Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos

2025, Haobo Yuan, Xiangtai Li, et al.,

論文：<https://arxiv.org/abs/2501.04001>

Project Page: <https://lxtgh.github.io/project/sa2va>

Github : <https://github.com/bytedance/Sa2VA>

Huggingface Model: <https://huggingface.co/ByteDance/Sa2VA-4B>　（旧バージョン）

Huggingface Model: <https://huggingface.co/collections/ByteDance/sa2va-model-zoo> (最新バージョン)

（TikTokなどを運営するByteDanceが出したもの, apache-2.0）

Abstract

本研究では、画像と動画の両方を緻密にグラウンディングした理解を実現する初の統合モデルであるSa2VAを紹介します。特定のモダリティやタスクに限定されることが多い既存のマルチモーダル大規模言語モデルとは異なり、Sa2VAは、ワンショットの指示チューニングを最小限に抑えながら、参照セグメンテーションや会話など、幅広い画像および動画タスクをサポートします。Sa2VAは、基礎的な動画セグメンテーションモデルであるSAM-2と、高度な視覚言語モデルであるLLaVAを組み合わせ、テキスト、画像、動画を共有LLMトークン空間に統合します。LLMを使用することで、Sa2VAはSAM-2が正確なマスクを生成するための指示トークンを生成し、静的および動的な視覚コンテンツの両方をグラウンディングしたマルチモーダル理解を可能にします。さらに、複雑な動画シーンにおける72,000以上のオブジェクト表現を含む自動ラベル付けデータセットであるRef-SAVを導入し、モデルのパフォーマンスを向上させるように設計されています。また、Ref-SAVデータセット内の2,000個の動画オブジェクトを手動で検証し、複雑な環境での参照動画オブジェクトセグメンテーションのベンチマークを実施します。実験では、Sa2VAが複数のタスク、特に参照動画オブジェクトのセグメンテーションにおいて最先端の性能を達成していることが示され、複雑な実世界アプリケーションへの応用可能性が強調されています。コード、データセット、モデルは、さらなる研究のために公開されています。

図1：提案するSa2VAの機能の図解。(a)。ビデオが与えられると、Sa2VAは参照オブジェクトをセグメント化し、シーン全体を理解できます。(b)。Sa2VAは、画像会話、ビデオ会話、画像参照セグメンテーション、ビデオ参照セグメンテーション、およびシングルショットの指示チューニングによるグラウンデッドキャプション生成をサポートします。(c)。Sa2VAは、GLaMM [66]やOMG-LLaVA [99]などの既存のMLLMと比較して、複数画像、ビデオ参照セグメンテーション、およびチャットベンチマークで優れた結果を達成しています。

カレンダー が含まれている画像

AI 生成コンテンツは誤りを含む可能性があります。

1. Introduction  
   マルチモーダル大規模言語モデル（MLLM）は、大規模言語モデル（LLM）[73, 33, 86]の急速な発展に伴い、大きな進歩を遂げてきました。MLLMは、視覚的質問応答（VQA）[1, 70]、物語生成[87, 26, 105]、対話型編集[76, 39, 30]など、様々な画像・動画レベルのタスクでその恩恵を受けています。  
   重要な方向性の一つは、ビデオコンテンツをきめ細かく理解することです。これには、言語記述を用いたピクセルのセグメント化と追跡、ビデオ内の視覚的なプロンプトに対するVQAの実行などが含まれます。特に、図1(a)に示すように、インタラクティブモードでビデオを再生する際にユーザーがループに参加できるため、プロンプト可能なビデオのファインゲイン分析を実現したいと考えています。これは、短編ビデオ編集[6, 75, 58]、ロボットナビゲーション[23, 24, 21]、監視分析[63]など、さまざまなアプリケーションにつながります。  
   しかし、最先端のビデオ知覚モデル [11, 48, 98, 67] もビデオ MLLM [95, 50, 55, 43] もこれを実現できません。前者は制約された意味概念によって制限されており、オープンエンド機能（ビデオ VQA または名前付きビデオチャットタスク）が欠けています。たとえば、最新の基礎モデルである SAM-2 [67] は、プロンプト可能なセグメンテーションと追跡を実現できます。ただし、言語表現の理解やビデオ会話などのテキスト認識タスクは実行できません。一方、ビデオ MLLM [51, 55, 42, 50] は長いビデオを理解し、VQA を実行できます。たとえば、最新の LLaVA [42] はビデオ VQA で優れた結果を達成できますが、知覚タスクを実行できず、視覚プロンプトも理解できません。いくつかの研究 [99, 18, 85] では、知覚モデルと MMLM の組み合わせが検討されています。しかし、ほとんどの研究は画像データセットを探索するか、特定の動画理解タスクを解決しようとしています。私たちの知る限り、両方の利点を活用できる研究は存在しません。  
   以上の分析を通して、本研究では、SAM-2 [67] と LLaVA のような MLLM [53] という 2つの視覚基盤モデルを 1つのフレームワークに統合するための第一歩を踏み出します。このようにして、私たちのモデルは SAM-2 の時空間認識能力と MLLM のオープンエンド能力を継承するだけでなく、きめ細かいトレーニング データから得られる追加の知識を学習することによる利点も享受します。ただし、この目標を達成するには、3つの主要な課題に対処する必要があります。(1) タスク定式化: 特にマルチモーダル入力に対して、ワンショット トレーニング設定でさまざまなタスクを効果的に定式化する方法。特にビデオ認識タスクとビデオ VQA タスクの場合。以前の研究 [85, 18] はそれらの一部しか解決していません。(2) パフォーマンスのバランス: MLLM の言語能力を犠牲にすることなく、強力な参照視覚理解能力を確保するなど、タスク間の競合をどのように解決するか。以前の研究 [99, 82] では、グラウンディング タスクを実行すると会話タスクのパフォーマンスが低下することがわかっていることに注意してください。 (3) 知識共有：SAM-2とMLLMの事前学習済み知識を活用し、堅牢で統合されたモデルを構築する方法。SAM-2は10億以上のマスクを用いて学習されており、ほとんどのMLLMは大規模な学習ペアを用いて学習されている。  
   SAM-2とLLaVAのようなMLLMを統合し、画像と動画の両方に対する統一されたグラウンデッドな理解を実現する初のモデル、Sa2VAを紹介します。画像と動画の両方を対象に、参照セグメンテーション、視覚的質問応答（VQA）、グラウンデッドな会話生成（GCG）など、様々なタスクをワンショットの視覚指示チューニング形式に定式化します。重要な知見は、LLMの柔軟なトークン長処理を活用し、入力画像、動画、視覚プロンプトをすべて視覚トークンとして扱い、追加の特別な設計を必要とせずに実現することです。共同学習を通して、グラウンディングタスクと会話タスク間の競合を効果的に解決できることを実証します。MLLMをエージェントまたはコネクタとして視覚専門家を呼び出す既存のアプローチとは異なり、本モデルはエンドツーエンドで学習され、モデルとデータセットの両方でスケーラビリティを実現します。さらに、SAM-2のデコーダーとメモリモジュールをフリーズする分離設計を採用することで、SAM-2の知覚および追跡機能を維持します。この設計により、私たちの方法はプラグアンドプレイ モジュールにもなり、モデルを最新の MLLM で更新できるようになります。  
   さらに、SA-Vデータセット[67]に基づく自動パイプラインによってキュレーションされた、挑戦的な参照ビデオセグメンテーションデータセットRef-SAVを紹介します。これは、既存の参照ビデオセグメンテーションデータセットが、オクルージョンの少ない小規模で短いクリップに限定されているという経験的観察に基づいています。このデータセットを用いていくつかの既存モデルをベンチマークし、改善とさらなる探究の余地が十分にあることがわかりました。  
   Sa2VAは、画像や動画データセットを含むマルチデータセットの共学習によって学習されています。Sa2VAは、図1(c)に示すように、余計な機能は一切追加することなく、6つの参照画像および動画セグメンテーションデータセットで最先端の結果を達成しながら、従来のMLLMと比較して強力な画像および動画チャット機能を維持しています。Ref-SAVでは、ゼロショットテスト設定で従来の方法の15%以上を達成し、提案するRef-SAVトレーニングセットを使用することでさらに強力な結果を達成しました。図1(c)に示すように、私たちの研究は、画像と動画の統一的で密度が高く、根拠のある理解のための新しい強力なベースラインを構築します。全体として、私たちの貢献は次のとおりです。  
   ・視覚トークンの観点から、画像チャット、画像参照セグメンテーション、ビデオチャット、参照ビデオオブジェクトのセグメンテーション、グラウンデッドキャプション生成といった複数のタスクを再考し、これらすべてのタスクを単一の命令チューニングプロセスとして定式化します。  
   ・我々は、SAM-2とLLaVA類似モデルを1つのモデルに統合し、時空間グラウンデッドな理解を実現する初のシンプルなフレームワークであるSa2VAを開発しました。共有SEGトークンを用いたシンプルな分離フレームワークを導入します。また、SAM-2の知識を再利用したシンプルなマスク追跡手法も提案します。  
   ・参照動画オブジェクトセグメンテーションの難易度の高いベンチマークであるRef-SAVを導入し、各例を手動でチェックします。このベンチマークでは、より厳しいオクルージョン、長いテキスト入力、モーションブラーなどを考慮しています。また、Ref-SAVトレーニングデータセットを構築するためのシンプルなデータアノテーションパイプラインも開発し、このトレーニングデータセットがRef-SAVにおけるモデルのパフォーマンスを向上させることを明らかにしました。  
     
   表1：様々な代表的モデルの能力比較。本手法は様々なタスクとモダリティに対応しています。Sa2VAは、ビデオ上のこれらのインタラクティブ機能を活用することで、図1(a)および(b)に示すように、ビデオ内で複数のプロンプト可能なタスクを実行できます。  
   テーブル

   AI 生成コンテンツは誤りを含む可能性があります。
2. Related Work  
   **Multi-modal Large Language Models.**先行研究 [44, 45, 31] では、より優れたマルチモーダル融合手法と特徴抽出器が検討され、特に視覚言語タスク向けのさまざまな融合アーキテクチャが設計されています。高度な LLM フレームワーク [4, 73, 33] により、LLM 上のマルチモーダル命令チューニング [2, 71, 53, 8] が主要なパイプラインになります。その後、さまざまなベンチマーク [32, 59, 49, 19, 72] が登場し、より優れたパフォーマンスを実現するためのより優れたソースとして機能します。これは、データが現在の MLLM で中心的な役割を果たしていることを意味します。代表的な研究の 1 つである LLaVA [53] は、視覚特徴を視覚トークンとして扱います。著者らは、広範な VQA タスクを統合するための統一データ形式も提供しています。その後、いくつかの研究 [92] では、LLaVA の視覚入力を強化するために、より強力な視覚的手がかりが検討されています。一方、いくつかの研究 [96, 94, 68, 97, 14, 15, 51, 95, 65, 29, 40] では、LLaVA を視覚的グラウンディング、検出、セグメンテーション、ビデオ VQA に適応させるための追加コンポーネントが追加されています。最近では、画像、ビデオ、およびマルチ画像分析を 1 つのフレームワークに統合するという新しい傾向もあります [42, 55]。 LLaVA-OneVision [42] は、4つの異なる入力ソースを処理する単一のモデルを設計しています。視覚知覚では、いくつかの研究 [84, 48, 67] も、マルチデータセットとマルチタスクの共同トレーニングを検討しています。 SAM-2 [67] は、画像とビデオの共同インタラクティブ セグメンテーション用の統合システムを設計しています。私たちのモデル Sa2VA は、SAM-2 を現在の VLM モデルに統合して 1つのエンドツーエンド モデルを作成し、セグメンテーション、チャット、キャプションを含む高密度でグラウンディングされた理解のために画像とビデオを統合することを目指しています。  
   **Referring Segmentation.**このタスクは、言語記述によって駆動される特定のマスク（画像入力の場合）または追跡されたマスク（ビデオ入力の場合）を出力することを目的としています。以前の研究 [91, 52, 38, 13, 80, 102] では、より良いパフォーマンスを達成するためにさまざまな融合モジュールが検討されています。次に、いくつかのより強力なモデル [84] は、ビデオの統一されたセグメンテーションと追跡を実現するために、トランスフォーマーベースの方法を採用しています。LLMを搭載した最近のいくつかの研究 [40, 99, 83, 65] では、推論参照やマスクとキャプションの共同生成など、より複雑な参照タスクが提案されています。たとえば、LISA [40] は推論ベースのセグメンテーションを行います。次に、GLaMM [66] は新しいデータセットに注釈を付け、領域レベルのキャプションとセグメンテーションタスクを提案しています。一方、最近のいくつかの研究では、参照セグメンテーションと会話に関する共同指示チューニングが検討されています [85, 101, 8]。私たちの方法は、SAM-2 [67]の助けを借りてこれらの研究をビデオ領域に拡張し、画像/ビデオ参照タスクとVQAタスクの両方でより強力なパフォーマンスを維持しています。  
   **Video Segmentation and Grounding.**  
   現在のビデオセグメンテーション手法 [47, 34, 106] は、閉じたセット内のピクセルをセグメント化して追跡することに重点を置いています。いくつかの研究 [25, 104] では、オープン語彙の設定が検討されています。ただし、概念は LLM の知識空間と比較するとまだ限られています。ビデオグラウンディングについては、最近の研究 [28] で、ビデオとオーディオの共同理解について LLM が検討されています。最近の研究 VISA [85] では、ビデオオブジェクトのセグメンテーションの推論も検討されています。ただし、これはビデオセグメンテーションデータセットのみでトレーニングされており、エンドツーエンドのトレーニングがないため、他のタスクでの能力とスケールアップの可能性が制限されています。私たちのモデル Sa2VA は、静的 (画像) と動的 (ビデオ) の両方の視覚コンテンツのきめ細かい時空間モデリングをサポートし、さまざまなタスクで優れたパフォーマンスを実現します。
3. Method
   1. UnifyingMulti-task Representations  
      様々な画像・動画理解タスクに対応する統一モデルの開発は、空間情報と時間情報が異なるため困難です。この課題に対処するため、まずタスクの定式化を再検討し、Sa2VA開発の基盤となる統一的な表現を提案します。  
      **Referring Image/Video Object Segmentation.**画像参照セグメンテーションでは、入力テキストトークン Ti ∈ RN× D が与えられると、モデルは入力画像 Ii ∈ RH× W× 3 を処理し、テキスト記述と一致するバイナリマスク Mo ∈ RH× W を生成します。ここで、N と D はテキストトークンの数と次元です。ビデオオブジェクト参照セグメンテーションでは、モデルは入力ビデオ Vi ∈ RT× H× W× 3 を受け取り、バイナリ時空間マスク（マスクレット） Mo ∈ RT× H× W を出力します。ここで、T、H、W はビデオフレーム番号、高さ、幅です。  
      **Image/Video Chat and Grounded Caption Generation.**画像およびビデオチャットタスクの場合、入力テキストトークンTiと対応する画像IiまたはビデオViが与えられると、モデルは回答テキストToを生成します。グラウンデッドキャプション生成タスクの場合、モデルはマスクMoと位置合わせされたテキストToを同時に出力します。  
      **Visual Prompt Understanding Tasks.**  
      視覚プロンプト理解タスクの場合、テキストトークンTiと画像Iiに加えて、モデルは追加の視覚プロンプトトークンV Pi（画像上のボックスまたはポイント）も入力として受け取り、対応するマスクMoと整列したテキスト回答Toを出力します。  
      **Unified Task Representation.**  
      既存の研究では、上記のタスクに対して、特定の設計モデル、あるいは部分的に統合されたモデル（タスクごとに異なるモデルの重み付けを用いる）を用いて対処しています。本研究では、LLMを活用して様々なビジュアルトークンを管理することで、前述のすべてのタスクをワンショットの命令チューニングプロセスに統合できると主張します。全体的なプロセスは以下のように定式化できます。  
        
      入力トークンはタスクによって異なります。チャットのみのタスクの場合、モデルはテキストトークンToのみを出力します。参照セグメンテーションタスクの場合、モデルはマスク（画像）またはマスクレット（動画）Moを出力します。グラウンデッドキャプション生成タスクの場合、モデルはToとMoの両方を出力します。モデルは主に画像/動画、テキスト、マスクを入力と出力として処理するため、これらのタスクを単一のフレームワークに統合し、モデルをエンドツーエンドでトレーニングすることが合理的です。このアプローチにより、単一のモデルで前述のすべてのタスクを効果的にサポートできます。
   2. Sa2VA Framework  
      Sa2VA の全体的なフレームワークを図 2 に示します。これは、MLLM モデルと SAM-2の2つのコンポーネントで構成されています。  
      **Pre-trained MLLMs.**  
      我々は、1つのビジュアルエンコーダー、1つのビジュアル投影層(MLP)、そして1つのLLMを含む事前学習済みのMLLMを採用する。ビジュアルエンコーダーは、入力画像、動画、サブ画像を受け取り、視覚的特徴を出力する。ビジュアル投影層は、特徴をトークンにマッピングする。これらのトークンは、入力テキストトークンとともに、LLMの入力として機能し、LLMはそれらに基づいてテキストトークン予測を生成する。先行研究[46, 76, 100, 114]で見られるように、事前学習済みのMLLMを利用することで、その強力な機能を活用する。画像会話タスクと動画会話タスクの両方において、元の事前学習済みMLLM（例：InternVL2 [14]、Qwen-VL [2]）と同じパイプラインを使用する。  
      **Decoupled Design.**  
      事前学習済みのMLLMに加えてSAM-2 [77] を追加します。SAM-2の出力トークン（視覚特徴やデコーダー出力）はLLMに取り込まないことに注意してください。この設計選択には3つの理由があります。第一に、追加の計算コストをかけずに組み合わせを簡素化することを目指しています。第二に、MLLMとSAM-2は整合していないため、余分なトークンを追加するとMLLMからの知識継承が妨げられます。第三に、この設計により、MLLMコミュニティの急速な進歩に合わせて、事前学習済みの進化型MLLMを活用する柔軟なフレームワークに完全に変換できます。  
      **Tuning SAM-2 Decoder via ‘‘[SEG]’’ Tokens.**  
      SAMを利用した以前の研究[46, 100]と同様に、SAM-2のデコーダーをMLLMに接続するために特別なトークン「[SEG]」を使用します。「[SEG]」トークンの隠れ状態は新しい種類のプロンプトとして機能し、SAM-2のデコーダーに送られ、セグメンテーションマスクにデコードされます。「[SEG]」の隠れ状態は、SAM-2の新しい時空間プロンプトとして考えることができます。SAM-2は、この時空間プロンプトに基づいて、画像や動画内の関連するオブジェクトマスクをセグメント化します。トレーニング中、SAM-2デコーダーは時空間プロンプトを理解するように調整でき、勾配は「[SEG]」トークンを介してMLLMにバックプロパゲーションされ、トレーニングデータセットから知識を得ることができます。  
      **Utilizing SAM-2 knowledge for mask tracking.**Ref-VOSタスク向けに、公開ベンチマークで優れた結果を達成するためのシンプルなフレームワークを開発しました。具体的には、まず動画からキーフレーム（最初の数フレーム）を抽出します。これらのキーフレームはMLLM（Multi-Length Motion Model）処理によって処理されます。MLLMは、キーフレームのマスクを生成するためのSAM-2プロンプトとして「[SEG]」トークンを生成します。次に、キーフレームの特徴によってエンコードされたメモリを利用して、残りのフレームのマスクを生成します。アルゴリズム1にデフォルトの推論アルゴリズムを示します。  
      ダイアグラム

      AI 生成コンテンツは誤りを含む可能性があります。  
      図2：提案するSa2VAモデル。このモデルはまず、入力テキスト、視覚プロンプト、画像、動画をトークン埋め込みにエンコードします。これらのトークンは、大規模言語モデル（LLM）によって処理されます。出力テキストトークンは、「[SEG]」トークンと関連する言語出力を生成するために使用されます。SAM-2デコーダーは、SAM-2エンコーダーから画像と動画の特徴と「[SEG]」トークンを受け取り、対応する画像と動画のマスクを生成します。
   3. Ref-SAV Dataset and Benchmark  
      **Data Annotation Pipeline.**  
      SA-Vデータセット[77]の参照オブジェクトのテキスト表現を生成するための自動アノテーションパイプラインを作成した。図3に示すように、このパイプラインは3つのステージから構成される。  
      (1) Object-level annotation.  
      まず、ビデオから物体領域が最も大きいフレームを選択し、物体以外のピクセルをマスクします。切り取られた画像（マスクなし）とフル画像（マスクあり）は、それぞれInternVL2-76B [14] に入力され、詳細な記述が生成されます。生成された記述はQwen2-72B [101] によって整合性チェックされ、矛盾する記述は破棄されます。この方法により、エラーが発生しやすいケースを除外することができます。  
      (2) Scene- level annotation.  
      画像と前段階のオブジェクトレベルの記述は両方ともInternVL2-76B [14]に送られ、シーンと周囲のオブジェクトとの関係を含む詳細なオブジェクト記述が生成されます。  
      (3) Video-level annotation.  
      ビデオから8フレームを均一にサンプリングし、各フレーム内のオブジェクトを強調するために黄色の枠線を適用します。これらのフレームはシーンレベルの記述とペアになっており、InternVL2-76B [14] によって処理され、オブジェクトの動きと動作を捉えたビデオレベルの記述が生成されます。  
        
      ダイアグラム

      AI 生成コンテンツは誤りを含む可能性があります。  
      図3 データアノテーションパイプライン。提案する自動データアノテーションパイプラインは、オブジェクトレベル、シーンレベル、ビデオレベルのテキスト表現アノテーションの3つの段階で構成されています。  
        
      **Ref-SAV Dataset.**  
      上記のパイプラインを使用して、SA-V データセットの詳細なオブジェクト表現に自動的に注釈を付け、Ref-SAVを作成しました。Ref-SAV のトレーニング セットには、人間によるラベル付けのない 37,311 本のビデオと 72,509 個のオブジェクト表現が含まれています。Ref-SAV 検証セットでは、検証セットとテスト セットに含まれるビデオの数が限られているため、SA-V データセットのトレーニングセットからビデオのサブセットを選択して、Ref-SAV 評価ベンチマークを作成しました。これらのビデオは、Ref-SAV トレーニング データセットとは完全に別のものです。評価ベンチマークは 2 つの部分で構成されています。1) 長い表現セットは、上記の自動注釈パイプラインを使用して生成され、人間の注釈者によって慎重にフィルター処理されます。2) 短い表現セットは、手動で注釈が付けられます。この評価ベンチマークには、1,147 本のビデオと 1,945 個のオブジェクト表現が含まれており、1,694 個の長い表現と 251 個の短い表現で構成されています。
   4. Sa2VA Training and Testing.

～～省略～～  
VQAタスクとセグメンテーションタスクの両方を一緒に学習しました。  
MLLMの学習は事前学習ではなく、LoRAでのファインチューニングを実施。  
その際にSAM-2のデコーダー部分もファインチューニングを行った。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。

1. Experiments  
   ～～省略～～
   1. Main Results

Comparison With SOTA MLLMs.  
表6に示すように、Sa2V-8BはRefCOCO、RefCOCO+、RefCOCOgでそれぞれ81.9、76.5、78.9のcIoUスコアを達成し、GLaMM-7Bを2.4、3.9、4.7 cIoU上回っています。Sa2VAは、RefCOCO+とRefCOCOgの最先端手法に対して良好なパフォーマンスを示し、LISA、GLaMM、PixelLLM、PSALM、OMG-LLaVAなどの既存のグラウンディングMLLMを大幅に上回っています。さらに、Sa2VAは強力な会話能力を示し、MME、MMbench、SEED-Benchで2229（1651/578）、82.4、75.5のスコアを達成しました。一方、既存のグラウンディングMLLMは壊滅的な忘却のために会話のパフォーマンスが低下しています。 Sa2VA は、画像 QA ベンチマークにおいて InternVL2 に匹敵するパフォーマンスを実現しており、Sa2VA がベース MLLM のチャット パフォーマンスをほぼ維持していることを示しています。  
Sa2VAはビデオベンチマークでも優れた性能を発揮します。MeVIS、Ref-DAVIS17、ReVOSでそれぞれ46.9、75.2、57.6のJ&Fスコアを達成し、従来のSOTA VISA-13Bを2.4、4.8、6.7のJ&Fスコアで上回りました。さらに、Sa2VA-8BはビデオQAベンチマークMMBench-Videoで1.34のスコアを獲得し、InternVL2-8Bの1.28を上回りました。Sa2VAはビデオセグメンテーションと理解の両方の機能を備えているため、ビデオチャットタスクでは優位な結果を達成していませんが、これは既存の手法では同時に処理できないためです。主な実験結果は、Sa2VAが汎用性が高く強力なMLLMであることを示しています。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。  
**Detailed Results on Image Benchmarks.**  
[46, 95, 114]で示されているように、命令チューニングを行うとチャットタスクのパフォーマンスが大幅に低下します。この問題に対処するために、我々は共学習を活用します。表7には、画像チャットデータセットにおける追加結果も示しています。本モデルは、複数の画像チャットベンチマークにおいて良好な結果を維持しながら、高い参照セグメンテーション性能を達成しており、その汎用性を実証しています。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。  
**Results on Ref-SAV validation set.**  
表10では、提案するRef-SAVベンチマークを用いて、いくつかの最先端のRef-VOSモデルをベンチマークした。基礎モデル[99]や最近のビデオMLLMモデル[100]でさえ、我々のベンチマークでは強い結果を達成できなかったことがわかった。これは、我々のベンチマークがSAM-2 [77]からのより多くのオクルージョン、より長いテキスト記述、そして多様な注釈を特徴としているためである。逆に、我々の手法Sa2VAは、Ref-SAVトレーニングセットの有無にかかわらず、強い結果をもたらすことができる。我々の手法は、我々のトレーニングセットを用いることでさらに強化することができ、Ref-SAVトレーニングセットがビデオ理解コミュニティにとって貴重な補足資料となることを示している。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。  
**Comparison with specific fine-tuned methods.**オリジナルのSa2VA設定は、すべてのタスクを完了するために同じモデルの重みを使用する必要があるため、より困難です。公平な比較のために、先行研究[76, 114]と同様に、特定の微調整されたデータセットでMLLMを使用して評価します。表11に示すように、共学習モデルと比較して3つのデータセット全体で1.5%の改善が見られ、Sa2VAは最近の手法の中で最高のパフォーマンスを発揮しています。別の観点から見ると、単一のデータセットでの微調整はあまり効果がありません（RefCOCOでは0.4）。しかし、これは各データセットを微調整して新しいモデルを作成する必要があります。したがって、汎用性と利便性から、Sa2VAのようなより一般的なモデルを推奨します。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。  
**InternVL-2.5 Results.**  
さらに、より強力なMLLMと追加のデータセットを用いたSa2VAのスケーラビリティを評価するために、アブレーション実験を行った。表8に示すように、MLLMをInternVL2 [14] からInternVL2.5 [13] にアップグレードすると、画像、動画、マルチモーダルベンチマーク全体で一貫して性能が向上する。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。

* 1. Ablation Studies

～～省略～～

* 1. More Comparison Results

**Comparison with recent video MLLM models on referral video segmentation**最近、いくつかの研究では、MLLMをビデオ参照セグメンテーションタスクにも組み合わせています。 表15に示すように、3つのデータセットでモデルをいくつかの最近の方法と比較しています。これらの競合には、VideoGLaMM [70]、VideoLISA [4]、HyperSeg [88]、InstructSeg [87]などが含まれます。また、[70]で報告された組み合わせた強力なベースライン（SAM2 [77] + GlaMM [76]）も含めます。 表15に示すように、Sa2VA-8Bモデルは3つのデータセットすべてで強力なパフォーマンスを実現しています。 Sa2VA-4Bモデルも強力な競争力を示し、MeViSで最高スコアを達成しながら、画像/ビデオチャット、GCG、および視覚プロンプト理解の能力を維持しています。 同時に、他の参照ビデオセグメンテーション手法はこのタスクにのみ焦点を当てており、他のタスクのトレーニングデータが不足しているため、得られたモデルはビデオや画像チャットなどの他のタスクではパフォーマンスが低下します。 **Comparison with vision expert models.**表16に示すように、本モデルを、参照セグメンテーションに特化して設計された最近のビジョンエキスパートモデルと比較しました。これらの専門モデルは一般的に軽量ですが、汎用26Bモデル（Sa2VA-26B）は依然として最も優れた結果を達成し、5つのデータセット全てにおいてこれらのモデルを上回っています。これは重要な点です。なぜなら、これらの専門モデルは会話機能を備えておらず、ほとんどの場合、単一のモダリティ（動画または画像）に限定されているのに対し、本モデルは汎用性が高いからです。  
**Effectiveness of training dataset on more methods.**さらに、代表的な方法である UniRef++ [92] を評価することにより、提案する Ref-SAV トレーニング データセットの有効性を実証します。表 17 に示すように、データの影響を分離するために 2 つの異なる設定を比較します。ゼロ ショット設定では、新しいトレーニングを行わずに、事前トレーニング済みの UniRef++ モデルを Ref-SAV 検証データセットで直接テストします。対照的に、ファイン チューニング設定では、最初に提案する Ref-SAV トレーニング データセットでモデルをトレーニングしてから、同じ検証セットで評価します。結果は、データセットでファイン チューニングされたモデルが、すべてのメトリックで大幅な改善を達成することを明確に示しています。最も顕著なのは、全体的な J&F スコアが 10.5 から 14.6 に大幅に増加していることです。この大幅な向上は、「Long」(12.5 から 17.2 J&F) と「Short」(8.6 から 12.0 J&F) の両方の検証分割で一貫しており、自動データ エンジンが Ref-VOS モデルのパフォーマンスを大幅に向上させる可能性を秘めていることを示しています。  
Results on visual prompts understanding tasks.  
先行研究[107, 114]に倣い、視覚プロンプト理解タスクについても報告する。具体的には、RefCOCOgデータセットを用いて領域キャプションの性能を評価する。表18に示すように、我々の手法Sa2VA4Bは、近年の視覚プロンプト理解モデルの中でも最高の結果を達成した。METEOR指標では17.3というスコアを獲得し、次点の手法であるOsprey [37]の16.6を大きく上回った。これは、Sa2VAが強力な領域認識キャプションを生成できることを示している。

* 1. Discussions

**Sa2VA with recent MLLM models.**  
表14は、Qwen2.5-VL [3]、Qwen3-VL [97]、およびInternVL3 [124]を含むさまざまな最近のベースマルチモーダル大規模言語モデル（MLLM）と統合されたSa2VAフレームワークの包括的な評価を示しています。結果は、ベースMLLMの選択がパフォーマンスに大きく影響し、この影響がさまざまなタスク間で異なることを明確に示しています。たとえば、Sa2VA-InternVL3-14Bモデルは、特にRefCOCO（83.6）やMeViS（59.2）などの画像およびビデオセグメンテーションベンチマークで最も強力な全体的なパフォーマンスを実現しますが、他のモデルも特定のドメインで競争力のある強さを示しています。特に、Sa2VA-Qwen3VL-4BモデルはMMBenchで最高スコア（86.3）を記録し、他の画像チャットデータセットでも非常に競争力のある結果を示しています。この違いは、さまざまなベースMLLMアーキテクチャが異なる機能を持っていることを強調しており、モデル選択はタスク固有の最適化にとって重要な要素となります。さらなる研究を促進し、コミュニティがこれらの調査結果に基づいて構築できるようにするために、トレーニング済みのすべてのモデルのチェックポイントを公開しました。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。  
  
Discussion on the inference speed.  
Sa2VAの計算オーバーヘッドは、各推論ステップでトークンを順次生成する自己回帰的な性質を持つMLLMコンポーネントによって大きく左右されます。対照的に、主要な追加モジュールであるSAM-2は、2億2000万パラメータ（10億から260億の範囲のMLLMと比較して）と比較的軽量で効率的であり、画像ごとに1回のフォワードパスのみでビジョンタスクで39.5 FPSを達成しています。MLLMの出力長が可変であるため、正確なレイテンシの推定は困難ですが、より具体的な比較を行うために、フレームと言語の長さを固定した制御ベンチマーク（テキストプロンプトとして「男性を分割してください」を5フレームで合計1280トークンで表示）も構築し、計算コストをより体系的に推定します。このベンチマークの結果は表19に詳細に示されています。推論時間は、Sa2VA-1Bで0.123秒、Sa2VA-4Bで0.282秒、Sa2VA-8Bで0.201秒、Sa2VA-26Bで0.463秒でした。推論速度は最適化に大きく依存し、4Bモデル（0.282秒）と8Bモデル（0.201秒）の差は、採用されているベースLLM（Qwen vs. InternLM）の違いによるものと考えられます。とはいえ、これらの結果は、モデルの計算コストの大部分が、効率的なシングルパスSAM-2モジュールではなく、自己回帰MLLMコンポーネントに起因していることを示しています。  
テーブル

AI 生成コンテンツは誤りを含む可能性があります。  
  
**Inference Strategy.**  
キーフレーム サンプリング戦略の影響も分析しました。最初の 5 フレームを使用するデフォルトのアプローチは、シンプルさと一貫性のために選択されましたが、この方法が、長距離の時間依存性やビデオの後半に登場するオブジェクトを捕捉するモデルの能力を制限するかどうかを調査しました。これをテストするために、Sa2VA-8B モデルを使用して MeVIS データセットでアブレーション スタディを実行しました。最初の 1、3、4、5つの連続するフレームを選択した場合のパフォーマンスを、ビデオ全体にわたって 5つのフレームを選択する均一サンプリング戦略と比較しました。この比較の結果の詳細は、表20 です。表が示すように、連続するフレームの数を増やすと、一般にパフォーマンスが向上し、1フレームの場合は55.1 J&F ですが、4フレームの場合は59.5 J&F になります。ただし、「最初の 5 フレーム」戦略 (58.9 J&F) は、「均一 5 フレーム」戦略 (62.9 J&F) を大幅に上回っています。注目すべきは、5フレーム法はどちらも推論時間（0.207秒）が同じで、処理する画像トークン数（1280）も同じであることです。この結果は、より洗練された間隔ベースのサンプリング戦略によって性能が向上することを示唆しています。しかしながら、Sa2VAは可能な限りシンプルなままにし、サンプリングに関する高度な手法の設計は今後の課題としたいと考えています。

* 1. Visualizations  
     **Results on image referring segmentation.**  
     図5に示すように、画像参照セグメンテーションタスクを視覚化しています。異なる言語記述を用いて、Sa2VAは記述を手がかりに異なるマスクをセグメンテーションできます。  
     グラフィカル ユーザー インターフェイス, Web サイト

     AI 生成コンテンツは誤りを含む可能性があります。  
     **Results on video referring segmentation.**

図6 に示すように、私たちの方法は、多様で困難な状況の処理に優れており、オクルージョン シーンや非常に動的な環境でも効果的に調整して実行できる優れた能力を示しています。  
座る, いっぱい, 異なる, 束 が含まれている画像

AI 生成コンテンツは誤りを含む可能性があります。  
  
**Results on visual prompt understanding.**図7に示すように、本手法は視覚的な手がかりに基づいて説明文を生成することができます。本手法によって生成された説明文は、高度な文脈認識を示しており、視覚的な手がかりの詳細を効果的に捉えています。  
テキスト

AI 生成コンテンツは誤りを含む可能性があります。  
**Results on GCG task.**  
図4は、グラウンデッドキャプション生成タスクにおける視覚的な例をさらにいくつか示しています。従来のSOTAモデルであるOMG-LLaVAと比較すると、本モデルはマスク品質とテキストマスクのアライメントの両方で優れた結果を達成しています。前者はセグメンテーション出力のアライメント精度が向上したことを示しており、後者はテキストと領域のアライメント精度が向上したことを示しています。  
タイムライン

AI 生成コンテンツは誤りを含む可能性があります。  
  
**Video demo.**  
動画などの動的な視覚コンテンツは、論文の静的な形式では効果的に提示できないため、補足資料に動画デモを追加しました。これらの動画は、本手法の性能をより明確かつ直感的に理解していただくことを目的としています。デモはZIPファイルをご覧ください。

1. Failure Cases and Future Work  
   **Failure cases of Sa2VA and Future work.**  
   Sa2VAは、さまざまな参照セグメンテーションおよびビデオ参照セグメンテーションデータセットでより優れた結果を達成しています。ただし、まだ調査が必要な失敗ケースがいくつかあります。今後の作業では、Sa2VAの2つの欠点を特定します。1つは、図8に示すように、困難で区別された参照例を含む長いビデオです。これは、Sa2VAが、長くて複雑なテキストを整列させるためにビデオコンテンツ全体を知らずにオンラインモードで動作するためです。もう1つは、参照セグメンテーションタスクに悪影響を与えずにVQAタスク（画像とビデオ）を改善する方法です。メインペーパーのアブレーション表に示されているように、VAQデータの規模が大きくなると、参照セグメンテーションタスクのパフォーマンスが低下します。したがって、両方のタイプのバランスを取り、根拠のある予測知識を維持する方法はまだ解決されていません。  
   **Future directions on Ref-SAV benchmarks.**  
   表3で述べたように、長いテキスト、オクルージョン入力、そしてカメラの動きを伴う長い動画は、ベンチマークにおいて重大な問題となります。そのため、新しいモデルでは、より堅牢なメモリ設計を含むいくつかの具体的な設計を用いて、これらの課題に対処するために、より堅牢なRef-VOSシステムを設計する必要があります。さらに、長いテキストを理解し、長いテキストのグラウンディング能力を向上させることも重要です。
2. Conclusion

本研究では、SAM-2とLLaVAのようなMLLMを統合し、画像と動画の両方に対して緻密でグラウンディングされた理解を実現する汎用フレームワークであるSa2VAを紹介します。この手法は、ワンショットの命令チューニングのみで、参照画像/動画セグメンテーションや画像/動画会話など、様々な画像・動画理解タスクを処理できます。LLaVAとSAM-2の両方の知識を活用することで、このモデルはマスク生成と言語生成の両方において強力な機能を備えています。提案手法の有効性を示すために、参照動画オブジェクトセグメンテーションのベンチマークであるRef-SAVを提案します。広範な実験結果から、Sa2VAは様々なタスクにおいてベンチマーク全体で優れた性能を発揮することが示されています。さらに、Sa2VAは様々な最新のVLMで拡張でき、ピクセルマルチモーダルシステムのためのより強力なベースラインを構築できる可能性があります。