

# Bag of Bags: Adaptive Visual Vocabularies for Genizah Join Image Retrieval

Sharva Gogawale, Gal Grudka, Daria Vasyutinsky-Shapira,  
Omer Ventura, Berat Kurar-Barakat, Nachum Dershowitz

School of Computer Science and AI, Tel Aviv University, Ramat Aviv, Israel

{sharvag, galgrudka, omerventura}@mail.tau.ac.il, {dariashap, berat, nachum}@tau.ac.il

## Abstract

*A join is a set of manuscript fragments identified as originally emanating from the same manuscript. We study manuscript join retrieval: Given a query image of a fragment, retrieve other fragments originating from the same physical manuscript. We propose Bag of Bags (BoB), an image-level representation that replaces the global-level visual codebook of classical Bag of Words (BoW) with a fragment-specific vocabulary of local visual words. Our pipeline trains a sparse convolutional autoencoder on binarized fragment patches, encodes connected components from each page, clusters the resulting embeddings with per-image  $k$ -means, and compares images using set-to-set distances between their local vocabularies. Evaluated on fragments from the Cairo Genizah, the best BoB variant (viz. Chamfer) achieves Hit@1 of 0.78 and MRR of 0.84, compared to 0.74 and 0.80, respectively, for the strongest BoW baseline (BoW-RawPatches- $\chi^2$ ), a 6.1% relative improvement in top-1 accuracy. We furthermore study a mass-weighted BoB-OT variant that incorporates cluster population into prototype matching and present a formal approximation guarantee bounding its deviation from full component-level optimal transport. A two-stage pipeline using a BoW shortlist followed by BoB-OT reranking provides a practical compromise between retrieval strength and computational cost, supporting applicability to larger manuscript collections. The code and dataset are available at [https://github.com/TAU-CH/midrash\\_bob](https://github.com/TAU-CH/midrash_bob).*

## 1. Introduction

The Cairo Genizah is a unique source of preserved fragmented medieval manuscripts, accumulated between the 11th and 19th centuries in the Ben Ezra Synagogue in Old Cairo and now dispersed across dozens of libraries and private collections worldwide. The manuscripts are mostly written in Hebrew, Aramaic, and Judeo-Arabic. Their study over the past century and a quarter has had enormous impact on our knowledge of medieval Mediterranean history,

literature, commerce, and culture.

To facilitate more thorough study of the Genizah texts, there is a strong need to group related fragments and reconstruct as much as possible of their original manuscripts. Over decades, scholars have spent considerable effort manually identifying such groups, known as *joins*: manuscript fragments that originally belonged to the same physical codex or scroll, but have since been separated, requiring researchers to examine them in person across institutions. Manual identification remains the gold standard for discovering joins; however, it does not scale well and cannot be applied to the entire collection.

Automating join retrieval is a challenging computer vision task because many manuscript fragments are often severely damaged, incomplete, and stained. The difficulty arises from the fact that fragments originating from the same manuscript may share only subtle, handwriting-style-specific visual patterns, while fragments from different manuscripts can still appear globally similar in aspects such as background texture, aging, and degradation level.

Standard retrieval architectures frequently rely on Bag of Words (BoW) paradigms. BoW maps images to histograms over a shared global codebook. It summarizes all fragments with respect to the same dictionary. However, this global quantization is fundamentally limited for paleographic analysis, as it causes critical image-specific handwriting style information to be lost during quantization. Two images with identical codeword frequencies but different geometric character style distributions are assigned distance zero by BoW, yet they may have been written by entirely different hands. Conversely, fragments from the same manuscript may differ in lighting and layout, causing BoW to miss them. BoW’s pooling over a global codebook erases the image-specific handwriting style signal.

Key contributions are as follows:

- We propose Bag of Bags (BoB), a fragment-specific representation for fragment retrieval that replaces a shared global visual-word codebook with per-fragment vocabularies of local visual words, constructed from sparse autoencoder embeddings of connected-component patches.

- We instantiate BoB with three set-to-set distances—bipartite assignment, symmetric Chamfer, and mass-weighted optimal transport (OT)—and show empirically that soft nearest-neighbor matching (Chamfer) is most robust for partially damaged fragments, while weighted OT provides a principled mass-aware alternative with formal approximation guarantees.
- We present a two-stage retrieval pipeline in which BoW-Cosine generates candidates and BoB-OT reranks only the shortlist, reducing online cost to  $\mathcal{O}(M \cdot K^3)$  independent of gallery size while closely preserving the accuracy of exhaustive BoB-OT.
- We provide a detailed ablation on vocabulary size, encoder dimension, activation sparsity, and component normalization, showing that BoB’s page-adaptive vocabulary structure contributes retrieval gains beyond what the encoder quality alone explains.

## 2. Related Work

The Cairo Genizah, stored in the Ben Ezra Synagogue in Fustat (Old Cairo) and discovered and retrieved at the end of the 19th century, is a large and historically significant collection of manuscript fragments dating mainly from the 10th to the 15th centuries. These documents were subsequently dispersed across more than fifty libraries and collections worldwide [14]. A fundamental challenge facing Genizah scholarship is that many leaves were discovered as loose or damaged fragments. Identifying joins has long been a central scholarly task. The results of these manual efforts can be found on the site of the [Friedberg Genizah Project](#), currently in transition to the [National Library of Israel](#). Though thousands of joins have already been documented, substantial manual effort is still required [14]. This motivates automated retrieval systems that, given a query fragment, rank likely join candidates for expert inspection.

Automated identification of joins was first explored in [21], who proposed a method for finding candidate matches between manuscript fragments. Their approach represents each image using local visual features and a bag of features representation, and then compares pairs of leaves using a learned similarity measure to determine whether they may belong to the same original work. However, this method requires computationally expensive pairwise comparisons during retrieval. Analogous fragment assembly problems arise in other historical document collections, including ancient papyri, where AI-based approaches have been surveyed for matching dispersed fragments via visual texture and learned similarity measures [16, 20].

More broadly, historical document retrieval has been studied through word and pattern spotting, where the goal is to retrieve repeated words or writer-specific patterns without full OCR. An efficient learning-free word spotting method based on local binary patterns was proposed



Figure 1. Representative join groups from the Cairo Genizah benchmark. Each row shows fragments originating from the same physical manuscript, held among different institutions worldwide. (The blue backgrounds were designed for easy algorithmic segmentation.)

for handwritten historical documents, demonstrating robustness to degradation and variations across multiple writers [7]. Bag of Visual Words (BoVW) and Bag of Words (BoW) models constitute a fundamental paradigm in image retrieval [19]. These approaches quantize local image descriptors into a shared codebook, representing each image as a histogram of visual words. Extending this to documents, Shekhar showed that recognition-free retrieval is possible by quantizing local descriptors and comparing fixed-dimensional histograms [18]. A two-stage system for word and pattern spotting in historical manuscripts combines BoVW for candidate generation with a spatial verification step based on the Longest Weighted Profile algorithm, enabling retrieval from a single query [8]. These methods are closely related, but they target repeated words or local patterns rather than page-level manuscript identity. This global quantization can be challenging for fine-grained scribal analysis: fragments with identical global word counts but entirely different geometric distributions in feature space receive identical representations. Optimal transport offers a more flexible alternative for comparing visual distributions. Earth Mover’s Distance (EMD) utilizes optimal transport to compare feature distributions [17], recently extended to match continuous GMMs [15]. We combine these threads: instead of assigning all pages to one global codebook, we learn a fragment-specific vocabulary and compare fragments through set-to-set matching.

A closely related research area is writer identifica-

tion [6], which seeks to determine the identity of the writer of a specific document based on handwriting out of a given set of writers for whom writing samples are supplied. Computer vision tools have been developed to assist paleographers directly in scribal hand identification, producing interpretable patch-level similarity maps that experts can interrogate alongside traditional paleographic evidence [12]. Existing methods are commonly divided into texture-based and grapheme-based approaches. Texture-based methods [1–3] characterize handwriting using statistical properties of the written trace, such as slant, curvature, and texture, whereas grapheme-based methods extract local structures and map them into a shared feature space. The latter can be formed using BoW, which relies on zero-order statistics by counting assignments of local descriptors to the nearest visual words in a predefined vocabulary [9, 22] or by richer aggregation schemes such as Fisher vectors, which encode first and second-order deviations of local descriptors from a probabilistic visual vocabulary [10, 13], and VLAD, which aggregates residuals between local descriptors and their nearest visual words [4, 5].

### 3. Methodology

#### 3.1. Problem Formulation

Let  $\mathcal{I} = \{I_i\}_{i=1}^N$  be a collection of manuscript fragment images partitioned into join clusters  $\mathcal{C} = \{C_j\}_{j=1}^M$ , where  $C_j \subseteq \mathcal{I}$  denotes a set of fragment images originating from the same physical manuscript. Given a query fragment  $I_q$ , the *retrieval task* is to rank all other images so that members of  $C_{j(q)}$ , where  $j(q)$  denotes the cluster index of  $I_q$ , appear at top ranks. We evaluate retrieval using Hit@ $k$ , mAP@ $k$ , mean reciprocal rank (MRR), and Macro-F1@1. The implemented system constructs two competing representations in a shared latent space: (i) an image-adaptive *Bag of Bags* (BoB) representation, and (ii) a global codebook *Bag of Words* (BoW) baseline. The BoB pipeline proceeds in three stages: (1) character-anchored patch extraction via connected component analysis; (2) sparse autoencoder encoding of each patch; and (3) per-image  $k$ -means clustering to produce a local prototype vocabulary. Fragment similarity is then measured as a set-to-set distance between per-image vocabularies.

#### 3.2. Dataset

Our dataset consists of 287 manuscript fragment images from the Cairo Genizah join benchmark, drawn from multiple institutions: Cambridge University Library (Taylor-Schechter and related collections), the Jewish Theological Seminary (JTS/ENA) in New York, the British Library, and the Alliance Israélite Universelle in Paris. See the samples in Fig. 1. We validate consistency against ground-truth labels across 100 join clusters. The dataset is highly im-

balanced, with cluster sizes ranging from 2–9 fragments. To account for this imbalance, in evaluation, we include Macro-F1@1 alongside Hit@1 and MRR; it averages the per-cluster F1 score, weighting each cluster equally regardless of size. The autoencoder is trained on individual patches without join-cluster supervision and is therefore unaffected by cluster-size skew. The manuscripts are, for the most part, written in Hebrew letters across a variety of languages and script modes, some more square (with mainly disconnected letters) and others more cursive (with many interconnected letters). All experiments use binarized (Otsu-thresholded) grayscale images. Our experiments are conducted on this manually annotated benchmark subset (287 images, 100 join clusters), drawn from a much larger Genizah collection containing approximately 250–300K fragment images overall. We focus on this annotated subset because reliable join labels are currently available at this scale, while the proposed retrieval pipeline is designed to support deployment over the full corpus.

#### 3.3. Character-Anchored Patch Extraction

For each image, we first binarize. On the resulting image  $I_i$ , we run 8-connected component analysis to obtain the component set  $\mathcal{C}_i$ . Images whose mean pixel intensity exceeds 128 are inverted so that text is always white (non-zero) for consistent detection.

**Area filtering.** Given an input manuscript fragment  $I$ , we isolate individual character instances using connected component analysis. We filter the resulting component set  $\mathcal{C}_i$  to exclude noise, background artifacts, and massive text blocks based on pixel area, yielding valid components

$$\tilde{\mathcal{C}}_i = \{c \in \mathcal{C}_i : A_{\min} \leq \text{area}(c) \leq A_{\max}\}.$$

Thresholds  $A_{\min} = 300$  and  $A_{\max} = 3000$  were chosen based on visual inspection of the component size distribution: Components below 300 pixels correspond primarily to noise, ink specks, and binarization artifacts, while those above 3000 pixels typically correspond to large ink blobs or partially merged letter clusters that do not represent individual letterforms.

**Fit-and-pad normalization.** Each retained component is scaled so that its longest side equals 60 pixels while preserving the aspect ratio and is then centered in a zero-padded  $64 \times 64$  patch. This yields scale-normalized patches while preserving character geometry. The resulting patch set for image  $I_i$  is  $\mathcal{P}_i = \{p_1^i, \dots, p_{n_i}^i\}$ .

**Quality filtering.** Each component bounding-box crop must contain at least 5% white pixels. After normalization, the final patch must also contain at least 2% white pixels. Pages yielding fewer than 200 valid components after all filters are excluded.

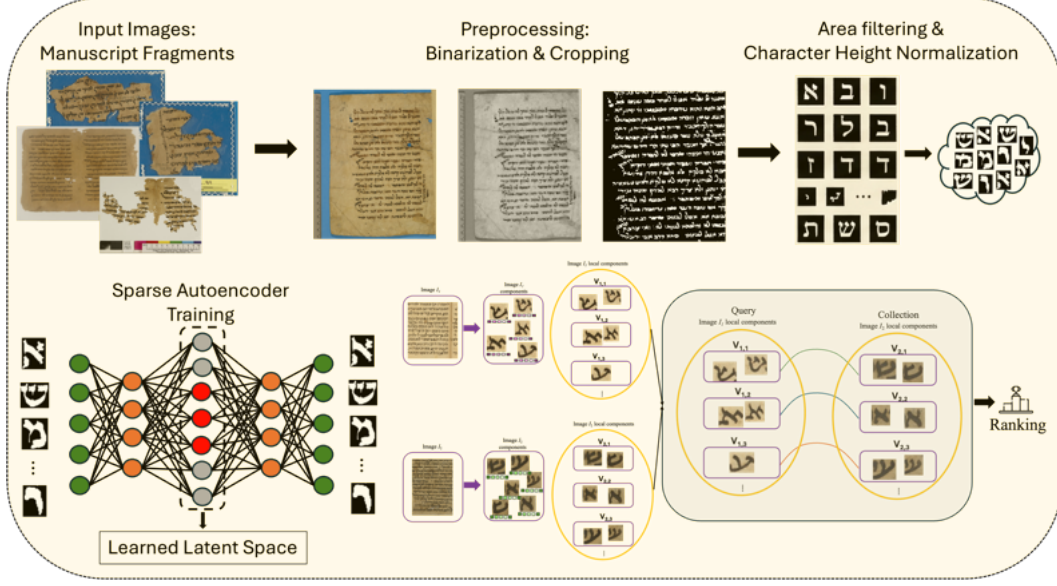


Figure 2. Overview of the proposed pipeline.

### 3.4. Sparse Convolutional Autoencoder

We encode normalized character patches into a dense feature space using a sparse convolutional autoencoder,  $f_\theta : \mathbb{R}^{64 \times 64} \rightarrow \mathbb{R}^{128}$ . The encoder  $f_\theta^{\text{enc}}$  consists of three stride-2 convolutional layers followed by a linear projection. The decoder  $f_\theta^{\text{dec}}$  mirrors the encoder using a linear expansion followed by three transposed convolutional layers to reconstruct the input patch. The full model has  $\approx 1.1\text{M}$  parameters at  $d = 128$ . The network is trained to minimize the mean-squared reconstruction error with an  $L_1$  sparsity penalty on the latent representation:

$$\mathcal{L} = \underbrace{\frac{1}{|\mathcal{P}|} \sum_p \|f_\theta^{\text{dec}}(f_\theta(p)) - p\|_2^2}_{\text{reconstruction}} + \lambda \underbrace{\frac{1}{|\mathcal{P}|} \sum_p \|f_\theta(p)\|_1}_{\text{sparsity}},$$

where  $\lambda = 10^{-5}$  balances reconstruction fidelity and feature sparsity. For a fragment  $I_i$ , this produces a set of continuous embeddings  $E_i = \{f_\theta(p) : p \in \mathcal{P}_i\} \subseteq \mathbb{R}^{128}$ .

The autoencoder is trained on patches drawn from manuscript images in the dataset. Patches are extracted identically for both training and inference using an aspect-ratio-preserving normalization (Sec. 3.3). Specifically, each connected component  $c \in \tilde{\mathcal{C}}_i$  is cropped to its bounding box, resized to a canonical extent while preserving aspect ratio, and centered on a zero-padded  $64 \times 64$  image patch. This ensures that the encoder learns representations directly from the isolated character patches used during retrieval and avoids any mismatch between training and deployment.

**Training details.** For each training image, we sample up to 300 valid component patches and discard patches with in-

sufficient foreground content. The model is trained from scratch for 50 epochs using the Adam optimizer with a learning rate of  $10^{-3}$  and a batch size of 256, with early stopping based on the reconstruction loss.

### 3.5. The Bag-of-Bags (BoB) Representation

Unlike BoW, which quantizes  $E_i$  against a global codebook, BoB generates a page-adaptive representation. We partition  $E_I$  into  $K = 20$  clusters using  $K$ -means in the main configuration used for Tab. 1 and the runtime analysis; Sec. 4.3 studies sensitivity to  $K$ . Let  $G_1, \dots, G_K$  denote these clusters, and let  $n_I = |E_I|$  be the number of component embeddings for page  $I$ . The page is represented by a set of prototypes and their corresponding masses:  $\mu_a = \frac{1}{|G_a|} \sum_{z \in G_a} z$ ,  $\pi_a = |G_a|/n_I$ . The final BoB representation is the “bag” of local prototypes:  $B(I) = \{(\mu_a, \pi_a)\}_{a=1}^K$ .

### 3.6. Distance Formulation

We write  $V(I) = \{\mu_a\}_{a=1}^K$  for the support set of centroids in the weighted BoB representation  $B(I) = \{(\mu_a, \pi_a)\}_{a=1}^K$ .

Given two manuscript pages  $I$  and  $J$  with local vocabularies,  $V(I) = \{\mu_a\}_{a=1}^K$  and  $V(J) = \{\nu_b\}_{b=1}^K$ , and normalized cluster-population weights  $\pi, \rho \in \Delta^K$ , we define three set-to-set distances over their page-adaptive prototypes.

**BoB-Chamfer.** Each prototype independently finds its nearest match in the opposing vocabulary:

$$d_{\text{Chamfer}}(I, J) = \frac{1}{2K} \left( \sum_{a=1}^K \min_b \|\mu_a - \nu_b\|_2 + \sum_{b=1}^K \min_a \|\nu_b - \mu_a\|_2 \right).$$



Unlike bipartite assignment, Chamfer imposes no global bijection between prototypes: each centroid independently contributes its nearest-neighbor cost. This rewards partial style overlap, since a fragment may share only a subset of character types with its join candidate while still receiving a low distance, without penalty for unmatched prototypes. This directly reflects the physical structure of damaged Genizah fragments, where material loss or truncation guarantees that only a subset of character styles will be mutually present. Chamfer is  $\mathcal{O}(K^2)$  per pair and achieves the strongest empirical retrieval performance across all metrics (Sec. 4). We note that while symmetric, it does not satisfy the triangle inequality in general.

**BoB-Hungarian.** As a formal geometric alternative, we enforce a strict one-to-one bipartite assignment:

$$d_{\text{Hung}}(I, J) = \frac{1}{K} \min_{\sigma \in S_K} \sum_{a=1}^K \|\mu_a - \nu_{\sigma(a)}\|_2,$$

where  $S_K$  denotes the permutation group on  $K$  elements. Solved via the Hungarian algorithm on the cost matrix  $C[a, b] = \|\mu_a - \nu_b\|_2$ , this is exactly the discrete Wasserstein-1 distance  $W_1(\mathcal{P}_I, \mathcal{P}_J)$  between uniform empirical distributions  $\mathcal{P}_I = \frac{1}{K} \sum_a \delta_{\mu_a}$ . Because  $W_1$  with an  $\ell_2$  ground metric is a valid metric,  $d_{\text{Hung}}$  satisfies symmetry, identity of indiscernibles, and the triangle inequality—properties Chamfer lacks. Its per-pair cost is  $\mathcal{O}(K^3)$ .

**BoB-OT (mass-weighted).** We further incorporate true scribal frequency by replacing uniform weights with normalized cluster populations:

$$d_{\text{OT}}(I, J) = \min_{T \in \Pi(\pi, \rho)} \sum_{a=1}^K \sum_{b=1}^K T_{ab} \|\mu_a - \nu_b\|_2,$$

where  $\Pi(\pi, \rho)$  is the set of non-negative transport plans with marginals  $\pi$  and  $\rho$ , solved via the POT library [11]. A prototype accounting for 40% of a page’s components contributes proportionally more to the distance than one accounting for 2%. BoB-OT improves over  $d_{\text{Hung}}$  (+5.6 pp Hit@1), confirming that prototype mass is informative, though both assignment-based distances are ultimately outperformed by Chamfer in this highly-degraded domain.

**Theoretical guarantee for BoB-OT.** While Chamfer excels empirically due to the partial-overlap structure of manuscript fragments, BoB-OT admits a formal approximation guarantee that grounds the use of prototype-level matching and motivates the vocabulary-size ablation.

**Proposition.** *For any two pages,  $I$  and  $J$ , let*

$$\begin{aligned} \mathcal{P}_I &= \frac{1}{n_I} \sum_{i=1}^{n_I} \delta_{z_i}, & \tilde{\mathcal{P}}_I &= \sum_{a=1}^K \pi_a \delta_{\mu_a}, \\ \varepsilon_I &= \frac{1}{n_I} \sum_{i=1}^{n_I} \|z_i - \mu_{a(i)}\|_2, \end{aligned}$$

*and define  $\mathcal{P}_J$ ,  $\tilde{\mathcal{P}}_J$ , and  $\varepsilon_J$  analogously for page  $J$ . Then  $|W_1(\mathcal{P}_I, \mathcal{P}_J) - W_1(\tilde{\mathcal{P}}_I, \tilde{\mathcal{P}}_J)| \leq \varepsilon_I + \varepsilon_J$ .*

*Proof.* Assigning each  $z_i$  to its centroid  $\mu_{a(i)}$  defines a feasible transport plan from  $\mathcal{P}_I$  to  $\tilde{\mathcal{P}}_I$ , so  $W_1(\mathcal{P}_I, \tilde{\mathcal{P}}_I) \leq \varepsilon_I$ . The triangle inequality for  $W_1$  gives the desired bound.  $\square$

This shows that BoB-OT approximates full component-level optimal transport with error controlled by the within-page  $K$ -means quantization quality. Larger  $K$  reduces  $\varepsilon_I$  and  $\varepsilon_J$ , tightening this approximation—as confirmed empirically (Tab. 3). BoB-Chamfer achieves stronger retrieval performance due to the partial-overlap structure of damaged fragments, while BoB-OT provides a principled mass-aware alternative with formal approximation guarantees.

### 3.7. Bag-of-Words (BoW) Baseline

We evaluate two BoW variants that share the same encoder as BoB but differ in how the global codebook is constructed. **BoW-centroids** pools the page-level local visual words (BoB prototypes, weighted by cluster population) into a global codebook. **BoW-RawPatches** constructs the global codebook directly from all raw component embeddings across the dataset, bypassing per-page clustering entirely—this is the standard Bag-of-Visual-Words formulation and serves as the primary BoW comparator for BoB. For each page  $I$ , let  $\{(\mu_a, m_a)\}_{a=1}^K$  denote its local prototypes and their cluster populations, where  $m_a$  is the number of component embeddings assigned to prototype  $\mu_a$  and  $n_I = \sum_{a=1}^K m_a$ . We fit a global  $k$ -means codebook,  $C = \{c_r\}_{r=1}^{K_g}$  ( $K_g = 100$ ), on the pooled set of local prototypes, weighted by their populations  $m_a$  (equivalently, by repeating each prototype according to its cluster size).

Each page is then represented by a tf-idf weighted histogram  $h_I \in \mathbb{R}^{K_g}$ . Its term-frequency component is

$$\text{tf}_I[r] = \frac{1}{n_I} \sum_{a=1}^K m_a \mathbb{1}(\text{nn}(\mu_a) = c_r),$$

where  $\text{nn}(\mu_a)$  denotes the nearest global codeword to  $\mu_a$ , and the indicator function  $\mathbb{1}(\cdot)$  equals 1 if its argument holds and 0 otherwise. The inverse-document-frequency term is

$$\text{idf}[r] = \log\left(\frac{N+1}{|\{I : \text{tf}_I[r] > 0\}| + 1}\right) + 1,$$

where  $N$  is the number of pages in the corpus. We use a smoothed inverse-document-frequency term, which avoids zero weights for codewords appearing in all pages and remains well-defined if a codeword has zero document frequency. The final representation is obtained by tf-idf weighting followed by  $\ell_2$  normalization.

We compare pages via distances between their final page-level BoW representations: Euclidean, cosine,  $\chi^2$ , and

Hellinger. For Hellinger distance, histograms are first renormalized to unit  $\ell_1$  mass. This baseline provides a meaningful shared-vocabulary comparator while discarding the page-specific prototype geometry retained by BoB.

### 3.8. Retrieval, Reranking and Complexity

For each query page, all other pages are ranked in ascending order of distance, excluding the query page itself. A page is considered relevant if it belongs to the same join cluster as the query. We report Hit@ $k$ , mAP@ $k$ , MRR, and Macro-F1@1. Query pages with no positive mate in the dataset are excluded from metric aggregation.

**Computational cost.** BoW comparison costs  $\mathcal{O}(K_g)$  per pair once histograms are built ( $K_g = 100$  operations). BoB-Chamfer costs  $\mathcal{O}(K^2)$  (400 operations at  $K = 20$ ) and BoB-Hungarian/OT cost  $\mathcal{O}(K^3)$  (8000 operations)—a  $4\times$  to  $80\times$  overhead per pair relative to BoW. This overhead is the price of expressiveness: without per-image quantization, directly matching all  $n_I \geq 200$  raw component embeddings via OT would cost  $\mathcal{O}(n^3) \approx 10^6\text{--}10^9$  operations per pair, making gallery-scale retrieval infeasible. At our operating point ( $K = 20$ ), the full precomputed distance matrix builds in under five seconds; precomputed lookup at query time is effectively free for all methods.

**Two-stage retrieval.** For larger manuscript collections, where computing or storing a full  $N \times N$  BoB-OT matrix becomes expensive—BoB-OT costs  $\mathcal{O}(K^3)$  per pair versus  $\mathcal{O}(K_g)$  for BoW—we propose a two-stage pipeline: BoW-Cosine retrieves the top- $M$  candidates efficiently, and BoB-OT reranks only that shortlist. We evaluate with  $M=30$ , reducing online cost to  $\mathcal{O}(M \cdot K^3)$  per query independent of gallery size. See Fig. 2. Table 1 shows that the two-stage pipeline matches or improves exhaustive BoB-OT across all reported ranking metrics on this benchmark.

## 4. Results

We evaluated retrieval strategies on our benchmark dataset.

### 4.1. Retrieval Performance

Table 1 reports the full retrieval comparison across three families of baselines. **Pooling baselines** (MeanPool, MaxPool) aggregate all component embeddings into a single page vector without any clustering; MaxPool performs poorly (Hit@1  $\leq 0.31$ ), confirming that the element-wise maximum is not an informative aggregation for manuscript embeddings. MeanPool-Cosine is competitive at 0.545 Hit@1, but substantially below BoB, showing that a simple global average of component embeddings cannot substitute for structured page-level vocabulary matching. **BoW-centroids** constructs a global codebook over page-level BoB prototypes; its best variant (L2) achieves 0.545 Hit@1. **BoW-RawPatches** is the standard Bag-of-Visual-Words formulation applied directly to raw component embeddings

without any per-page clustering stage; its strongest variant ( $\chi^2$ ) reaches 0.739 Hit@1 and 0.800 MRR, making it the most competitive baseline.

The main result is consistent across all ranking metrics: page-adaptive BoB representations outperform every baseline. BoB-Chamfer achieves the highest overall performance at 0.784 Hit@1 and 0.841 MRR, improving over the strongest BoW baseline (BoW-RawPatches- $\chi^2$ , 0.739 Hit@1) by +4.5% absolute (+6.1% relative), and over the pooling baselines by a larger margin. Crucially, BoW-centroids is substantially weaker than BoW-RawPatches despite sharing the same encoder, showing that the gain of BoB does not come from the encoder alone, but from the combination of page-adaptive clustering and set-level prototype matching.

Among adaptive distances, BoB-Chamfer outperforms both BoB-Hungarian-L2 and BoB-OT by 0.079 and 0.023 at Hit@1 respectively. This suggests that in the fragment setting, soft nearest-neighbor matching is more robust than strict one-to-one assignment: Paired fragments often share only a subset of local patterns due to truncation and damage, and Chamfer’s partial-overlap formulation rewards shared style without penalizing unmatched prototypes.

### 4.2. Distance-Space Separation

To probe whether fragment adaptive vocabularies induce a more separable retrieval geometry, we compare the distributions of intra-cluster and inter-cluster distances. Table 2 reports this analysis for BoW-Cosine and BoB-Hungarian-L2. BoB-Hungarian-L2 shows a larger intra/inter-cluster mean gap (0.343 vs. 0.270), slightly stronger KS separation (0.665 vs. 0.641), and nearly identical AUC separation (0.902 vs. 0.899). We use Hungarian-L2 for this analysis because, as a valid metric (Sec. 3.6), it provides a geometrically principled distance space for comparing intra- and inter-cluster distance distributions. We view this as supporting diagnostic evidence for the broader BoB design, rather than as a complete explanation of the strongest retrieval results, which are achieved by BoB-Chamfer.

### 4.3. Ablation Studies

We examine sensitivity of retrieval performance to the main design choices: vocabulary size  $K$ , latent dimension  $d$ , sparsity regularization, and component normalization. All ablations use BoB-Hungarian-L2 unless stated otherwise.

**Local vocabulary size  $K$ .** Table 3 varies the number of page-adaptive visual words for  $K = 8..64$  at fixed  $d = 128$ , sparsity enabled. Performance improves until  $K = 32$ , then degrades at  $K = 64$ . On this benchmark subset,  $K = 32$  yields the strongest retrieval metrics. The main results and runtime analysis use  $K = 20$ ; we treat Tab. 3 as a sensitivity study rather than as evidence of a single universally optimal operating point, since the best  $K$  depends

Table 1. Full retrieval comparison on the Genizah join-retrieval benchmark. Methods are grouped into pooling baselines, shared-vocabulary BoW baselines (BoW-RawPatches: global codebook over raw component embeddings; BoW-centroids: global codebook over page-level BoB prototypes), and page-adaptive BoB distances. Higher is better for all metrics. **Boldface** indicates the best value per column.

| Family         | Method                                        | Hit@1        | mAP@1        | Hit@5        | mAP@5        | Hit@10       | mAP@10       | MRR          | MacroF1@1    |
|----------------|-----------------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Pooling        | MeanPool-Cosine                               | 0.545        | 0.545        | 0.773        | 0.469        | 0.841        | 0.475        | 0.642        | 0.384        |
|                | MaxPool-L2                                    | 0.307        | 0.307        | 0.534        | 0.288        | 0.648        | 0.300        | 0.424        | 0.187        |
| BoW-centroids  | L2                                            | 0.545        | 0.545        | 0.682        | 0.506        | 0.784        | 0.516        | 0.619        | 0.397        |
|                | Cosine                                        | 0.534        | 0.534        | 0.739        | 0.509        | 0.841        | 0.520        | 0.617        | 0.354        |
|                | $\chi^2$                                      | 0.511        | 0.511        | 0.716        | 0.510        | 0.830        | 0.525        | 0.622        | 0.352        |
|                | Hellinger                                     | 0.500        | 0.500        | 0.716        | 0.505        | 0.830        | 0.519        | 0.616        | 0.344        |
| BoW-RawPatches | L2                                            | 0.659        | 0.659        | 0.795        | 0.562        | 0.852        | 0.564        | 0.730        | 0.543        |
|                | Cosine                                        | 0.670        | 0.670        | 0.841        | 0.605        | 0.898        | 0.611        | 0.752        | 0.516        |
|                | $\chi^2$                                      | 0.739        | 0.739        | 0.852        | 0.649        | 0.898        | 0.655        | 0.800        | 0.584        |
|                | Hellinger                                     | 0.705        | 0.705        | 0.841        | 0.631        | 0.909        | 0.640        | 0.774        | 0.560        |
| BoB (ours)     | Hungarian                                     | 0.705        | 0.705        | 0.864        | 0.665        | 0.909        | 0.660        | 0.771        | 0.580        |
|                | OT (mass-weighted)                            | 0.761        | 0.761        | 0.875        | 0.671        | 0.886        | 0.668        | 0.814        | 0.615        |
|                | <b>Chamfer</b>                                | <b>0.784</b> | <b>0.784</b> | <b>0.898</b> | <b>0.734</b> | <b>0.932</b> | <b>0.736</b> | <b>0.841</b> | <b>0.675</b> |
| Two-stage      | BoW-Cosine $\rightarrow$<br>BoB-OT ( $M=30$ ) | 0.771        | 0.771        | 0.875        | 0.691        | 0.901        | 0.668        | 0.814        | 0.615        |

Table 2. Distance-separation statistics for intra-cluster vs. inter-cluster pairs. Larger mean gap, KS statistic, and AUC separation indicate stronger distributional separation.

| Method           | Intra | Inter | Gap   | KS    | AUC   | Cohen's $d$ |
|------------------|-------|-------|-------|-------|-------|-------------|
| BoW-Cosine       | 0.347 | 0.617 | 0.270 | 0.641 | 0.899 | 1.897       |
| BoB-Hungarian-L2 | 1.735 | 2.078 | 0.343 | 0.665 | 0.902 | 1.831       |

Table 3. Ablation on local vocabulary size  $K$  for BoB-Hungarian-L2 ( $d = 128$ , sparsity enabled). (MF1 = MacroF1)

| $K$       | Hit@1        | mAP@1        | Hit@5        | mAP@5        | MRR          | MF1@1        |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|
| 8         | 0.648        | 0.648        | 0.818        | 0.600        | 0.739        | 0.541        |
| 16        | 0.705        | 0.705        | 0.841        | 0.650        | 0.767        | 0.569        |
| 20        | 0.705        | 0.705        | 0.864        | 0.665        | 0.771        | 0.580        |
| <b>32</b> | <b>0.773</b> | <b>0.773</b> | <b>0.864</b> | <b>0.689</b> | <b>0.813</b> | <b>0.647</b> |
| 64        | 0.705        | 0.705        | 0.830        | 0.658        | 0.770        | 0.599        |

on script complexity and collection scale (see Section 5).

**Latent dimension  $d$ .** Table 4 varies the encoder dimension while retraining from scratch. Performance peaks at  $d = 128$ ; increasing to 256 does not improve and slightly degrades retrieval, suggesting that 128-dimensional embeddings already capture the discriminative structure of the character embedding space. We use  $d = 128$  as the default.

**Activation sparsity regularization.** Table 5 compares training with and without the  $\ell_1$  latent penalty. Enabling sparsity improves Hit@1 (+0.012), mAP@1 (+0.012), Hit@5 (+0.023), mAP@5 (+0.017), and MRR (+0.008). MacroF1@1 also improves substantially under sparsity (+0.052). Overall, the sparse model is superior across all

Table 4. Ablation on latent dimension  $d$  ( $K=32$ , sparsity enabled).

| $d$        | Hit@1        | mAP@1        | Hit@5        | mAP@5        | MRR          | MF1@1        |
|------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 64         | 0.705        | 0.705        | 0.807        | 0.642        | 0.756        | 0.578        |
| <b>128</b> | <b>0.773</b> | <b>0.773</b> | <b>0.864</b> | <b>0.689</b> | <b>0.813</b> | <b>0.647</b> |
| 256        | 0.648        | 0.648        | 0.807        | 0.604        | 0.719        | 0.546        |

Table 5. Ablation on  $\ell_1$  sparsity regularization for BoB-Hungarian-L2 ( $K = 32$ ,  $d = 128$ ).

| Sparsity  | Hit@1        | mAP@1        | Hit@5        | mAP@5        | MRR          | MF1@1        |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|
| <b>On</b> | <b>0.773</b> | <b>0.773</b> | <b>0.864</b> | <b>0.689</b> | <b>0.813</b> | <b>0.647</b> |
| Off       | 0.761        | 0.761        | 0.841        | 0.672        | 0.805        | 0.595        |

reported metrics; we retain the  $\ell_1$  penalty in the default configuration.

**Aspect-ratio-preserving normalization.** Table 6 compares aspect-ratio-preserving patch normalization against direct isotropic resizing to  $64 \times 64$ . Preserving aspect ratio improves Hit@1, Hit@5, mAP@5, and MRR. Direct resizing stretches characters to fill the patch uniformly, distorting width-to-height ratios that are discriminative for Hebrew letterforms. We adopt aspect-ratio-preserving normalization as the default preprocessing strategy for both training and retrieval.

#### 4.4. Computational Trade-Offs

Table 7 reports query-time cost. Precomputed matrix lookup is effectively free ( $< 0.01$  ms/query) for all methods. On-the-fly BoB-Hungarian costs  $\approx 10$  ms/query due to the

Table 6. Ablation on component patch normalization for BoB-Hungarian-L2. *Preserved*: aspect-ratio-preserving scale-and-pad. *Stretched*: direct isotropic resize to  $64 \times 64$ .

| Method    | Hit@1        | mAP@1        | Hit@5        | mAP@5        | MRR          | MF1@1        |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|
| Preserved | <b>0.739</b> | <b>0.739</b> | <b>0.841</b> | <b>0.671</b> | <b>0.789</b> | 0.592        |
| Stretched | 0.716        | 0.716        | 0.830        | 0.652        | 0.778        | <b>0.603</b> |

Table 7. Query-time runtime profiling. Precomputed lookup is free for all methods; on-the-fly assignment costs scale with  $K$ .

| Method                                     | ms/query | Scales w/ $N$ |
|--------------------------------------------|----------|---------------|
| BoW-Cosine (precomputed)                   | 0.006    | No            |
| BoB-Hungarian (precomputed)                | 0.006    | No            |
| BoB-Hungarian (on-the-fly)                 | 10.30    | Yes           |
| BoW-Cosine $\rightarrow$ BoB-OT ( $M=30$ ) | 11.33    | No            |

$\mathcal{O}(K^3)$  assignment, making exhaustive online matching impractical for large galleries. The two-stage pipeline (BoW-Cosine  $\rightarrow$  BoB-OT rerank,  $M=30$ ) costs  $\approx 11$  ms/query but scales as  $\mathcal{O}(M \cdot K^3)$  independent of gallery size, providing a practical deployment mode for large collections without requiring a fully precomputed BoB distance matrix.

## 5. Discussion

We have presented Bag-of-Bags (BoB), a manuscript join retrieval framework that replaces a shared global codebook with page-adaptive vocabularies built from sparse autoencoder embeddings of connected-component patches. Pages are compared through set-based distances over these local prototypes, enabling retrieval that is sensitive to page-specific visual structure rather than global frequency statistics alone. We further showed that the mass-weighted BoB-OT variant admits a formal approximation guarantee: its deviation from full component-level optimal transport is bounded by the sum of the two pages’  $k$ -means quantization errors. On our Genizah benchmark, the best variant, BoB-Chamfer, achieves a Hit@1 of 78.4%, a relative improvement of 6.1% over the strongest BoW baseline. Overall, the results show that, for manuscript join retrieval, modeling how a page organizes its local visual modes is more effective than forcing all pages into a single shared vocabulary. More broadly, adaptive set-based representations appear promising for retrieval problems involving partial, noisy, and heterogeneous visual evidence. These results are consistent with the intuition that page-adaptive vocabularies preserve page-specific local structure that may be blurred by shared-codebook frequency representations.

**Chamfer vs. assignment-based distances.** Among BoB variants, Chamfer outperforms both Hungarian and OT despite having weaker formal structure. A plausible explanation

is that fragment pairs often share only a subset of local character patterns: Chamfer’s nearest-neighbor formulation rewards this partial overlap without penalizing unmatched prototypes, whereas one-to-one assignment forces every prototype to contribute regardless of whether it has a meaningful counterpart on the opposite page. This suggests that for partial-document retrieval, soft set matching may be more robust than globally constrained correspondence.

**Limitations.** The evaluation is constrained to a manually annotated subset of 287 images forming 100 join clusters, reflecting the difficulty of obtaining verified, high-confidence ground-truth joins from domain experts. The complete Genizah corpus contains over 250,000 fragments, and the pipeline is designed with that full-corpus setting in mind. We expect the per-image vocabulary construction to remain advantageous at scale: adapting to each fragment’s local character distribution rather than a fixed global codebook becomes more valuable as stylistic diversity increases. Empirical validation on a larger annotated sample remains an important direction for future work, but awaits the accumulation of ground truth. We focus on BoW-family baselines sharing the same sparse encoder so as to isolate the effect of page-adaptive vocabulary construction; comparison against more recent learned retrieval models is another important direction for the future. Extending the pipeline to smaller fragments and more heavily degraded pages is a natural direction for future work. On our benchmark test set,  $K = 32$ , the approximate number of graphemes for those texts, yielded the strongest retrieval metrics, at non-negligible cost compared to the  $K = 20$  used in most of our experiments. It remains to be seen what the ideal number is for the complete corpus. It may in fact be best to form  $K' > K$  clusters and retain only the  $K$  most populated prototypes for matching, decoupling vocabulary richness from matching cost.

It would be interesting to investigate to what extent image-adaptive prototype vocabularies are also beneficial in other computer-vision settings that depend on subtle local visual cues, such as fine-grained recognition. In such problems, instances may be globally similar while differing in small local patterns, and matching may rely on partial correspondences rather than full global appearance agreement. This makes them a natural setting in which to study whether set-to-set matching over image-specific prototype vocabularies might provide an advantage over shared global-codebook representations.

## Acknowledgments

We are grateful to the Friedberg Genizah Project (FGP) and Dr. Roni Shweka for their contribution to the construction of the dataset. This research was funded in part by the European Union (ERC, MiDRASH, Project No. 101071829). Principal investigators: Nachum Dershowitz, Tel Aviv Uni-



versity; Judith Olszowy-Schlanger, EPHE-PSL; Avi Shmidman, Bar-Ilan University; and Daniel Stökl Ben Ezra, EPHE-PSL). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

## References

- [1] A. A. Brink, J. Smit, M. L. Bulacu, and Lambert R. B. Schomaker. Writer identification using directional ink-trace width measurements. *Pattern Recognition*, 45(1):162–171, 2012. [3](#)
- [2] Marius Bulacu and Lambert Schomaker. Automatic handwriting identification on medieval documents. In *14th International Conference on Image Analysis and Processing (ICIAP 2007)*, pages 279–284. IEEE, 2007.
- [3] Marius Bulacu, Lambert Schomaker, and Louis Vuurpijl. Writer identification using edge-based directional features. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR)*, pages 937–941, 2003. [3](#)
- [4] Vincent Christlein, David Bernecker, and Elli Angelopoulou. Writer identification using VLAD encoded contour-Zernike moments. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 906–910. IEEE, 2015. [3](#)
- [5] Vincent Christlein, Martin Gropp, Stefan Fiel, and Andreas Maier. Unsupervised feature learning for writer identification and writer retrieval. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–997. IEEE, 2017. [3](#)
- [6] Shaveta Dargan and Munish Kumar. Writer identification system for Indic and non-Indic scripts: State-of-the-art survey. *Archives of Computational Methods in Engineering*, 26(4):1283–1311, 2019. [3](#)
- [7] Sounak Dey, Anguelos Nicolaou, Josep Lladós, and Uma-pada Pal. Local binary pattern for word spotting in handwritten historical document. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 574–583, Cham, 2016. Springer International Publishing. [2](#)
- [8] Vladislavs Dovgalecs, Alexandre Burnett, Pierrick Tra-nouez, Stéphane Nicolas, and Laurent Heutte. Spot it! Finding words and patterns in historical documents. In *12th International Conference on Document Analysis and Recognition (DAR)*, pages 1039–1043, 2013. [2](#)
- [9] Stefan Fiel and Robert Sablatnig. Writer retrieval and writer identification using local features. In *10th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 145–149. IEEE, 2012. [3](#)
- [10] Stefan Fiel and Robert Sablatnig. Writer identification and writer retrieval using the Fisher vector on visual vocabularies. In *12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 545–549. IEEE, 2013. [3](#)
- [11] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. [5](#)
- [12] Samuel Grieggs, C. E. M. Henderson, Sebastian Sobecki, Alexandra Gillespie, and Walter Scheirer. The paleographer’s eye ex machina: Using computer vision to assist humanists in scribal hand identification. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7216–7225, 2024. [3](#)
- [13] Rajiv Jain and David Doermann. Combining local features for offline writer identification. In *14th International Conference on Frontiers in Handwriting Recognition*, pages 583–588. IEEE, 2014. [3](#)
- [14] Heidi G. Lerner and Seth Jerchow. The Penn/Cambridge Genizah fragment project: Issues in description, access, and reunification. *Cataloging & Classification Quarterly*, 42(1): 21–39, 2006. [2](#)
- [15] Peihua Li, Qilong Wang, and Lei Zhang. A novel earth mover’s distance methodology for image matching with Gaussian mixture models. In *IEEE International Conference on Computer Vision*, pages 1689–1696, 2013. [2](#)
- [16] Isabelle Marthot-Santaniello, Manh Tu Vu, Olga Serbaeva, and Marie Beurton-Aimar. Stylistic similarities in Greek papyri based on letter shapes: A deep learning approach. In *Proceedings of the Document Analysis and Recognition Workshop (DAR)*, pages 307–323, Berlin, Heidelberg, 2023. Springer-Verlag. [2](#)
- [17] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. [2](#)
- [18] Ravi Shekhar and C. V. Jawahar. Word image retrieval using bag of visual words. In *10th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 297–301, 2012. [2](#)
- [19] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003. [2](#)
- [20] Eugenio Vocaturo and Ester Zumpano. Assembling fragments of ancient papyrus via artificial intelligence. In *Pervasive Knowledge and Collective Intelligence on Web and Social Media*, pages 3–13, Cham, 2023. Springer Nature. [2](#)
- [21] Lior Wolf, Rotem Littman, Naama Mayer, Tanya German, Nachum Dershowitz, Roni Shweka, and Yaacov Choueka. Identifying join candidates in the Cairo Genizah. *International Journal of Computer Vision*, 94(1):118–135, 2011. [2](#)
- [22] Yu-Jie Xiong, Ying Wen, Patrick SP Wang, and Yue Lu. Text-independent writer identification using SIFT descriptor and contour-directional feature. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 91–95. IEEE, 2015. [3](#)