Needle in a Patched Haystack: Evaluating Saliency Maps for Vision LLMs

Bastien Zimmermann, Matthieu Boussard

Craft AI, Paris, France
bastien.zimmermann@craft.ai, matthieu.boussard@craft.ai

Abstract

ColPali recently proposed a method for explaining multimodal retrieval-augmented generation (RAG) by visualizing how vision–language models (VLMs) connect image patches to text tokens. However, our theoretical analysis and experiments show that these similarity-based saliency maps are fragile and often misleading. We therefore caution against relying solely on intuitive visualizations and present a principled patch-level dissection technique that traces how vision LLMs actually accumulate evidence across modalities. To address this issue, we introduce *Needle-in-a-Patched-Haystack*: a patch-centered dataset and metric suite that quantifies transparency by benchmarking localization performance in vision LLMs. Together, our analysis and toolkit establish a stricter standard for VLM interpretability and provide a drop-in evaluation protocol for future research on robust, multimodal explanations.

Introduction

Retrieval-Augmented Generation (RAG (Lewis et al. 2020)) combines the strengths of information retrieval and generative language modeling to enhance performance in tasks such as question answering and document summarization. In multimodal contexts, RAG leverages both textual and visual data, which is crucial for applications like medical-imaging diagnostics and multimedia content retrieval. This approach promises more accurate and contextually relevant results by drawing on multiple data modalities.

Recent advances include the *ColPali* model developed by Faysse et al. (2025). This model addresses the challenges of visually rich document retrieval. Traditional systems, while effective at query-to-text matching, often fail to fully exploit visual cues, such as tables, figures, and page layouts, that are essential for comprehensive document understanding. *ColPali* leverages the document-understanding capabilities of recent vision-language models to derive contextualized embeddings directly from images of document pages. It integrates a late-interaction matching mechanism into dense retrieval (Zhao et al. 2024): queries and documents are encoded separately, into dense vectors, which then undergo a nuanced interaction phase where pairwise similarity scores are computed and aggregated to refine relevance estimation. *ColPali* not only outperforms modern document-retrieval pipelines but provides a visual alignment heatmap.

ColPali improves transparency with an interpretability method that superimposes a late-interaction heatmap on the input image, highlighting patches most salient to each query token. This fosters user trust and validation, as it clarifies how visual information influences model outputs.

Nevertheless, the complexity of multimodal RAG models such as *ColPali* poses significant challenges for explainability. Their black-box nature makes it difficult to trace outputs back to input features, particularly when visual elements are involved. This opacity hinders user trust and model validation, underscoring the need for robust methods to interpret model predictions.

Overview of Contributions:

- **Theoretical Analysis of Cosine Similarity Limitations:** We critically examine the use of cosine similarity for generating saliency maps in multimodal RAG. Our analysis shows that, although widely used, cosine similarity often fails to accurately reflect the contribution of input features to model predictions, potentially leading to misleading interpretations.
- **Transparency in Vision LLMs:** We introduce a method that improves transparency in Vision Language Models by dissecting image patch processing. This method provides deeper insights into the decision-making processes of these models, and clarifies how visual information influences outputs.
- **Introduction of *Needle-in-a-Patched-Haystack* Metrics and Datasets:** Addressing the need for specialized tools to assess the interpretability of multimodal RAG, we introduce *Needle-in-a-Patched-Haystack*, a package of metrics and datasets tailored to assess how well vision LLMs localize and identify the image patches contributing to the model's decision.

Background and Related Work

Recent advancements in Information Retrieval (IR) have shifted from sparse representations (e.g., TF-IDF, BM25) to dense representations, where queries and documents are embedded into a shared continuous vector space. This evolution, often referred to as *dense retrieval*, is driven by its abil-

ity to capture richer semantic relationships between query and document tokens (Khattab and Zaharia 2020).

Dense Retrieval

In dense retrieval settings, both queries and documents are encoded into low-dimensional vectors through a shared neural encoder. Similarity is typically computed via inner product or cosine similarity, allowing for efficient large-scale retrieval. This strategy has been shown to significantly improve retrieval performance over classical bag-of-words approaches, or emb, due to its ability to generalize beyond exact lexical matches. (Zhao et al. 2024)

Late-Interaction Mechanism

Despite its effectiveness, dense retrieval faces the challenge of balancing computational efficiency with expressive power. Early-interaction models, which capture token-level interactions during encoding, can be prohibitively expensive for large corpora. To mitigate this cost, the *late-interaction mechanism* decouples encoding from interaction, maintaining crucial token-level granularity without sacrificing retrieval speed (Khattab and Zaharia 2020). Specifically:

- **Encoding:** Both queries and documents are passed through a shared neural encoder, yielding token-level embeddings for each token. Since document embeddings are independent of specific queries, they can be precomputed, reducing overall inference time.
- **Interaction:** Once query embeddings have been computed, a pairwise similarity matrix is constructed between query tokens and document tokens. An aggregation function (e.g., max-pooling) is then applied, yielding a final relevance score that captures finer semantic nuances across tokens.

This two-stage design enables efficient large-scale retrieval by avoiding document re-encoding for each query.

Needle in a Haystack

Building on the need for targeted evaluation methods that capture how Vision LLMs localize critical information, we now turn to the widely used *Needle in a haystack* paradigm.

The Needle-In-A-Haystack benchmark (Kamradt 2023) evaluates long-context models by embedding a specific set of words ("needle") within a larger text ("haystack"). Models are assessed based on their ability to accurately retrieve the needle. (Gemini-team 2024) adapted this benchmark to the visual domain by overlaying the needle onto a random video frame, resembling movie subtitles placed in the upper part of the image. Other approaches extend the benchmark by using visual concepts as the needle (Wu et al. 2025).

Multimodal Retrieval-Augmented Generation

Moving beyond purely textual domains, Multimodal RAG (Lewis et al. 2020) extends these concepts to settings where queries and documents may consist of multiple modalities, such as text and images. A typical pipeline retrieves relevant data from diverse sources and subsequently employs a generative model (e.g., a Transformer) to synthesize an output that leverages the retrieved information.

The Interpretability of *ColPali*

ColPali (Faysse et al. 2025) introduced a strategy for multimodal RAG that focuses on visually rich document retrieval. Unlike traditional systems that rely primarily on text-based representations, *ColPali* leverages the powerful feature extraction capabilities of modern Vision-Language Models to encode entire document pages as images. Subsequently, a late interaction mechanism computes pairwise similarity between query tokens and image patches, yielding refined relevance estimates.

In tandem with its strong retrieval performance, *ColPali* provides *interpretability* through a saliency-like visualization: by superimposing a heatmap over the original document image, it highlights which patches most strongly align (in cosine similarity terms) with each query token. This process ostensibly offers users an intuitive window into how the model prioritizes different visual regions within a document.

Their visualization reveals subtle image features the model deems relevant, highlighting latent semantic links beyond object detection. Notably, *ColPali* uses interpolation to smoothly blend attention heatmaps with the image, upscaling saliency maps to match image dimensions.

By integrating these heatmaps along the late-interaction pipeline, *ColPali* presents a mechanism to inspect the inner workings of dense, multimodal systems. This approach enhances user trust but also facilitates model debugging.

Vision-Language Architectures and Transparent Processing

Vision–Language Models (VLMs) fuse visual and textual signals into a unified embedding space, enabling multimodal reasoning, retrieval-augmented generation (RAG), and document understanding (Dosovitskiy et al. 2021). Current state-of-the-art VLMs employ a patch-token pipeline, where images are partitioned into non-overlapping patches. Each patch is projected into a *visual token*, positionally encoded, and subsequently merged with textual tokens.

Here, we propose a structured visualization methodology to inspect internal VLM processes. Figures 1, 2, and 3 (appendix) illustrate three distinct model architectures: *ColPali*, *ColQwen*, and *Gemma*-3. Each figure comprises six panels generated using our visualization function: the original image, resized image, visual encoder grid, raw embedding of the visual encoder, embedding post-VLM forward pass, and a similarity map with respect to a given textual query. These visualizations expose the intermediate representational stages of each architecture, enabling detailed model inspection and interpretability.

ColPali *ColPali* leverages a 27-layer SigLIP vision transformer configured with (32×32) convolutional patch embedding, a 1,152-dimensional hidden representation, and GELU-Tanh activations, combined with a PaLI language head. The vision encoder processes a fixed grid of (32×32) patches, later merged with text. The model's rigid stride preserves computational predictability; at the same time, explicit patch-to-token correspondence makes heat-map attribution straightforward, facilitating clear visual interpretations as illustrated in Figure 1.

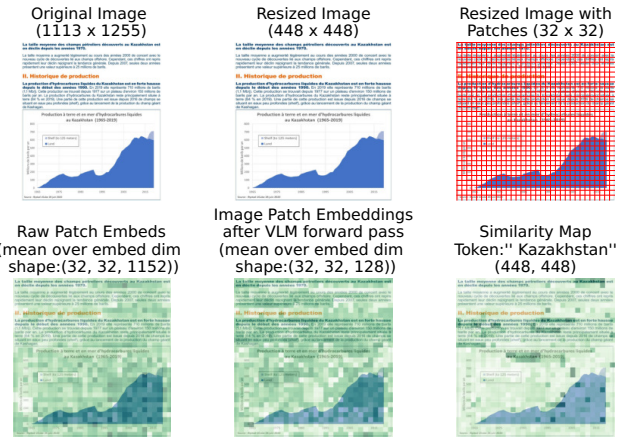


Figure 1: *ColPali*'s image processing pipeline.

ColQwen *ColQwen* adopts *ColPali*'s vision transformer architecture but integrates the *Qwen2-VL* language model. It utilizes a deeper vision tower of 32 transformer blocks (1280-hidden dimension), rotary positional embeddings, and employs a learnable hierarchical patch merger. A distinctive feature of *ColQwen* is its *dynamic patching* strategy: it dynamically adjusts stride to ensure the visual token count remains below 768, accommodating varying image aspect ratios while preventing quadratic computational overhead from attention operations. The resulting variable-length visual sequence is aligned with textual embeddings through the same late-interaction mechanism, enabling token-level retrieval on large-scale images (cf Figure 2).

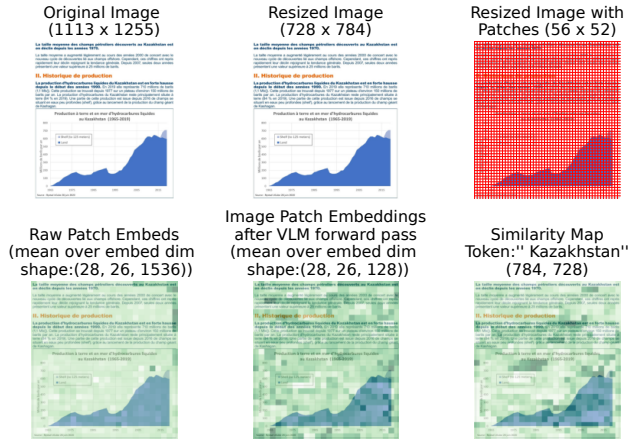


Figure 2: *ColQwen*'s image processing pipeline.

***Gemma* Approach** Gemma-3 integrates a SigLIP visual encoder followed by global average pooling, compressing 4,096 patch embeddings into a fixed-size representation of $(16 \times 16) = 256$ visual tokens. These tokens are then projected into the language model embedding space and concatenated directly with textual tokens. Unlike previous architectures, multimodal integration in *Gemma* occurs inside the decoder via bidirectional cross-attention, rather than through late-stage merging. (cf Figure 3).

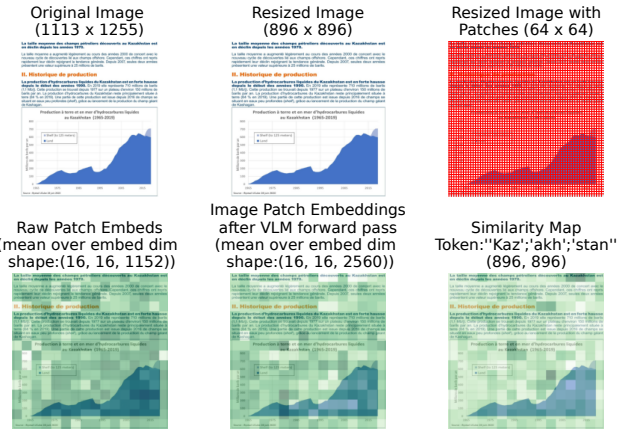


Figure 3: *Gemma*'s image processing pipeline.

Limitations of Cosine Similarity for Saliency Maps

Although *ColPali*'s overlaid heatmap method is appealing for its simplicity, recent analyses highlight potential pitfalls in relying solely on raw cosine similarity to interpret model behavior (Steck, Ekanadham, and Kallus 2024). While it provides a close representation of the late-interaction mechanism, the resulting explanations can nevertheless be *misleading*. In particular, mapping the late-interaction outputs back to the patched inputs is non-trivial, and cosine similarity merely captures representational overlap that *seems* to explain model decisions but may fail to show whether the model is attending to truly relevant features.

Late Interaction. Given a query q and a document d, we denote their multi-vector representations in the common embedding space $\mathbb{R}^D$ as

$$\mathbf{E}_q \in \mathbb{R}^{N_q \times D}, \quad \mathbf{E}_d \in \mathbb{R}^{N_d \times D},$$

where N_q and N_d are the numbers of embedding vectors for the query and the document, respectively. The late-interaction operator $\mathrm{LI}(q, d)$ is defined as the sum, over all query vectors $\mathbf{E}_q(i)$, of their *maximum* inner product with the N_d document vectors $\mathbf{E}_d(j)$:

$$\mathrm{LI}(q, d) = \sum_{i=1}^{N_q} \max_{j=1}^{N_d} \langle \mathbf{E}_q(i) \mid \mathbf{E}_d(j) \rangle. \tag{1}$$

Intuitively, each query vector "selects" its best matching document vector, and these maxima are aggregated to form a retrieval score.

Dot Product and Cosine Similarity: The dot product between two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^D$ is closely related to the cosine similarity measure. By definition,

$$\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos(\theta).$$

Under normalization, maximizing the dot product is equivalent to maximizing the cosine similarity. Under ℓ_2-normalization, maximizing the dot product is therefore equivalent to maximizing the cosine similarity, making it a natural choice for similarity-based retrieval.

Limits of Cosine Similarity Between Token and Patch Embeddings

Computing cosine similarities between a *reference* token (e.g., a specific patch embedding) and other patch tokens in a Vision Transformer essentially measures how close the patch embeddings are to the reference in the model's latent space. At first glance, it might seem like this could serve as a simple saliency map—*which patches are most similar to this reference?*—but there are several pitfalls:

1. **Embeddings Are Not Necessarily Interpretable as Saliency:** Each patch embedding depends on contextual information from the entire image (due to self-attention). As a result, the similarity between a reference patch and others might partly reflect that patch's influence on the others (or vice versa) rather than isolate how critical it is to the model's prediction.

   In Vision Transformers, self-attention mechanisms compute

   $$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\Big)\mathbf{V}.$$

   Although $\mathbf{Q}\mathbf{K}^\top$ is a dot product, the final attention weight for a token depends on the entire context (softmax across all patches). Thus, even if patch j is locally similar to the query token, the model may *route* attention elsewhere, making patch j unimportant for the final decision (Serrano and Smith 2019).
2. **No Direct Link to the Model's Decision:** In opposition to standard saliency metrics—which assess how perturbations in specific input regions affect a model's final decision—raw cosine similarity is agnostic to the output. Even if two embeddings are closely aligned in direction, this does not reflect the *effect* of a patch on the prediction. By contrast, methods such as gradient-based attribution (Sundararajan, Taly, and Yan 2017) or perturbation tests (e.g., occluding a patch) explicitly capture causal or mechanistic contributions, thereby clarifying *why* the model reaches a particular conclusion and how *sensitive* it is to individual features.
3. **Attention $\neq$ Similarity:** In Vision Transformers, attention weights are not simple cosine similarities between token embeddings. Instead, they are dynamic, context-dependent scores computed via the scaled dot-product between queries (e.g., the class token) and keys (e.g., patch embeddings), followed by a softmax. These weights depend on the model's learned projections and token interactions, not just geometric alignment. Thus, high cosine similarity doesn't imply strong attention, and vice versa. Empirical studies (e.g., (Jain and Wallace 2019)) show that attention often fails to reflect true feature importance, reinforcing that attention is not a direct similarity measure.

Although the cosine similarity of patch embeddings may reflect feature similarity, the associated saliency is flawed because it does not directly measure the contribution or importance of those patches to the model's output. If the goal is interpretability or highlighting regions that are crucial for a decision, methods directly linked to the model's gradients (e.g., gradient-based saliency), attention distributions, or other attribution techniques provide more faithful explanations.

Methodology

Our methodology systematically evaluates the saliency maps of VLMs in terms of both localization and text-matching capabilities within patch-based document images. To this end, we introduce four synthetic datasets that progressively increase in difficulty and realism, reflecting common scenarios in RAG pipelines. Furthermore, we introduce a set of metrics, based on per-patch similarity scores, to quantify a model's ability to accurately focus at the patch level and retrieve text.

Patch-Based Datasets for Vision-Language Models

To assess both raw localization capabilities and text-based retrieval within document images, we create specialized datasets that evaluate specific capabilities involved in this task (Figure 4). Each dataset is designed around the notion of a *special patch*. This patch is the only patch we are interested in and we want to evaluate the ability of the model to localize it.

Because different backbones discretize images into patches of different sizes and grids, we dynamically set the dataset's image dimensions so that its grid aligns exactly with every model's internal patch representation—avoiding partial overlaps or mismatched cells.

The concrete patch sizes, grid resolutions, and font sizes used for every model (under both *in-grid* and *out-grid* settings) are listed in the appendix.

Patch Dataset: Assessing Raw Localization Capabilities Images are divided into a grid of size (n_patches_x $\times$ n_patches_y), with each patch sized (*patch_size* $\times$ *patch_size*) pixels. All patches are white, except for one *special patch* coloured black. This setup evaluates the model's capacity to detect and focus on visually distinct elements, representing a basic yet crucial localization task.

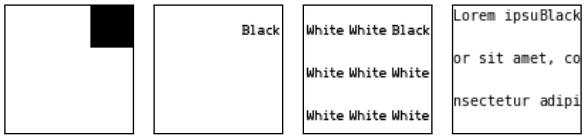
Single-word Dataset: Simulating Text-Rich Document Settings Next, the *special patch* is augmented with high-contrast text, challenging VLMs to integrate textual and visual cues effectively. This scenario emulates PDF-like documents, where a salient text region must be accurately localized within the broader image.

Multi-words Dataset: Further Text Complexity Non-*special patches* are now populated with a different word than the special-patch one. Here, the model must isolate relevant textual information in the *special patch* while ignoring distracting, irrelevant content in other patches. We constitute two sets of Multi-words datasets, depending on how we

build the pair of words between the interesting and confusing ones. The first set is comprised of unrelated pairs and the second one of related one.

We first construct a length-restricted vocabulary by drawing $n_{\text{real}} = 2{,}000$ authentic English adjectives from a cached MIT list and augmenting it with $n_{\text{fake}} = 10$ synthetically generated non-words of identical character length produced by *Faker* (Faraglia and Contributors 2025); a fixed random seed guarantees replicability of this sampling procedure. Each unique token w is mapped to a $\mathbb{R}^{300}$ embedding via the pretrained English fastText model (`cc.en.300`), then ℓ_2-normalized to unit length. We exhaustively evaluate cosine similarity for every unordered pair (w_i, w_j), assign a *positive* label when $\cos(\theta_{ij}) > 0.7$, and mark a pair *negative* when $|\cos(\theta_{ij})| \leq 0.1$, thereby isolating, respectively, semantically coherent and quasi-orthogonal word pairs. Finally, we then sample 10 positives and 10 negatives word pairs.

Text Dataset: Lorem Ipsum Finally, non-*special patches* are also populated with unrelated text, substantially increasing the difficulty. Here, the model must isolate relevant textual information in the *special patch* while ignoring distracting, irrelevant content in other patches.



(a) Patch (b) Single-word (c) Multi-words (d) Text

Figure 4: Visual representations of datasets used to assess VLMs, with the *special patch* at position (2, 0) inside a 3 × 3 grid.

Evaluation Metrics for Image Similarity Maps

We assess model performance by measuring how the saliency or similarity map aligns with the *special patch*'s known location, focusing on how well saliency maps capture relevant image features, especially textual regions.

Let the flattened similarity map be $\mathbf{s} \in \mathbb{R}^n$ with entries s_i and denote by

$$i_{\max} \;=\; \arg\max_i s_i$$

the index of the most salient patch. We further let $\mathcal{I} \subseteq \{1, \ldots, n\}$ collect the indices of all ground-truth *interesting* patches. For readability concerns, in some cases, the special word spans multiple patches. (the font size required for it to fit within a single patch would be too small to be read) For each map we report four complementary metrics:

- **Accuracy:** A binary success indicator,

$$\text{Acc} \;=\; \mathbb{1}\big(i_{\max} \in \mathcal{I}\big),$$

  which equals 1 iff the model's top-ranked patch coincides with *any* interesting patch.

- **Score:** The mean similarity assigned to the interesting regions,

$$\text{Score} \;=\; \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} s_i,$$

  capturing the absolute response strength the model allocates to the target content.

- **Rank:** The 1-based ordinal position of the best interesting patch in the global ordering of similarities. The rank is then normalized to remove the dependency on the grid size, giving a value in $[0, 1)$:

$$\widehat{\text{Rank}} \;=\; \frac{1}{HW} \sum_{j=1}^{HW} \mathbb{1}\big(s_j > \max_{i \in \mathcal{I}} s_i\big),$$

  where $H \times W$ is the shape of the similarity map. A score of 0 means an interesting patch is the *global* maximum; values nearer 1 indicate poorer localisation.

- **Distance (*normalised*):** The Euclidean distance between the predicted peak patch and the *nearest* interesting patch, scaled by the grid diagonal so the result lies in $[0, 1]$:

$$\widehat{\text{Dist}} \;=\; \frac{1}{\sqrt{(H-1)^2 + (W-1)^2}} \min_{i \in \mathcal{I}} \big\| \mathbf{p}_{\max} - \mathbf{p}_i \big\|_2,$$

  with $\mathbf{p}_i = (r_i, c_i)$ and $\mathbf{p}_{\max}$ the coordinates of the predicted peak. Thus 0 denotes a perfect hit and 1 a miss at the opposite corner of the grid.

Focusing on per-patch metrics instead of pixel-level signals, our evaluation captures a VLM's saliency map's ability to select the correct patch in line with its internal processing. This patch-centric approach avoids misinterpretations from salient features spilling into neighboring patches.

Needle in a Patched Haystack: Grid-Based Saliency Evaluation

To quantify how reliably a vision–language model pinpoints textual cues, we first build a *grid result map*: for every spatial index (x, y) in each dataset grid we implant the *special patch*, compute the full suite of metrics from Section , and log the outcome, producing a two-dimensional performance surface conditioned on patch location. Averaging these per-location scores across multiple iterations for the same query yields the *Needle-in-a-Patched-Haystack* map, a smoothed saliency landscape that simultaneously exposes the regions where the model consistently attends to the correct patch and the areas where it is prone to spurious activations, thereby offering a concise yet incisive diagnostic of its localisation prowess.

Experimental Results

Using the *Needle-in-a-Patched-Haystack* benchmark of Section , we probe cosine-similarity saliency maps along three axes that matter for retrieval-augmented vision–language systems: *patch-level grounding*, *robustness to clutter and distribution shift*, and *susceptibility to lexical interference*. This yields the following research questions.

1. **Patch-level localisation.** Across the four synthetic datasets that move from blank grids (PATCH) to text-rich pages (TEXT), how accurately does a model elevate the *special patch* to the top of its similarity map?
2. **Robustness to realism and spatial bias.** As distractors become progressively more realistic, and regardless of patch position, how do accuracy, rank, and localisation error degrade?
3. **Lexical interference.** When the distractor word is *semantically related*—rather than unrelated—to the query term, to what extent do similarity maps misattribute importance, and does late-interaction fine-tuning amplify this confusion?

We first describe our evaluation protocol, then answer the research questions in the following sections. A qualitative analysis and limitations are then presented.

Experimental Setup

Models. We evaluate three publicly available VLMs that span different design decisions: *ColPali* (vidore/colpali-v1.2), *ColQwen* (vidore/colqwen2-v1.0), and *Gemma* (google/gemma-3-4B-it), which was *not* fine-tuned for retrieval tasks.

Datasets. The four synthetic datasets introduced in Section —PATCH, SINGLE-WORD, MULTI-WORD and TEXT—form an ascending difficulty ladder.

Each configuration is generated with 20 iterations, with each iteration featuring a different set of word pairs when relevant.

Baseline. We report a single **random** baseline that characterizes chance-level performance: it records (i) the average Euclidean distance between two randomly selected patches, (ii) the expected rank of a randomly selected patch, and (iii) the probability of correctly guessing the patch. This baseline explicitly captures the performance achievable without any spatial reasoning.

Metrics and statistics. We follow the per-patch metrics described in Section , reporting *Accuracy*, *Score*, *Rank*, and *Distance*. Unless otherwise stated, results are aggregated across random seeds and reported as *mean ± 95% confidence interval*. The half-width of each two-sided interval is $t_{0.975,\,n-1}\, s/\sqrt{n}$, where s is the sample standard deviation over the n seeds and $t_{0.975,\,n-1}$ is the critical value of Student's t-distribution.

Implementation details. All experiments were conducted on a NVIDIA A10G-24GB GPU.

Performance Evolution Across Dataset Types

Figure 5 summarizes the behavior of the *ColPali*, *ColQwen*, and *Gemma* models across all datasets introduced in Section . Taken together, the results highlight that the closer the dataset resembles a real text document, the better the models perform at locating pertinent information, as evidenced by steadily rising accuracy. Furthermore, even when incorrect, the models consistently approach the patch of interest as the dataset becomes more realistic, reflected by a reduction in the distance metric. The closer the samples resemble the training dataset distribution, the more effectively the similarity maps highlight salient regions.

While the score distribution does not change drastically, the rank metric reveals different trends across models. *ColPali*'s rank steadily improves, *ColQwen*'s rank deteriorates except for the text dataset, where it significantly improves, and no clear trend emerges for *Gemma*. Considering the other metrics, this indicates for *ColQwen* that although localization improves, the similarity map's distribution of scores shifts toward higher values, even in non-relevant regions.

Gemma achieves the highest performance on single-word datasets, whereas *ColQwen* and *ColPali* perform better on text-rich images. Suggesting that RAG-oriented fine-tuning has enhanced their performance on precise this task.

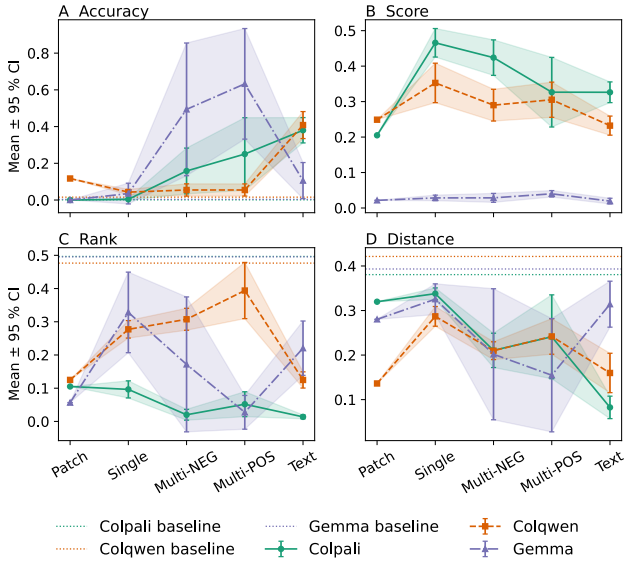


Figure 5: Mean ± 95% CI for four performance metrics across dataset types. Accuracy and Score are higher-better; rank and distance are lower-better. ‘

Qualitative Analysis

O-shaped anomaly Figure 6 presents a visualization of the metric results when the *ColPali* model is applied to the Patch Dataset. Despite exhibiting a strong signal in terms of *score*, the model consistently fails to correctly identify the special patch. In other words, the top-scoring patch never coincides with the special patch. Nevertheless, the *rank* of the special patch is very low (i.e., it is among the top-ranked patches), suggesting that although the special patch does not emerge as the single brightest region, it still receives notable attention.

A notable pattern is the emergence of an ”O-shaped” region of patches near the bottom-left corner, which behaves differently from the rest of the grid. This region frequently exhibits distinct similarity scores, potentially indicating a

spurious visual or embedding cue detected by the *ColPali* model. Further investigation revealed that the vision embedding component of *ColPali* is faulty in this region (See Figure 1 or Figure 4 in the appendix). The model also struggles along the image borders.

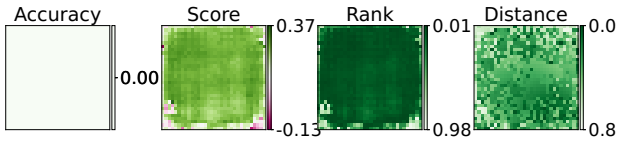
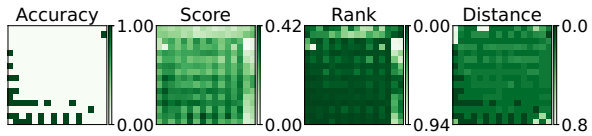
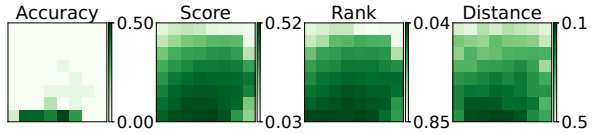


Figure 6: Visualization of patch embeddings for the Patch Dataset using the *ColPali* model.

Bottom-left gradient bias In Figure 7, *ColQwen* displays a distinct spatial bias gradient concentrated in the bottom-left area across different datasets. A distinct grid pattern emerges (Figure 7a), with certain squares exhibiting very high similarity scores. Upon further inspection on another dataset (Figure 7b), we observe this bias manifesting as a bottom-left gradient. This indicates a strong spatial bias. The model's localization capabilities appear uneven depending on the position of the interesting patch in the image.



(a) Patch dataset.



(b) Single-word dataset.

Figure 7: Performance of *ColQwen* on two distinct datasets.

Semantic Interference: Positive vs. Negative Pairs

To isolate the effect of lexical similarity we compare pairs of words where the distractor word is either semantically *related* or *unrelated* to the needle (Figure 8).

Changing from semantically related distractors to unrelated ones impacts every model. Performance slightly drops for both *ColPali* and *ColQwen* when distractors are related, but improves for *Gemma*. This aligns with the intuition that more similar words are more confounding. We hypothesise that late-interaction fine-tuning encourages token-level matching and therefore amplifies lexical interference, whereas *Gemma*'s weaker retrieval inductive bias reduces false positives.

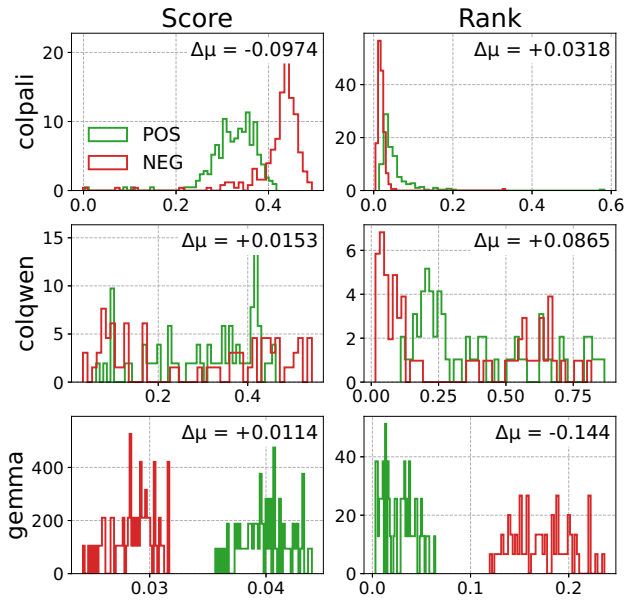


Figure 8: Distribution of *score* and *rank* for **positive** (green) and **negative** (red) pairs across the evaluated models. Numbers in the upper-right of each cell report the change in mean value from (NEG) to (POS).

Discussion and Limitations

Our investigation critically evaluates the reliability of cosine similarity-based saliency maps in multimodal RAG systems. While our experiments confirm that VLMs leverage patch-level visual information, several key limitations emerge:

Spatial Biases. We observe notable spatial artifacts—such as the "O-shaped anomaly" in *ColPali* and the "bottom-left gradient bias" in *ColQwen*, indicating non-uniform processing of patch representations. These artifacts may stem from positional encoding schemes, dataset-induced biases, or architectural constraints inherent to Transformers. Addressing them may require improved positional encodings or targeted regularization techniques.

Modality Gap. A persistent semantic gap between image and text embeddings, commonly referred to as the "modality gap phenomenon", continues to challenge effective multimodal alignment. Despite embedding both modalities in a shared space, residual separation impairs retrieval, clustering, and classification tasks (Role, Meyer, and Amblard 2025). Our work shows that, despite the efforts of *ColPali* and *ColQwen* to reduce this gap, it persists when the image strays from the fine-tuning distribution. Bridging this gap is essential for improving the interpretability and functional accuracy of saliency representations.

Spatial Reasoning Limitations. Our findings echo broader evidence of spatial reasoning limitations in LLMs, including recent work (Cohn and Blackwell 2024). While both VLMs and LLMs exhibit some spatial understanding, they often struggle with precise spatial relationships, such as directionality or topology—further undermining localization accuracy in vision-only contexts.

Future research focus on evaluating models using real-world scanned documents and developing fine-tuning objectives that mitigate positional and modality-related biases. These directions will enhance both the interpretability and performance of multimodal saliency methods.

Conclusion

We present a in-depth critique of cosine similarity-based saliency techniques in multimodal RAG systems. Through theoretical analysis and systematic experiments with our *Needle-in-a-Patched-Haystack* benchmark, we uncover key limitations in current interpretability practices. Our results caution against over-reliance on raw similarity maps and emphasize the need for more robust, context-aware saliency attribution methods. The dataset and evaluation toolkit released with this work aim to foster further research toward more transparent and trustworthy VLM.

Impact Statement

This work aims to advance the transparency of multimodal RAG by introducing the Needle-in-a-Patched-Haystack benchmark, which exposes weaknesses in patch-level saliency maps.

Trust. A rigorous, open-source test suite lets practitioners better evaluate that VLMs truly ground their outputs in relevant image regions, reducing the risk of hidden failure modes in high-stakes settings such as medical imaging or legal document review.

Responsibility. Because all datasets are procedurally generated and contain no personal data, the benchmark poses no direct privacy risks and is freely extensible, encouraging reproducible and ethical research.

Cost and footprint. Benchmarking VLMs is compute-intensive: repeated forward passes incur substantial energy and financial costs, which can hinder smaller labs and enlarge the field's carbon footprint. Providing trusted runs and releasing open results can mitigate these drawbacks. Precise evaluation is essential for safe deployment, but it must be balanced against environmental and economic costs.

References

Cohn, A. G.; and Blackwell, R. E. 2024. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions (Short Paper). In *16th International Conference on Spatial Information Theory (COSIT 2024)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Faraglia, D.; and Contributors, O. 2025. Faker. https://github.com/joke2k/faker. Version 24.15.0, accessed 22 May 2025.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. In *The Thirteenth International Conference on Learning Representations*.

Gemini-team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv:2403.05530 [cs].

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics.

Kamradt, G. 2023. LLMTest_NeedleInAHaystack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack/blob/main/README.md. GitHub repository. 1, 2.

Khattab, O.; and Zaharia, M. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, 39–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380164.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Role, F.; Meyer, S.; and Amblard, V. 2025. Fill the Gap: Quantifying and Reducing the Modality Gap in Image-Text Representation Learning. arXiv:2505.03703.

Serrano, S.; and Smith, N. A. 2019. Is Attention Interpretable? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931–2951. Florence, Italy: Association for Computational Linguistics.

Steck, H.; Ekanadham, C.; and Kallus, N. 2024. Is Cosine-Similarity of Embeddings Really About Similarity? In *Companion Proceedings of the ACM Web Conference 2024*, 887–890. ArXiv:2403.05440 [cs].

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3319–3328. JMLR.org.

Wu, T.-H.; Biamby, G.; Quenum, J.; Gupta, R.; Gonzalez, J. E.; Darrell, T.; and Chan, D. 2025. Visual Haystacks: A Vision-Centric Needle-In-A-Haystack Bencahmark. In *The Thirteenth International Conference on Learning Representations*.

Zhao, W. X.; Liu, J.; Ren, R.; and Wen, J.-R. 2024. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.*, 42(4): 89:1–89:60.