Novelty Assessment Report

Paper: MetaEmbed: Scaling Multimodal Retrieval at Test-Time with Flexible Late Interaction
PDF URL: https://openreview.net/pdf?id=yKDqg9HwZX
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Universal multimodal embedding models have achieved great success in capturing semantic relevance between queries and candidates. However, current methods either condense queries and candidates into a single vector, potentially limiting the expressiveness for fine-grained information, or produce too many vectors that are prohibitively expensive for multi-vector retrieval. In this work, we introduce MetaEmbed, a new framework for multimodal retrieval that rethinks how multimodal embeddings are constructed and interacted with at scale. During training, a fixed number of learnable Meta Tokens are appended to the input sequence. At test-time, their last-layer contextualized representations serve as compact yet expressive multi-vector embeddings. Through the proposed Matryoshka Multi-Vector Retrieval training, MetaEmbed learns to organize information by granularity across multiple vectors. As a result, we enable test-time scaling in multimodal retrieval where users can balance retrieval quality against efficiency demands by selecting the number of tokens used for indexing and retrieval interactions. Extensive evaluations on the Massive Multimodal Embedding Benchmark (MMEB) and the Visual Document Retrieval Benchmark (ViDoRe) confirm that MetaEmbed achieves state-of-the-art retrieval performance while scaling robustly to models with 32B parameters.

Core Task Landscape

This paper addresses: **Multimodal Retrieval with Flexible Late Interaction Embeddings**

A total of **28 papers** were analyzed and organized into a taxonomy with **10 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Late Interaction Architecture Design and Optimization**
- **Domain-Specific Multimodal Retrieval Applications**
- **Multimodal Fusion Strategies and Cross-Modal Interaction**
- **General Multimodal Interface and Interaction Frameworks**

Complete Taxonomy Tree

- Multimodal Retrieval with Flexible Late Interaction Embeddings Survey Taxonomy
- Late Interaction Architecture Design and Optimization
  - Token-Level Interaction and Matching Mechanisms ★ (5 papers)
  - [0] MetaEmbed: Scaling Multimodal Retrieval at Test-Time with Flexible Late Interaction (Anon et al., 2026) View paper
  - [1] Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering (Lin Weizhe, 2023) View paper
  - [2] Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers (Lin Weizhe, 2024) View paper
  - [3] CLaMR: Contextualized Late-Interaction for Multimodal Content Retrieval (Wan, 2025) View paper
  - [4] Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval (Arun Reddy, 2025) View paper
  - Embedding Efficiency and Scalability (2 papers)
  - [16] PyLate: Flexible Training and Retrieval for Late Interaction Models (Antoine Chaffin, 2025) View paper
  - [27] SLIM: Sparsified Late Interaction for Multi-Vector Retrieval with Inverted Indexes (Minghan Li, 2023) View paper
  - Unified Multimodal Embedding Models (3 papers)
  - [6] Llama nemoretriever colembed: Top-performing text-image retrieval model (Xu Mengyao, 2025) View paper
  - [21] jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval (GÃ¼nther, 2025) View paper
  - [22] ModernVBERT: Towards Smaller Visual Document Retrievers (Paul Teiletche, 2025) View paper
- Domain-Specific Multimodal Retrieval Applications
  - Visual Question Answering and Knowledge Retrieval (2 papers)
  - [11] Multimodal Retrieval-Augmented Generation Question-Answering System (Chen, 2025) View paper
  - [13] Developing Visual Augmented Q&A System using Scalable Vision Embedding Retrieval & Late Interaction Re-ranker (Saxena, 2025) View paper
  - Document and Visual Content Retrieval (2 papers)
  - [7] ColMate: Contrastive Late Interaction and Masked Text for Multimodal Document Retrieval (Ahmed Masry, 2025) View paper
  - [9] ArtSeek: Deep artwork understanding via multimodal in-context reasoning and late interaction retrieval (Vessio Gennaro, 2025) View paper
  - E-Commerce and Recommendation Systems (4 papers)
  - [5] Notellm-2: Multimodal large representation models for recommendation (Chao Zhang, 2025) View paper
  - [17] Rethinking Convolutional Neural Network in Multimodal Sequential Recommendation (Z Zhou, 2025) View paper
  - [18] UniECS: Unified Multimodal E-Commerce Search Framework with Gated Cross-modal Fusion (Zihan Liang, 2025) View paper
  - [28] MTSTRec: Multimodal Time-Aligned Shared Token Recommender (MY Hong, n.d.) View paper

- Multimodal Fusion Strategies and Cross-Modal Interaction
  - Adaptive and Query-Dependent Fusion (2 papers)
  - [14] Inter-modal Interactions in Multimodal Learning: A Comprehensive Survey (Yunfeng Fan, 2025) View paper
  - [15] Query-adaptive late fusion for hierarchical fine-grained video-text retrieval (Wentao Ma, 2022) View paper
  - Late Fusion and Multi-Level Integration (6 papers)
  - [8] Multi-modal and multi-scale temporal fusion architecture search for audio-visual video parsing (Jiayi Zhang, 2023) View paper
  - [10] Multimodal fusion for multimedia analysis: a survey (P. Atrey, 2010) View paper
  - [19] A Late Fusion Approach Using CSNNs for Multi-Modal Toxicity Detection in Online Media (Thu N. Nguyen, 2025) View paper
  - [20] A Comparison of Late-Fusion Training Strategies for Quad-Modal Joint Embeddings (Domenic Luca FÃ¼rer, 2024) View paper
  - [23] Late Fusion and Multi-Level Fission Amplify Cross-Modal Transfer in Text-Speech LMs (Cuervo, 2025) View paper
  - [26] Video Memorability Prediction Via Late Fusion Of Deep Multi-Modal Features (Roberto Leyva, 2021) View paper
  - Cross-Modal Adapters and Prompt Learning (2 papers)
  - [24] CROME: Cross-Modal Adapters for Efficient Multimodal LLM (Ebrahimi, 2024) View paper
  - [25] Bilateral Adaptive Cross-Modal Fusion Prompt Learning for CLIP (Qiang Wang, 2024) View paper
- General Multimodal Interface and Interaction Frameworks (1 papers)
  - [12] Multimodal interfaces (Sharon Oviatt, 2007) View paper

Narrative

Core task: multimodal retrieval with flexible late interaction embeddings. This field centers on architectures that defer the fusion of text, image, video, and other modalities until a late stage, enabling fine-grained token-level matching rather than collapsing representations into single vectors. The taxonomy reflects four main branches: Late Interaction Architecture Design and Optimization focuses on the mechanics of token-level interaction and matching mechanisms, exploring how to efficiently compute cross-modal similarities at a granular level (e.g., Fine-grained Late Interaction[1], Preflmr[2], CLaMR[3]). Domain-Specific Multimodal Retrieval Applications adapts these techniques to specialized contexts such as video search (Video-ColBERT[4]) or art retrieval (ArtSeek[9]). Multimodal Fusion Strategies and Cross-Modal Interaction examines broader fusion paradigms, including query-adaptive and bilateral approaches (Query-adaptive Late Fusion[15], Bilateral Adaptive Fusion[25]), while General Multimodal Interface and Interaction Frameworks addresses user-facing systems and retrieval-augmented generation pipelines (Multimodal RAG[11]).

Recent work has concentrated on balancing expressiveness with computational efficiency: some studies push toward richer token-level alignments to capture nuanced semantic correspondences, while others seek lightweight architectures suitable for large-scale deployment (PyLate[16], Jina Embeddings v4[21]). MetaEmbed[0] sits squarely within the Token-Level Interaction and Matching Mechanisms cluster, emphasizing flexible embedding strategies that adapt interaction granularity based on query and document characteristics. Compared to neighbors like CLaMR[3], which also targets fine-grained matching, MetaEmbed[0] appears to prioritize meta-learning or parameterized flexibility in how tokens are aligned, rather than a fixed interaction schema. This contrasts with Video-ColBERT[4], which specializes in temporal video retrieval, illustrating how the same late-interaction principle can be tailored to different modalities and application constraints. Open questions remain around optimal token budget allocation, cross-modal attention mechanisms, and generalization across diverse retrieval benchmarks.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering

Authors: Lin Weizhe, Chen Jinghong, Weizhe Lin, Jinghong Chen, Jingbiao Mei, et al. (9 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Knowledge-based Visual Question Answering (KB-VQA) requires VQA systems to utilize knowledge from external knowledge bases to answer visually-grounded questions. Retrieval-Augmented Visual Question Answering (RA-VQA), a strong framework to tackle KB-VQA, first retrieves related documents with Dense Passage Retrieval (DPR) and then uses them to answer questions. This paper proposes Fine-grained Late-interaction Multi-modal Retrieval (FLMR) which significantly improves knowledge retrieval in RA-VQ...

Relationship Analysis

Both papers belong to the Token-Level Interaction and Matching Mechanisms category, focusing on fine-grained token-wise interactions for multimodal retrieval. They overlap in using late interaction mechanisms with multi-vector embeddings to compute similarity scores through token-level matching (MaxSim operations). The key difference is that the original paper (MetaEmbed) introduces learnable Meta Tokens with Matryoshka-style nested training for flexible test-time scaling, while the candidate paper (FLMR) focuses on aligning vision models with text retrievers through a mapping network and emphasizes cross-modality token interactions for knowledge-based VQA retrieval.

---

2. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers

Authors: Lin Weizhe, Weizhe Lin, Chen Jinghong, Jingbiao Mei, Byrne, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Large Multimodal Models (LMMs) excel in natural language and visual understanding but are challenged by exacting tasks such as Knowledge-based Visual Question Answering (KB-VQA) which involve the retrieval of relevant information from document collections to use in shaping answers to questions. We present an extensive training and evaluation framework, M2KR, for KB-VQA. M2KR contains a collection of vision and language tasks which we have incorporated into a single suite of benchmark tasks for t...

Relationship Analysis

Both papers belong to the Token-Level Interaction and Matching Mechanisms category, focusing on fine-grained late interaction approaches for multimodal retrieval. They overlap in using token-wise similarity computations (MaxSim operations) between query and document representations to enable more expressive matching than single-vector methods. However, MetaEmbed introduces learnable Meta Tokens with Matryoshka-style nested training for flexible test-time scaling, while PreFLMR extends ColBERT-style late interaction by incorporating vision encoders with cross-attention mapping structures and focuses on multi-task knowledge-based VQA retrieval across diverse datasets.

---

3. CLaMR: Contextualized Late-Interaction for Multimodal Content Retrieval

Authors: Wan, David, Wang Han, David Wan, Stengel-Eskin, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Online video web content is richly multimodal: a single video blends vision, speech, ambient audio, and on-screen text. Retrieval systems typically treat these modalities as independent retrieval sources, which can lead to noisy and subpar retrieval. We explore multimodal video content retrieval, where relevance can be scored from one particular modality or jointly across multiple modalities simultaneously. Consequently, an effective retriever must dynamically choose which modality (or set of mo...

Relationship Analysis

Both papers belong to the Token-Level Interaction and Matching Mechanisms category, focusing on fine-grained token-wise or patch-wise interaction schemes for multimodal retrieval. They overlap in their use of late interaction mechanisms to compute similarity between multimodal representations, with both employing token-level matching rather than single-vector approaches. However, MetaEmbed introduces learnable Meta Tokens with Matryoshka Multi-Vector Retrieval for test-time scaling flexibility, while CLaMR focuses on contextualized joint encoding of multiple modalities (video, audio, OCR, metadata) with modality-aware contrastive learning to dynamically select relevant modalities during retrieval.

---

4. Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval

Authors: Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In this work, we tackle the problem of text-to-video retrieval (T2VR). Inspired by the success of late interaction techniques in text-document, text-image, and text-video retrieval, our approach, Video-ColBERT, introduces a simple and efficient mechanism for fine-grained similarity assessment between queries and videos. Video-ColBERT is built upon three main components: a fine-grained spatial and temporal token-wise interaction, query and visual expansions, and a dual sigmoid loss during trainin...

Relationship Analysis

Both papers belong to the Token-Level Interaction and Matching Mechanisms category, employing fine-grained token-wise or patch-wise interaction schemes for multimodal retrieval. While MetaEmbed introduces learnable Meta Tokens to create compact multi-vector embeddings with Matryoshka-style nested training for flexible test-time scaling, Video-ColBERT applies ColBERT-style MaxSim operations specifically to text-to-video retrieval by performing dual-level interactions on both static frame features and temporally contextualized video features. The key distinction is that MetaEmbed focuses on general multimodal retrieval with controllable embedding granularity, whereas Video-ColBERT specializes in video retrieval with spatial-temporal interaction design.

Contributions Analysis

Overall novelty summary. The paper introduces MetaEmbed, a framework that uses learnable Meta Tokens to produce compact multi-vector embeddings for multimodal retrieval, combined with Matryoshka Multi-Vector Retrieval training to organize information by granularity. It resides in the Token-Level Interaction and Matching Mechanisms leaf, which contains five papers total, including the original work. This leaf sits within the broader Late Interaction Architecture Design and Optimization branch, indicating a moderately populated research direction focused on fine-grained matching strategies rather than single-vector compression or domain-specific applications.

The taxonomy reveals three sibling leaves within Late Interaction Architecture Design: Embedding Efficiency and Scalability (two papers on sparsification and compression), and Unified Multimodal Embedding Models (three papers on joint text-image architectures). Neighboring branches address Domain-Specific Applications (e.g., visual question answering, e-commerce) and Multimodal Fusion Strategies (adaptive fusion, cross-modal adapters). MetaEmbed's focus on flexible token-level interaction distinguishes it from efficiency-centric methods like PyLate and from unified embedding models that prioritize single-vector objectives, positioning it as an architectural innovation within the late-interaction paradigm.

Among twenty-seven candidates examined via semantic search, none were flagged as clearly refuting any of the three contributions. The MetaEmbed framework and Meta Tokens concept examined ten candidates with zero refutable overlaps; the Matryoshka Multi-Vector Retrieval training method also examined ten candidates with no refutations; and the test-time scaling mechanism examined seven candidates, again with no refutations. This suggests that within the limited search scope—top-K semantic matches plus citation expansion—no prior work was found to substantially overlap with the proposed techniques, though the analysis does not claim exhaustive coverage of the entire field.

Given the moderate density of the Token-Level Interaction leaf and the absence of refutable prior work among the examined candidates, the contributions appear relatively novel within the scope analyzed. However, the search examined only twenty-seven papers, and the taxonomy contains twenty-eight total papers across ten leaves, so broader or deeper literature searches might uncover additional related work. The assessment reflects what is visible in the current taxonomy structure and candidate set, not a definitive claim about the entire multimodal retrieval landscape.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: MetaEmbed framework with Meta Tokens for multimodal retrieval

Description: The authors propose MetaEmbed, a framework that appends a small number of learnable Meta Tokens to input sequences. Their last-layer contextualized representations serve as compact yet expressive multi-vector embeddings for retrieval, reducing the number of vectors needed while maintaining quality.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning Compact Vision Tokens for Efficient Large Multimodal Models

URL: View paper

Brief Assessment

Compact Vision Tokens[51] focuses on reducing spatial redundancy in vision tokens for efficient LMM inference, not on multimodal retrieval with learnable meta tokens for compact multi-vector embeddings as in the original paper.

---

2. Visual Semantic Contextualization Network for Multi-Query Image Retrieval

URL: View paper

Brief Assessment

Visual Semantic Contextualization[47] focuses on multi-query image retrieval with a [cls] token approach, not on learnable meta tokens for flexible multi-vector embeddings as in MetaEmbed.

---

3. Representation Learning for Visual Tasks: A Study of Attention and Information Selection

URL: View paper

Brief Assessment

Attention Information Selection[55] focuses on visual representation learning using register tokens in vision transformers for image retrieval, not on multimodal (text-image) retrieval with learnable meta tokens appended to input sequences as in MetaEmbed.

---

4. Multi-vector attention models for deep re-ranking

URL: View paper

Brief Assessment

Multi-vector Attention[52] focuses on text-based document retrieval using multi-vector attention mechanisms, not on multimodal retrieval with learnable meta tokens. The candidate addresses a different problem domain (text passage retrieval) with a different technical approach (attention-based pooling vs. learnable meta tokens).

---

5. Efficient token-guided image-text retrieval with consistent multimodal contrastive training

URL: View paper

Brief Assessment

Efficient Token-guided Retrieval[48] focuses on image-text retrieval using pre-extracted region features from Faster R-CNN and BERT embeddings, processed through dual transformers. It does not propose learnable meta tokens appended to input sequences as compact multi-vector embeddings.

---

6. Cross-Modal Retrieval and Semantic Refinement for Remote Sensing Image Captioning

URL: View paper

Brief Assessment

Cross-Modal Remote Sensing[49] focuses on remote sensing image captioning using cross-modal retrieval for sentence retrieval and semantic refinement, not on general multimodal retrieval with learnable meta tokens for compact multi-vector embeddings. The technical approaches and problem domains are fundamentally different.

---

7. HybridToken-VLM: Hybrid Token Compression for Vision-Language Models

URL: View paper

Brief Assessment

HybridToken-VLM[46] addresses vision-language model efficiency through visual token compression for VLMs, not multimodal retrieval embeddings. The candidate compresses visual tokens for generative VLM inference, while the original paper focuses on retrieval-specific multi-vector embeddings.

---

8. ACE: A Generative Cross-Modal Retrieval Framework with Coarse-To-Fine Semantic Modeling

URL: View paper

Brief Assessment

ACE[53] focuses on generative cross-modal retrieval using discrete token sequences from K-means and RQ-VAE for identifier construction, not learnable meta tokens appended to input sequences for multi-vector embeddings. The technical approaches are fundamentally different.

---

9. ReMatch: Boosting Representation through Matching for Multimodal Retrieval

URL: View paper

Brief Assessment

ReMatch[54] focuses on a different technical approach: it uses learnable tokens with chat-style generative matching and orthogonal regularization, rather than the matryoshka multi-vector retrieval framework proposed in the original paper. The candidate does not demonstrate that similar prior work exists for the original's specific nested multi-vector approach with test-time scaling.

---

10. VisionSelector: End-to-End Learnable Visual Token Compression for Efficient Multimodal LLMs

URL: View paper

Brief Assessment

VisionSelector[50] focuses on visual token compression for efficient MLLM inference through learnable selection mechanisms, not on multimodal retrieval with multi-vector embeddings. The candidate addresses a different problem domain (compression for generation tasks) rather than retrieval-oriented embedding construction.

Contribution 2: Matryoshka Multi-Vector Retrieval (MMR) training method

Description: The authors introduce MMR, a training approach that organizes embeddings into hierarchical nested groups. By performing contrastive learning across parallel nested groups, the model learns coarse-to-fine multi-vector embeddings that enable flexible retrieval at different granularities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Enhanced hierarchical contrastive learning for recommendation

URL: View paper

Brief Assessment

Enhanced Hierarchical Contrastive[38] focuses on recommendation systems using topic and semantic graphs with contrastive learning for user-item matching, not on hierarchical nested embeddings for flexible multi-vector retrieval at different granularities.

---

2. HierVision: Standardized and Reproducible Hierarchical Sources for Vision Datasets

URL: View paper

Brief Assessment

HierVision[39] focuses on standardizing and organizing hierarchical metadata for vision datasets, not on training methods for multi-vector retrieval or contrastive learning across nested embedding groups.

---

3. Hierarchical contrastive learning with multiple augmentations for sequential recommendation

URL: View paper

Brief Assessment

Hierarchical Contrastive Sequential[40] focuses on sequential recommendation with user behavior sequences, not multimodal retrieval. While both use hierarchical structures and contrastive learning, the candidate applies augmentations to create low-level and high-level

view pairs for sequential patterns, whereas the original paper organizes embeddings into nested groups for flexible retrieval granularity in multimodal contexts.

---

4. Self-Supervised Learning of Dense Hierarchical Representations for Medical Image Segmentation

URL: View paper

Brief Assessment

Dense Hierarchical Representations[44] focuses on medical image segmentation using hierarchical feature pyramids for voxel-wise representations, not multi-vector retrieval with nested embeddings for information retrieval tasks.

---

5. Hihpq: Hierarchical hyperbolic product quantization for unsupervised image retrieval

URL: View paper

Brief Assessment

Hihpq[41] focuses on hierarchical hyperbolic product quantization for image retrieval, not multi-vector retrieval with nested embeddings. The hierarchical structure in Hihpq[41] refers to semantic clustering levels, not nested embedding groups for flexible retrieval granularity.

---

6. HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention

URL: View paper

Brief Assessment

HiCLIP[43] focuses on hierarchical attention mechanisms for vision-language pretraining but does not address multi-vector retrieval with nested embeddings or test-time scaling capabilities that characterize MMR.

---

7. HGCL: Hierarchical Graph Contrastive Learning for User-Item Recommendation

URL: View paper

Brief Assessment

HGCL[36] focuses on hierarchical graph contrastive learning for user-item recommendation systems, not on multi-vector retrieval with nested embeddings for flexible granularity. The hierarchical structures in HGCL[36] are built through item clustering for recommendation tasks, which is fundamentally different from the nested multi-vector embeddings with coarse-to-fine retrieval proposed in the original paper.

---

8. Hierarchicalcontrast: A coarse-to-fine contrastive learning framework for cross-domain zero-shot slot filling

URL: View paper

Brief Assessment

HierarchicalContrast[42] focuses on slot filling in task-oriented dialogue using coarse-to-fine contrastive learning with entity-level and token-level labels. The ORIGINAL paper's MMR organizes embeddings into hierarchical nested groups for flexible multi-vector retrieval at different granularities, which is a fundamentally different application and technical approach.

---

9. Use all the labels: A hierarchical multi-label contrastive learning framework

URL: View paper

Brief Assessment

Hierarchical Multi-label Learning[37] focuses on hierarchical multi-label classification with contrastive learning applied to label hierarchies in classification tasks, not on organizing multi-vector embeddings for flexible retrieval at different granularities as in MMR.

---

10. Retrieval-style in-context learning for few-shot hierarchical text classification

URL: View paper

Brief Assessment

Retrieval-style Hierarchical Classification[45] focuses on hierarchical text classification with label-aware representations using contrastive learning for similar labels, not on organizing embeddings into nested groups for flexible multi-granularity retrieval as in MMR.

Contribution 3: Test-time scaling mechanism for multimodal retrieval

Description: The authors enable a test-time scaling capability where users can dynamically adjust the number of Meta Embeddings used during retrieval. This allows flexible trade-offs between retrieval accuracy, index size, and latency without retraining the model.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Towards Multi-Granularity Memory Association and Selection for Long-Term Conversational Agents

URL: View paper

Brief Assessment

Multi-Granularity Memory[31] focuses on conversational memory management with adaptive granularity selection for dialogue contexts, not multimodal retrieval with flexible late interaction embeddings.

---

2. Hvlad: Point Cloud Based Large-Scale Place Recognition Network in Dusty Environment

URL: View paper

Brief Assessment

Hvlad[35] focuses on point cloud-based place recognition in dusty environments, not multimodal retrieval with dynamic embedding granularity. The candidate addresses a completely different domain (3D point clouds vs. multimodal text-image) and does not discuss test-time scaling or adaptive retrieval mechanisms.

---

3. Enhancing Retrieval-Augmented Generation via Dual-Granularity Document Indexing

URL: View paper

Brief Assessment

Dual-Granularity Indexing[33] focuses on text segmentation strategies (dual chunking with different segment sizes) for document retrieval, not on dynamic adjustment of embedding vectors during retrieval. The candidate does not address test-time scaling of multi-vector embeddings or flexible trade-offs in multimodal retrieval systems.

---

4. Dynamic embedding size search with minimum regret for streaming recommender system

URL: View paper

Brief Assessment

Dynamic Embedding Size[29] addresses streaming recommender systems with dynamic embedding size selection for users/items, not multimodal retrieval with test-time scaling of meta embeddings. The domains (recommendation vs. retrieval) and mechanisms (embedding size search vs. multi-vector granularity) are fundamentally different.

---

5. Towards Efficient and Robust Moment Retrieval System: A Unified Framework for Multi-Granularity Models and Temporal Reranking

URL: View paper

Brief Assessment

Moment Retrieval System[30] focuses on video moment retrieval with ensemble search strategies and temporal reranking, not on dynamic embedding granularity adjustment for multimodal retrieval systems.

---

6. Towards a Smaller Student: Capacity Dynamic Distillation for Efficient Image Retrieval

URL: View paper

Brief Assessment

Capacity Dynamic Distillation[34] focuses on dynamic model compression during training for efficient image retrieval using knowledge distillation, not on test-time scaling of multimodal embeddings with flexible granularity adjustments.

---

7. Progressively optimized bi-granular document representation for scalable embedding based retrieval

URL: View paper

Brief Assessment

Bi-granular Document Representation[32] focuses on text-based ad-hoc search with a two-stage sparse-dense retrieval architecture for memory efficiency, not on multimodal retrieval or test-time scaling with adjustable embedding granularity. The candidate addresses corpus-scale memory constraints through bi-granular indexing (sparse in-memory, dense on-disk), while the original paper enables dynamic adjustment of multi-vector embeddings at test-time for accuracy-efficiency trade-offs in multimodal settings.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] MetaEmbed: Scaling Multimodal Retrieval at Test-Time with Flexible Late Interaction View paper
- [1] Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering View paper
- [2] Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers View paper
- [3] CLaMR: Contextualized Late-Interaction for Multimodal Content Retrieval View paper
- [4] Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval View paper
- [5] Notellm-2: Multimodal large representation models for recommendation View paper
- [6] Llama nemoretriever colembed: Top-performing text-image retrieval model View paper
- [7] ColMate: Contrastive Late Interaction and Masked Text for Multimodal Document Retrieval View paper
- [8] Multi-modal and multi-scale temporal fusion architecture search for audio-visual video parsing View paper
- [9] ArtSeek: Deep artwork understanding via multimodal in-context reasoning and late interaction retrieval View paper
- [10] Multimodal fusion for multimedia analysis: a survey View paper
- [11] Multimodal Retrieval-Augmented Generation Question-Answering System View paper
- [12] Multimodal interfaces View paper
- [13] Developing Visual Augmented Q&A System using Scalable Vision Embedding Retrieval & Late Interaction Re-ranker View paper
- [14] Inter-modal Interactions in Multimodal Learning: A Comprehensive Survey View paper
- [15] Query-adaptive late fusion for hierarchical fine-grained video-text retrieval View paper
- [16] PyLate: Flexible Training and Retrieval for Late Interaction Models View paper
- [17] Rethinking Convolutional Neural Network in Multimodal Sequential Recommendation View paper
- [18] UniECS: Unified Multimodal E-Commerce Search Framework with Gated Cross-modal Fusion View paper
- [19] A Late Fusion Approach Using CSNNs for Multi-Modal Toxicity Detection in Online Media View paper
- [20] A Comparison of Late-Fusion Training Strategies for Quad-Modal Joint Embeddings View paper
- [21] jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval View paper
- [22] ModernVBERT: Towards Smaller Visual Document Retrievers View paper
- [23] Late Fusion and Multi-Level Fission Amplify Cross-Modal Transfer in Text-Speech LMs View paper
- [24] CROME: Cross-Modal Adapters for Efficient Multimodal LLM View paper
- [25] Bilateral Adaptive Cross-Modal Fusion Prompt Learning for CLIP View paper
- [26] Video Memorability Prediction Via Late Fusion Of Deep Multi-Modal Features View paper
- [27] SLIM: Sparsified Late Interaction for Multi-Vector Retrieval with Inverted Indexes View paper
- [28] MTSTRec: Multimodal Time-Aligned Shared Token Recommender View paper
- [29] Dynamic embedding size search with minimum regret for streaming recommender system View paper
- [30] Towards Efficient and Robust Moment Retrieval System: A Unified Framework for Multi-Granularity Models and Temporal Reranking View paper
- [31] Towards Multi-Granularity Memory Association and Selection for Long-Term Conversational Agents View paper
- [32] Progressively optimized bi-granular document representation for scalable embedding based retrieval View paper
- [33] Enhancing Retrieval-Augmented Generation via Dual-Granularity Document Indexing View paper
- [34] Towards a Smaller Student: Capacity Dynamic Distillation for Efficient Image Retrieval View paper
- [35] Hvlad: Point Cloud Based Large-Scale Place Recognition Network in Dusty Environment View paper
- [36] HGCL: Hierarchical Graph Contrastive Learning for User-Item Recommendation View paper
- [37] Use all the labels: A hierarchical multi-label contrastive learning framework View paper
- [38] Enhanced hierarchical contrastive learning for recommendation View paper
- [39] HierVision: Standardized and Reproducible Hierarchical Sources for Vision Datasets View paper

- [40] Hierarchical contrastive learning with multiple augmentations for sequential recommendation View paper
- [41] Hihpq: Hierarchical hyperbolic product quantization for unsupervised image retrieval View paper
- [42] Hierarchicalcontrast: A coarse-to-fine contrastive learning framework for cross-domain zero-shot slot filling View paper
- [43] HiCLIP: Contrastive language-image pretraining with hierarchy-aware attention View paper
- [44] Self-Supervised Learning of Dense Hierarchical Representations for Medical Image Segmentation View paper
- [45] Retrieval-style in-context learning for few-shot hierarchical text classification View paper
- [46] HybridToken-VLM: Hybrid Token Compression for Vision-Language Models View paper
- [47] Visual Semantic Contextualization Network for Multi-Query Image Retrieval View paper
- [48] Efficient token-guided image-text retrieval with consistent multimodal contrastive training View paper
- [49] Cross-Modal Retrieval and Semantic Refinement for Remote Sensing Image Captioning View paper
- [50] VisionSelector: End-to-End Learnable Visual Token Compression for Efficient Multimodal LLMs View paper
- [51] Learning Compact Vision Tokens for Efficient Large Multimodal Models View paper
- [52] Multi-vector attention models for deep re-ranking View paper
- [53] ACE: A Generative Cross-Modal Retrieval Framework with Coarse-To-Fine Semantic Modeling View paper
- [54] ReMatch: Boosting Representation through Matching for Multimodal Retrieval View paper
- [55] Representation Learning for Visual Tasks: A Study of Attention and Information Selection View paper