AdaVideoRAG: Omni-Contextual Adaptive Retrieval-Augmented Efficient Long Video Understanding

Abstract

マルチモーダル大規模言語モデル（MLLM）は、動画理解において優れた性能を示しているが、固定長のコンテキストと長期的な依存関係のモデリングにおける脆弱性のために、長い動画を処理する際の有効性が低下するという問題を抱えている。検索拡張生成（RAG）技術は、動的な知識拡張を通じてこれらの制限を軽減できるが、動画理解のための既存のRAG方式は、入力クエリの難易度に関係なく均一な構造を使用する固定検索パラダイムを採用している。これにより、単純なクエリ（フレームレベルの物体認識など）に対して冗長な計算オーバーヘッドと遅延（複雑なグラフトラバーサル操作など）が発生する一方で、マルチホップ推論のための検索粒度が不十分であることから、重大な情報損失が発生する可能性がある。このようなシングルステップの検索メカニズムは、モデルにおけるリソース効率と認知の深さのバランスを著しく制限する。これに対処するため、まず長時間動画理解のための新しいAdaVideoRAGフレームワークを提案する。このフレームワークは軽量な意図分類器を用いて、クエリの複雑さに基づいて、さまざまな動画理解タスクに、最も単純なものから最も高度なものまで、適切な検索スキームを動的かつ適応的に割り当てます。コンテキストモデリングのためにマルチモーダル信号から貴重な情報を抽出し、対応するデータベース（クリップキャプション、ASR、OCRからのテキストベース、ビジュアルベース、深い意味理解のためのグラフ）を構築するためのオムニナレッジインデックスモジュールを導入する。これにより、単純な検索からグラフ検索までの階層的な知識アクセス、統合、生成が可能になり、リソース消費と動画理解機能の最適なバランスが実現される。最後に、深い理解を評価するためのHiVUベンチマークを構築する。広範囲にわたる実験の結果、本フレームワークは長編動画のビデオQA（Video-QA）の全体的な効率と精度を向上させ、軽量API呼び出しを介して既存のMLLMとシームレスに統合できることが示されました。これにより、動画分析における適応型検索拡張の新たなパラダイムが確立されます。コードは近日中にオープンソース化されます。

1 Introduction

マルチモーダル大規模言語モデル（MLLM）は、その強力なマルチモーダル知覚および一般化能力により、多様なシナリオに対応するための普遍的な技術パラダイムとなり、ビデオ理解において強力な生成能力を発揮しています[31, 48, 29, 1]。しかし、特定の領域に適用する場合、知識の固定化（最新の知識を動的に更新できない）、制御不能な推論（幻覚のリスク）、弱い一般化（追加の微調整コストと時間コストが必要）などの課題によって制約され、マルチホップ質問やクロスモーダル関連付け要件（特に長時間のビデオシナリオ）の処理が困難になり、パフォーマンスの低下につながります[36, 32]。検索拡張生成 (RAG) は、事前トレーニング済みの知識に限定されることなく、外部知識ベースと生成モデルの協調推論を統合することにより、プライベート ドメイン データ シナリオに簡単に適応でき、大規模言語モデルの事実の精度とドメイン適応性を向上させるための中核的なパラダイムになっています。

現在の RAG 研究は主にテキスト モダリティ [25, 10, 19]、静止画像 [7]、表形式 [6] に焦点を当てており、マルチモーダル知識キャリアとしてのビデオ独自の価値が見過ごされています。長時間ビデオ理解の人気が高まるにつれて、ビデオ モダリティ入力をサポートする RAG モデルに対する新たな需要が生じています。長時間ビデオに関する既存の RAG 研究のほとんどは、ビデオから得られたマルチモーダル情報から知識ベースを構築および取得することによって、質問応答生成を強化しようとしています。たとえば、Luo ら [32] は、光学式文字認識 (OCR)、自動音声認識 (ASR)、およびオブジェクト検出からの視覚的に整列した補助テキスト特徴を組み込み、ビデオ知識ベースを作成して、長時間ビデオに対する質問応答を可能にしています。ただし、この方法は、Fig. 1 に示すように、データベース全体のグローバルな理解を必要とする意味作成クエリやマルチホップの質問をサポートしていません。最近のVideoRAG [36]は、グラフデータベースを構築することで長時間動画のコンテキスト情報の精度を大幅に向上させましたが、階層的なグラフデータベースを維持する必要があり、膨大な計算時間と時間を要するだけでなく、新しいシナリオへの移行時にも高いコストが発生します。動画理解のための実用的なRAGは、動画の種類やクエリの難易度に応じて適切な処理方法を柔軟に割り当て、精度を維持しながら効率を向上させる必要があると考えています。

現実世界のビデオ理解タスクにはさまざまな複雑さのコンテンツ理解ニーズが含まれることを考慮すると、異なる難易度の質問に対する問題解決戦略には明確な優先順位が設定されます。簡単な常識を必要とする短いビデオの QA では検索は必要なく、MLLM に問い合わせることで正しい回答を直接取得できますが、複雑な長いビデオの質問では、有効な情報をフィルタリングするために RAG による検索に依存します。複数段階の推論が必要な質問や複数の種類の知識に依存する質問など、より複雑な質問の場合、正しい回答を導き出すにはグラフベースの RAG が必要です。したがって、結果を取得してから返すという画一的なアプローチは最適ではありません。この問題を解決するために、本稿では、Fig. 1に示すように、AdaVideoRAG と呼ばれる適応型 RAG ベースのビデオ理解スキームを提案します。このスキームは、まずユーザークエリを難易度に分類し、次に難易度に基づいて最も適切な検索戦略を適応的に割り当てます。さらに、動画に含まれる視覚的特徴、クリップキャプション、音声認識、シーンテキストといった複合情報フローをさらに統合し、外部検索から取得した関連テキスト情報をデータ拡張に活用します。質問の難易度に応じて、クエリは異なるレベルのデータベース検索モード（ナイーブ検索とグラフ検索）にルーティングされます。これらのマルチモーダルな知識入力と検索戦略により、きめ細かな文脈表現能力をより効果的に提供することができ、最終的にはMLLMの長時間動画や複雑な質問応答タスクにおける処理能力の上限をさらに高めることができます。

提案されたAdaVideoRAGフレームワークの有効性を示すため、動画理解におけるフルスタック能力評価のための初のオープンベンチマークデータセットであるHiVUを正式にリリースしました。このデータセットは、短いクリップ（1分）から非常に長い動画（106分）まで、連続した持続時間スペクトルをカバーする120本の動画サンプルを画期的に統合しており、知識教育（講義、金融、法律、心理学、ドキュメンタリー）、情報（ニュース、インタビュー）、エンターテインメント（スポーツ、料理、メイク、フィットネス、テレビドラマ、アニメーション）という3つの主要テーマにわたる高頻度シーンカテゴリーを網羅しています。質問設計に関しては、革新的な3段階の難易度定量化システムを開発しました。1) 基本レベル1（L1）は、フレームレベルのコンテンツ認識に焦点を当てています（例：「動画の5秒目に現れるオブジェクトはどれですか？」）。2) 上級レベル2（L2）は、時相論理推論を必要とします（例：「話者はいつグラフニューラルネットワークの説明を始めますか？」）。 3) エキスパートレベル3（L3）は、クロスモーダル因果推論（例：「15分目のナレーションを削除すると、プロット展開にどのような影響が出るか？」）に挑戦します。ActivityNet [2]（単一動作認識）やMovieQA [37]（自由記述QA）といった従来のデータセットと比較して、このベンチマークは初めて異なるレベルでの認知複雑性評価を実現し、映像理解研究のための階層的な評価フレームワークを提供します。これは、長尺映像モデリング、複雑な推論タスク、そして現実世界のシナリオの一般化におけるモデルの体系的な最適化をサポートします。

要約すると、私たちの貢献は以下のとおりです。

1) クエリの複雑さに基づいて、さまざまなビデオ理解タスクに適した検索スキーム（最も単純なものから最も高度なものまで）を動的かつ適応的にルーティングする、新しいAdaVideoRAGフレームワークを提案します。これにより、リソース消費とビデオ理解能力の最適なバランスを実現します。

2) コンテキストモデリングのためのマルチモーダル信号から貴重な情報を抽出し、対応するデータベースを構築するためのオムニナレッジインデックスモジュールを導入します。軽量な意図分類モデルを用いて入力クエリの難易度を決定することで、ナイーブな検索からグラフ検索に至るまで、階層的な知識アクセス、統合、生成を可能にし、リソース消費とビデオ理解能力のバランスを保ちます。

3) ビデオ理解モデルの多階層推論能力を評価する階層型ビデオ理解ベンチマークHiVUを初めて公開します。広範な比較実験とアブレーション研究により、長時間ビデオの深層理解におけるAdaVideoRAGの利点が実証されています。

図1：様々なビデオ理解フレームワークの比較：i) MLLMは効率的だが、単純な問題しか扱えない。ii) VideoRAG [32]はナイーブな検索によって外部知識を統合するが、難しい推論問題には依然として苦労する。iii) 最近のVideoRAG [36]はグラフ検索を用いて複雑な問題に取り組むが、効率が低い。我々の新しいAdaVideoRAGフレームワークは、クエリの意図分類に基づいてクエリを異なる検索パスに適応的にルーティングすることで、有効性と効率性のより良いトレードオフを実現する。

2 Method

長時間動画理解のためのMLLM中心の適応型RAGフレームワーク「AdaVideoRAG」を紹介します。このフレームワークは、精度を確保しながら効率を大幅に向上させることができます。図2に示すように、本手法は4つの部分から構成されます。1) クエリ意図分類（2.1節）、2) オムニナレッジインデックス作成（2.2節）、3) 適応型検索パラダイム（2.3節）、4) 統合と生成（2.4節）。

2.1 Query Intent Classification

すべてのユーザーリクエストの複雑さは一様ではありません。単純なユーザーリクエストであれば、シンプルなソリューションで計算電力消費を抑え、ユーザーが感じるレイテンシの感覚を軽減できます。一方、複雑な質問に対しては、複雑なマルチモデル、マルチモーダル、マルチステップのクエリを用いることで、より高い精度を実現します。上記の目標を達成するために、軽量な意図分類モデルを用いて、入力側でクエリの難易度を分類することを提案します。具体的には、動画理解の難易度をきめ細かく評価するシステムを定義し、構築しました。

Level-1: Straightforward reasoning. (レベル 1: 率直な推論。)

質問には基本的に論理的な関係性はほとんどなく、質問に答えるために必要な知識は動画コンテンツ内に直接提供されています。例えば、「5秒目に登場する女性はどんな色の服を着ていますか？」といった質問です。このような質問に対して、既存のMLLMモデルは既に非常に成熟した解法を備えています。このような単純な質問に複雑な処理を適用すると、不要な計算オーバーヘッドが発生します。

Level-2: Simple reasoning. (レベル2: 単純な推論。)

これは、基本的な時空間的・因果関係に関するシングルステップ推論を伴い、モデルが局所的なイベント間の論理的な関連性を確立することを必要とします。例えば、「雨のシーンが始まる前に女性はなぜ泣いたのか？」という問いには、2段階の推論が必要です。1) 時間的な位置特定によって「雨」の開始時点を特定する。2) この時点以前の登場人物の行動（口論の音声など）とシーンの変化（天気予報の字幕など）を検索し、その動機を説明する因果連鎖を構築する。このようなタスクは、既存のMLLMS手法におけるクロスモーダルな時間的手がかりに関する統合上の欠陥を露呈し、検索粒度の不一致により重要な中間証拠が欠落する傾向があります。

Level-3: Hard reasoning. (レベル3: 難しい推論。)

最も難易度の高い動画理解では、長い文脈から様々な主題と関係性を抽出し、時間的および意味的次元にわたってエンティティと関係性をマッピングする知識グラフを構築し、それを強力なMLLM推論機能と組み合わせて判断を下す必要があります。例えば、「この映画はどのような人生教訓を伝えているのか？」といった質問です。このような質問では、モデルは動画内に含まれる深い意味的関係性をマイニングし、マルチホップ推論を実行して正しい答えを導き出す必要があります。

Intent classification model. (意図分類モデル。)

レベル 1 からレベル 3 までの基本定義と例が与えられたので、Chain-of-Thought (CoT) 推論を備えた大規模言語モデルを使用してクエリ Q を分類します。 これは、プラグアンドプレイ API として RAG (検索拡張生成) アーキテクチャに統合でき、微調整を必要とせずに適切なプロンプトを通じて意図分類結果を提供します。 分類結果に基づいて、無検索から単純なナイーブ検索、さらに複雑なグラフ検索に至るまで、段階的な知識検索戦略を自動的にトリガーできます。 意図分類結果 L の計算は、次のように定式化できます。 L = LLMintent(Q, promptintent)。 ここで、LLM は軽量の CoT モデルです。 この論文では、全体のプロセスと比較して時間の消費割合が非常に小さい (平均 ≤ 5%) Qwen2.5-7B [18, 45] を採用しています。

図2: AdaVideoRAGフレームワークの概要。1) クエリ意図分類（セクション2.1）。2) オムニナレッジインデックス作成（セクション2.2）。3) 適応型検索パラダイム（セクション2.3）。4) 統合と生成（セクション2.4）。

2.2 Omni-Knowledge Indexing for Long-Context Understanding

RAGを搭載したMLLMは、動画理解タスクを実行する際に、外部知識ベースを動的に活性化することでコンテキストモデリングを実現できます。これにより、長いコンテキストのウィンドウ長制限がある程度緩和され、動画全体の意味理解が向上します。この目的のために、我々はオムニナレッジインデックスモジュールを提案します。このモジュールは、コンテキストモデリングに必要な情報を複数のモーダル信号から抽出し、対応するデータベースを構築することで、RAGシステムが最も関連性の高い情報をより正確に取得し、高品質な生成を実行できるようにします。

2.2.1 Omni-Knowledge Text-Base Establishment

長時間の動画理解タスクでは、MLLMのコンテキストウィンドウサイズ制限のため、ハードウェア制約下で動画のフレームサンプリングとリサイズを行う必要があります。しかし、これにより動画に含まれる豊富な視覚情報や、未使用の音声およびテキストのマルチモーダル情報が失われることになります。そこで、外部の正規化モジュールを用いて動画からマルチモーダル情報を抽出し、独自のテキストベースを構築します。

Auxiliary text extraction and database construction. (補助テキスト抽出とデータベース構築。)

入力される長いビデオVは、一定の時間間隔 (この論文では 1 クリップあたり 30 秒) で、N個の連続する意味的に完全なクリップ V = ( C1 、C2 、…、CN ) = { Cn } に分割されます。各クリップ Cn に対して、均一なフレームサンプリングが実行され、キーフレームが抽出されます。この論文では、マルチモーダル表現プリミティブ Fn として5フレームを選択します。これは、フレーム数を増やしてもパフォーマンスが大幅に向上するのではなく、計算能力とモデルの複雑さが増すためです。具体的には、補助テキスト抽出には3つのカテゴリが含まれます。1) 量子化 MiniCPM-V [46] (VLM モデルとして使用) は、サンプリングされたフレームに対して、キャラクターの属性や時空間アクションなどの意味的要素を含むきめ細かいテキスト記述 TC を生成し、最終的にキャプション データベース DC を構築します。2) オーディオは、ビデオの中で最も直接的な情報伝達手段であり、そこで、音声抽出器としてFastWhisper[33]を使用し、各クリップの音声をテキスト形式TAに変換し、埋め込みモデルを介してベクトルとして保存してASR(Automatic Speech Recognition: 自動音声認識)データベースDAを生成します。3)各フレームの文字TOはEASYOCR[22]を介して抽出され、OCRデータベースDOが構築され、MLLMの認識能力の不足を補います。

Knowledge graph construction. (ナレッジグラフの構築。)

レベル3の複雑な推論クエリに対処するために、クリップキャプション（TC）、ASR（TA）、OCR（TO）に基づいて知識グラフを構築します。具体的には、BGE-M3 [5]はテキストチャンクからエンティティと関係性を抽出します。1) エンティティは、ビデオ内の最小限のドメイン固有の意味解釈単位を表し、<エンティティタイプ、エンティティ名、時空間属性>の3つで特徴付けられます。2) 関係性は、時空間関係、因果関係、機能関係など、エンティティ間のさまざまな意味的関連性を包含し、ビデオテキスト情報を体系的に構造化します。

2.2.2 Vision-Base Establishment

クリップのキャプション、ASR、OCRから抽出​​したテキスト情報だけに頼っていては、最適な知識インデックスを構築することが困難です。マルチモーダルデータの典型的なキャリアであるビデオには、オブジェクトの外観の変化、シーンの空間レイアウト、人間の顔の表情や動きなど、テキストでは正確に記述するのが難しい詳細が豊富な視覚的特徴が含まれています。これらの視覚情報は、複雑な知識推論および検索タスクに不可欠な役割を果たしています。そのため、私たちはキーフレームから特徴を抽出し、それらを最終的な特徴として連結するImageBind [16]画像エンコーダ（図2のEnc.）を導入します。このモデルは、画像、テキスト、オーディオなどの異種のモーダルデータを同じ高次元意味空間にマッピングできる高度なクロスモーダルアライメントアルゴリズムに基づいているためです。

2.3 Adaptive Retrieval Paradigm

意図分類（第2.1節）後、ユーザークエリ（Q）は難易度に応じて異なる検索パスにルーティングされ、有効性を確保することを前提に総合的な効率が向上します。

None retrieval with direct MLLM. (直接 MLLM による取得はありません。)

レベル1のシナリオでは、モデルはクエリ（Q）とビデオ全体（Cn）をMLLMに直接入力し、直接的な応答を取得します。このアプローチは、外部の知識ベースを導入することなく、MLLMが持つ固有の知識と推論機能を活用するため、単純な質問に対する全体的な効率が大幅に向上します。

Naive retrieval with simple reasoning. (単純な推論による単純な検索。)

レベル2の検索シナリオにおいて、本研究では、補助テキスト（クリップキャプション、ASR、OCR）と視覚モダリティ間の意味的整合を共同で最適化することにより、単純な論理的質問を扱う長時間動画の検索効率と精度を大幅に向上させるマルチモーダル協調グラウンディングフレームワークを提案する。具体的には、まず元のクエリを、異なるモダリティ特性に適応させたサブクエリに分離する。1) クリップキャプション検索の場合、クエリを平叙文(「主語で始まりピリオドで終わる」普通の文のこと)に書き換え、選択肢の干渉を除去し、シーンに適した説明コンテンツを追加する。2) ASR認識テキストの場合、クエリから口語表現を抽出し、コアアクションとイベントを保持し、断片的な音声セグメントに一致するように文脈修飾語句を追加する。3) 離散OCRテキストの場合、クエリから特定のエンティティ情報を抽出します。典型的な例として、入力クエリが「30番選手のパフォーマンスはどうでしたか？」の場合、書き換えられた出力は次のようになります。i)「クリップキャプション」：「30番選手のパフォーマンス。」、ii)「ASRテキスト」：「30番選手のパフォーマンスはどうですか。」、iii)「OCRテキスト」：「30番選手」。クエリ書き換えは、異なるセマンティクス間の分布のずれを効果的に緩和します。クロスモーダル類似度計算により、クエリに関連する候補コンテンツと、各テキストブロックに対応するビデオクリップを迅速に見つけることができます。

本研究では、視覚特徴データベースDVから意味的に最も関連性の高いビデオコンテンツをさらに検索してクエリします。具体的には、クリップキャプションの書き換え結果を視覚検索のためのセマンティックアンカーとして再利用します。事前学習済みのクロスモーダル セマンティック アライメント エンコーダー ImageBind [16]を使用して、ビデオをテキスト埋め込み空間 { Fn } にマッピングします。テキストと視覚埋め込み間のコサイン類似度を計算することで、類似度スコアが閾値（本稿では0.5に設定）を超える候補セグメントが除外されます。これらのセグメントは、最も高い信頼性を持つ上位K個の視覚的証拠を保持するようにランク付けされます。このアプローチは、統一されたセマンティック埋め込み空間を活用することで、ビジュアルとテキストのアライメントにおけるモダリティギャップを大幅に削減し、長いビデオでのローカルな詳細の損失の問題を効果的に軽減します。最後に、視覚的特徴とテキスト空間のアライメントを通じて取得されたビデオ { Cv } は、補助テキスト検索を通じて特定されたビデオチャンク { C c , a, o } とマージされ、レベル2での単純な推論のための検索証拠プールが構築されます。

Graph retrieval in hard reasoning. (ハード推論におけるグラフ検索。)

補助テキストと視覚的特徴検索から得られる情報のみに頼るだけでは、MLLMがより複雑な意味付けクエリシナリオに取り組むには不十分です。したがって、MLLM推論を容易にするためには、複数のイベントと時間的ノードをモデル化できる、より豊富で意味的に正確な補助情報が必要です。この課題に対処するため、単純なクエリに使用されるナイーブな検索手法に代えて、Light-RAG [19] に基づくより深い検索アプローチを採用し、難しいクエリを処理します。具体的には、リソース制約を考慮し、補助テキスト埋め込みを再利用してグラフを構築します。次に、書き換えられたクリップキャプションとエンティティ/リレーションシップの説明間の類似度スコアを計算し、最も関連性の高いエンティティとリレーションシップを返します。グラフマップ内では、取得されたエンティティとリレーションシップに関連するその他の情報を収集し、クエリ中心の思考マップに組み合わせることができます。この取得されたグラフマップは、MLLMがグローバルで多層的な情報を考慮するのを支援し、イベント内の時空間的および因果関係をより適切にモデル化するのに役立ちます。さらに、グラウンディング(特定の知識や情報源（例えばナレッジベースなど）に接続することで、言語モデルの回答や生成内容を裏付けるプロセスを指す)から得られた視覚的証拠を表現するために、統一された意味的埋め込み空間を用いることで、検索精度を向上させます。検索された動画 { Cv } にグラフ検索結果 { Cg } を重ね合わせることで、レベル3における難推論のための多階層検索証拠プールを構築します。

Filtering then sorting evidences. (証拠をフィルタリングして分類します。)

予備的な検索結果を得た後、検索結果に対して粗から細への情報精製を行う。まず、異なるデータベースから取得された重複するビデオ情報ブロックを除外する。次に、ビデオブロックのコンテンツ記述（クリップのキャプション、ASR、OCRテキストを含む）とクエリを、小規模LLM（論文ではQwen2.5-7B [45, 18]）に同時に入力し、きめ細かなフィルタリングを行うことで、無関係な検索結果を除外する。最後に、選択されたビデオクリップを元のビデオ時間に基づいて再ランク付けし、時間的な因果関係情報を保持する。

2.4 Multimodal Information Integration and Generation

MLLMにクエリ精度を高めるためのより包括的な情報を提供するために、クリップキャプション、ASR、OCRコンテキストから得られる補助テキスト情報（単純推論の場合は{T c , a , o }、難解推論の場合は{T g }と表記）と、ビジュアル・テキスト・グラウンディングからの視覚情報{C v }を取得します。取得したコンテキストと対応するビデオクリップ{C c , a , o }を統合した後、結合された入力はMLLMに送られ、推論と生成が行われ、最終出力Rが生成されます。

テキスト, 手紙

AI 生成コンテンツは誤りを含む可能性があります。

2.5 HiVU: Hierarchical Video Understanding Benchmark

既存のビデオ理解データセットは、再生時間が不十分 [14] か、コンテンツが魅力的でない [43, 53] ため、深い理解を必要とするクエリを生成できません。長いビデオでの堅牢な推論タスクをサポートし、さまざまな手法を評価するために、階層的ビデオ理解 (HiVU) ベンチマークを構築しました。この目的のために、知識教育 (講義、金融、法律、心理学、ドキュメンタリー)、情報 (ニュース、インタビュー)、エンターテイメント (スポーツ、料理、メイク、フィットネス、テレビドラマ、アニメーション) の3つのジャンルを選択しました。YouTube から知識コンテンツが豊富な 120 の長いビデオ データセット (合計 60 時間) を手動で収集しました。分布は図3に示されています。さらに、2.1 節で説明されているように、簡単なものから難しいものまで 3 層のクエリ推論を設計しました。この階層的なクエリ設計により、さまざまな難易度にわたってモデルの推論能力を包括的かつ詳細に評価できます。

Evaluation metrics.

HiVUデータセットの自由回答形式の質問応答タスクでは、RAG分野でモデルの能力を評価するために広く使用されている勝率メトリックシステム[10, 19]からインスピレーションを得ています。具体的には、大規模言語モデル（LLM）を判断基準として使用し、モデル出力を通じて2つのスキームの比較結果を定量化し、最終的に競合スコアをパーセンテージ形式で提示します。勝率比較では、クエリを5つの側面から包括的に検討します。1) 包括性：この側面は、モデルの応答がクエリを完全にカバーし、重要な情報を見逃したり、一方的な回答を提供したりすることを避けているかどうかに焦点を当てています。2) エンパワーメント：主に、モデルの応答がユーザーに実用的な価値とインスピレーションを提供できるかどうかを検査します。3) 信頼性：この側面は、モデルの出力コンテンツの信頼性と信憑性を重視します。4) 深さ：モデルが表面的な現象を超えて、クエリの背後にある本質的な問題を明らかにし、詳細な分析と議論を行うことができるかどうかを評価します。 5) 密度: 冗長、空虚、または冗長な表現を避けながら、モデルの応答の情報内容とコンパクトさに焦点を当てます。

図 3: さまざまな視点から見た HiVU の統計分布。

3 Experiments

3.1 Experimental Setup

提案されたAdaVideoRAG法と各モジュールの有効性を、主に新たに提案されたHiVUベンチマークセクション2.5で包括的に評価し、さらに徹底的に評価するために公開されているビデオ理解ベンチマークも紹介します。具体的には、1）HiVUは3つのドメインにまたがる10以上のサブジャンルを含み、合計60時間にわたる120の知識豊富な長時間ビデオデータセットで構成されています。2）Video-MME [14]は、多様なビデオとマルチモーダルデータを特徴とする、ビデオ分析におけるMLLMのフルスペクトルマルチモーダル評価ベンチマークです。900本のビデオ（11秒から1時間まで、短、中、長に分類）と、6つの主要な視覚ドメイン（知識、映画、スポーツなど）と30のサブドメインをカバーする2,700の多肢選択式の質問が含まれており、ビデオ分析におけるマルチモーダル大規模言語モデル（MLLM）の知覚、推論、要約機能の評価に重点を置いています。 3) MLVU [53]は、多様なジャンル・長時間の長時間動画の理解能力を評価するためのマルチタスクベンチマークである。3分から2時間以上（平均12分）の長時間動画を対象とし、映画、監視、ゲームなど多様な動画タイプにわたって9つのマルチタスク（例：単一詳細／複数詳細理解）を設定し、長時間動画理解能力を総合的に評価することを目指している。

3.2 Experimental Results

Improving open-sourced MLLMs with AdaVideoRAG on MLVU\_test [53] benchmark.

MLVUテストセットで調査されたすべてのマルチモーダル大規模言語モデルの全体的な評価結果を表1に示します。これらの結果には、ベースラインモデルであるVideo-LLaVA [29]と、最近リリースされた2つの評価の高いオープンソースモデルであるQwen2.5-VLシリーズ[1]とVideoLLaMA3 [47]が含まれています。評価結果は、提案したAdaVideoRAG戦略が各MLLMの質問応答精度を大幅に向上させることを明確に示しています。そして、それは特に2つの主要な種類のタスクで際立っています。1つ目は、動画に関するマルチホップ推論を必要とするトピック推論（TR）などのタスク、2つ目は全体論的推論を伴うアクションカウント（AC）などのタスクです。これは、AdaVideoRAGが基本的な質問応答能力を強化できるだけでなく、MLLMが複雑な推論やマルチ詳細処理タスクでブレークスルーを達成するのを効果的に支援できることを示しています。注目すべきは、MLVUデータセットでは比較的低いパフォーマンスを示したQwen2.5-VL-7Bモデルですが、AdaVideoRAGを導入することで精度が顕著に向上し、約40%向上し、Qwen2.5-VL-32Bのような大パラメータモデルの精度にまで達したことです。さらに、AdaVideoRAGを搭載したオープンソースモデルVideoLLaMAは、Qwen2.5-VL-32Bよりもパラメータ数が少ないにもかかわらず、長時間動画で優れたパフォーマンスを示し、GPT-4oに匹敵する性能を発揮しました。これらの実験結果は、MLLMの推論能力向上におけるAdaVideoRAGの普遍性と有効性を完全に証明しています。

表1：MLVU\_testにおけるAdaVideoRAGの設定の有無と教師ありベースラインの比較。フレーム：サンプリングフレームレートまたは制限される画像数。「2fps-768」は、ビデオが2fpsでサンプリングされ、上限が768フレームであることを示します。M-Avg：多肢選択タスクの平均パフォーマンス。

Comparison with state-of-the-art VideoRAG [32] on Video-MME [14] dataset.

表1の実験結果から、AdaVideoRAGはMLLMの推論性能を効果的に向上できることが十分に検証されたため、同じ数のパラメーターを持つ後続の制御実験の基本モデルとして、VideoLLaMA3とQwen2.5-VL-7Bモデルを選択しました。 表2では、Video-MMEデータセットでAdaVideoRAGとVideo-RAG [32]の水平比較テストを実施しています。 実験結果は、両方のRAG手法がベースMLLMのビデオ理解能力を大幅に向上できることを示しています。 ただし、長いビデオの処理を含むタスクでは、AdaVideoRAGがより明確な利点を示しています。 これは主に、AdaVideoRAGが長いビデオの検索中により複雑で合理的な知識マップを構築できるため、ビデオの正確な理解と効率的な推論が可能になるためです。

表2:Video-MME [14]データセットにおけるAdaVideoRAGとVideoRAG [32]の比較。

Impact of LLM arbiters.

さまざまな難易度の意味付けタスクにおける検索戦略のパフォーマンスを調査するために、提案された階層的ビデオ理解ベンチマーク（HiVU）に基づいて比較実験を実施し、最終的な回答の品質を評価するための評価審判としてLLMを使用します。特定のLLMの選択に関しては、表3に示すように、それぞれ異なるパラメーターと推論能力を持つモデルを表す、Deepseek-R1-7B [45, 18]とDeepseek-R1-32B [45, 18]、Qwen2.5-32B [45]とQwQ-32B [39]の2セットのコントロール実験を実施します。実験結果は、パラメーターが大きくChain-of-Thought（CoT）推論メカニズムを備えたモデルが、他のモデルの性能を評価する際に、より強力な識別能力を示すことを示しています。これらの調査結果に基づいて、評価結果の正確性と信頼性を確保するために、HiVU ベンチマーク評価の評価アービターとして DeepSeek-32B モデルを選択しました。

表 3: LLM アービタ構成 (パラメータ スケールと推論機能) の HiVU ベンチマーク評価への影響。

Comparison with state-of-the-art VideoRAG [32] on HiVU dataset.

HiVUデータベンチマークには、推論の難易度によって分類された3つのタスクがあります。簡単（L1）、シンプル（L2）、難しい（L3）。異なるレベルに対して、AdaVideoRAGは、検索なし、ナイーブ検索、階層的な強化メカニズムを形成するグラフ検索まで、さまざまな検索戦略を採用しています。そして、次の一連の実験は推論能力の向上を検証するように設計されています。表4に示すように、ハードレベルのビデオ理解タスクでは、AdaVideoRAGと統合されたマルチモーダル大規模言語モデルは、元のモデルと比較してより顕著な利点を示し、タスクの難易度が上がるにつれて、2つの間のギャップがより顕著になります。この結果は、複雑な推論シナリオにおけるAdaVideoRAGの有効性を確認するだけでなく、HiVUベンチマークの3レベルの難易度区分の合理性と科学性を間接的に検証し、モデルの推論能力を定量的に評価するための信頼できる基盤を提供します。

一方、表4に示すように、HiVUベンチマークにおいてVideoRAG [32]との水平比較を行った。予想通り、AdaVideoRAGはレベル1およびレベル2においてVideoRAG [32]と同等の性能を示した。しかし、グローバル推論とマルチホップ推論を必要とするレベル3においては、本手法の方がより顕著な優位性を示した。

表4：HiVUでのパフォーマンス。左：AdaVideoRAGの有無による結果の比較。右：VideoRAG [32] とAdaVideoRAGの有無による結果の比較。

Ablation Study

以下の分析では、提案手法の主要コンポーネントを正確に評価するために、3つのアブレーション研究を実施します。これらは以下の通りです。1) グラフなし：グラフマップ内のエンティティと関係性の検索をキャンセルします。2) 視覚検索なし：視覚からテキストへのグラウンディングにおける特徴検索を削除します。3) ナイーブテキスト検索なし：キャプション、OCR、ASRデータベースからの検索をキャンセルします（表5の比較を参照）。各モジュールの設計は効果的であり、モデルの理解能力を向上させることができることがわかります。

表5：グラフベースの知識検索、ビジョンベースの埋め込み検索、補助テキスト検索コンポーネントのアブレーション。

4 Conclusion

当社のAdaVideoRAGは、マルチホップ思考と深い推論を必要とする難解なビデオ理解タスクにおいて、卓越したパフォーマンスを発揮します。同時に、AdaVideoRAGはオムニ情報を効率的に統合し、画像、動画、テキストなどのマルチソースデータの価値を最大限に活用します。また、基本的な質問応答タスクと高次意味理解タスクを柔軟に切り替えることで、効率と精度のバランスを効果的にとり、マルチモーダル大規模言語モデルの汎化能力と応用の普遍性を大幅に向上させます。

Limitations, broader impact and social impact.

ローカル展開の計算要件により制限されるため、本稿では32B規模までのモデルのパフォーマンスのみを調査し、671Bのような大規模モデルは検討していません。ただし、671Bはより良い結果をもたらす可能性があります。また、ルーティングは3レベルのみに設定していますが、実際のアプリケーションシナリオではよりきめ細かい分類が必要になる可能性があります。社会的影響の観点から見ると、この技術は、情報の誤解を招くという新たなリスク（例えば、RAGで生成されたフェイクビデオニュースを悪用して世論を歪めるなど）を生み出す可能性があり、技術倫理、法的規範、業界標準を網羅した協調的なガバナンスシステムの確立が緊急に必要です。