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

Bastien Zimmermann [1] Matthieu Boussard [1]

[1]Craft AI, Paris, France

Motivation

Saliency maps provide a window on multimodal RAG, yet cosine-similarity maps (e.g. ColPali) 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**
  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

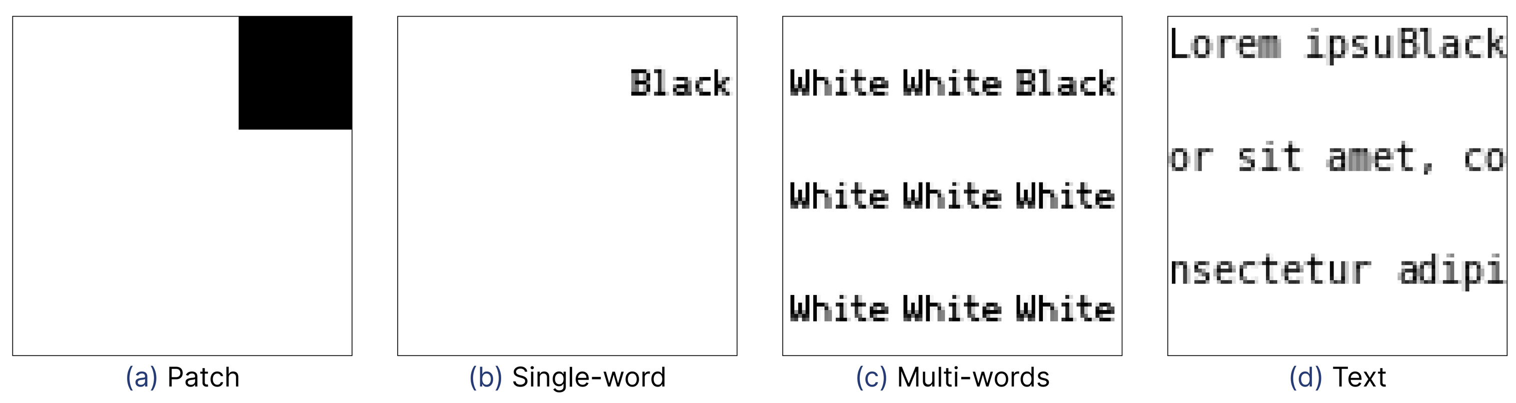

Figure 1. Visualisations of the datasets used to assess VLMs, with the special patch at position (2,0) inside a 3×3 grid.

- **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** ⇒ raw visual localization.
  2. **Single-word** ⇒ joint vision–text cue.
  3. **Multi-words** ⇒ isolate relevant text amid distractors.
  4. **Text (Lorem Ipsum)** ⇒ hardest real-world case.

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,

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

which equals 1 iff the model's top-ranked patch coincides with any interesting patch.
- **Score:** The mean similarity assigned to the interesting regions,

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

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

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

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

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

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.

Transparent processing of VLM for visual RAG

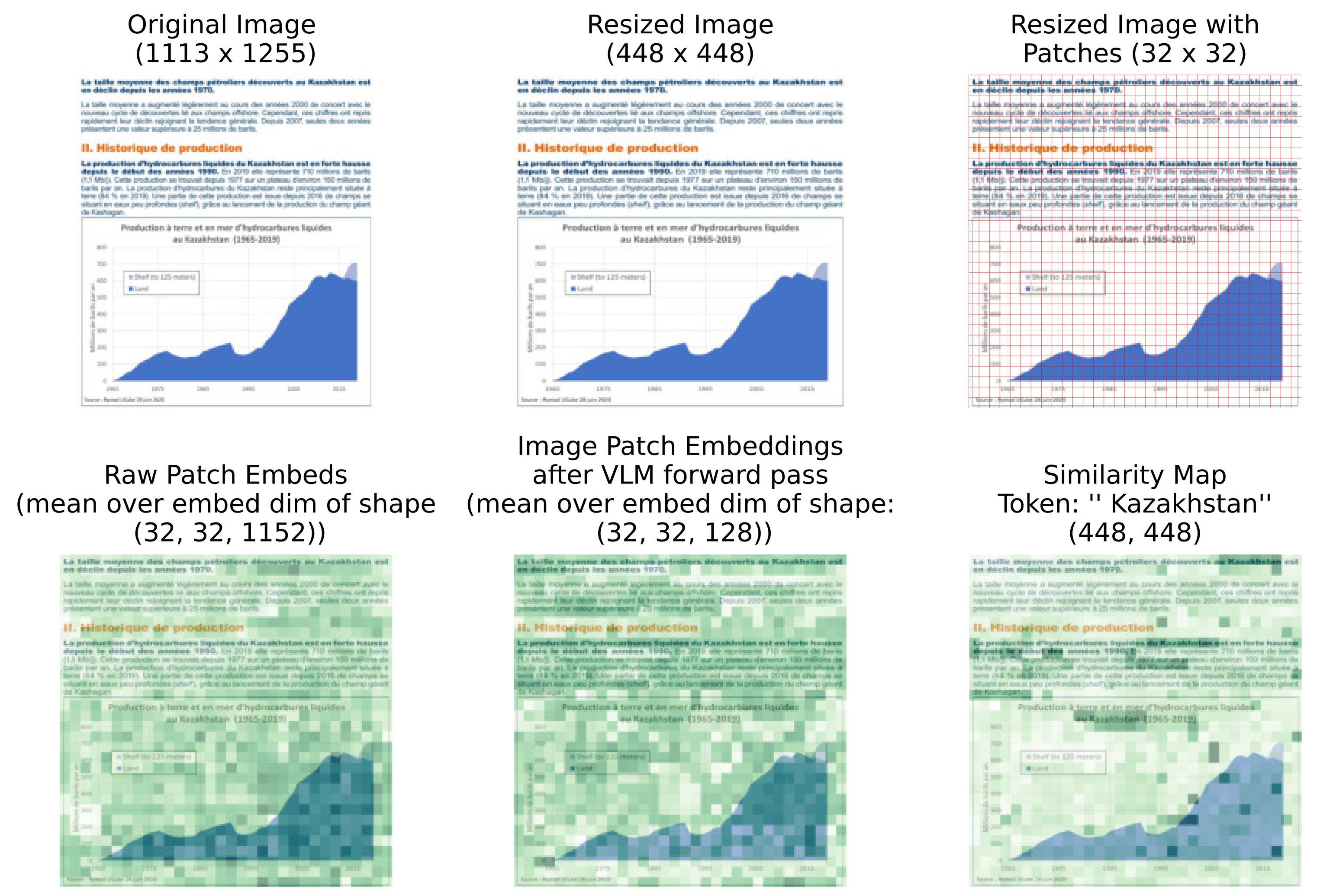

Figure 2. Visualization of ColPali's image processing pipeline.

Results

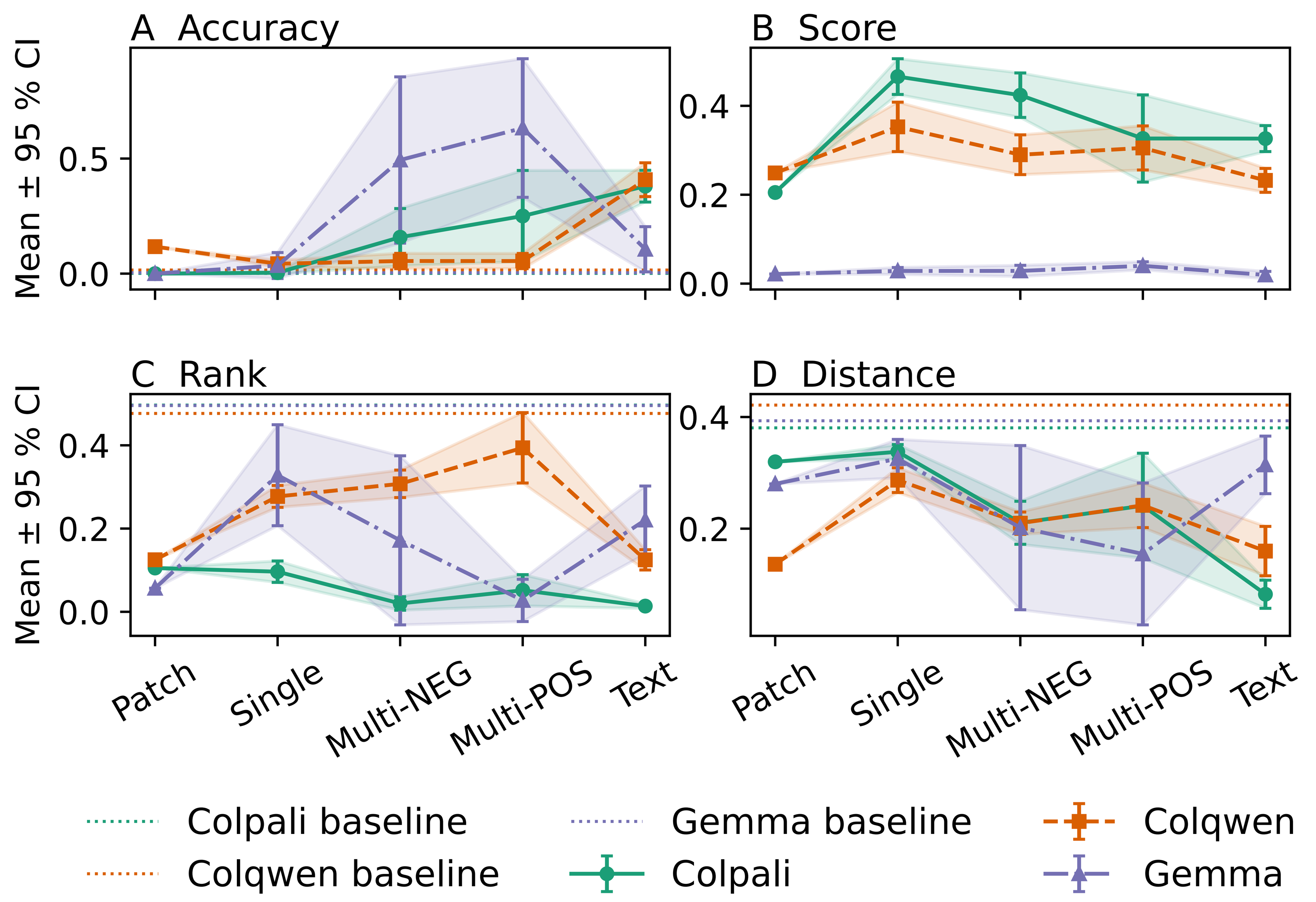

Figure 3. Mean ± 95% CI for the four performance metrics across dataset types. Accuracy and Score are higher-better; Rank and Distance are lower-better.

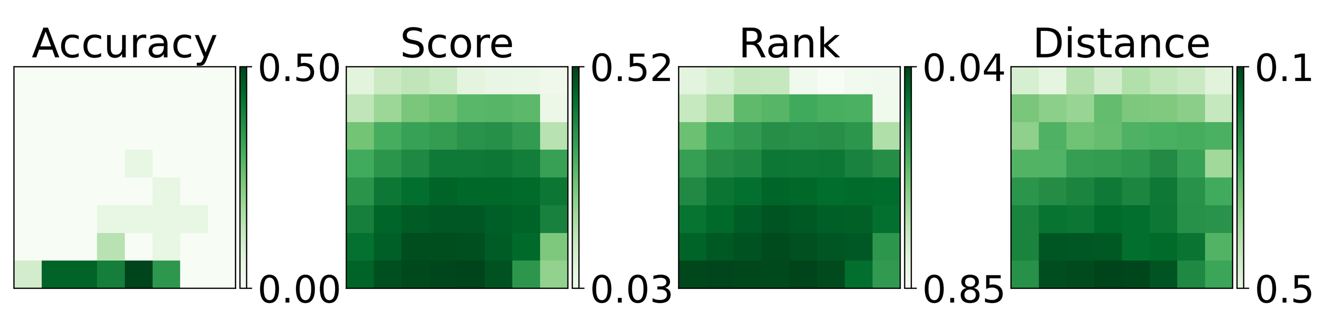

Figure 4. Bottom left gradient bias of *ColQwen* on the Single-word dataset.

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.
- **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.