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

CVPR DENVER COLORADO JUNE 3-7, 2026

Funded by the European Union
European Research Council
Established by the European Commission

Agenda

- Cairo Genizah
- Motivation & Background
- Limitations of Traditional Methods
- Annotated Dataset
- Bag of Bags Representation
- Results & Ablations
- Conclusion and Future Work

Cairo Genizah

“It is a battlefield of books, and the literary production of many centuries had their share in the battle, and their disjecta membra are now strewn over its area"   -   Solomon Schechter

The Cairo Genizah

- Collection of ~400,000 medieval manuscript fragments (11th–19th CE.)
- Written in Hebrew, Aramaic, and Judeo-Arabic scripts
- Now dispersed across 50+ libraries and private collections worldwide
- Scholars manually identify “joins” between fragments from the same manuscript
- Manual join discovery is expert-driven, slow, and does not scale
- Automated retrieval can prioritize candidate joins for human experts

Motivation: Reconstructing Fragmented Manuscripts

Join: A set of manuscript fragments originating from the same physical manuscript, now separated and held across different institutions.

Retrieval Task: Given a query image of a fragment, rank all other fragments so that those from the same manuscript appear at the top.

Query image

Severe Damage

Fragments are damaged, stained, incomplete, and torn

Subtle Cues

Same-manuscript signal is in fine-grained handwriting style

Global Similarity

Different manuscripts share aging, texture, degradation

Limitations of Standard Approaches

- Standard retrieval architectures frequently rely on Bag of Words (BoW) paradigms, mapping local descriptors to a shared global codebook.

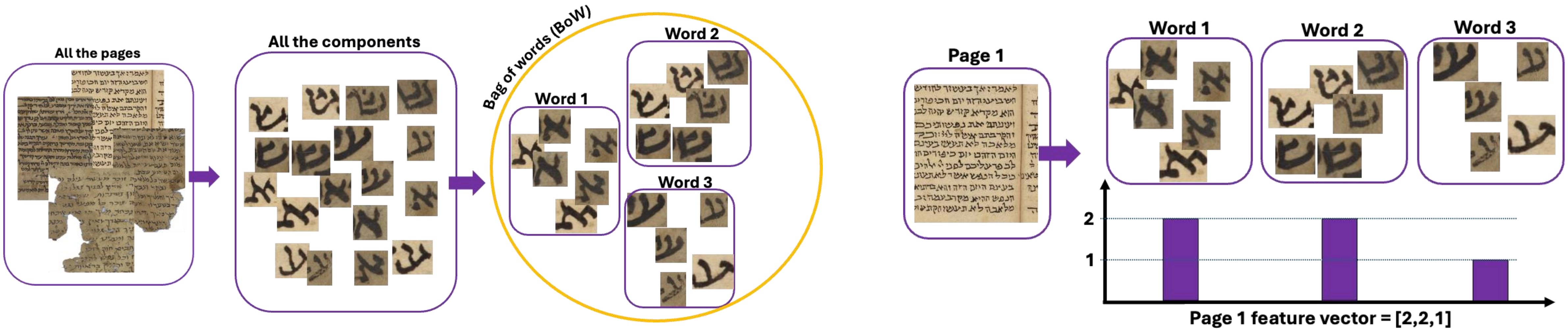
- Global quantization causes critical image-specific handwriting style information to be lost.

- Two images with identical codeword frequencies but entirely different geometric character styles are given a distance of zero by BoW- yet maybe written by entirely different hands.

- Differences in lighting and layout can cause BoW models to miss fragments from the same manuscript.

Background: Bag of Words (BoW)

- Classical BoW maps all fragments to histograms over a shared global codebook
- Classical BoW represents each image as a histogram over a shared global visual codebook
- Local patches/components are assigned to their nearest global “visual word”
- The final vector records how often each visual word appears and discards the geometry and page-specific structure of local handwriting patterns



Key Contributions

1. Propose **Bag of Bags**, a fragment-specific representation for manuscript retrieval.
2. Build local vocabularies from connected-component patch embeddings. Use a sparse convolutional autoencoder to encode character-like patches.
3. Compare fragments using set-to-set distances between local vocabularies.
4. Show improved retrieval over pooling and strong Bag-of-Words baselines.
5. Comprehensive Ablation Study showing BoB's gains are structural.

Annotated Dataset

- Manually verified join 100 join clusters consisting of 287 Cairo Genizah manuscript fragment images
- Cluster sizes range from 2 to 9 fragments
- Images are drawn from multiple institutions, including Cambridge, JTS, British Library, and Alliance Israelite Universelle

Example of a Join

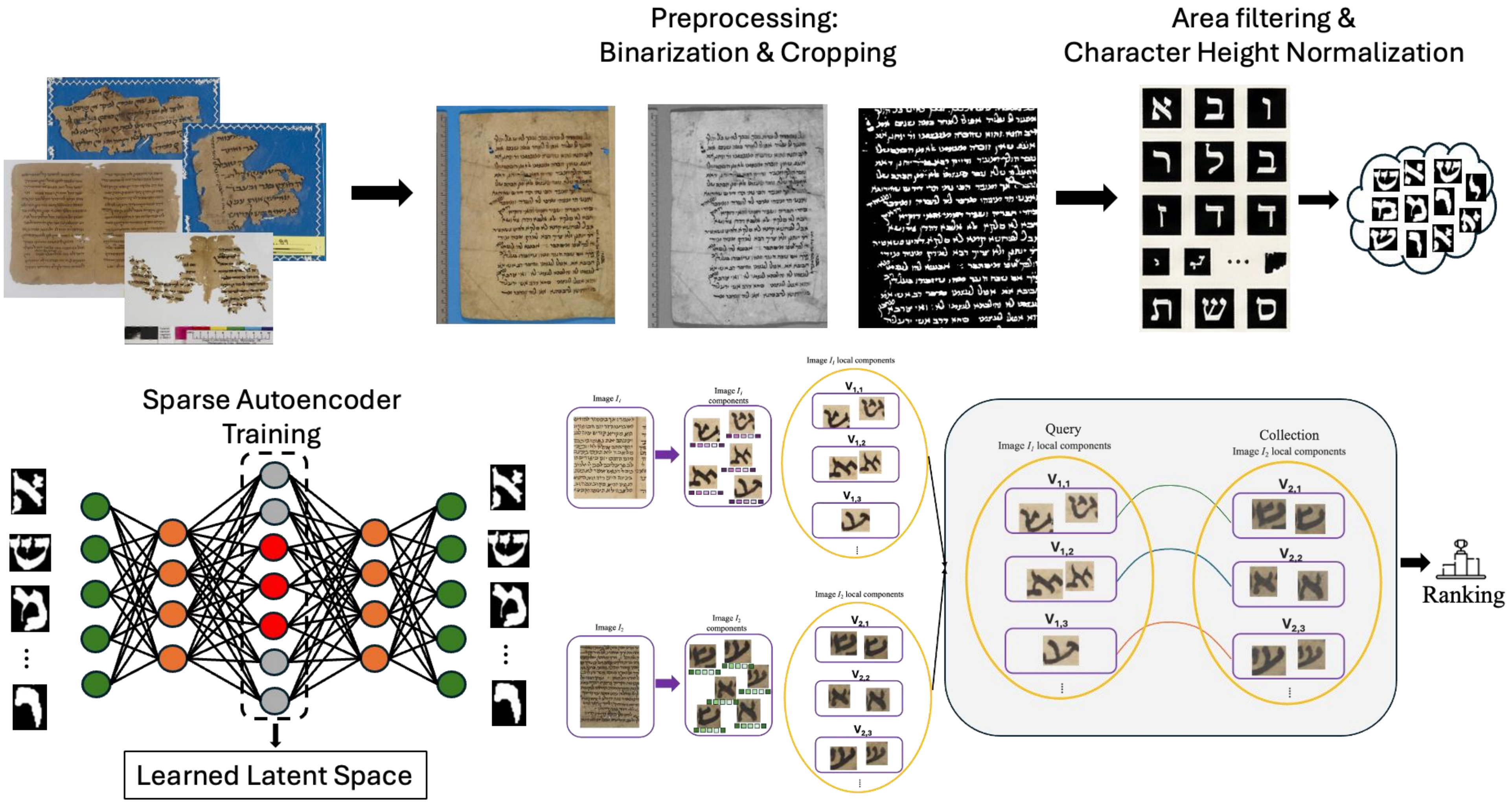
Jewish Theological Seminary
Ms. ENA 1492.21

Hungarian Academy Sciences
Ms. Kaufmann GEN 441

Jewish Theological Seminary
Ms. ENA 2070.1

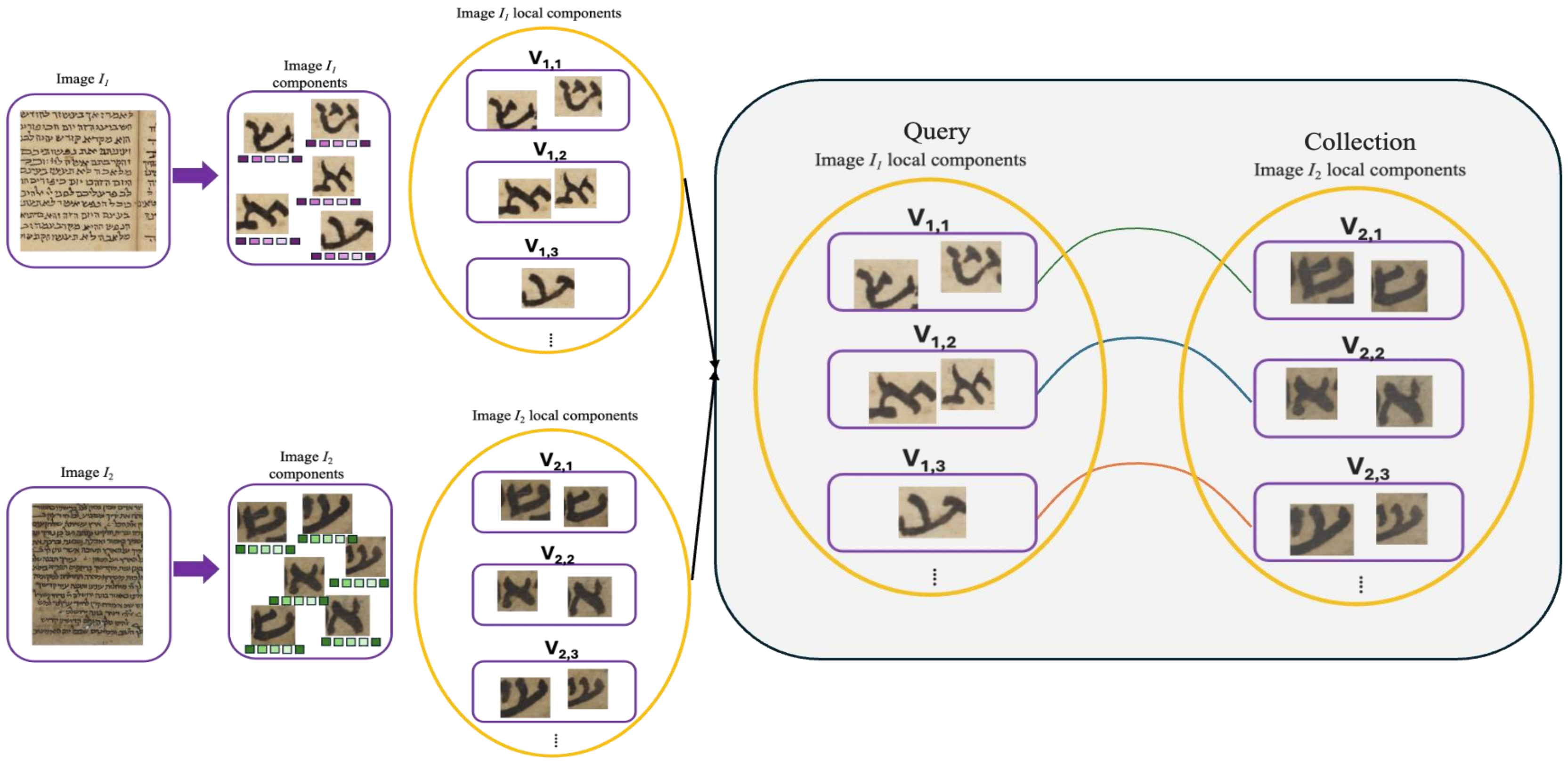
Cambridge University Library
Ms. T-S Ar. 48.19

Overview of the Pipeline



The Bag of Bags Representation

- For each fragment, collect all patch embeddings
- Run per-fragment k-means clustering
- Cluster centroids become local visual words; Cluster populations become prototype masses
- Final representation: B(I) = {(prototype, mass)}

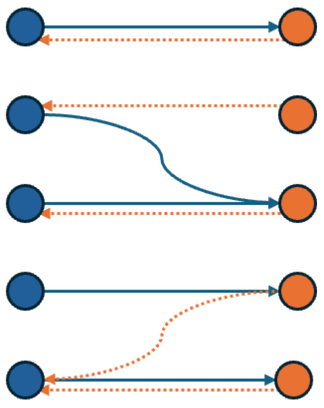


Set-to-Set Distance Formulations

BoB-Chamfer

Soft bidirectional nearest-neighbor matching
$O(K^2)$

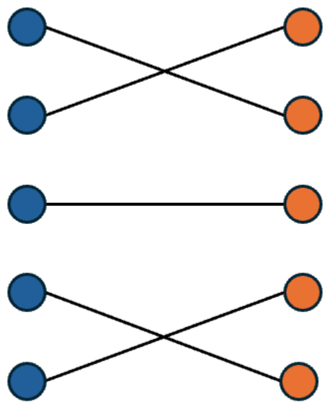
- Each prototype finds its nearest match
- Rewards partial overlap; robust to damaged fragments



BoB-Hungarian

Strict one-to-one bipartite assignment
$O(K^3)$

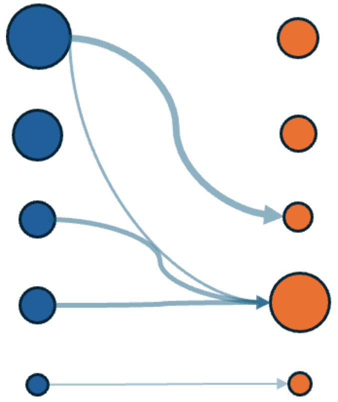
- Each prototype finds its nearest match
- Rewards partial overlap; robust to damaged fragments



BoB-OT (Mass Weighted)

Mass weighted optimal transport
$O(K^3)$

- Each prototype finds its nearest match
- Rewards partial overlap; robust to damaged fragments



● = prototypes from *Page I* ● = prototypes from *Page J*

Results: Retrieval Performance

Method	Hit@1	Hit@5	mAP@5	MRR	Macro-F1@1
BoW-Centroids-Cosine	0.61	0.84	0.42	0.70	0.55
BoW-Centroids-χ^2	0.65	0.85	0.45	0.73	0.59
BoW-RawPatches-Cosine	0.72	0.90	0.53	0.79	0.67
BoW-RawPatches-χ^2	0.74	0.91	0.55	0.80	0.68
BoB-Hungarian	0.68	0.89	0.49	0.76	0.62
BoB-OT	0.72	0.91	0.53	0.80	0.67
BoB-Chamfer	**0.78**	**0.93**	**0.60**	**0.84**	**0.73**

- **BoB-Chamfer achieves 6.1% relative improvement in Hit@1 over best BoW baseline**
- BoB-OT matches BoW-RawPatches-χ^2 while adding mass-aware matching + formal guarantees
- BoW-RawPatches consistently outperforms BoW-Centroids → global codebook from raw embeddings is a stronger baseline

Results: Retrieval Performance

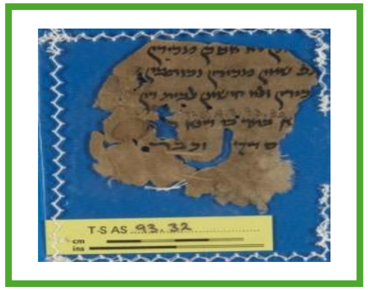
Query

Top 1

Top 2

Top 3

Top 4



Top 5

Query

Top 1

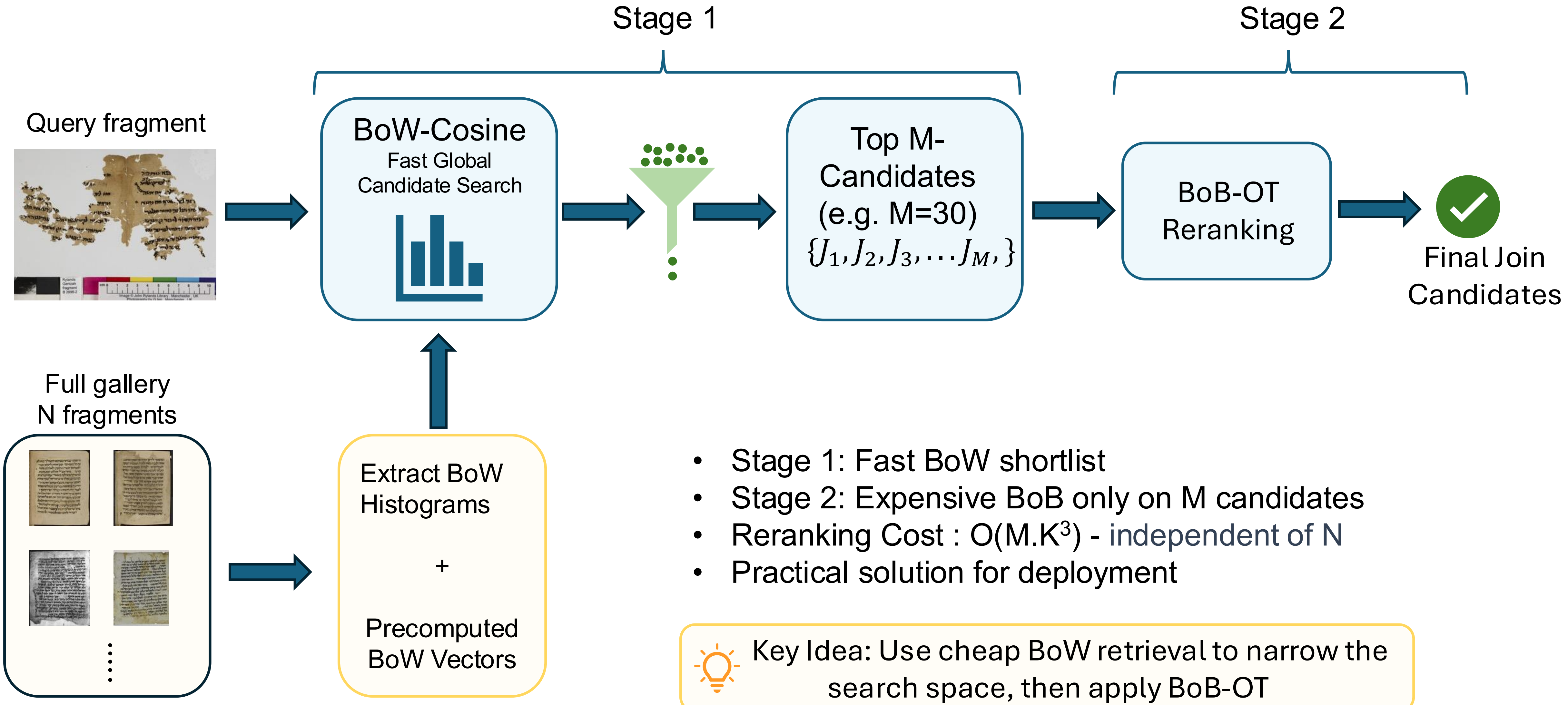
Top 2

Top 3

Top 4

Top 5

The Two-Stage Retrieval Pipeline



- Stage 1: Fast BoW shortlist
- Stage 2: Expensive BoB only on M candidates
- Reranking Cost : O(M.K^3) - independent of N
- Practical solution for deployment

Key Idea: Use cheap BoW retrieval to narrow the search space, then apply BoB-OT

Ablation Results

Local Vocabulary Size K

K=20 is optimal. Smaller K underrepresents style diversity; larger K overfits to noise. BoB gains are not just from richer k-means — structure matters.

Latent Dimension d

d=128 best balances representation power and clustering quality. d=64 is too compressed; d=256 overfits on this dataset size.

Activation Sparsity (λ)

$\lambda=10^{-5}$ optimal. Sparsity improves cluster separability; too much sparsity degrades reconstruction and downstream retrieval.

Aspect-Ratio Normalization

Preserving aspect ratio during patch normalization yields +2–3pp Hit@1 over naive resize. Character geometry is informative.

Conclusion

- Manuscript join retrieval requires matching subtle local visual evidence.
- Global BoW histograms can discard page-specific handwriting structure.
- Bag of Bags builds adaptive vocabularies for each fragment.
- Set-to-set matching improves retrieval over strong BoW baselines.
- Chamfer distance is especially effective for partial, damaged fragments.
- Adaptive local vocabularies are promising for fine-grained visual retrieval.

Future Work

- Evaluation uses a small manually annotated benchmark. Ground-truth joins are expensive and require expert validation. Larger-scale validation on the full Genizah corpus remains future work.
- Explore hybrid systems combining BoB with learned global descriptors
- Adapt the method to smaller and more degraded fragments.
- Test adaptive vocabularies in other fine-grained retrieval domains

Thank You!
☺

Sharva Gogawale email: sharvag@mail.tau.ac.il

Prof. Nachum Dershowitz (PI) email: nachumd@tauex.tau.ac.il

Code: github.com/TAU-CH/midrash_bob

Bag of Bags

TAU-CH Lab