

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



Bastien Zimmermann¹ Matthieu Boussard²

¹Craft AI, Paris, France

Motivation

Saliency maps provide a window on multimodal RAG, yet cosine-similarity maps (e.g. Col-Pali) are fragile—lighting up spurious patches and collapsing under lexical noise. Lack of rigorous evaluation. No standard way to measure whether a vision–language model

truly localizes the patches that drive its predictions.

High-stakes applications need transparency. Critical domains require knowing why a model retrieved specific regions.

Key Contributions

- Theoretical critique of cosine similarity
- We prove it misaligns with true patch influence in multimodal RAG, revealing systemic failure modes.
- Novel transparency method for Vision LLMs

Transparent processing of VLM for visual RAG



moyenne des champs pétrollers décenverts au Kazabhstan est n depuis les années 1970.

2000 de concert avec le La taille i nt, ses atriffres ont repris 2007, seulos deux années repleteme

ction

La production d'hydrosurisans liquides du Kazakhstan est en forte has depuis le debut des années 1990, En 2010 elle repriserte 110 million de (U.) Maj). Catte production es trouvel dopuis (971 na un plateau d'environ 158 millio baris par an. La production d'hydrocarbures du Kazakhstan reste principaisment du terre (Hi % en 2018). Une partie de catte production est issue depuis 2016 de share baris no esus peu protocées (phol?), gritos es lancement de la production du champ de Kastegan.





I. Historique de production a production d'hydrocarbures impeldes du Kazakhetan est enferte la profue le délart des années 1990. En 2018 ette représente 710 millions o 1 May, Cotté production se travaré depuis 1977 sur un plateau d'environ 150 millions et la par av. La production d'hydrocarbures du Kazakhetan reste principalement en da té es 19118 : Line samé de cette conductions des mes devices 1918 de la cette et la partieux des 1918 : Line samé de solte productions est teures devices 1918 de la cette

Resized Image





Resized Image with

Patches (32 x 32)

Raw Patch Embeds aft (mean over embed dim of shape (mean o

Image Patch Embeddings after VLM forward pass (mean over embed dim of shape:

Similarity Map Token: '' Kazakhstan''

- Our method decomposes visual token flows to produce a faithful depiction of the mechanisms that generate saliency maps.
- Needle-in-a-Patched-Haystack benchmark
- A dataset & metric suite that probes localisation fidelity for vision-language tasks.

Patch-Based Datasets for Vision–Language Models

	Black	White White Black	Lorem ipsuBlack
		White White White	or sit amet, co
		White White White	nsectetur adipi
(a) Patch	(b) Single-word	(c) Multi-words	(d) Text

Figure 1. Visualisations of the datasets used to assess VLMs, with the special patch at position (2,0) inside a 3×3 grid.

- Goal Probe localization and text-conditioned retrieval by centering every image on a single special patch.
- Grid alignment Each image is resized so its patch grid exactly matches the resolution and patch size of the tested model, avoiding partial overlaps.
- Datasets (in increasing difficulty)
- **1. Patch** \Rightarrow raw visual localization.
- 2. Single-word \Rightarrow joint vision-text cue.
- **3.** Multi-words \Rightarrow isolate relevant text amid distractors.
- 4. Text (Lorem lpsum) \Rightarrow hardest real-world case.







Figure 2. Visualization of ColPali's image processing pipeline.



Cosine Similarity ≠ Saliency

Representational overlap is not causal importance.

High cosine similarity between patch embeddings merely shows they occupy nearby directions in latent space—it does not prove those patches drive the model's prediction.

 Context Entangles Patches. Each embedding already mixes information from all patches, so similarity may reflect shared context rather than true relevance.

Faithful explanations must link to the output.

Gradient-based or perturbation methods quantify how changing a patch alters the final score, providing a more reliable saliency signal.

Evaluation Metrics

• Accuracy: A binary success indicator,

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

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

Score: The mean similarity assigned to the interesting regions,

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

• Rank (normalised): Share of patches scoring higher (0 = best):

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

- where $H \times W$ is the shape of the similarity map.
- **Distance (normalised):** The Euclidean distance between the predicted peak patch and

Figure 3. Mean \pm 95% CI for the four performance metrics across dataset types. Accuracy and Score are higher-better; Rank and Distance are lower-better.



Figure 4. Bottom left gradient bias of *ColQwen* on the Single-word dataset.

Key Takeaways

- **Realism matters**: Greater realism increases accuracy and reduces localisation error.
- Model-specific behaviour: Trends differ across models.
- Spatial biases: "O-shaped" (ColPali) and bottom-left gradient (ColQwen) anomalies.

the nearest interesting patch, scaled by the grid diagonal:

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

Needle in a Patched Haystack: Grid-Based Saliency Evaluation

- Grid result map: Iteratively plant the special patch at every grid index (x, y); compute localisation metrics for each placement.
- Aggregation: Average per–location scores over all runs to obtain a smoothed 2-D surface.
- Outcome: The resulting Needle-in-a-Patched-Haystack map visualises where the model consistently locks onto the correct patch and where spurious activations arise, offering a concise diagnostic of localisation skill.

References

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Celine Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In The Thirteenth International Conference on Learning Representations, 2025.

- Lexical interference: Semantically related distractors hurt ColPali/ColQwen but slightly benefit Gemma.
- Bottomline: Raw cosine-similarity maps can mislead; our benchmark and toolkit enable more transparent, trustworthy saliency attribution.

Discussion, Limitations & Future Work

- Spatial biases persist: Vision-LLMs exhibit location-specific artefacts, likely rooted in positional encodings, data biases, or architectural limits—calling for better encoding schemes or regularisation.
- Modality gap: Image/Text embeddings remain partially separated—closing this gap is essential for reliable retrieval.
- VLM have a weak spatial reasoning: Directionality and topology remain challenging for current models.
- Next steps:
- Validate on real-world scanned documents.
- Design fine-tuning objectives that penalize positional and modality bias.
- Develop context-aware saliency methods beyond raw cosine scores.

Actionable Interpretability Workshop at ICML 2025

bastien.zimmermann@craft.ai