

Unbalanced Optimal Transport for Efficient Visual Document Retrieval

Hoyeon Shin^{1*}, Jeongyeon Kim^{1,2*}, Yeong Jun Koh³, Yeoneung Kim¹,
and Hanul Kim^{1**}

¹ Seoul National University of Science and Technology, Seoul, Korea
{hy_shin, jy_kim, yeoneung, hukim}@seoultech.ac.kr

² SIZL, Seongnam, Korea

³ Chungnam National University, Daejeon, Korea {yjkoh}@cnu.ac.kr

Abstract. Vision-Language Models (VLMs) have advanced Visual Document Retrieval (VDR) by enabling OCR-free, multi-vector representations. However, these models generate excessive tokens to preserve dense visual details, which severely limits retrieval scalability. Existing compression methods mitigate this overhead but rely on isolated heuristics that fail to balance token importance, semantic redundancy, and coverage capacity. We propose a visual token compression framework formulated as an Unbalanced Optimal Transport (UOT) problem. Our approach unifies these three objectives into a single alternating optimization process. Furthermore, we theoretically show that minimizing the transport cost upper-bounds the MaxSim retrieval error, ensuring ranking stability. Extensive experiments on the ViDoRe V1 and V3 benchmarks demonstrate that our method preserves fine-grained document semantics and improves the efficiency-effectiveness trade-off over state-of-the-art baselines. Code is available at <https://github.com/shhhhhyy/Unbalanced-Optimal-Transport-for-EVDR>.

Keywords: Visual Document Retrieval · Visual Token Compression · Unbalanced Optimal Transport

1 Introduction

Visual Document Retrieval (VDR) finds relevant document images for a given text query. Early VDR systems [13, 32] relied on optical character recognition (OCR) [37, 39] pipelines to feed parsed page information to standard text retrievers. However, these pipelines incur high indexing latency and struggle on visually complex pages featuring dense tables, figures, and irregular layouts [15]. Therefore, recent studies [15, 26] have shifted to OCR-free paradigms driven by Vision-Language Models (VLMs) [2, 24, 35, 43]. By mapping visual documents into a unified embedding space, VLM-based retrievers bypass OCR bottlenecks and inherently preserve structural semantics.

* Equal contribution.

** Corresponding author.

Within VLM-based retrievers, encoding visual documents requires balancing search speed and fine-grained accuracy. Compressing a page into a single global vector enables fast retrieval but often sacrifices localized visual details critical for complex queries. To address this limitation [26], multi-vector models like ColPali [15] represent each page as N patch-level visual tokens. They evaluate relevance through late interaction [19], using a MaxSim operation to match each query token with the most similar page token. However, retaining all visual tokens creates a severe scalability bottleneck. Because scoring M query tokens against N document tokens requires $\mathcal{O}(NM)$ operations, index size and query latency become too costly for large-scale corpora.

To mitigate this overhead, recent approaches [8, 27, 45] compress the N visual tokens into a compact set of K representatives. Standard techniques include pruning, which discards tokens based on attention scores, and merging, which aggregates similar patches through clustering. Despite such efficiency, these methods typically treat token importance and semantic redundancy as separate objectives. They also overlook the coverage capacity of the merged tokens. Unconstrained merging often collapses too many distinct patches into a single vector, degrading the fine-grained diversity needed for late interaction. Therefore, effective compression must balance three objectives: minimizing semantic redundancy, retaining informative tokens, and regulating coverage capacity.

In this paper, we propose a visual token compression framework for efficient VDR formulated as an Unbalanced Optimal Transport (UOT) [7, 16, 23] problem. Our formulation naturally addresses three main objectives: the transport cost minimizes semantic redundancy by clustering similar patches, the source marginals prioritize informative tokens, and the target marginals regulate coverage capacity to prevent representation collapse. By solving this UOT problem through an efficient alternating optimization process, our method achieves a highly compact yet diverse set of representative tokens. In addition, we analyze the theoretical properties of the UOT objective and prove that its minimization effectively upper-bounds the MaxSim retrieval error. Extensive experiments on VDR benchmarks [15, 25] validate the effectiveness of our approach, delivering higher retrieval performance at the same compute and comparable accuracy, and demonstrating a superior Pareto frontier over recent state-of-the-art methods [15, 27], as depicted in Fig. 1.

Our main contributions are summarized as follows:

- We present a VDR-specific compression principle that balances token importance, semantic redundancy, and coverage capacity.
- We propose a visual token compression framework formulated as a UOT problem, unifying these three objectives into an efficient alternating optimization process.
- We provide rigorous theoretical guarantees proving that minimizing the unbalanced transport cost upper-bounds the MaxSim retrieval error degradation, alongside asymptotic and complexity analyses.
- Extensive experiments on standard VDR benchmarks demonstrate that our method consistently improves the efficiency–effectiveness trade-off compared to state-of-the-art baselines.

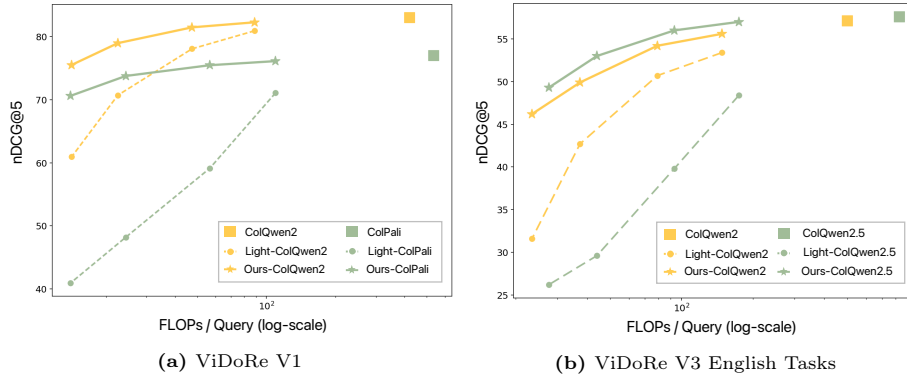


Fig. 1: Efficiency–effectiveness trade-off. nDCG@5 vs. FLOPs/query (log-scale) on ViDoRe benchmarks [15, 25]. FLOPs are estimated via a late-interaction proxy proportional to corpus size, expected query-document token interactions, and embedding dimension. Our method establishes a superior Pareto frontier compared to recent state-of-the-art baselines [15, 27].

2 Related Work

2.1 Visual Document Retrieval

Visual Document Retrieval (VDR) aims to retrieve relevant document images from large-scale corpora. Early VDR systems [13, 32] relied on OCR to convert images into text for standard retrieval. Although later layout-aware methods [11, 18, 44] integrated spatial cues, their continued reliance on the OCR pipeline caused high indexing latency and difficulties parsing complex structures like tables and diagrams. These constraints led to OCR-free frameworks that leverage Vision-Language Models (VLMs) to encode pages directly. For instance, DSE [26] proposed representing document images as single vectors to prioritize efficiency, though this approach often lacks the localized detail needed for precise matching. In contrast, ColPali [15] introduced a multi-vector late interaction [19] to preserve these details, which yields higher accuracy. However, the resulting visual token explosion presents a significant bottleneck for storage and retrieval speed.

2.2 Visual Token Compression

To mitigate visual token overhead, various token reduction techniques have been proposed. General-purpose methods exploit semantic redundancy in vision transformers via pruning and merging [36, 47], or saliency-driven selection [6, 50]. Beyond simple pruning heuristics, FitPrune [48] uses attention-distribution divergence for training-free pruning, while TopV [46] estimates token contribution via inference-time transport optimization. However, document images are notably dense, where task-critical information—such as table cells and captions—is often

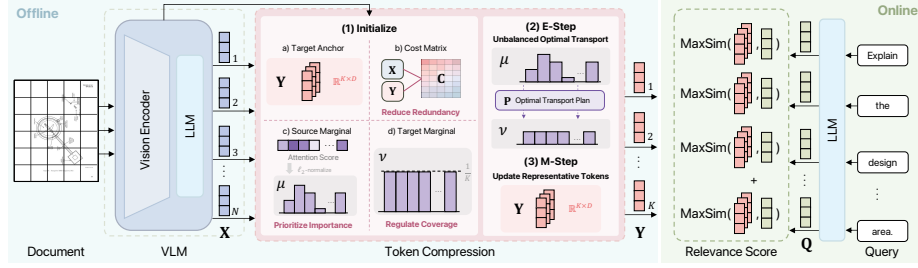


Fig. 2: Overview: Visual Token Compression. Offline, N visual tokens $\mathbf{X}$ are compressed into K anchors $\mathbf{Y}$ ($K \ll N$). Our Unbalanced Optimal Transport (UOT) formulation jointly optimizes semantic redundancy (cost $\mathbf{C}$), token importance (source μ), and anchor capacity (target ν) to enable efficient online MaxSim retrieval.

concentrated in small regions, making aggressive reduction risky. In the visual document domain, DocKylin [49] proposed efficient visual slimming at both the pixel and token levels to handle high-resolution, text-dense pages. To address this, VDR-specific efforts have explored clustering-based pooling [8], semantic clustering [27], and attention-based pruning [45]. Despite these advancements, existing approaches typically rely on a single heuristic—similarity or importance—and often fail to control merge imbalance. Such imbalance can lead to token over-collapse, ultimately degrading the coverage required for effective late interaction. Orthogonal to these methods, two-stage pipelines [20] balance efficiency by combining filtering with multi-vector reranking.

2.3 Optimal Transport in Computer Vision

Optimal Transport (OT) [33, 34, 41] provides a principled framework for aligning distributions by minimizing the cost of mass redistribution. In computer vision, it serves as a robust alternative to nearest-neighbor assignments by finding structured correspondences between feature sets, with applications ranging from local matching [38] to object detection [17]. Beyond matching, OT enables balanced clustering by enforcing marginal constraints that prevent samples from collapsing onto a few centers [5, 42]. This ensures that all prototypes are effectively utilized, maintaining the diversity required for compact representations. However, classical OT strictly enforces mass conservation, which can be undesirable in practical vision scenarios involving noisy features, redundant tokens, or partial correspondences. Unbalanced Optimal Transport (UOT) [7, 16, 23] relaxes these marginal constraints by introducing soft penalties, allowing mass to be partially discarded or redistributed. This relaxation improves robustness in cases where exact matching is neither necessary nor semantically meaningful [9, 21, 28].

3 Proposed Method

3.1 Problem Setup

VLM-based Visual Document Retrieval (VDR) frameworks [15] encode documents and queries into multi-vector representations. These models leverage the ViT architecture [14] to represent a document image as N visual tokens $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$, where each D -dimensional token vector is ℓ_2 -normalized. Similarly, a text query is encoded into M ℓ_2 -normalized tokens $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_M]^\top \in \mathbb{R}^{M \times D}$. Late interaction [19] then computes the relevance score via the MaxSim operation:

$$s(\mathbf{Q}, \mathbf{X}) = \sum_{i=1}^M \max_{1 \leq j \leq N} \langle \mathbf{q}_i, \mathbf{x}_j \rangle. \quad (1)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Due to the large number of tokens N necessary for high-resolution images, this matching requires $\mathcal{O}(NM)$ operations per query-document pair, imposing severe computational and storage bottlenecks. To this end, token merging [4, 27] compresses $\mathbf{X}$ into a compact set of K representatives $\mathbf{Y} \in \mathbb{R}^{K \times D}$ ($K \ll N$) while preserving retrieval performance. We interpret this compression task as aligning dense patch-level features with a smaller set of representative anchors under the late-interaction objective. This perspective motivates a transport-based formulation, where token assignments are modeled as a mass redistribution problem.

3.2 Preliminaries: Unbalanced Optimal Transport

Optimal Transport (OT) provides a principled framework for aligning discrete measures. Given source and target mass vectors $\boldsymbol{\mu} \in \mathbb{R}_+^N, \boldsymbol{\nu} \in \mathbb{R}_+^K$ and a cost matrix $\mathbf{C} \in \mathbb{R}^{N \times K}$, a transport plan $\mathbf{P} \in \mathbb{R}_+^{N \times K}$ maps mass from source to target elements to minimize the total cost. However, the strict mass conservation is often too restrictive when dealing with redundant or noisy data. To address this, Unbalanced Optimal Transport (UOT) [16, 23] relaxes the marginal constraints via soft penalties:

$$\min_{\mathbf{P} \geq 0} \langle \mathbf{P}, \mathbf{C} \rangle_F - \gamma H(\mathbf{P}) + \rho \text{KL}(\mathbf{P}\mathbf{1} \parallel \boldsymbol{\mu}) + \rho \text{KL}(\mathbf{P}^\top \mathbf{1} \parallel \boldsymbol{\nu}), \quad (2)$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product, $H(\mathbf{P})$ is the entropy regularization weighted by $\gamma > 0$, and $\rho > 0$ controls the degree of marginal relaxation. This formulation allows for flexible mass transport, which the generalized Sinkhorn algorithm [7] optimizes efficiently.

3.3 Token Compression via Unbalanced Optimal Transport

We formulate the compression of N visual tokens $\mathbf{X}$ into K representatives $\mathbf{Y}$ as a UOT-based mass redistribution problem. We address three objectives to

ensure high-quality retrieval: minimizing semantic redundancy, preserving token importance, and regulating coverage capacity. While existing methods aggregate similar patches or prune less informative ones, they often neglect the coverage capacity of individual representatives. Unconstrained merging can collapse excessive patches into a single token, degrading the fine-grained localized features necessary for the MaxSim operation in (1). Our framework unifies these objectives by mapping them to the transport cost and the relaxed marginal constraints, respectively.

Initialization Let $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_K]^\top \in \mathbb{R}^{K \times D}$ represent the K tokens to optimize. We initialize these anchors, denoted as $\mathbf{Y}^{(0)}$, via a k -means++ strategy [1]. We then apply ℓ_2 normalization to these sampled anchors to project them onto the unit hypersphere. This seeding captures the visual distribution and accelerates convergence. We then define the transport cost and marginal distributions as follows:

Cost Matrix. We penalize transport between dissimilar tokens to reduce semantic redundancy. Each entry c_{ij} of $\mathbf{C} \in \mathbb{R}^{N \times K}$ is the cosine distance between the source token $\mathbf{x}_i$ and target anchor $\mathbf{y}_j$. Since these are ℓ_2 -normalized for all i, j , we have:

$$c_{ij} = 1 - \langle \mathbf{x}_i, \mathbf{y}_j \rangle. \quad (3)$$

Minimizing this cost encourages semantically redundant patches to collapse into the same representative token.

Source Marginal. To prioritize salient regions, we use the VLM’s self-attention weights as an importance proxy, since they naturally emphasize informative regions [12, 15]. Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be the attention matrix averaged over all transformer layers. The importance score $\tilde{a}_i$ is defined as the total incoming attention, $\tilde{a}_i = \sum_{n=1}^N a_{ni}$, where a_{ni} denotes the element at the n -th row and i -th column. The i -th element μ_i of the source marginal distribution $\boldsymbol{\mu}$ is then given by:

$$\mu_i = \frac{\tilde{a}_i + \epsilon}{\sum_{n=1}^N (\tilde{a}_n + \epsilon)}, \quad (4)$$

where ϵ is a small constant for numerical stability.

Target Marginal. To regulate the coverage of representative tokens, we define the target marginal $\boldsymbol{\nu} \in \mathbb{R}_+^K$ as a uniform distribution, where each element ν_j is given by:

$$\nu_j = \frac{1}{K} \quad (5)$$

This establishes an equal capacity for each representative token, encouraging an equitable assignment of source mass. The KL penalty in (2) relaxes strict capacity, allowing the plan to balance the uniform capacity prior with semantic density of document regions, maintaining localized features while avoiding rigid fixed marginals.

Algorithm 1 UOT-based Token Compression (PyTorch-style)

```

# X: Visual tokens (N, D)
# A: Attention matrix (N, N)
# K: Number of representative tokens
# T: Number of EM loop iterations

# 1. Initialize Marginal
a_tilde = A.sum(dim=0)
mu = l1_normalize(a_tilde, dim=0)
nu = l1_normalize(torch.ones(K), dim=0)

# 2. Initialize Anchors
Y = init_kmeans++(X, K)
Y = l2_normalize(Y, dim=-1)

# 3. Alternating Optimization (EM Loop)
for _ in range(T):
    # E-Step: Solve Unbalanced Optimal Transport
    C = 1.0 - (X @ Y.T)
    P = generalized_sinkhorn(C, mu, nu)

    # M-Step: Update Representative Tokens
    Y = P.T @ X
    Y = l2_normalize(Y, dim=-1)

```

Algorithm 2 Generalized Sinkhorn Algorithm (PyTorch-style)

```

# C: Cost matrix (N, K)
# mu: Source marginal (N,), nu: Target marginal (K,)
# gamma: Entropy weight, rho: KL penalty weight
# L: Number of Sinkhorn iterations

log_mu, log_nu = mu.log(), nu.log()
log_u, log_v = torch.zeros_like(mu), torch.zeros_like(nu)

log_K = -C / gamma
fi = rho / (rho + gamma)

for _ in range(L):
    # Row Update
    t1 = log_K + log_v.unsqueeze(0)
    log_u = fi * (log_mu - torch.logsumexp(t1, dim=1))

    # Column Update
    t2 = log_K + log_u.unsqueeze(1)
    log_v = fi * (log_nu - torch.logsumexp(t2, dim=0))

return torch.exp(log_u.unsqueeze(1) + log_K + log_v.unsqueeze(0))

```

Alternating Optimization Since the representative tokens $\mathbf{Y}$ and their optimal assignments $\mathbf{P}$ are interdependent, we jointly solve for them using an Expectation-Maximization (EM)-style alternating optimization framework [10]. The overall token compression procedure is summarized in Alg. 1.

E-step. At iteration t , we fix the current target anchors $\mathbf{Y}^{(t)}$ and compute the cost matrix $\mathbf{C}^{(t)}$ using (3). Given the predefined marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we solve the UOT objective in (2) via the generalized Sinkhorn algorithm [7]. To ensure numerical stability, we perform this optimization in the log domain, as detailed

in Alg. 2. This yields the optimal transport plan $\mathbf{P}^{(t)} \in \mathbb{R}_+^{N \times K}$. This soft assignment matrix balances three objectives: minimizing semantic distance, respecting attention-based source token importance, and adhering to the relaxed capacity limits of each cluster.

M-step. Next, we fix the transport plan $\mathbf{P}^{(t)}$ and update the target anchors to minimize the expected transport cost. Given the cosine distance cost matrix, we update each anchor j using a mass-weighted sum and ℓ_2 normalization:

$$\mathbf{y}_j^{(t+1)} = \frac{\sum_{i=1}^N p_{ij}^{(t)} \mathbf{x}_i}{\left\| \sum_{i=1}^N p_{ij}^{(t)} \mathbf{x}_i \right\|_2}, \quad \forall j \in \{1, \dots, K\}, \quad (6)$$

where $p_{ij}^{(t)}$ is the (i, j) -th entry of $\mathbf{P}^{(t)}$. The transport plan $\mathbf{P}^{(t)}$ encodes both the attention-based source mass $\boldsymbol{\mu}$ and the relaxed uniform target capacity $\boldsymbol{\nu}$. Consequently, the updated tokens $\mathbf{Y}$ achieve a critical balance: the source mass ensures that informative tokens dominate the representations, while the regulated capacity prevents excessive aggregation into any single anchor. This dual effect preserves the localized feature diversity required for late interaction and effectively suppresses redundant tokens.

4 Theoretical Analysis

We theoretically analyze our UOT-based compression to ensure it preserves retrieval performance. Specifically, we bound the MaxSim approximation error, prove geometric robustness against worst-case queries, and demonstrate that these token-level bounds strictly preserve system-level document rankings. Detailed proofs for all theorems are provided in the supplementary material.

4.1 MaxSim Approximation Error Bound

We define the retrieval error as the difference between the original MaxSim score $s(\mathbf{Q}, \mathbf{X})$ and the compressed score $s(\mathbf{Q}, \mathbf{Y})$ defined in (1). The following theorem establishes an upper bound on this error using a transport plan $\mathbf{P}$.

Theorem 1 (MaxSim Approximation Error Bound). *Assume all token vectors in $\mathbf{Q}, \mathbf{X}, \mathbf{Y}$ are ℓ_2 -normalized. Let $\mathbf{C} \in \mathbb{R}^{N \times K}$ be the cosine-distance cost matrix where $c_{ij} = 1 - \langle \mathbf{x}_i, \mathbf{y}_j \rangle$. For a transport plan $\mathbf{P} \in \mathbb{R}_+^{N \times K}$, let $i_m^* = \arg \max_{1 \leq i \leq N} \langle \mathbf{q}_m, \mathbf{x}_i \rangle$ be the best-matching original token for query token $\mathbf{q}_m$. If the row mass $r_m = \sum_{j=1}^K p_{i_m^* j}$ is positive, the MaxSim degradation is bounded by:*

$$s(\mathbf{Q}, \mathbf{X}) - s(\mathbf{Q}, \mathbf{Y}) \leq \sqrt{2} \sum_{m=1}^M \sqrt{\frac{1}{r_m} \sum_{j=1}^K p_{i_m^* j} c_{i_m^* j}}. \quad (7)$$

Remark. This bound justifies our formulation. The term $\sum_j p_{i_m^* j} c_{i_m^* j}$ measures the transport cost for tokens that dominate the MaxSim score. Minimizing $\langle \mathbf{P}, \mathbf{C} \rangle$ therefore reduces the approximation error in informative regions. Additionally, our importance-weighted marginal $\boldsymbol{\mu}$ assigns more mass to these likely maximizers, which further tightens the bound.

4.2 Worst-Case Query Robustness

We analyze the robustness of our framework against any possible unit-norm query. Let $\Delta(\mathbf{X}, \mathbf{Y})$ denote the worst-case MaxSim gap for a single query token $\mathbf{q}$. The following theorem connects this retrieval gap to the geometric covering radius between the original tokens and their representatives.

Theorem 2 (Worst-case Query Robustness). *Assume all token vectors are ℓ_2 -normalized, and let $c_{ij} = 1 - \langle \mathbf{x}_i, \mathbf{y}_j \rangle$ be the cosine cost. The worst-case MaxSim gap is bounded by the geometric covering radius:*

$$\Delta(\mathbf{X}, \mathbf{Y}) := \sup_{\|\mathbf{q}\|_2=1} \left(\max_i \langle \mathbf{q}, \mathbf{x}_i \rangle - \max_j \langle \mathbf{q}, \mathbf{y}_j \rangle \right) \leq \max_i \min_j \|\mathbf{x}_i - \mathbf{y}_j\|_2. \quad (8)$$

Furthermore, for any transport plan $\mathbf{P} \in \mathbb{R}_+^{N \times K}$ with row mass $r_i = \sum_{j=1}^K p_{ij} > 0$, this gap is upper-bounded by the transport cost:

$$\Delta(\mathbf{X}, \mathbf{Y}) \leq \sqrt{2} \max_i \sqrt{\frac{1}{r_i} \sum_{j=1}^K p_{ij} c_{ij}}. \quad (9)$$

Remark. This bound links the worst-case retrieval error to the geometric covering radius. While unconstrained compression may discard minor visual details and expand this radius, Theorem 2 proves that minimizing the UOT cost strictly bounds it. By ensuring every original token $\mathbf{x}_i$ remains geometrically close to a representative, our method prevents search blind spots, preserving fine-grained features necessary for robust long-tail retrieval.

4.3 Ranking Stability under Compression

We show that bounding the per-document MaxSim degradation preserves retrieval rankings. For a query $\mathbf{Q}$ and a document collection $\mathcal{D}$, let s_d and $\tilde{s}_d$ denote the original and compressed scores for document d , respectively. Assume a deviation bound $|s_d - \tilde{s}_d| \leq \varepsilon_d(\mathbf{Q})$. As established, the UOT transport cost directly governs this bound $\varepsilon_d(\mathbf{Q})$.

Theorem 3 (Ranking and Top- k Stability). *Assume the deviation bound $|s_d - \tilde{s}_d| \leq \varepsilon_d(\mathbf{Q})$ holds.*

1. **Pairwise Stability:** *For any two documents a and b , if their original score gap satisfies $s_a - s_b > \varepsilon_a(\mathbf{Q}) + \varepsilon_b(\mathbf{Q})$, their relative ranking is preserved after compression ($\tilde{s}_a > \tilde{s}_b$).*

2. **Top-k Stability:** Let $\mathcal{T}_k$ be the set of the top- k documents under the original scores. Define the top- k gap as $\Delta_k(\mathbf{Q}) = s_{(k)} - s_{(k+1)}$, where $s_{(k)}$ is the k -th highest original score. If $\Delta_k(\mathbf{Q}) > \max_{d \in \mathcal{T}_k} \varepsilon_d(\mathbf{Q}) + \max_{d \notin \mathcal{T}_k} \varepsilon_d(\mathbf{Q})$, the top- k set $\mathcal{T}_k$ is exactly preserved under the compressed scores.

Remark. Theorem 3 bridges our token-level transport bounds with the system-level retrieval objective. Specifically, realizing the deviation $\varepsilon_d(\mathbf{Q})$ via Theorem 1 yields an explicit sufficient condition for ranking preservation based entirely on transport costs. Therefore, minimizing the UOT cost maximizes the theoretical guarantees of ranking stability.

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate our framework on ViDoRe V1 [15] and V3 [25], standard benchmarks for visually rich document retrieval. ViDoRe V1 contains multiple subtasks covering diverse document types, languages, and real-world scenarios. We report results on the standard academic and practical tasks. ViDoRe V3 contains domain-specific datasets with graded relevance annotations, allowing multiple relevant pages per query instead of V1’s single-positive setup.

ViDoRe V1 Academic Tasks. This split adapts VQA benchmarks by treating questions as queries and pages as documents. InfoVQA (InfoQ) [30] features dense infographics, while arXivQA (ArxivQ) [22] focuses on scientific figures and trends. TabFQuAD (TabF) provides French table queries, and TAT-DQA (TATQ) [51] mixes tables with text. DocVQA (DocQ) [31] covers real-world free-form documents (*e.g.*, receipts) requiring key-field localization.

ViDoRe V1 Practical Tasks. This split is designed to evaluate retrieval in realistic settings. It includes five domain-specific subsets: Energy, Government (Gov.), Healthcare (Health.), Artificial Intelligence (AI.), and French environmental reports (Shift.). Each subset contains 1,000 document pages and 100 queries. The topic corpora are intentionally narrow to increase similarity among candidate pages, making the ranking problem more challenging and representative of real-world retrieval scenarios.

ViDoRe V3 English-Only Tasks. This benchmark contains 26K pages and 3.1K queries translated into 6 languages. We evaluate on ViDoRe V3 using English queries only, covering five domain-specific subsets: Computer Science, Finance, HR report, Industrial, and Pharmaceuticals. These splits contain 13K document pages and 1.5K queries in a domain-focused retrieval setting. The benchmark includes diverse query types ranging from simple fact-seeking questions to queries requiring structural or semantic understanding of documents. Due to high similarity among candidate pages and the increased query complexity, the retrieval

task becomes more challenging. To keep the main evaluation focused on domain-specific retrieval, we use English queries in the main experiments and report results on the remaining language queries in the supplementary material.

Evaluation Metrics. We evaluate retrieval performance using nDCG@5 and Recall@1. Specifically, nDCG@5 measures ranked relevance within the top-5 results, and Recall@1 evaluates whether the ground-truth page is ranked first. Because ViDoRe V1 uses binary single-relevance labels while V3 provides multiple graded labels, we ensure consistent evaluation by adopting the capped Recall@ k from BEIR [40]. This metric computes the fraction of relevant documents retrieved within the top- k , capped by the total relevant documents per query. For efficiency, we report the merging factor (MF), the ratio of original visual tokens N to representative tokens K , which directly reflects memory and scoring cost reductions. Also, we measure average query latency as the mean time to compute query-document relevance scores.

Implementation Details We evaluate our framework using Qwen2-VL-2B (Qwen2) [43] and PaliGemma-3B (Pali) [3] for ViDoRe V1, and Qwen2/Qwen2.5-VL-3B (Qwen2.5) for ViDoRe V3. For a fair comparison, all compression methods are evaluated under the merging factors (MF) $\in \{5, 10, 25, 50\}$. In Algs. 1 and 2, we use $T = 20$ EM iterations, $L = 3$ Sinkhorn iterations, and $\rho = 5$ for all models. The entropic regularization γ is set to 0.04 for Qwen2/Qwen2.5 and 0.01 for Pali. All experiments are conducted on a single NVIDIA A6000 GPU.

5.2 Main Results

Quantitative Comparison. We set the uncompressed multi-vector retrievers [15] (ColQwen2, ColPali, ColQwen2.5) as upper bounds. We compare our method against two training-free compression baselines: (i) attention-based pruning [15] (e.g., ColQwen2+Prune), which discards low-attention tokens, and (ii) distance-based clustering [27] (e.g., Light-ColQwen2), which merges redundant tokens via hierarchical clustering. For a fair comparison, all methods utilize the same frozen VLM backbones and operate under identical token budgets (MF $\in \{5, 10, 25, 50\}$).

Tables 1 to 3 summarize the results across the ViDoRe V1 and V3 benchmarks. While all methods perform comparably under lenient budgets (MF=5), heuristic baselines degrade severely at extreme compression (MF=50), as unconstrained clustering forcibly merges disparate visual elements and induces representation collapse. In contrast, our framework remains highly robust by explicitly regulating token capacity and adaptively redistributing mass to preserve fine-grained localized features. Consequently, our method achieves substantial gains over Light-ColQwen baselines at MF=50, outperforming them by +19.4 nDCG@5 on the V1 Practical tasks (Qwen2) and +23.1 nDCG@5 on the challenging V3 english tasks (Qwen2.5). These consistent trends across multiple VLM backbones confirm that principled capacity regulation is essential for complex retrieval scenarios.

Table 1: Results on ViDoRe V1 academic tasks. nDCG@5/Recall@1 ($\uparrow$) and Time (ms/q). Best compression results per MF are bolded.

Method	MF	Time	ArxivQ	DocQ	InfoQ	TabF	TATQ	Avg.
<i>Base model: Qwen2-VL-2B</i>								
ColQwen2	-	0.63	86.8/80.4	56.0/48.0	90.6/86.2	88.7/81.4	78.5/66.0	80.1/72.4
ColQwen2+Prune	5	0.16	83.6/76.8	54.7/46.4	86.7/80.2	85.8/78.2	71.9/59.2	76.5/68.2
Light-ColQwen2			85.8/79.0	54.5/46.2	87.4/81.6	87.3/79.6	75.3/61.8	78.1/69.6
Ours			86.3/79.6	56.4/48.4	89.0/84.2	89.5/83.2	77.0/64.0	79.6/71.9
ColQwen2+Prune	10	0.11	80.8/74.6	51.1/42.4	80.2/72.2	82.3/73.6	64.9/51.9	71.9/62.9
Light-ColQwen2			84.9/78.6	53.8/ 45.6	82.4/74.2	87.3/79.3	70.6/57.4	75.8/67.0
Ours			85.8/79.2	54.1/44.6	88.5/83.2	90.0/84.6	74.6/61.3	78.6/70.6
ColQwen2+Prune	25	0.08	75.2/68.6	42.2/34.2	72.3/62.2	76.8/67.9	53.1/40.4	63.9/54.7
Light-ColQwen2			81.8/75.6	48.7/39.4	72.9/63.0	86.5/78.6	56.5/45.1	69.3/60.3
Ours			83.5/76.6	50.8/41.4	85.8/79.2	87.7/81.4	69.0/55.5	75.4/66.8
ColQwen2+Prune	50	0.07	66.3/58.6	35.8/28.6	65.7/54.4	65.0/54.3	43.1/31.9	55.2/45.6
Light-ColQwen2			77.4/69.2	39.4/30.4	65.8/55.8	81.1/71.8	44.5/34.4	61.6/52.3
Ours			82.2/75.4	45.2/36.4	81.2/73.6	86.8/79.6	60.1/47.1	71.3/62.4
<i>Base model: PaliGemma-3B</i>								
ColPali	-	0.80	77.6/71.0	49.6/41.6	79.1/72.2	85.3/77.9	65.2/52.8	71.4/63.1
ColPali+Prune	5	0.20	76.0/70.0	47.9/39.2	76.1/68.2	85.0/77.9	61.3/48.6	69.3/60.8
Light-ColPali			75.8/ 70.2	42.3/34.8	76.5/68.8	83.9/76.1	57.4/43.7	67.2/58.7
Ours			76.4/70.2	49.2/41.6	77.9/71.6	80.5/72.5	63.9/51.9	69.6/61.6
ColPali+Prune	10	0.13	72.9/66.4	45.0/36.6	72.8/63.4	82.2/72.1	54.0/41.7	65.4/56.0
Light-ColPali			72.9/66.4	35.0/27.0	59.3/51.2	80.8/71.4	41.1/28.7	57.8/48.9
Ours			75.7/69.8	48.1/40.2	77.4/70.8	81.3/74.6	62.7/49.5	69.0/61.0
ColPali+Prune	25	0.08	67.4/59.4	37.4/29.6	64.6/56.4	71.4/60.7	43.7/31.8	56.9/47.6
Light-ColPali			63.1/53.8	25.2/19.2	58.6/47.8	71.4/60.7	27.3/17.3	49.1/39.8
Ours			75.0/68.2	45.4/39.0	76.3/69.6	78.7/71.1	58.0/45.3	66.7/58.6
ColPali+Prune	50	0.07	60.1/51.0	29.3/23.0	57.7/49.6	60.9/49.6	33.0/24.1	48.2/39.5
Light-ColPali			51.0/43.4	21.8/16.8	45.1/35.4	62.9/51.1	21.0/12.2	40.4/31.8
Ours			73.8/67.4	42.8/37.2	73.6/65.6	76.5/68.6	51.9/38.4	63.7/55.4

5.3 Experimental Analysis

We analyze our framework’s core mechanisms using Qwen2 on ViDoRe V1. An ablation study validates our three design objectives by comparing heuristic base-lines against optimal transport. Qualitative visualizations then show how our approach preserves geometric density and prevents representation collapse. Additional results are in the supplement.

Ablation Study. We perform this ablation study under $MF \in \{25, 50\}$, where performance disparities are most pronounced. As noted in Sec. 3.3, effective compression requires balancing three objectives: minimizing token redundancy, preserving token importance, and regulating coverage capacity. We incrementally integrate them into a Hierarchical Agglomerative Clustering (HAC) baseline. We heuristically model token redundancy as centroid Euclidean distance, token importance as average attention, and coverage capacity as the total attention sum per cluster. The supplementary material details these criteria.

Table 4 shows that relying solely on feature redundancy (HAC: Dist) yields the lowest performance (57.9 nDCG@5 at $MF=50$). Incrementally integrating attention-based importance (HAC: Dist+Attn) and bounding coverage capacity (HAC: Dist+Attn+Cap.) prevents excessive merging, progressively boosting scores to 62.6 and 65.4. This confirms all three objectives are crucial for maintaining the representation diversity required by MaxSim. While these heuristics

Table 2: Results on ViDoRe V1 practical tasks. nDCG@5/Recall@1 ($\uparrow$) and Time (ms/q). Best compression results per MF are bolded.

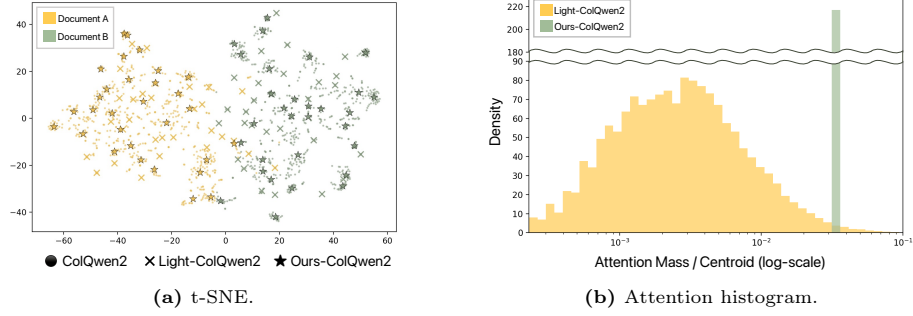
Method	MF	Time	Shift.	AI.	Energy.	Gov.	Health.	Avg.
<i>Base model: Qwen2-VL-2B</i>								
ColQwen2	-	1.06	87.5/77.0	86.7/67.0	85.6/67.0	84.3/63.0	85.2/64.0	85.9/67.6
ColQwen2+Prune Light-ColQwen2 Ours	5	0.27	78.1/60.0 81.7/67.0 86.9/74.0	85.2/64.0 85.6/65.0 84.7/62.0	85.7/69.0 83.3/64.0 83.2/62.0	85.9/67.0 82.7/61.0 85.6/67.0	82.7/60.0 85.2/63.0 84.2/61.0	83.5/64.0 83.7/64.0 84.9/65.2
ColQwen2+Prune Light-ColQwen2 Ours	10	0.19	70.0/49.0 75.1/58.0 83.5/66.0	82.7/61.0 83.4/62.0 84.5/63.0	82.2/ 66.0 79.1/59.0 83.7/64.0	84.8/65.0 80.5/57.0 86.1/67.0	81.9/58.0 83.5/62.0 83.9/62.0	80.3/59.8 80.3/59.6 84.3/64.4
ColQwen2+Prune Light-ColQwen2 Ours	25	0.15	54.9/37.0 58.1/42.0 81.5/70.0	73.7/51.0 76.7/57.0 83.4/63.0	73.0/51.0 73.7/50.0 83.3/68.0	75.8/55.0 78.1/54.0 82.7/61.0	79.4/54.0 74.1/52.0 81.6/58.0	71.4/49.6 72.1/51.0 82.5/64.0
ColQwen2+Prune Light-ColQwen2 Ours	50	0.13	41.4/29.0 38.0/23.0 72.2/57.0	66.3/46.0 64.6/42.0 84.0/66.0	65.1/48.0 64.0/43.0 77.9/58.0	70.2/49.0 69.2/43.0 82.4/62.0	74.9/50.0 65.5/44.0 82.2/60.0	63.6/44.4 60.3/39.0 79.7/60.6
<i>Base model: PaliGemma-3B</i>								
ColPali	-	0.80	74.6/58.0	85.2/66.0	85.8/69.0	85.9/67.0	81.6/57.0	82.6/63.4
ColPali+Prune Light-ColPali Ours	5	0.32	68.5/54.0 58.6/42.0 77.0/62.0	81.6/ 61.0 83.2/60.0 83.1/ 61.0	84.2/64.0 81.1/62.0 85.3/70.0	84.5/64.0 76.3/54.0 86.2/66.0	80.3/58.0 75.1/50.0 81.4/59.0	79.8/60.2 74.9/53.6 82.6/63.6
ColPali+Prune Light-ColPali Ours	10	0.22	59.1/50.0 37.4/27.0 75.2/60.0	80.6/58.0 66.4/42.0 84.7/67.0	77.5/56.0 67.8/46.0 83.2/63.0	78.6/57.0 63.9/41.0 84.1/63.0	78.7/56.0 66.3/43.0 82.2/61.0	74.9/55.4 60.4/39.8 81.9/63.6
ColPali+Prune Light-ColPali Ours	25	0.15	46.4/37.0 25.1/15.0 69.8/57.0	73.3/51.0 54.8/33.0 82.7/62.0	63.5/46.0 51.1/31.0 85.2/71.0	69.9/49.0 51.6/33.0 84.1/63.0	68.3/43.0 53.8/29.0 82.3/63.0	64.3/45.2 47.3/28.2 80.8/63.2
ColPali+Prune Light-ColPali Ours	50	0.14	32.9/24.0 21.1/10.0 62.8/44.0	62.2/39.0 50.0/25.0 79.9/59.0	51.8/38.0 46.1/27.0 81.1/63.0	55.0/36.0 40.4/24.0 83.7/64.0	56.2/35.0 50.0/30.0 79.9/59.0	51.6/34.4 41.5/23.2 77.5/57.8

Table 3: Results on ViDoRe V3 english-only tasks. nDCG@5/Recall@1 ($\uparrow$) and Time (ms/q). Best compression results per MF are bolded.

Method	MF	Time	C.S.	Fin.	H.R.	Ind.	Phar.	Avg.
<i>Base model: Qwen2-VL-2B</i>								
ColQwen2	-	2.50	67.7/75.0	54.5/54.5	55.4/55.4	49.3/53.2	58.8/59.1	57.1/59.4
ColQwen2+Prune Light-ColQwen2 Ours	5	0.63	67.7/75.8 67.3/ 76.4 67.5/75.5	52.7/53.7 48.8/50.0 52.3/ 55.2	51.5/52.2 48.8/47.5 54.0/55.0	44.7/45.9 45.9/48.6 46.7/48.2	55.9/56.6 56.4/55.8 57.6/58.2	54.5/56.8 53.4/55.7 55.6/58.4
ColQwen2+Prune Light-ColQwen2 Ours	10	0.41	62.5/67.9 64.8/73.6 67.2/76.4	46.3/48.5 46.3/48.7 48.8/51.0	48.1/ 51.9 45.4/45.9 53.1/50.9	37.7/38.9 41.8/42.6 45.9/48.6	52.1/50.8 55.4/ 54.9 56.0/54.7	49.3/51.6 50.7/53.1 54.2/56.3
ColQwen2+Prune Light-ColQwen2 Ours	25	0.29	56.7/60.5 58.7/66.2 64.6/71.8	35.8/36.9 34.8/38.4 43.5/45.8	39.5/38.4 35.0/36.8 48.5/47.2	32.0/31.1 35.4/35.6 40.0/42.6	43.1/43.1 49.8/48.9 53.0/52.2	41.4/42.0 42.7/45.2 49.9/51.9
ColQwen2+Prune Light-ColQwen2 Ours	50	0.26	51.0/57.7 49.0/56.5 60.9/69.0	27.7/29.4 19.2/21.0 38.9/41.0	32.6/33.3 24.2/25.5 45.8/47.5	27.4/27.2 27.2/25.7 35.0/35.6	35.7/34.3 38.4/38.2 50.5/51.4	34.9/36.4 31.6/33.4 46.2/48.9
<i>Base model: Qwen2.5-VL-3B</i>								
ColQwen2.5	-	2.39	69.8/75.5	52.7/56.8	55.8/56.9	48.9/51.1	60.6/61.8	57.6/60.4
ColQwen2.5+Prune Light-ColQwen2.5 Ours	5	0.60	66.0/71.6 63.5/69.9 69.3/77.3	48.7/54.4 43.5/44.8 52.2/57.4	51.9/52.8 44.5/46.2 55.7/55.3	41.8/42.4 42.0/45.4 48.0/50.4	57.0/58.8 56.3/56.0 59.7/60.4	53.1/56.0 48.4/51.6 57.0/60.2
ColQwen2.5+Prune Light-ColQwen2.5 Ours	10	0.41	62.8/68.4 50.5/55.1 69.7/74.1	39.5/44.0 32.3/34.5 50.8/56.1	45.8/46.9 31.4/29.9 53.5/53.1	37.4/36.0 32.7/28.9 45.8/47.5	53.4/56.0 51.9/52.5 60.1/62.9	47.8/50.3 39.8/40.2 56.0/58.7
ColQwen2.5+Prune Light-ColQwen2.5 Ours	25	0.30	55.4/62.8 44.6/48.1 69.4/78.2	34.2/35.6 23.2/21.0 46.1/50.3	36.1/36.5 20.1/17.3 51.2/53.8	29.3/27.9 21.6/21.1 41.5/37.0	46.2/46.2 38.6/36.8 56.7/56.9	40.2/41.8 29.6/28.9 53.0/55.2
ColQwen2.5+Prune Light-ColQwen2.5 Ours	50	0.26	51.4/60.0 45.2/50.5 66.4/73.1	24.9/23.9 19.6/18.4 40.6/45.2	26.1/28.6 21.6/22.0 47.3/50.3	19.0/18.0 13.6/11.6 37.8/36.6	38.3/37.4 31.0/25.5 54.3/55.2	31.9/33.6 26.2/25.6 49.3/52.1

Table 4: Ablation study on ViDoRe V1 academic tasks. nDCG@5/Recall@1 ($\uparrow$) and Time (ms/q). Best compression results per MF are bolded.

Method	MF	ArxivQ	DocQ	InfoQ	TabF	TATQ	Avg.
(1) HAC: Dist	25	82.8/76.4	47.4/38.4	72.9/63.0	86.5/78.2	52.2/38.7	69.3/60.3
(2) HAC: Dist+Attn		82.9/75.4	48.7/39.6	80.5/72.6	87.2/79.6	58.3/44.6	71.5/62.4
(3) HAC: Dist+Attn+Cap.		83.4/75.8	48.3/38.6	81.2/73.4	88.5/82.1	61.3/47.4	72.5/63.5
(4) Balanced OT		84.2/77.2	50.4/41.4	84.9/77.6	88.1/81.4	67.8/54.7	75.1/66.5
(5) Unbalanced OT (Ours)		83.5/76.6	50.8/41.4	85.8/79.2	87.7/81.4	69.0/55.5	75.4/66.8
(1) HAC: Dist	50	75.9/67.4	37.8/30.2	65.8/55.8	81.3/72.5	35.7/25.2	61.6/52.3
(2) HAC: Dist+Attn		78.3/71.4	38.5/29.2	72.3/62.2	79.6/68.6	44.5/32.2	62.6/52.7
(3) HAC: Dist+Attn+Cap.		79.9/73.0	41.7/33.0	76.4/67.8	82.8/73.9	46.4/32.9	65.4/56.1
(4) Balanced OT		80.7/73.0	44.9/35.6	80.9/73.8	87.2/80.7	61.2/47.9	71.0/62.1
(5) Unbalanced OT (Ours)		82.2/75.4	45.2/36.4	81.2/73.6	86.8/79.6	60.1/47.1	71.3/62.4

**Fig. 3: Qualitative analysis (MF=25).** (a) t-SNE: embedding regions of original tokens ($\bullet$), Light-ColQwen2 ($\times$), and Ours ($\star$). (b) Attention histogram: distribution of mass per representative token; green bar indicates the uniform share (≈ 0.033).

validate our design choices, their greedy nature remains sub-optimal. Unifying these objectives into an Optimal Transport (OT) framework theoretically bounds MaxSim degradation, improving performance. Our Unbalanced OT formulation relaxes target marginals; allowing representative tokens to adapt their capacities to local semantic density minimizes transport cost further, achieving the highest overall performance.

Geometric Preservation in Embedding Space. Fig. 3a visualizes token embeddings via t-SNE [29] at MF=25. While original ColQwen2 tokens ($\bullet$) form dense semantic regions, Light-ColQwen2 representatives ($\times$) often deviate into low-density areas because unconstrained clustering skews large cluster centroids and overfits small ones. In contrast, our UOT anchors ($\star$) remain within the original high-density regions. This aligns with Theorem 2’s geometric covering guarantee: penalizing the transport cost restricts representatives near their source tokens, thereby minimizing semantic distortion and preserving accurate MaxSim retrieval.

Attention Mass Distribution. Fig. 3b shows the density histogram of attention mass per representative token (log-scale). Light-ColQwen2 yields a highly

skewed distribution where a few centroids dominate, indicating representation collapse from unconstrained merging. Conversely, our capacity regulation distributes mass more uniformly, successfully concentrating it near the expected uniform share (≈ 0.033 for $K = 30$ tokens at MF=25). This ensures all compressed tokens contribute effectively to the semantic coverage, minimizing redundancy and dead capacity.

6 Conclusion

This work addressed the scalability bottleneck in VLM-based Visual Document Retrieval caused by the token explosion in late-interaction models. We demonstrated that effective token compression requires the principled integration of semantic redundancy, token importance, and coverage capacity. By reformulating this task as an Unbalanced Optimal Transport (UOT) problem, we proposed a unifying framework that optimizes these objectives within a single formulation. Supported by theoretical guarantees that bound MaxSim approximation errors, our approach establishes a superior efficiency–effectiveness trade-off on the ViDoRe V1 and V3 benchmarks, consistently outperforming state-of-the-art baselines across diverse VLM architectures.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (Grant No. RS-2024-00352566), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2026-25510490), and Seoul National University of Science and Technology.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: SODA (2007)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint (2025)
3. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint (2024)
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: ICLR (2023)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
6. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: ECCV. pp. 19–35. Springer (2024)
7. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.* **87**(314), 2563–2609 (2018)

8. Clavié, B., Chaffin, A., Adams, G.: Reducing the footprint of multi-vector retrieval with minimal performance impact via token pooling. arXiv preprint (2024)
9. De Plaen, H., De Plaen, P.F., Suykens, J.A., Proesmans, M., Tuytelaars, T., Van Gool, L.: Unbalanced optimal transport: A unified framework for object detection. In: CVPR (2023)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B* **39**(1), 1–22 (1977)
11. Denk, T.I., Reisswig, C.: Bertgrid: Contextualized embedding for 2d document representation and understanding. arXiv preprint (2019)
12. Djilali, Y.A.D., McGuinness, K., O’Connor, N.E.: Vision transformers are inherently saliency learners. In: BMVC (2023)
13. Doermann, D.S.: The indexing and retrieval of document images: A survey. *CVIU* **70**(3), 287–298 (1998). <https://doi.org/10.1006/CVIU.1998.0692>
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
15. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. In: ICLR (2025)
16. Frogner, C., Zhang, C., Mobahi, H., Araya, M., Poggio, T.A.: Learning with a wasserstein loss. In: NeurIPS (2015)
17. Ge, Z., Liu, S., Li, Z., Yoshie, O., Sun, J.: OTA: Optimal transport assignment for object detection. In: CVPR. pp. 303–312 (2021)
18. Katti, A.R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., Faddoul, J.B.: Chargrid: Towards understanding 2d documents. In: EMNLP. pp. 4459–4469. Association for Computational Linguistics (2018). <https://doi.org/10.18653/v1/D18-1476>
19. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: ACM SIGIR (2020)
20. Kim, J., Lee, G., Choi, D., Kim, T., Shin, K.: Hybrid-vector retrieval for visually rich documents: Combining single-vector efficiency and multi-vector accuracy. arXiv preprint (2025)
21. Lee, T., Choi, J., Choi, J., Kang, M.: Unpaired point cloud completion via unbalanced optimal transport. In: ICML (2025)
22. Li, L., Wang, Y., Xu, R., Wang, P., Feng, X., Kong, L., Liu, Q.: Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In: ACL. pp. 14369–14387 (2024)
23. Liero, M., Mielke, A., Savaré, G.: Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Invent. Math.* **211**(3), 969–1117 (2018)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. vol. 36, pp. 34892–34916 (2023)
25. Loison, A., Macé, Q., Edy, A., Xing, V., Balough, T., Moreira, G., Liu, B., Faysse, M., Hudelot, C., Viaud, G.: Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. arXiv preprint (2026)
26. Ma, X., Lin, S.C., Li, M., Chen, W., Lin, J.: Unifying multimodal retrieval via document screenshot embedding. In: EMNLP. pp. 6492–6505 (2024)
27. Ma, Y., Li, J., Zang, Y., Wu, X., Dong, X., Zhang, P., Cao, Y., Duan, H., Wang, J., Cao, Y., et al.: Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. In: Findings of ACL (2025)
28. Ma, Z., Wei, X., Hong, X., Lin, H., Qiu, Y., Gong, Y.: Learning to count via unbalanced optimal transport. In: AAAI. pp. 2319–2327 (2021)

29. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *JMLR* **9**(11) (2008)
30. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Info-graphicvqa. In: *WACV*. pp. 1697–1706 (2022)
31. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images. In: *WACV* (2021)
32. Mitra, M., Chaudhuri, B.B.: Information retrieval from documents: A survey. *IR* **2**(2-3), 141–163 (2000). <https://doi.org/10.1023/A:1009950525500>
33. Montesuma, E.F., Mboula, F.M.N., Souloumiac, A.: Recent advances in optimal transport for machine learning. *IEEE TPAMI* **47**(2), 1161–1180 (2024)
34. Peyré, G., Cuturi, M.: Computational optimal transport. *Found. Trends Mach. Learn.* **11**(5-6), 355–607 (2019)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *ICML*. pp. 8748–8763. *PmLR* (2021)
36. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: *NeurIPS* (2021)
37. S., M., C.Y., S., K., Y.: Historical review of ocr research and development. *Proc. IEEE* **80**(7), 1029–1058 (07 1992). <https://doi.org/10.1109/5.156468>, <https://cir.nii.ac.jp/crid/1362825895220049152>
38. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: *CVPR* (2020)
39. Smith, R.: An overview of the tesseract OCR engine. In: *ICDAR*. pp. 629–633. *IEEE Computer Society* (2007). <https://doi.org/10.1109/ICDAR.2007.4376991>, <https://doi.org/10.1109/ICDAR.2007.4376991>
40. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint* (2021)
41. Villani, C.: *Optimal Transport: Old and New*, vol. 338. Springer (2009)
42. Wang, J., Dai, T., Zhang, B., Yu, S., Lim, E.G., Xiao, J.: POT: Prototypical optimal transport for weakly supervised semantic segmentation. In: *CVPR* (2025)
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint* (2024)
44. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: *KDD*. pp. 1192–1200. *Association for Computing Machinery* (2020). <https://doi.org/10.1145/3394486.3403172>
45. Yan, Y., Xu, G., Zou, X., Liu, S., Kwok, J., Hu, X.: Docpruner: A storage-efficient framework for multi-vector visual document retrieval via adaptive patch-level embedding pruning. *arXiv preprint* (2025)
46. Yang, C., Sui, Y., Xiao, J., Huang, L., Gong, Y., Li, C., Yan, J., Bai, Y., Sadayappan, P., Hu, X., et al.: Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In: *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 19803–19813. *IEEE Computer Society* (2025)
47. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: *CVPR* (2025)
48. Ye, W., Wu, Q., Lin, W., Zhou, Y.: Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In: *AAAI*. pp. 22128–22136 (2025)

49. Zhang, J., Yang, W., Lai, S., Xie, Z., Jin, L.: Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. In: AAAI (2025)
50. Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., Zhang, S.: Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In: ICCV. pp. 20857–20867 (2025)
51. Zhu, F., Lei, W., Feng, F., Wang, C., Zhang, H., Chua, T.S.: Towards complex document understanding by discrete reasoning. In: ACM MM. pp. 4857–4866 (2022)