Indexing Multimodal Language Models for Large-scale Image Retrieval

Bahey Tharwat[1*] Giorgos Kordopatis-Zilos[2*] Pavel Suma[2] Ian Reid[1] Giorgos Tolias[2]
[1]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2]VRG, FEE, Czech Technical University in Prague, Prague, Czech Republic

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated strong cross-modal reasoning capabilities, yet their potential for vision-only tasks remains underexplored. We investigate MLLMs as training-free similarity estimators for instance-level image-to-image retrieval. Our approach prompts the model with paired images and converts next-token probabilities into similarity scores, enabling zero-shot re-ranking within large-scale retrieval pipelines. This design avoids specialized architectures and fine-tuning, leveraging the rich visual discrimination learned during multimodal pre-training. We address scalability by combining MLLMs with memory-efficient indexing and top-k candidate re-ranking. Experiments across diverse benchmarks show that MLLMs outperform task-specific re-rankers outside their native domains and exhibit superior robustness to clutter, occlusion, and small objects. Despite strong results, we identify failure modes under severe appearance changes, highlighting opportunities for future research. Our findings position MLLMs as a promising alternative for open-world large-scale image retrieval.
code: http://github.com/gkordo/mllm-ilr

1. Introduction

Multimodal Large Language Models (MLLMs) [4, 13, 32, 36, 56] have recently emerged as powerful tools for bridging vision and language. Their ability to jointly process images and text enables rich cross-modal reasoning. From a computer vision perspective, MLLMs offer a unique advantage: they use natural language both as supervision during training and as an interaction mechanism at inference, providing remarkable flexibility. This makes them suitable for diverse tasks such as visual question answering (VQA) [1, 11, 33], image captioning [17, 23], object recognition [10, 42, 65], and scene understanding [24, 47, 50]. In these settings, language serves as an additional source of context, enhancing interpretability and adaptability.

The appeal of MLLMs is twofold: (i) natural language supervision is expressive and diverse, enabling learning from large multimodal corpora without task-specific annotations, and (ii) text-based interaction allows zero-shot adaptation to novel tasks through prompting [1, 33]. This combination of scalable training and flexible deployment motivates their use beyond traditional multimodal tasks.

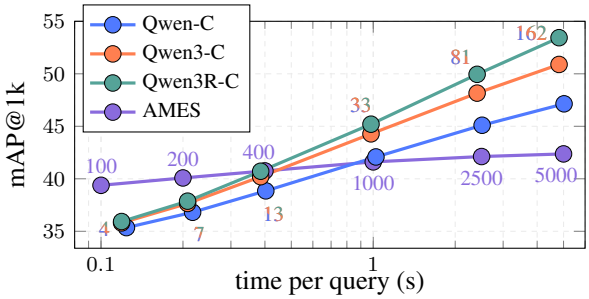

Figure 1. **Performance *vs*. re-ranking time.** MLLM-based re-rankers have higher per-image inference time than methods trained specifically for instance-level image retrieval with re-ranking (AMES), but under a fixed query-time budget, they achieve better retrieval performance, already with as few as 20 re-ranked images, indicated by the colored numbers. All reported methods have roughly the same memory footprint.

In this work, we explore MLLMs in a training-free manner for a vision-only task, *i.e.*, large-scale image retrieval. Our focus is on instance-level retrieval [37, 49], where the goal is to find the exact object rather than a category. Specifically, we provide the query and candidate images to the model and prompt it with a simple question, *e.g.* "Do the images show the same object?", requiring a binary (yes/no) response. We convert the next-token probabilities for "yes" and "no" into a similarity score, effectively turning the model into a zero-shot classifier for object identity.

This constitutes an unconventional application of MLLMs, as image similarity does not inherently involve language understanding. Our goal is to repurpose their reasoning capabilities, avoiding specialized architectures [3, 8], domain-specific training data [40, 62], and costly fine-tuning typically used for instance-level retrieval. We leverage the extensive pre-training of MLLMs across diverse domains, which implicitly equips them with strong visual discrimination in an open-world setting. Interestingly, their internal design often involves comparing image patches, reminiscent of similarity-learning architectures [52, 54], albeit without explicit supervision for this task.

*Equal contribution.

We integrate this similarity estimation into a large-scale retrieval pipeline as a re-ranking step on top of an efficient initial ranking using compact image descriptors. Since MLLMs require raw, often high-resolution images, direct application to large collections is impractical. To address this, we adopt indexing-friendly design choices: database images are pre-processed into patch-level features, quantized or pruned, and stored in memory-efficient structures. During retrieval, only the top-k candidates are processed by the MLLM, reducing computational overhead.

Our findings show that MLLMs, when prompted appropriately, serve as effective training-free similarity estimators for instance-level retrieval, offering not only a favorable accuracy-memory tradeoff but also a favorable accuracy-speed tradeoff when operating under a sufficiently generous time budget (see Figure 1). Compared to task-specific models, MLLMs perform better outside their native training domains, demonstrating superior generalization across diverse object categories. We show that our indexing scheme is effective not only for training-free usage of generic MLLMs but also for MLLM models specifically fine-tuned as re-rankers.

Qualitatively, MLLMs outperform global descriptors and task-specific re-rankers in challenging cases involving small objects, heavy clutter, and occlusion, showcasing strong partial matching capabilities. Despite these advantages, we identify unexpected failure modes in robustness analysis, paving the way for future research to enhance the reliability and scalability of MLLMs for similarity estimation and large-scale retrieval.

2. Related work

Image retrieval. Image retrieval has largely relied on learned global descriptors. Early deep methods repurpose classification networks by pooling activation maps [5, 40], while later approaches train networks explicitly for retrieval using metric learning objectives [43, 44] and advanced sampling strategies [35, 51]. Although effective, these methods require domain-specific training data, which limits adaptability. Foundation vision models such as CLIP [41] introduced generic representations with strong cross-domain generalization, enabling efficient large-scale retrieval. Yet, global descriptors often lack the precision needed for fine-grained and instance-level similarity estimation.

To improve precision, recent work has incorporated re-ranking based on local descriptors. State-of-the-art methods such as RRT [54], R^2Former [69], and CVNet [27] refine similarity estimation by matching dense features or modeling cross-image correlations with transformer or convolutional architectures. Since such dense approaches often incur high indexing costs, AMES [52] is designed for an asymmetric setup with binarized descriptors [25], while dense-to-sparse CVNet [28] sparsifies database descriptors. LOCORE [63] uses only a fraction of local descriptors and departs from pairwise similarity by fine-tuning an LLM for listwise re-ranking. Despite their strong performance, these methods remain computationally and memory-intensive and rely on domain-specific training. In contrast, we propose a training-free re-ranking approach based on MLLMs, leveraging their open-world generalization while avoiding specialized architectures and domain-dependent training.

MLLM for image retrieval. MLLMs [6, 13] offer reasoning capabilities beyond traditional embedding approaches. In retrieval, they have been studied mainly in settings where text is part of the query, such as composed image retrieval (CIR). Although some methods address CIR in a zero-shot manner [14, 15, 55], instance-level granularity is argued to require dedicated training [18, 31] or additional corpora with positive and negative samples [38].

Another line of work trains MLLMs to produce universal multimodal embeddings, including for image-to-image retrieval. These approaches rely on extensive fine-tuning, typically through contrastive objectives on large-scale paired data [20] or massive multimodal collections [21, 30, 61, 67]. This trend is motivated by the view that pre-trained MLLMs are weak at retrieval. For example, LamRA [34] trains a LoRA [16] module for retrieval and re-ranking, reporting poor performance when MLLMs are used directly, while MM-Embed [30] introduces modality-aware hard-negative fine-tuning, yet still relies on the pre-trained MLLM for re-ranking. To the best of our knowledge, we are the first to show that a pre-trained MLLM can serve directly as a strong zero-shot re-ranker for pure image-to-image, instance-level retrieval in a fully training-free setting.

Improved MLLM efficiency. While recent MLLMs can process high-resolution images for fine-grained retrieval [58, 60], indexing large-scale databases with them remains memory-intensive. FastV [9] attributes much of this cost to redundant visual tokens processed within transformer layers. To address this, prior work has explored visual token pruning within the transformer by identifying and removing less informative tokens, for example, by maximizing token diversity in DivPrune [2] or exploiting attention sparsity in LLaVA-PruMerge [46].

Other methods compress visual content into a smaller token set by modifying the feature extractor. TokenPacker [29] merges low-resolution queries with high-resolution details while BRAVE [22] uses a Multi-Encoder Querying Transformer to fuse and resample features from multiple vision encoders. In contrast, our work mainly focuses on training-free token-set compression to reduce the memory footprint.

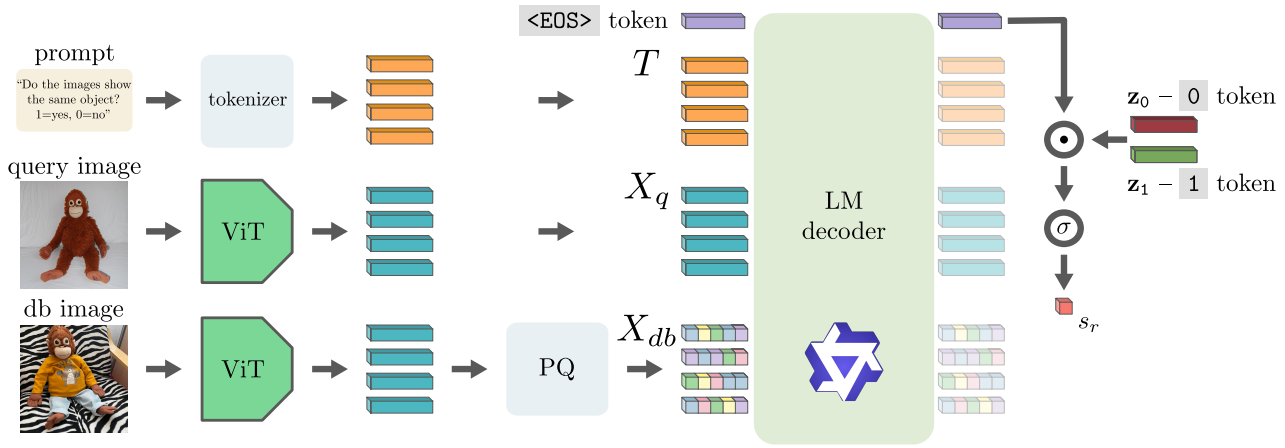

Figure 2. **Overview of the proposed MLLM-based re-ranking approach** for instance-level image retrieval. A query image, a database image, and a task-specific textual prompt are jointly used as input for similarity estimation. A vision encoder extracts visual tokens from both images (X_q, X_{db}), which are concatenated with the textual prompt tokens (T) and the end-of-sentence (EOS) token to form the multimodal input to the MLLM. Scalable retrieval is achieved by applying compression and indexing schemes, *e.g.* Product Quantization (PQ), to the database image tokens, enabling memory-efficient storage. The MLLM outputs the updated EOS token representation, which is the joint representation of all inputs and is then compared via dot product with the classifier vectors $\mathbf{z}_1$ and $\mathbf{z}_0$, corresponding to vocabulary tokens "1" and "0", respectively. The final similarity score (s_r) is given by the softmax (σ) probability assigned to token "1".

3. Method

3.1. Problem Formulation

The goal of image retrieval is to search a database of images given a query image q and return the most relevant results. We focus on the most fine-grained notion of relevance, *i.e.* instance-level retrieval, where two images are considered relevant if they depict the same specific object. Formally, retrieval requires a similarity function $s(q, x) \in \mathbb{R}$ that measures the similarity between the query and a database image x, enabling ranking of all candidates by similarity.

In large-scale settings with millions of images, efficiency in both speed and memory is critical. We adopt a two-stage pipeline: (i) an initial exhaustive search using compact global descriptors extracted by a vision encoder, producing a coarse ranking (global similarity $s_g(q, x)$) (ii) a re-ranking step applied to the top-k candidates using a more accurate but computationally intensive similarity estimation (re-ranking similarity $s_r(q, x)$).

Our re-ranking leverages an MLLM to estimate instance-level similarity between image pairs. This approach improves precision while controlling cost through k. However, re-ranking typically increases memory requirements, as it demands richer representations than a single global descriptor. Direct access to raw database images, often assumed by MLLMs, is impractical due to storage overhead and prohibitive query-time latency. To address this, we design an indexing-friendly solution that enables memory-efficient storage of image representations while exploiting the reasoning capabilities of MLLMs for re-ranking.

3.2. MLLM architecture overview

Given a set of images $\{x_1, \ldots, x_C\}$ and a text prompt p, the MLLM extracts visual tokens V and textual tokens T, and estimates the probability distribution over an output token sequence $y_{1:N}$ from a vocabulary Z as:

$$P(y_{1:N} \mid V, T) = \prod_{n=1}^{N} P\big(y_n \mid y_{1:n-1}, V, T\big), \qquad (1)$$

where y_n denotes the n-th generated token, and $P(\cdot)$ is a categorical distribution over Z.

Visual tokens are obtained via a vision encoder $g : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{M \times D}$, which maps an input image of height H and width W into a sequence of M D-dimensional embeddings. Each image x_i is mapped to a feature matrix $X_i = g(x_i)$. We employ a Vision Transformer (ViT) [12] for this purpose, which divides each image into non-overlapping patches and processes them through a transformer network [57]. The visual token sequence is formed by concatenating the encoded representations of all input images, *i.e.* $V = [X_1, \ldots, X_c] \in \mathbb{R}^{L_v \times D}$, where $L_v = C \cdot M$ denotes the total number of visual tokens.[1]

The textual prompt p is tokenized, and the corresponding token embeddings are retrieved from a lookup table, forming $T \in \mathbb{R}^{L_t \times D}$, where L_t is the number of textual prompt tokens. Generated tokens y_n are also mapped into the same embedding space as the prompt tokens.

We denote the language model by $f : \mathbb{R}^{(n+L_v+L_t) \times D} \rightarrow \mathbb{R}^{D}$, which processes the concatenated multimodal token sequence, comprising visual tokens, textual prompt tokens,

[1] For simplicity, we assume that all images produce the same number of tokens, which may not hold in practice due to varying image sizes.

previously generated tokens, and a special *end-of-sentence* (EOS) token, and maps them to a sequence of vector representations. The model is implemented as a transformer decoder, outputting one vector per input token. All outputs except the one corresponding to the EOS token are discarded; the EOS token's representation serves as the joint embedding of the entire input. Therefore, we represent f as a mapping to a single vector.

This representation is then passed through a linear softmax classifier with $|Z|$ class vectors, one per vocabulary token. Let $\mathbf{z}_i \in \mathbb{R}^D$ denote the class vector for the i-th token. The probability of generating token i at position n is given by:

$$P_i(t_n \mid t_{1:n-1}, V, T) = \frac{e^{\mathbf{z}_i^\top f([V,T,t_{1:n-1}])}}{\sum\limits_{j \in 1 \ldots |Z|} e^{\mathbf{z}_j^\top f([V,T,t_{1:n-1}])}} \quad (2)$$

3.3. Image-to-image similarity with MLLMs.

We repurpose the MLLM to estimate the similarity between a query image q and a database image x_{db}. The model is prompted with a textual instruction p that asks whether the two images depict the same object instance:

```
You are given two images:  a query and a
candidate.  Determine whether the exact
same object instance from the query image
is present in the candidate image.  Output
strictly a single digit, do not output
anything else.:
• 0 = the object instance does not appear.
• 1 = the object instance appears in the
  candidate.
```

The model is expected to generate the token "`1`" if the images show the same object, and "`0`" otherwise. Thus, the vocabulary of interest is restricted to these two tokens only, represented by vectors $\mathbf{z}_0$ and $\mathbf{z}_1$. In this setting, the model does not perform auto-regressive generation, and no multiple output tokens are considered. Consequently, the MLLM operates as a binary classifier that predicts if the two images depict the same object instance.

The input to the function f consists of the visual tokens extracted from the query and database images, along with the textual prompt tokens. Specifically, the input sequence is $u = [X_q, X_{\text{db}}, T]$, where $X_q = g(x_q)$ and $X_{\text{db}} = g(x_{\text{db}})$ are the visual token sequences for the query and database images, respectively, and T represents the tokenized prompt. The joint representation $\mathbf{u} = f(u)$ is obtained by processing this input through the MLLM.

To estimate similarity, we evaluate the probability assigned to token "`1`" using the softmax classifier

$$s_r(q, x_{\text{db}}) = \frac{e^{\mathbf{z}_1^\top \mathbf{u}}}{e^{\mathbf{z}_0^\top \mathbf{u}} + e^{\mathbf{z}_1^\top \mathbf{u}}}. \quad (3)$$

This probability serves as the re-ranking similarity score between the query image q and the database image x_{db}.

3.4. Compression and indexing.

During re-ranking, the representations of the query image and the textual prompt, denoted as X_q and T, are computed once per query and reused across all k candidate images. In contrast, the representations of the top-k database images, $X_{\text{db}} \in \mathbb{R}^{M \times D}$, must be accessed individually. To enable scalable retrieval, we explore strategies for indexing that involve compressing and storing these representations in advance, in a memory-efficient manner.

Let $\mathbf{x}_m \in \mathbb{R}^D$, for $m = 1, \ldots, M$, denote the representation of the m-th patch (token) extracted from a database image x_{db}. Our goal is to reduce both the memory footprint of each token and the total number of tokens per image. We investigate the following indexing strategies:

Product Quantization (PQ). Instead of storing full-precision token embeddings $\mathbf{x}_m \in \mathbb{R}^D$, we apply Product Quantization (PQ) [19] to obtain compact codes. The D-dimensional space is divided into d-dimensional subspaces, each associated with a codebook of K centroids, computed offline on a generic image set. Each vector is represented by indices of the nearest centroids in each subspace, resulting in D/d indices per vector. We adopt a standard configuration with $K = 256$ centroids per subspace, allowing each index to be stored in one byte, and each vector in D/d bytes. Prior to feeding into the model, compressed tokens are reconstructed by concatenating the corresponding centroids from each subspace, yielding an approximate full-precision vector. PQ enables a trade-off between memory usage and reconstruction fidelity by adjusting d.

Token pruning. We adapt DivPrune [2], a recent token pruning method designed for MLLMs. The algorithm implements a diverse selection, similar to Farthest Point Sampling (FPS). It begins by selecting the token furthest from its closest neighbor, then iteratively adds the token that is furthest from the set of tokens already selected.

k-means clustering for token selection. We apply k-means clustering to the tokens of each image, grouping them into K clusters. The centroids are the stored representations, while to define the spatial position of the centroid tokens when fed to the network, we pick the position of the medoids, *i.e.* tokens closest to the cluster centroids.

Uniform token sampling. To reduce the token count, we uniformly sample tokens from the original set. Specifically, we retain one token per non-overlapping 2×2 window on the 2D token grid, effectively downsampling by a factor of 4. The rest tokens are discarded. Selected tokens are preserved in their original form and accompanied by the original spatial positions to retain spatial context.

Average pooling with a sliding window. Instead of selecting a single token, we compute the average of the four tokens within each non-overlapping 2×2 window. This reduces the number of tokens by a factor of 4 while preserving local information in a compact form.

Resolution reduction as token reduction. The number of tokens M is proportional to the input image resolution. For a ViT with patch size P, the number of tokens scales as $M \sim \frac{H}{P} \times \frac{W}{P}$. Particularly, the Qwen model we use applies an additional 2×2 pooling on the 2D token grid, reducing the number of tokens by a factor of 4. Thus, reducing the input resolution directly decreases the number of tokens.

4. Experiments

4.1. Evaluation setting

Datasets and metric. We evaluate our method on several established datasets for instance-level image retrieval. Our main evaluation is conducted on ***ILIAS*** [26], a large-scale, multi-domain dataset designed for instance-level image retrieval. Due to its size, diversity, and challenging nature, ILIAS naturally reflects open-world settings. In addition, we evaluate on three other instance-level datasets, *i.e.* ***INSTRE*** [59], ***$\mathcal{R}$OP+1M*** [39], and ***Product1M*** [66]. Further details regarding the datasets are in the supplementary material. Performance is measured using mean Average Precision at rank k (mAP@k), which computes the mean of the average precision across all queries, truncated at the top k retrieved results. Unless specified otherwise, we evaluate mAP@$1k$.

Implementation details. As the default choice for the MLLM, we use the Qwen model [6], specifically the `Qwen2.5-VL-7B-Instruct` variant. In this configuration, the vision encoder is a 32-layer ViT with a patch size of 14. The Qwen2.5 family incorporates a token merger module that merges every four tokens before propagation to the LLM, resulting in an effective patch size of 28. An input image of resolution 560×420 produces $M = 300$ visual tokens, each of dimensionality $D = 3{,}584$. Hence, in the uncompressed setting, these tokens are stored in `float16` format, amounting to approximately 2MB of memory per image. We evaluate the indexing strategies using the following configurations: PQ with subspace dimensions $d \in \{4, 8, 16, 32, 64, 128\}$ learned on a set on 2M vectors extracted from LAION [45] images, input image resolutions $\{280, 392, 560, 756, 1008, 1232\}$, k-means clustering with $K \in \{150, 70\}$ clusters, and token pruning resulting in final token sets of $\{150, 70\}$ tokens. Following prior work [8, 27, 52], we use global-local similarity ensembling with fixed factor $\lambda = 0.5$ for all datasets. We re-rank the top-1k neighbors based on global similarity. Unless stated otherwise, we use 756px for the no-compression setting and PQ_{16} with 560px for the compression setting as the default configurations. The use of compression is denoted by C, *e.g.* Qwen-C.

Other methods. As a global descriptor we use the best-performing ones on ILIAS[2], *i.e.* Perception Encoder (PE) with linear adaptation [26] and DINOv3 [48]. Additionally, we evaluate and compare with AMES [52], a state-of-the-art re-ranking method for instance-level image retrieval, which is trained on GLDv2 [62], an instance-level dataset of landmarks. For a fair comparison, we train AMES with more powerful contemporary local descriptors extracted with DINOv3, using the public implementation provided by the authors. We use the fp variant that stores full-precision descriptors. AMES is evaluated with sets of $\{50, 100, 200, 400, 600\}$ local descriptors. We also benchmark the more recent MLLMs, InternVL3.5 [60] and Qwen3 [64], used in a training-free way. As MLLM-based models that are trained for re-ranking, we evaluate LamRA [34] (same variant as Qwen, provided by the authors) and Qwen3-Reranker (Qwen3R) [68]. Note that Qwen3R is not training-free as a re-ranking method, but its use remains training-free with respect to the instance-level nature of our retrieval task. In the supplementary material, we apply the proposed indexing to Qwen3 and Qwen3R to demonstrate its universal applicability and effectiveness.

4.2. Results

Performance *vs*. memory comparison. In Fig. 3, we compare performance against memory footprint across different methods. Various re-ranking models improve retrieval performance on top of global descriptor baselines, at the expense of extra memory. Uncompressed Qwen variants across multiple input resolutions provide consistent gains, with the 1008px variant achieving over 22 points higher mAP compared to global, but with three orders of magnitude higher memory footprint. The performance saturates for larger resolutions. Although InternVL achieves slightly higher performance even at a smaller resolution of 560px, its memory cost remains significantly larger (by almost 50%), due to the image tiling mechanism used for token extraction. Notably, with a more recent model, *i.e.* Qwen3, performance is higher with a lower memory footprint, highlighting the value of training-free methods benefiting from MLLM advances in a straightforward way. AMES, which is explicitly trained for this task on landmark images, yields smaller improvements but with a much lower memory footprint. Interestingly, the MLLM architecture, specifically the function f, closely resembles that of AMES, and is identical to RRT [54], a re-ranking model preceding AMES. These models are designed to cross-compare sparse sets of local descriptors within and across images, analogous to how the MLLM performs cross-image reasoning, but with a dense and rich set of local descriptors. Still, Qwen benefits from its broad generalization capabilities, having been exposed to diverse visual domains, beyond landmarks, during training.

[2] https://vrg.fel.cvut.cz/ilias/

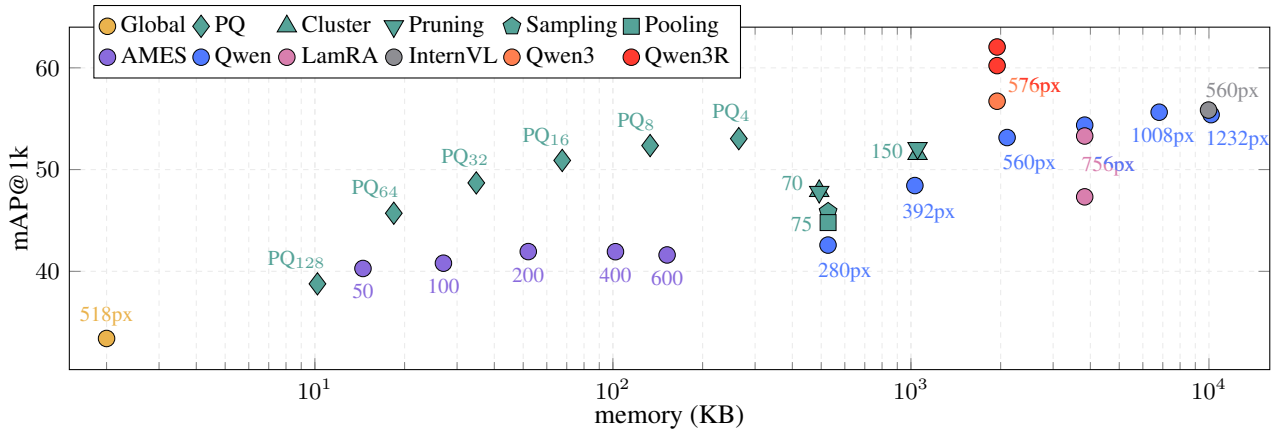

Figure 3. **Performance *vs*. memory.** Memory footprint per image is approximated assuming a fixed resolution, where the longer image side is set to the target value and the shorter side is scaled to 3/4 of it, reflecting a typical aspect ratio. Five models for re-ranking, operating on top of global similarities obtained from the PE model with linear adaptation. All indexing strategies are applied on the Qwen model with 560px image resolution. Text indicates: subspace dimension for PQ (*e.g*. PQ_{16} has 16 subspace dimensions), pixels for different resolutions (*e.g*. 560px), and number of tokens (*e.g*. 70) for the rest of the indexing methods. AMES used with descriptor sets of increasing cardinalities. LamRA and Qwen3R are shown both with their default prompts and also with our prompt, which performs better.

LamRA model performs comparably to Qwen when used with our prompt, indicating that general-purpose fine-tuning is not necessarily beneficial for instance-level retrieval. Nevertheless, results of Qwen3R show that its specialized architecture and fine-tuning can yield further improvements compared to the training-free variants.

Using indexing strategies on the representation of Qwen, *i.e*. PQ in particular, allows Qwen to provide a much better trade-off than AMES. None of the token selection methods, namely pooling, sampling, clustering, or pruning, stands out in terms of a good trade-off. Interestingly, DivPrune, a previously proposed method for a similar task, *i.e*. token pruning, is not better than our simple clustering-based baseline.

ELViS [53] is a concurrent work to ours, which achieves 52.0 mAP@1k, evaluated with 400 local descriptors from DINOv3, same as AMES. Its performance is much higher than AMES, and comparable to Qwen with PQ_8 for roughly the same memory. Its drawback is that it requires supervised training, but its advantage is speed, by re-ranking 1000 images in 0.1 seconds.

Inference latency. Our method is designed as a re-ranking approach that explicitly controls execution time via the shortlist size. While lighter re-rankers, such as AMES, achieve lower latency, their performance saturates quickly as top-k increases. Figure 1 reports performance as a function of re-ranking time. For short re-ranking time, AMES provides a better trade-off by re-ranking more database images; however, our MLLM-based approach scales substantially better as time increases, especially with the more recent models, yielding significantly larger performance gains. For fairness, AMES runtime is measured using a batch size that matches the GPU VRAM usage of Qwen-C for a single pair (≈16 GB).

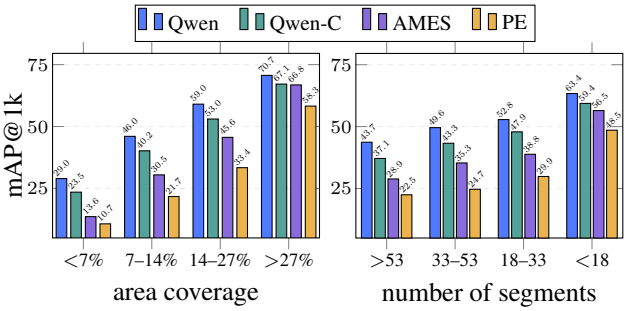

Figure 4. **Performance comparison across different amounts of object area coverage and background clutter.** Positives across all queries are jointly ranked based on coverage or clutter and split into 4 equal-sized groups. mAP@1k averaged over the queries in each group. Comparison between Qwen with and without compression (Qwen-C), AMES, and PE.

Robustness to scale and clutter on ILIAS. Following the original work [26], we group query–positive pairs based on object area coverage and the degree of background clutter, and report average performance over queries that contain at least one positive example within each group. As shown in Fig. 4, the MLLM-based re-ranking improves consistently across all cases. The improvement margin widens notably in challenging conditions – specifically, when objects occupy a small fraction of the image or when scenes exhibit substantial clutter. This trend holds for both the uncompressed and compressed Qwen variants, where compression compromises the performance in a similar way across the whole range. While AMES, which relies on local descriptors, also consistently improves over global, the MLLM demonstrates superior robustness, particularly for small objects and heavily cluttered scenes.

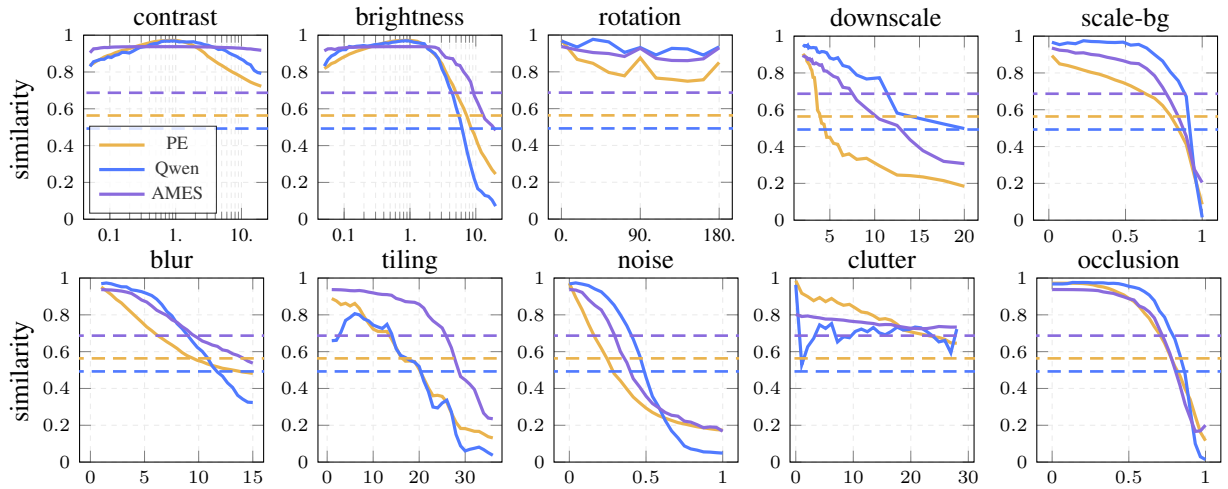

Figure 5. **Robustness analysis of MLLMs under controlled transformations.** Similarity scores for positive image pairs are shown across ten types of visual perturbations: contrast, brightness, rotation, scale, background scaling, blur, tiling, noise, clutter, and occlusion. 50 queries from ILIAS are used to generate positive pairs. Dashed lines indicate the average hard-negative similarity of the 5th percentile of the negatives in the re-ranking list, acting as a baseline. The crossing between the solid and dashed lines indicates the robustness point. Comparison between Qwen without compression, AMES, and PE. The horizontal axis represents the strength of each transformation.

Robustness analysis. To better understand the behavior of the MLLMs, we perform a robustness analysis by artificially generating positive image pairs through controlled transformations. Each transformation has a continuous *factor* that determines its strength – the larger the factor, the stronger the applied distortion. Details of the transformations and their factor ranges are provided in the supplementary material. Additionally, to contextualize the absolute scores and enable model-specific relative comparisons, we establish a hard-negative baseline for each model. For each query, we compute its similarity scores against all negative database images and extract the 5th percentile. The average of these 5th-percentile values across all queries serves as the indicative negative baseline for each model. Fig. 5 illustrates the estimated similarity scores across ten types of transformations for each model, alongside each model's corresponding negative baseline, visualized as dashed lines. Robustness is evaluated based on the factor value at which the model's similarity for positive pairs drops below the negative baseline. Qwen exhibits strong robustness in several challenging cases, including scale variation, noise addition, Gaussian blur, and occlusion, consistently outperforming the other methods. The only exceptions are *brightness* and *tiling* transformation, which simulate illumination changes and occlusion combined with background clutter, respectively, and where AMES demonstrates significantly better robustness. This indicates that drastic appearance changes pose a challenge for MLLMs, exposing a limitation that paves the way for future work. The rest of the cases are easier, where all models demonstrate robustness.

Furthermore, to better understand the model's behavior, we test the effect of geometric consistency between the patches of the two images on Qwen's decoder, by shuffling positional encodings. Specifically, we shuffle position IDs after extracting visual tokens, just before applying function f. Shuffling only the query reduces performance on ILIAS to 52.6 (-1.8), while shuffling both the query and database further drops it to 51.6 (-2.8). The relatively small degradation indicates that Qwen primarily depends on appearance cues, with geometric consistency playing a minor role.

Method	ILIAS	INSTRE	$\mathcal{R}$OP	Prod1M
PE [7] with linear adaptation [26]				
No re-rank	33.4	92.6	60.7	66.5
AMES [52]	41.9	91.9	**79.5**	66.8
LamRA [34]	47.8	95.3	67.5	72.9
Qwen [6]	41.0	94.6	63.0	72.3
LamRA	53.3	**96.6**	66.5	73.1
Qwen	**54.4**	96.4	68.1	**74.5**
Qwen-C	50.9	95.2	65.8	73.6
DINOv3 [48]				
No re-rank	26.5	82.2	67.5	52.2
AMES [52]	34.2	85.1	**82.1**	54.1
LamRA [34]	39.4	89.1	67.2	62.7
Qwen [6]	32.8	89.1	64.5	62.4
LamRA	41.2	90.5	67.9	62.2
Qwen	**41.7**	**90.7**	71.1	**64.4**
Qwen-C	38.6	89.7	68.8	63.8

Table 1. **Performance (mAP@1k) comparison on multiple datasets** using two different global descriptors. We present four instance-level to show the different behavior between training-free Qwen and LaMRA, which is fine-tuned but for more generic retrieval tasks. AMES is reported using 400 local descriptors (102kB), which achieves the best performance on average. Gray indicates the use of our task-specific prompt; otherwise, a generic prompt is used.

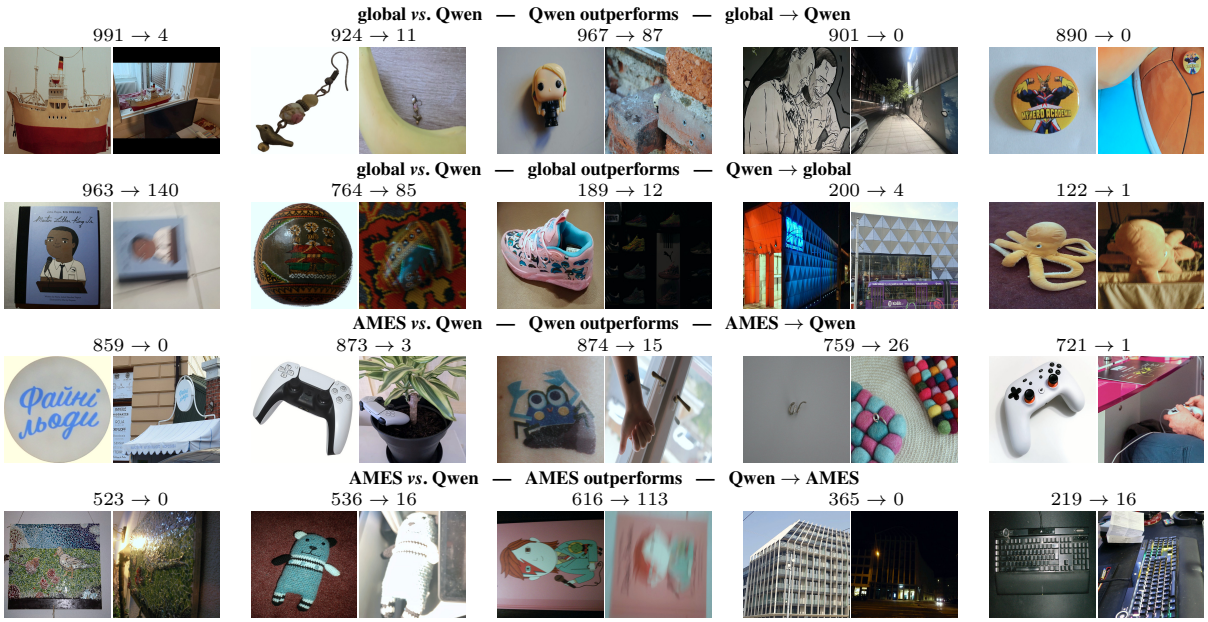

Figure 6. **Qualitative examples where one method benefits the most compared to another.** We compare global (PE) and AMES *vs.* Qwen by showing pairs of query and positive image. → indicates the number of negative images ranked before the positive for two models, and it goes from the weaker to the stronger model for each pair.

Performance on multiple datasets. In Tab. 1, we compare various re-ranking methods across multiple datasets. AMES achieves the best performance only on the landmarks dataset, *i.e.* its training domain, while, apart from ILIAS, it provides minimal or no improvements over global due to the domain gap. LamRA, when used with its original prompt and despite being trained for retrieval, consistently underperforms compared to the training-free Qwen, prompted appropriately. With our revised prompts, LamRA exhibits a substantial performance increase; however, even under this setting, Qwen generally outperforms LamRA. This suggests that although LamRA's training on a diverse set of queries and domains may compromise its performance relative to the training-free variant in some scenarios, it compromises its effectiveness on the most fine-grained task, such as instance-level retrieval. A similar trend is observed when DINOv3 is used for computing global similarities, indicating that the conclusions hold regardless of the model used for the selection of the top-k candidates.

Qualitative analysis. Delving deeper into the model comparisons, we conduct a qualitative comparison between global ranking with PE, MLLM-based re-ranking with Qwen, and AMES re-ranking. Figure 6 presents examples where one method outperforms the others, measured by the number of negative images ranked before the positive one.

Global vs. Qwen: Qwen significantly improves upon the global ranking, particularly in scenarios involving objects at varying scales, small objects, cluttered scenes, and partial occlusions. This suggests that Qwen possesses strong spatial reasoning, fine-grained recognition capabilities, and the ability to perform partial matching. However, there are instances where Qwen degrades performance, especially in cases affected by motion blur or significant appearance changes due to factors such as illumination. This aligns with our observations from the robustness analysis.

AMES vs. Qwen: Qwen consistently outperforms AMES in handling small objects, occlusions, and scenes with severe visual clutter, reinforcing the observations from the previous comparison. Conversely, AMES demonstrates superior performance in scenarios involving illumination changes, motion blur, and large viewpoint variations; similarly, weaknesses are exposed in our robustness analysis.

5. Conclusion

We present a training-free approach for instance-level image retrieval that leverages MLLMs as universal similarity estimators. By prompting them with task-oriented instructions, we repurpose them into effective image-to-image comparators capable of detecting object instances and translating their output to similarity estimation without any additional training. To enable scalability, we investigate indexing strategies that significantly reduce memory cost with marginal loss in performance. Digging deeper, we reveal that MLLMs, particularly Qwen, exhibit strong robustness to common visual transformations such as scale variation, occlusion, and clutter; however, they are vulnerable to drastic appearance changes, such as illumination changes and motion blur. Overall, our findings highlight the untapped potential of MLLMs in instance-level image retrieval, being able to scale to databases with millions of images.

Acknowledgments

This work was supported by the Junior Star GACR (grant no. GM 21-28830M), Horizon MSCA-PF (grant no. 101154126), CTU in Prague (grant no. SGS23/173/OHK3/3T/13), and the Czech National Recovery Plan—CEDMO 2.0 NPO (MPO 60273/24/21300/21000) provided by the Ministry of Industry and Trade. We acknowledge VSB – Technical University of Ostrava, IT4Innovations National Supercomputing Center, Czech Republic, for awarding this project (OPEN-36-1) access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (grant ID: 90254).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, and Malcolm el al. Reynolds. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *CVPR*, 2025. 2, 4

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 1

[4] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. MiniGPT4-Video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. In *arXiv*, 2024. 1

[5] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. In *ICCV*, 2015. 2

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. In *arXiv*, 2025. 2, 5, 7

[7] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception Encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025. 7

[8] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020. 1, 5

[9] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024. 2

[10] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *CVPR*, 2024. 1

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICML*, 2021. 3

[13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. In *arXiv*, 2024. 1, 2

[14] Nikos Efthymiadis, Bill Psomas, Zakaria Laskar, Konstantinos Karantzalos, Yannis Avrithis, Ondřej Chum, and Giorgos Tolias. Composed image retrieval for training-free domain conversion. In *WACV*, 2025. 2

[15] Bohan Hou, Haoqiang Lin, Xuemeng Song, Haokun Wen, Meng Liu, Yupeng Hu, and Xiangyu Zhao. FiRE: Enhancing MLLMs with fine-grained context learning for complex image retrieval. In *SIGIR*, 2025. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[17] Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang, Zhe Lin, and Jiebo Luo. FINECAPTION: Compositional image captioning focusing on wherever you want at any granularity. In *CVPR*, 2025. 1

[18] Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. CoLLM: A large language model for composed image retrieval. In *CVPR*, 2025. 2

[19] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *PAMI*, 2011. 4

[20] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. In *arXiv*, 2024. 2

[21] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*, 2025. 2

[22] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. BRAVE: Broadening the visual encoding of vision-language models. In *ECCV*, 2024. 2

[23] Taehoon Kim, Pyunghwan Ahn, Sangyun Kim, Sihaeng Lee, Mark Marsden, Alessandra Sala, Seung Hwan Kim, Bohyung Han, Kyoung Mu Lee, and Honglak et al. Lee. NICE: CVPR 2023 challenge on zero-shot image captioning. In *CVPR*, 2024. 1

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1

[25] Giorgos Kordopatis-Zilos, Christos Tzelepis, Symeon Papadopoulos, Ioannis Kompatsiaris, and Ioannis Patras. DnS: Distill-and-select for efficient and accurate video indexing and retrieval. *IJCV*, 2022. 2

[26] Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Šuma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiří Matas, Ondřej Chum, and Giorgos Tolias. ILIAS: Instance-level image retrieval at scale. In *CVPR*, 2025. 5, 6, 7

[27] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *CVPR*, 2022. 2, 5

[28] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval and its memory footprint optimization. *PAMI*, 2024. 2

[29] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient visual projector for multimodal LLM. *IJCV*, 2025. 2

[30] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. In *ICLR*, 2025. 2

[31] Bangwei Liu, Yicheng Bao, Shaohui Lin, Xuhong Wang, Xin Tan, Yingchun Wang, Yuan Xie, and Chaochao Lu. IDMR: Towards instance-driven precise visual correspondence in multimodal retrieval. In *ICCVW*, 2025. 2

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1

[34] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LamRA: Large multimodal model as your advanced retrieval assistant. In *CVPR*, 2025. 2, 5, 7

[35] Jing Lu, Chaofan Xu, Wei Zhang, Ling-Yu Duan, and Tao Mei. Sampling wisely: Deep image embedding by top-k precision optimization. In *ICCV*, 2019. 2

[36] OpenAI. Gpt-4o system card. In *arXiv*, 2024. 1

[37] James Philbin, Ondřej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1

[38] Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *NeurIPS*, 2025. 2

[39] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 5

[40] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *PAMI*, 2019. 1, 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[42] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. DINO-X: A unified vision model for open-world object detection and understanding. In *arXiv*, 2024. 1

[43] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019. 2

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 5

[46] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *ICCV*, 2025. 2

[47] Hengcan Shi, Son Duy Dao, and Jianfei Cai. LLMFormer: Large language model for open-vocabulary semantic segmentation. *IJCV*, 2025. 1

[48] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. In *arXiv*, 2025. 5, 7

[49] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1

[50] Matteo Sodano, Federico Magistri, Lucas Nunes, Jens Behley, and Cyrill Stachniss. Open-world semantic segmentation including class similarity. In *CVPR*, 2024. 1

[51] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016. 2

[52] Pavel Suma, Giorgos Kordopatis-Zilos, Ahmet Iscen, and Giorgos Tolias. AMES: Asymmetric and memory-efficient similarity estimation for instance-level retrieval. In *ECCV*, 2024. 1, 2, 5, 7

[53] Pavel Suma, Giorgos Kordopatis-Zilos, Yannis Kalantidis, and Giorgos Tolias. Elvis: Efficient visual similarity from local descriptors that generalizes across domains. In *ICLR*, 2026. 6

[54] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *CVPR*, 2021. 1, 2, 5

[55] Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *CVPR*, 2025. 2

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. In *arXiv*, 2023. 1

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[58] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. In *arXiv*, 2024. 2

[59] Shuang Wang and Shuqiang Jiang. INSTRE: A new benchmark for instance-level object retrieval and recognition. *TOMM*, 2015. 5

[60] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. In *arXiv*, 2025. 2, 5

[61] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *ECCV*, 2024. 2

[62] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020. 1, 5

[63] Zilin Xiao, Pavel Suma, Ayush Sachdeva, Hao-Jen Wang, Giorgos Kordopatis-Zilos, Giorgos Tolias, and Vicente Ordonez. LOCORE: Image re-ranking with long-context sequence modeling. In *CVPR*, 2025. 2

[64] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. In *arXiv*, 2025. 5

[65] Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. DetCLIPv3: Towards versatile generative open-vocabulary object detection. In *CVPR*, 2024. 1

[66] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *ICCV*, 2021. 5

[67] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *CVPR*, 2025. 2

[68] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. In *arXiv*, 2025. 5

[69] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *CVPR*, 2023. 2