ミニグPT-4:
視覚言語理解の強化
高度な大規模言語モデルを搭載

Deyao Zhuキング 、ジュン•チェン 、シャオチェン•シェン、シャン•リー、モハメド•エルホセイニー
•アブドラ科学技術大学 {deyao.zhu,jun.chen,xiaoqian.shen,
xiang.li.1,mohamed.elhoseiny}@kaust.edu.sa

抽象的な

近年のGPT-4は、手書きテキストから直接ウェブサイトを生成したり、画像内のユーモラスな要素を識別したりするなど、並外れたマルチモーダル能力を発揮しています。これらの機能は、これまでの視覚言語モデルではほとんど見られませんでした。しかし、GPT-4の技術的な詳細は依然として明らかにされていません。GPT-4の強化されたマルチモーダル生成能力は、洗練された大規模言語モデル（LLM)の活用に起因していると考えています。この現象を検証するために、我々はMiniGPT-4を紹介します。MiniGPT-4は、固定された視覚エンコーダーと、固定された高度なLLMであるVicunaを1つの投影層でアラインメントさせます。私たちの研究は、視覚的特徴を高度な大規模言語モデルと適切にアラインメントさせることで、詳細な画像記述の生成や手描きの下書きからのウェブサイト作成など、GPT-4が実証した数多くの高度なマルチモーダル能力を実現できることを初めて明らかにしました。さらに、 MiniGPT-4には、与えられた画像からインスピレーションを得て物語や詩を書いたり、食べ物の写真に基づいてユーザーに料理の仕方を教えたりするなど、他の新たな能力も現れています。

実験では、短い画像キャプションペアで学習したモデルが、不自然な言語出力（例 :繰り返しや断片化)を生成する可能性があることがわかりました。この問題に対処するため、第2段階で詳細な画像記述データセットをキュレートし、モデルを微調整しました。これにより、モデル生成の信頼性と全体的なユーザビリティが向上しました。コード、事前学習済みモデル、および収集したデータセットは、 https://minigpt-4.github.io/で公開しています。

1はじめに

近年、大規模言語モデル（LLM)は急速な進歩を遂げています（Ouyang et al., 2022; OpenAI, 2022; Brown et al., 2020; Scao et al., 2022a; Touvron et al., 2023; Chowdhery et al., 2022; Hoffmann et al., 2022）。これらのモデルは優れた言語理解能力を備えており、複雑な言語タスクをゼロショット方式で実行できます。特に、大規模マルチモーダルモデルであるGPT-4が最近導入され、視覚言語の理解と生成に関するいくつかの印象的な機能を実証しました（OpenAI, 2023）。たとえば、GPT-4は詳細かつ正確な画像の説明を生成したり、珍しい視覚現象を説明したり、手書きのテキスト指示に基づいてWebサイトを構築したりすることもできます。

GPT-4は驚くべき視覚言語能力を示していますが、その並外れた能力の背後にある方法はまだ謎に包まれています（OpenAI、 2023）。これらの優れた能力は、より高度な大規模言語モデル（LLM)の活用に起因している可能性があると考えています。LLMは、GPT-3のfew-shot prompting設定（Brown et al., 2020)やWei et al.（2022)の調査結果に見られるように、様々な創発能力を発揮しています（Wei et al., 2022）。このような創発特性は、小規模モデルでは見つけるのが困難です。これらの創発能力はマルチモーダルモデルにも適用可能であり、GPT-4の優れた視覚記述能力の基盤となっている可能性があると推測されています。

この仮説を実証するために、MiniGPT-4という新しい視覚言語モデルを提示する。これは、高度な大規模言語モデル（LLM）であるVicuna （Chiang et al., 2023)を利用している。VicunaはLLaMA （Touvron et al., 2023)を基盤としており、GPT-4のChatGPTの90%の品質を達成すると報告されている。

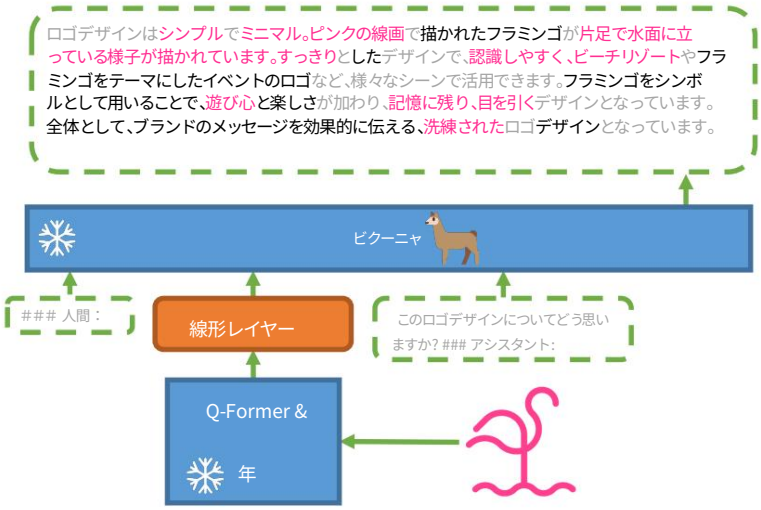


図1： MiniGPT-4のアーキテクチャ。事前学習済みのViTとQ-Formerを備えたビジョンエンコーダ、単一の線形投影層、そして高度なVicuna大規模言語モデルで構成されています。MiniGPT-4では、視覚的特徴をVicunaと整合させるために線形投影層の学習のみが必要です。

言語デコーダーとして、視覚評価を担う。視覚知覚に関しては、 EVA-CLIP (Fang et al., 2022)のViT-G/14とQ-Formerネットワークで構成されるBLIP-2 (Li et al., 2023)と同じ事前学習済みの視覚コンポーネントを採用している。MiniGPT-4は、エンコードされた視覚特徴をVicuna言語モデルと整合させるための単一の投影層を追加し、その他の視覚および言語コンポーネントはすべて固定している。
MiniGPT-4は、最初に4つのA100 GPUでバッチサイズ256を使用して20,000ステップのトレーニングを受け、LAION （Schuhmann et al., 2021）、 Conceptual Captions （Changpinyo et al., 2021; Sharma et al., 2018）、およびSBU （Ordonez et al., 2011)の画像を含む複合画像キャプションデータセットを活用して、視覚的特徴をVicuna言語モデルと整合させます。しかし、視覚的特徴を言語モデル（LLM)と整合させるだけでは、チャットボットに似た堅牢な視覚会話機能を確保するには不十分です。生の画像とテキストのペアに根本的なノイズが存在すると、標準以下の言語出力につながる可能性があります。そのため、さらに3,500の詳細な画像説明ペアを収集し、設計された会話テンプレートを使用してモデルをさらに微調整し、生成された言語の自然さと使いやすさを向上させます。

実験では、MiniGPT-4がGPT-4で実証された機能と同様の多くの機能を備えていることがわかりました。たとえば、MiniGPT-4は複雑な画像の説明を生成したり、手書きのテキスト指示に基づいてウェブサイトを作成したり、珍しい視覚現象を説明したりすることができます。さらに、私たちの調査結果から、MiniGPT-4にはGPT-4のデモでは紹介されなかったさまざまな興味深い機能もあることが明らかになりました。たとえば、MiniGPT-4は、食べ物の写真から詳細な料理のレシピを直接生成したり、画像に触発されて物語や詩を書いたり、画像内の商品の広告を書いたり、写真に示されている問題を特定して対応する解決策を提供したり、人物、映画、芸術に関する豊富な事実を画像から直接取得したりすることができます。これらの機能は、それほど強力ではない言語モデルを使用するKosmos-1 （Huang et al.、 2023)やBLIP-2 （Li et al.、2023)などの以前の視覚言語モデルには存在しません。これは、視覚的特徴と高度な言語モデルを統合することが、視覚言語モデルを強化するための鍵の 1 つであることをさらに証明しています。

主な調査結果の概要を以下に示します。

- 私たちの研究では、視覚的特徴をVicunaのような高度な大規模言語モデルと連携させることで、MiniGPT-4はGPT-4のデモで示されたものと同等の高度な視覚言語機能を実現できることを説得力のある証拠とともに明らかにしました。
- 我々の調査結果は、1つの投影層のみを学習させるだけで、事前学習済みのビジョンエンコーダーを大規模言語モデルと効果的に整合させることができることを示唆しています。MiniGPT-4では、 4つのA100 GPUで約10時間の学習のみが必要です。
- 短い画像キャプションペアを使って視覚的特徴を大規模な言語モデルと単純に整合させるだけでは、高性能なモデルを開発するには不十分であり、

不自然な言語生成。小規模ながらも詳細な画像記述ペアを用いた更なる微調整により、この制限に対処し、ユーザビリティを大幅に向上させることができます。

関連作品2

大規模言語モデルは、近年、学習データのスケールアップとパラメータ数の増加により、驚異的な成功を収めています。BERT （Devlin et al., 2018）、 GPT-2 （Radford et al., 2019）、 T5 （Raffel et al., 2020)といった初期のモデルが、この進歩の基盤を築きました。その後、 1750億パラメータという大規模なGPT-3 （Brown et al., 2020)が導入され、数多くの言語ベンチマークにおいて大きな進歩を示しました。

この発展は、Megatron- Turing NLG (Smith et al., 2022)、 Chinchilla (Hoffmann et al., 2022)、 PaLM (Chowdhery et al., 2022)、 OPT (Zhang et al., 2022)、 BLOOM (Scao et al., 2022b)、 LLaMA (Touvron et al., 2023)など、様々な大規模言語モデルの創出に影響を与えました。Wei et al. (Wei et al., 2022)はさらに、大規模モデルにのみ発現するいくつかの創発能力を発見しました。これらの能力の出現は、大規模言語モデルの開発におけるスケールアップの重要性を強調しています。さらに、InstructGPT （Ouyang et al., 2022)やChatGPT （OpenAI, 2022)は、事前学習済みの大規模言語モデルGPT-3を人間の意図、指示、そして人間のフィードバックと整合させることで、人間との会話的なインタラクションを可能にし、多様で複雑な幅広い質問に答えることができます。最近では、 Alpaca （Taori et al., 2023)やVicuna （Chiang et al., 2023)など、 LLaMA （Touvron et al., 2023)をベースに開発されたオープンソースモデルもいくつかあり、同様の性能を示しています。

視覚言語タスクにおける事前学習済みLLMの活用。近年、視覚言語タスクにおいて自己回帰言語モデルをデコーダーとして使用する傾向が大きく高まっています（ Chen et al., 2022; Huang et al., 2023; Yang et al. , 2022; Tiong et al., 2022 ; Alayrac et al ., 2022; Li et al., 2023 ; 2022; Driess et al., 2023）。このアプローチは、クロスモーダル転移を活用し、言語領域とマルチモーダル領域間で知識を共有できるようにします。VisualGPT （Chen et al., 2022)やFrozen （Tsimpoukelli et al., 2021)などの先駆的な研究は、事前学習済み言語モデルを視覚言語モデルデコーダーとして使用することの利点を実証しています。その後、ゲーテッドクロスアテンションを用いて事前学習済みのビジョンエンコーダと言語モデルを整合させるためにFlamingo （ Alayrac et al., 2022)が開発され、数十億の画像とテキストのペアで学習され、優れた文脈内少数ショット学習能力を示しました。その後、 Q-Formerを備えたFlan-T5 （Chung et al., 2022)を採用したBLIP-2（Li et al., 2023)が導入され、視覚的特徴を言語モデルと効率的に整合させました。最近では、 5620億のパラメータを備えたPaLM-E （Driess et al., 2023)が開発され、実世界の連続センサーモダリティをLLMに統合することで、実世界の知覚と人間の言語とのつながりを確立しました。GPT-4 （OpenAI, 2023)も最近リリースされ、膨大な画像とテキストデータの事前学習により、より強力な視覚理解•推論能力を示しました。

ChatGPTなどのLLMは、他の専門モデルと連携することで、視覚言語タスクのパフォーマンスを向上させる強力なツールであることが証明されています。たとえば、Visual ChatGPT （Wu et al., 2023)とMM-REACT （Yang* et al., 2023)は、ChatGPTがコーディネーターとして機能し、さまざまな視覚基盤モデルと統合して、より複雑な課題に取り組むための連携を促進する方法を示しています。ChatCaptioner （Zhu et al., 2023)は、ChatGPTを質問者として扱い、BLIP-2が回答するためのさまざまな質問を促します。複数ラウンドの会話を通じて、ChatGPTはBLIP-2から視覚情報を抽出し、画像コンテンツを効果的に要約します。Video ChatCaptioner （ Chen et al., 2023)はこのアプローチを拡張し、ビデオの時空間理解に適用します。

ViperGPT (Sur´ s et al., 2023)は、LLM と様々な視覚モデルを組み合わせることで、複雑な視覚クエリをプログラム的に処理できる可能性を示しています。一方、MiniGPT-4 は視覚情報と言語モデルを直接連携させることで、外部の視覚モデルを使用せずに多様な視覚言語タスクを実現します。

3方法

MiniGPT-4は、事前学習済みの視覚エンコーダからの視覚情報を、高度な大規模言語モデル（LLM)と整合させることを目指しています。具体的には、言語デコーダとしてVicuna （Chiang et al., 2023)を利用します。VicunaはLLaMA （Touvron et al., 2023)を基盤として構築されており、幅広い複雑な言語タスクを実行できます。視覚知覚には、MiniGPT-4で使用されているものと同じ視覚エンコーダを使用します。

BLIP-2 (Li et al., 2023) は、 ViT バックボーン(Fang et al., 2022)であり、事前トレーニング済みの Q-Former と結合されています。言語モデルと視覚モデルはどちらもオープンソースです。私たちは、線形投影層を用いて視覚エンコーダとLLM間のギャップを埋めることを目指しています。モデルの概要を図1に示します。

効果的なMiniGPT-4を実現するために、2段階の学習アプローチを提案します。第1段階では、画像とテキストの大規模な対応付け済みペアのコレクションを用いてモデルを事前学習し、視覚と言語に関する知識を獲得します。第2段階では、小規模ながらも高品質な画像とテキストのデータセットと、設計された会話テンプレートを用いて事前学習済みモデルを微調整し、生成の信頼性と使いやすさを向上させます。

3.1最初の事前トレーニング段階

初期の事前学習段階では、モデルは大規模な画像とテキストのペアのコレクションから視覚言語知識を獲得するように設計されています。注入された投影層からの出力は、 LLMへのソフトプロンプトとみなされ、対応する正解テキストを生成するよう促します。

事前学習プロセス全体を通して、事前学習済みのビジョンエンコーダとLLMは両方とも固定された状態のままで、線形投影層のみが事前学習されます。モデルの学習には、 Conceptual Caption (Changpinyo et al., 2021; Sharma et al., 2018)、 SBU (Ordonez et al., 2011) 、LAION (Schuhmann et al., 2021)を組み合わせたデータセットを使用しました。モデルは、バッチサイズ256で20,000ステップの学習を行い、約500万組の画像とテキストのペアをカバーしました。このプロセス全体は、4基のA100 (80GB) GPUを使用して約10時間で完了します。

最初の事前学習段階における問題点最初の事前学習段階を経て、MiniGPT-4は豊富な知識を有し、人間の質問に対して適切な回答を提供する能力を示しました。しかしながら、単語や文の繰り返し、断片的な文、無関係な内容など、支離滅裂な言語出力を生成する例も確認されています。これらの問題は、MiniGPT-4が人間と円滑な視覚的会話を行う上で障害となっています。

GPT-3でも同様の課題が見られました。GPT -3は広範な言語データセットで事前学習されているにもかかわらず、ユーザーの意図に正確に合致する言語出力を生成するのに苦労しています。GPT-3は、人間からのフィードバックによる指示の微調整と強化学習のプロセスを経て、GPT-3.5 (Ouyang et al., 2022; OpenAI, 2022)へと進化し、より人間に優しい出力を生成できるようになります。この現象は、MiniGPT-4が初期の事前学習段階を終えた現在の状態に似ています。そのため、この段階で私たちのモデルが流暢で自然な人間言語出力を生成するのに苦労するのは当然のことです。

3.2視覚言語ドメイン向けの高品質アライメントデータセットのキュレーション。

生成言語の自然さを高め、モデルの使いやすさを向上させるには、第二段階のアライメントプロセスが不可欠です。NLP分野では、指示の微調整データセット（Taori et al., 2023)や会話データセット（sha, 2023)は容易に入手可能ですが、視覚言語領域においては同等のデータセットは存在しません。この欠点を補うため、視覚言語アライメント用に特別に調整された詳細な画像記述データセットを慎重に作成しました。このデータセットは、第二段階のアライメントプロセスにおいてMiniGPT-4を微調整するために利用されます。

初期位置合わせ画像テキスト生成初期段階では、最初の事前学習段階で得られたモデルを用いて、入力画像の包括的な説明を生成します。モデルがより詳細な画像説明を生成できるように、以下に示すように、 Vicuna (Chiang et al., 2023)言語モデルの会話形式に準拠したプロンプトを設計しました。このプロンプトでは、<ImageFeature> は線形投影層によって生成された視覚的特徴を表します。

###人間: <Img><ImageFeature></Img>この画像を詳しく説明してください。できるだけ詳しく、あなたが見ているものをすべて話してください。###アシスタント:

不完全な文を識別するために、生成された文が80トークンを超えているかどうかを検査します。超えていない場合は、追加のプロンプト「###人間: 続行 ###アシスタント:」を組み込み、 MiniGPT-4に生成プロセスを延長させます。両方のステップからの出力を連結することで、より包括的な画像の説明を作成できます。このアプローチにより、詳細で有益な画像の説明を含む画像とテキストのペアを生成できます。

概念キャプションデータセット(Changpinyo et al.、 2021; Sharma et al.、 2018)を作成し、事前トレーニング済みモデルを使用して各画像に対応する言語の説明を生成します。

データ後処理：上記の自動生成された画像の説明には、単語や文の繰り返し、断片的な文、無関係なコンテンツなど、ノイズの多い、または支離滅裂な説明が含まれています。これらの問題を修正するために、ChatGPTを用いて以下のプロンプトで説明を修正します。

指定された段落の誤りを修正してください。重複した文、意味のない文字、英語ではない文などを削除してください。不要な繰り返しを削除してください。不完全な文は書き直してください。
説明なしで結果を直接返します。入力段落が既に正しい場合は、説明なしでそのまま返します。

後処理段階が完了すると、各画像の説明の正確性を手動で検証し、高品質を保証します。具体的には、まず頻繁に表示されるエラー（「申し訳ありませんが、間違いを犯しました…」や「申し訳ありません…」など）を特定し、それらを自動的にフィルタリングするためのルールをハードコードしました。また、ChatGPTが検出できない冗長な単語や文を削除することで、生成されたキャプションを手動で調整します。最終的に、 5,000個の画像とテキストのペアのうち、約3,500個のみが要件を満たし、これらのペアは第2段階のアライメントプロセスに使用されます。

3.3第2段階の微調整

第2段階では、厳選された高品質な画像とテキストのペアを用いて、事前学習済みモデルを微調整します。この微調整では、以下のテンプレートに定義されているプロンプトを使用します。

###人間: <Img><ImageFeature></Img><Instruction>###アシスタント:

このプロンプトでは、 <Instruction> は、定義済みの命令セットからランダムに抽出された命令を表します。この命令セットには、「この画像を詳しく説明してください」や「この画像の内容を説明してください」といった、様々な形式の命令が含まれています。この特定のテキスト画像プロンプトについては、回帰損失を計算していないことに注意してください。

その結果、MiniGPT-4 はより自然で信頼性の高い言語出力を生成できるようになりました。
さらに、この微調整プロセスは驚くほど効率的で、バッチサイズ 12 でわずか 400 のトレーニング ステップしか必要とせず、単一の A100 GPU で約 7 分しかかからないことがわかりました。

4つの実験

この実験では、様々な定性的な例を通して、MiniGPT-4モデルの多様かつ新たな能力を示すことを目指します。これらの能力には、詳細な画像の説明の生成、ミーム内の面白い側面の特定、写真からの料理レシピの提供、画像に詩を添えるといったものが含まれます。さらに、画像キャプション作成タスクにおける定量的な結果も示します。

4.1 MINIGPT-4を用いた定性的な能力評価による創発的能力の発見

例

MiniGPT-4 は、従来の視覚言語モデルと比較して多くの高度な機能を備えています。
たとえば、画像を詳細に説明したり、特定のミームのユーモラスな側面を解釈したりすることができます。
ここでは、それぞれ異なる能力を強調する 8 つの異なる例を使用して、私たちのモデルを主要な視覚言語モデルの 1 つである BLIP-2 (Li et al., 2023) と定性的に比較しました。

図2の例では、 MiniGPT-4が画像内の様々な要素（賑やかな街路、時計台、商店、レストラン、バイク、人、街灯、雲など）を効果的に識別していることがわかります。一方、BLIP-2は画像キャプション生成において、街路、人、バイクしかカバーできません。図4aに示す別の例では、MiniGPT-4がミームがなぜユーモラスなのかをうまく説明していることがわかります。嘘をついている犬は、多くの人が月曜日に感じるのと同じ気持ちだと解釈しています。月曜日は、週の中で最も嫌な日とされることが多いです。一方、BLIP-2は画像の内容を簡潔に説明するだけで、画像の面白さを理解できていません。

図2: 詳細な説明

図3: 広告プロモーション

また、他の際立った能力を実演することで、MiniGPT-4の他の能力も紹介します。これには、特定の画像に基づいて広告プロモーションを作成する（図3）、映画の写真から事実情報を取得する（図8）、食べ物の画像から料理のレシピを生成する（図11）、植物の病気を診断して治療計画を提案する（図12）、手書きの下書きからウェブサイトを作成する（図4b）、画像に触発されて詩を書く（図10)などが含まれます。これらの能力は、それほど強力ではない言語モデル（LLM)を使用するBLIP-2（言語モデルとしてFlan-T5 XXL （Chung et al., 2022)を利用)などの従来の視覚言語モデルには存在しません。この対比は、これらの高度な視覚言語能力は、視覚的特徴がVicuna （Chiang et al., 2023)などの高度なLLMと適切に調整された場合にのみ出現することを示しています。

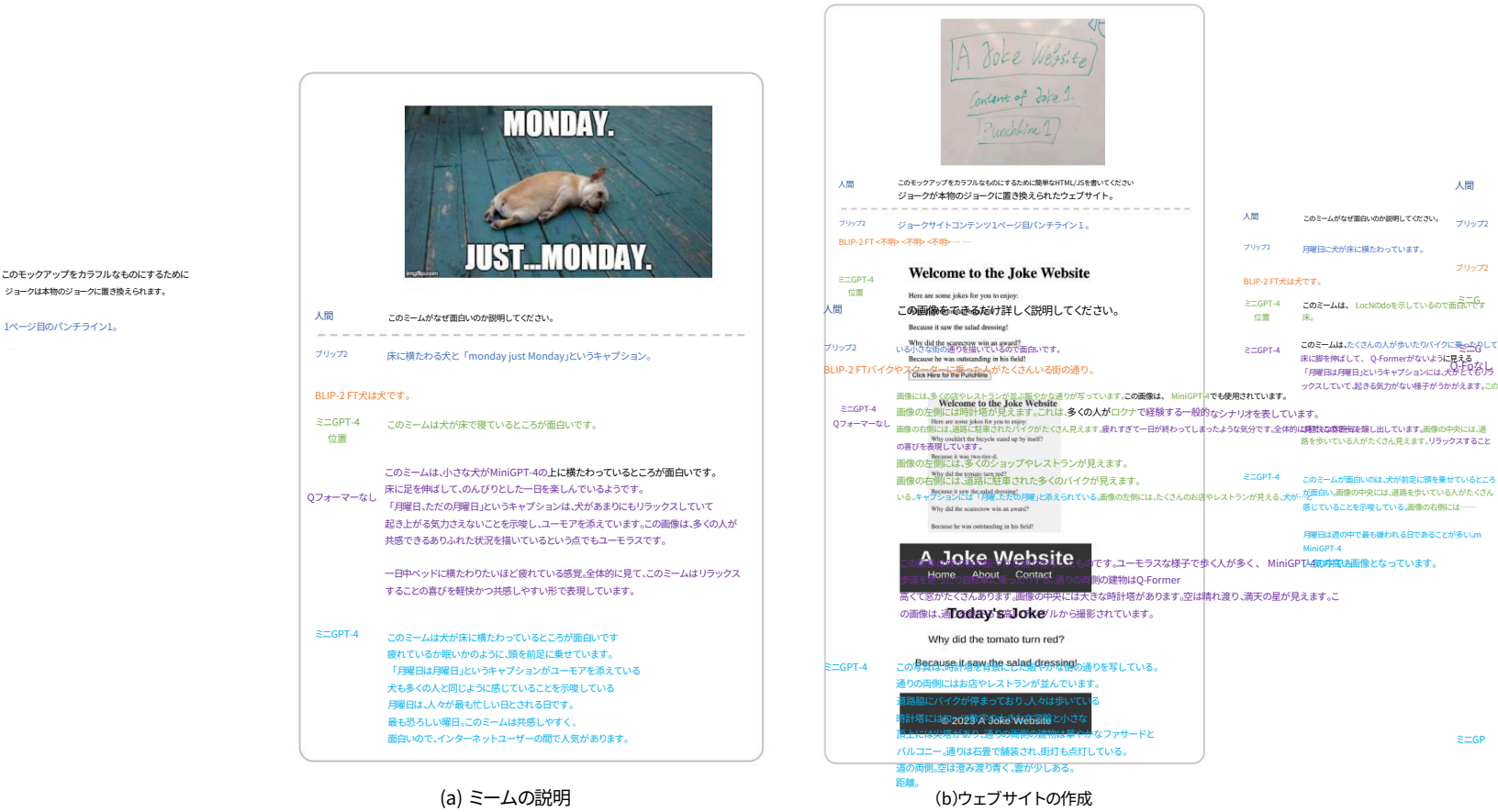
4.2定量分析

高度な能力高度な視覚言語タスクでのパフォーマンスを定量化するために、 4つのタスクで構成される小規模な評価データセットを作成しました。「このミームが面白い理由を説明してください。 」という質問によるミーム解釈、 「このようなものはどのように作ればよいでしょうか。 」という質問によるレシピ生成、 「これのためのプロフェッショナルな広告の下書きを手伝ってください。 」というプロンプトによる広告作成、 「この画像について美しい詩を作ってもらえますか。 」という詩の作成です。合計で100枚のさまざまな画像を収集し、各タスクに25枚の画像を割り当てました。人間の評価者に、モデル生成が要求を満たしているかどうかを判断してもらいました。結果をBLIP-2 (Li et al., 2023)と比較し、結果を表1に示します。ミーム解釈、詩の作成、広告の作成において、BLIP-2は要求を満たすのに苦労しています。レシピ生成では、BLIP-2は25件中4件で成功しています。対照的に、MiniGPT-4は、レシピ、広告、詩の生成において、約80%のインスタンスで要求に対応しています。さらに、MiniGPT-4 は、 25 件中 8 件でミーム内の難しいユーモア理解を正しく理解します。

画像キャプションCOCOキャプションベンチマークにおけるMiniGPT-4の性能を評価し、BLIP-2 (Li et al., 2023)と比較する。本モデルが生成するキャプションは、典型的には豊富な視覚的詳細を含んでいる。そのため、従来の類似性に基づく画像キャプション評価指標では、本モデルの正確な評価が困難である。この点について、本モデルは、以下の方法で性能を評価する。

表1: 高度な視覚言語タスクにおける定量的な結果。MiniGPT-4は高いパフォーマンスを示し、要求の65%に正常に応答しました。

	ミーム	レシピ	広告	詩	平均
ブリップ-2	0/25	4/25	1/25	0/25	5/100
ミニGPT-4	8/25	18/25	19/25	20/25	65/100



(a) ミームの説明 (b)ウェブサイトの作成

図4: BLIP-2からのモデル生成、BLIP-2は第2段階のデータ（BLIP-2 FT)を微調整した、
ローカルナラティブデータを使用して微調整された MiniGPT-4 (MiniGPT-4 LocNa)、MiniGPT-4
Q-Former なしのモデル (MiniGPT-4 No Q-Former)、および MiniGPT-4。

ローカルナラティブデータセットで微調整されたオリジナルのMiniG
バリアントとの質的比較（生成されたキャプションがすべてのグラン
ドトゥルースキャプション情報をカバーしているかどうかを確認）
言語モデル FlanT5 XXL を搭載したオリジナルの BLIP-2。
ChatGPTと詳細は付録A.3を参照。表2の結果は、MiniGPT-4が優れていることを示している。
BLIP-2は、実際の視覚的オブジェクトとより一致するキャプションを生成する。
関係性。MiniGPT-4の成功率は66.2%で、BLIP-2よりもかなり正確である。
わずか27.5%しか達成していません。従来のVQAタスクの詳細な評価については、付録A.2をご覧ください。

4.3第2段階の微調整に関する分析

オリジナルのMiniGPT-4、Q-FormerなしのMiniGPT-4バリアント（MiniGPT)との比較。第2段階の微調整の有効性。第2段階の微調整後に事前学習されたモデルのみの利用。

ローカルナラティブデータセット（MiniGPT-4 LocNa)で調整されたBLIP-2（FlanT5 XXL)は、最初の事前学習段階で微調整されたが、繰り返しの単語や文の発生などの失敗
につながる可能性がある。
言語モデル FlanT5 XXL を搭載したオリジナル BLIP-2。
第二段階の微調整プロセスを経て、このことが図5で確認できる。MiniGPT -4

表2: COCOキャプションの評価。
ChatGPTは生成されたキャプションが
すべての視覚的オブジェクトと関係を網羅する
真実のキャプション。

	BLIP-2	ミニGPT-4
正確性	1376/5000	3310/5000
割合	27.5%	66.2%

表3: 詳細キャプションの失敗率
詩生成タスクの前後
第二段階の微調整。微調整段階
生成障害を大幅に削減します。

故障率	詳細なキャプション	詩
ステージ2の前	35%	32%
ステージ2以降	2%	1%

図5: 第二段階の微調整前のMiniGPT-4では完全なテキストを出力できなかった。微調整後は生成が改善された。

図 6: MiniGPT-4 の制限の例。MiniGPT-4 は存在しないテーブルクロスを幻覚し、ウィンドウを正しく見つけることができません。

第二段階の微調整前は不完全なキャプションが生成されます。しかし、第二段階の微調整後、MiniGPT-4は完全で流暢なキャプションを生成できるようになります。本節では、第二段階の微調整アプローチの重要性と有効性を検証します。

第 2 段階のファインチューニングの影響を定量化するために、 COCO テスト セットから 100 枚の画像をランダムにサンプリングし、詳細な説明生成と詩の作成という 2 つのタスクでモデルのパフォーマンスを調査しました。使用したプロンプトは、「画像を詳細に説明してください。」と「この画像について美しい詩を書いていただけますか。」でした。これらのタスクは、第 2 段階のファインチューニングの前後の両方のモデルで実行されました。各段階で、モデルの失敗生成数を手動でカウントしました。結果を表 3 に示します。第 2 段階のファインチューニング前は、生成された出力の約 1/3 がグランド トゥルースのキャプションまたは詩と一致しませんでした。対照的に、第 2 段階のファインチューニング後のモデルでは、両方のタスクで 100 枚のテスト画像のうちの失敗ケースが 2 件未満です。

これらの実験結果は、第二段階のファインチューニングによって生成される出力の品質が大幅に向上することを示しています。第二段階のファインチューニング前後のモデル生成の定性的な例を図5に示します。

オリジナルのBLIP-2は、第二段階のデータの恩恵を受けることができるでしょうか ?本研究では、 MiniGPT-4と同様に、 BLIP-2 (Li et al., 2023)を第二段階のデータで微調整し、 MiniGPT-4と同様の高度な能力を獲得できるかどうかを検証します。微調整されたBLIP-2はBLIP-2 FTと表記されます。MiniGPT-4はBLIP-2と同じ視覚モジュールを使用していますが、BLIP-2は言語モデルとしてFlanT5 XXL (Chung et al., 2022)を使用しています。これは、MiniGPT-4モデルで使用されているVicuna (Chiang et al., 2023)モデルほど強力ではありません。モデルの高度な能力を評価するために、同じプロンプトを使用しています。

定性的な結果は図4、13、14に示されています。BLIP -2 FTは依然として短い応答を生成し、ミームの説明やウェブサイトのコーディングなどの高度なタスクには一般化できないことがわかります（図4 ）。
私たちの研究結果は、BLIP-2 の比較的弱い言語モデル FlanT5 XXL ではこのような小規模なデータセットから得られる恩恵が少ないことを示唆しており、VLM システムにおけるより高度な LLM の有効性を強調しています。

第二段階 :Localized Narrativesを用いたデータセット Localized Narratives (Pont-Tuset et al., 2020)は、注釈者が画像を記述すると同時に、対応する領域をローカライズする詳細な画像記述データセットです。ここでは、第二段階において、自ら収集したデータセットをLocalized Narrativesデータセットに置き換えて、モデルの性能をテストします。このモデルは、

表4: 建築設計におけるアブレーション

モデル	AOK-VQA GQA	
MiniGPT-4 (a)	58.2	32.2
MiniGPT-4（Q-Formerなし）(b)	56.9	33.4
MiniGPT-4 + 3レイヤー (c) MiniGPT-4	49.7	31.0
+ Finetune Q-Former	52.1	28.0

表5:幻覚評価

	CHAIRi平均長さ	
ブリップ2	1.3	6.5
MiniGPT-4（ショート）	7.2	28.8
MiniGPT-4（ロング）	9.6	175

MiniGPT-4 LocNaとして。図4、13、14の定性的な結果は、 MiniGPT - 4 LocNaが長い画像説明を生成できることを示しています（図14）。しかし、生成された出力は単調な表現で品質が低くなっています。さらに、MiniGPT-4 LocNaは、ミームがなぜ面白いのかを説明するなどの複雑なタスクにおいて、オリジナルのMiniGPT-4ほど汎化能力が高くありません（図4a）。このパフォーマンスの差は、Localized Narrativesにおける単調で繰り返しの多い画像説明に起因する可能性があります。

4.4建築設計におけるアブレーション

単一の線形層を使用して視覚特徴をLLMに合わせることの有効性をさらに実証するために、(a) Q- Formerを削除し、VITの出力をVicunaの埋め込み空間に直接マッピングする（つまり、Q-Formerなし）、 (b) 1つの層の代わりに3つの線形層を使用する、(c) ビジョンモジュールでQ-Formerをさらに微調整するなど、さまざまなアーキテクチャ設計で実験を行いました。すべてのバリアントは、元の設計と同じ方法でトレーニングされています。表4のAOK-VQA （Schwenk et al.、 2022）およびGQA （Hudson&Manning、 2019)データセットの結果は、バリアント（a） Q-FormerなしのMiniGPT-4が元の設計と同様のパフォーマンスを示していることを示しています。

図4、13、14に示すこのバリアントの定性的な結果も、同様の高度なスキルを示しています。これは、BLIP-2のQ-Formerが高度なスキルにおいて重要な役割を果たしていないことを示しています。さらに、(b) MiniGPT-4+ 3層と(c) MiniGPT-4 + finetuning Q-Formerのどちらのバリアントも、オリジナルのMiniGPT-4よりもわずかにパフォーマンスが劣っています。これは、限られた学習データ設定において、単一の投影層でビジョンエンコーダと大規模言語モデルを整合させるのに十分であることを示しています。

4.5限界分析

幻覚MiniGPT-4はLLM上に構築されているため、存在しない知識を幻覚するなど、LLMの制限を継承しています。図6の例では、MiniGPT-4は画像に白いテーブルクロスが存在しないにもかかわらず、誤って存在すると識別しています。ここでは、メトリクスCHAIRi （Rohrbach et al., 2018)を使用して、生成の幻覚率を測定し、モデル生成の長さを制御するための2つの異なるプロンプトを使用します。 MiniGPT-4（長文）：この画像をできるだけ詳細に説明してください。 MiniGPT-4（短文）：20語未満で、画像を簡潔かつ正確に説明してください。

表5の結果は、キャプションが長いほど幻覚率が高くなる傾向があることを示しています。例えば、MiniGPT-4（長)は平均175語のキャプションを生成し、幻覚率が高いのに対し、 MiniGPT-4（短)は平均28.8語のキャプションを生成し、幻覚率は低いです。BLIP-2は平均6.5語で、幻覚は少ないものの、表2に示すように、カバーするオブジェクトは少ないです。詳細な画像の説明における幻覚は未解決の問題です。幻覚検出モジュールを備えたAIフィードバックと強化学習を組み合わせることが、潜在的な解決策となる可能性があります。

空間情報理解： MiniGPT-4の視覚認識能力は依然として限られており、空間的な位置特定を区別するのが難しい場合があります。例えば、図6のMiniGPT-4は窓の位置を識別できません。この限界は、空間情報理解用に設計された画像とテキストの対応データが不足していることに起因している可能性があります。RefCOCO (Kazemzadeh et al., 2014)やVisual Genome (Krishna et al., 2017)などのデータセットを用いた学習によって、この問題を軽減できる可能性があります。

5議論

MiniGPT-4はどのようにしてこれらの高度な能力を獲得するのでしょうか？ GPT-4が示す高度な視覚言語能力の多くは、画像理解と言語生成という2つの基礎スキルに根ざした構成スキルとして理解できます。画像ベースの詩作タスクを例に挙げてみましょう。ChatGPTやVicunaのような高度なLLMは、既にユーザーの指示に基づいて詩を作成できます。画像を理解する能力を獲得すれば、訓練データに画像と詩のペアがなくても、画像ベースの詩作タスクへの構成的一般化が可能になります。

MiniGPT-4は、最初の事前学習段階では、画像キャプションデータセットから画像と短い画像説明との相関関係をモデル化することで画像理解を学習します。しかし、これらの画像キャプションデータセットの言語スタイルは、現代のLLM生成の言語スタイルとは異なり、歪んだ言語生成につながり、構成的汎化の成功を妨げています。そこで、言語生成能力を回復するための第2段階の微調整を導入します。2段階学習後のMiniGPT-4は、下書きからのウェブサイトコーディングやミーム解釈など、多くの高度な構成的視覚言語能力に汎化することに成功し、私たちの仮説を検証しました。今後の研究

構成的一般化のメカニズムをさらに深く掘り下げ、それを強化する方法を模索するかもしれません。私たちの研究は、視覚に基づくLLM機能の早期探究として、この分野におけるさらなる研究を促進するものとなることを期待しています。

参考文献

シェアポイント。 https://github.com/domeccleston/sharegpt, 2023年。

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.

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。

Soravit Changpinyo、Piyush Sharma、Nan Ding、Radu Soricut. Conceptual 12m: ウェブスケールの画像テキスト事前学習によるロングテール視覚概念の認識への取り組み。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、pp. 3558–3568、2021年。

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、pp. 18030–18040、2022年。

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.

Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、 Siyuan Zhuang、Yonghao Zhuang、Joseph E. Gonzalez、Ion Stoica、Eric P. Xing。 Vicuna： 2023年3月、90%*のチャットGPT品質でGPT-4を圧倒するオープンソースのチャットボット。URL https://vicuna.lmsys.org 。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、他「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。

チョン•ヒョンウォン、レ•ホウ、シェイン•ロングプレ、バレット•ゾフ、イー•テイ、ウィリアム•フェダス、エリック•リー、シュエジ•ワン、モスタファ•デガニ、シッダールタ•ブラフマー、他。命令を微調整した言語モデルのスケーリング。 arXiv プレプリント arXiv:2210.11416、

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の事前学習。arXivプレプリント arXiv:1810.04805、2018年。

ダニー•ドリース、フェイ•シア、メディ•SM•サジャディ、コーリー•リンチ、アーカンクシャ•チョウドリー、ブライアン•イクター、アイザーン•ワヒド、ジョナサン•トンプソン、クアン•ヴオン、ティアンヘ•ユー 他Palm-e: 具体化されたマルチモーダル言語モデル。 arXiv プレプリント arXiv:2303.03378、2023。

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学習の限界を大規模に調査します。 arXiv プレプリント arXiv:2211.07636、2022。

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、トレバー•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 他
計算最適化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。

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

Shaohan Huang、Li Dong、Wenhui Wang、Yaru Hao、Saksham Singhal、Shuming Ma、Tengchao Lv、Lei Cui、Owais Khan Mohammed、Qiang Liu他「言語だけが必要なわけではない：知覚と言語モデルの整合」arXivプレプリントarXiv:2302.14045、2023年。

ドリュー•A•ハドソン、クリストファー•D•マニング。GQA：実世界視覚推論と構成的質問応答のための新たなデータセット。IEEE/CVFコンピュータビジョン•パターン認識会議論文集、pp. 6700–6709、2019年。

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. 自然言語処理における経験的手法に関する2014年会議論文集(EMNLP)、pp. 787–798、2014年。

Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalantidis、Li-Jia Li、David A Shamma、他「ビジュアルゲノム：クラウドソーシングによる高密度画像アノテーションを用いた言語と視覚の連携」International journal of computer vision、 123:32–73、2017年。

Junnan Li、Dongxu Li、Caiming Xiong、Steven Hoi. Blip:統合的な視覚•言語理解•生成のための言語•画像のブートストラッピング事前学習。国際機械学習会議、pp. 12888–12900。PMLR、2022年。

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

OpenAI. chatgptの紹介。https://openai.com/blog/chatgpt、2022年。

OpenAI。GPT-4技術レポート、2023年。

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. Advances in Neural Information Processing Systems, 24, 2011.

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.Advances in Neural Information Processing Systems, 35: 27730–27744, 2022.

Jordi Pont-Tuset、Jasper Uijlings、Soravit Changpinyo、Radu Soricut、Vittorio Ferrari.視覚と言語をローカライズされた物語で繋ぐ. Computer Vision–ECCV 2020: 第16回ヨーロッパ会議、英国グラスゴー、2020年8月23日～28日、議事録、パートV 16、pp. 647–664. Springer、 2020年。

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever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。

コリン•ラフェル、ノアム•シャジーア、アダム•ロバーツ、キャサリン•リー、シャラン•ナラン、マイケル•マテナ、ヤンチー•ゾウ、ウェイ•リー、ピーター•J•リウ。統合テキスト•ツー•テキスト変換を用いた転移学習の限界の探究。機械学習研究ジャーナル、21(1):5485–5551、2020年。

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。オブジェクト画像キャプションにおける幻覚。arXivプレプリントarXiv:1809.02156、2018年。

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•パブリック、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Bloom: 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、2022a。

Teven Le Scao、Angela Fan、Christopher Akiki、Ellie Pavlick、Suzana Ilic、Daniel Hesslow、Roman Castagne、Alexandra Sasha Luccioni、Fran c¸ ois Yvon、Matthias Galle他「Bloom: 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, 2022b.

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4億点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。

ダスティン•シュウェンク、アプールフ•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のベンチマーク。European Conference on Computer Vision、pp. 146–162。Springer、2022年。

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）、pp. 2556–2565、2018年。

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。

D´ dac Sur´ s、 Sachit Menon、Carl Vondrick。Vipergpt: Python実行による視覚的推論推論。arXivプレプリントarXiv:2303.08128、2023。

Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang、B. 橋本辰則。スタンフォード アルパカ: 指示に従うラマ モデル。 https://github.com/tatsu-lab/stanford_alpaca, 2023年。

Anthony Meng Huat Tiong、Junnan Li、Boyang Li、Silvio Savarese、Steven CH Hoi。プラグアンドプレイVQA：大規模な事前学習済みモデルをゼロ学習で結合することによるゼロショットVQA。arXivプレプリントarXiv:2210.08773、2022年。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothee´ Lacroix、Baptiste Roziere、Naman Goyal、Eric Hambro、Faisal Azhar他「Llama：オープンで効率的な基礎言語モデル」arXivプレプリントarXiv:2302.13971、2023年。

マリア•ツィンポケリ、ジェイコブ•L•メニック、セルカン•カビ、SMエスラミ、オリオール•ヴィニャルズ、フェリックス•ヒル。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年。

ジェイソン•ウェイ、イー•テイ、リシ•ボンマサニ、コリン•ラフェル、バレット•ゾフ、セバスチャン•ボルゴー、ダニ•ヨガタマ、マールテン•ボスマ、デニー•チョウ、ドナルド•メッツラー、エド•H•チー、橋本達規、オリオール•ヴィニャルズ、パーシー•リャン、ジェフ•ディーン、ウィリアム•フェダス。大規模学習モデルの新たな能力。機械学習研究のトランザクション、2022年。ISSN 2835-8856。 URL https://openreview.net/forum?id=yzkSU5zdwD.調査認証。

Chenfei Wu、Shengming ying 、Weizhen Qi、Xiaodong Wang、Zecheng Tang、および Nan Duan。ビジュアルチャット: ビジュアル基盤モデルを使用して会話、描画、編集します。 arXiv プレプリントarXiv:2303.04671、2023。

アントワーヌ•ヤン、アントワーヌ•ミエシュ、ヨゼフ•シヴィック、イヴァン•ラプテフ、コーデリア•シュミット。固定双方向言語モデルによるゼロショットビデオ質問応答。arXivプレプリント arXiv:2206.08155, 2022.

Zhengyuan Yang*、Linjie Li*、Jianfeng Wang*、Kevin Lin*、Ehsan Azarnasab*、Faisal Ahmed*、 Zicheng Liu、Ce Liu、Michael Zeng、Lijuan Wang。 Mm-react: マルチモーダルな推論とアクションのためにチャットグループを促します。 2023年。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona Diab、Xian Li、Xi Victoria Lin他。Opt: オープンな事前学習済みTransformer言語モデル。arXivプレプリントarXiv:2205.01068、2022年。

Deyao Zhu、Jun Chen、Kilichbek Haydarov、Xiaoqian Shen、Wenxuan Zhang、Mohamed Elho- seiny. Chatgptが質問し、blip-2が回答：視覚的記述の強化に向けた自動質問。arXivプレプリント arXiv:2303.06594, 2023.

付録

A.1より定性的な結果

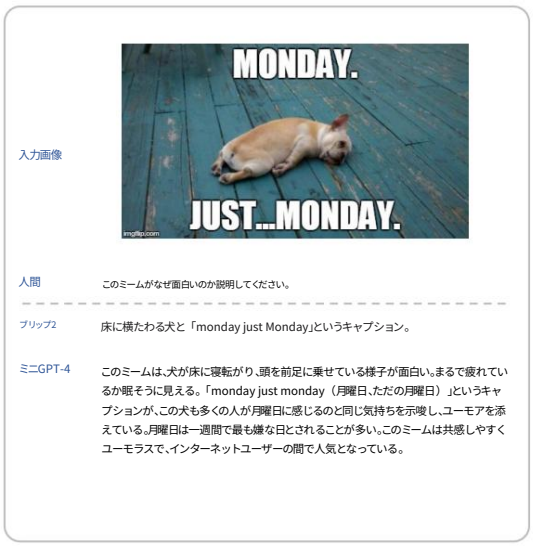


図7: 広告プロモーション

図8: 事実検索

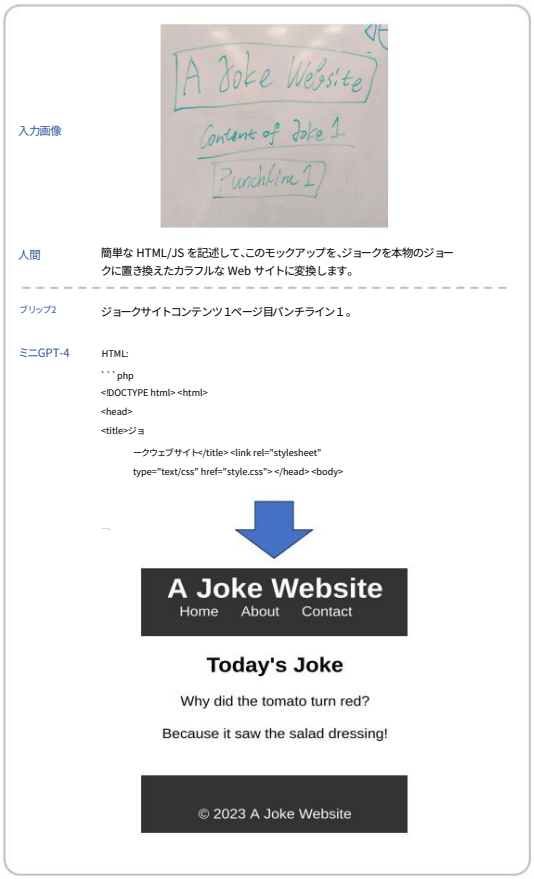


図9: ウェブサイトの作成

図10: 詩の書き方

図11: 料理レシピ生成

図12: 植物栽培

A.2従来のVQAベンチマークによる評価

この研究の目的は、GPT-4で実証された驚くべきマルチモーダル機能を再現することである。
詳細な画像の説明を生成したり、手描きの下書きからウェブサイトを作成したりといった機能があります。
高度な視覚言語スキルの最も重要な要素である方法論を強調する
MiniGPT-4は意図的に最小限に抑えられています。例えば、学習可能なモデルの容量は限られています（
MiniGPT-4はわずか500万ペアで訓練されているのに対し、BLIP-2は129ペアで訓練されている。
数百万の画像とテキストのペア。このような簡略化されたアプローチでは、
従来のベンチマーク。これは私たちの主な目的ではありませんが、VQAの定量分析を提供します。
データセットA-OKVQA（多肢選択式） （Schwenk et al.、 2022)およびGQA （Hudson&Manning、 2019）。
さらに、従来のベンチマークでMiniGPT-4の潜在能力を示すために、
単純なアブレーション研究。ここでは、LoRAを用いてLLMを単純に解凍する（Hu et al., 2021）
VQAv2、OKVQA、A-OKVQAデータセットからより多くのトレーニングデータを取り込み、
第二段階の微調整。表6の結果は、オリジナルのMiniGPT-4がBLIP-2より遅れていることを示している。
学習能力と訓練データを増やすだけで、
大幅なパフォーマンス向上が見られ、これは私たちの期待を裏付けるものです。私たちのモデルの
従来の視覚ベンチマークのパフォーマンスは、慎重に設計されたトレーニングによって向上できる。
戦略（例 :データセットのサンプル比率、学習率スケジュールなど）、より多くのトレーニングデータ/データセット、および
追加の学習可能なパラメータ。従来の視覚ベンチマークのパフォーマンスを向上させることは
このプロジェクトの目的のため、この側面は将来の研究のために留保します。

モデル	トレーニングデータ	AOK-VQA	GQA
ブリップ2	1億2900万の画像とテキストのペア	80.2	42.4
ミニGPT-4	500万個の画像とテキストのペア	58.2	32.2
MiniGPT-4 (Finetune Vicuna)	500 万個の画像とテキストのペア	67.2	43.5

表6: BLIP-2とMiniGPT-4のパフォーマンス比較

A.3キャプション評価の詳細

ChatGPTを使用して、ベースラインモデルがすべてのオブジェクトと視覚的特徴をカバーしているかどうかを判断します。グランドトゥルースのキャプションに示されている関係。COCO評価データセットでは、ランダムに1つの真実のキャプションを選択し、それを参照キャプションとして扱います。次のプロンプトを適用します。評価を実行します。

画像キャプション1 「{ground-truth caption}」が1つあり、もう1つの画像キャプション2があります「{比較キャプション}」。画像キャプション2は、表示されているすべてのオブジェクトと視覚的な関係を網羅していますか？画像キャプション1には何がありますか? 説明なしで「はい」または「いいえ」のみで答えてください。

A.4より定性的なアブレーション結果

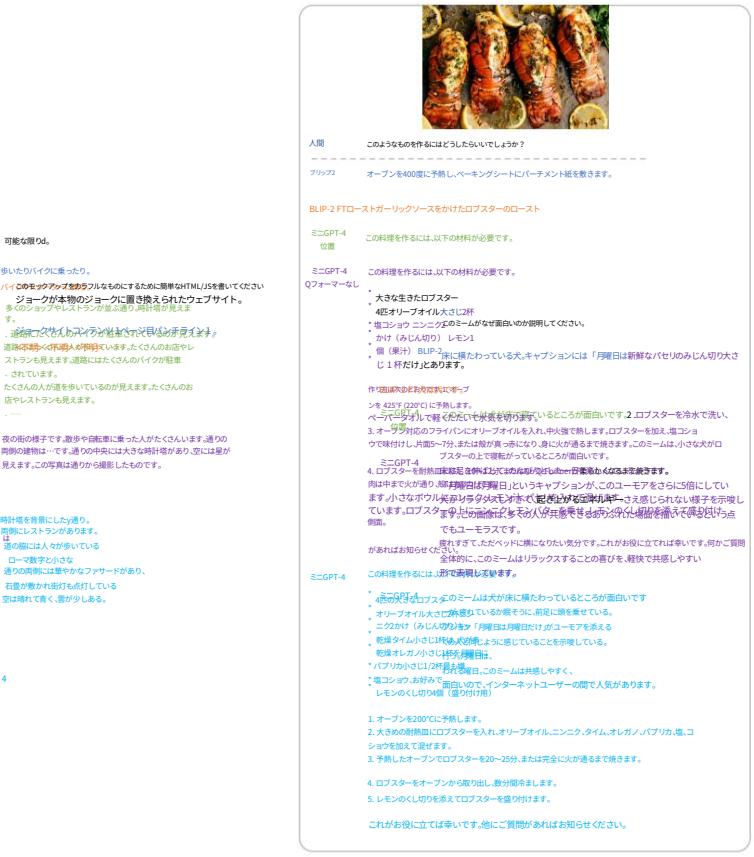


図14: アブレーション研究の詳細な説明

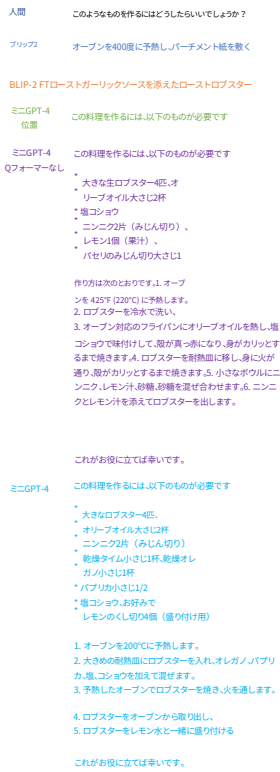


図13: Q-Formerなしの MiniGPT-4 のローカル研究

Q-Formerなし（MiniGPT-4 No Q-Former）は、事前トレーニングされたビジョンとFlanT5 XXL)をデータセット(BLIP-2 FT)で微調整し、

オリジナルの MiniGPT-4、Q-Former なしの MiniGPT-4 バリアント (MiniGPT-4 No Q-Former)、Local Narrative データセットで微調整された BLIP-2 (FlanT5 XXL)、データセットで微調整された BLIP-2 (FlanT5 XXL) (BLIP-2 と言語モデル FlanT5 XXL) の間の比較。