Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling

Abstract

我々は、InternVL 2.0をベースに構築された高度なマルチモーダル大規模言語モデル（MLLM）シリーズであるInternVL 2.5を紹介します。このシリーズは、コアモデルアーキテクチャを維持しながら、トレーニングとテスト戦略、そしてデータ品質の大幅な強化を導入しています。この研究では、モデルのスケーリングとパフォーマンスの関係を深く掘り下げ、ビジョンエンコーダー、言語モデル、データセットサイズ、テスト時の構成におけるパフォーマンスの傾向を体系的に調査します。多分野推論、文書理解、マルチ画像/ビデオ理解、現実世界の理解、マルチモーダル幻覚検出、ビジュアルグラウンディング、多言語機能、純粋言語処理など、幅広いベンチマークでの広範な評価を通じて、InternVL 2.5はGPT-4oやClaude-3.5-Sonnetなどの主要な商用モデルに匹敵する競争力のあるパフォーマンスを示しています。特筆すべきは、当社のモデルがオープンソースMLLMとして初めてMMMUベンチマークで70%を超えるスコアを達成したことです。Chain-of-Thought（CoT）推論によって3.7ポイントの改善を達成し、テスト時間のスケーリングにおける大きな可能性を示しています。HuggingFaceのデモはhttps://huggingface.co/spaces/OpenGVLab/InternVLをご覧ください。

1. Introduction  
   近年、マルチモーダル大規模言語モデル（MLLM）[60, 137, 246, 36, 35, 248, 140, 228, 192, 275, 143, 54, 170]が人工知能における重要な技術として登場し、テキスト、画像、動画といった複数のモダリティからの情報を処理・理解することが可能となっている。これらのモデルは、自然言語処理、コンピュータビジョン、ヒューマンコンピュータインタラクションといった分野における画期的な進歩を期待できる。しかしながら、大規模MLLMの開発は依然として困難な課題であり、膨大な計算リソース、高度なアーキテクチャ、そして多様なデータタイプをスケーラブルに効果的に統合する能力が必要となる。  
   これらの課題に対処するため、モデルアーキテクチャの強化 [220, 232, 5, 172, 157, 210]、ビジョンエンコーダのスケーリング [252, 66, 36, 293, 185] および言語モデル [231, 235, 64, 19, 229, 221, 62]、より多様で高品質なデータセットの組み込み [124, 234, 25, 155]、テスト時間のスケーリングプロセスの改良 [215, 249, 230] など、さまざまな試みが行われてきました。GPT-4o [192] やClaude-3.5-Sonnet [8] などの著名な商用モデルは優れた性能を示していますが、そのクローズドな性質が透明性とアクセス性を制限し、オープンソースコミュニティにギャップを残しています。 InternVLシリーズ[36, 35, 71]やQwen-VLシリーズ[13, 246]などのオープンソースのマルチモーダルモデルは、高性能で透過的な代替手段を提供していますが、望ましいレベルのパフォーマンスと効率を達成するという点では依然として不十分です。  
   本研究では、InternVL 2.0の基礎アーキテクチャを基盤とする、高度な大規模MLLMシリーズであるInternVL 2.5を紹介します。InternVLシリーズ全体の目標を継承し、商用のクローズドソースモデルとオープンソースのマルチモーダルモデル間のパフォーマンスギャップを埋めることを目指しています。InternVL 2.5では、ビジョンエンコーダ、言語モデル、データセットサイズ、推論時間の変更がモデル全体のパフォーマンスにどのように影響するかなど、MLLMのさまざまな要素を体系的に調査し、マルチモーダルモデルにおけるスケーリングとパフォーマンスの関係を実証しています。具体的には、いくつかの興味深い発見がありました。(1) 大規模なビジョンエンコーダは、MLLMをスケールアップする際に、トレーニングデータへの依存度を大幅に低減します。表3に示すように、600Mビジョンエンコーダを搭載したQwen2-VL-72B [246]と比較して、6Bビジョンエンコーダを搭載したInternVL2.5-78Bは、わずか1/10のトレーニングトークンを使用して、より良いパフォーマンスを達成できます。これにより、MLLMをスケールアップする際の探索コストが大幅に削減されます。 (2) データの品質が重要です。InternVLを2.0から2.5にアップグレードすると、データセットのサイズが2倍になりましたが、厳密なフィルタリングにより品質が大幅に向上しました。たとえば、異常なサンプル（繰り返しパターンなど）を慎重に除外し、MMMU [289]などのChain-of-Thought（CoT）推論タスクやOlympiadBench [80]のような複雑な課題で大幅な改善を実現しました。既存のオープンソースMLLMのほとんどは、CoT [249]を使用するとパフォーマンスが低下する傾向があることに注意してください。 (3) テスト時間のスケーリングは、困難なマルチモーダルQAに有益です。 MMMUのような高難度タスクにおいて、CoTを用いたInternVL2.5-78Bは70.1%の正解率を達成し、これは直接的な回答よりも3.7ポイント高い結果となりました。その後、CoTを多数決と組み合わせることで、さらなる改善が期待できることを検証しました。  
   私たちの貢献は以下のようにまとめられます。  
   (1) InternVL 2.5をオープンソースコミュニティに公開し、マルチモーダルAIシステムの開発と応用のための強力なツールを提供し、この分野におけるさらなる研究を促進します。  
   (2) MLLMの様々なコンポーネント（ビジョンエンコーダ、言語モデル、データセットサイズ、推論時間など）のスケーリングがパフォーマンスにどのような影響を与えるかを調査します。  
   (3) 多分野推論、文書理解、複数画像/動画理解、実世界理解、マルチモーダル幻覚検出、ビジュアルグラウンディング、多言語対応能力、純粋言語処理など、多様なベンチマークを用いた広範な評価により、InternVL 2.5はGPT-4o [192]やClaude-3.5-Sonnet [8]などの主要な商用モデルに匹敵する競争力のあるパフォーマンスを示します。これは、MMMU検証セットで70％を超えた最初のオープンソースMLLMであり[289]、新たなベンチマークを設定し、マルチモーダルAIの進歩におけるオープンソースソリューションの可能性を浮き彫りにしました。  
     
   図1：OpenCompassリーダーボードにおける様々なMLLMのパフォーマンス。InternVL 2.5は強力なマルチモーダル機能を備えており、GPT-4o [192]やClaude-3.5-Sonnet [8]といったクローズドソースモデルに匹敵する。しかし、OpenCompassのスコアは8つの学術VQAベンチマークから算出されており、全体的な機能のサブセットのみをカバーしているため、クローズドソースモデルと同等のパフォーマンスを実現するには、さらなる努力が必要である。
2. Model Architecture  
   図2：全体アーキテクチャ。InternVL 2.5は、InternVL 1.5 [35] およびInternVL 2.0と同じモデルアーキテクチャ、すなわち、広く用いられている「ViT-MLP-LLM」パラダイムを継承しています。これは、事前学習済みのInternViT-300MまたはInternViT-6Bと、MLPプロジェクターを介して様々なサイズのLLM [19, 229]を組み合わせたものです。以前のバージョンと同様に、ピクセルアンシャッフル操作を適用して、448×448の画像タイルごとに生成される1024個のビジュアルトークンを256個のトークンに削減します。さらに、InternVL 1.5と比較して、InternVL 2.0および2.5では、既存の単一画像およびテキストのみのデータに加えて、複数画像およびビデオデータを組み込むなど、追加のデータタイプが導入されています。  
     
   2.1 Overall Architecture  
   図2と表2に示すように、InternVL 2.5は、様々なMLLM研究で広く採用されている「ViT-MLP-LLM」パラダイム[150, 151, 36, 316, 162, 246, 124, 256]に従って、前身のInternVL 1.5[35]およびInternVL 2.0と同じモデルアーキテクチャを保持しています。  
   この新バージョンでは、このアーキテクチャの実装において、新たに段階的に事前学習されたInternViT-6BまたはInternViT-300Mと、InternLM 2.5 [19]やQwen 2.5 [229]を含む、サイズとタイプの異なる様々な事前学習済みLLMを、ランダムに初期化された2層MLPプロジェクターを用いて統合しています。以前のバージョンと同様に、高解像度処理のスケーラビリティを向上させるために、ピクセルアンシャッフル操作を適用し、ビジュアルトークンの数を元の4分の1に削減しました。その結果、このモデルでは、448×448の画像タイルは256個のビジュアルトークンで表現されます。  
   入力データの前処理に関しては、InternVL 1.5と同様の動的解像度戦略を採用し、入力画像のアスペクト比と解像度に基づいて画像を448×448ピクセルのタイルに分割します。InternVL 2.0からの主な違いは、図2(b)に示すように、複数の画像と動画データのサポートを追加したことです。データの種類によって前処理の設定が異なり、詳細はセクション3.1で説明します。  
   2.2 Vision Encoder  
   InternVLはビジョンエンコーダとしてInternViT [36] を採用しています。InternViTの学習過程をより詳細に記録するために、表1に詳細な情報を示しています。InternViTには現在、InternViT-6BとInternViT-300Mの2種類のモデルサイズがあります。  
   InternViT-6B.  
   InternViT-6B-224pxは、私たちのCVPR論文[36]で初めて紹介され、その構造はバニラViT [61]を踏襲していますが、QK-Norm [53]とRMSNorm [294]を組み込んだ小さな調整が加えられています。59億のパラメータ、48層、隠れサイズ3200、25ヘッドを持ち、対照損失[195]を使用して学習されました。当時のゲインは限られていたため、重みを継続的に改良するために増分事前学習戦略を採用しました。具体的には、MLPプロジェクターを介してInternViT-6BをLLMに接続し、簡単なMLPウォームアップの後、次のトークン予測損失（図4(a)に示す）を使用してInternViT-6Bを共同学習し、視覚的特徴抽出機能を強化しました。 V1.0およびV1.2では、学習に448×448の固定解像度を使用していましたが、それ以降のバージョンでは、高解像度処理を改善するために動的解像度学習に切り替えました。InternVL 1.5レポート[35]に詳述されているように、InternViT-6B-448px-V1.2の最後の3層を削除し、その深度を48層から45層に減らしました。これは、これらの層がCLIP損失目標に合わせて調整され、局所情報よりもグローバルアライメントを優先したためです。その結果、最新のInternViT-6B-448px-V2.5を含む以降のすべてのバージョンは、45層と55億のパラメータを備えています。  
     
   InternViT-300M.  
   InternViT-300M-448px-Distillは、コサイン蒸留損失を利用した教師モデルInternViT-6B-448px-V1.5の蒸留版である。このモデルは、0.3Bのパラメータ、24層、1024の隠れサイズ、16個のアテンションヘッドで構成される。6Bバージョンとは異なり、0.3BバリアントはQK-Norm [53]を使用せずに標準のLayerNorm [11]を採用している。蒸留コストを削減するために、多少のアーキテクチャ上の違いはあるものの、適用可能な場合はCLIP-ViT-Large-336px [195]を使用してこのモデルを初期化した。蒸留後、このモデルをLLMと統合し、上記と同様の手順に従って、動的高解像度とNTP損失を使用してビジョンエンコーダーをトレーニングした。その後、ビジョンエンコーダーを抽出し、InternViT-300M-448pxとしてリリースした。本レポートでは、NTP 損失を使用して、より多様なデータ混合物で以前の重みを段階的に事前トレーニングすることで、InternViT-300M をさらに改良し、強化された InternViT-300M-448px-V2.5 が完成しました。  
     
   2.3 Large Language Model  
   表2は、InternVL 1.5、InternVL 2.0、そして最新のInternVL 2.5を含む、InternVLの異なるバージョンで使用されている言語モデルの概要を示しています。表に示すように、以前のバージョンは主にInternLM 2 [19]、Qwen 2 [268]、Phi 3 [1]、Yi [279]、Llama 3 [64]などの言語モデルに基づいて構築されていました。より優れた性能を実現するために、InternVL 2.5シリーズでは、言語バックボーンをInternLM 2.5 [19]やQwen 2.5 [229]などの最新の最先端モデルに全面的にアップグレードしました。  
   図3：さまざまなデータタイプのデータ形式の図解。（a）単一画像のデータセットの場合、入力の最大解像度を確保するために、最大タイル数n maxが1つの画像に割り当てられます。（b）複数画像のデータセットの場合、タイルの総数n maxはサンプル内のすべての画像に比例して分配されます。（c）ビデオデータセットの場合、この方法はn max = 1に設定し、個々のフレームのサイズを448×448の固定解像度に変更することでアプローチを簡素化します。
3. Training Strategy  
   3.1 Dynamic High-Resolution for Multimodal Data  
   InternVL 2.0および2.5では、InternVL 1.5 [35] で導入された動的高解像度学習アプローチを拡張し、複数の画像および動画データセットを処理する能力を強化しました。このプロセスは主に以下のステップで構成されます。  
   Closest Aspect Ratio Matching.  
   入力画像 I （寸法 W × H ）が与えられた場合、アスペクト比は r = W/ H として計算されます。目的は、歪みを最小化する最も近いアスペクト比を選択しながら、画像を S × S （ただし S = 448 ）のタイルにリサイズすることです。タイルの数 n tiles は、事前定義された範囲 [ n min , n max ] に制限されます。  
   サイズ変更に最適なアスペクト比を見つけるために、ターゲットアスペクト比のセット ℛ を次のように定義します。  
     
   元のアスペクト比rと各目標アスペクト比r targetとの差を最小化することで、最も近いアスペクト比r bestが選択されます。  
     
   Image Resizing and Splitting.  
   最適なアスペクト比が決定されると、画像は新しい寸法 W new × H new にサイズ変更されます。ここで、 i best と j best は r best に対応する係数です。  
     
   次に、画像はサイズ S × S のタイルに分割されます。タイルの数は、n tiles = i best × j best として計算されます。各タイルは、サイズが一定になるように、サイズ変更された画像から切り取られます。  
   Thumbnail Generation.  
   オプションとして、タイルの数 n tiles > 1 の場合、元の画像 I は S × S の正方形にサイズ変更され、追加のサムネイル I thumb が生成されます。このサムネイルはタイルのリストに追加され、ローカライズされたタイルと並べて全体図を表示します。 n tiles = 1 の場合、追加するサムネイルが存在しないため、この手順は自動的にスキップされます。  
   Data Formats for Different Data Types.  
   図 3 に示すように、InternVL 2.0 および 2.5 の動的高解像度方式は、単一画像データセットを超えて、複数画像およびビデオデータセットもサポートします。  
   単一画像のデータセットの場合、最大タイル数 n max が単一画像に割り当てられ、可能な限り高い解像度で処理されます。このシナリオでは、ビジュアルトークンは <img> タグと </img> タグで囲まれ、追加の補助タグは使用されません。  
   複数画像データセットの場合、タイルの総数n max は1つのサンプル内のすべての画像に分配されます。各画像は、個々の画像を明確に識別するために、Image-1のような補助タグで識別されます。画像自体は、画像データの開始と終了を示す<img>タグと</img>タグで囲まれています。各画像に割り当てられるタイルの数I i は、画像の総数N image に比例し、次の式に従います。  
     
   動画データの場合、このアプローチは n max = 1 と設定することで簡素化されます。各動画フレームは448 × 448の固定解像度にリサイズされるため、タイリングは不要になります。これは、トレーニング中に通常、1つの動画から多数のフレーム（例：32または64）が抽出されるためです。このモデルでは、高解像度の入力がない場合でも、8,192または16,384個のビジュアルトークンが生成されます。各動画フレームは、Frame-1などのタグでラベル付けされ、画像データと同様に<img>タグと</img>タグで囲まれています。  
   3.2 Single Model Training Pipeline  
   図4：トレーニングパイプラインとプログレッシブスケーリング戦略の図解。(a) 単一モデルのトレーニングパイプライン。トレーニングプロセスは、ステージ1（MLPウォームアップ）、オプションのステージ1.5（ViT増分学習）、ステージ2（フルモデル命令チューニング）の3つのステージに分かれています。このマルチステージ設計により、ビジョンと言語のアライメントが段階的に強化され、トレーニングが安定化し、モジュールをより大規模なLLMとの統合に向けて準備できます。(b) プログレッシブスケーリング戦略。初期段階で小規模なLLMでトレーニングされたViTモジュールは、より大規模なLLMと容易に統合できるため、リソースオーバーヘッドを抑えながらスケーラブルなモデルアライメントが可能になります。  
     
   InternVL 2.5の単一モデルの学習パイプラインは3つのステージに分かれており、モデルの視覚認識能力とマルチモーダル能力を強化するように設計されています。各ステージでは、視覚と言語のモダリティを段階的に統合し、パフォーマンスの最適化と学習効率のバランスを取ります。  
   Stage 1: MLP Warmup.  
   図4(a)に示すように、学習は視覚表現と言語表現を繋ぐ最初の橋渡しとなるMLPプロジェクターのウォームアップから始まります。この段階では、視覚エンコーダ（InternViT [36]）と言語モデルは両方とも固定された状態で、MLPプロジェクターのみが学習されます。最適な性能を得るために、学習コストは増加しますが、この段階から動的高解像度学習戦略を採用します。  
   このフェーズでは、表4に示す事前学習データの組み合わせを利用します。このデータは構造化されたChatML形式でフォーマットされ、NTP損失を考慮して最適化されています。さらに、収束を加速するために高い学習率を適用することで、MLPはLLMの入力空間に迅速に適応し、堅牢なクロスモーダルアライメントを確立できます。MLPウォームアップフェーズでは、後段で追加の学習可能なコンポーネントをアンロックする前に、モデルがマルチモーダルタスクを処理できるように十分に準備されていることを確認し、学習の安定性を向上させます。  
   Stage 1.5: ViT Incremental Learning (Optional).  
   図4(a)に示すように、ステージ1.5では、ビジョンエンコーダに増分学習を導入します。このステージでは、ビジョンエンコーダとMLPプロジェクターの両方が学習可能であり、ステージ1と同じ事前学習データとNTP損失を用いて学習が行われます。このステージの目的は、ビジョンエンコーダの視覚特徴抽出能力を強化し、より包括的な情報を取得できるようにすることです。特に、多言語OCRデータや数式チャートなど、Webスケールデータセット（例：LAION-5B [203]）では比較的稀な領域において、より包括的な情報を取得できるようになります。  
   表3に示すように、この段階では学習率を低く設定することで壊滅的な忘却を防ぎ、エンコーダが以前に学習した能力を失わないようにします。さらに、新しいドメイン要件やデータが導入されない限り、ビジョンエンコーダの学習は一度だけで済みます。一度学習すれば、再学習なしで異なるLLMで再利用できます（図4(b)およびセクション3.3を参照）。そのため、ステージ1.5はオプションとなります。これは、エンコーダが既に特定のタスク向けに最適化されている場合に特に有益であり、大きな追加コストをかけずに様々なサイズのLLMに統合できます。  
   Stage 2: Full Model Instruction Tuning.  
   最終段階では、図4(a)に示すように、ViT、MLP、LLMで構成されるモデル全体が、高品質のマルチモーダル指示データセットを用いて学習されます。最終的なユーザー向け出力を生成するLLMが学習可能になるため、データ品質は特に重要です。ノイズの多い少量のデータ（例えば数千サンプル）であっても、繰り返し出力や特定の誤った結果など、モデルの異常な動作につながる可能性があります。LLMの劣化を軽減するため、この段階では厳格なデータ品質管理を実施しています。  
   さらに、このステージのトレーニングハイパーパラメータはシンプルに保たれ、各コンポーネントに異なる学習率を適用するのではなく、モデル全体に​​統一された学習率が適用されます。このステージを完了すると、InternVL 2.5の完全なトレーニングプロセスが終了します。ステージ3（より高品質なデータを用いたトレーニング後、または他のトレーニング手法（例：選好最適化））を通じてさらなる改善を行うこともできますが、これは将来の課題として残しておく予定です。  
   3.3 Progressive Scaling Strategy  
   図4に示すように、ビジョンエンコーダ（例：InternViT）をLLMに効率的に適合させるための漸進的スケーリング戦略を提案します。以前、InternVL 1.5および2.0のトレーニングで同様の戦略を採用しましたが、今回はこのアプローチを明確に定義された方法論として形式化しました。この戦略では、小規模でリソース効率の高いLLMから開始し、徐々に大規模なLLMへとスケールアップしていく段階的なトレーニングアプローチを採用しています。このアプローチは、ViTとLLMをNTP損失を用いて共同トレーニングした場合でも、得られる視覚特徴は他のLLMによって容易に理解できる一般化可能な表現であるという私たちの観察に基づいています。  
   具体的には、ステージ1.5では、InternViTはより小規模なLLM（例：20B）と並行して学習され、基本的な視覚能力とクロスモーダルアライメントの最適化に重点が置かれます。このフェーズでは、大規模なLLMで直接学習する場合に伴う高い計算コストを回避できます。共有重みメカニズムを使用することで、学習済みのInternViTは、再学習を必要とせずに、より大きなLLM（例：72B）に簡単に転送できます。したがって、より大きなモデルを学習する場合、以前のステージで最適化されたInternViTモジュールが再利用されるため、ステージ1.5をスキップできます（表3を参照）。これにより、学習が加速されるだけでなく、ビジョンエンコーダーの学習済み表現が保持され、より大きなモデルに効果的に統合されます。  
   この漸進的スケーリング戦略を採用することで、大規模なMLLM学習に通常伴うコストのほんの一部で、スケーラブルなモデル更新を実現できます。例えば、Qwen2-VL [246] は累計1.4兆トークンを処理しますが、私たちのInternVL2.5-78Bは約1200億トークンで学習されており、これはQwen2-VLの10分の1にも満たない量です。このアプローチは、事前学習済みコンポーネントの再利用を最大化し、冗長な計算を最小限に抑え、複雑な視覚言語タスクに対応できるモデルの効率的な学習を可能にするため、リソースが限られた環境において特に有利です。  
   3.4 Training Enhancements  
   モデルの現実世界のシナリオへの適応性と全体的なパフォーマンスを向上させるために、2つの重要な手法が導入されています。これらの最適化は、ユーザーエクスペリエンスとモデルのベンチマークパフォーマンスの向上に不可欠です。  
   Random JPEG Compression.  
   学習中の過学習を回避し、モデルの実世界におけるパフォーマンスを向上させるため、空間情報を保持するデータ拡張技術、すなわちJPEG圧縮を適用します。具体的には、品質レベル75～100のランダムJPEG圧縮を適用し、インターネット上の画像によく見られる劣化をシミュレートします。この拡張により、ノイズの多い圧縮画像に対するモデルの堅牢性が向上し、さまざまな画質においてより一貫したパフォーマンスを実現することでユーザーエクスペリエンスが向上します。  
   Loss Reweighting.  
   トークン平均化とサンプル平均化は、NTP損失の重み付けに広く適用されている2つの戦略です。トークン平均化はすべてのトークンにおけるNTP損失の平均を計算しますが、サンプル平均化はまず各サンプル内（トークン全体）のNTP損失を平均化し、次にサンプル全体にわたって平均化します。これらの戦略は、統一された形式で表現できます。  
   テキスト

   AI によって生成されたコンテンツは間違っている可能性があります。  
   ここで、ℒ i と w i はそれぞれトークン i の損失と重みを表し、x はトークン i が属する応答内のトークンの数を表します。  
   トークン平均化を使用すると、各トークンが最終的な損失に均等に寄与するため、トークン数の多いレスポンスに傾き、ベンチマークパフォーマンスが低下する可能性があります。一方、サンプル平均化では各サンプルが均等に寄与することが保証されますが、モデルが短いレスポンスを優先し、ユーザーエクスペリエンスに悪影響を与える可能性があります。トレーニング中に長いレスポンスまたは短いレスポンスへの偏りを軽減するために、 w i = 1/ x ^0.5 となる再重み付け戦略を適用します。この手法は「平方平均化」と呼ばれ、長さの異なるレスポンスの寄与をバランスさせます。
4. Data Organization  
   4.1 Dataset Configuration  
   InternVL 2.0 および 2.5 では、図 5 に示すように、トレーニング データの構成はいくつかの主要なパラメーターによって制御され、トレーニング中のデータセットのバランスと分散が最適化されます。  
   図5：データセットの構成。InternVL 2.0および2.5では、データ拡張は選択的に適用され、画像データセットでは有効、動画とテキストでは無効になります。最大タイル数（n max ）は入力の解像度を制御し、複数画像データセットの場合は高い値、動画の場合は低い値になります。繰り返し係数（r ）は、各データセットのサンプリング頻度を調整することでデータセットのサンプリングバランスを調整し、堅牢でバランスの取れた学習を実現します。  
     
   Data Augmentation.  
   まず、データ拡張（3.4節で導入したJPEG圧縮）を条件付きで適用し、データセットの特性に基づいて拡張手法を有効または無効にすることで、堅牢性を向上させます。具体的には、すべての画像データセットに対してこの拡張を有効にし、すべてのビデオデータセットに対しては無効にすることで、異なるビデオフレームで同じ画質が得られるようにします。  
   Maximum Tile Number.  
   パラメーター n max は、データセットごとに許容されるタイルの最大数を定義し、モデルに入力される画像または動画フレームの解像度を効果的に制御します。これにより、複雑さや種類が異なるデータセットを柔軟に処理できます。例えば、複数画像データセット、高解像度ドキュメント、インフォグラフィックの場合は n max = 24 または 36 に設定し、その他のほとんどの低解像度画像データセットの場合は n max = 6 または 12 を使用し、動画データセットの場合は n max = 1 に設定できます。この調整は InternVL 2.0 で初めて導入されましたが、InternVL 1.5 では、すべてのデータセットに n max = 12 という一律の値が適用されていました。  
   Repeat Factor.  
   最後に、繰り返し係数 r は各データセットのサンプリング頻度を決定します。 r ∈ ( 0 , 4 ] の場合、このパラメーターは r < 1 の場合にダウンサンプリングを可能にし、トレーニング中にデータセットの重みを減らし、 r > 1 の場合にアップサンプリングを可能にし、そのデータセットのエポック数を効果的に増やします。 このメカニズムはデータセットの相対的な割合を細かく調整し、トレーニングデータ全体にバランスのとれた分散を保証します。 特にマルチタスク学習では、 r を調整することにより、各ドメインまたはタスクからのデータが適切にトレーニングされ、単一のデータセットの過剰適合または不足適合が防止され、よりバランスの取れたモデルパフォーマンスにつながります。

(中略)

1. Evaluation on Multimodal Capability  
   InternVLのマルチモーダルタスクにおけるパフォーマンスを包括的に評価するために、我々は多様なベンチマークセットを採用しました。これには、確立された古典的なデータセットと、VLMEvalKit [63]によって提供される新しいデータセットの両方が含まれます。これらのベンチマークは幅広いカテゴリにまたがっており、様々なマルチモーダルタスクにおけるInternVLの能力を徹底的かつバランスよく評価することを目指しています。  
   5.1 Multimodal Reasoning and Mathematics  
   5.1.1 Benchmarks  
   私たちは、さまざまな分野に関連するベンチマークにわたる包括的な評価を通じて、InternVL のマルチモーダルな数学および推論能力を評価します。  
   MMMU [289]:  
   MMMUは、6つの分野にわたる大学レベルの課題においてMLLMを評価するベンチマークであり、特定分野における専門家レベルの推論能力と高度な知覚能力をテストします。MMMUの検証セットとテストセットにおいて、直接回答型推論アプローチとCoT推論アプローチの両方で達成された最高精度を報告します。  
   MMMU-Pro [290]:  
   MMMU-Proは、MMMUベンチマークのアップグレード版であり、幅広い学問分野におけるモデルのマルチモーダル理解および推論能力をより正確かつ厳密に評価するように設計されています。本研究では、標準（10項目）、視力、そして総合（標準と視力の平均）の3つの指標を報告します。ここで「標準」と「vision」は、CoTと直接回答の設定における最高スコアであり、これは元の論文と一致しています。  
   MathVista [163]:  
   MathVistaは、視覚的な文脈におけるMLLMの数学的推論能力を評価するためのベンチマークであり、代数、幾何学、統計といった推論の種類を網羅しています。本稿では、testminiセットにおけるスコアを報告します。  
   MATH-Vision [245]:  
   MATH-Visionは、実際の競技会で出題された3,040問の視覚的に文脈化された数学問題からなる高品質なデータセットです。testminiとフルセットの両方でのパフォーマンスを報告します。  
   MathVerse [299]:  
   MathVerseは、図表ベースの数学問題を解くMLLMを評価するための視覚的な数学ベンチマークです。2,612の高品質な多科目数学問題で構成されており、それぞれが視覚情報とテキスト情報の程度が異なる6つの異なるバージョンに変換されています。testminiセットにおけるパフォーマンスを報告します。  
   OlympiadBench [80]:  
   OlympiadBenchは、オリンピック競技大会や高考（Gaokao）の高難度の数学と物理の問題を収録した、バイリンガル対応のマルチモーダルベンチマークです。各問題には専門家レベルの段階的な推論が注釈として付与されており、論理的推論と問題解決能力を詳細に評価できます。このベンチマークは難易度が高く、明確に定義されたCoTプロンプトによってパフォーマンスが大幅に向上する可能性があります。  
   5.1.2 Evaluation Results  
   多分野的推論能力は、モデルが抽象的な概念を理解、処理、操作する能力を反映しており、複雑な問題解決や意思決定タスクにおいて極めて重要です。表6の左側のセクションでは、MMMU [289] や MMMU-Pro [290] を含む多分野的推論関連ベンチマークにおけるInternVL 2.5のパフォーマンスを比較しています。  
   ここでは、直接回答とCoT推論の両方の性能をテストし、より高いスコアを報告しています。結果は、我々のモデルがLLaVA-OneVision [124]、NVLM [50]、VILA 1.5 [143]、Qwen2-VL [246]などの既存のオープンソースモデルに対して有望な改善を達成していること、およびInternVL2シリーズの以前のバージョンと比較して顕著な進歩を達成していることを示唆しています。具体的には、InternVL2.5-78BはMMMU検証セットで70を超えるスコアを達成しており、InternVL2-Llama3-76Bに対して7.4ポイントの改善を示しています。これらの結果は、我々のモデルのパフォーマンスがGPT-4o [192]、Claude-3.5-Sonnet [8]、Gemini-1.5-Pro [200]などのいくつかの高度なクローズドソースモデルの性能に近づいていることを示しています。さらに、多数決投票により、CoT使用時のMMMUベンチマークにおけるInternVL2-Llama3-76Bのスコアは62.7から65.3に向上しました。InternVL 2.5でも同様の現象が見られ、テスト時間のスケーリングがMLLMのCoT推論を改善できることが示されています。  
   数学的推論は、より高次の推論能力を反映し、科学技術アプリケーションにおける MLLM の可能性を高めます。表 6 の右側のセクションでは、4 つのマルチモーダル数学ベンチマークでの InternVL 2.5 のパフォーマンスを示しています。これらの結果は、InternVL 2.0 に対して大幅な進歩を示しています。特に、InternVL2.5-78B は、MathVista テストミニセット [163] で 72.3% の精度を達成しました。さらに、挑戦的な OlympiadBench [80] では、InternVL 2.5 シリーズは 2.0 シリーズと比較して全体的な改善を示しました。この進歩の一部は、データフィルタリングパイプラインによるものと考えています。具体的には、2.0 モデルでは CoT 推論中に頻繁にデッドロックが発生し、正しい最終回答に到達できないことがわかりましたが、2.5 シリーズではこの問題が軽減されています。  
   5.2 OCR, Chart, and Document Understanding  
   5.2.1 Benchmarks  
   さまざまな OCR 関連データセットの包括的な評価を通じて、InternVL の OCR、チャート、ドキュメント理解機能を評価します。  
   AI2D [109]:  
   AI2Dは、5,000枚を超える小学校の理科の図表を収録したデータセットです。各図表には詳細な注釈と多肢選択問題が付されています。公平な比較のため、テストセットでは「マスクあり」と「マスクなし」の両方の設定で結果を報告しています。  
   ChartQA [181]:  
   ChartQAは、チャートやグラフなどのデータ視覚化を用いてモデルが解釈・推論する能力を評価することに焦点を当てたデータセットです。評価指標は、ChartQAにおける人間によるテストセットと拡張現実によるテストセットの両方における平均緩和精度です。  
   TextVQA [212]:  
   TextVQAは、画像内のテキストに基づく視覚的推論のベンチマークを目的として設計されたデータセットです。画像内のテキストを読み取って解釈し、関連する質問に正確に答えるモデルを必要とします。TextVQA検証セットにおけるVQAの精度を報告します。  
   DocVQA [42]:  
   DocVQAは、文書画像内のテキストから情報を理解・取得するモデルの能力を評価することを目的としたデータセットです。テストセットのパフォーマンスは、テキストの類似性を測定することで回答精度を測るANLS指標を用いて報告されます。  
   InfoVQA [183]:  
   InfographicVQAは、テキスト、グラフィック、視覚要素を組み合わせた複雑なインフォグラフィックを解釈・推論するモデルの能力を評価することを目的としたデータセットです。パフォーマンスは、テストセットにおいてANLS指標を用いて測定されます。  
   OCRBench [158]:  
   OCRBench は、テキスト認識、シーン テキスト VQA、ドキュメント VQA、キー情報抽出、手書きの数式認識の 5つのタスクにわたって MLLM の OCR 機能を最大 1000 スコアで評価します。  
   SEEDBench-2-Plus [125]:  
   SEED-Bench-2-Plusは、グラフ、地図、ウェブを対象とした2,300の人間による注釈付き質問を用いて、テキストリッチな視覚タスクにおけるMLLMを評価します。このデータセットにおける平均精度を報告します。  
   CharXiv [257]:  
   CharXivは、科学論文から2,323点のグラフを収録した包括的な評価ツールです。複雑な視覚情報の統合を必要とする推論問題（RQ）と、グラフの基本的な要素の理解を評価する記述問題（DQ）の2種類の質問が含まれています。  
   VCR [302]:  
   ビジュアルキャプション修復（VCR）は、画像内の部分的に隠されたテキストを、視覚コンテンツとテキストの両方を理解することで復元するタスクです。VCR-EN-EasyサブセットにおけるExact Match（EM）スコアとJaccard類似度を報告します。  
   5.2.2 Evaluation Results  
   表7は、OCR関連タスクにおけるInternVL 2.5とその前身であるInternVL 2.0、他の代表的なオープンソースモデル（例：Qwen2-VL [246]、LLaVA-OneVision [124]）、およびクローズドソースモデル（例：GPT-4o [192]、Claude-3.5-Sonnet [8]）との詳細な比較を示しています。ほとんどのベンチマークにおいて、InternVL 2.5はすべてのモデルスケールにおいてInternVL 2.0を大幅に改善し、現在の最先端モデルであるQwen2-VL-72B [246]に匹敵する性能を示しており、学習戦略とデータ品質の改善の有効性を反映しています。  
   しかし、2Bスケールでは、InternVL2.5-2BはQwen2-VL-2Bと比較して、TextVQA [212]、DocVQA [184]、InfoVQA [183]​​などのベンチマークにおいて性能が劣ります。データと学習戦略の違いに加え、モデルアーキテクチャも重要な役割を果たしている可能性があると考えられます。具体的には、Qwen2-VL-2Bは600Mのビジョンエンコーダと15Bの言語モデルを搭載しているのに対し、InternVL2.5-2Bは300Mのより小型のビジョンエンコーダと18Bの言語モデルを採用しています。より小規模なMLLM（例えば2B）では、同じ総パラメータバジェットを考えると、ビジョンエンコーダのサイズがOCR性能に比較的重要な役割を果たすようです。  
   さらに、InternVL 2.5は、視覚キャプション復元（VCR）タスクで優れたパフォーマンスを示しています[302]。2.5シリーズは、このタスクでInternVL 2.0よりも大幅に改善されており、2BモデルはEM / Jaccardスコア93.2 / 97.6に達し、前世代の32.9 / 59.2を大きく上回っています。この改善は、VCRトレーニングセットのごく一部（約22Kサンプル）を導入したことによるものです。VCRタスクでのモデルのパフォーマンスが低かったのは、OCR機能が不十分だったためではなく、タスク固有の指示に対する指示追従能力が不十分だったためであることがわかりました。これらの少数ながらも焦点を絞ったサンプルを活用することで、InternVL 2.5はVCRタスクの指示追従能力が著しく向上し、大幅なパフォーマンス向上をもたらしました。  
   5.3 Multi-Image Understanding  
   5.3.1 Benchmarks  
   さまざまなマルチイメージベンチマークにわたって、InternVL のマルチイメージ関係の認識と理解の能力を評価します。  
   BLINK [70]:  
   BLINKベンチマークは、古典的なコンピュータービジョンの課題に着想を得た14のタスクを通じて、MLLMの中核的な視覚認識能力を評価します。問題の半数以上は複数の画像に関するものです。結果は検証セットに基づいて報告されています。  
   Mantis-Eval [100]:  
   Mantis-Evalは、複数の画像を対象としたMLLMの推論能力を評価するための、綿密にキュレーションされた小規模ベンチマークです。サイズ認識や重量比較といったトピックを網羅し、人間による注釈が付けられた217の難易度の高い問題で構成されています。  
   MMIU [186]:  
   MMIUは、マルチイメージタスクにおけるMLLMのパフォーマンスを厳密に評価するために開発された、包括的なベンチマークスイートです。7つの異なるタイプのマルチイメージ関係と52の多様なタスクを網羅し、包括的な評価フレームワークを提供します。  
   MuirBench [241]:  
   MuirBenchは、MLLMの複数画像理解能力を評価するための包括的なベンチマークです。12のタスクと10種類の複数画像関係を網羅し、回答不可能なインスタンスバリアントを用いたモデル評価を強化します。  
   MMT-Bench [277]:  
   MMT-Benchは、運転やナビゲーションといったマルチモーダルタスクにおいてMLLMを評価します。特に、認識、推論、計画といったタスクに重点を置き、複数の画像理解を必要とするサブタスクを多数含みます。テストを迅速化するため、結果は検証セットで報告されます。  
   MIRB [308]:  
   MIRBは、MLLMが複数の画像を理解し推論する能力を評価するために設計されたベンチマークです。MIRBには、知覚、視覚世界知識、推論、マルチホップ推論の4つのタスクカテゴリが含まれます。報告されるパフォーマンスは、これら4つのカテゴリの平均スコアです。  
   5.3.2 Evaluation Results  
   インターネット上での情報交換において、複数画像コンテンツがますます一般的になるにつれ、モデルが複数の画像間の関係性を同時に理解・分析する能力を備えることが不可欠になっています。表8の左側では、InternVL 2.5の複数画像理解能力を、BLINK [70]、Mantis-Eval [100]、MMIU [186]、MuirBench [241]、MMT-Bench [277]、MIRB [308]という6つの異なるベンチマークで評価しています。これらのベンチマークは、画像間の推論、情報の統合、タスク固有の要件への対応など、幅広いスキルをテストします。  
   InternVL 2.5は、すべてのモデルスケールにおいてInternVL 2.0と比較して一貫した改善を達成しており、推論能力の向上と複数画像情報の統合の改善を反映しています。例えば、2Bスケールでは、InternVL2.5-2BはMantis-Eval（54.8 vs. 48.4）およびMuirBench（40.6 vs. 32.5）で大幅な改善を達成しています。これらの進歩は、セクション4.5で詳述するように、追加の複数画像データセットが追加されたことに大きく起因しています。これらのデータセットは慎重にキュレーションされ、高品質であり、複数の視覚入力を理解および推論するモデルの能力を向上させる上で重要な役割を果たしました。  
   より大規模な環境では、InternVL 2.5は大幅な進歩を示し、高度なクローズドソースモデルと競合する性能を達成しています。例えば、InternVL2.5-78BはMMIUで55.8のスコアを記録し、GPT-4oの55.7にほぼ匹敵しています。また、MMT-Benchでは70.8のスコアを記録し、GPT-4oの65.4を上回っています。これらの結果は、モデルサイズのスケーリングと、マルチイメージタスク向けに特別に調整された高品質のトレーニングデータを組み込むことの重要性を強調しています。しかし、BLINKとMuirBenchでは、私たちのモデルはGPT-4o [192]と比較して約5ポイントのパフォーマンスギャップを示しており、追加の高品質のマルチイメージトレーニングデータを組み込むなど、さらなる改善が必要であることを示唆しています。  
   5.4 Real-World Comprehension  
   5.4.1 Benchmarks  
   私たちは、現実的で複雑なタスクにおけるその能力を評価するために設計された一連の実際のベンチマークで、InternVL のパフォーマンスを評価します。  
   RealWorldQA [47]:  
   RealWorldQAは、MLLMの現実世界の空間理解能力を評価するために設計されたベンチマークです。様々な現実世界のシナリオから抽出された700枚以上の画像が含まれており、それぞれに質問と検証可能な回答が添付されています。  
   MME-RealWorld [306]:  
   MME-RealWorldは、5つの領域における43の実世界シナリオにわたる複雑な高解像度画像タスクにおけるMLLMを評価するためのベンチマークです。ここでは、データセットの英語版全セットをテストします。  
   WildVision [171]:  
   WildVision-Benchは、人間の嗜好を考慮したMLLMの実環境における評価を目的としたベンチマークです。実際のユーザーQAインタラクションから厳選された500件の高品質サンプルで構成されています。このベンチマークは、勝率指標を用いてモデルのパフォーマンスを定量化し、実用アプリケーションにおいて人間の期待に応える能力に関する洞察を提供します。  
   R-Bench [126]:  
   R-Benchは、MLLMの実世界の画像歪みに対する堅牢性を評価するために設計されたベンチマークであり、実用的なシナリオにおける破損画像の処理における耐性を測定します。MCQタスクの絶対的な堅牢性の総合スコアを報告します。これは、VLMEvalKitの「R-Bench-Dis」に相当する、低、中、高の難易度における平均スコアです。  
   5.4.2 Evaluation Results  
   現実世界環境の複雑性と動的性質を考慮すると、モデルは様々な困難な状況に対応できる堅牢性を備えていなければなりません。表8の右側に示すように、InternVL 2.5は、RealWorldQA [47]、MME-RealWorld [306]、WildVision [171]、R-Bench [126]を含む4つの現実世界理解ベンチマークにおいて最高のパフォーマンスを達成し、前バージョンのInternVL 2.0を大幅に上回っています。これは、InternVL 2.5が複雑で絶えず変化する現実世界のシナリオにおいて、より実用的に適用できる可能性を秘めていることを示しています。  
   RealWorldQA、MME-RealWorld、R-Benchなどの多肢選択式の質問を含むベンチマークでは、InternVL 2.5は強力な現実世界の知覚・理解能力を示しています。一方、WildVisionベンチマークでは、GPT-4o [192]を審査モデルとして使用し、さまざまなMLLMのパフォーマンスを参照モデルであるClaude-3-Sonnet [8]と比較して評価します。このベンチマークでは、モデルの出力品質とユーザーエクスペリエンスが重要な指標です。InternVL2.5-78Bは簡潔な回答を提供することには優れていますが、人間の好みに合わせてより長い回答を生成する際には依然としてギャップが見られます。具体的には、InternVL2.5-78Bのスコアは71.4であるのに対し、GPT-4oは80.6であり、ユーザーエクスペリエンスに顕著な違いが見られます。  
   これらの結果は、InternVL 2.5がほとんどのタスクにおいて正確かつ簡潔な回答を提供している一方で、よりパーソナライズされた詳細な回答を生成する点においては改善の余地があることを示しています。今後の研究では、自由回答形式のタスクや複雑なインタラクションにおけるモデルのパフォーマンス向上に重点を置き、人間の嗜好により合致し、GPT-4oとのユーザーエクスペリエンスのギャップを埋めることを目指します。  
   5.5 Comprehensive Multimodal Evaluation  
   5.5.1 Benchmarks  
   当社は、以下のベンチマークを通じて、InternVLの包括的なマルチモーダル機能を評価します。  
   MME [68]:  
   MMEは、MLLM向けに設計された初の包括的な評価ベンチマークです。物体の存在、計数、位置、色認識に加え、常識推論、数値計算、テキスト翻訳、コード推論を含む14のサブタスクにおいて、モデルの知覚能力と認知能力を評価します。全タスクの総合スコアを報告します。  
   MMBench [156]:  
   MMBenchは、20次元にわたる約3,000の多肢選択式質問を通じて、MLLMのマルチモーダル理解を評価します。英語版と中国語版の両方をサポートしており、テストセットにおけるモデルのパフォーマンススコアを提示します。  
   MMBench v1.1 [156]:  
   MMBenchと比較すると、MMBench v1.1は、ノイズや品質の低い質問を少数削除した改良データセットを採用しており、全体的なデータ品質がわずかに向上しています。英語版テストセットにおけるモデルのパフォーマンスを報告します。  
   MMVet [283]:  
   MMVetは、複雑なタスクにおけるMLLMの統合能力を評価するために設計されたベンチマークです。認識、知識、空間認識、言語生成、OCR、数学という6つのコアコンピテンシーを、16の統合タスクにわたって評価します。VLMEvalKitはこのベンチマークのスコアリングモデルとしてGPT-4-Turboを使用しており、公式評価サーバーと比較してスコアがわずかに低いことに注意してください。  
   MMVet v2 [284]:  
   MMVetを拡張したMMVet v2では、画像とテキストのシーケンス理解という新機能を備えた強化ベンチマークが導入され、インターリーブされたコンテンツを処理するモデルの能力を評価できるようになりました。ここでは、GPT-4-0613をスコアリングモデルとして採用した公式評価サーバーをスコアリングに利用しています。  
   MMStar [28]:  
   MMStarは、MLLMのマルチモーダル機能を評価するベンチマークです。高度な視覚・言語理解、データ漏洩の最小化、視覚依存性の重視に重点を置いた、厳選された1,500個のサンプルが含まれています。  
   5.5.2 Evaluation Results  
   MME [68]、MMBenchシリーズ[156]、MMVetシリーズ[283, 284]、MMStar [28]などの包括的なマルチモーダル評価ベンチマークは、多様なマルチモーダルタスクにわたってモデルのパフォーマンスを評価するための貴重で広く採用されているフレームワークを提供します。  
   表9の左側に示すように、InternVL 2.5モデルは、様々なモデルサイズにおいて、特に1B～8Bパラメータの小型モデルにおいて、従来のInternVL 2.0シリーズを一貫して上回る性能を示しています。例えば、英語と中国語の両方のタスクを評価するMMBench v1.0ベンチマークでは、InternVL 2.5モデルは大幅な改善を示しています。InternVL2.5-4Bは81.1/79.3というスコアを達成し、InternVL2-4Bの78.6/73.9を上回りました。また、InternVL2.5-8Bは84.6/82.6というスコアを達成し、InternVL2-8Bの81.7/81.2を上回りました。  
   MMVetシリーズのベンチマークにおいて、小規模モデルの性能は大幅に向上した一方で、最大モデルであるInternVL2.5-78Bは依然としてQwen2-VL-72B [246]を上回っていないことも注目に値します。現在、MMVet v2の最先端モデルは、GPT-4o [192]やClaude-3.5-Sonnet [8]といったクローズドソースモデルのままです。これは、マルチモーダル統合機能におけるオープンソースモデルとクローズドソースモデルのギャップを浮き彫りにしています。私たちは、これを将来の開発における重要な方向性と認識しています。  
   5.6 Multimodal Hallucination Evaluation  
   5.6.1 Benchmarks  
   私たちは、以下の4つの異なるベンチマークでInternVLの幻覚傾向を評価します。  
   HallusionBench [77]:  
   HallusionBenchは、言語幻覚や視覚錯覚といった課題に焦点を当て、はい/いいえの判断質問形式を用いてMLLMにおける画像文脈推論を評価するベンチマークです。パフォーマンスは、aAcc、fAcc、qAccの3つの指標の平均スコアを用いて報告します。  
   MMHal-Bench [223]:  
   MMHal-Benchは、MLLMにおける幻覚を評価するために設計されたベンチマークです。OpenImagesデータセットの画像から抽出された96個の難問と、それに対応する正解と画像コンテンツが含まれています。スコアリングはGPT-4oを用いて行われ、スコアは0から6の範囲です。  
   CRPE [250]:  
   CRPEは、多肢選択式の質問を用いて、オブジェクト間の関係性における幻覚レベルを測定するベンチマークです。このベンチマークの関係性サブセットにおける精度を報告します。  
   POPE [139]:  
   POPEは、MLLMにおける物体幻覚を評価するためのベンチマークであり、2値質問を用いて幻覚傾向を定量化・分析します。ランダム、ポピュラー、敵対的の3つのカテゴリーにおける平均F1スコアを報告します。  
   5.6.2 Evaluation Results  
   InternVLの性能を、4つの主要な幻覚評価ベンチマークであるHallusionBench [77]、MMHal [223]、CRPE [250]、POPE [139]で評価しました。これらのベンチマークは、マルチモーダルタスクにおける幻覚、つまり事実の不正確さの頻度を評価し、テキストや画像などの複雑な入力を処理する際のモデルの信頼性を測る指標を提供します。  
   InternVL 2.5モデルは、特に小規模なモデル（例：1B-8Bパラメータ）において、InternVL 2.0シリーズと比較して大幅な進歩を示しています。例えば、InternVL2.5-1BとInternVL2.5-2Bは、すべての幻覚ベンチマークでスコアが向上しており、1BモデルはHallusionBenchで39.0というスコアを達成しました。これは、以前のバージョンの34.0から大幅に向上しています。同様に、2Bモデルは42.6に向上し、以前の2Bモデルを約5ポイント上回りました。これらの結果は、マルチモーダルデータ処理における幻覚の低減において大幅な改善が見られたことを示しています。  
   最大規模のモデルであるInternVL2.5-78Bも改善が見られ、以前のバージョンや他の主要モデルと比較して幻覚が減少しています。HallusionBenchのスコアは57.4で、Qwen2-VL-72B（58.1）やGPT-4o（55.0）といった上位モデルと競合しています。InternVL2.5-78Bはこれらの幻覚評価ベンチマークにおいて比較的低い幻覚率を示していますが、実用上、長い応答を生成する際には、依然として多少の幻覚が避けられません。これは今後の課題として取り組む予定です。  
   5.7 Visual Grounding  
   5.7.1 Benchmarks  
   RefCOCO、RefCOCO+、およびRefCOCOgデータセットでの参照表現理解（REC）を介して、InternVLの視覚的グラウンディング機能を評価します。モデルは、与えられた説明から画像内のターゲットオブジェクトを識別します。  
   RefCOCO [108]:  
   COCO 上に構築されたこのデータセットには、50,000 個のオブジェクトに対する 142,210 個の参照表現を含む 19,994 枚の画像が含まれており、REC タスクのテスト A (人物中心) やテスト B (その他のオブジェクト) などのサブセットに分割されています。  
   RefCOCO+ [108]:  
   RefCOCOに似ていますが、絶対的な位置情報を伴わない属性ベースの説明を重視しています。19,992枚の画像と141,564の表現が含まれており、モデルは記述的な属性に重点を置く必要があります。  
   RefCOCOg [177]:  
   25,799 枚の画像と 95,010 の表現を含むこのデータセットには、より長く複雑な表現と、REC タスクで複雑な言語を管理するための難しいモデルが含まれています。  
   5.7.2 Evaluation Results  
   視覚的グラウンディングは、テキスト記述と視覚コンテンツを結び付け、正確なマルチモーダルインタラクションを可能にするために重要です。表10は、8Bスケールおよび78BスケールにおけるInternVL 2.5とその前身であるInternVL 2.0を、他の主要なMLLM（例：CogVLM-Grounding-17B [248]、Qwen2-VL [246]）および特殊なグラウンディングモデル（例：Grounding-DINO-L [153]、UNINEXT-H [267]、ONE-PEACE [247]）とRefCOCO [108]、RefCOCO+ [108]、およびRefCOCOg [177]データセットで評価した比較です。  
   InternVL2.5-8Bは前任者のパフォーマンスを向上させ、平均スコアが82.9から87.6に上昇し、Qwen2-VL-7B（87.6対87.9）と同等の結果を達成しましたが、グラウンディングの微調整と大規模モデル化の恩恵を受けているFerret-v2-13B [297]とCogVLM-Grounding-17B [248]にはわずかに遅れをとりました。大規模では、InternVL2.5-78Bは平均スコア92.3という最先端のパフォーマンスを達成し、InternVL2-Llama3-76Bより2.3ポイント向上し、Qwen2-VL-72B [246]を上回りました。これらの向上は、データとトレーニングの最適化の有効性を強調し、ローカリゼーション機能を大幅に向上させました。  
   5.8 Multimodal Multilingual Understanding  
   5.8.1 Benchmarks  
   InternVLのマルチモーダル多言語理解能力を、以下の3つの代表的なベンチマークを使用して評価します。  
   MMMB and Multilingual MMBench [218]:  
   MMMBは、6言語、15カテゴリ、12,000の質問を備えた大規模な多言語マルチモーダルベンチマークです。評価対象となる言語は、英語（en）、中国語（zh）、ポルトガル語（pt）、アラビア語（ar）、トルコ語（tr）、ロシア語（ru）です。Multilingual MMBenchは、多言語理解評価にGPT-4翻訳を用いることで、MMBench [156]をこれら6言語に拡張します。  
   MTVQA [227]:  
   MTVQAは、テキスト中心のビジュアル質問応答向けにカスタマイズされた多言語ベンチマークです。9言語にわたる高品質な専門家による注釈が含まれており、特に多言語環境における「ビジュアルとテキストの不整合」という課題に対処しています。MTVQAの平均スコアを報告します。  
   5.8.2 Evaluation Results  
   多言語対応能力はMLLMにとって極めて重要であり、その応用範囲の拡大と言語間コミュニケーションの向上に寄与する。グローバル展開においては、MLLMはリソースの多い言語と少ない言語の両方を効果的に処理する必要がある。表11に示すように、我々は3つの多言語ベンチマーク、MMMB [218]、Multilingual MMBench [218]、MTVQA [227]でモデルのパフォーマンスを評価した。  
   InternVL2.5-78BとQwen2-VL-72B [246]の比較では、学習データ、モデルアーキテクチャ、学習戦略の違いにもかかわらず、多言語性能は非常に類似していることが明らかになりました。これは、MLLMの多言語性能が、基盤となる言語モデルから大きく継承されていることを示唆しています。両モデルは同じLLMを共有しており、強力な多言語LLMがMLLMにおける効果的な多言語性能の基盤を形成していることを示唆しています。  
   5.9 Video Understanding  
   5.9.1 Benchmarks  
   Video-MME [69]:  
   Video-MMEは、フルスペクトル動画分析におけるMLLMを評価するためのベンチマークです。動画、字幕、音声などのマルチモーダル入力に加え、様々なドメインと期間にわたる多様な動画タイプに対応しています。このベンチマークでは、16フレーム、32フレーム、48フレーム、64フレームの4つの設定でテストを行い、最大値を報告します。「字幕あり」と「字幕なし」の両方の設定で結果を報告します。  
   MVBench [131]:  
   MVBenchは、オープンワールドにおけるMLLMの時間認識能力を包括的に評価するために設計された動画理解ベンチマークです。知覚から認知に至るまで、単一フレームでは効果的に解決できない20の難解な動画タスクを網羅しています。このベンチマークは16フレームでテストします。  
   MMBench-Video [65]:  
   MMBench-Videoは、MLLMのビデオ理解能力と時間的推論能力を評価するための定量的なベンチマークです。多様な領域、マルチショットの長尺ビデオ、幻覚、常識的推論、時間的推論といった機能をカバーしています。このベンチマークでは、16フレーム、32フレーム、48フレーム、64フレームの4つの異なる設定でテストを行い、最大スコアを報告します。  
   MLVU [315]:  
   MLVUは、3分から2時間までの長時間動画理解タスクにおけるMLLMを評価するために設計された包括的なベンチマークです。MLVUには、全体理解、単一詳細理解、複数詳細理解の3つのカテゴリーに分類された9つの評価タスクが含まれています。16フレーム、32フレーム、48フレーム、64フレームの4つの設定で評価を行い、最も高い「M-Avg」結果を報告します。  
   LongVideoBench [262]:  
   LongVideoBenchは、長フレーム入力を伴う参照推論タスクに焦点を当てており、モデルは参照クエリに基づいて詳細なマルチモーダル情報を正確に取得し、推論する必要があります。16フレーム、32フレーム、48フレーム、64フレームの4つの設定をテストし、検証セットで最良の結果を報告します。  
   CG-Bench [7]:  
   CG-Benchは、MLLMにおける長尺動画理解を評価するためのベンチマークです。既存のベンチマークとは異なり、CG-Benchは、質問への回答に必要な手がかりをモデルが検索する能力に焦点を当てています。1,219本の厳選された動画と12,000以上の質問と回答のペアが含まれています。真の動画理解を評価するために、2つの新しい手がかりベースの評価手法が導入されています。このベンチマークは32フレームでテストしました。  
   5.9.2 Evaluation Results  
   ビデオ理解は、MLLMの時間的およびマルチモーダル情報処理能力を評価する上で不可欠です。これを包括的に評価するために、我々はVideo-MME [69]、MVBench [131]、MMBench-Video [65]、MLVU [315]、LongVideoBench [262]、CG-Bench [7]の6つのベンチマークをテストしました。これらのベンチマークは、短いビデオの理解から長いビデオの推論まで、多様なタスクをカバーしています。  
   表 12 に示すように、InternVL 2.5 は、すべてのベンチマークで InternVL 2.0 よりも一貫して改善されています。たとえば、最小のモデルである InternVL2.5-1B では、Video-MME スコアが 42.9/45.4 から 50.3/52.3 に、MVBench スコアが 57.5 から 64.3 に向上しています。さらに、図 10 に示すように、InternVL 2.5 は、増加する入力フレームを処理する際のスケーラビリティが前身よりも優れていることがわかります。これらの改善は、次の 2つの主な機能強化によるものです。(1) より高品質のビデオ データが追加されたことで、モデルのビデオ理解機能が大幅に向上しました。(2) トレーニング フレームのサンプリング戦略を 4～24 フレームから 8～32 フレームに調整することで (図 5(c) に示すように)、より豊富なビデオ情報を処理するモデルの能力が向上しました。その結果、InternVL 2.0 モデルは通常 16 フレームまたは 32 フレームで最高のパフォーマンスを発揮しますが、入力フレームが増えるとパフォーマンスが低下します。一方、InternVL 2.5 は入力フレームを増やすことでメリットが得られ、長いビデオの理解においてより優れたスケーラビリティを示します。  
   私たちの最大のモデルであるInternVL2.5-78Bは、オープンソースモデルの中でトップクラスの性能を達成し、クローズドソースシステムのパフォーマンスに近づいています。オープンソースモデルと比較して、InternVL2.5-78BはMVBench（76.4 vs. 73.6）およびMMBench-Video（1.97 vs. 1.70）でQwen2-VL-72Bを上回っていますが、字幕付きのVideo-MMEスコアはわずかに低くなっています（74.0 vs. 77.8）。GPT-4o [192]やGemini-1.5-Pro [138]などのクローズドソースモデルに対して、InternVL2.5-78Bは競争力のあるパフォーマンスを示しています。Video-MMEでは72.1/74.0のスコアを獲得し、GPT-4o（71.9/77.2）やGemini-1.5-Pro（75.0/81.3）とほぼ同等のスコアを獲得しています。しかし、LongVideoBenchでは63.6という結果に終わり、Gemini-1.5-Pro（64.0）やGPT-4o（66.7）にわずかに及ばない結果となりました。これは、オープンソースモデルによる長時間動画理解には依然として課題が残っており、さらなる改善の余地があることを示しています。
2. Evaluation on Language Capability  
   LLM（法学修士）とMLLM（数学修士）の言語能力を徹底的に評価するため、多様なデータセットを用いて、5つのコアとなる側面からパフォーマンスを評価します。これらのベンチマークには、総合的な試験、言語と知識、推論、数学、コーディングといった課題が含まれます。  
   6.1 Benchmarks  
   Comprehensive Examination.  
   我々は様々な試験関連データセットを用いてLLMとMLLMを徹底的に評価した。(1) MMLU [85]には、人文科学、社会科学、STEMなど多様なトピックを網羅する57のサブタスクが含まれており、5ショットアプローチで評価される。(2) 中国の文脈に焦点を当てたCMMLU [127]は、一般領域と中国特有の領域にまたがる67のサブタスクを特徴としており、これも5ショット設定でテストされている。(3) C-Eval [96]には、4つの難易度にわたる52のサブタスクが含まれており、5ショット設定で評価される。(4) 中国の大学入試から派生したGAOKAO-Bench [304]は、主観的および客観的な質問タイプの両方を包括的にカバーし、客観的な質問は0ショット設定で評価される。  
   Language and Knowledge.  
   言語および知識ベースの評価については、能力をテストするために設計されたさまざまなデータセットを使用します。(1) TriviaQA [103]は、読解と、質問ごとに複数の回答があるオープンドメインQAタスクの両方を含み、0ショット設定で評価されます。(2) NaturalQuestions [117]は、専門家によって検証されたユーザー生成の質問を特徴としており、これも0ショット方式で評価されます。(3) C3 [219]は、0ショットの結果が報告されている、自由形式の多肢選択式の中国語機械読解データセットです。(4) RACE [118]は、12歳から18歳の中国の中学生と高校生の英語試験問題を含む読解データセットで、高校のサブセットの結果が0ショット設定で報告されています。  
   Reasoning.  
   推論能力を測定するために、我々は（1）WinoGrande [202]のようなデータセットを使用します。これは、代名詞の曖昧さ回避を必要とする44,000の多肢選択式の質問を通じて常識的な推論をテストし、0ショット設定で評価されます。（2）HellaSwag [292]は、自然言語推論のシナリオと4つの結果オプションを使用してモデルに課題を与え、最も論理的な結論を選択することを要求し、これも0ショット方式で評価されます。（3）BigBench Hard（BBH）[224]は、人間のパフォーマンスを超えるのが難しいように特別に選ばれた23のタスクで構成されており、推論の深さをさらに評価し、0ショットの結果が報告されています。  
   Mathematics.  
   数学の分野では、(1)GSM8K-Test[43]が約1,300の初等レベルの状況問題を提供しており、4ショット設定で評価されます。(2)MATH[86]は、代数や微積分などの科目にわたって12,500の高校競技レベルの問題を提示し、それぞれに詳細な解答があり、これも4ショット方式で評価されます。(3)TheoremQA[33]は、数学、物理学、金融などの分野で定理の適用を必要とする800のSTEMに焦点を当てた問題を紹介しており、0ショットの結果が報告されています。  
   Coding.  
   コーディング能力を評価するために、我々は以下のベンチマークを用いる。(1) HumanEval [31]：このベンチマークには164のPythonプログラミングタスクが含まれており、それぞれが詳細な仕様と対になっていて、コーディング性能を評価するための基準となる。これは4ショット設定で評価される。(2) MBPP [9]：974のエントリーレベルのプログラミングタスクで構成されるMBPPは、単純な算術問題からより複雑なシーケンス定義まで、幅広い課題をカバーしており、3ショット設定で評価される。(3) MBPP-CN [9]：多言語プログラミング能力を評価するために設計されたMBPPの中国語版。この拡張により評価範囲が広がり、言語的および文脈的多様性が含まれるようになり、0ショットの結果が報告されている。  
   6.2 Evaluation Results  
   MLLMの開発において、強力な純粋言語能力を維持することは依然として極めて重要です。InternLM2 [19]のアプローチに倣い、OpenCompassツールキット[46]を用いて、17の純粋言語ベンチマークにおいてモデルの性能を包括的に評価しました。これらのベンチマークは5つの主要グループに分類され、モデルの純粋言語能力を徹底的に評価しています。  
   結果によると、InternVL 2.0は、基礎LLM試験と比較して、純粋言語能力においてわずかに低下しています。例えば、InternVL2-2Bの平均スコアは39.2で、InternLM2-1.8B-Chatと比較して2.1ポイント低下しました。同様に、InternVL2-8Bの平均スコアは67.2で、InternLM2.5-7B-Chatと比較して2.3ポイント低下しました。  
   この問題を解決するため、高品質なオープンソースの純粋言語教育データを大規模に収集し、厳格なフィルタリングパイプラインを適用して低品質のサンプルを排除することで、データ全体の品質を向上させました。InternVL 2.5におけるこれらの改善により、言語パフォーマンスの低下が効果的に緩和され、モデルはいくつかのタスクにおいてオリジナルのLLMに匹敵、あるいは凌駕する性能を実現しました。これは、高品質な言語データによる補完と最適化によって、MLLMの純粋言語能力を維持できるだけでなく、マルチモーダルタスクのためのより強固な基盤を構築できることを示しています。  
   7 Evaluation on Vision Capability  
   このセクションでは、様々なドメインとタスクにおけるビジョンエンコーダの性能を包括的に評価します。評価は、(1) 画像分類（グローバルビューにおけるセマンティック品質の表現）、(2) セマンティックセグメンテーション（ローカルビューにおけるセマンティック品質の表現）という2つの主要カテゴリーに分かれています。このアプローチにより、InternViT の連続的なバージョンアップデートにおける表現品質を評価することができます。  
   7.1 Image Classification  
   7.1.1 Benchmarks  
   多様な画像分類データセットに対する包括的な評価を通じて、InternViT の全体的な意味品質を評価します。  
   ImageNet-1K [56]:  
   1,000 クラスにわたる 100 万枚を超える画像を含む、広く使用されている大規模なデータセット。画像分類モデルのベンチマークによく使用されます。  
   ImageNet-ReaL [16]:  
   ImageNet の検証セットの再ラベル付けバージョン。強化されたラベル付けプロトコルに従って、より正確で堅牢なマルチラベル注釈を提供します。  
   ImageNet-V2 [199]:  
   ImageNet-1K でトレーニングされたモデルの堅牢性を評価するために設計されたデータセット。オリジナルの ImageNet 方法論を使用して収集された新しいテスト画像が特徴です。  
   ImageNet-A [87]:  
   ResNetモデルによって誤分類されることが多い、自然発生的で加工されていない画像の難解なデータセットです。実世界における敵対的評価が難しい例に遭遇した場合のモデルの限界を浮き彫りにします。  
   ImageNet-R [84]:  
   芸術作品、スケッチ、玩具、彫刻、その他の創造的な表現からなる、ImageNetの200クラスにわたる3万枚の画像を含むレンディションデータセット。一般的なオブジェクトの抽象的な表現を認識するモデルの堅牢性を評価します。  
   ImageNet-Sketch [242]:  
   このデータセットには51,000枚のスケッチ画像が含まれており、ImageNetクラスごとに約50枚のスケッチが含まれています。クラス名の後に「sketch of」を付けてGoogle画像検索を行い、モデルが抽象的な手描き表現に一般化できるかどうかをテストしています。  
   7.1.2 Settings  
   この研究では、線形プロービング[32]と注意プーリングプロービングの2つの評価方法を採用してInternViTモデルの性能を評価した。  
   Linear Probing [32]:  
   この手法では、事前学習済みモデルを固定し、その上に線形分類器のみを学習させます。バックボーンを更新することなく学習済み特徴の品質を評価することで、画像分類などの下流タスクにおいて、単純な線形分類器で利用可能な意味情報を事前学習済みモデルがどれだけ効果的に捕捉しているかに関する知見が得られます。  
   Attention Pooling Probing:  
   一方、Attention Pooling Probingは、固定された特徴の上にAttention Pooling層を追加することでモデルを評価します。このアプローチにより、Attention Poolingはタスクに関連する特徴を動的に選択し、無関係な情報に干渉されることなく分類できるため、ビジョンエンコーダは最終層でより豊富な情報を保持できます。  
     
   どちらの実験でも、ImageNet-1K [56] をトレーニングセットとして使用し、ImageNet-1K検証セットといくつかのImageNetバリアント（ImageNet-ReaL [16]、ImageNet-V2 [199]、ImageNet-A [87]、ImageNet-R [84]、ImageNet-Sketch [242]）でモデルを評価し、ドメイン一般化能力をベンチマークしました。  
   モデルはSGDを最適化器として用い、ピーク学習率は0.2、モメンタムは0.9、重み減衰は無しで学習されました。コサイン学習率減衰スケジュールは、10回の学習エポック（ウォームアップエポックは1回）にわたって適用されました。入力解像度は448×448、パッチサイズは14、バッチサイズは合計1024です。学習中は、ランダムサイズクロッピングや水平反転などのデータ拡張技術が用いられています。これらの分類実験のコードとログは、GitHubリポジトリ1で公開されます。  
   7.1.3 Evaluation Results  
   表14に示すように、結果はInternViTのバージョン更新全体にわたって興味深い傾向を示しています。モデルの進化に伴い、線形プロービングの性能は大幅に低下し、すべてのバージョンで平均がグレーベースラインを下回っています。一方、アテンションプーリングプロービングは、多少の変動はあるものの、一貫してグレーベースラインを上回っています。この結果、InternViTのバージョン更新全体にわたって、平均スコア差（3.5から6.7）が拡大傾向にあり、Δで示されています。  
   これは、モデルの最終層の特徴が線形分離性が低下することを示唆しており、これはおそらく、表現がより複雑でオープンエンドな意味情報を捉えるよう進化するにつれて生じていると考えられます。注意プーリング機構は、この強化された表現空間から関連する特徴を効果的に選択し、線形分離性の低下による課題を相殺します。さらに、これらの知見は、InternViTが反復的な更新を通じて、壊滅的な忘却を生じることなく、主要な事前学習特性を維持していることを示唆しています。バージョンが進むにつれて、その表現はより多様化し、オープンセットな意味を捉え、汎化能力を高めます。これは、現実世界のタスクのために高度な抽象度が求められるMLLMにとって特に有益な利点です。  
   7.2 Semantic Segmentation  
   7.2.1 Benchmarks  
   2 つの代表的なセマンティック セグメンテーション データセット、ADE20K と COCO-Stuff-164K を使用して、InternViT のローカル ビューのセマンティック品質を評価します。  
   ADE20K [313]:  
   150のオブジェクトと背景のカテゴリーにわたる2万枚以上の画像とアノテーションを含む包括的なデータセット。シーン解析に広く使用されています。オブジェクトとパーツの両方に詳細なピクセルレベルのラベルが付与されており、様々なきめ細かなセグメンテーションタスクを容易に実行できます。  
   COCO-Stuff-164K [18]:  
   オリジナルのCOCO画像にピクセルレベルのアノテーションを追加し、91の「もの」クラス（草や空など）と80の「物」カテゴリ（人や車など）を追加することで、合計172のクラスをカバーしています。これらの包括的なラベルにより、このデータセットはシーン解析やセマンティックセグメンテーションのタスクをサポートし、画像分析におけるより豊かなコンテキスト理解を可能にします。  
   7.2.2 Settings  
   この研究では、線形プロービング、ヘッドチューニング、フルチューニングの3つの評価方法を採用して、セマンティックセグメンテーションタスクにおけるInternViTモデルのパフォーマンスを評価しました。  
   Linear Probing:  
   線形プロービングは、線形セグメンテーションヘッドを備えた固定バックボーンを適用し、学習した特徴の線形分離可能性に関する洞察を提供します。この手法は、最小限の適応でピクセルレベルの意味情報を評価するためのベースラインを提供しますが、エンコーダの複雑な特徴に対する能力を完全には捉えられない可能性があります。  
   Head Tuning:  
   ヘッドチューニングでは、UperNet [264] のヘッドは学習可能な状態のまま、InternViT は固定されます。これにより、モデルはより強力なヘッドを利用でき、線形分離性への依存を低減できます。この設定により、複雑でオープンエンドな特徴によって引き起こされる線形分離性の低下が緩和され、ビジョンエンコーダの能力をより正確に評価できるようになります。  
   Full Tuning:  
   完全なチューニングには、InternViTバックボーンとUperNet [264]セグメンテーションヘッドの両方を学習可能にすることが含まれます。これにより、モデルはすべてのレイヤーをターゲットタスクに適応させ、既存の線形分離可能性への依存を最小限に抑えることができます。この設定は、ビジョンエンコーダの視覚特徴抽出能力を評価するための別の視点を提供します。  
   AdamW [161] を用い、ピーク学習率は4e-5、多項式減衰スケジュールを採用した。フルチューニングでは、層ごとの学習率減衰（0.95）を適用する。重み減衰は、ヘッドチューニングとフルチューニングの両方で0.05に設定し、線形プローブでは減衰なしとした。入力解像度は504 × 504、パッチサイズは14、バッチサイズは16である。トレーニングは、1.5Kウォームアップ反復と80K合計反復で構成される。フルチューニングでは、ドロップパス率0.4を適用する。MMSegmentation [45] のデフォルトのデータ拡張を利用する。これらの実験に関連するすべてのコードとログはGitHub2で公開される。  
   7.2.3 Evaluation Results  
   表15に示すように、InternViTモデルのセマンティックセグメンテーション性能は、ADE20K [313]とCOCO-Stuff-164K [18]において、線形プロービング、ヘッドチューニング、フルチューニングの3つの構成で評価されている。結果は、モデルの特徴表現がバージョンアップごとにどのように進化するかについて明確な傾向を示している。  
   線形プロービングの結果は、モデルのバージョンが進むにつれてmIoUスコアが低下していることを示しています。平均スコアは、InternViT-6B-224pxの45.0から、InternViT-6B-448px-V2.5では37.5に低下しています。これは、InternViTのアップデートに伴い、特徴量の線形分離性が低下し、より複雑でオープンエンドな情報の捕捉へと移行していることを反映しています。  
   ヘッドチューニングでは、モデルは線形プロービングとは異なる傾向を示しています。他のすべてのバージョンのInternViTは、ベースラインのInternViT-6B-224pxのmIoUスコア51.9を上回り、パフォーマンスの低下は見られません。これによりΔ1値が上昇し、InternViT-6B-224pxの6.9からInternViT-6B-448px-V2.5の15.1に増加しました。Δ1の上昇は、特徴量の線形分離性は低下しているものの、その品質は維持され、複雑な情報を効果的に捉えていることを示唆しています。同様に、フルチューニングではΔ2値に見られるように、一貫した結果が得られます。Δ2がInternViT-6B-224pxの10.2からInternViT-6B-448px-V2.5の17.7に増加したことも、この傾向をさらに裏付けています。  
   全体的に見て、モデルバージョン間でΔ1とΔ2の値が増加していることは、単純で線形分離可能な特徴から、より複雑で非線形な表現への移行を示唆しています。この進化は、InternVLの開発においてバージョンが進化するにつれて、InternViTの視覚情報抽出能力が向上していることと一致しています。これは、ViTの増分学習戦略が、ビジョンエンコーダのオープンエンド特徴抽出能力を向上させる上で有効であることを示しています。  
   8 Conclusion  
   本研究では、InternVL 2.0のアーキテクチャを基盤とし、トレーニング、テスト戦略、データ品質を大幅に改善した、高度なオープンソースのマルチモーダル大規模言語モデル（MLLM）シリーズであるInternVL 2.5を紹介します。ビジョンエンコーダー、言語モデル、データセットサイズ、テスト時の構成を分析し、モデルのスケーリングとパフォーマンスの関係を体系的に調査します。多様なベンチマークによる広範な評価により、InternVL 2.5は、多分野推論、文書理解、ビデオ理解、多言語処理などのタスク全体で競争力のあるパフォーマンスを実現することが実証されています。特に、MMMUベンチマークで70％を超えた最初のオープンソースMLLMであり、オープンソースとOpenAI o1などの商用モデルとのギャップを縮めています。InternVL 2.5をコミュニティと共有することで、マルチモーダルAIの研究と応用を推進するための強力なツールに貢献したいと考えています。この研究を基にした今後の開発を楽しみにしています。

(補足)

Dynamic High Resolution

InternVL1.5から使用されている。

UReader [127]に着想を得て、入力画像の様々な解像度とアスペクト比に効果的に適応する動的高解像度学習アプローチを採用しました。この手法は、画像をタイルに分割する柔軟性を活用し、多様な画像解像度に対応しながら詳細な視覚情報を処理するモデルの能力を向上させます。この手法は主に以下のステップで構成されます。

Dynamic Aspect Ratio Matching.

処理中に自然なアスペクト比を維持するために、事前に定義されたアスペクト比のセットから最適なアスペクト比を動的にマッチングします。計算リソースが限られているため、トレーニング中に許可するタイルは最大12個です。したがって、このセットには、{1:1、1:2、2:1、3:1、…、2:6}など、1～12個のタイルで形成されるアスペクト比の35通りの組み合わせすべてが含まれます。マッチングプロセスでは、入力画像ごとにアスペクト比を計算し、絶対差を測定することで35通りの定義済みアスペクト比と比較します。複数の定義済みアスペクト比が一致する場合（たとえば、1:1と2:2）、入力画像の面積の2倍を超えないアスペクト比を優先して、低解像度画像の過度な拡大を防ぎます。

Image Division & Thumbnail.

適切なアスペクト比が決定されると、画像は対応する解像度にリサイズされます。例えば、800 × 1300 の画像は 896 × 1344 にリサイズされます。リサイズされた画像は、448 × 448 ピクセルのタイルに分割されます。タイルに加えて、画像全体のサムネイルも含まれており、全体的なコンテキストを捉えることができます。このサムネイルは 448 × 448 に縮小され、モデルがシーン全体を理解するのに役立ちます。そのため、トレーニング中はビジュアルトークンの数は 256 から 3,328 の範囲になります。テスト中はタイルの数が最大 40 に増加し、ビジュアルトークンの数は 10,496 になります。

Pixel unshuffle

ピクセルアンシャッフルは、ピクセルシャッフル（Shi et al. 2016）の逆操作です。ピクセルアンシャッフルは特徴を複数のサブ特徴に分割します。

サブ特徴は元の特徴の完全な情報を含みますが、解像度は低くなります。そのため、ピクセルアンシャッフルは情報損失を防ぎながら特徴のサイズを縮小するために用いられ、より少ない消費電力で複雑な処理が可能になります。

⇄

Pixel shuffleは、低解像度の画像を高解像度化する際に使用される。

Deconvolution(逆畳み込み)もPixel shuffleと用途は同じだが、Deconvolutionでは、格子状のノイズが発生することがある。Pixel shuffleはこの問題を解消させる手法。

Deconvolutionは、「不均一な重なり」が生じやすく、比喩的な意味でのペイントが、ある場所に他の場所よりも多く塗られてしまうことがあります。特に、カーネルサイズ（出力ウィンドウのサイズ）がストライド（上面の点の間隔）で割り切れない場合、逆畳み込みは不均一な重なりを引き起こします。原理的には、ニューラルネットワークは重みを慎重に学習することでこれを回避できますが（後ほど詳しく説明します）、実際にはニューラルネットワークがこれを完全に回避するのは困難です。

Pixel shuffleの特徴は出力がすべて元のテンソル成分によって決まっておりかつ元のテンソル要素が出力に寄与する部分がひとつしかない (i.e. overlap がない)という点です。これによって出力がゆがんでしまうことを防ぐことが出来ます。

例えば、画像を二倍のサイズにしたい時には、出力の直前で四枚の画像を作りその四枚を決まった位置関係で組み合わせて出力にします。

[ESPCNを使ってみた #Python - Qiita](https://qiita.com/oki_uta_aiota/items/74c056718e69627859c0)

[DeepLearningを用いた超解像手法/ESPCNの実装 #Python - Qiita](https://qiita.com/morimoris/items/08ec250ceb09a0004768)

[Deconvolution and Checkerboard Artifacts](https://distill.pub/2016/deconv-checkerboard/)

[[2203.08921] Hybrid Pixel-Unshuffled Network for Lightweight Image Super-Resolution](https://arxiv.org/abs/2203.08921)