Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video

arxiv : [[2510.03458] Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video](https://arxiv.org/abs/2510.03458)

hugging face: [nvidia/omni-embed-nemotron-3b · Hugging Face](https://huggingface.co/nvidia/omni-embed-nemotron-3b)

(NVIDIAが出したやつ)

**Abstract.**

我々は、現実世界の情報ニーズの複雑化に対応するために開発された、統合マルチモーダル検索埋め込みモデルであるOmni-Embed-Nemotronを紹介します。検索拡張生成（RAG）は外部知識を取り込むことで言語モデルを大きく進歩させましたが、既存のテキストベースの検索ツールはクリーンで構造化された入力に依存しており、PDF、スライド、動画などの現実世界の文書に見られる視覚的および意味的に豊かなコンテンツの取り扱いに苦労しています。ColPaliなどの最近の研究では、画像ベースの表現を用いて文書レイアウトを維持することで検索品質を向上できることが示されています。これに基づき、Qwen2.5-Omniなどの最近のマルチモーダルモデルの機能に着想を得て、我々は検索をテキストと画像だけでなく、音声と動画のモダリティにも拡張します。Omni-Embed-Nemotronは、単一のモデルでクロスモーダル検索（例：テキスト→動画）とジョイントモーダル検索（例：テキスト→動画+音声）の両方を可能にします。 Omni-Embed-Nemotron のアーキテクチャ、トレーニング設定、評価結果について説明し、テキスト、画像、ビデオ検索におけるその有効性を示します。

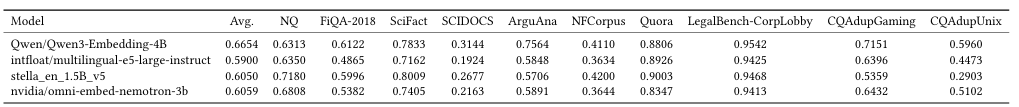
1. Introduction  
   検索拡張生成（RAG）は、外部知識検索を統合することで言語モデルを強化し、事実の精度と文脈認識を向上させながら、モデルが大規模なコーパスを推論できるようにします。近年のテキスト検索の進歩により、NV-Embed（Lee et al., 2024）、NV-Retriever（Moreira et al., 2024）、Qwen3-Embedding（Zhang et al., 2025a）、e5-mistral（Wang et al., 2023）といった高性能モデルが開発され、Massive Text Embedding Benchmark（MTEB）（Muennighoff et al., 2022; Chung et al., 2025）などのベンチマークで優れた結果を達成しています。MTEBは、幅広い検索、分類、クラスタリングタスクにおけるテキスト埋め込みモデルの性能を評価するために設計された、広く使用されている評価ベンチマークです。しかし、これらのモデルは、整然とフォーマットされたテキスト入力を必要としますが、現実世界の文書（PDFやスライドなど）には複雑なビジュアルレイアウトが含まれていることがよくあります。この問題に対処するため、ColPali（Faysse et al., 2024）は、文書を画像に変換し、テキストとビジュアルの両方を保存して検索できるようにすることを提案しています。 Qwen-VL（Wang et al., 2024a）、LLaMA-3.1-Nemotron-Nano-VL（Bercovich et al., 2025）、Eagle2（Li et al., 2025; Chen et al., 2025）などの視覚言語モデル（VLM）は、CLIP（Radford et al., 2021）、SigLIP（Tschannen et al., 2025）、C-RADIO（Ranzinger et al., 2024）などのエンコーダーを使用して、結合マルチモーダル表現を学習します。これらの開発により、文書をネイティブの視覚形式で理解し、テキストと画像の共同表現を学習できるマルチモーダル検索システムが促進され、Colpali (Faysse et al., 2024)、Llama Nemoretriever Colembed (Xu et al., 2025)、Colnomic Embed Multimodal (Team, 2025) など、多くの視覚検索モデルが存在します。  
   現実世界のコンテンツは、テキストや画像にとどまらず、音声や動画といったリッチなモダリティを包含しています（図1参照）。そのため、従来のユニモーダル検索手法では、多様なクエリ形式や異種データソースへの対応がますます困難になっています。この限界に対処するため、私たちは、テキスト、画像、音声、動画など、様々なモダリティを横断して関連文書やメディアを検索できる統合検索モデル、Omni-Embed-Nemotronを提案します。このアプローチは、クロスモーダル検索とジョイントモーダル検索の両方をサポートし、動画のみ、あるいは動画＋音声＋画像といった組み合わせのクエリを可能にします。  
     
   図 1. テキスト、画像、ビデオの3つの入力様式を備えたマルチモーダル検索アーキテクチャ。
2. Model  
   私たちの検索システムは、図2に示すように、クエリとコーパス項目がそれぞれ独立して共有の埋め込み空間にエンコードされるバイエンコーダモデルとして構築されています。推論時には、ドット積やコサイン類似度などの類似度指標を用いて関連性を効率的に計算します。このアーキテクチャは高いスケーラビリティを提供し、大規模コーパスからの効率的な検索をサポートします。  
   当モデルの重要な特徴は、クロスモーダルおよびジョイントモーダル検索のサポートです。クエリとコーパスエントリは、任意の単一モーダル（例：テキスト、画像、音声、動画）またはそれらの任意の組み合わせ（例：テキスト+音声→画像+動画）から取得できます。これにより、音声による説明とテキストキーワードを用いて動画データベースを検索したり、マルチモーダルプロンプトを用いて関連する音声・映像コンテンツを検索したりするなど、豊富で柔軟な検索ユースケースが可能になります。  
   私たちのモデルは、視覚、言語、音声、ビデオを横断した入力を処理するように設計されたマルチモーダル基盤モデルであるQwen-Omni / Qwen2.5-Omni-3B (Jin Xu, 2025) を基盤としています。具体的には、クロスモーダル理解を担うQwen-OmniアーキテクチャのThinkerバックボーンを活用しています。本モデルは検索タスクに重点を置き、音声出力を生成しないため、Talkerコンポーネントは削除しました。この設計選択により、強力なマルチモーダル表現機能を維持しながら、アーキテクチャを簡素化しています。  
   音声トークンと動画トークンをインターリーブし、Time-aligned Multimodal RoPE (TMRoPE) を適用して単一のトークンシーケンス内で同期タイムスタンプを強制する Qwen-Omni モデル (Jin Xu, 2025) (図 3 (a)) とは対照的に、私たちの検索エンコーダーは2つのストリームを分離しています (図 3 (b))。音声と動画を独立してエンコードし、各モダリティのネイティブな時間構造を、クロスモーダルトークンインターリーブなしで保持します。私たちの実験によると、この非インターリーブ設計により、表2および3に示すように検索パフォーマンスが向上しました。生成タスクとは異なり、検索タスクでは音声情報と動画情報を分割せずに完全な状態に保つことで、モデルが各モダリティの完全なコンテキストをキャプチャできるようになり、関連するコンテンツの一致と検索が容易になると考えています。  
     
   図2. クエリとコーパスの両方が同じLLMでエンコードされ、その後にプーリング層が続く、バイエンコーダ検索システム。得られた表現を比較して類似度スコアを計算する。  
   図3. 2つの融合戦略の比較：(a) Qwen Omniモデルのインターリーブ融合戦略。音声トークンと動画トークンはTMRoPEを使用して順次整理され、同期されます。(b) 弊社の検索モデルの個別ストリーム融合戦略。音声と動画はトークンインターリーブなし（＝交互に処理しない）で独立してエンコードされます。
3. Training
   1. Contrastive learning  
      対照学習を活用し、クエリと正のコーパス間の埋め込み類似度を最大化し、クエリと負のコーパス間の埋め込み類似度を最小化します。InfoNCEの対照損失（Chen et al., 2020）を採用し、モデルが共有埋め込み空間内の正と負のペアを区別できるように学習します。  
      ～中略～  
      対照学習の有効性を高めるため、ハードネガティブマイニングにNV-Retriever (Moreira et al., 2024) の上位kサンプルとポジティブ閾値の割合戦略を採用しました。閾値は0.95に設定し、クエリとの類似度がクエリとポジティブの類似度スコアの95%未満である、最も関連性の高いK個のネガティブサンプルを選択します。実験ではK=2としました。これにより、モデルは難しいネガティブサンプルから学習するようになり、高い類似度スコアを持つ潜在的な偽陰性サンプルを排除できます。
   2. Training  
      学習中は、音声と映像の両方のエンコーダをフリーズし、LoRA (Hu et al., 2022) のチューニングを言語モデル（LLM）のみに適用します。このアプローチにより、計算オーバーヘッドを最小限に抑えながらモデルのクロスモーダル機能を効率的に適応させ、マルチモーダル入力を処理する能力を効果的に一般化できます。適応能力と学習効率のバランスをとるため、ハイパーパラメータ LoRA ランクと LoRA スケーリング係数をそれぞれ r=16 と α=32 に設定しました。  
      モデルアーキテクチャを改良し、因果的注意(causal attention)を双方向注意(bidirectional attention)に置き換え、大規模なテキスト-テキストおよびテキスト-画像ペアで学習します。これにより、クエリや文書に対する強固な基礎検索モデルを構築し、学習した表現を視覚入力と整合させることで、異なるモダリティ間での整合が可能になります。
   3. Datasets  
      Llama Nemoretriever Colembed (Xu et al., 2025) の方法論に従い、テキストと画像のペアと、ColPali トレーニング セット (Faysse et al., 2024)、Wiki-SS-NQ (Xueguang Ma, 2024; Ma et al., 2025)、VDR (Cimolai and Markewich, 2025)、Docmatix (Laurençon et al., 2024)、Natural Questions (NQ) (Kwiatkowski et al., 2019)、Stack Exchange (Stack Exchange, Inc., 2023)、SQuAD (Rajpurkar et al., 2016)、HotpotQA (Yang et al., 2018) などの高品質で小規模なテキスト対テキスト検索データセットを組み合わせてモデルを微調整します。これらの多様で困難な例は、特に視覚検索タスクにおいて、効果的なマルチモーダルアライメントを促進します。  
      学習中に音声や動画データを使用しなくても、本モデルはこれらのモダリティで優れたパフォーマンスを発揮します。クロスモーダル機能をさらに拡張するために、FineVideoデータセット（Farré et al., 2024）のテキストと動画、およびテキストと音声のペアを組み込んだ追加の学習段階を実験しました。FineVideo学習セットを追加することで、FineVideo評価ベンチマークでのパフォーマンスが向上しました。実験結果を表4に示します。ただし、他のモダリティ間での検索性能の向上や、未知の動画データセットへの一般化の向上は見られませんでした。全体として、テキスト、画像、動画、音声のためのモデルの統合された埋め込み空間は、この追加段階によってほとんど影響を受けませんでした。
4. Results  
   私たちは、ビデオ検索、画像検索、テキストのみの検索という 3つの方法でモデルを評価します。
   1. Video Retrieval  
      **FineVideo Benchmark.**FineVideo (Farré et al., 2024) には、平均再生時間4.7分の43,751本の動画が含まれており、122の多様なカテゴリにまたがっています。現実的な検索ベンチマークを構築するため、content\_fine\_categoryのクラス分布がテストセットでも維持されるようにしながら、10,000本の動画を検索コーパスとしてランダムにサンプリングしました。これらの動画には46,553件の関連する質問が含まれており、検索タスクのテキストクエリとして使用しました。各クエリは、コーパスから抽出した正解動画とペアになっています。  
      **LPM Benchmark.**また、教育コンテンツに焦点を当てたLPMデータセット（Lee et al., 2022）から2つ目のベンチマークを構築しました。LPM​​の動画は、ライブスライドデッキと講演者の頭部で構成される録画された講義で、スライドには手動で注釈が付けられ、セグメント化されたタイムスタンプが付与されています。データセットは、音声コンテンツのGoogle ASR（automatic speech recognition）書き起こしを提供しています。これらの書き起こしを対応するスライド画像と対応させ、これらのテキストと画像のペア入力を用いてLlama-3.2-90B-Vision-Instruct（AI, 2024）に指示し、評価用の合成質問1,000件を生成しました。  
      **Baselines.**テキスト画像検索におけるViDoReやテキストテキスト検索におけるMTEBのような標準的なテキスト・ビデオ検索ベンチマークは存在しないため、FineVideoで提供されている音声テキスト化トランスクリプトとLPMから生成したトランスクリプトを標準的なテキスト検索モデルの入力コーパスとして用い、独自の評価環境を構築した。ベースラインとして、MTEBリーダーボード(Muennighoff et al., 2022)において4Bパラメータ下でトップランクのモデルであるQwen/Qwen3-Embedding-4B (Zhang et al., 2025a)と、他の2つの高性能検索モデル、intfloat/multilingual-e5-large-instruct (Wang et al., 2024b)およびstella\_en\_1.5B\_v5 (Zhang et al., 2025b)を選択した。この設定により、当社のマルチモーダル検索モデルと、文字起こしされたビデオコンテンツのみで動作する最先端のテキストのみの検索システムを直接比較できるようになります。  
      **Results.**表1は、両ベンチマークにおけるモデルとベースラインの検索性能をまとめたものです。両データセットにおいて、モデルはFineVideoで最も強力で、NDCG@10で最高値の0.5662を達成し、テキストのみのベースライン（Qwen3-Embedding-4B）の最高値をそれぞれ2.6ポイント上回りました（表1）。これは、ASRトランスクリプトにノイズが多い場合や不完全な場合、マルチモーダルな手がかり（映像フレームと音声）を活用することが有益であることを示しています。 LPMの講義形式の動画では、クエリがトランスクリプトから合成され、テキスト品質も高いため、Qwen3-Embedding-4Bは0.8634/0.8518でリードしています。一方、本モデルは0.8465/0.8355で競争力を維持しており、それぞれ1.7ポイントと1.6ポイント差で、stella\_en\_1.5B\_v5よりわずかに遅れをとりましたが、intfloat/multilingual-e5-large-instructよりは優位です。全体的に、マルチモーダル検索はオープンドメイン動画において明確な優位性を示していますが、テキスト中心のシナリオではその優位性は縮小しています。平均値を見ると、本モデルはNDCG@10で0.7064、NDCG@5で0.6921と、他のモデルをわずかに上回っています。  
      グラフィカル ユーザー インターフェイス, アプリケーション, 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。  
      表1. LPMとFineVideoデータセットにおけるNDCG@10およびNDCG@5メトリクスを用いたモデル性能比較。平均列はLPMとFineVideoの平均を示す。最良の平均値は太字で強調表示されている。1～3行目はトランスクリプトで評価したテキストのみのベースラインであり、4行目はビデオで評価したオムニモデルである。  
      (NDCG@k：上位k番目までの実際のランキングに対して予測結果の正しさを測る。)  
        
      Modality Breakdown.  
      表2と表3は、異なる入力様式でのパフォーマンスを示しています。LPMでは、トランスクリプト+OCRをモデルで使用した場合(0.8636)は、最良のテキストベースライン(0.8634)とほぼ一致しましたが、音声のみ(0.8238)とビデオのみ(0.7365)はそれに遅れをとりました。音声とビデオを融合すると0.8373に向上し、2つのストリームを後で融合して個別にスコアリングすると0.8465に達し、テキストとの差が縮まりました。FineVideoでは、テキストのみのバリアント(0.6082)がすべてのテキストベースラインを上回りました。非テキストチャネルでは、音声のみ(0.5407)がビデオのみ(0.4488)を明らかに上回っています。音声とビデオの単純な初期融合(0.4700)は、音声のみを下回りますが、後期の融合(「音声とビデオを個別に」)では0.5662まで回復し、融合戦略の重要性を示しています。全体的に、講義形式のコンテンツではテキストが中心となりますが、オープンドメインのビデオでは、オーディオはフレームのみよりも多くの信号を伝達し、テキストが不完全または利用できない場合は、後期融合によって非テキストの堅牢性が最善になります。  
        
      表2. LPMデータセット（NDCG@10）におけるマルチモーダル検索性能。現在のベースラインモデルはテキスト入力のみをサポートしており、マルチモーダル機能（音声のみ、ビデオのみ、音声+ビデオ）は利用できません。  
      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。  
        
      表3. FineVideoデータセット（NDCG@10）におけるマルチモーダル検索性能。現在のベースラインモデルはテキスト入力のみをサポートしており、マルチモーダル機能（音声のみ、ビデオのみ、音声+ビデオ）は利用できません。  
      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。  
        
      **Domain Specific Fine-Tuning**このモデルはビデオ データセットで明示的にトレーニングされていないにもかかわらず、結果はビデオ検索タスクで優れたパフォーマンスを示し、モダリティ全体にわたるモデルの一般化能力を強調しています。  
      ドメイン固有適応の利点をさらに評価するため、FineVideoのトレーニングセットでモデルを微調整し、対応するテストセットで評価しました。データセットはビデオレベルでランダムに分割され、トレーニングと評価の間でビデオが重複しないようにしています。この微調整により、表4に示すようにパフォーマンスが向上し、ndcg@5は0.55から0.61に向上しました。この結果は、ドメイン内データに対するタスク固有の微調整が、ビデオ領域における検索精度を効果的に向上できることを裏付けています。  
        
      表4. FineVideoテストセットにおけるビデオ検索性能。本モデルはビデオ領域への強力なゼロショット汎化を示しており、領域固有の微調整によりさらなる性能向上が実現されています。  
      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。
   2. Image Retrieval  
      我々は、視覚理解と検索を評価するために設計された大規模マルチモーダルベンチマークであるViDoReベンチマーク（Faysse et al., 2024）を用いて、モデルの画像検索性能を評価しました。ViDoReは、多様な画像コレクションと自然言語クエリの組み合わせで構成されており、テキストと画像表現間のクロスモーダルアライメントをテストするのに適しています。  
      強力なベースラインを確立するため、4Bパラメータスケールにおいて、本モデルを複数の高性能画像検索モデルと比較しました。特に、このカテゴリにおいて4Bパラメータスケールにおけるすべてのモデルの中でトップランクのモデルであるnvidia/llama-nemoretriever-colembed-1b-v1 (Xu et al., 2025) を含めました。このモデルは、軽量マルチモーダル検索における現在の最先端技術を代表するものです。  
      この評価の結果は表5に示されています。主要なベースライン（llama-nemoretriever-colembed-3B）の平均スコアは91.0であるのに対し、オムニモデルの平均スコアは85.7です。このモデルはArxivQAでは競争力があり、幅広い分野（AI、政府報告書、エネルギー）で強みを発揮していますが、Shift Project、TAT-DQA、DocVQAでは劣っています。全体として、この結果は、このモデルが視覚文書の検索において優れた能力を持っていることを示しています。  
        
      表5. ViDoRe V1におけるベースラインモデルと提案モデルの評価（7月30日時点）。結果はnDCG@5メトリクスを用いて示されている。  
      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。
   3. Text Retrieval  
      本モデルのテキスト検索能力を評価するため、Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022) からいくつかのベンチマークタスクを選択しました。これらのタスクは、オープンドメインの質問応答、科学文献検索、議論型テキストマッチングなど、様々な領域をカバーしており、様々な検索シナリオにおけるモデルのパフォーマンスを包括的に把握できます。  
      表6に示すように、このモデルはテキストのみに特化した埋め込みと比べると最高の結果には至りませんが、データセット全体にわたって堅実かつ競争力のあるパフォーマンスを発揮します。最も重要なのは、このモデルがモダリティ全体にわたる汎用的な検索システムとして機能し、テキスト検索において良好なパフォーマンスを維持しながら、統一された埋め込み空間内で画像、動画、音声の検索もサポートできることを示していることです。  
        
      表6. テキスト検索ベンチマークにおける埋め込みモデルの評価。結果はnDCG@10を用いて報告されている。  
      
5. Impact of Modality and Preprocessing on Sequence Length and Retrieval Performance  
   私たちのアブレーション研究では、プロセッサの引数がシーケンスの長さと結果として得られる検索パフォーマンスの両方に強く影響することが示されています。計算コストと精度のバランスをとるには、慎重な構成が必要です。私たちの経験に基づくと、表7 に示す構成設定は、さまざまなモダリティにわたって安定したパフォーマンスを提供します。説明のために、再生時間が1050.67秒でファイルサイズが28 MBのMP4ビデオを例に挙げます。表7のプロセッサ設定を使用して、前処理後のシーケンスの長さを計算します。これは、プロセッサ設計の選択がトークン化された入力の長さを直接決定し、したがってマルチモーダル検索の計算要件を決定する方法を示しています。さらに、モダリティの選択は検索パフォーマンスに影響します。表2 と3に示すように、ビデオとオーディオを組み合わせると、どちらかのモダリティを単独で使用するよりも良い結果が得られます。あるいは、ASRを適用してビデオからテキストを抽出することは、Finevideo のようなコンテンツが豊富なオーディオの場合は効率的ですが、追加のASR 処理が必要になります。全体として、これらの調査結果は、効率的でスケーラブルな検索システムを構築するためのモダリティを考慮した前処理の重要性を強調しています。  
     
   表7. アブレーション研究で使用したプロセッサ引数設定。これらの値は、異なるモダリティ間でシーケンスの長さとパフォーマンスのバランスをとっています。  
   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。  
     
   表8. 異なるモダリティ設定におけるシーケンス長。この結果は、モダリティの違いによってシーケンス長が異なり、それが検索時の計算要件に直接影響を与えることを示しています。  
   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。
6. Conclusion  
   Qwen-Omniバックボーン上に構築されたマルチモーダル検索モデルOmni-Embed-Nemotron-3Bを発表しました。このモデルは、テキスト、画像、音声、動画を統合された埋め込み空間で処理できます。アーキテクチャの簡素化、モダリティ固有のエンコーディング、そして対照学習により、このモデルは複数のベンチマークで優れた性能を発揮します。必ずしも特化した単​​一モダリティシステムを上回る性能ではありませんが、テキスト、画像、動画の検索において堅牢で競争力のある結果を示しています。これらの結果は、多様なモダリティに対する統合検索モデルの有効性を強調するものであり、ドメイン固有の微調整とより洗練されたマルチモーダルアライメント戦略による今後の改善を示唆しています。