できない可能性があります。
例えば、「赤いジャケットを着ている人の空間的な位置を教えてください」と質問すると、モデルはバウンディングボックス形式（例： < Xleft >< Ytop >< Xright >< Ybottom >)で位置を返答するか、自然言語（例 :右上隅)で物体の位置を説明します。このような曖昧さを軽減し、各タスクを容易に区別できるようにするために、トレーニング用に設計されたマルチタスク指示テンプレートに、タスク固有のトークンを導入しました。以下では、このマルチタスク指示テンプレートについて詳しく説明していきます。

一般的な入力形式。LLaMA -2会話テンプレートの設計を踏襲し、マルチモーダル指導テンプレートに適応させます。テンプレートは以下のように表記されます。

[INST] <Img> < ImageFeature> </Img> [タスク識別子] 命令 [/INST]

このテンプレートでは、[INST] はユーザー ロールと見なされ、[/INST] はアシスタント ロールと見なされます。
ユーザー入力は3つの部分に構造化されます。最初の部分は画像特徴、2番目の部分はタスク識別子トークン、3番目の部分は指示入力です。

タスク識別子トークン。本モデルは、様々なタスク間の曖昧性を低減するため、各タスクに固有の識別子を使用します。表1に示すように、視覚的な質問応答、画像キャプション、グラウンディングされた画像キャプション、参照表現の理解、参照表現の生成、フレーズ解析およびグラウンディングのそれぞれについて、6つの異なるタスク識別子を提案しました。視覚に関連しない指示については、本モデルはタスク識別子トークンを使用しません。

タスク識別子	VQA [vqa]	キャプション [キャプション]	グラウンド別]	キャプション	REC	REG	オブジェクト解析とグラウンド [グラウンド]	[参照]	[識別]	[検出]

表1: 視覚的な質問応答、画像キャプション、グラウンディングされた画像キャプション、参照表現の理解(REC)、参照表現の生成(REG)、オブジェクトの解析とグラウンディング(モデルが入力テキストからオブジェクトを抽出し、その境界ボックスの位置を決定する)の6つの異なるタスクのタスク識別子トークン。

空間位置表現。参照表現理解（REC）、参照表現生成（REG）、グラウンデッドイメージキャプションなどのタスクでは、モデルは参照対象物の空間位置を正確に特定する必要があります。本モデルでは、空間位置を境界ボックスのテキストフォーマット、具体的には「{< Xlef t >< Ytop >< Xright >< Ybottom >}」で表現します。X座標とY座標は、[0,100]の範囲に正規化された整数値で表されます。 < Xlef t >と< Ytop >は、生成された境界ボックスの左上隅のX座標とY座標を示し、< Xright >と< Ybottom >は、右下隅のX座標とY座標を示します。

3.3 マルチタスク指示訓練

設計したマルチタスク指示テンプレートを指示トレーニングに適応させます。基本的な考え方は、タスク固有の識別子トークンを持つ指示をMiniGPT-v2のタスク指向指示トレーニングの入力として取り込むことです。入力指示にタスク識別子トークンが含まれている場合、モデルはトレーニング中にマルチタスクを理解しやすくなります。視覚的アライメントを向上させるために、タスク識別子指示を使用してモデルを3段階でトレーニングします。最初の段階では、多くの弱ラベル付き画像テキストデータセットと、高品質の細粒度視覚言語アノテーションデータセット（弱ラベル付き画像テキストデータセットには高いデータサンプリング比率を割り当てます)を通じて、MiniGPT-v2が幅広い視覚言語知識を構築できるようにします。2番目の段階では、モデルを

複数のタスクのためのきめ細かなデータ。第3段階は、よりマルチモーダルなモデルを微調整することです。
多様なマルチモーダル指示に適切に応答し、行動するための指示および言語データセット
マルチモーダルチャットボットとして。各段階でのトレーニングに使用したデータセットは表2に示されています。

データ型	データセット	ステージ1	ステージ2	ステージ3
弱くラベル付けされた	GRIT-20M（RECおよびREG）、LAION、CC3M、SBU ✓			
グラウンドキャプション	グリット-20M ✓			
キャプション	COCOキャプション、テキストキャプション✓			✓
REC	RefCOCO、RefCOCO+、RefCOCOg、ビジュアルゲノム ✓			✓
登録	RefCOCO、RefCOCO+、RefCOCOg ✓		✓	✓
VQA	GQA、VQAv2、OCR-VQA、OK-VQA、AOK-VQA ✓		✓	✓
マルチモーダル指導 LLaVA データセット、Flickr30k、マルチタスク会話			✓	✓
言語データセット	不自然な指示		✓✓	✓

表 2: モデルの 3 段階トレーニングに使用したトレーニング データセット。

ステージ1：事前学習。幅広い視覚言語知識を得るために、モデルは以下のようなものを組み合わせて学習します。
弱くラベル付けされたデータセットと細粒度データセット。弱くラベル付けされたデータセットには高いサンプリング比率を与える。
第一段階でより多様な知識を得ること。

弱ラベルデータセットについては、LAION [42]、 CC3M [44]、 SBU [33] 、GRIT-20Mを使用します。
Kosmos v2 [36]は、参照表現理解（REC)のためのデータセットを構築し、
表現生成（REG）、およびグラウンデッドイメージキャプション作成。

細粒度データセットとしては、COCOキャプション[24]やテキストキャプション[45]などのデータセットを使用する。
画像キャプション、 RECではRefCOCO [20]、 RefCOCO+ [56]、 RefCOCOg [29]が用いられた。REGでは、
ReferCOCOとその変種のデータを再構築し、フレーズ→境界の順序を逆にした。
ボックスをバウンディングボックスに変換する→フレーズ。VQAデータセットの場合、トレーニングにはさまざまなデータセットを使用します。
GQA [19]、VQA-v2 [14]、OCR-VQA [31]、OK-VQA [30]、AOK-VQA [43]など。

ステージ2：マルチタスクトレーニング。各タスクにおけるMiniGPT-v2のパフォーマンスを向上させるために、
この段階では、細粒度データセットを用いてモデルを学習することに焦点を当てます。弱教師あり学習は除外します。
GRIT-20MやLAIONなどのデータセットをステージ1から取得し、それに応じてデータサンプリング比率を更新します。
各タスクの頻度に応じて、モデルは高品質のタスクを優先的に実行します。
さまざまなタスクで優れたパフォーマンスを実現する画像テキスト データ。

ステージ3：マルチモーダル命令の調整。その後、より多くのモデルをチューニングすることに焦点を当てます。
マルチモーダル指示データセットを構築し、チャットボットとしての会話能力を強化しています。
第2段階のデータセットを使用し、LLaVA [26]を含む指導データセットを追加する。
Flickr30kデータセット[37]、我々が構築した混合マルチタスクデータセット、および言語データセット、Unnatural
指示[17]。ステージ2の細粒度データセットに対しては低いデータサンプリング比率を与え、
新しい命令データセットのデータサンプリング比率が向上します。

– LLaVA命令データ。詳細な命令を含むマルチモーダル命令チューニングデータセットを追加します。
LLaVA [26]から抽出した23,000件と58,000件のデータ例から、特徴量による説明と複雑な推論を抽出した。

– フリッカー30k。第2段階のトレーニング後、MiniGPT-v2は効果的にグラウンデッドな
画像のキャプション。しかし、これらの説明は短く、多くの場合、非常に少数の
視覚的オブジェクト。これは、我々のモデルが使用されたKOSMOS-v2 [36]のGRIT-20Mデータセットが
訓練されたモデルは、各キャプションに限られた数の視覚的オブジェクトを特徴としており、
より多くの視覚的対象を認識できるようにするための適切なマルチモーダル指示の調整が欠けている。改善するには
そこで、より文脈的な情報を提供するFlickr30kデータセット[37]を使用してモデルを微調整した。
キャプション内のエンティティの根拠。

モデルを訓練するために、Flickr30kデータセットを2つの異なる形式で用意し、グラウンデッドなパフォーマンスを実現した。
画像のキャプションと新しいタスク「オブジェクトの解析とグラウンディング」：

1)根拠のある画像キャプション。最低5つの根拠のあるフレーズを含むキャプションを選定します。
約2.5kのサンプルがあり、モデルに直接指示して、グラウンド画像のキャプションを生成します。例：
部屋の中央に 木製のテーブル {<Xlef t><Ytop><Xright><Ybottom>}が置かれています。

2)オブジェクトの解析とグラウンディング。この新しいタスクは、入力されたキャプションからすべてのオブジェクトを解析することです。
そして各物体を接地します。これを可能にするために、タスク識別子[検出]を使用してこれを区別します
他のタスクからの能力。また、Flickr30kを使用して2種類の指示データセットを構築します。

方法	接地	OKVQA	GQA	VSR (ゼロショット)	IconVQA (ゼロショット)	ビズウィズ (ゼロショット)	HM (ゼロショット)
フラミンゴ-9B		44.7	-	31.8	-	28.8	57.0
ブリップ-2（13B）		45.9	41.0	50.9	40.6	19.6	53.7
インストラクトBLIP（13B）		-	49.5	52.1	44.8	33.4	57.5
ミニGPT-4（13B）		37.5	30.8	41.6	37.6	-	-
LLaVA（13B）		54.4	41.3	51.2	43.0	-	-
シクラ（13B）	✓	47.2	-	-	-	-	-
私たちのもの（7B）	✓	56.9	60.3	60.6	47.7	32.9	58.2
私たちの（7B)チャット	✓	57.8	60.1	62.9	51.5	53.6	58.8

表3 :複数のVQAタスクの結果。各タスクのトップ1の精度を報告します。グラウンディング列
モデルに視覚的な位置推定機能が組み込まれているかどうかを示します。
各ベンチマークは太字で表示されます。

方法	モデルの種類	RefCOCO ヴァル	RefCOCO テストA	RefCOCO テストB	RefCOCO+ ヴァル	RefCOCO+ テストA	RefCOCO+ テストB	RefCOCOg値 ヴァル	RefCOCOg値 テスト	平均
ユニネクスト	専門モデル	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37	88.90
G-DINO-L		90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02	86.60
ビジョンLLM-H	ジェネラリストモデル	-	86.70	-	-	-	-	-	-	-
OFA-L		79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	72.65
シクラ（7B）		87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	82.93
シクラ（13B）		87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	83.96
私たちのもの（7B）		88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66	84.29
私たちの（7B)チャット		88.06	91.29	84.30	79.58	85.52	73.32	84.19	84.31	83.70

表4:参照表現理解タスクの結果。MiniGPT -v2は
VisionLLM [53]、 OFA [52] 、Shikra [7]などの多くのVLジェネラリストモデルと互換性があり、
UNINEXT [54]やG-DINO [27]などの専門モデルと比較した精度のギャップ。

キャプション→グラウンドフレーズとフレーズ→グラウンドフレーズ、それぞれ約2.5kと3kを含む
サンプル。次にモデルに[検出]の説明という指示を与えると、モデルは
入力画像の説明からオブジェクトを直接解析し、オブジェクトを境界に落とし込む
箱。

– マルチタスクデータセットの混合。単一ラウンドの指示と回答のペアで広範囲にトレーニングした後、
モデルは、コンテキストが複数回の会話中に複数のタスクをうまく処理できない可能性がある。
より複雑になります。この状況を緩和するために、新しい複数ラウンドの会話データセットを作成します。
異なるタスクからのデータを混合することで、このデータセットを第3段階のモデルトレーニングに組み込みます。

– 不自然な指導。言語モデルの会話能力は、長時間の学習後に低下する可能性がある。
視覚言語訓練では、この問題を修正するために、言語データセットUnnatural Instruction [17]を学習データに加える。
言語生成能力の回復を助けるモデルの第 3 段階のトレーニング。

4つの実験

このセクションでは、実験の設定と結果を示します。私たちは主に、
（詳細な)画像/グラウンディングキャプション、視覚質問応答、視覚グラウンディングタスクについて
参照表現の理解を含む。定量的および定性的な結果の両方を提示する。

実装の詳細。トレーニングプロセス全体を通して、MiniGPT-v2の視覚的なバックボーンは
凍結されたままです。線形投影層のトレーニングと言語の効率的な微調整に焦点を当てます。
LoRA [18]を用いたモデル。LoRAでは、低ランク適応によってWqとWvを微調整する。
実装では、ランクをr = 64に設定しました。モデルは448x448の画像解像度でトレーニングしました。
すべての段階で、私たちは設計したマルチモーダル指導テンプレートを使用して、
モデルのトレーニング中にさまざまな視覚言語タスクを実行します。

トレーニングとハイパーパラメータ。コサイン学習率スケジューラを備えたAdamWオプティマイザーを使用する。
モデルを訓練するために、初期段階では、8基のA100 GPUで40万ステップの訓練をグローバル
バッチサイズは96、最大学習率は1e-4です。この段階は約90時間かかります。
第2段階では、モデルは最大学習率で4xA100 GPUで50,000ステップのトレーニングを受ける。

	OKVQA	GQA	WizViz	VSR	IconVQA	HM	平均
タスク識別子なしの当社	50.5	53.4	28.6	57.5	44.8	56.8	48.6
私たちの	52.1	54.6	29.4	59.9	45.6	57.4	49.8

表5: VQAベンチマークにおけるタスク識別子除去研究。モデル中にタスク識別子ありトレーニングにより、複数のVQAベンチマークからVQAパフォーマンスを全体的に向上させることができます。

方法	チェアリ↓	チェア↓	レン
MiniGPT-4	9.2	31.5	116.2
mPLUG-Owl	30.2	76.8	98.5
LLaVA	18.8	62.7	90.7
マルチモーダルGPT	18.2	36.2	45.7
MiniGPT-v2（ロング）	8.7	25.3	56.5
MiniGPT-v2（接地）	7.6	12.5	18.9
MiniGPT-v2（ショート）	4.4	7.1	10.3

表6:幻覚に関する結果。MiniGPT -v2の幻覚を異なる方法で評価した。

指導テンプレートを作成し、評価用に3つのバージョンのキャプションを出力します。「長い」バージョンについては、プロンプトを使って、与えられた画像についての簡単な説明を生成します。「grounded」バージョンでは、指示は[グラウンディング]このイメージをできるだけ詳細に説明してください。「短縮版」では、プロンプトは[キャプション]画像を簡単に説明してください。

1e-5、グローバルバッチサイズ64を採用し、このトレーニングステージは約20時間かかります。最後のこの段階では、4xA100 GPUで24のグローバルバッチサイズを使用して、さらに35,000ステップのトレーニングが実行される。このトレーニング段階には約 7 時間かかりましたが、最大学習率は 1e-5 のままでした。

4.1 定量評価

データセットと評価指標。VQAと視覚的グラウンディングのさまざまな手法でモデルを評価します。ベンチマーク。VQAベンチマークとして、OKVQA [43]、 GQA [19]、視覚空間推論(VSR) [25]、 IconVQA [28]、 VizWiz [15]、 HatefulMemes、(HM) [21]などが挙げられる。視覚的な根拠については、RefCOCO [20]とRefCOCO+[56]、RefCOCOg[29]ベンチマークでモデルを評価した。

VQAベンチマークを評価するために、貪欲なデコード戦略を用いたオープンエンドアプローチを採用した。各VQA質問を次の指示テンプレート「[vqa]質問」で評価します。以前の方法[10]では、モデルの応答をグラウンドトゥルースとトップ1の精度の報告。視覚的なグラウンディングベンチマークには、テンプレートを使用します。「[参照] 参照表現の場所を教えてください」と各参照表現の理解について質問と予測された境界ボックスは、そのIOUが報告精度に適切であるとみなされます。予測値と真実値の間の差は 0.5 より大きい。

視覚的質問応答の結果。表3は、複数のVQAを用いた実験結果を示す。ベンチマーク。我々の結果はMiniGPT-4 [59]、 Shikra [7]などのベースラインと比較して良好である。LLaVA [26]、 InstructBLIP [10]を全てのVQAタスクで上回った。例えばQKVQAでは、MiniGPT- v2はMiniGPT-4、Shikra、LLaVA、BLIP-2をそれぞれ20.3%、10.6%、3.4%、11.9%上回った。結果は、私たちのモデルの視覚的な質問応答能力の高さを示している。さらに、MiniGPT-v2（チャット）版は、2回目以降にトレーニングしたバージョンよりも高いパフォーマンスを示した。ステージ。OKVQA、VSR、IconVQA、VizWiz、HMでは、MiniGPT-v2（チャット）がMiniGPT-v2よりも優れている。それぞれ0.9%、2.3%、4.2%、20.7%、0.6%増加しました。この好調な業績は、第三段階の訓練中に向上した言語スキルは、視覚的な質問に役立つことができる理解と応答が向上し、特に VizWiz ではトップ 1 の精度が 20.7% 向上しました。

表現理解の結果を参照する。表4は、RECのベースラインと我々のモデルを比較したものである。ベンチマーク。MiniGPT-v2はRefCOCO、RefCOCO+、Ref- COCOgで優れたRECパフォーマンスを示し、他の視覚言語ジェネラリストモデルよりも優れたパフォーマンスを発揮しました。MiniGPT-v2はOFA-L [52]はRefCOCO/RefCOCO+/RefCOCOgの全タスクにおいて8%以上の精度で優れていた。強力なベースラインであるShikra（13B） [7]と比較して、我々のモデルは依然としてより良い結果を示している（例えば、84.29%対83.96%）。平均精度は低い。これらの結果は、競合する視覚的根拠の直接的な証拠となる。

図3: MiniGPT-v2の様々なマルチモーダル機能の例。
モデルは、表現理解の参照、参照などの複数のタスクを完了することができます。
表現生成、詳細な画像キャプション、視覚的な質問回答、詳細な画像
説明、および指定された入力テキストからフレーズとグラウンディングを直接解析します。

MiniGPT-v2の優れた機能。私たちのモデルは専門モデルよりも劣るものの、有望な
このパフォーマンスは、視覚的な根拠を把握する能力が向上していることを示しています。

タスク識別子のアブレーション。タスク識別子がタスク識別子に与える影響についてアブレーション研究を実施しています。
MiniGPT-v2の性能。タスク識別子を使用しないモデルとの比較
VQAベンチマークで、両モデルとも4xA100 GPUで24時間、同数の
複数の視覚言語課題に対する訓練ステップ。表5の結果は、
複数のVQAベンチマークでトークン識別子のトレーニングが全体的な品質向上に効果があることを一貫して示しています。
MiniGPT-v2のパフォーマンス。具体的には、タスク指向の指示学習を備えたMiniGPT-v2
平均で1.2%のトップ1精度向上を達成しました。これらのアブレーション結果は、明確な
タスク識別子トークンを追加する利点と、マルチタスクのためのマルチタスク識別子の使用をサポートする
学習効率。

幻覚。画像説明生成におけるモデルの幻覚を測定し、その結果をMiniGPT-4 [59]、 mPLUG- Owl [55]、 LLaVA [26]、MultiModal-GPT [13]などの他の視覚言語ベースラインと比較した。 [23]の方法論に従い、 CHAIR [40]を使用してオブジェクトレベルと文レベルの両方で幻覚を評価した。表6に示すように、 MiniGPT-v2は他のベースラインと比較して幻覚が少ない画像説明を生成する傾向があることがわかった。MiniGPT-v2では3種類のプロンプトを評価した。まず、特定のタスク識別子なしで、与えられた画像の簡単な説明を生成するプロンプトを使用し、より詳細な画像説明が生成される傾向がある。次に、グラウンディングされた画像キャプションを評価するために、[グラウンディング]この画像をできるだけ詳しく説明してくださいという指示プロンプトを提供する。最後に、 [キャプション]画像を簡単に説明するようにモデルに指示する。これらのタスク識別子を用いることで、MiniGPT-v2は様々なレベルの幻覚を伴う多様な画像記述を生成することができます。その結果、これら3つの命令バリアント全てにおいて、特に[caption]と[grounding]のタスク指定子を用いた場合、ベースラインよりも幻覚の度合いが低くなっています。

4.2 定性的な結果

本稿では、本モデルのマルチモーダル機能を補完的に理解するための定性的な結果を示します。いくつかの例を図3に示します。具体的には、これらの例において、a) 物体識別、b) 詳細なグラウンディングされた画像キャプション、c) 視覚的な質問応答、d) 参照表現の理解、e) タスク識別子を用いた視覚的な質問応答、f) 詳細な画像説明、g) 入力テキストからの物体解析とグラウンディングといった様々な能力を示しました。

より定性的な結果は付録に記載されています。これらの結果は、本モデルが競合する視覚言語理解能力を備えていることを示しています。さらに、第3段階では物体解析とグラウンディングタスクに関する数千の指示サンプルのみを用いてモデルを学習しましたが、モデルは指示に効果的に従い、新しいタスクにおいて汎化能力を発揮できることに注目してください。これは、本モデルが多くの新しいタスクに適応できる柔軟性を備えていることを示しています。

なお、画像の説明や視覚的なグラウンディングを生成する際に、モデルが幻覚を示すことが時々あることにご注意ください。例えば、モデルが存在しない視覚的オブジェクトの説明を生成したり、グラウンディングされたオブジェクトの視覚的位置を不正確に生成したりすることがあります。より高品質な画像とテキストが整合したデータを用いた学習と、より強力な視覚バックボーンや大規模言語モデルとの統合により、この問題を軽減できる可能性があると考えています。

5 結論

本稿では、様々な視覚言語マルチタスク学習のための統一インターフェースとして機能するマルチモーダルLLMであるMiniGPT-v2を紹介します。複数の視覚言語タスクを処理できる単一のモデルを開発するために、学習および推論中に各タスクに異なる識別子を使用することを提案します。これらの識別子は、モデルが様々なタスクを容易に区別し、学習効率を向上させるのに役立ちます。MiniGPT-v2は、多くの視覚的質問応答および参照表現理解ベンチマークにおいて最先端の結果を達成しています。また、このモデルは新しい視覚言語タスクにも効率的に適応できることがわかり、MiniGPT-v2が視覚言語コミュニティにおいて多くの潜在的な応用の可能性を秘めていることを示唆しています。

参考文献

[1] 共有部。 https://github.com/domeccleston/sharegpt、2023年。

[2] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、 Arthur Mensch、Katherine Millican、Malcolm Reynolds他「Flamingo：少量学習のための視覚言語モデル」Advances in Neural Information Processing Systems、2022年。

[3] Jinze Bai、Shuai Bai、Shusheng Yang、Shijie Wang、Sinan Tan、Peng Wang、Junyang Lin、Chang Zhou、 Jingren Zhou。 Qwen-vl: 多彩な機能を備えたフロンティアの大型ビジョン言語モデル。 arXiv プレプリントarXiv:2308.12966、2023。

[4] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell他。言語モデルは少数ショット学習器である。 神経情報処理システムの進歩、33:1877–1901、2020年。

[5] Jun Chen, Han Guo, Kai Yi, Boyang Li, Mohamed Elhoseiny. Visualgpt: 画像キャプション作成のための事前学習済み言語モデルのデータ効率の高い適応. IEEE/CVFコンピュータビジョンおよびパターン認識会議論文集, 18030–18040ページ, 2022年.

[6] Jun Chen、Deyao Zhu、Kilichbek Haydarov、Xiang Li、Mohamed Elhoseiny。ビデオチャットキャプション：より豊かな時空間記述に向けて。arXivプレプリントarXiv:2304.04227, 2023。

[7] Keqin Chen、Zhao Zhang、Weili Zeng、Richon Zhang、Feng Zhu、Rui Zhao。シクラ:マルチモーダル llm の参照対話魔法を解き放ちます。 arXiv プレプリント arXiv:2306.15195、2023。

[8] Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E. Gonzalez、Ion Stoica、および Eric P. Xing。 Vicuna: 90%* の chatgpt 品質を備えた gpt-4 を印象づけるオープンソースチャットボット、2023 年 3 月。

[9] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、 Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、et al. Palm：パスウェイによる言語モデルのスケーリング。arXivプレプリント arXiv：2204.02311、2022年。

[10]ダイ•ウェンリャン、リー•ジュンナン、リー•ドンシュウ、アンソニー•メン•フア•ティオン、チャオ•ジュンチー、ワン•ウェイシェン、リー•ボーヤン、パスカル•フォン、スティーブン•ホイ。 Instructblip: 命令チューニングによる汎用ビジョン言語モデルに向けて、2023 年。

[11] Jacob Devlin、Ming-Wei Chang、Kenton Lee、Kristina Toutanova. Bert:言語理解のための深層双方向変換器の事前学習.arXivプレプリントarXiv:1810.04805, 2018.

[12] Yuxin Fang、Wen Wang、Binhui Xie、Quan Sun、Ledell Wu、Xinggang Wang、Tiejun Huang、Xinlong Wang、 Yue Cao。 Eva: マスクされた視覚表現の大規模な学習の限界を探ります。 arXiv プレプリントarXiv:2211.07636、2022。

[13] Tao Gong、Chengqi Lyu、Shilong Zhang、Yudong Wang、Miao Zheng、Qian Zhao、Kuikun Liu、Wenwei Zhang、Ping Luo、Kai Chen。 Multimodal-gpt: 人間との対話のための視覚および言語モデル。 arXiv プレプリント arXiv:2305.04790、2023。

[14] Yash Goyal、Tejas Khot、Douglas Summers-Stay、Dhruv Batra、Devi Parikh. 「vqaのvを重要視する：視覚的な質問応答における画像理解の役割を高める」 IEEEコンピュータビジョンおよびパターン認識会議論文集、6904-6913ページ、2017年。

[15] Danna Gurari、Qing Li、Abigale J Stangl、Anhong Guo、Chi Lin、Kristen Grauman、Jiebo Luo、Jeffrey P Bigham. Vizwizグランドチャレンジ：視覚障碍者からの視覚的な質問に答える. IEEEコンピュータビジョンとパターン認識会議論文集、3608-3617ページ、2018年。

[16]ジョーダン•ホフマン、セバスチャン•ボルゴー、アーサー•メンシュ、エレナ•ブチャツカヤ、トレヴァー•カイ、エリザ•ラザフォード、ディエゴ•デ•ラス•カサス、リサ•アン•ヘンドリックス、ヨハネス•ウェルブル、エイダン•クラークほか。コンピューティングに最適な大規模言語モデルのトレーニング。 arXiv プレプリント arXiv:2203.15556、2022。

[17] Or Honovich、Thomas Scialom、Omer Levy、Timo Schick「不自然な指示：（ほぼ）人間の労力を必要とせずに言語モデルを調整する」 arXivプレプリントarXiv:2212.09689、2022年。

[18] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen。 Lora: 大規模な言語モデルの低ランク適応。 arXiv プレプリント arXiv:2106.09685、2021 。

[19] Drew A HudsonとChristopher D Manning. Gqa: 実世界視覚推論と構成的質問応答のための新しいデータセット. IEEE/CVFコンピュータービジョンおよびパターン認識会議論文集、6700-6709ページ、2019年。

[20] Sahar Kazemzadeh、Vicente Ordonez、Mark Matten、Tamara Berg、「Referitgame：自然風景の写真に写った物体の参照」、2014年自然言語処理における経験的手法に関する会議（EMNLP)の議事録、787-798ページ、2014年。

[21] Douwe Kiela、Hamed Firooz、Aravind Mohan、Vedanuj Goswami、Amanpreet Singh、Pratik Ringshia、 Davide Testuggine. ヘイトミームの課題：マルチモーダルミームにおけるヘイトスピーチの検出.神経情報処理システムの進歩, 33:2611–2624, 2020.

[22] Junnan Li、Dongxu Li、Silvio Savarese、Steven Hoi。Blip-2：凍結画像エンコーダと大規模言語モデルを用いたブートストラップ言語画像事前学習。arXivプレプリントarXiv:2301.12597、2023年。

[23] Yifan Li、Yifan Du、Kun Zhou、Jinpeng Wang、Wayne Xin Zhao、Ji-Rong Wen。大規模な視覚言語モデルにおける物体幻覚の評価。 arXiv プレプリント arXiv:2305.10355、2023。

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick. Microsoft coco: コンテキストにおける共通オブジェクト. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[25] Fangyu Liu、Guy Emerson、Nigel Collier. 視覚空間推論. 計算言語学会誌、11:635–651、2023年。

[26] Haotian Liu、Chunyuan Li、Qingyang Wu、Yong Jae Lee。視覚的な指示のチューニング。 arXiv プレプリント arXiv:2304.08485, 2023.

[27] Shilong Liu、Zhaoyang Zeng、Tianhe Ren、Feng Li、Hao Zhang、Jie Yang、Chunyuan Li、Jianwei Yang、 Hang Su、Jun Zhu 他。接地恐竜: オープンセットの物体検出のための接地事前トレーニングを受けた恐竜と結婚します。 arXiv プレプリント arXiv:2303.05499、2023。

[28] Pan Lu、Liang Qiu、Jiaqi Chen、Tony Xia、Yizhou Zhao、Wei Zhang、Zhou Yu、Xiaodan Liang、Song-Chun Zhu。 Iconqa: 抽象的な図の理解と視覚的な言語推論のための新しいベンチマーク。 arXiv プレプリント arXiv:2110.13214、2021。

[29] Junhua Mao、Jonathan Huang、Alexander Toshev、Oana Camburu、Alan L Yuille、Kevin Murphy。明確な物体記述の生成と理解。IEEEコンピュータビジョンおよびパターン認識会議論文集、2016年、11～20ページ。

[30]ケネス•マリノ、モハマド•ラステガリ、アリ•ファルハディ、ルーズベ•モッタギ「Ok-vqa：外部知識を必要とする視覚的質問応答ベンチマーク」IEEE/cvfコンピュータービジョンおよびパターン認識会議論文集、3195-3204ページ、2019年。

[31] Anand Mishra、Shashank Shekhar、Ajeet Kumar Singh、Anirban Chakraborty「OCR-VQA：画像内のテキスト読み取りによる視覚的な質問応答」2019年国際文書分析認識会議（ICDAR）、947-952ページ。IEEE、2019年。

[32] OpenAI.chatgptの紹介.https://openai.com/blog/chatgpt, 2022年.

[33] Vicente Ordonez、Girish Kulkarni、Tamara Berg「Im2text：キャプション付き写真100万枚を用いた画像記述」神経情報処理システムの進歩、24、2011年。

[34] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、 Sandhini Agarwal、Katarina Slama、Alex Ray、他「人間のフィードバックによる指示に従う言語モデルの学習」神経情報処理システムの進歩、35:27730–27744、2022年。

[35]ギリェルメ•ペネド、クエンティン•マラルティッチ、ダニエル•ヘスロー、ルクサンドラ•コジョカル、アレッサンドロ•カッペリ、ハムザ•アロベイドリ、バティスト•パニエ、エブテサム•アルマズロウエイ、ジュリアン•ローネイ。 falcon llm 用の洗練された Web データセット: Web データおよび Web データのみを含む厳選されたコーパスよりも優れたパフォーマンスを発揮します。 arXiv プレプリント arXiv:2306.01116、2023 。

[36] Zhiliang Peng、Wenhui Wang、Li Dong、Yaru Hao、Shaohan Huang、Shuming Ma、Furu Wei。 Kosmos-2:マルチモーダルな大規模言語モデルを世界に発信します。 arXiv プレプリント arXiv:2306.14824、2023。

[37] Bryan A Plummer、Liwei Wang、Chris M Cervantes、Juan C Caicedo、Julia Hockenmaier、Svetlana Lazebnik. Flickr30kエンティティ：より豊富な画像から文への変換モデルのための領域とフレーズの対応関係の収集. IEEE国際コンピュータビジョン会議論文集、2015年、2641-2649ページ。

[38]アレック•ラドフォード、ジェフリー•ウー、レウォン•チャイルド、デビッド•ルアン、ダリオ•アモデイ、イリヤ•スツケヴァー他 言語モデルは教師なしマルチタスク学習者である。OpenAIブログ、1(8):9、2019年。

[39] Jack W Rae、Sebastian Borgeaud、Trevor Cai、Katie Millican、Jordan Hoffmann、Francis Song、John Aslanides、Sarah Henderson、Roman Ring、Susannah Young、et al. 言語モデルのスケーリング：方法、分析、およびGopherのトレーニングからの洞察。arXivプレプリントarXiv:2112.11446、2021年。

[40]アンナ•ローバッハ、リサ•アン•ヘンドリックス、ケイリー•バーンズ、トレバー•ダレル、ケイト•サエンコ。物体幻覚画像キャプションにおいて。arXivプレプリントarXiv:1809.02156、2018年。

[41]テブン•ル•スカオ、アンジェラ•ファン、クリストファー•アキキ、エリー•パヴリック、スザナ•イリッチ、ダニエル•ヘスロー、ロマン•カスターニュ、アレクサンドラ•サーシャ•ルッチョーニ、フランソワ•イヴォン、マティアス•ガレ、他。ブルーム: 176b パラメータのオープンアクセス多言語モデル。 arXiv プレプリント arXiv:2211.05100、2022。

[42] Christoph Schuhmann、Richard Vencu、Romain Beaumont、Robert Kaczmarczyk、Clayton Mullis、Aarush Katta、Theo Coombes、Jenia Jitsev、Aran Komatsuzaki。Laion-400m：クリップフィルタリングされた4億の画像とテキストのペアのオープンデータセット。arXivプレプリントarXiv:2111.02114、2021年。

[43]ダスティン•シュウェンク、アプーヴ•カンデルワル、クリストファー•クラーク、ケネス•マリノ、ルーズベ•モッタギ「A- okvqa：世界知識を用いた視覚的質問応答ベンチマーク」ヨーロッパコンピュータビジョン会議、146-162ページ。シュプリンガー、2022年。

[44] Piyush Sharma、Nan Ding、Sebastian Goodman、Radu Soricut. 「概念キャプション：自動画像キャプション作成のための、クリーニング済み•上位概念化された画像の代替テキストデータセット」第56回計算言語学会年次会議論文集（第1巻：長編論文）、2556-2565ページ、 2018年。

[45] Oleksii Sidorov、Ronghang Hu、Marcus Rohrbach、Amanpreet Singh. Textcaps：読解力のある画像キャプション作成のためのデータセット. 2020年。

[46]シェイデン•スミス、モストファ•パトワリー、ブランドン•ノリック、パトリック•ルグレスリー、サムヤム•ラジバンダリ、ジャレッド•キャスパー、ズン•リュー、シュリマイ•プラブモエ、ジョージ•ゼルビアス、ヴィジェイ•コルティカンティ、他。 deepspeed と megatron を使用して、大規模な生成言語モデルである megatron-turing nlg 530b をトレーニングします。 arXiv プレプリント arXiv:2201.11990、2022 。

[47] Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang、およびTatsunori B. Hashimoto. Stanford alpaca: 指示に従うラマモデル. https://github.com/tatsu-lab/stanford_alpaca , 2023.

[48] MosaicML NLPチーム。mpt-7bの紹介：オープンソースで商用利用可能なLLMの新しい標準、 2023年。アクセス日：2023年5月5日。

[49]ユーゴ•トゥヴロン、ティボー•ラブリル、ゴーティエ•イザカール、ザビエル•マルティネ、マリー＝アンヌ•ラショー、ティモシー•ラクロワ、バティスト•ロジエール、ナマン•ゴヤル、エリック•ハンブロ、ファイサル•アズハルほか。 Llama: オープンで効率的な基盤

言語モデル。arXivプレプリントarXiv:2302.13971、2023。
[50] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale、et al. Llama 2:オープンな基盤と微調整されたチャットモデル。arXivプレプリントarXiv:2307.09288、2023。
[51] Maria Tsimpoukelli、Jacob L Menick、Serkan Cabi、SM Eslami、Oriol Vinyals、Felix Hill.固定言語モデルを用いたマルチモーダル少数ショット学習. ニューラル情報処理システムの進歩, 34:200–212, 2021.

[52]ペン•ワン、アン•ヤン、ルイ•メン、ジュンヤン•リン、シュアイ•バイ、ジカン•リー、ジャンシン•マ、チャン•ジョウ、ジンレン•チョウ、ホンシア•ヤン。 Ofa: シンプルなシーケンス間の学習フレームワークを通じて、アーキテクチャ、タスク、およびモダリティを統合します。機械学習に関する国際会議、23318 〜 23340 ページ。PMLR、2022年。
[53] Wenhai Wang、Zhe Chen、Xiaokang Chen、Jiannan Wu、Xizhou Zhu、Gang Zeng、Ping Luo、Tong Lu、Jie Zhou、Yu Qiao 他。 Visionllm: 大規模言語モデルは、ビジョン中心のタスクのためのオープンエンドのデコーダーでもあります。 arXiv プレプリント arXiv:2305.11175、2023。
[54] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, Huchuan Lu.物体発見•検索としての普遍インスタンス知覚. IEEE/CVFコンピュータビジョンおよびパターン認識会議論文集, 15325-15336ページ, 2023年.

[55] Qinghao Ye、Haiyang Xu、Guohai Xu、Jiabo Ye、Ming Yan、Yiyang Zhou、Junyang Wang、Anwen Hu、Pengcheng Shi、Yaya Shi 他。mplug-owl: モジュール化により、大規模な言語モデルにマルチモーダリティが与えられます。 arXiv プレプリント arXiv:2304.14178、2023。

[56] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, Tamara L Berg. 参照表現におけるコンテキストのモデリング. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 69–85. Springer, 2016.
[57] Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin、et al. Opt: オープンな事前学習済みトランスフォーマー言語モデル。arXivプレプリント arXiv:2205.01068、2022年。

[58]デヤオ•ジュー、ジュン•チェン、キリチベク•ハイダロフ、シャオチェン•シェン、ウェンシュアン•チャン、モハメド•エルホセイニー。Chatgpt が質問し、blip-2 が答える: 視覚的記述を豊かにするための自動質問。arXiv プレプリントarXiv:2303.06594、2023。

[59] Deyao Zhu、Jun Chen、Xiaoqian Shen、Xiang Li、Mohamed Elhoseiny。Minigpt-4 :高度な大規模言語モデルによる視覚言語理解の強化。arXivプレプリント arXiv:2304.10592、 2023年。

[60]諸葛ミンチェン、リュウ•ハオジェ、フランチェスコ•ファッチョ、ディラン•R•アシュリー、ロバート•チョルダス、アナンド•ゴパラクリシュナン、アブドラ•ハムディ、ハサン•アベド•アル•カデル•ハムード、ヴィンセント•ハーマン、入江一樹、他。自然言語に基づいた心の社会におけるマインドストーム。 arXiv プレプリント arXiv:2305.17066、2023。

付録

補足として、私たちのモデルから生成された、視覚言語マルチタスク機能を実証するより定性的な結果を示します。

A.1 さまざまな視覚言語課題の指導テンプレート

RefCOCO/RefCOCO+/RefCOCOg: [参照] 質問の場所を教えてください

VizWiz: [vqa] 画像に基づいて、この質問に1つの単語またはフレーズで回答してください。提供された情報が不十分な場合は、「回答不可」と回答してください。

憎悪的なミーム: [vqa] これは「質問」と書かれた画像です。憎悪的ですか？ 回答: VSR: [vqa] 画像に基づいて、この文は正しいですか、それとも誤りですか？ 質問IconQA、GQA、OKVQA: [vqa] 画像に基づいて、この質問に単語またはフレーズで回答してください: 質問

A.2 追加の定性的な結果

私たちのモデルが視覚入力を受け取り、タスク指向の識別子に基づいて質問に答えることができるかどうかを調査するために、図 4、図 5、図 6、図 7 のグラウンデッド イメージ キャプション、図 8、図 9、図 10、図 11 のオブジェクト解析とグラウンディング、図12、図 13、図 14、図 15 の参照表現の理解、図 16、図 17、図 18、図 19 のオブジェクト識別など、複数の視覚言語タスクをモデルを使用して実行しました。

各タスクについて、モデルの視覚言語機能を示す4つの例を共有します。デモの結果は、複数の視覚言語タスクにおけるMiniGPT-v2の競合する視覚理解機能の直接的な証拠を示しています。たとえば、グラウンデッドキャプションの場合、モデルはオブジェクトの詳細な空間位置を含む正しいグラウンデッドイメージキャプションを提供できます。識別の場合、モデルは予想されるオブジェクト名も生成します。MiniGPT-v2は新しいシーンを理解し、質問識別子に従って応答できます。ただし、モデルにはまだいくつかの幻覚があることにも注意する必要があります。たとえば、図6では、複数の人物が正確にグラウンデッドされておらず、図7では、画像に花瓶が存在しません。

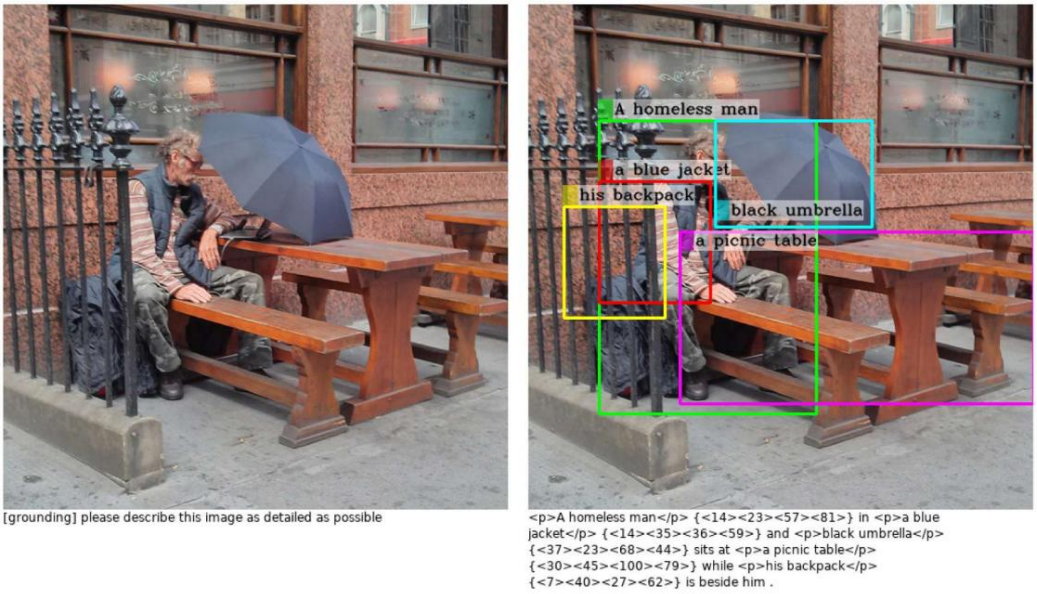
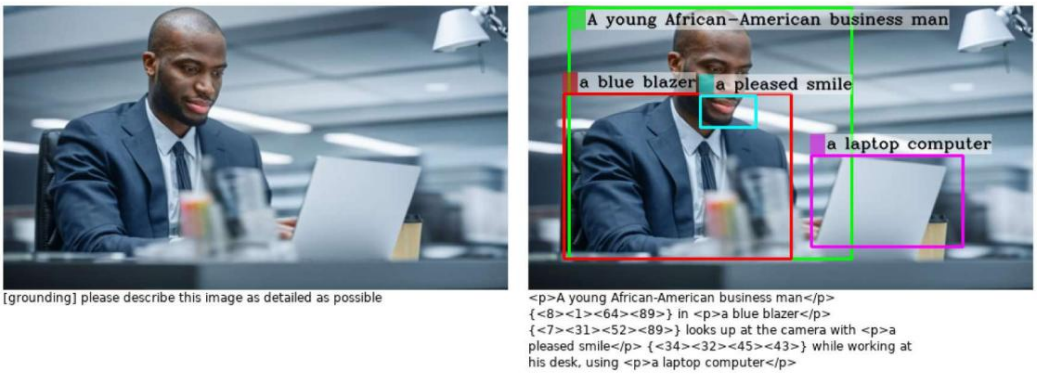


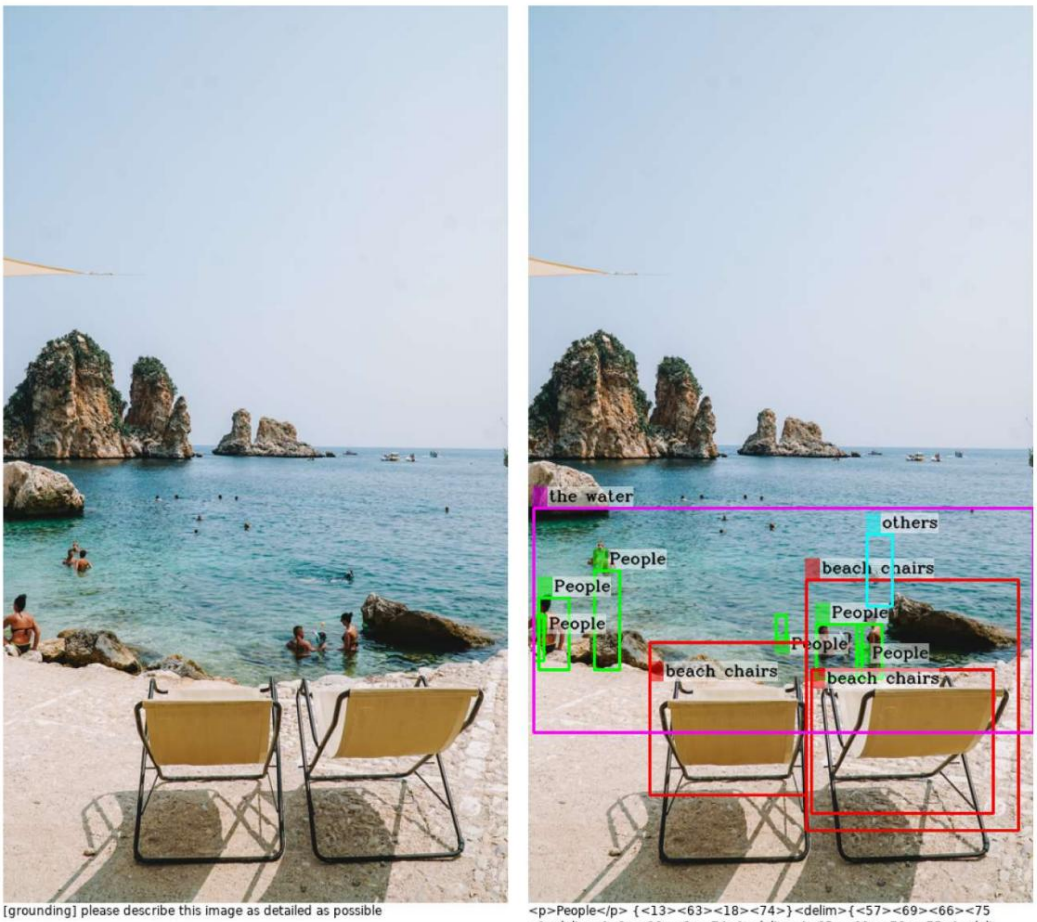
図 4: 詳細な接地画像キャプションの例。



[grounding] please describe this image as detailed as possible

A young African-American business man {<8><1><64><89>} in a blue blazer {<7><31><52><89>} looks up at the camera with a pleased smile {<34><32><45><43>} while working at his desk, using a laptop computer

図5: 詳細な地面画像キャプションの例



[grounding] please describe this image as detailed as possible

People {<13><63><18><74>}<delim>{<57><69><66><75>}<delim>{<2><66><8><74>}<delim>{<65><69><70><75>}<delim>{<1><68><3><73>}<delim>{<49><68><51><71>} sit and lay on beach chairs {<55><64><97><92>}<delim>{<24><71><55><88>}<delim>{<56><74><92><90>} while

図6: 詳細な地面画像キャプションの例

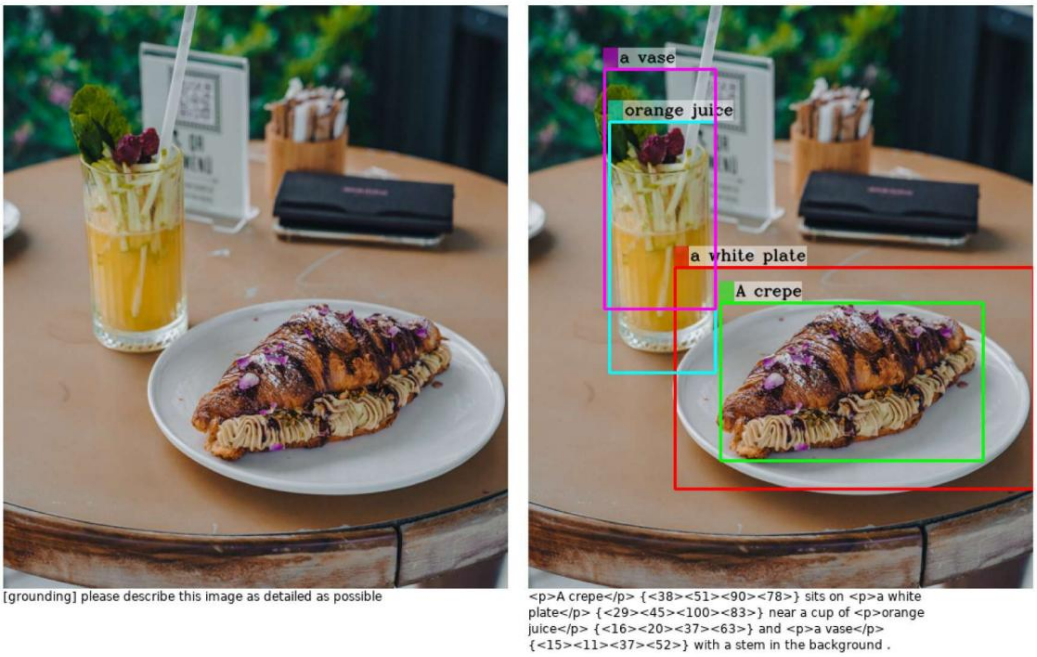


図7: 詳細な画像キャプションの例

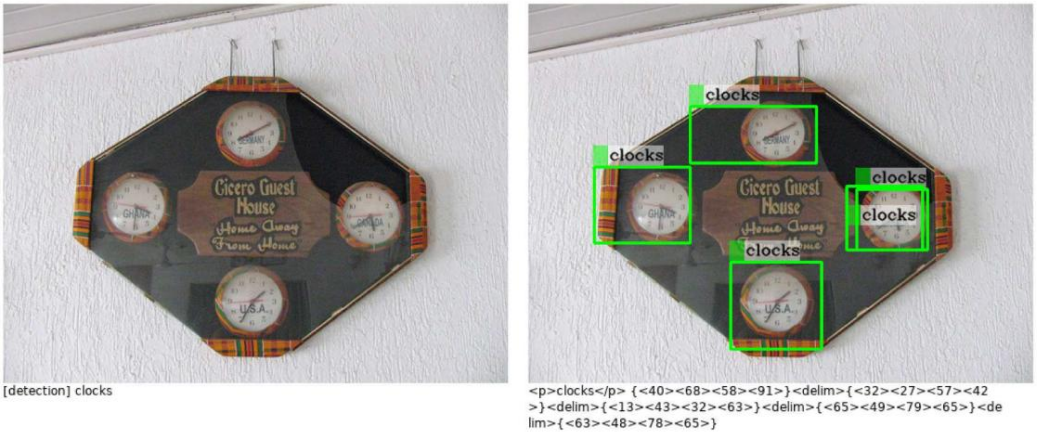


図8: オブジェクト解析とグラウンディングの例

図9: オブジェクト解析とグラウンディングの例

図10: オブジェクト解析とグラウンディングの例

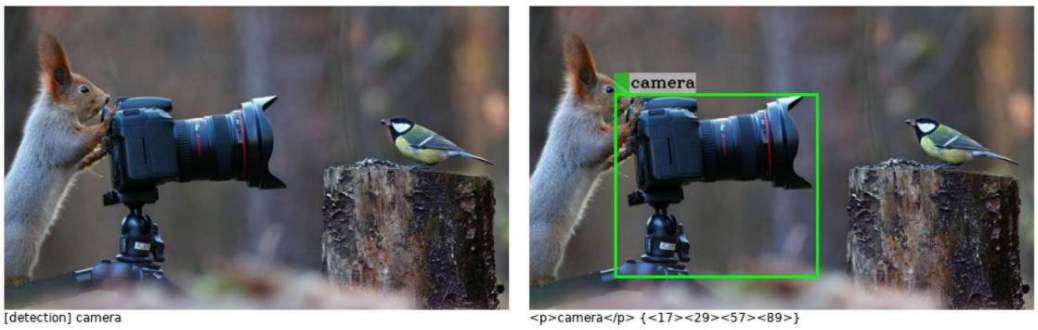


図11: オブジェクト解析とグラウンディングの例

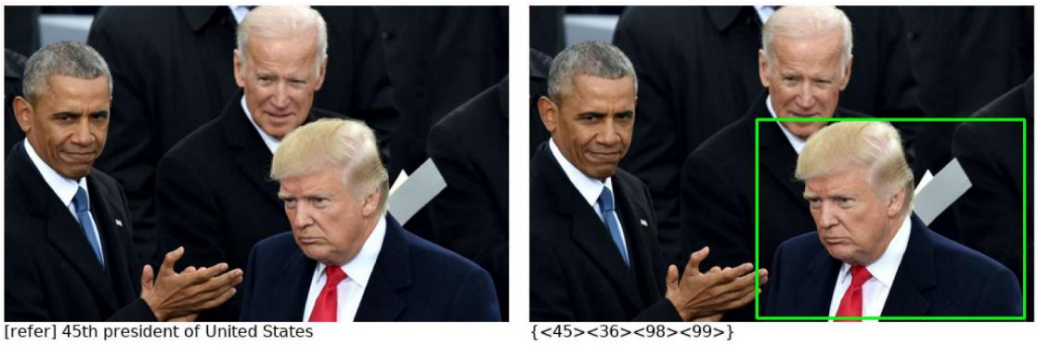


図12: 参照表現の理解例

図13: 参照表現の理解例

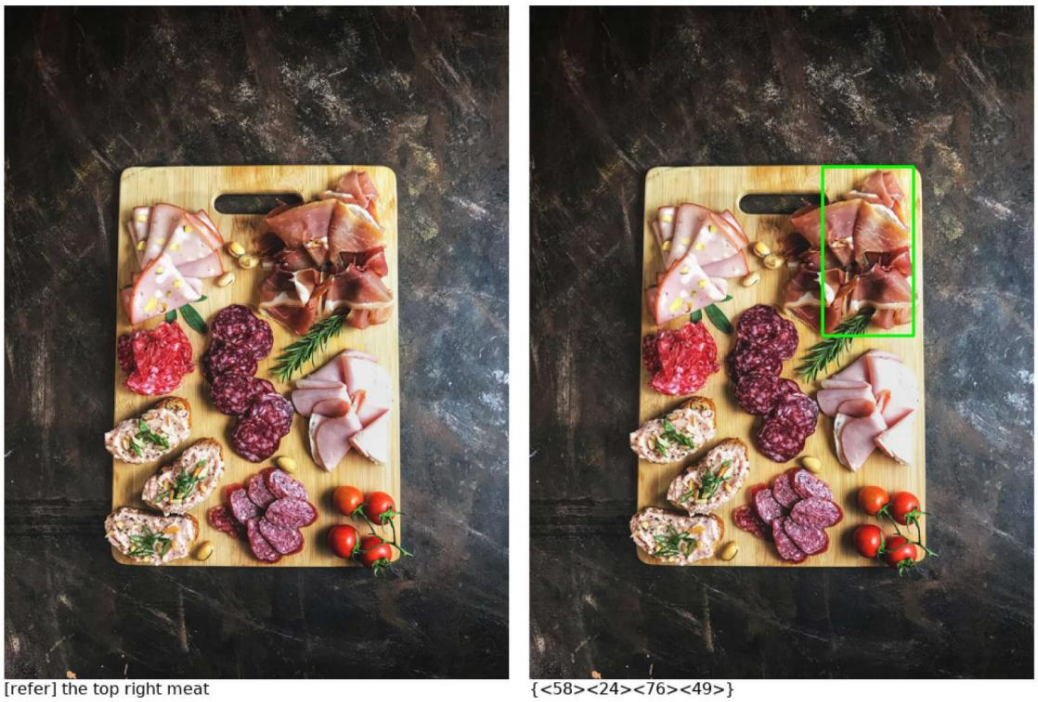


図14: 参照表現の理解例

図15: 参照表現の理解例

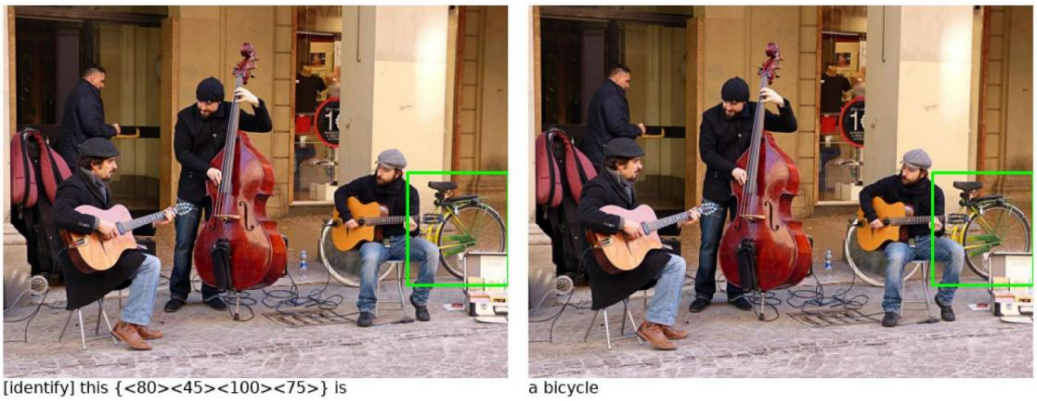


図16: オブジェクト識別の例

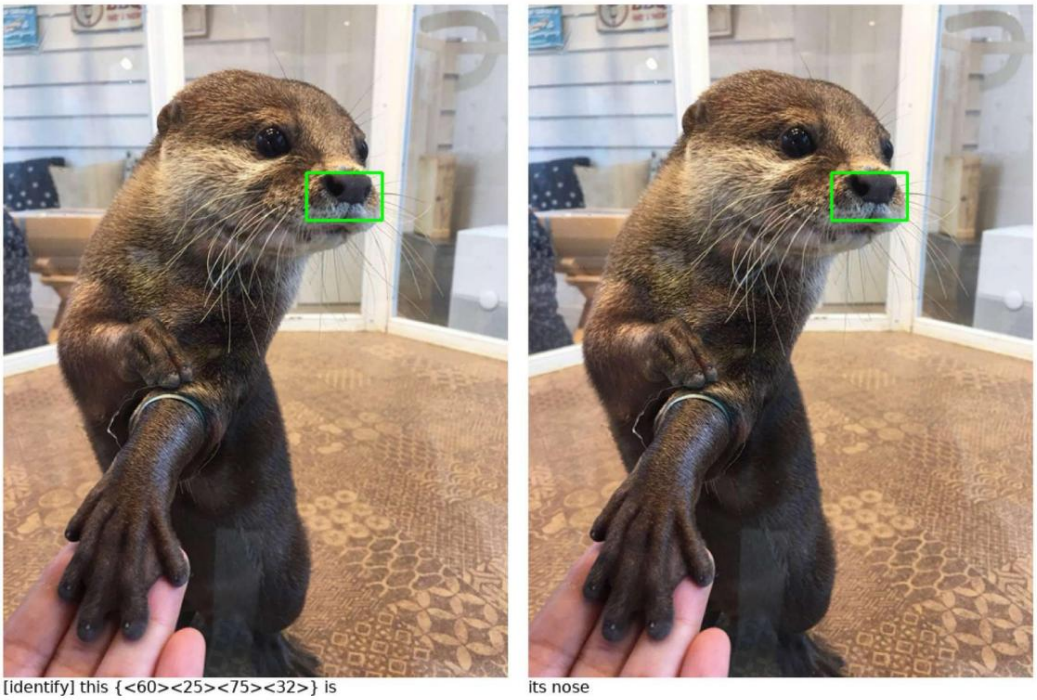


図17: オブジェクト識別の例

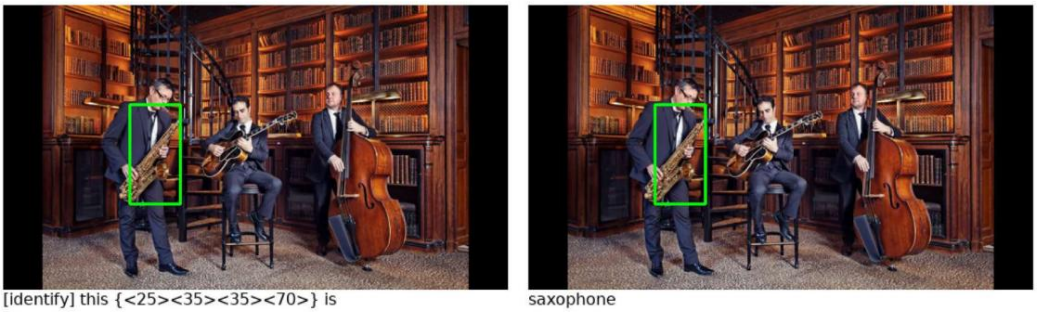


図18: オブジェクト識別の例

図19: オブジェクト識別の例