WHERE BINDING BEGINS: ATOMIC FEATURE MAPS IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) describe complex scenes fluently, yet they fail at conjunctive search, where the target is defined by a combination of features, such as a particular color and shape, and each distractor shares one of them. This failure is a signature of the binding problem: keeping track of which features belong to which object. Humans show the same failure when they must respond before attention can select objects one at a time. Recent mechanistic interpretability work suggests that VLMs bind features through content-independent spatial retrieval mechanisms, but it remains unclear how the feature representations supplied to these mechanisms are themselves constructed. By intervening on the internal activations of Qwen2.5-VL-7B, we find that the searched-for features are gathered while the model reads the question (e.g., "Is there a green letter L?") by attention heads at the color word and at the shape word that build maps of where each feature appears in the image. Patching these maps between displays makes the model attach a feature to the wrong object and report an absent target. These illusory conjunctions grow as objects get closer together, as in human crowding, and this proximity effect is carried by shape heads, not color heads. Supplying a clean shape map raises search accuracy from 32% to 94%, whereas a clean color map has little effect. Ablating the maps impairs counting but spares spatial-relation judgments, consistent with a separate indexing mechanism. These results suggest that, within a single forward pass, the model implements a parallel, feature-based stage of search but lacks the serial, object-by-object selection that reliable binding requires.

1 INTRODUCTION

Binding requires a vision system to keep track of which features belong to which object (Treisman & Gelade, 1980; Greff et al., 2020; Roelfsema & Serre, 2026). Human vision handles this in part by attending to objects one at a time, and when observers must respond before attention can do so, they fail in a characteristic way. In disjunctive search, where the target differs from every distractor along a single feature, they find it however many distractors surround it. In conjunctive search, where only a combination of features defines the target, performance degrades as distractors accumulate. Vision-language models (VLMs) describe complex scenes fluently, yet they show the same signature (Campbell et al., 2024), with accuracy near ceiling in disjunctive search and falling toward chance in conjunctive search (Figure 1). VLMs thus behave like pre-attentive human vision, even though they are built on self-attention.

Recent mechanistic work has begun to explain how VLMs bind. Assouel et al. (2026) show that attention heads assign content-independent spatial indices to objects and later use them to retrieve object features. Binding by location is what human vision attributes to attention, which sits uneasily with the search results above, where VLMs behave like pre-attentive vision (Campbell et al., 2024). The two findings come from different regimes: the indexing analyses use displays of two to nine well-separated objects, whereas search fails as distractors accumulate. Representational analyses link binding errors to overlap between competing object and attribute directions (Savietto et al., 2026; Li et al., 2025), and training models to point to objects through explicit coordinates removes these errors (Haputhanthri et al., 2026). None of this work asks what changes between sparse and cluttered displays. An indexing mechanism can only bind what it is given. Where are the per-feature representations it operates over constructed, and do they themselves fail under clutter?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

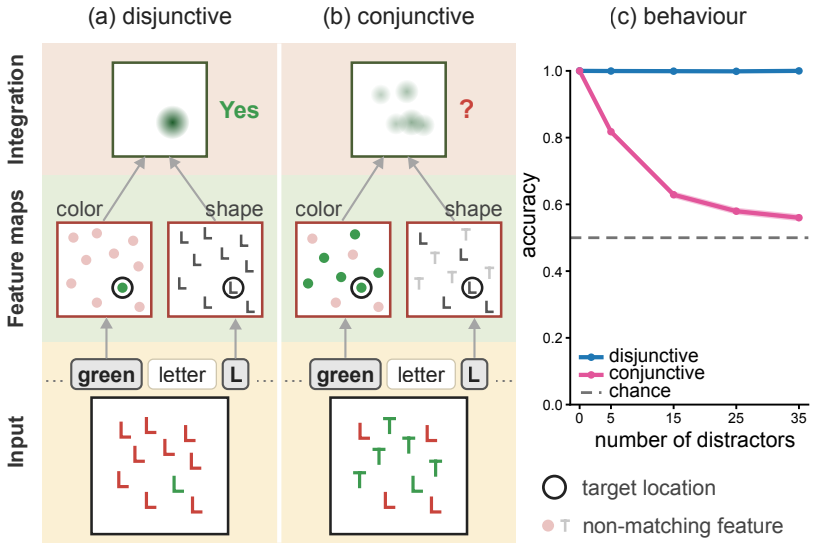


Figure 1: **Visual search in Qwen2.5-VL-7B, framed by Feature Integration Theory.** Each column shows the input display (bottom), the feature maps gathered at the query's feature words, and their integration into a decision, for the query *"Is there a {color} letter {shape} in the image?"*. **(a)** Disjunctive display (a green L among red Ls): the color map alone peaks at the target, so no binding is needed. **(b)** Conjunctive display (green Ts added): both maps contain several candidates, and only their intersection identifies the target. The target location is the same in (a) and (b), which differ only in distractor identity. **(c)** Accuracy versus number of distractors (2250 images per bin, half target-present; dashed line: chance).

Feature Integration Theory (FIT) offers an answer from human vision: separate, spatially organized maps for each elementary feature are computed pre-attentively, and attention later binds them into objects (Treisman & Gelade, 1980). Staged "perceive-then-relate" computation has also been found in vision transformers (Lepori et al., 2024). In a VLM, however, there is no reason this stage must live where FIT places it, in a bottom-up visual stream computed before the task is known. The model reads the question after the image, and every query token can attend back to every image token.

We find that this is where the feature maps are constructed. Using causal mediation analysis on Qwen2.5-VL-7B (Bai et al., 2025; Yang et al., 2024), we show that the conjunctive–disjunctive performance gap is mediated largely by a sparse set of attention heads operating at the positions of the query's feature words, the color word and the shape word. We call them color heads and shape heads, or feature-map heads together. Each word position gathers an atomic feature map, a spatial representation of where its feature occurs in the image, computed without regard to the object's other features. The query word acts as a probe that reads a feature-specific map out of the image. Binding is thus deferred to a later step that must bring two separately constructed, word-anchored maps together.

We test this account with causal interventions. Transplanting feature maps across displays induces illusory conjunctions; supplying a clean shape map largely restores conjunctive search; and shape heads, but not color heads, carry the increase in binding errors with object proximity. These results identify a source of binding failure upstream of spatial indexing: the feature maps themselves become unreliable in clutter, with the shape map forming the main bottleneck. Our contributions are:

- **Feature maps are built on the query.** Causal tracing localizes the conjunctive–disjunctive gap to attention heads at the query's feature-word positions (Section 3).
- **These heads compute atomic spatial feature maps.** Transplanting feature-map heads between minimal-pair displays induces illusory conjunctions in 81% of trials for shape heads and 99% for color heads, compared with 0% for random heads. Position-randomized sources reduce these rates

to 25% and 38%, respectively, showing that the heads encode where a feature occurs rather than whether the target conjunction is present (Section 4.1).

- **The shape map is the bottleneck for binding.** Supplying a clean shape map raises conjunctive-search accuracy from 32% to 94%, whereas the corresponding color intervention has little effect. Shape heads, but not color heads, also carry the proximity dependence of illusory conjunctions (Sections 4.2–4.3).
- **Feature maps and spatial indices support distinct binding regimes.** Ablating the feature maps disrupts color, shape, and conjunction counting (e.g., 55.8% $\rightarrow$ 11.3%) while sparing grid-based spatial-relation judgments. This points to at least two binding routes: word-anchored feature maps that become unreliable in clutter, and content-independent spatial indexing (Assouel et al., 2026) that supports binding in sparse, structured displays (Section 5). "'

Together, these results help reconcile two lines of prior work. VLMs behave like pre-attentive vision in cluttered search (Campbell et al., 2024), yet they have a spatial indexing mechanism that binds features in sparse displays (Assouel et al., 2026). We show that the difference arises upstream of indexing: the feature maps that indexing operates over are built at the query's feature words and break down as distractors accumulate. Whether the indices themselves also degrade with set size remains to be tested. In this view, self-attention reproduces the parallel, feature-based stage of human visual attention but not its serial, object-by-object stage.

These results help reconcile two lines of prior work. In cluttered search, VLMs behave like pre-attentive vision (Campbell et al., 2024), yet in sparse displays they bind features through a spatial indexing mechanism (Assouel et al., 2026). Our results point to a source of failure upstream of indexing. Before indexing can bind features, the model must first build a map of where each feature appears, and it does so at the query's feature words. In cluttered displays, these maps, especially the shape map, become unreliable: supplying a clean shape map restores search. Whether the indices themselves also degrade as distractors increase remains to be tested directly. Within a single forward pass, the model therefore appears to implement a parallel, feature-based stage of search, but not the serial, object-by-object selection that supports reliable binding in human vision.

2 Related Work

2.1 Binding Problem

VLM failures on visual search and other multi-object tasks have been attributed to the binding problem, in which objects that share attributes interfere during feedforward processing (Campbell et al., 2024). Haputhanthri et al. (2026) show that VLMs trained to point to objects through explicit spatial coordinates attend to objects one at a time, and that fine-tuning models to point before answering eliminates binding errors. Representational analyses show that VLMs develop compositional representations whose geometry predicts downstream failures (Li et al., 2025; Savietto et al., 2026). Assouel et al. (2026) identify content-independent spatial indexing mechanisms: attention heads assign index-like representations to objects and later use them to retrieve object features, so that errors in spatial retrieval produce binding failures. These accounts explain how already-represented features are associated with objects. We instead localize where those features are represented and why they fail under clutter.

The clearest precedent for a staged account comes from vision transformers solving relational tasks. Lepori et al. (2024) show that ViTs decompose same–different judgments into a perceptual stage, which extracts object attributes into disentangled subspaces, and a relational stage that operates over them. Their concern is the relation between two objects; ours is the binding of two features within one. Both presuppose per-attribute representations built before the combinatorial step, and it is those representations that we localize here. In diffusion transformers, Wang et al. (2026) find cross-attention heads that translate relational text into spatial gradients over image tokens, while separate heads route object identity to those locations, suggesting that a division between what and where is not peculiar to discriminative VLMs.

2.2 Functional Specialization of Attention Heads

Mechanistic interpretability has identified attention heads with reproducible, causally important roles, such as induction heads that reuse earlier token transitions (Olsson et al., 2022), the name-mover and S-inhibition heads of indirect-object identification (Wang et al., 2023), and copy-suppression heads (McDougall et al., 2024). These head families are defined by the operation they perform or their role in a circuit. The heads we identify are instead selective for a feature dimension, carrying spatial information about shape or about color, so their causal role depends on what kind of perceptual content they represent. Feature-selective units are familiar from vision encoders (Bau et al., 2017), and color is represented in language models in a perceptually structured way (Abdou et al., 2021; Patel & Pavlick, 2022), but neither line localizes a causal mechanism in the attention heads of the language backbone, which a binding account requires. Recent work has begun to localize the mechanisms by which the language backbone of VLMs grounds, tracks, and verbalizes visual content (Cui et al., 2026; Gandikota & Bau, 2026; Feucht et al., 2026).

3 Localizing Mechanisms of Visual Search

3.1 Task Setup

To isolate mechanisms responsible for binding in visual search, we construct a dataset similar to that of Campbell et al. (2024). The model is asked to find a target that is a conjunction of a color and a letter. Surrounding the target, we develop two sets of distractors. The disjunctive set surrounds the target with distractors of the same shape, but a different color, and in the conjunctive set, distractors can share either the target's shape or its color, but not both. An example of these displays is shown in Figure 1. Our dataset includes 6 different colors and 6 letters. We include a varying number of distractor counts from 0-35. Each distractor count includes 1125 pairs of target-present and target-absent cases for both the conjunctive and disjunctive settings. Furthermore, across disjunctive and conjunctive pairs, we ensure the target and distractor locations are fixed. Throughout our experiments we use Qwen2.5-VL-7B (Bai et al., 2025; Yang et al., 2024). Our results verify that the disjunctive setting shows consistent near-perfect performance regardless of the number of distractors, whereas the conjunctive setting shows a consistent decline in performance as the number of distractors increases.

3.2 Causal Tracing

To understand the drop in performance and the difference in computation between the conjunctive and disjunctive settings, we follow the causal mediation approach of Vig et al. (2020); Meng et al. (2022) by intervening on attention blocks and measuring the indirect effect on the model's output.

$$\text{CMM} = \begin{cases} \text{logit}_{\text{True}} - \text{logit}_{\text{False}}, & \text{if the ground truth is True,} \\ \text{logit}_{\text{False}} - \text{logit}_{\text{True}}, & \text{if the ground truth is False,} \end{cases} \tag{1}$$

For each individual run, we quantify how confident the model is in the correct answer using the logits of the outputs True and False. We term this the Correct Margin Metric (CMM), which is shown in Equation 1. We then patch a span of 5 attention blocks at the last 15 token positions from the disjunctive to the corresponding conjunctive setting and measure the change in CMM. Our scenes include 35 distractors and our dataset has 450 target present and absent conditions (900 total) across the various colors and shapes to ensure these components generalize.

Our results, displayed in Figure 2, highlight that the tokens that refer to the target ("shape" and "color") alongside the question mark token and the final token have a significant causal effect on the final output. Overall, we notice a very localized computation, as the majority of components show no causal effect. Additionally, we notice that over the target shape token, the intervention has a positive causal effect when the target is present and a negative causal effect otherwise. We notice that the attention blocks over the "shape" token result in the largest causal effects.

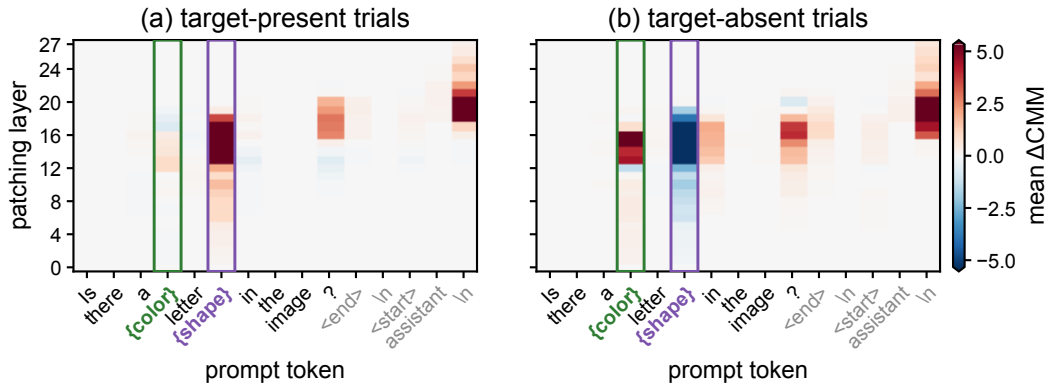


Figure 2: **Causal tracing results when patching disjunctive attention blocks into the conjunctive setting.** Each point patches up to five consecutive layers starting from the layer shown on the y-axis. Results split by (a) target present and (b) target absent.

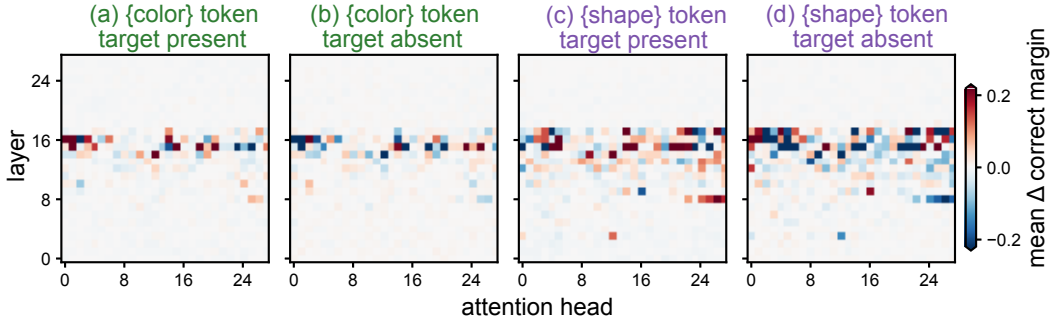


Figure 3: **Localizing the causally influential heads by patching from disjunctive to conjunctive visual search.** Component outputs are patched from the 35-distractor disjunctive condition into the corresponding 35-distractor conjunctive condition, and we report the mean change in the Correct Margin Metric (Eq. 1) over 900 matched pairs. (a) and (b) report the target present and absent splits for the color tokens. (c) and (d) report these for the shape token.

3.3 Localizing Attention Heads

To understand the role of individual attention heads over causally relevant tokens, we conduct a similar patching experiment as the previous section. Specifically, we patch attention heads over the shape tokens from the disjunctive setting to the conjunctive one and measure the difference in CMM. We conduct a similar patching experiment over the color tokens, except we source activations from a disjunctive setting with distractors that are the same color but different shapes, as sourcing from this setting showed larger causal effects (Section A.2).

In Figure 3 we see that the effect of patching individual heads varies greatly in opposite directions for target-present and target-absent cases. This difference suggests that the set of heads supporting target identification can be harmful for target rejection when sourced from the disjunctive setting. Regardless, the magnitude of their causal effect remains consistently large. Furthermore, fewer heads show large individual effects at the color token than at the shape token. Inducing illusory conjunctions nonetheless requires patching more color heads (10%) than shape heads (1%; Section 4.1), which suggests that the color signal is spread across many heads with small individual effects. Interestingly, for both the shape and color tokens, the causally relevant heads stop at layers 17 and 16, respectively.

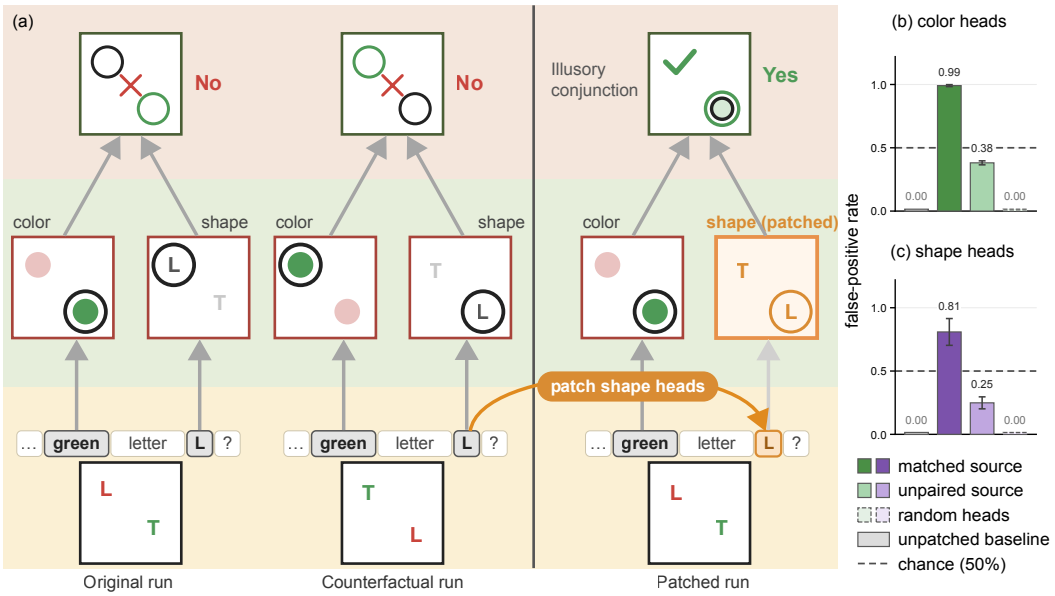


Figure 4: **Inducing illusory conjunctions through a minimal-pair swap. (a)** Schematic for the query *"Is there a green letter L?"*. The original display (left) contains a red L and a green T; the color and shape maps peak at different locations, so the model answers No. The counterfactual (middle) swaps the two objects and is also answered No. Patching the shape heads from the counterfactual into the original run (right) moves the shape-map peak onto the green object, and the model reports an illusory conjunction (Yes). **(b, c)** False-positive rate on target-absent scenes after patching the top 10% of color heads (b) or the top 1% of shape heads (c). Matched: the source swaps the two objects' locations. Unpaired: source object locations are randomized. Random heads: an equal number of random heads at the same token. Baseline: the unpatched original scene. Dashed line: chance.

4 Feature-Map Heads

4.1 Characterizing Heads

We find that the causally relevant attention heads encode parallel feature maps for shape and color. Our findings in the previous section showed that the heads do not encode information about target presence, as they have a positive effect when patching between target-present cases, and have the opposite effect otherwise. Motivated by these findings, we test whether the heads with the largest absolute causal effects encode atomic features (shape/color in isolation), rather than the conjunction of two features, by devising a simple patching experiment. We ask the model if a conjunction such as "green letter L" exists in a scene and provide it with an image of a green T and red L. To understand whether these heads encode a single feature (color or shape), we run a counterfactual setting where the green T and red L are swapped and patch either the shape heads or color heads into the original setting. If the model can perceive a green L, the heads likely encode atomic features; otherwise, they may encode the conjunction of the two concepts. In Figure 4, we include a visualization of the experiment.

We find that under patching, the model consistently hallucinates the target, suggesting that these heads encode the color and shape features in isolation. Patching 1% of the attention heads over the shape token results in the model hallucinating the target 81% of the time. Similarly, patching 10% of the top causally ranked heads over the color token results in the model hallucinating the target 99% of the time. The effects of patching more heads are discussed in Section A.3. To ensure the hallucination is due to a swap of spatial positions, we patch these heads sourced from a run where the positions of the two objects are randomly placed. We find that this significantly reduces the hallucination rate, implying the heads encode the spatial attributes of the features. Furthermore, as a control, we patch a random 1% of heads over the shape token and a random 10% of heads over the color token and find that they have no effect, ensuring that the ranked heads identified mediate

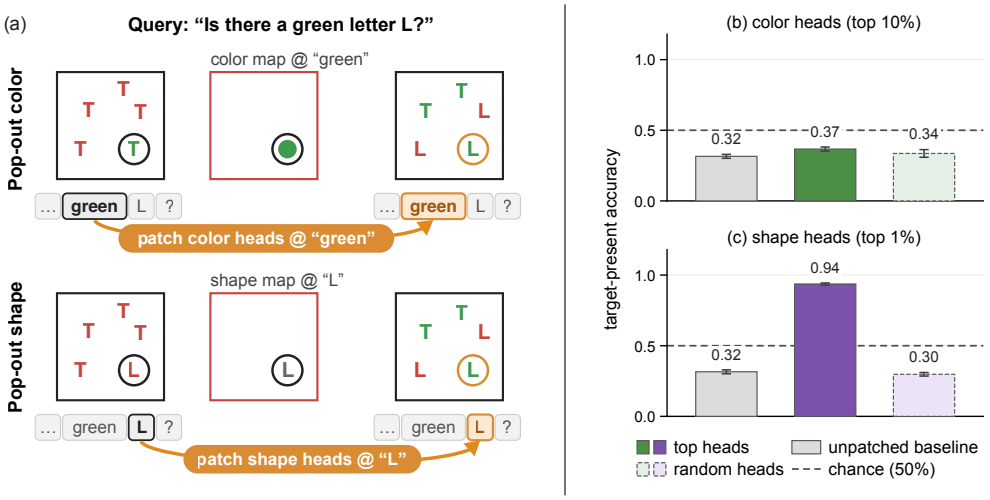


Figure 5: **A clean shape map alone restores conjunctive search.** **(a)** Query *"Is there a green letter L?"*. Heads at the color or shape word are patched into a conjunctive, target-present scene (right) from a pop-out source with matched object positions. The color pop-out has a green T among red Ts; the shape pop-out has a red L among red Ts. No source object carries both target features, so only single-feature spatial information is transferred. **(b, c)** Target-present accuracy on 15-distractor conjunctive trials after patching the top 10% of color heads (b) or top 1% of shape heads (c), averaged over 1,125 trials. Error bars: SEM. Dashed line: chance.

this effect. In Section A.4, we find that these characterized feature maps are necessary for binding in visual search by ablating the heads on conjunctive and disjunctive settings.

4.2 Recovering Performance Through Shape and Color Heads

While our analysis highlights that a sparse set of attention heads can induce an illusory conjunction, we question whether intervening on color or shape heads can improve performance in the conjunctive setting. We construct counterfactuals for the conjunctive setting by building scenes in which the object at the target position pops out from the surrounding distractors by either color or shape alone. For example, if the conjunctive scene contains a green L target among green T and red L distractors, the color pop-out places a green T at the target location and replaces all distractors with red Ts. Analogously, the shape pop-out places a red L at the target location and again replaces all distractors with red Ts. All object positions are kept identical to the original scene. Notably, in both settings the target location carries only the target's color or only its shape, never both. This ensures that the patched heads transfer single-feature (feature-map) information rather than porting over the final decision observed in the source. Furthermore, because no distractor contains either target feature, any recovered performance is likely to arise from transferring information at the target position rather than from a distractor. A visualization of this patching experiment is provided in Figure 5. We intervene on the 15-distractor, target-present conjunctive trials by patching the top 1% of shape heads from the shape pop-out scene and the top 10% of color heads from the color pop-out scene, using the same text prompt as the target. In Figure 5 we see that patching shape heads results in a larger performance increase, to 93.7% from 31.6% accuracy, whereas color heads have little to no effect. In Section A.6 we further investigate limitations of feature-map heads. These results suggest that the drop in performance under the conjunctive setting can be causally linked to the shape heads' inability to encode the target location.

4.3 Illusory Conjunctions Are Mediated By Shape Heads

Failures of feature binding are commonly attributed to illusory conjunctions, that is, when features of separate objects are incorrectly bound together. To observe whether the feature maps causally mediate these emergent binding failures, we conduct a simple patching experiment. In Section A.5, we find that as two objects get closer together, the chance of illusory conjunctions steadily increases.

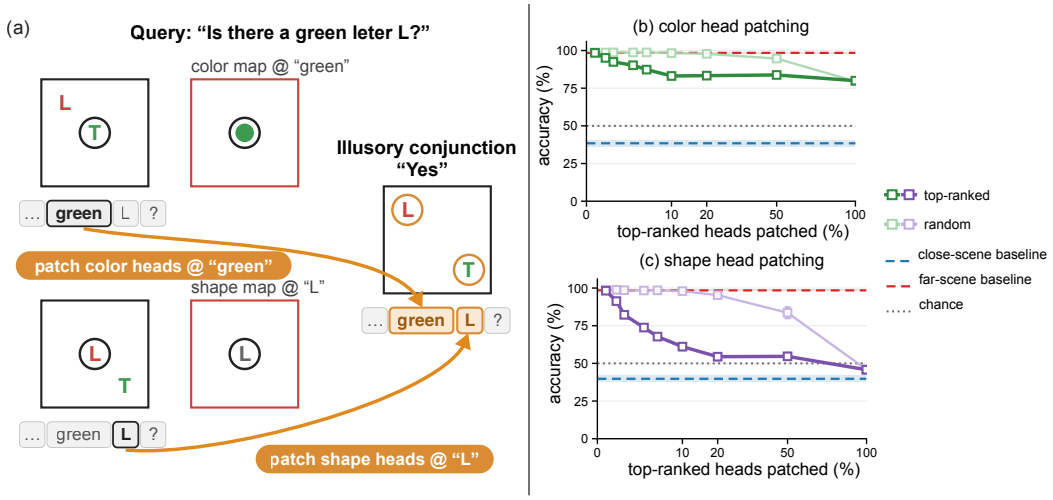


Figure 6: **Shape heads mediate the proximity dependence of illusory conjunctions. (a)** Query *"Is there a green letter L?"* on target-absent scenes with a red L and a green T. In the far base scene (∼70 px apart), the model correctly answers No. Source scenes move one object closer: the green object for color heads, the L for shape heads. **(b, c)** Accuracy on far scenes versus the percentage of top-ranked color (b) or shape (c) heads patched from the close scene. Dashed lines: unpatched close (blue) and far (red) accuracy. Faded lines: random heads.

We hypothesize that this could be due to either shape or color heads lacking fine-grained spatial encodings of objects. We therefore ask whether a target conjunction (e.g., a green L) is present when its color and shape are split across two objects (e.g., a green T and a red L). In the baseline image the objects are spaced 1.4 plot units (approximately 70 pixels) apart on a random axis, such that the model never hallucinates the target. We then patch the shape heads from a run where the target shape is moved closer to the target color object along that axis. Similarly, we patch the color heads from a run where the target color is moved closer to the target shape and record model performance. As shown in Figure 6, patching the heads over the "shape" token results in the model behaving nearly identically to the close objects baseline, whereas patching the heads over the "color" token has a much smaller shift, implying it does not mediate this behavior. Overall, this suggests that illusory conjunctions may arise out of a failure of shape heads to precisely encode spatial information.

5 Generalization of Feature-Map Heads

5.1 Feature-Map Heads Are Necessary for Counting

While the shape and color heads play a significant causal role in visual search, we aim to understand whether they are necessary for other tasks which require binding. To this end we investigate the necessity of these heads in counting and spatial reasoning. For the counting task we construct scenes with 6 objects varying along two random colors and two random letters. For each image we ask the model three questions: 1) "How many {color} letters are in the image?" 2) "How many letter {shape} are in the image?" 3) "How many {color} letter {shape}s are in the image?". In Figure 7, we find that both the top 5% of color and shape heads are causally necessary for completing the tasks. We observe that for color and shape, counting performance drops by over 50%, and for conjunction counting, the drop is even more significant. As a control, we ablated 5% random heads over these tokens and found that they had no drop, hence showing the importance of these feature-map heads in binding across tasks.

5.2 Feature-Map Heads Are Not Necessary for Spatial Relations

Another task that requires binding is spatial relation tasks. For example, in a task where there are 4 objects in a grid of varying colors and shapes, we ask the model four questions: What is the

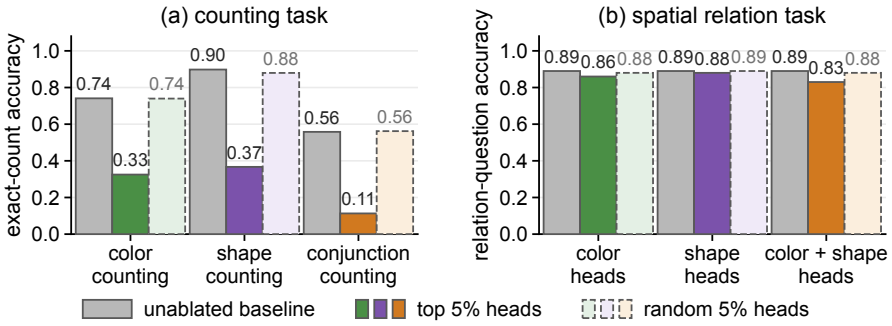


Figure 7: **Necessity of feature-map heads in other tasks.** Accuracy on (a) color, shape, and conjunction counting and (b) spatial-relation questions after ablating the top 5% of color heads, shape heads, or both, compared with the unablated baseline and an equal number of random heads.

color/shape of the object to the right/bottom of the {color} {shape}? Figure 7 shows that ablating the identified feature-map heads leaves spatial-relation accuracy near baseline. Consistent with Assouel et al. (2026), the model may instead rely on a separate spatial-indexing mechanism when reasoning over these more structured displays. This dissociation suggests that the mechanisms supporting binding may depend on both the structure of the input and the demands of the downstream task.

6 Discussion and Conclusion

We asked why VLMs, although built on self-attention, fail at binding much like pre-attentive human vision. Our answer is that the model builds a separate map for each feature as it reads the question, and that the shape map is the bottleneck for binding in clutter. Although transformers differ from the brain in data, objectives, and architecture, these mechanisms parallel those of primate vision. The query words act like a search template, much as neurons in monkey visual cortex respond more strongly to every item sharing the target's color or shape (Bichot et al., 2005), guided by prefrontal representations of the searched-for feature (Bichot et al., 2015). Binding errors also depend on distance in both systems. In human vision, crowding mixes the features of nearby objects and produces illusory conjunctions (Pelli et al., 2004; Pelli & Tillman, 2008). Crowding has neural correlates in visual cortex (Kim & Pasupathy, 2024) and is consistent with lateral connections that link nearby locations (Bosking et al., 1997; Angelucci et al., 2002). In the VLM, this distance dependence is carried by the shape heads (Section 4.3), and it is not simply a tokenization effect, since accuracy changes smoothly across separation increments much smaller than an image token.

One reason VLMs look pre-attentive may lie in self-attention itself. Because it weights tokens by how well their content matches the query, heads at the shape word likely spread over every object of that shape, much like feature-based attention, and softmax normalization may divide attention among all matches. Cortical responses are also normalized (Carandini & Heeger, 2012), but there, attention can bias normalization toward a single object (Reynolds & Heeger, 2009). Humans resolve conjunctive search by selecting one location at a time (Treisman & Gelade, 1980), a serial process that depends on feedback beyond the initial feedforward sweep (Kreiman & Serre, 2020); models of attention likewise distinguish binding within a single feedforward pass from recurrent binding over several task-tuned passes (Tsotsos et al., 2008). The VLM uses the task, since its maps are built at the query words, but only within a single forward pass. Location-based indices (Assouel et al., 2026) may support binding in sparse, structured displays (Section 5), whereas training models to point to objects through explicit coordinates, which turns generated tokens into a feedback loop, reduces binding errors (Haputhanthri et al., 2026). Within a single pass, VLMs thus appear to implement the parallel, feature-based stage of attention but not the serial, feedback-dependent one. Our analyses use a single model, Qwen2.5-VL-7B, and whether other VLMs build the same maps remains to be tested. Because a small set of heads carries these maps and can be patched and ablated directly, VLMs offer a model system for testing hypotheses about binding that are difficult to test in the brain.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability, including correcting typos and grammar. All authors reviewed the final text, and we take responsibility for the full content of this work, including any text produced with the aid of generative AI.

REFERENCES

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? A case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132, 2021.

Alessandra Angelucci, Jonathan B. Levitt, Emma J. S. Walton, Jean-Michel Hupé, Jean Bullier, and Jennifer S. Lund. Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19):8633–8646, 2002.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *International Conference on Learning Representations*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017.

Narcisse P. Bichot, Andrew F. Rossi, and Robert Desimone. Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, 308(5721):529–534, 2005.

Narcisse P. Bichot, Matthew T. Heard, Ellen M. DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–844, 2015.

William H. Bosking, Ying Zhang, Brett Schofield, and David Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of Neuroscience*, 17(6):2112–2127, 1997.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor W. Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113436–113460, 2024.

Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

Kelly Cui, Nikhil Prakash, Shoval Messica, Ayush Raina, David Bau, Antonio Torralba, and Tamar Rott Shaham. The dual mechanisms of spatial variable binding in vision-language models, 2026. URL https://arxiv.org/abs/2603.22278.

Sheridan Feucht, Benno Krojer, Sarah Wang, Henry Abrahamsen, Byron C. Wallace, and David Bau. Using ocr heads to verbalize image semantics, 2026. URL https://arxiv.org/abs/2609.18823.

Rohit Gandikota and David Bau. Gaze heads: How vlms look at what they describe, 2026. URL https://arxiv.org/abs/2606.14703.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

Udith Haputhanthri, Declan Campbell, Rim Assouel, Jonathan D. Cohen, and Taylor W. Webb. Binding visual features point by point. *arXiv preprint arXiv:2605.25427*, 2026.

Taekjun Kim and Anitha Pasupathy. Neural correlates of crowding in macaque area V4. *Journal of Neuroscience*, 44(24):e2260232024, 2024.

Gabriel Kreiman and Thomas Serre. Beyond the feedforward sweep: Feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464(1):222–241, 2020.

Michael A. Lepori, Alexa R. Tartaglini, Wai K. Vong, Thomas Serre, Brenden M. Lake, and Ellie Pavlick. Beyond the doors of perception: Vision transformers represent relations between objects. In *Advances in Neural Information Processing Systems*, volume 37, pp. 131503–131544, 2024.

Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P. Kording. Does object binding naturally emerge in large pretrained vision transformers? In *Advances in Neural Information Processing Systems*, volume 38, pp. 3783–3812, 2025.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–363, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022.

Denis G. Pelli and Katharine A. Tillman. The uncrowded window of object recognition. *Nature Neuroscience*, 11(10):1129–1135, 2008.

Denis G. Pelli, Melanie Palomares, and Najib J. Majaj. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, 4(12):1136–1169, 2004.

John H. Reynolds and David J. Heeger. The normalization model of attention. *Neuron*, 61(2): 168–185, 2009.

Pieter R. Roelfsema and Thomas Serre. Feature binding in biological and artificial vision. *Trends in Cognitive Sciences*, 30(3):190–191, 2026.

Daniele Savietto, Declan Campbell, André Panisson, Marco Nurisso, Giovanni Petri, Jonathan D. Cohen, and Alan Perotti. The geometry of representational failures in vision language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

John K. Tsotsos, Antonio J. Rodríguez-Sánchez, Albert L. Rothenstein, and Eugene Simine. The different stages of visual recognition need different attentional binding strategies. *Brain Research*, 1225:119–132, 2008.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Binxu Wang, Jingxuan Fan, and Xu Pan. Circuit mechanisms for spatial relation generation in diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23612–23621, 2026.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

A Appendix

A.1 Dataset Details

A.1.1 Visual Search Dataset

To study why VLMs fail on visual search, we generate a synthetic dataset including both conjunctive and disjunctive settings. For each setting we test across varying distractor counts including 0, 5, 15, 25, 35. Our dataset includes 3 shape configurations (L/T, O/N, X/S), as well as 3 color configurations (green/red, blue/orange, purple/yellow) corresponding to opposite ends of the color wheel. This results in 3 shape x 3 color = 9 total configurations. For each configuration we have 250 examples (125 target present and 125 target absent). In total, this yields 2,250 scenes per search setting at each distractor count. Furthermore, for each conjunctive trial there exists a "counterfactual" disjunctive trial. In the counterfactual setting, the target (if present) and distractors are placed at the same exact locations. For causal tracing and head localization, we use 50 target-present and 50 target-absent pairs from each of the 9 configurations in the 35-distractor setting, yielding 900 pairs in total. In Figure A1, we show the prompt used for all visual search tasks.

A.1.2 Minimal-Pair Swap Dataset

For the minimal-pair swap dataset used in Section 4.1 we generated a dataset of controlled image pairs. Each scene contains exactly two objects: one that shares the target's color but has the distractor's shape, and one that shares the target's shape but has the distractor's color. For example, when the target is a green L, each scene contains a green T and a red L. Neither object matches the target conjunction, so every scene is a target-absent conjunctive trial (which the model can correctly answer). We use the same 3 shape configurations (L/T, O/N, X/S) and 3 color configurations (green/red, blue/orange, purple/yellow) as the visual search dataset, which gives 3 shape x 3 color = 9 total configurations. For each configuration we generate 100 pairs, yielding 900 pairs (1,800 images) in total. Each pair consists of an "original" image and a "counterfactual" image. In the counterfactual, the two objects swap positions. For each configuration, the model is asked a binary question about the target conjunction, which once again does **not** exist in the image (e.g., "Is there a green letter L in the image?").

A.1.3 Performance Recovery Dataset

The performance recovery experiment in Section 4.2 uses the 15-distractor, target-present conjunctive trials from our visual search dataset as targets. As before, we use 3 shape configurations (L/T, O/N, X/S) and 3 color configurations (green/red, blue/orange, purple/yellow), giving 9 total configurations with 125 target-present trials each, or 1,125 target scenes in total. For every target scene we generate two matched pop-out sources, a *color pop-out* and a *shape pop-out*. In the color pop-out, the object at the target location takes the target's color and the distractor's shape, and in the shape pop-out it takes the target's shape and the distractor's color. In both cases every remaining object is replaced by a distractor-color, distractor-shape item. For example, for a green L target among green T and red L distractors, the color pop-out is a green T among red Ts, and the shape pop-out is a red L among red Ts. All object positions are identical to the original conjunctive scene. As a result, each source differs from its target only in the features of its objects, and the target location carries exactly one of the two target features. Additionally, because no distractor in a source contains either target feature, the only target-relevant information in a source is at the target location.

A.1.4 Illusory Conjunction Dataset

For the illusory conjunction patching experiment in Section 4.3, we generate a dataset of two-object scenes in which the target's color and shape are split across two separate objects. Each scene contains one object with the target's color and the distractor's shape, and one object with the distractor's color and the target's shape. For example, for a green L target, each scene contains a green T and a red L. Since no object matches the target conjunction, the correct answer is always that the target is absent. We use the same 3 shape configurations (L/T, O/N, X/S) and 3 color configurations (green/red, blue/orange, purple/yellow) as the visual search dataset, giving 9 total configurations. For each of the 9 configurations we include 50 sets of scenes. For each set we render three conditions. In the

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>Answer the following question
with exactly one word:

True

or

False

Do not provide any explanation.

Question:
You are presented with an image containing letters.

Is there a color letter shape in the image?<|im_end|>
<|im_start|>assistant
```

Figure A1: Prompt used for all visual search experiments, including visual search, minimal-pair swap, performance recovery, and illusory conjunctions. The **bold** color and shape words are replaced with the target features of each configuration.

far condition, which serves as the patching target, the two objects are separated by 1.4 plot units (approximately 70 pixels), a distance at which the model does not report the target. In the *shape-close* condition, the color object is held fixed and the shape object is moved 0.7 plot units toward it along the same axis, halving the separation to 0.7 plot units (approximately 35 pixels). In the *color-close* condition, the shape object is instead held fixed and the color object is moved toward it by the same amount. Consequently, the two close conditions have the same separation, and each differs from the far scene only in the position of one object. During patching, we patch shape heads at the shape token from the shape-close scene, and color heads at the color token from the color-close scene, into the far scene, using the same text prompt for all scenes (e.g., "Is there a green letter L in the image?").

A.1.5 Counting Dataset

To test whether the color and shape heads identified in visual search generalize to other tasks that require feature binding, we generate a synthetic counting dataset. Each scene contains 6 letters. For each scene we sample 2 of 6 colors (green, red, blue, orange, purple, yellow) and 2 of 6 letters (L, T, O, N, X, S), and each object is assigned one of the two selected colors and one of the two selected letters. As a result, every scene contains between 2 and 4 distinct color–shape conjunctions, and the counts for any single feature and for any conjunction differ across scenes. In total the dataset contains 1,000 scenes. For each scene we ask three questions about the first selected color and shape: a *color* question ("How many green letters are in the image?"), a *shape* question ("How many letter Ls are in the image?"), and a *conjunction* question ("How many green letter Ls are in the image?"). Full prompts are shown in Figure A2. This yields 3,000 questions in total. The model is instructed to answer with exactly one word from zero to six, and we score its response over these seven answer tokens.

A.1.6 Spatial Relation Dataset

To test whether the color and shape heads identified in visual search are also needed for relational reasoning, we generate a synthetic spatial relation dataset. Each scene contains 4 letters arranged in a 2×2 grid, one letter per cell. Each letter's color is sampled independently from 6 colors (green, red, blue, orange, purple, yellow), and its shape independently from 6 letters (L, T, O, N, X, S). Because colors and shapes are sampled independently, the other objects in a scene frequently share a color or a shape with one another. In total the dataset contains 1,000 scenes. In every scene, the top-left letter serves as the reference object, and we refer to it by its conjunction of color and shape (e.g., "the green letter L"). For each scene we ask four questions about the reference object's neighbors: the color of the letter to its right ("What is the color of the letter to the right of the green letter L?"), the identity of the letter to its right ("What letter is to the right of the green letter L?"), and the same

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>Answer the following question
with exactly one word:

zero, one, two, three, four, five, or six.

Do not provide any explanation.

Question:
You are presented with an image containing letters.

How many color letter shapes are in the image?<|im_end|>
<|im_start|>assistant
```

Figure A2: Prompt used for the counting task conjunction question. For the single-feature questions, the final line is replaced with "How many **color** objects are in the image?" or "How many letter **shape**s are in the image?". The **bold** color and shape words are replaced with the first selected color and shape of each scene.

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>Answer the following question
with exactly one word:

green, red, blue, yellow, purple, or orange.

Do not provide any explanation.

Question:
You are presented with an image containing colored letters arranged in
a 2x2 grid.

What is the color of the letter to the right of the color letter
shape?<|im_end|>
<|im_start|>assistant
```

Figure A3: Prompt used for the spatial relation task and the right-color question. The remaining questions replace the final line with "What letter is to the right of the **color** letter **shape**?", "What is the color of the letter below the **color** letter **shape**?", and "What letter is below the **color** letter **shape**?". We ask the model to report either the letter or color to observe robustness to prompt. For the letter questions, the answer options are instead "L, T, X, S, O, or N." The **bold** color and shape words are replaced with the features of the top-left reference letter in each scene.

two questions for the letter below it. The full prompt is shown in Figure A3. This yields 4,000 questions in total. In 66% of scenes, at least one other object shares either the reference's color or its shape. In these scenes the reference can only be located by binding both of its features, so answering correctly requires both locating the reference by its conjunction and then retrieving a feature of a spatially adjacent object.

A.2 Localizing Color Heads

In our disjunctive-to-conjunctive causal tracing experiment, we found that the most causally influential heads existed over the shape tokens. This could be explained by the disjunctive setting containing only one shape and having color differ at the target. To test for this, we run the same causal tracing experiment, except we source from a disjunctive setting where the distractors are the same color as the target but a different shape. For example, if the original disjunctive scene had a target green L and the distractors were red Ls, then this alternative causal tracing scene contains the same target except the distractors are green Ts. As shown in Figure A4, the color token now has the largest causal effect. Furthermore, we see that the trend of improving performance under target

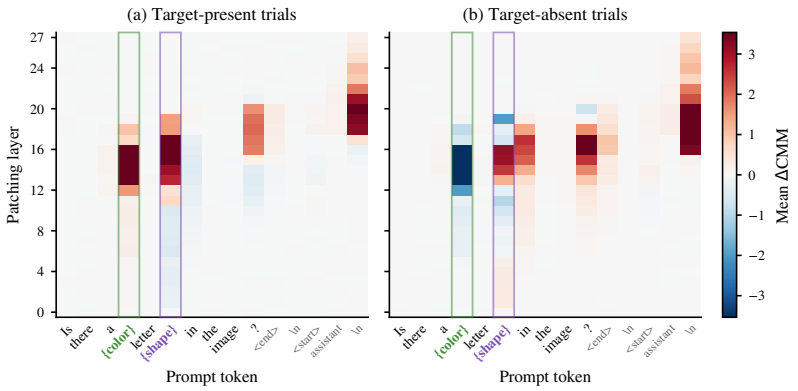


Figure A4: **Causal tracing results when patching disjunctive attention blocks into the conjunctive setting.** Each point patches up to five consecutive layers starting from the layer shown on the y-axis. Results split by (a) target present and (b) target absent.

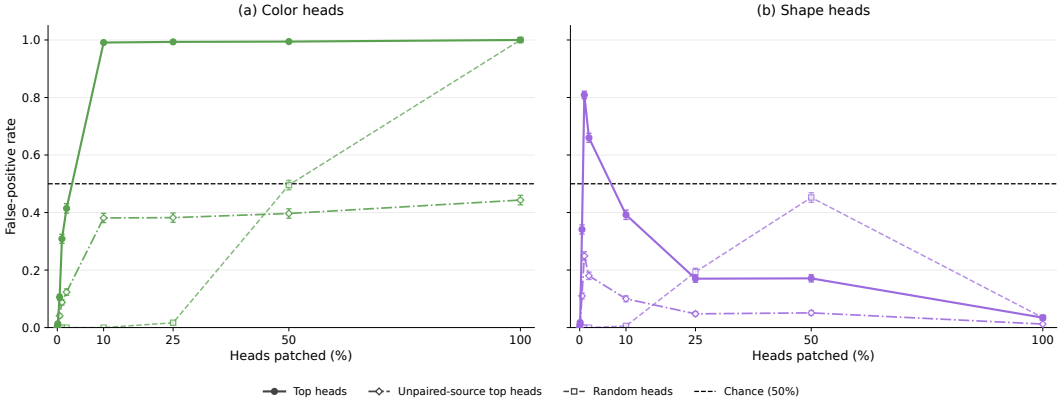


Figure A5: **Minimal-pair swap experiment over top ranked heads.** Illusory conjunction rate over the top shape and color heads identified by absolute causal score in disjunctive to conjunctive patching.

present and harming performance under the absent case occurs at the color token. Together with the results in the rest of the paper, this likely occurs because these blocks encode features atomically. For example, in the original disjunctive source every object is an L, so the shape map at the shape token marks every object location; patching it into the conjunctive case makes every green T read as a green L, leading the model to always answer true. While this behavior is asymmetrical, it is a very useful setting for discovering feature-map heads. Throughout our causal experiments we set up settings and counterfactuals such that performance improvements could not be due to "non-target-hallucination".

A.3 Percentage Of Heads

To determine which heads encode atomic features rather than their conjunction, we need to ensure they correctly induce a hallucination in the experiment described in Section 4.1. As shown in Figure A5, we see that for color heads, the heads induce the illusory conjunction once 10% of the heads are patched, whereas the shape heads peak in inducing the illusory conjunction at 1%, and then taper off afterwards, implying less causally relevant heads for mediating search behavior introduce some type of conjunction or alternatively interference. One reason for this could be that when a scene has a larger amount of distractors, the model is not able to consistently use heads which encode conjunctions and relies on these feature-map heads, although we leave verification of this hypothesis to future work. Nevertheless, we use the observations to motivate why we define the 1% of shape

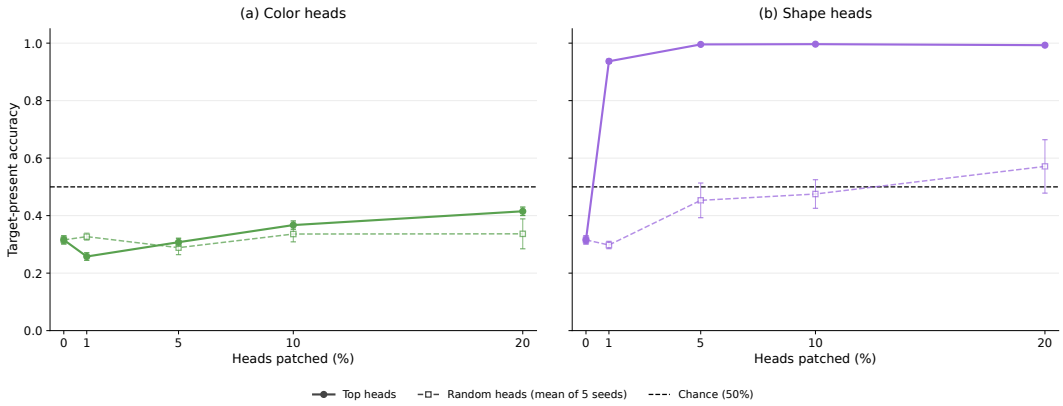


Figure A6: **Pop-out patching experiment over top ranked heads.** Accuracy of patched conjunctive setting. The top shape and color heads identified by absolute causal score in disjunctive to conjunctive patching.

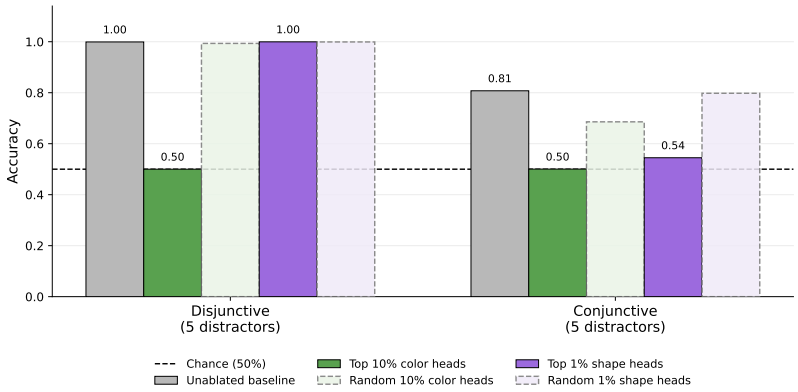


Figure A7: **Feature-specific heads are necessary for visual search.** We ablate the top-ranked color heads (10%; green) and shape heads (1%; purple), alongside equally sized random-head controls (light, dashed bars) over the corresponding tokens, on disjunctive and conjunctive trials with five distractors. The dashed horizontal line denotes chance accuracy (50%).

heads and 10% of color heads as the feature-map heads used in Section 4.2. However, we provide performance from patching more percentages of these heads in Figure A6.

A.4 Shape and Color Heads Are Necessary for Search

To understand whether the feature-map heads we identified are necessary for allowing the model to do visual search, we ablate the heads over the 5-distractor conjunctive and disjunctive search datasets. In Figure A7, we see that for the conjunctive setting, ablating either the color heads or the shape heads reduces the model's performance to chance, whereas ablating a random set of heads over those tokens has little effect. In the disjunctive setting, color heads reduce performance to chance, whereas shape heads cause no performance decrease. This asymmetry is expected because every object in the disjunctive display has the same shape, so the target can be identified by locating the uniquely colored green object, making the color heads—but not the shape heads—necessary. Overall, these results show that the identified feature maps are necessary computational components of binding in visual search.

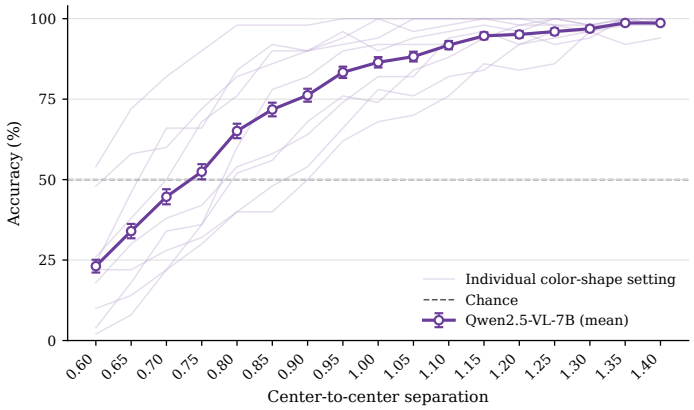


Figure A8: **Illusory conjunctions depend on spatial separation.** Accuracy of Qwen2.5-VL-7B in target recognition when the target is absent. Distance between objects is shown on the x-axis.

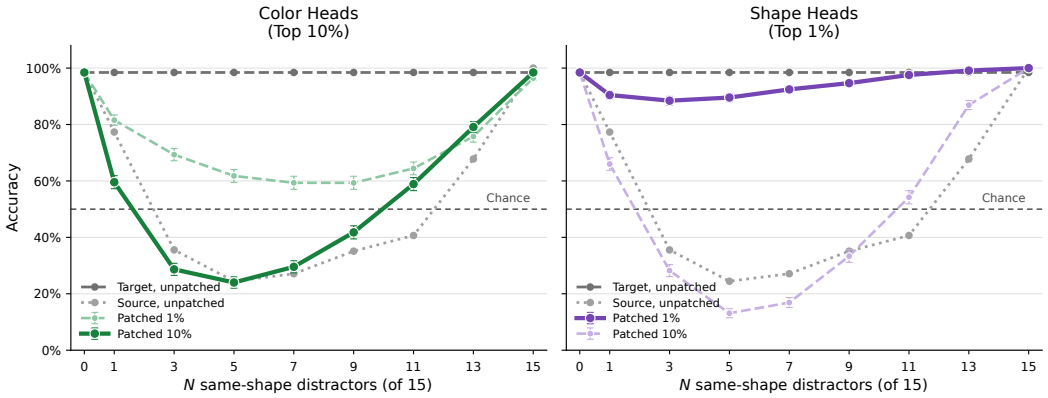


Figure A9: **Interference from increasingly conjunctive source scenes.** We patch color and shape heads from source scenes containing N same-shape/different-color distractors into a fixed 15-distractor color-disjunctive target scene. At $N = 0$ and $N = 15$, the source target is identifiable by shape and color, respectively; intermediate values require feature binding. Thick lines denote the primary set of heads used in previous experiments (top 10% color heads; top 1% shape heads); shaded lines show the alternate patching percentage.

A.5 Illusory Conjunctions In VLMs

Given these spatial atomic feature maps, we hypothesize that illusory conjunctions arise when their positional encoding is inaccurate. We start by behaviorally asking whether distance between two objects correlates with accuracy. We ask the model "Is there a {color} letter {shape} in the image?" and provide it with a scene of an object of the same color but different shape, and same shape but different color. We test across increasing increments of 0.05 plot units from 0.60 plot units (∼30 pixels) to 1.40 plot units (∼70 pixels). In Figure A8, we find that as the distance between the two distractors increases, the illusory conjunction rate consistently decreases. This suggests that tokenization alone may not explain the error, and a lack of fine-grained spatial encoding from the feature-map heads could be to blame, as we explore in Section 4.3.

A.6 Feature-Map Heads Are Susceptible To Interference

While performance can be recovered through the shape heads, we sought to understand fundamental limitations of these feature-map heads. Specifically, whether they face interference under more complex scenes. To study this, we use a 15-distractor scene with a colored target and distractors of the same color but a different shape. Under this condition, the model can always identify the target.

We then replace the top shape and color heads with runs from more complex scenes. Specifically, we successively replace distractors with those that are the same shape but a different color than the target. Notably, at the two endpoints, the target is identifiable by a single feature: shape when no distractors are replaced and color when all distractors are replaced. In contrast, intermediate scenes contain both same-color/different-shape and same-shape/different-color distractors, so neither feature alone identifies the target. The binding demand is therefore greatest near the midpoint, where the two distractor types are approximately balanced. After patching the heads, we utilize the resulting accuracy decline as a proxy to identify potential capacity limitations in these heads: as the scene contains more target shape/colors distractors, the heads must represent an increasing number of feature locations. As shown in Figure A9, the set of color heads that we have identified seems to be more susceptible to interference compared to the shape heads. This seems to be partially explained by the differing number of heads in the two sets. Interestingly, however, we previously noted that shape heads bottlenecked conjunctive recovery, such that a stable patching of them could recover full performance. Taken together with these results, this could suggest a functional hierarchy in which color heads provide a context-sensitive priority signal that is vulnerable to competition, whereas a sparse subset of shape heads carries the spatial information on which the final target decision depends. Interference in the color pathway may therefore prevent the model from selecting or constructing the relevant shape representation; once a clean shape representation is supplied directly, it is sufficient to restore performance.