InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models

Abstract

ネイティブマルチモーダル事前学習パラダイムを備えた、InternVLシリーズの大幅な進歩であるInternVL3を紹介します。テキストのみの大規模言語モデル（LLM）を視覚入力をサポートするマルチモーダル大規模言語モデル（MLLM）に適応させるのではなく、InternVL3は、単一の事前学習段階で、多様なマルチモーダルデータと純粋なテキストコーパスの両方からマルチモーダル機能と言語機能を共同で獲得します。この統合学習パラダイムは、従来のMLLMの事後学習パイプラインで一般的に発生する複雑さとアライメントの課題に効果的に対処します。パフォーマンスとスケーラビリティをさらに向上させるために、InternVL3は、拡張されたマルチモーダルコンテキストをサポートするために可変視覚位置エンコーディング（V2PE）を組み込み、教師あり微調整（SFT）や混合選好最適化（MPO）などの高度な事後学習手法を採用し、最適化された学習インフラストラクチャとともにテスト時間のスケーリング戦略を採用しています。広範な実証評価により、InternVL3は幅広いマルチモーダルタスクにおいて優れた性能を発揮することが実証されています。特に、InternVL3-78BはMMMUベンチマークで72.2というスコアを達成し、オープンソースMLLMにおける新たな最先端技術となりました。その性能は、ChatGPT-4o、Claude 3.5 Sonnet、Gemini 2.5 Proといった主要な独自モデルと高い競争力を維持しながら、高い純粋言語能力も維持しています。オープンサイエンスの原則に基づき、次世代MLLMの研究開発を促進するため、学習データとモデルの重みを全て公開していきます。

1. Introduction  
   （中略）  
   本稿では、InternVLシリーズ[21, 20, 18]の最新のマイルストーンであるInternVL3を紹介します。InternVL3は、ネイティブなマルチモーダル事前学習戦略を特徴としています。テキストのみの大規模言語モデルを事前に学習し、その後、視覚処理をサポートするためにマルチモーダルアライメントによって後付けするのではなく、InternVL3は、テキストのみのコーパスと多様なマルチモーダルデータセットの両方に同時に曝露することで、事前学習段階からマルチモーダル機能を学習します。この統合的なアプローチにより、モデルは言語能力とマルチモーダル能力をより効率的かつ統合的に同時に獲得することが可能になります。  
   InternVL3は、パフォーマンスとスケーラビリティの両方を強化する複数のイノベーションによって、さらに優れた性能を発揮します。より長いマルチモーダルコンテキストに対応するため、可変視覚位置符号化（V2PE）メカニズム[42]を採用しています。さらに、教師あり微調整（SFT）と混合選好最適化（MPO）[124]からなる高度な学習後戦略、テスト時間スケーリング戦略[125]、そして最適化された学習インフラストラクチャ[15]を組み合わせることで、InternVL3の効率とパフォーマンスが大幅に向上します。  
   包括的な実証的評価により、InternVL3は、多分野推論、文書理解、複数画像/ビデオ理解、現実世界の理解、マルチモーダル幻覚検出、視覚的グラウンディング、多言語機能など、幅広いタスクで前身モデル（例：InternVL2.5 [18]）を凌駕することが実証されています。特に、拡張されたドメイン固有のデータセットを組み込むことで、InternVL3はツールの使用、GUIエージェント、産業用画像分析、および空間推論においても顕著な改善を示し、InternVLシリーズが対処するマルチモーダルシナリオを大幅に拡張しています。Qwen2.5-VL [7]などの他のオープンソースMLLMと非常に競争力があり、クローズドソースモデル（例：ChatGPT-4o [98]、Claude-3.5 Sonnet [3]、Gemini-2.5 Pro [117]）と同等です。この汎用性は、MMMUベンチマーク[141]における72.2ポイントのパフォーマンスによって証明されており、オープンソースMLLMの新たな基準を確立しました。さらに、InternVL3は、同規模の他の高度なLLMに匹敵する言語機能を備えています。  
   オープンソース コミュニティ内でのさらなる進歩を促進するために、この作業と並行してトレーニング データ 1 とモデルの重みを公開し、次世代 MLLM の継続的な開発の透明性と再現性を確保します。
2. InternVL3  
   Model Architecture  
   InternVL3のアーキテクチャは、先行モデルと同じ基本的なフレームワークを踏襲しており、「ViT-MLP-LLM」パラダイム[66, 18, 41, 20]に準拠しています。詳細なアーキテクチャ仕様は表1にまとめられています。  
   後述するネイティブ事前学習パラダイムを使用すればMLLMをゼロから学習させることもできますが、ここでは計算コストを削減するため、ViTおよびLLMコンポーネントを学習済みモデルの重みで初期化することにしました。ビジョンエンコーダーには、InternViT-300MとInternViT-6Bの2つの構成があります。言語モデルについては、学習済みの大規模言語モデル（LLM）、具体的にはQwen2.5シリーズとInternLM3-8Bを活用します。重要なのは、LLMコンポーネントが、命令調整されたバリアントを使用せず、学習済みの基本モデルのみから初期化されることです。モデルで利用される多層パーセプトロン（MLP）は、ランダム初期化を備えた2層ネットワークです。InternVL2.5で採用されたアプローチに沿って、InternVL3は高解像度画像を処理するため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組み込んでいます。この操作により、ビジュアル トークンの数が元の値の 4 分の 1 に削減され、448×448 の各イメージ タイルが 256 個のビジュアル トークンで表されます。  
   Variable Visual Position Encoding.  
   InternVL3は、視覚トークンの位置増分をより小さく柔軟な単位で扱う可変視覚位置符号化（V2PE）[42]も統合しています。この改良により、位置ウィンドウを過度に拡張することなく、より長いマルチモーダルコンテキストの処理が容易になります。具体的には、MLLMの各学習サンプルは以下のように表現されます。  
     
   （中略）  
   Native Multimodal Pre-Training  
   言語事前学習とマルチモーダルアライメント学習を単一の事前学習段階に統合する、ネイティブマルチモーダル事前学習アプローチを提案する。従来のパラダイム（言語のみの大規模モデルを最初に学習し（通常は言語事前学習に続いて言語事後学習）、その後、追加のモダリティに対応するように適応させる）とは異なり、本手法は、事前学習プロセス中にマルチモーダルデータ（画像とテキスト、動画とテキスト、またはインターリーブされた画像とテキストのシーケンスなど）を大規模テキストコーパスとインターリーブすることで、統合最適化を実現する。この統合学習スキームにより、事前学習済みモデルは言語機能とマルチモーダル機能の両方を同時に学習することができ、最終的には追加のブリッジモジュールや後続のモデル間アライメント手順を導入することなく、視覚言語タスクを処理する能力を向上させることができる。  
   Multimodal Autoregressive Formulation.  
   （省略）  
   Joint Parameter Optimization.  
   （省略）  
   Data.  
   InternVL3で利用される事前学習データは、マルチモーダルデータと純粋言語データの2つのカテゴリーに大別されます。マルチモーダルデータセットは、既存のデータセットと新たに取得した実世界データの合成から構成されます。具体的には、画像キャプション、一般的な質問応答、数学、チャート、光学式文字認識（OCR）、知識グラウンディング、文書理解、マルチターン対話、医療データなど、多様な領域をカバーするInternVL2.5の事前学習コーパスを活用しています。全体的なデータ規模は増加していませんが、MLPモジュールの重みだけでなく、ViTおよびLLMコンポーネントに関連する重みも更新することで、このデータセットの有用性が大幅に向上しました。さらに、実世界アプリケーションにおけるモデルの一般化能力を高めるため、グラフィカルユーザーインターフェース（GUI）、ツールの使用、3Dシーンの理解、ビデオの理解に関連するタスクからの追加データが組み込まれています。  
   マルチモーダルデータセットに典型的に見られる比較的短く多様性に乏しいテキストコンテンツを補うため、純粋な言語データを事前学習プロセスに統合しました。これにより、モデルの言語理解および生成能力を維持し、強化することができます。言語コーパスは主にInternLM2.5の事前学習データに基づいて構築され、さらに様々なオープンソーステキストデータセット[8, 77, 79]で拡張されています。この拡張は、知識集約型タスクにおけるモデルのパフォーマンス、および数学的および推論タスクにおける熟練度を向上させることを目的としています。  
   これらの異種データソースのバランスをとることは複雑であるため、適切なサンプリング戦略を決定することは容易ではありません。InternVL3 では、マルチモーダルデータと言語データ間の最適なサンプリング比率を確立するために 2 段階の戦略を採用しています。最初に、マルチモーダルデータセットと言語データセットで別々のモデルをトレーニングし、対応するベンチマークでパフォーマンスを評価して、各モダリティ内で最適なサンプリング比率を特定できるようにします。次に、固定された合計トレーニング予算の下で、2 つのモダリティを組み合わせて、相対的なサンプリング比率を決定します。経験的研究によると、言語データとマルチモーダルデータの比率が 1:3 の場合、ユニモーダルベンチマークとマルチモーダルベンチマークの両方で最高の全体的なパフォーマンスが得られます。この構成では、トレーニングトークンの総数は約 2,000 億で、言語データから 500 億、マルチモーダルデータから 1,500 億になります。  
     
   Post-Training  
   ネイティブマルチモーダル事前学習の後、モデルのマルチモーダル会話および推論能力をさらに強化するために、2段階の事後学習戦略を適用します。この戦略は、教師あり微調整（SFT）と混合選好最適化（MPO）で構成されます。SFTフェーズでは、モデルは正の教師信号の下で高品質の応答を模倣するように学習されます。続くMPOフェーズでは、正と負の両方のサンプルからの追加の教師信号を導入することで、全体的な能力をさらに向上させます。  
   Supervised Fine-Tuning.  
   このフェーズでは、InternVL2.5 [18] で提案されたランダムJPEG圧縮、平方損失再重み付け、およびマルチモーダルデータパッキングといった技術が、InternVL3シリーズでも採用されています。InternVL3のSFTフェーズにおけるInternVL2.5と比較した主な進歩は、より高品質で多様な学習データの使用にあります。具体的には、ツールの使用、3Dシーンの理解、GUI操作、ロングコンテキストタスク、ビデオ理解、科学的なダイアグラム、創造的な文章作成、そしてマルチモーダル推論といった学習サンプルをさらに拡張しました。  
   Mixed Preference Optimization.  
   事前学習と SFT では、モデルは以前の真のトークンを条件として次のトークンを予測するように学習されます。しかし、推論中、モデルは各トークンをそれ自身の以前の出力に基づいて予測します。真のトークンとモデル予測トークン間のこの不一致により分布のシフトが生じ、モデルの Chain-of-Thought (CoT) 推論能力が損なわれる可能性があります。この問題を軽減するために、我々は混合選好最適化 (MPO) [124] を採用しています。これは、正サンプルと負サンプルの両方から追加の監督を導入し、モデルの応答分布を真の分布に合わせ、推論パフォーマンスを向上させます。具体的には、MPO の学習目的は、選好損失 ℒ p 、品質損失 ℒ q 、および生成損失 ℒ g の組み合わせであり、次のように定式化できます。  
     
   （中略）  
   Data.  
   SFTデータについては、InternVL2.5 [18] で使用されたコーパスをベースに、ツールの使用、3Dシーンの理解、GUI操作、科学的な図表、創造的な文章、マルチモーダル推論のサンプルを追加して学習コーパスを構築しました。その結果、学習サンプル数はInternVL2.5の1630万からInternVL3では2170万に増加しました。 MPOデータについては、MMPR v1.2 [124]で提案されたデータパイプラインとサンプルに基づいて選好ペアを構築します。これらのサンプルは、一般的なビジュアル質問応答（VQA）[43, 50, 90, 83, 127, 126]、科学[57, 16, 82]、チャート[91, 54, 11]、数学[72, 104, 10, 81, 55, 40, 147, 106]、OCR[92, 107, 9, 49, 96]、文書[24]など、幅広い分野をカバーしています。ロールアウトの生成には、SFTバージョンのInternVL3-8B、38B、78Bを使用しています。MPOフェーズでは、すべてのモデルが約30万サンプルで構成される同じデータセットでトレーニングされます。  
     
   Test-Time Scaling  
   テスト時間スケーリングは、LLM（学習者言語能力）およびMLLM（学習者言語能力）の推論能力を向上させる効果的な方法であることが示されている[108, 94, 87, 70, 120, 36, 152, 125]。本研究では、Best-of-N評価戦略を用い、推論および数学的評価における最適な応答を選択するための批評モデルとしてVisualPRM-8B [125]（マルチモーダルプロセスにて報酬を与えるのに特化したモデル）を採用した。  
   Visual Process Reward Model.  
   VisualPRMは、まず与えられたソリューションの各ステップに品質スコアを割り当て、次にこれらのスコアを平均して、このソリューションの全体スコアを取得します。このプロセスは、MLLMの生成能力を効果的に活用できるように、マルチターンチャットタスクとして定式化されています。画像 I 、質問 q 、およびこの質問に対するステップバイステップのソリューション s = { s 0 、 s 1 、 ⋯ 、 s n } ∈ 𝒮 の最初のステップ s 0 は最初のターンに含まれ、後続の各ターンで新しいステップが提示されます。トレーニング段階では、モデルは各ターンで与えられたステップの正確さを次のように予測する必要があります。  
     
   （中略）  
   Data.  
   VisualPRM400K [125]は、MMPR v1.2 [124]から収集されたマルチモーダル質問に基づいて構築されたVisualPRMの学習に使用されます。VisualPRM400Kのデータパイプラインに従い、InternVL3の8Bおよび38Bバリアントからサンプルロールアウトを取得することでVisualPRM400Kをさらに拡張します。
3. Experiments  
   （省略）
4. Conclusion  
   ネイティブマルチモーダル事前学習パラダイムを実装した、InternVLシリーズの大幅な進化版であるInternVL3を発表しました。事前学習段階で言語能力とマルチモーダル能力を共同学習することで、InternVL3は事後MLLM学習パイプラインに典型的に伴って発生する学習の複雑さと最適化の課題を回避します。拡張マルチモーダルコンテキストのための可変視覚位置符号化（V2PE）、教師あり微調整や混合選好最適化といった高度な事後学習戦略、そしてテスト時間のスケーリングを組み込むことで、InternVL3は堅牢な言語能力を維持しながら、幅広いマルチモーダルタスクにわたる新たなオープンソースベンチマークを確立します。特に、InternVL3-78BはMMMUベンチマークで72.2ポイントのスコアを達成し、従来のオープンソースMLLMを上回り、主要なプロプライエタリMLLM（Gemini-2.5 Proなど）とのパフォーマンス格差を縮小しました。マルチモーダル大規模言語モデルにおけるコミュニティ主導のイノベーションを促進するという当社のコミットメントに沿って、InternVL3のトレーニングデータとモデルの重みを公開し、急速に進化するこの分野におけるさらなる研究開発を促進していきます。