AdaVideoRAG\_まとめ

[[2506.13589] AdaVideoRAG: Omni-Contextual Adaptive Retrieval-Augmented Efficient Long Video Understanding](https://arxiv.org/abs/2506.13589)

＜論文の概要＞

* クエリの難易度に基づいて、さまざまな動画理解タスクに適した検索スキームを動的かつ適応的に割り当てる、新しいAdaVideoRAGフレームワークを提案する。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する。
* 動画内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Omni-Knowledge Indexing moduleを導入する。軽量なintent classification model (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Knowledgeアクセス、統合、生成を可能にし、リソース消費と動画理解能力のバランスを保つ。
* 動画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を初めて公開する。  
  （今回はHiVUの内容については割愛。）

※コードは近日公開予定とのこと。

[GitHub - xzc-zju/AdaVideoRAG](https://github.com/xzc-zju/AdaVideoRAG)

＜手法の説明＞

AdaVideoRAGは長時間動画理解のためのMLLM中心の適応型RAGフレームワークである。

本手法は以下の4つの部分から構成される。

1) Query Intent Classification (クエリ意図分類)

2) Omni-Knowledge Indexingの作成

3) Adaptive Retrieval Paradigm (適応型検索パラダイム)

4) 統合と生成

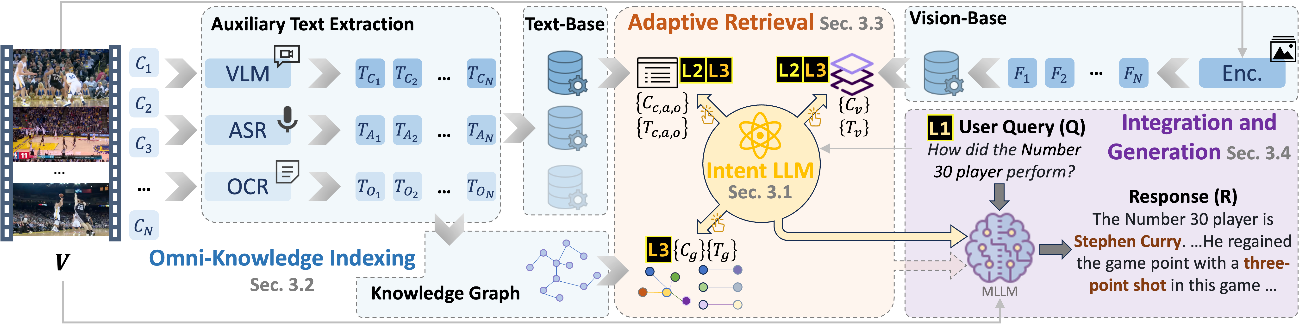


図2: AdaVideoRAGフレームワークの概要

1) Query Intent Classification (クエリ意図分類)

ユーザーからのクエリの難易度を分類することで、より適切な検索スキームを割り当てることができるようにする。

本手法では、Qwen2.5-7Bを用いて以下の3段階にクエリの難易度を分類している。

Level-1: Straightforward reasoning.

論理的な関係性はほとんどなく、回答す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る質問。

例）「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。

（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る。）

Level-2: Simple reasoning.

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する質問。

例）「雨のシーンが始まる前に女性はなぜ泣いたのか？」という質問。

　　これには以下の2段階の推論が必要となる。  
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。

Level-3: Hard reasoning.

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る質問。

例）「この映画はどのような人生教訓を伝えているのか？」とう質問。

　　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る。

2) Omni-Knowledge Indexingの作成

長いコンテキストのウィンドウ長制限をある程度緩和し、動画全体の意味理解を向上させるためにOmni-Knowledge Indexingを用いる。

これにより、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する。

テキストデータベースの構築：

入力される長い動画を一定の時間間隔 (本手法では1クリップあたり30秒) で、N個の連続するクリップ V = ( C1, C2, …, CN ) = { Cn } に分割する。

各クリップ Cn に対して、均一なフレームサンプリング(本手法では、5フレーム分)が実行され、キーフレーム{ Fn }が抽出される。

この各クリップやキーフレームに対して、以下の方法でテキストデータベースを構築する。

* VLM モデルとしてMiniCPM-Vの量子化モデルを使用し、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 を構築する。
* 音声抽出器としてFastWhisper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を構築する。
* 各フレームの文字TOはEASYOCR使用して抽出し、OCRデータベースDOを構築する。これによりMLLMの認識能力の不足を補う。

ナレッジグラフの構築：

Level-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）に基づいてナレッジグラフを構築する。

具体的には、BGE-M3 を用いてテキストチャンクからエンティティとリレーションシップを抽出する。

* エンティティは、動画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つで特徴付けられる。
* リレーションシップは、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、動画におけるテキスト情報を体系的に構造化する。

ビジョンデータベースの構築：

テキスト情報だけでは最適なKnowledge Indexを構築することは困難である。

テキストでは正確に記述するのが難しい視覚的特徴（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）も検索に用いることができるように、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できるImageBind の画像エンコーダ（図のEnc.）を導入する。

ImageBindによってキーフレームから抽出した特徴のデータベースDVを構築する。

※ImageBind の画像エンコーダによって、画像、テキスト、オーディオなどの異なる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。

3) Adaptive Retrieval Paradigm (適応型検索パラダイム)

意図分類の後、ユーザークエリの難易度に応じて以下の３つの異なる検索パスのいずれか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を向上させる。

None retrieval with direct MLLM：

クエリの難易度がLevel-1の場合、モデルはクエリ（Q）とビデオ全体（Cn）をMLLMに直接入力し、回答を取得する。

このアプローチは、上述したデータベースを用いることなく、MLLMが持つ固有の知識と推論機能を活用するため、全体的な効率が大幅に向上する。

Naive retrieval with simple reasoning：

クエリの難易度がLevel-2の場合、テキスト情報（クリップキャプション、ASR、OCR）とビジョン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、単検索効率と精度を大幅に向上させるmultimodal collaborative grounding frameworkを用いる。

具体的には、まず元のクエリを、以下のように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。

* クリップキャプション検索の場合、クエリを平叙文に書き換え、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。
* 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。
* 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する。

例）

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、書き換えた結果は次のようになる。

「クリップキャプション」：「30番選手のパフォーマンス。」

「ASRテキスト」：「30番選手のパフォーマンスはどうですか。」

「OCRテキスト」：「30番選手」

次に、ImageBindを用いて、書き換えたクリップキャプションとビジョンデータベースDV中のデータで類似度を計算する。

類似度が閾値(本手法では0.5)を超える候補セグメントのうち信頼度の高い上位K個のデータを視覚的エビデンスとして保持する。

最後に、上記を通じて取得された動画{ Cv } は、対応する動画チャンク{ Cc, a, o }とマージされ、Level-2における回答のための検索エビデンスプールが構築される。

Graph retrieval in hard reasoning：

クエリの難易度がLevel-3の場合、Light-RAG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する。

具体的には、まずテキスト情報をもとにナレッジグラフを構築する。

※ナレッジグラフにより、テキストチャンクからエンティティとリレーションシップを抽出する。

次に、書き換え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(=グラフ検索結果)を返す。

最後に、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、Level-3における回答のための多階層検索エビデンスプールを構築する。

エビデンスのフィルタリングと分類：

“Naive retrieval with simple reasoning”や“Graph retrieval in hard reasoning”で得た検索結果の情報精製を行う。

まず、異なるデータベースから取得された重複する動画情報ブロックを除外する。

次に、動画ブロックのコンテンツ記述（クリップキャプション、ASR、OCRテキストを含む）とクエリを、Qwen2.5-7Bに同時に入力し、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、選択された動画クリップを元の動画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。

4) 統合と生成

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テキスト情報（simple reasoningの場合は{Tc, a, o}、hard reasoningの場合は{Tg }と表記）と、視覚情報{Cv}を取得する。取得したコンテキストと対応するビデオクリップ{Cc, a, o}を統合した後、結合された入力はMLLMに送られ、最終出力®が生成される。

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。

＜AdaVideoRAGの効果＞

MLVUによる評価：

MLVU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。

MLVUの詳細は以下参照。

[**[2406.04264] MLVU: Benchmarking Multi-task Long Video Understanding**](https://arxiv.org/abs/2406.04264)

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。

上の表1は、MLVUにおけるAdaVideoRAGの有無による制度の比較結果である。

表1よりAdaVideoRAGを用いることで、各MLLM(Video-LLaVA, Qwen2.5-VL, VideoLLaMA3)の質問応答精度を大幅に向上させることが示されている。

特にQwen2.5-VL 7Bモデルでは、AdaVideoRAGを導入することで精度が約40%も向上する結果となった。

※Paramsの欄で72Bとなっているのは32Bの誤記。

Video-MMEによる評価：

Video-MMEは、多様な動画とマルチモーダルデータを特徴とする、動画分析におけるMLLM(Multimodal Large Language Model)の包括的なマルチモーダル評価ベンチマークである。900本の動画（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、動画分析におけるMLLMの知覚、推論、要約機能の評価に重点を置いている。

Video-MMEの詳細は以下参照。

[[2405.21075] Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis](https://arxiv.org/abs/2405.21075)

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。

上の表2は、Qwen2.5-VL 7BとVideoLLaMA3に対して従来の手法であるVideo-RAGと今回の提案手法であるAdaVideoRAGをそれぞれ用いた際の比較結果である。

表２より、長いビデオの処理を含むタスク(Medium, Long)では、AdaVideoRAGがより明確な利点を示している。 これは主に、AdaVideoRAGが長い動画の検索中に、より複雑で合理的な知識マップを構築でき、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ある。

＜まとめ＞

AdaVideoRAGは、マルチホップ思考と深い推論を必要とする難解な動画理解タスクにおいて、卓越したパフォーマンスを発揮する。

同時に、AdaVideoRAGは全ての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する。

また、基本的な質問応答タスクと高度な意味理解タスクで検索パスを柔軟に切り替えることで、効率性と精度のバランスを効果的に取り、マルチモーダル大規模言語モデルの汎化能力と様々な応用分野に適用できる普遍性を大幅に向上させる。