A Spatially-Aware Search Engine for Textual Content in Images

PRANAV RAMESH, Harvard University, USA
MOHAMED ZIDAN CASSIM, Harvard University, USA
GIOVANNI D'ANTONIO, Harvard University, USA

Standard image search engines often treat text within images as secondary metadata or ignore its spatial location. This limits users' ability to find images based on text appearing in specific visual areas. We present a spatially-aware textual image search engine designed to address this limitation. Our approach utilizes an inverted index mapping text n-grams to their normalized bounding box coordinates within images. Queries consist of text and an optional spatial region. Relevance scoring combines spatial factors (Intersection over Union - IoU and proximity) with n-gram length, weighted according to configurable parameters. To facilitate development and evaluation, we developed a pipeline for generating synthetic datasets with controlled text placement and ground truth. We evaluated our system against non-spatial baselines (keyword-only and n-gram-only) using Mean Average Precision (MAP) and Precision@k (P@k) on this synthetic data. Results demonstrate statistically significant improvements in ranking quality for both n-gram usage over keywords (MAP 0.21 vs 0.03) and spatial awareness over n-grams alone (MAP 0.67 vs 0.21), validating the effectiveness of incorporating both n-grams and spatial context. A visualization tool was also developed to aid in understanding search results.

CCS Concepts: • **Information systems** → **Information retrieval**; **Retrieval tasks and goals**; *Indexing*; Search engine architectures and scalability; • **Computing methodologies** → *Computer vision*; Document analysis and representation; Image processing.

Additional Key Words and Phrases: image search, text localization, spatial search, n-grams, OCR, information retrieval

ACM Reference Format:
Pranav Ramesh, Mohamed Zidan Cassim, and Giovanni D'Antonio. 2025. A Spatially-Aware Search Engine for Textual Content in Images. 1, 1 (April 2025), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Problem Statement

Images frequently contain rich textual information, such as signs, labels, headlines, logos, or embedded text in documents and screenshots. Traditional image search systems primarily focus on visual features or global textual tags, often failing to leverage the specific content and location of text within the image. Users cannot easily query for images containing specific text within a particular visual region (e.g., "find photos with 'SALE' in the top-right corner" or "show screenshots where 'error message' appears near the bottom"). This lack of spatial awareness limits the precision and utility of text-based image retrieval.

Authors' Contact Information: Pranav Ramesh, pranavramesh@college.harvard.edu, Harvard University, Cambridge, MA, USA; Mohamed Zidan Cassim, mzcassim@college.harvard.edu, Harvard University, Cambridge, MA, USA; Giovanni D'Antonio, giovannidantonio@college.harvard.edu, Harvard University, Cambridge, MA, USA.

Manuscript submitted to ACM

1.2 Motivation

The ability to search for text within specific spatial regions of images unlocks numerous applications across multiple domains. Document analysis benefits by enabling users to find specific sections or figures in scanned documents based on headings or captions in known layout areas. In retail and e-commerce, such technology facilitates locating product images where price or discount tags appear in particular locations relative to the product. Scene understanding applications can identify street signs, shop names, or specific labels within photographs of complex scenes with greater accuracy when spatial relationships are considered. UI/UX researchers can leverage spatial text search to analyze screenshots and find instances where specific labels or error messages appear in certain interface elements. Additionally, accessibility is enhanced by enabling visually impaired users to query not just what text appears in an image, but where it is located.

To illustrate the practical utility more concretely, consider a case study in **Automated Data Entry from Scanned Documents**. Imagine processing a high volume of scanned invoices or receipts for accounting. While Optical Character Recognition (OCR) can extract all text from an image, accurately identifying the *semantic role* of specific text snippets (e.g., distinguishing the "Total Amount" from line item prices, or the "Invoice Date" from a "Payment Due Date") is a significant challenge due to the wide variety of document layouts. A traditional keyword search for terms like "Total" or for date patterns might yield multiple ambiguous candidates scattered across the document.

This is where spatially-aware search offers a distinct advantage. By leveraging common layout conventions, the system can target queries to specific regions. For example, a query searching for text matching a monetary amount pattern (e.g., `\$\d+\.\d{2}`) *specifically within the normalized coordinates corresponding to the bottom-right quadrant* of the document is highly likely to isolate the final **Total Amount**. Similarly, querying for date-like text primarily within the *top-right quadrant* can reliably extract the **Invoice Date**. This targeted spatial querying drastically reduces ambiguity compared to context-agnostic text search, significantly simplifying the development of robust automated data entry pipelines without requiring complex template matching or sophisticated layout analysis models for every document variant. It demonstrates how spatial awareness can act as a powerful heuristic for semantic disambiguation in structured documents.

Existing methods often rely on whole-image tags or complex scene understanding models that may not precisely capture localized text queries. A dedicated system focusing on spatial text search promises higher precision and user control for these tasks.

1.3 Proposed Solution & Contributions

To address the limitations of traditional methods, we propose and implement **SATIAS (Spatially-Aware Textual Image Search)**, a search engine designed to retrieve images based not only on *what* text they contain but also *where* that text is located. The core idea is to move beyond simple keyword matching by creating an index that explicitly links textual content (represented as n-grams, typically sequences of 1 to 3 words) to its precise spatial location within each image. This is achieved by building an inverted index where keys are text n-grams and values are lists of occurrences, each storing the `image_id` and the n-gram's **normalized bounding box** coordinates (percentages of image width/height) to ensure scale and aspect-ratio invariance. User queries can specify both a `query_text`, which is parsed into n-grams, and an optional target **spatial region**, also represented as a normalized bounding box. Candidate images containing matching n-grams are retrieved via the index, and each potential match is evaluated using a novel, configurable scoring mechanism. This scoring combines **spatial relevance**—calculated as a weighted sum of **Intersection over Union**

(IoU) for overlap and **centroid proximity** for nearness between the query region and the n-gram box—with **textual relevance**, where matches involving longer n-grams contribute more significantly. The relative importance of IoU versus proximity can be tuned via configuration weights. Finally, scores are aggregated per image, and the results are ranked to provide the user with images where the desired text appears in the specified location.

This paper details the design, implementation, and rigorous evaluation of the SATIAS system. Our primary contributions include: (1) the **novel algorithm design** itself, particularly the use of normalized coordinates for indexing and the hybrid spatial scoring function combining weighted IoU and proximity; (2) a robust and parallelized **synthetic data generation pipeline** (`data_generation/`) that creates large datasets with precise ground-truth bounding boxes and targeted queries, crucial for controlled offline evaluation independent of OCR errors; (3) a modular Python **system implementation** (`search_engine/`) encompassing indexing, flexible query parsing, spatial calculations, and search logic; (4) a **rigorous quantitative evaluation** framework (`evaluate.py`) using MAP@k and P@k metrics to compare SATIAS against keyword-only and n-gram-only baselines on the synthetic data, including statistical significance testing (Wilcoxon signed-rank test); and (5) an **interactive visualization tool** (`visualize_search.py`) with a GUI that allows users to execute searches and inspect results with overlays showing query regions and color-coded n-gram bounding boxes based on IoU, aiding qualitative analysis and debugging.

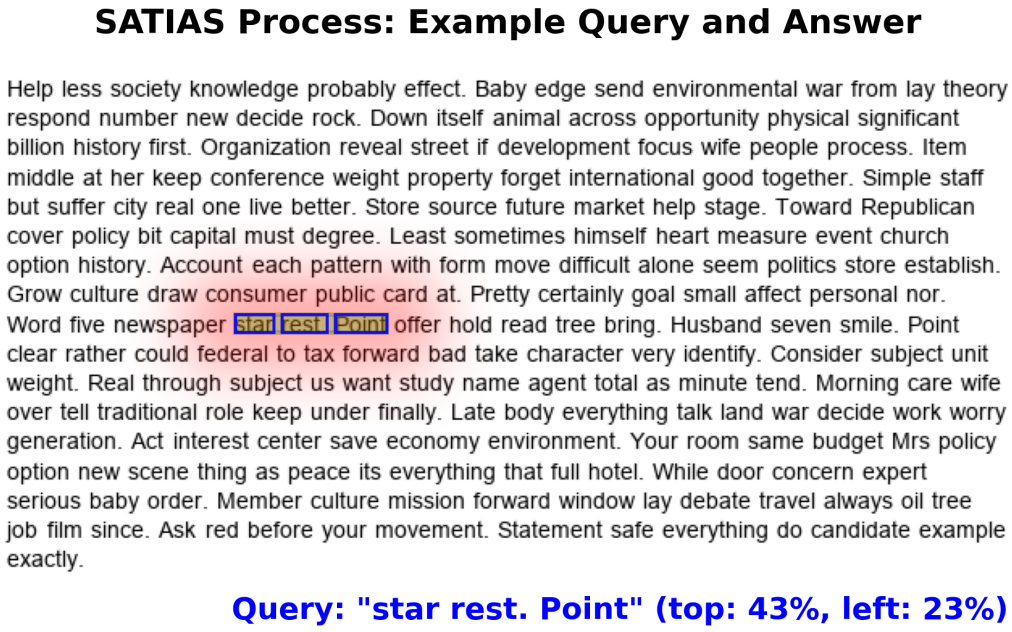

Fig. 1. Illustration of the SATIAS process. The blue boxes indicate the n-gram query region, the heat overlay shows the answer area, and the query text at bottom right demonstrates a location-aware search.

2 Prior Work

The challenge of searching for textual content within images, potentially constrained by location, has been explored from various perspectives. Our work draws upon foundational concepts while offering a specific, geometrically focused solution.

2.1 Foundational and Explicit Spatial Methods

Foundational Text-in-Image Search: Early research, such as that by Manmatha et al. (UMass CIIR, 2000) [6], focused on the fundamental problems of detecting, extracting (via OCR), and indexing text found within images for keyword-based retrieval. These systems laid the groundwork but typically treated text as document-level metadata or lacked mechanisms for precise spatial querying. Our system builds on this by explicitly indexing the *location* of extracted text (n-grams) and enabling queries against these locations.

Spatial-Semantic Approaches: More recent work has integrated spatial reasoning with semantic understanding. Mai et al. (CVPR 2017) [5] proposed a spatial-semantic image search framework where users define semantic layouts on a canvas, and a CNN synthesizes corresponding visual features for retrieval. This differs from our approach, which focuses narrowly on matching the precise geometric location (bounding box) of specific text n-grams provided in the query, rather than interpreting broader semantic layouts.

2.2 The Rise of MLLMs in Spatial Grounding

Recent years have seen a significant shift towards utilizing Multimodal Large Language Models (MLLMs) for tasks involving spatial grounding. These models, such as KOSMOS-2 [4] and Groma [8], aim to integrate visual perception, language comprehension, and spatial reasoning within unified architectures, often trained on web-scale datasets [4].

Approach: Instead of explicit geometric indexing and scoring like our system, MLLMs typically handle spatial information implicitly through learned mechanisms. These mechanisms include using location tokens where continuous bounding box coordinates are discretized into special tokens integrated into the language model's vocabulary [4], leveraging cross-modal attention mechanisms to learn correlations between text and image regions, and utilizing joint embedding spaces that align visual regions and textual descriptions, implicitly encoding spatial relationships.

Comparison to Our System: Compared to our system, MLLMs offer distinct advantages and disadvantages. MLLMs possess strong semantic understanding derived from their underlying LLMs, enabling them to handle synonyms, paraphrasing, and complex natural language queries describing spatial relations (e.g., "the book to the left of the lamp") [4, 8], a capability our exact n-gram matching system lacks. However, the reasoning process of MLLMs is often opaque ("black box"), whereas our system, using explicit IoU and proximity calculations, offers greater interpretability and direct control via tunable weights. Additionally, training state-of-the-art MLLMs for grounding requires massive datasets (like the GRIT dataset used for KOSMOS-2 [4]) and significant computational resources for pre-training and fine-tuning [8]. Our indexing approach, while requiring OCR, may have different data dependencies, primarily needing the images themselves rather than extensive grounded text-region pairs for initial training. This trend towards MLLMs highlights a different paradigm for spatial understanding, trading explicit geometric control for learned semantic richness and query flexibility, albeit with associated challenges in interpretability and data requirements.

2.3 Enabling Technologies

Text Spotting: Accurate detection and bounding box generation are critical prerequisites for any text-in-image search system. The field has advanced to handle arbitrary text shapes using segmentation, contour embedding, Bezier curves (ABCNet), Mask R-CNN, or sequential deformation. However, bounding box inaccuracy remains a challenge for real-world geometric scoring.

Indexing: Scalability requires efficient indexing structures. While our approach uses an in-memory inverted index, spatial databases traditionally use R-Trees/Quadtrees, often combined with inverted indexes in hybrid structures. Recent Learned Sparse Retrieval (LSR) methods (e.g., STAIR [3], Cao et al. [2], Bai et al. [1]) map dense embeddings to sparse lexical vectors compatible with inverted indexes, offering a promising direction for scalable multimodal retrieval.

2.4 Contributions

Our work occupies a niche focused on precise, spatially constrained retrieval of specific text n-grams. Compared to the prior work, our contributions are:

(1) The use of an efficient inverted index mapping n-grams directly to normalized bounding boxes
(2) A tunable spatial scoring function explicitly combining geometric overlap (IoU) and centroid proximity, offering direct control over spatial relevance criteria
(3) A dedicated synthetic data generation pipeline and evaluation methodology designed to rigorously assess the performance of spatial text localization, isolating it from OCR errors and providing targeted spatial query scenarios

Our approach provides a simple, interpretable method for precise geometric localization of exact text n-grams within rectangular regions. Its strengths are direct geometric control and the synthetic data pipeline for evaluation. Key limitations include dependence on OCR accuracy, lack of semantic understanding (unlike VSE/attention models), limited query expressiveness (compared to canvas/trace/relational queries), and scalability issues addressed by spatial/hybrid/LSR indexing. It represents a valuable baseline but stands apart from dominant deep learning trends emphasizing semantics and learned alignments.

2.5 Comparison of Approaches

Table 1 provides a comprehensive comparison of various spatially-aware image-text retrieval approaches, highlighting the distinctive positioning of our system among existing methods.

3 Methodology

Our system comprises two main phases: offline indexing and online query processing/search.

3.1 Core Algorithm Overview

The system first preprocesses a collection of images (or uses pre-computed metadata in our synthetic case) to build an inverted index. This index maps text n-grams to a list of all locations (image ID and normalized bounding box) where they appear. During online search, a user query (text + optional region) is processed. N-grams are extracted from the query text. The inverted index is used to retrieve candidate image locations matching these n-grams. Each match is scored based on n-gram length and spatial relevance relative to the query region. Scores are aggregated per image, and results are ranked.

Table 1. Comparison of Spatially-Aware Image-Text Retrieval Approaches

Approach	Query Type	Key Characteristics	Strengths/Weaknesses
Foundational Text-in-Image (Manmatha et al., 2000)	Keywords	Text as document metadata; Inverted Index; Keyword matching	(+) Established base approach (-) No spatial awareness
Explicit Spatial-Semantic (Mai et al., 2017)	Text-boxes on canvas	User-defined layout; Visual Feature Index; Feature similarity	(+) Flexible canvas input (-) Less precise text matching
VSE / Attention Models	Text Query	Implicit spatial via embeddings; Learned attention mechanisms	(+) Strong semantic understanding (-) No explicit spatial queries
MLLMs (KOSMOS-2, Groma)	Natural Language	Learned mechanisms; End-to-end approach	(+) Semantic flexibility (-) Black-box reasoning
Our Approach	N-grams + Optional Region	Normalized Bounding Boxes; Explicit IoU + proximity scoring	(+) Interpretable; Precise (-) Limited semantics; OCR dependent

3.2 Indexing Phase

Objective. Create an efficient lookup structure for n-gram occurrences and their spatial locations.

Process. The indexing phase creates a mapping between textual content and its spatial location across all images in the collection. It begins with the processing of image metadata, which can be derived from OCR output or, in our case, from pre-calculated metadata containing words, n-grams, and their precise pixel bounding boxes. For each n-gram in each image, the system normalizes its pixel bounding box coordinates from absolute pixel values `[t, l, b, r]` to percentage-based coordinates `[norm_t, norm_l, norm_b, norm_r]` relative to the image dimensions. This normalization step is crucial as it ensures that spatial comparisons remain consistent across images of varying resolutions and aspect ratios, enabling reliable spatial querying across diverse image sources.

The core data structure employed is an inverted index mapping n-grams to their occurrences. Each key in this index is a text n-gram string, and the corresponding value is a list of occurrences, where each occurrence contains an image identifier and normalized bounding box coordinates. As the system processes each n-gram from the input data, it appends a new entry containing the image identifier and normalized bounding box to the list associated with that n-gram text. Once constructed, the entire index is serialized and saved to persistent storage, allowing efficient loading in subsequent search sessions without rebuilding the index.

3.3 Query Processing

Objective. Convert user input into a format suitable for searching the index.

Process. The query processing stage transforms raw user input into structured data that can be efficiently matched against the inverted index. This process handles two key components: the query text and an optional spatial region specification. For the textual component, the system divides the input query text into individual words and generates all possible n-grams within the configured range (from the minimum to maximum n-gram length). This n-gram extraction mirrors the approach used during indexing, ensuring consistency between indexed content and query terms.

For the spatial component, the system parses an optional region string parameter, which can specify a target area within images using various formats (e.g., `"top: 10-30, left: 50-70"`). This parsing interprets different notation styles for specifying top, left, bottom, and right boundaries or ranges as percentages of image dimensions. The parser performs validation on these values and handles edge cases gracefully. If the spatial region string is missing, invalid, or cannot be parsed, the system defaults to using the full image area represented as normalized coordinates `[0.0, 0.0, 100.0, 100.0]`, effectively conducting a whole-image search. After processing both components, the function returns the complete query representation: a list of extracted query n-grams and a single normalized query bounding box ready for the search phase.

3.4 Search and Ranking

Objective. Retrieve and rank images based on textual and spatial relevance.

Process. The search process begins with the initialization of an image score accumulator that will track the relevance score for each candidate image. For each query n-gram, the system performs a lookup in the inverted index to retrieve all matching occurrences across the indexed images. When a query n-gram is found, the algorithm iterates through each occurrence, represented as a tuple of image identifier and normalized bounding box, and calculates its contribution to the overall image score.

The scoring mechanism first determines the appropriate spatial relevance component. For baseline non-spatial searches or when the query doesn't specify a region of interest (using the default full-image bounding box), this component is set to 1.0, effectively ignoring spatial factors. However, when processing spatially-aware queries with specific target regions, the system employs a sophisticated dual-metric approach. It calculates the Intersection over Union (IoU) between the query region and the n-gram bounding box, which quantifies the degree of overlap. Simultaneously, it computes a proximity score based on the distance between centroids of the query and n-gram regions, using an exponential decay function that rewards closer matches. These two metrics are then combined into a single spatial score using configurable weights (typically equal weights of 0.5 each). This hybrid approach effectively addresses the limitations of using either metric in isolation—IoU fails to reward nearby non-overlapping matches, while proximity alone would ignore the extent of overlap and relative sizes.

The algorithm then weights this spatial relevance by the n-gram length, recognizing that longer matching phrases should contribute more significantly to relevance than shorter ones. This weighted score is added to the accumulated score for the corresponding image. After processing all query n-grams and their occurrences, the system sorts the image scores in descending order and returns a ranked list of images with their relevance scores, representing the most relevant images for the given query.

This scoring approach provides a balance between textual and spatial relevance, with the configurable weights offering flexibility to adjust the importance of exact overlap versus general proximity based on specific application needs. The n-gram length weighting further enhances discrimination, favoring images containing more specific, longer matching phrases over those with only short, potentially more ambiguous matches.

3.5 Synthetic Dataset Generation

Motivation. Generating a synthetic dataset provided several key advantages for our research. The approach allowed us precise control over text content, layout, and repetition within the images. We were able to obtain perfect pixel-level bounding boxes for every word and n-gram, effectively eliminating OCR errors as a confounding variable during algorithm development. This dataset generation also facilitated automatically creating queries with known ground truth target images and specific spatial relationships (overlap, proximity, etc.) to systematically test different scoring scenarios. Additionally, the synthetic approach offered scalability by efficiently generating large datasets (thousands of images, tens of thousands of queries) using parallel processing.

Process. The synthetic dataset generation process began with centralized configuration controlling parameters like image dimensions, number of images, font settings, text density, repetition control, word distinctiveness, n-gram range, and query generation. The core logic first created a pool of unique sentences to ensure controlled repetition of words and phrases across different images. For each image, we created a blank canvas and selected a random subset of sentences from the pool. We probabilistically injected specific test phrases (e.g., "special offer") multiple times at random locations within the selected text, and replaced some common words with more distinctive words to aid later visual inspection and analysis. Words were then drawn onto the image sequentially (top-down, left-right), handling line wrapping based on margins and word width, with text allowed to bleed off the bottom edge to ensure full vertical coverage. Crucially, we calculated the precise pixel bounding box `[top, left, bottom, right]` for each individual word before drawing and stored this information.

After generating the words and their positions, we calculated all n-grams within the configured range (e.g., 1 to 3 words) for each image. For each n-gram, we determined the union bounding box (in pixels) based on the exact pixel bounding boxes of its constituent words. The query generation process then created a set number of queries for each image by selecting a random n-gram already placed in that image as the textual target, with its location serving as the ground truth. We randomly chose a query region type based on the configured distribution (e.g., No Region, Exact Match, High IoU, Low IoU, Nearby, Distant) and generated a corresponding normalized query region based on the target n-gram's location and the chosen type. To maximize efficiency, we parallelized the generation process across multiple CPU cores.

Output. The generation pipeline produced several essential outputs: the synthetic images themselves, comprehensive metadata containing details about each image and lists of all words and n-grams within it along with their exact pixel bounding boxes, and a queries dataset containing all generated queries. Each query record includes a query identifier, ground truth image identifier, query text, normalized target region coordinates, ground truth bounding boxes for the text, and information about how the query region was created relative to the target text. This structured output provided all necessary information for training and evaluating our spatial search algorithms.

4 Evaluation

We conducted a quantitative evaluation to assess the performance of the spatially-aware search engine compared to relevant non-spatial baselines, using the generated synthetic dataset.

Table 2. Performance Comparison of Search Methods

Metric	Spatial N-gram	N-gram Baseline	Keyword Baseline
MAP@10 (Max: 1.0)	0.6711	0.2110	0.0294
P@10 (Max: 0.1)	0.0795	0.0324	0.0054

4.1 Evaluation Setup

The evaluation utilized a large synthetic dataset consisting of 50,000 queries derived from 2,000 generated images. For each query, we defined the single "relevant" image as the one specified in the query record—specifically, the image from which the query's target n-gram was originally sampled. All other images were considered non-relevant for that query. We compared three distinct configurations in our evaluation: (1) the full Spatial N-gram algorithm, which combines spatial scoring (IoU + proximity with balanced weights, typically 0.5/0.5) and n-gram length weighting; (2) an N-gram Baseline that ignores the query region and effectively ranks based only on text match (n-gram presence) and n-gram length to isolate the effect of n-grams compared to simple keywords; and (3) a Keyword Baseline representing a rudimentary non-spatial search that breaks the query text into unique words and scores images based simply on the count of matching words found anywhere in the image, ignoring both n-grams and location to serve as a fundamental baseline. All metrics were calculated using a cutoff of k=10.

4.2 Evaluation Metrics

Our evaluation employed several standard information retrieval metrics to comprehensively assess the effectiveness of the retrieval approaches. Mean Reciprocal Rank (MRR) measured the average of the reciprocal ranks of the first relevant result across all queries, providing insight into how quickly a user would find the first relevant result. Mean Average Precision (MAP) calculated the mean of the average precision scores for each query, offering a single-figure measure of quality across recall levels. Precision at k (P@k) measured the proportion of relevant documents in the top k results, directly reflecting precision from a user's perspective when viewing a limited number of results. We also calculated recall at various cutoffs to measure the fraction of relevant documents retrieved, and normalized Discounted Cumulative Gain (nDCG) to evaluate the usefulness of the ranking based on the graded relevance of results. Additionally, we computed Success@k, representing the percentage of queries where at least one relevant document appeared in the top k results. For this application, Success@1 was particularly significant as it indicated the proportion of queries where the algorithm successfully retrieved the exact desired image as the top result.

4.3 Results

The evaluation was conducted on a dataset containing 50,000 queries with a cutoff of k=10. Table 2 summarizes the performance metrics for all three approaches.

Statistical Analysis. To assess the significance of these results, we performed pairwise Wilcoxon signed-rank tests on the Average Precision (AP) scores for each query. The following p-values were obtained:

- **Spatial N-gram vs. N-gram Baseline:** p-value = 0.0000
- **N-gram Baseline vs. Keyword Baseline:** p-value = 0.0000
- **Spatial N-gram vs. Keyword Baseline:** p-value = 0.0000

All p-values are less than 0.0001, indicating that the observed differences are highly statistically significant. The 95% confidence intervals for MAP@10 were [0.669, 0.673] for Spatial N-gram, [0.209, 0.213] for N-gram Baseline, and [0.029, 0.030] for Keyword Baseline. For P@10, the 95% confidence intervals were [0.0789, 0.0801] for Spatial N-gram, [0.0321, 0.0327] for N-gram Baseline, and [0.0052, 0.0056] for Keyword Baseline.

In summary, the Spatial N-gram approach achieved a MAP@10 of 0.6711 (95% CI: [0.669, 0.673]) and a P@10 of 0.0795 (95% CI: [0.0789, 0.0801]), both of which are significantly higher than the N-gram Baseline (MAP@10: 0.2110, P@10: 0.0324) and the Keyword Baseline (MAP@10: 0.0294, P@10: 0.0054), with all pairwise differences being highly statistically significant ($p < 0.0001$).

5 Visualization Tool

To complement the quantitative evaluation, an interactive GUI tool was developed using Tkinter and Pillow. This tool allows users to enter query text and specify spatial regions using percentage inputs, execute searches using the implemented backend, and view ranked results (Top, Middle, and Last sections) in a scrollable grid. Users can inspect individual result images with overlays showing the specified query region (semi-transparent blue) and bounding boxes around all found occurrences of the query n-grams within that image. The bounding boxes are color-coded based on their IoU with the query region (Red=0 to Green=1), providing immediate visual feedback on spatial relevance according to overlap. This tool proved invaluable for debugging the region parsing, understanding the scoring behavior (IoU vs. proximity), and visually verifying search results.

6 Conclusion and Future Work

6.1 Summary of Findings

We successfully designed, implemented, and evaluated a spatially-aware textual image search engine. By indexing text n-grams with their normalized spatial coordinates and employing a scoring function that weights both spatial overlap (IoU) and proximity, the system demonstrates a statistically significant improvement in ranking quality (MAP) compared to relevant non-spatial baselines on a large synthetic dataset. The developed synthetic data pipeline and visualization tool were crucial for iterative development and analysis.

6.2 Limitations

The current work relies on synthetic data with perfect text bounding boxes. Real-world application would require integrating an actual OCR engine, which introduces challenges like OCR errors, inaccurate bounding boxes, and confidence scoring. The current relevance model in the evaluation is binary and based on a single ground-truth source image per query. Real-world evaluation would need human-annotated graded relevance. The current scoring weights (IOU=0.5, Proximity=0.5) were chosen as a balance but may not be optimal for all use cases. Furthermore, the system has several other limitations, including index scalability, as the entire inverted index is loaded into memory, potentially posing challenges for extremely large datasets. It also lacks semantic understanding, relying on exact n-gram matching, which prevents it from handling synonyms or paraphrasing. Additionally, the spatial query language is simple, restricted to single rectangular regions and unable to support more complex spatial relationships.

While effective for its targeted task, these limitations are notable when compared to the capabilities of recent MLLM-based approaches [4, 8], which excel at semantic understanding and handling complex natural language queries. However, a key advantage of our system lies in its **Interpretability and Control**. Unlike the often opaque reasoning

of large MLLMs, our scoring is directly tied to explicit geometric calculations (IoU, proximity), making results easier to understand, debug, and tune via weighting parameters.

6.3 Future Directions

Several avenues exist for future work to address the current limitations and enhance the system's capabilities. Integrating a real-world OCR engine (e.g., Tesseract) is a primary next step, which necessitates incorporating OCR confidence scores into indexing and ranking and developing strategies to handle noisy or inaccurate bounding boxes. Transitioning to real-world evaluation is also crucial, requiring the collection or utilization of datasets with human annotations for spatial text queries and employing graded relevance metrics like nDCG, moving beyond the binary relevance used with the synthetic data. Exploring more sophisticated scoring models could further improve performance, potentially incorporating semantic text similarity, visual context features, or adaptive weighting schemes.

To address scalability, investigating dedicated spatial indexing structures (e.g., R-trees) alongside the inverted index could accelerate spatial filtering for large datasets. Enhancements to the user interface, such as allowing users to draw query boxes directly on an image, would improve usability. Exploring hybrid systems that combine the precision and interpretability of our approach with the semantic power of MLLMs [4, 8] offers a promising direction; this could involve using our system for candidate generation followed by MLLM re-ranking. Semantic augmentation techniques, such as incorporating text embeddings to handle synonyms, could also bridge the gap towards MLLM capabilities.

Further research could focus on advanced evaluation metrics specifically designed for grounding tasks, inspired by recent work like the SMuDGE framework [7], to measure spatial accuracy more directly. Finally, extending the query language to support more complex spatial relationships, such as relative positioning or non-rectangular areas, would significantly increase the system's expressiveness and utility.

Acknowledgments

References

[1] Yi Bai, Zhihe Yu, Xin Xu, Xi Yang, Xinbo Wang, and Bing Qin. 2024. Efficient Text-Image Sparse Retrieval via Bernoulli Random Variables Controlled Query Expansion. *arXiv preprint arXiv:2402.17535* (2024). https://arxiv.org/pdf/2402.17535

[2] Moning Cao, Yi Bai, Jingjing Wang, Zhengchen Cao, Liqiang Nie, and Min Zhang. 2023. Efficient Image-Text Retrieval via Keyword-Guided Pre-Screening. *arXiv preprint arXiv:2303.07740* (2023). https://arxiv.org/pdf/2303.07740

[3] Zinan Chen, Yunming Zhu, Wenxian Zhang, Shafiq R. Joty, and Lidong Bing. 2023. STAIR: Learning Sparse Text and Image Representation in Grounded Tokens. In *ICLR 2023 Workshop*. https://openreview.net/forum?id=HXUdnYIe8r

[4] Zhengyuan Huang, Feng Lv, Wenhui Bai, Xiaojun Wang, Jingzhou Liu, Haoran Yang, et al. 2024. Grounding Multimodal Large Language Models to the World. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/pdf/0ea36b222b82ac76c018c9aa7a47f9f978c705b2.pdf

[5] Long Mai, Hanqing Zhang, and Zuxuan Feng. 2017. Spatial-Semantic Image Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5589–5598. https://openaccess.thecvf.com/content_cvpr_2017/papers/Mai_Spatial-Semantic_Image_Search_CVPR_2017_paper.pdf

[6] R. Manmatha, Tushar M. Rath, and Fangfang Feng. 2000. Searching Text in Images. In *CIIR Technical Report*. University of Massachusetts Amherst. https://ciir-publications.cs.umass.edu/getpdf.php?id=317

[7] Hoang D. Nguyen, Andrew N. Bull, and Vinod Nair. 2025. Where is this coming from? Making groundedness count in the evaluation of Document VQA models. *arXiv preprint arXiv:2503.19120* (2025). https://arxiv.org/html/2503.19120v1

[8] Zhenfei Yin, Conghui Chen, Manolis Savva, and Fangbo Sung. 2024. Groma: Grounded Multimodal Large Language Model with Localized Visual Tokenization. In *European Conference on Computer Vision (ECCV)*.

Received 20 April 2025; revised 12 May 2025; accepted 5 June 2025