記事内容

カメラ越しに映るストリーミングビデオをリアルタイムで理解できる新しいAIシステム「InternLM-XComposer2.5-OmniLive」（IXC2.5-OL）を開発しました。

このシステムは3つの主要なモジュールで構成されています。1つ目は「ストリーミング認識モジュール」で、リアルタイムで映像や音声を処理し、重要な情報をメモリに保存します。

2つ目は「マルチモーダル長期メモリモジュール」で、短期記憶と長期記憶を統合し、より効率的な情報検索を可能にします。3つ目は「推論モジュール」で、ユーザーからの質問に対して記憶と認識の情報を組み合わせて回答を生成します。

IXC2.5-OLは、従来のAIシステムが抱えていた「認識と思考を同時に行えない」という制限を克服し、人間のように同時並行で処理できることを目指しています。

IXC2.5-OLは、長時間の映像理解においても優れた性能を発揮し、オープンソースモデルの中で最高水準の結果を達成しています。具体的な性能評価では、中国語音声認識のWenetspeech、英語音声認識のLibriSpeechといったベンチマークで競争力のある結果を示しました。また、映像理解のベンチマークMLVUでは66.2%、StreamingBenchでは73.79%という高いスコアを記録し、オープンソースモデルとしては最高性能を達成しています。

InternLM-XComposer2.5-OmniLive: A Comprehensive Multimodal System for Long-term Streaming Video and Audio Interactions

Abstract

人間の認知と同様に、長期間にわたって環境と対話できる AI システムを作成することは、長年の研究目標でした。マルチモーダル大規模言語モデル (MLLM) の最近の進歩により、オープンワールドの理解が大きく前進しました。ただし、継続的かつ同時のストリーミング知覚、記憶、推論の課題は、ほとんど未開拓のままです。現在の MLLM はシーケンスツーシーケンス アーキテクチャによって制約されており、入力を処理して応答を同時に生成する能力が制限され、知覚しながら考えることができないのと同じです。さらに、履歴データを保存するために長いコンテキストに依存することは、すべての情報を保持するとコストがかかり非効率的になるため、長期的な対話には実用的ではありません。したがって、このプロジェクトでは、すべての機能を実行する単一の基盤モデルに依存するのではなく、専門のジェネラリスト AI の概念からインスピレーションを得て、ストリーミング知覚、推論、記憶の分離メカニズムを導入し、ストリーミング ビデオとオーディオ入力とのリアルタイムの対話を可能にします。提案されたフレームワーク InternLM-XComposer2.5-OmniLive (IXC2.5-OL) は、3つの主要モジュールで構成されています。(1) ストリーミング知覚モジュール: マルチモーダル情報をリアルタイムで処理し、重要な詳細をメモリに保存し、ユーザーのクエリに応じて推論をトリガーします。(2) マルチモーダル長期メモリ モジュール: 短期メモリと長期メモリを統合し、短期メモリを長期メモリに圧縮して、効率的な検索と精度の向上を実現します。(3) 推論モジュール: 知覚モジュールとメモリ モジュールと連携して、クエリに応答し、推論タスクを実行します。このプロジェクトは、人間のような認知をシミュレートし、マルチモーダル大規模言語モデルが長期にわたって継続的かつ適応的なサービスを提供できるようにします。 InternLM-XComposer2.5-OmniLive (IXC2.5-OL) のすべてのコードとモデルは、https://github.com/InternLM/InternLM-XComposer/tree/main/InternLM-XComposer-2.5-OmniLive で公開されています。

1 Introduction

人間の認知と同様に、長期間にわたって環境を理解し、環境と対話できるAIシステム[55]を開発するという目標は、何十年にもわたって研究の中心となってきました。大規模データコーパス[69、54、95、112]とマルチモーダル大規模言語モデル[83、84、107]の台頭により、自由形式のマルチモーダル質問応答が大きく進歩しました。Mini-Omni [123]、VideoLLM-Online [12]、VITA [38]などの最近の開発は、より自然で没入感のあるオンラインインタラクションを可能にするための顕著な進歩を遂げました。しかし、単一のデコーダーのみの大規模言語モデルアーキテクチャの固有の制限により、継続的なインタラクションが可能なシステムを作成するという課題が残っています。

既存のアーキテクチャ [149, 123, 12, 38] は、リアルタイムおよび長期のストリーミング知覚、推論、およびメモリにおいて重大な制限に遭遇します。現在の MLLM で使用されているシーケンスツーシーケンス デコーダーのみのアーキテクチャでは、知覚 (たとえば、見る、聞く) と思考の切り替えが強制され、入力と出力の同時処理が制限されます。さらに、既存の研究 [145, 118, 33] は、コンテキスト ウィンドウ内でのマルチモーダル メモリの統合に依存しています。履歴情報を保存するために長いコンテキストに依存することは、特に継続的な AI 支援を必要とするシナリオでは、長期使用には実用的ではありません。ビデオ ストリームなどのマルチモーダル データは、数時間以内に数百万のトークンをすぐに蓄積できるため、複数日間のサービスにわたってコンテキストを維持することは実用的ではありません。コンテキスト内のすべての履歴の手がかりを保存するコストと非効率性により、継続的かつ長期的なサービスを提供するシステムの能力がさらに制限されます。対照的に、人間の脳は知覚と認知を難なく統合し、長期にわたるマルチモーダル記憶を保持することができます。これは、人間の脳皮質の機能分割設計と密接に関係していると考えられており、皮質の異なる領域が知覚、記憶、認知などの異なるタスクを担当しています。

我々は、特化型ジェネラリストAI[146]のパラダイムに着想を得て、ストリーミング知覚、推論、記憶のそれぞれに特化したジェネラリストモデルを融合したシステムInternLM-XComposer2.5-OmniLive（IXC2.5-OL）を提案する。このシステムは、AIモデルが環境と継続的に関わりながら、時間の経過とともに観察を保持できるように設計されています。短期および長期のマルチモーダルメモリを統合することにより、私たちのアプローチは人間のような認知をエミュレートし、より動的で持続的なインタラクションを可能にすることを目指しています。

図 1 に示すように、IXC2.5-OL システムは 3つの主要モジュールで構成されています。(1) ストリーミング知覚モジュール: このモジュールは、マルチモーダル情報ストリームをオンザフライ（休む間もなく）で処理します。知覚の精度と効率を確保するために、ビデオ ストリームとオーディオ ストリームは別々に処理されます。ライブ ビデオ知覚モデルがビデオ ストリームを処理し、情報をエンコードして重要な詳細をメモリに保存します。一方、オーディオ モデルは、人間の発話の内容や、吠え声、ノック、口笛などの他の音を認識します。人間のクエリが発生すると、推論プロセスがトリガーされます。(2) マルチモーダル長期記憶モジュール: このコンポーネントは、長期記憶と短期記憶の両方を統合し、詳細な短期情報と長期履歴キューの検索を可能にします。短期記憶をより情報量の多い長期記憶に継続的に圧縮して、検索の効率と精度を高めます。(3) 推論モジュール: 知覚モジュールによってアクティブ化される推論モジュールは、クエリを処理し、推論タスクを実行します。最も多くのモデルパラメータを持つコンポーネントとして、システムの深い認知プロセスの中核として機能します。

提案されたシステムは、AI に知覚、思考、記憶を同時に行う能力を与えます。交互に行う知覚と推論の限界を克服することで、IXC2.5-OL は継続的かつ適応的なサービスと長期的な AI サービスの提供を目指します。提案されたシステムは、AI アシスタントのパフォーマンスを向上させるだけでなく、動的な環境と継続的に対話して適応できるより幅広い AI アプリケーションにも貢献します。

IXC2.5-OLは、オーディオとビデオの両方のベンチマークで強力なパフォーマンスを発揮します。オープンソースモデルの中で、IXC2.5-OLは、中国語のWenetspeech [140]や英語のLibriSpeech [87]などのオーディオ認識（ASR）ベンチマークで競争力のある結果を達成しています。ビデオ理解ベンチマークでは、IXC2.5-OLは10B未満のパラメータを持つモデルの中で最先端の結果を達成し、MLVU [155]で66.2％のM-Avg、MVBench [62]で68.7％の全体精度を達成しました。さらに、Video-MME [37]（60.6％）とMMBench-Video [34]（1.42）で競争力のあるパフォーマンスを発揮します。最近のストリーミングビデオベンチStreamingBench [67]では、IXC2.5-OLはオープンソースモデルで新しいSOTA結果（73.79％）を達成し、リアルタイムビデオインタラクションの優れた能力を際立たせています。

マルチモーダル ストリーミング インタラクション コミュニティの開発を促進するために、モデル パラメータとともに、Web フロントエンドとバックエンドの両方のコードを含む推論およびデプロイメント ソース コードもリリースされました。IXC2.5-OL のすべてのコードとモデルは、https://github.com/InternLM/InternLM-XComposer/tree/main/InternLM-XComposer-2.5-OmniLive で公開されています。

図 1: 人間のような認知と専門的汎用 AI にヒントを得て、InterLM-XComposer2.5-OmniLive (IXC2.5-OL) を紹介します。これは、(1) ストリーミング ビデオとオーディオの入力をサポートするストリーミング認識モジュール、(2) 短期記憶を長期記憶に圧縮するマルチモーダル長期記憶モジュール、(3) 検索された記憶に基づいてクエリに回答する推論モジュールとのリアルタイムのインタラクションを可能にするシステムです。

2 Related Works

MLLMs for Text-Image Conversation.

大規模言語モデル（LLM）[7、86、81、24、51、109、110、46、108、136、5、90、9]は、言語の理解と生成における優れた能力により大きな注目を集めています。この成功を基に、LLMとビジョンエンコーダー[91、141、104、138、85、135、74、21、14、70、4、115、29、10、75、22、150]を統合した大規模視覚言語モデル（LVLM）[147、82、19、17、18、31、36、156、28、147、6、56、88、132、3、30、68]が開発され、視覚コンテンツを理解する能力が拡張され、テキストと画像の会話などのアプリケーションが可能になりました。以前のLVLMは主に単一画像、複数ラウンドの会話用に設計されていましたが、最近の進歩[1、4、153、103、68、48、30、148、58]により、複数画像入力を処理および理解する機能が拡張されました。

MLLMs for Video Understanding.

画像理解の進歩に加えて、MLLM の分野ではビデオ解析の取り組みが拡大しています [61, 80, 73, 32, 34, 98, 100, 127, 113]。ビデオ入力の複雑さに対処するために、既存のアプローチでは、スパースサンプリングや時間プーリング [66, 79, 77, 44, 133]、圧縮ビデオトークン [60, 144, 49, 119, 63, 94, 16]、メモリバンク [98, 43, 100, 89, 145, 118, 33] などの技術を活用しています。さらに、いくつかの方法では、ビデオ理解の橋渡しとして言語を利用しています [50, 45, 142]。これらのビデオ固有の戦略以外にも、ビデオ分析は、サンプリングされたビデオフレームから生成された高解像度の合成画像を解釈することとして捉えることもできます [52、126、149]。最近の進歩 [12、117、120、145] は、オンラインビデオ理解にますます重点が置かれており、AIがビデオストリームをリアルタイムで処理して環境をその場で理解するという現実世界のシナリオをシミュレートすることを目指しています。ただし、既存のソリューションでは、知覚、記憶、推論を同時に実行する機能がまだ不足しており、一貫性のある長期的な人間とAIの相互作用への適用が制限されています。

MLLMs for Audio Understanding.

音声理解は、シーケンスツーシーケンス（Seq2Seq）タスク [93] として効果的にモデル化することができ、オーディオトークナイザーとエンコーダー [105, 143, 25, 137] を組み込むことで、大規模な言語モデルとの強力な統合が可能になります。最近の研究では、オーディオ入力の受信に加えて、話者が自由に中断できるストリーミングデュプレックス音声モデル [114, 134, 78, 116] が調査されています。オーディオテキストモデルを超えて、新しい研究では、オーディオビジュアルモデル [96, 59] と、オーディオ、ビジュアル、テキストのモダリティを処理する統合アーキテクチャ [139, 38, 64] を詳しく調べています。

MLLMs for Omni-Modal Understanding.

複数のモダリティを単一のオムニモーダル基盤モデルに統合することは、有望な研究方向です。既存の研究 [42、139、121、64、38、13、124、102] では、通常はビデオとオーディオを組み合わせたオムニモーダル入力を処理して、さまざまな形式の出力を生成できるモデルが検討されています。これらの出力には、テキスト [42、64、38]、オーディオ [13、102、124]、オムニモーダルコンテンツ [139、121] が含まれます。IXC2.5-OL の現在の設計では、オーディオとビデオのモダリティを個別に処理して、共同トレーニング中の潜在的な影響を軽減しています。将来のバージョンでは、すべてのモダリティにわたる共同トレーニングがモデルに組み込まれ、シームレスなオムニモーダリティ統合が可能になります。

3 Method

図 2: InternLM-XComposer2.5-OmniLive のパイプライン (IXC2.5-OL)。IXC2.5-OL は、1) ストリーミング認識モジュール、2) マルチモーダル ロング メモリ モジュール、3) 推論モジュールという 3つの同時モジュールによって構成されるリアルタイム インタラクション システムです。

セクション 1 で簡単に紹介したように、IXC2.5-OL には 3つの独立したモジュールがあります。1) オンザフライの視覚および音声情報処理用のストリーミング知覚モジュール、2) メモリの統合と検索用のマルチモーダルロングメモリモジュール、3) 知覚およびメモリモジュールから情報を収集し、クエリを処理して推論タスクを実行する推論モジュールです。すべてのモジュールは同時に動作し、非同期で相互作用します。

3.1 Streaming Perception Module

IXC2.5-OL は、自然言語のほかに、ビデオとオーディオをネイティブに処理できます。これを実現するために、ストリーミング認識モジュールには、オーディオ翻訳モジュールとビデオ認識モジュールが含まれています。

オーディオ翻訳モジュールには、オーディオエンコーダー、オーディオプロジェクター、およびSmall Language Model (SLM) が含まれています。オーディオエンコーダーは、入力されたオーディオサンプルを高次元機能にエンコードし、オーディオプロジェクターはさらにその機能をSLMの入力空間にマッピングします。SLMは、オーディオのクラス（笑い、拍手、雨など）とオーディオ内の自然言語（自動音声認識）の両方を出力します。実際には、オーディオエンコーダーとしてWhisper [92]モデルを使用し、SLMとしてQwen2-1.8B [128]を使用します。トレーニングは2段階で構成され、トレーニングデータは表1に示されています。

ビデオ知覚モジュールは、マルチモーダルロングメモリモジュールに粗粒度の視覚情報を提供します。リアルタイムのビデオ入力ストリームを処理し、各フレームを意味的特徴にエンコードします。効率性を高めるために、実際にはOpenAI CLIP-L/14 [91]を使用しています。

表 1: オーディオ翻訳モジュールの事前トレーニングと教師あり微調整 (SFT) で使用されるデータセットの概要。事前トレーニング ステージでは、GigaSpeech および WenetSpeech データセットを使用して、自動音声認識 (ASR) タスクのみに焦点を当てています。SFT ステージには、さまざまなデータセットを活用した ASR タスクとオーディオ分類 (CLS) タスクの両方が含まれます。CommonVoice の場合、英語と中国語の分割のみを使用します。さらに、CLS タスクには、独自に構築した 475 個の「無音」サンプルが使用されます。

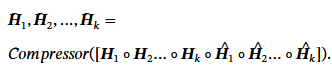
3.2 Multi-modal Long Memory Module

マルチモーダルロングメモリモジュールは、非常に長いビデオ入力を処理するためのコア設計であり、推論モジュールがコンテキストウィンドウから何百万ものトークンを取り除くのに役立ちます。これは、ビデオクリップを短期記憶にエンコードし、それを長期記憶に統合するビデオストリーミング[89]と同様のアイデアを共有しています。与えられた質問で、推論モジュールに最も関連性の高いビデオクリップを取得しました。正式には、マルチモーダルロングメモリモジュールは3つのタスクでトレーニングされます。

**Video Clip Compression(ビデオクリップの圧縮)**:知覚モジュール𝑭k ∈ ℝTN×Cから抽出されたk番目のビデオクリップの特徴を使用して、空間ダウンサンプリングによって短期メモリ𝑯k ∈ ℝTP×Cを初期化し、グローバルメモリ𝑯k^ ∈ ℝ1×Cを初期化します。LLMの自己回帰と特徴集約の性質によって圧縮を実現します。



**Memory Integration(メモリ統合)**:短期メモリは各短いビデオ クリップの詳細情報を表しますが、モデルにはビデオのマクロ ビューがまだありません。このため、ビデオ クリップのリストの短期メモリとグローバル メモリを、次の形式でコンプレッサーによって長期メモリに統合します。



𝑯¯=[𝑯¯1, 𝑯¯2, …, 𝑯¯k]∈ℝk×Cはビデオを高圧縮して表し、これを長期記憶と呼びます。

**Video Clip Retrieval(ビデオ クリップの取得)**:ユーザーが質問すると、マルチモーダル ロング メモリ モジュールは質問に関連するビデオ クリップを取得し、ビデオ クリップとその短期メモリの両方を推論モジュールに提供します。実際には、まず質問をメモリの特徴空間にエンコードします。長期メモリをトークン化された質問と連結してコンプレッサー入力とし、出力特徴の最後のトークンをメモリ空間に合わせた質問特徴と見なします。次に、質問特徴と各ビデオのグローバル メモリの類似性を計算し、推論モジュール用に最も関連性の高いクリップを選択します。

**Implementation Detail(実装の詳細)**:LLMとしてQwen2-1.8B [128]を使用し、前述の3つのタスク用に数種類のトレーニングデータを構築します。表2に示すように、VideoStreaming [89]で設計されたものと同じプレフィックスキャプションタスクを使用して、複数のソースからの短いビデオキャプションデータでビデオクリップ圧縮タスクをトレーニングします。メモリ統合タスクとビデオクリップ検索タスクについては、既製のビデオグラウンディングデータの他に、「意味暗黙の質問」と「参照暗黙の質問」という2つの独自のタスクのデータも構築します。

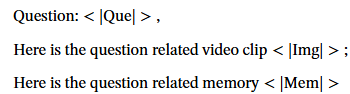
「セマンティクス暗黙の質問」とは、質問が直接オブジェクトを指し示すのではなく、オブジェクトの使用法や意味に言及しており、モデルは暗黙の質問を理解することでオブジェクトを見つけ出す必要があることを意味します。たとえば、ユーザーが「今日の天気はどうですか？」と質問すると、モデルは過去のビデオ ストリームで傘やサングラスなどの天気関連のオブジェクトを見つけ出す必要があります。別の例としては、「お腹が空いた。サンドイッチを温められる場所は？」という質問の場合、モデルは以前に見たことのある電子レンジを見つける必要があります。

「参照暗黙の質問」とは、質問が名詞ではなく代名詞を使用することを意味します。たとえば、「これは何ですか」は、正確なオブジェクトについては言及していませんが、モデルが現在のフレームを取得する必要があることを意味します。

どちらの種類の暗黙の質問も現実世界のコミュニケーションでよく使用されますが、現在のモデルではこれらを処理できなかったため、対応するトレーニング データを構築して、モデルにこれらの機能を強化しました。

3.3 Reasoning Module

推論モジュールは、改良版の InternLM-XComposer2.5 (以下、簡略化のため IXC2.5 と表記) によって初期化され、メモリ プロジェクターを追加してメモリ機能を IXC-2.5 と整合させます。特定の質問と、メモリ モジュールによって提供される視覚情報とメモリ情報の両方に対して、入力を次のように定式化します。



実際の使用では、回答すべきではないノイズの多い入力（ユーザーが「enn…」または「ok…」と言うなど）が存在するため、モデルは顕著性を維持し、次の質問を待つ必要があります。これを実現するために、各質問に回答する必要があるかどうかを判断するための「指示予測」プロセスを追加します。

3.4 System Pipeline

図 3 に示すように、システムはフロントエンド、SRS サーバー、およびバックエンド サーバーで構成されます。

図 3: IXC2.5-OL のシステム パイプライン。システムは、フロントエンド、SRS サーバー、およびバックエンド サーバーで構成されます。フロントエンドは、ビデオおよびオーディオ ストリームのキャプチャと、バックエンド サーバーからのオーディオの再生に使用されます。SRS サーバーは、ライブ ストリームの管理に使用されます。バックエンド サーバーは、オーディオとビデオの読み取り、メモリの抽出、および質問への回答を担当します。図の緑色のボックスは、スレッドまたはプロセスを表します。

Frontend.

JavaScript で開発されたフロントエンド アプリケーションにより、カメラとマイクがビデオとオーディオのストリーム入力をキャプチャし、それを SRS サーバーにプッシュできるようになります。同時に、バックエンドとの WebSocket 接続を確立して、オーディオ出力と割り込み信号をリッスンします。オーディオ データが受信されると、フロントエンドはそれを再生します。割り込み信号を受信すると、フロントエンドはオーディオの再生を一時停止し、保留中のオーディオを破棄します。

SRS Server.

SRS (Simple Realtime Server) は、RTMP、WebRTC、HLS、HTTP-FLV、SRT などの多数のリアルタイム ストリーミング プロトコルをサポートする、シンプルで効率的なリアルタイム ビデオ サーバーです。オーディオ ストリームとビデオ ストリームを確実に受信および配信できることで知られています。

Backend Server.

フロントエンドとの WebSocket 接続を確立した後、バックエンドは SRS サーバーからストリーミングをプルし、オーディオとビデオを読み取るための個別のスレッドを開始します。

オーディオ読み取りスレッドは、オーディオ ストリームを 4096 ビットのチャンクに分割し、オーディオ キューにエンキューします。音声アクティビティ検出 (VAD) [40] スレッドは、オーディオ キューからデータを継続的に読み取り、音声アクティビティの開始と終了を検出します。音声アクティビティの開始を検出すると、バックエンドはフロントエンドに割り込み信号を送信して現在再生中のオーディオを一時停止し、同時にビデオ プロセスにバックアップ信号を送信して、現在のメモリ状態を保存するように指示します。音声アクティビティの終了を検出すると、音声セグメント全体が ASR Todo キューにエンキューされます。ASR スレッドは、ASR Todo キューからオーディオ セグメントを継続的に読み取り、それらに対してバックグラウンド ノイズ分類と音声認識を実行し、結果を LLM Todo キューにエンキューして LLM で使用します。

ビデオ読み取りスレッドは、1 秒あたり 1 フレームの速度でビデオ フレームを読み取り、フレーム キューにエンキューします。コンプレッサー プロセスは、キューからビデオ フレームを読み取り、認識し、関連するメモリを抽出して保存します。VAD スレッドからバックアップ信号を受信すると、コンプレッサー プロセスは、後で取得できるように現在のメモリ状態を保存します。

LLMプロセスは、LLM Todoキューからテキストを読み取り、それがモデルからの応答を必要とする命令であるかどうかを判断します。命令として識別されたテキストの場合、コンプレッサープロセスは、現在の命令とバックアップされたメモリを使用してメモリグラウンディングを実行し、命令に関連するメモリを取得します。次に、LLMプロセスは、取得したメモリと命令に基づいて応答を生成し、結果の出力をTTS Todoキューにエンキューします。追加のTTSスレッド（例：F5-TTS [20]、MeloTTS [154]）は、TTS Todoキューのテキストをオーディオに変換し、フロントエンドに送信します。

4 Experiments

このセクションでは、オーディオとビデオの両方のベンチマークを含む、InternLM-XComposer2.5-OmniLive (IXC2.5-OL) のベンチマーク パフォーマンスを検証します。

4.1 Audio Benchmarks

私たちは、2つの有名な自動音声認識（ASR）ベンチマークである中国語（CN）用のWenetspeech [140]と英語（EN）用のLibriSpeech [87]でオーディオモデルを評価しました。WenetSpeechには、高品質で比較的きれいな中国語の音声を表すTest\_Netと、より難しい会話シナリオをキャプチャするTest\_Meetingの2つのテストセットが含まれています。LibriSpeechは、きれいで高品質の英語の音声を含むDev\_cleanとTest\_clean、およびノイズが多く複雑な発話を含むDev\_otherとTest\_otherの4つの分割で構成されています。

表 3 に示すように、IXC2.5-OL は、VITA や Mini-Omni などの最近のストリーミング オーディオ LLM と比較して優れたパフォーマンスを示し、特に軽量の 1.5B LLM のみで CN と EN の両方のベンチマークでより低いワード エラー レート (WER) を達成しています。

表3：ASRタスクの評価結果：「CN」は中国語の音声、「ENG」は英語の音声を表します。パフォーマンスはWER↓（単語誤り率）を使用して測定されます。

4.2 Video Benchmarks

表4、5、7、8では、MLVU [155]、Video-MME [37]、MMBench-Video [34]、MVBench [62]などの従来のビデオ理解ベンチマークで、IXC2.5-OLをクローズドソースAPIおよびオープンソースモデルと比較しています。さらに、リアルタイムビデオインタラクションのパフォーマンスをより適切に評価するために設計された、最近提案されたStreamingBench [67]でさまざまなモデルのパフォーマンスも評価します。この比較の結果を表6に示します。ビデオベンチマークの場合、基本モデルは評価中に各ビデオに対して64のサンプリングフレームを使用します。

表 4: MLVU ベンチマークの評価結果。IXC2.5-OL は、オープンソース モデルとクローズド ソース API の両方を上回る優れたパフォーマンスを示し、7B モデル スケールで SOTA を達成しました。

表 5: Video-MME ベンチマークの評価結果。IXC2.5-OL は、オープンソースの SOTA に近いパフォーマンスを示します。

表 6: StreamingBench によるリアルタイム視覚理解の評価結果。評価基準には、オブジェクト認識 (OP)、因果推論 (CR)、クリップ要約 (CS)、属性認識 (ATP)、イベント理解 (EU)、テキストリッチ理解 (TR)、予測推論 (PR)、空間理解 (SU)、アクション認識 (ACP)、およびカウント (CT) が含まれます。IXC2.5-OL は、すべてのオープンソース モデルの中では優れており、クローズド ソース API である Gemini 1.5 Pro にはわずかに及びません。

表 7: MMBench-Video の評価結果。タスクには、粗い知覚 (CP)、単一インスタンスの細粒度知覚 (FP-S)、インスタンス間の細粒度知覚 (FP-C)、幻覚 (HL)、論理推論 (LR)、属性推論 (AR)、関係推論 (RR)、常識推論 (CSR)、および時間推論 (TP) が含まれます。

表 8: MVBench での評価結果。タスクには、アクション シーケンス (AS)、アクション予測 (AP)、アクション反意語 (AA)、きめ細かいアクション (FA)、予期しないアクション (UA)、オブジェクトの存在 (OE)、オブジェクトの相互作用 (OI)、オブジェクトのシャッフル (OS)、移動方向 (MD)、アクションのローカリゼーション (AL)、シーン遷移 (ST)、アクション数 (AC)、移動数 (MC)、移動属性 (MA)、状態変更 (SC)、きめ細かいポーズ (FP)、文字順序 (CO)、自己中心的ナビゲーション (EN)、エピソード推論 (ER)、および反事実的推論 (CI) が含まれます。

MLVU

MLVU は、長時間ビデオ理解タスクにおけるマルチモーダル大規模言語モデル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。ビデオは 3 分から 2 時間までの範囲で、9つの異なる評価タスクが含まれています。ここでは、Topic Reasoning, Anomaly Recognition, Needle QA, Ego Reasoning, Plot QA, Action Order, Action Countの 7つのマルチチョイスタスクを評価します。詳細な比較は表 4 に示されています。IXC2.5-OL は、クローズドソース API と、パラメーターが 100 億未満のオープンソース モデルの中で最先端 (SOTA) のパフォーマンスを示し、Video-XL では以前の SOTA を 1.3 %、GPT-4o では 1.6 % 上回っています。

Video-MME

Video-MME は、高品質のビデオ ベンチマークです。ビデオは、11 秒から 1 時間までの短期、中期、長期のビデオを網羅する幅広いシナリオの一般化を保証するために、30 のサブフィールドを持つ 6つの主要な視覚領域から収集されます。表 5 に示すように、IXC2.5-OL はこのベンチマークで競争力のあるパフォーマンスを示し、以前の SOTA MiniCPM-V 2.6 に匹敵します。

StreamingBench

StreamingBench は、リアルタイム ビデオ評価用に設計されたストリーミング ビデオ ベンチマークです。18 のタスクで構成され、900 本のビデオと 4,500 の人間がキュレートした QA ペアが紹介されています。このコンテキストでは、リアルタイムでの視覚的理解の評価に焦点を当てています。表 6 は比較分析を示しており、IXC2.5-OL がすべてのオープン ソース モデルの中で優れており、以前の最先端モデルである LLaVA-OneVision よりも 2.67 % 向上し、クローズド ソース API である Gemini 1.5 Pro にはわずかに及ばないことを示しています。このパフォーマンスは、リアルタイム ビデオ インタラクションにおける IXC2.5-OL の優れた能力を裏付けています。

MMBench-Video

MMBench-Video は、600 本のビデオと 2000 の QA ペアで構成される自由形式の QA ビデオ ベンチマークです。各ビデオの長さは 30 秒から 6 分までです。回答が自由形式であることを考慮して、ベンチマークでは GPT-4 ベースの評価を利用して、精度、一貫性、人間の判断との整合性の点で品質を向上させます。結果を表 7 に示します。IXC2.5-OL は、知覚タスクで最先端のパフォーマンスを示し、全体的な評価でも同等のパフォーマンスを示します。

MVBench

MVBench は、時間的理解を重視したビデオ ベンチマークです。1つのフレームでは効果的に対処できない 20 の難しいビデオ タスクを網羅しています。表 8 に示すように、IXC2.5-OL は、7B のパラメータ サイズが小さいにもかかわらず、GPT-4 シリーズと 72B のオープン ソース モデル LLaVA-OneVision の両方を上回り、ビデオの時間的ダイナミクスを理解する優れた能力を示しています。

5 Conclusion

私たちは、長期記憶を備えたマルチモーダルテキスト、オーディオ、ビジュアル機能を向上させるリアルタイムストリーミングモデルである IXC2.5-OL を発表しました。IXC2.5-OL により、ユーザーは動的でインタラクティブな体験をすることができます。私たちのモデルのリアルタイム処理により、滑らかで応答性の高いインタラクションが可能になり、ユーザーはマルチモーダルデータの絶えず変化する環境にシームレスに関与できるようになり、より直感的で効率的なユーザーエクスペリエンスを提供できます。私たちの今後の取り組みでは、システム遅延を減らしてシームレスなユーザーエクスペリエンスを提供することに重点を置きます。