ATTENTION SINKS IN VLMS ARE BORROWED, NOT SEEN

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how to detect and mitigate visual attention sinks in vision-language models (VLMs). Visual attention sinks are image tokens that absorb a large share of attention over the image regardless of the query or task, and they mislead methods that treat attention maps as a record of where the model "looked." Prior work has detected sinks through massive activation dimensions and proposed redistributing their attention, but none have addressed what makes a token become a sink. Comparing six candidate detectors on ten VLMs from four families, we find that only the centered cosine similarity between an image token's key and the key of the language model's text-sink detects visual sinks. Thus, a visual sink is an image token whose key aligns with the text-sink's key. Moving a sink's key component along this direction to the median of the image keys sends its attention back to the text-sink (**Are Borrowed**) while leaving model answers unchanged (**Not Seen**). Building on this, we propose a training-free, entropy-triggered re-look that first cleans the attention map, then crops the most-attended region of the cleaned map and inserts it into the reasoning trace. Cleaning adds negligible computation to the forward pass, and the re-look improves accuracy with a mean gain of $+4.3$ points on fine-detail VQA benchmarks across model families.

1 INTRODUCTION

The attention map of a vision–language model is widely read as a record of where the model looks. It is used to trace how object information flows from image tokens to text (Kaduri et al., 2025; Neo et al., 2025; Jiang et al., 2025) and to guide model behavior: cropping or highlighting the most-attended region (Zhang et al., 2025a; Shen et al., 2025; Liu et al., 2026), reading out grounding heads (Kang et al., 2025a), pruning the least-attended tokens (Chen et al., 2024; Zhang et al., 2025d; Yang et al., 2025), and reweighting attention against hallucination (Huang et al., 2024; Liu et al., 2024; Che et al., 2025). However, this record is often misleading as a few image tokens, the visual attention sinks (Kang et al., 2025b), take a large share of the image attention regardless of the question. Any method that interprets or uses VLM attention can be misled by these tokens.

Attention sinks were first observed in language models as a few tokens that absorb much of the attention regardless of their content (Xiao et al., 2024; Su et al., 2026). These tokens carry massive activations (Sun et al., 2024) and their keys act as a bias rather than as content (Gu et al., 2025; Barbero et al., 2025). Every VLM inherits such text-sinks from its language model. Visual attention sinks have been identified through massive hidden dimensions (Kang et al., 2025b), attributed to the vision encoder, language model, rotary position encoding (Darcet et al., 2024; Luo et al., 2026; Choi et al., 2026; Zhang et al., 2026), and either removed or exploited by downstream methods (Li et al., 2026; Shukla & Kira, 2026). However, none of these studies has examined what a visual attention sink is or what makes an image token become one. Without this, it is unclear whether the sinks reported in different VLMs are the same phenomenon and whether they carry useful information.

We study visual attention sinks in ten VLMs from four families and find that they are identified by their keys. The features prior work associates with sinks, the hidden-state norm (Kang et al., 2025b) and the value and key norms (Sun et al., 2024), each detect sinks in only some model families. Since a text-sink's key acts as a bias (Gu et al., 2025), we build three detectors that test whether image sinks share this bias: the spread of the attention logits a token receives across queries, and the raw and centered cosine similarity between an image token's key and the text-sink key. Only the centered

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.

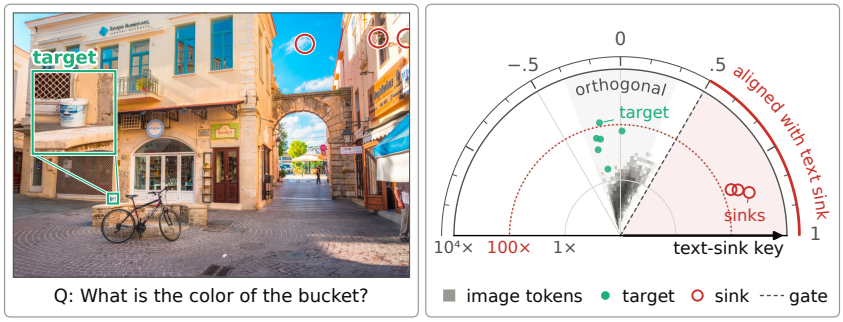


Figure 1: **Visual attention sinks point the same way as the text-sink.** Left: visual attention sinks (red, $\geq 100\times$ uniform attention outside the target) and the target (green) in a V* image. Right: key direction (angle) vs. attention (radius) of every image token

cosine detects sinks in every sink-bearing model (AUC ≥ 0.99). Thus, a visual attention sink is an image token whose key points along the text-sink key.

This definition also gives a way to remove visual attention sinks. At every layer, we flag the image tokens whose centered cosine with the text-sink key exceeds a threshold, and move the component of each flagged key along the text-sink direction to the median over image keys, leaving queries and values unchanged. In most sink-bearing models, this removal reveals two properties. First, visual attention sinks *are borrowed*. The attention the visual attention sinks held returns to the text-sink rather than spreading over the image. Second, visual attention sinks are *not seen*. The model's output distribution barely changes, with low KL divergence from the original distribution. The attention map, however, does change, with the most-attended image token often landing on the queried object.

With the visual attention sinks removed and the model's behavior preserved, the attention map is again a record of where the model looks. Reasoning VLMs increasingly revisit the image mid-trace by cropping or zooming, but they learn when and where to do so through reinforcement learning Zheng et al. (2026); Su et al. (2025); Zhang et al. (2025b); Lai et al. (2026) or rely on a separate search or tool calling procedure (Wu & Xie, 2024; Shen et al., 2025; Hu et al., 2024). We introduce a training-free re-look for revisiting the image only when the model becomes uncertain. At the first decoding step where the next-token entropy exceeds a threshold (i.e., the model is uncertain; Wang et al., 2025a), we crop the most-attended region of the cleaned map from the original image and insert it into the trace. Across 10 models V* (Wu & Xie, 2024), HR-Bench (4K and 8K) (Wang et al., 2025c), and MME-RealWorld-Lite (Zhang et al., 2025c), our method improves accuracy on 39 with a mean gain of $+4.3$ points. Because the key edit is cheap and leaves the model's outputs essentially unchanged, the cleaned map could replace the raw one wherever attention guides a VLM, from interpreting where the model looks to providing annotation-free grounding signals in RL pipelines.

2 Related Work

2.1 Attention sinks

Attention sinks are the few tokens that absorb much of the attention regardless of content (Xiao et al., 2024). They carry massive activations (Sun et al., 2024), and their key acts as a bias while their value is near zero (Gu et al., 2025). Vision transformers show the same on high-norm patches (Darcet et al., 2024). A VLM inherits the text sink from its language model, and in some families image tokens behave like one as well. Kang et al. (2025b) find these visual attention sinks through massive activations in a few hidden dimensions. Later work attributes them to rotary decay (Zhang et al., 2026), separates encoder-born from language-model-born sinks (Choi et al., 2026), or reads them as carriers of semantics (Luo et al., 2026).

Prior work treats the attention on a visual sink as misplaced and redistributes it over the image (Kang et al., 2025b) or intervenes on it against hallucination (Huang et al., 2024; Liu et al., 2024). We find otherwise: in the full-attention models, removing a sink returns its attention to the text sink, not to the target, and the answers stay nearly the same. What changes is the attention map, which matters

to every method that reads one, from attention-guided cropping to token pruning and grounding (Zhang et al., 2025a; Chen et al., 2024; Kang et al., 2025a).

2.2 Visual Tools in Multi-modal Reasoning

Attention to the image fades as the trace grows (Sun et al., 2025), and a self-reflective phrase in the trace does not make the model look again (Shi et al., 2026). A phrase injected from outside helps only over the tokens already in context (Bi et al., 2026). Other methods act on the image. Visual programming composes vision experts and code around a language model (Gupta & Kembhavi, 2023; Surís et al., 2023), Visual Sketchpad adds a drawing canvas (Hu et al., 2024), and a line of work trains a crop or zoom action into the trace (Zheng et al., 2026; Su et al., 2025; Zhang et al., 2025b). Training-free methods pick the region from the model's confidence (Shen et al., 2025) or attention (Zhang et al., 2025a; Gao et al., 2025). only Look Twice (Morini et al., 2026) filters sink tokens first with the score of Kang et al. (2025b).

Our re-look differs in where, what, and when. It reads the attention map after the key edit removes the sinks. It inserts a fresh crop as a second image block rather than encoded tokens (Gao et al., 2025) or a phrase (Bi et al., 2026). And it fires at the first uncertain step of the trace, since a few high-entropy tokens set its direction (Wang et al., 2025a).

3 Attention Sinks in VLMs Are Borrowed, Not Seen

We first identify which VLMs have visual attention sinks (§ 3.2) and show that a sink is an image token whose key aligns with the text-sink key (§ 3.3). We then remove the sinks and measure where their attention goes and how the model's output changes (§ 3.4).

All analyses use ten models from the Qwen3-VL (QVL; Bai et al., 2025), InternVL (IVL; Zhu et al., 2025; Wang et al., 2025b), Gemma-4 (G4 Team et al., 2026), and Pixtral (Pix Agrawal et al., 2024) families on V* (Wu & Xie, 2024), whose target boxes let us separate sinks from the queried object. HR-Bench 4K and 8K (Wang et al., 2025c) have no target boxes but give matching results (§ A.12).

3.1 Definitions

A prompt ($\mathcal{IQ}$) is the n image tokens $\mathcal{I}$ and its text question tokens $\mathcal{Q}$. For layer ℓ and head h, $\alpha_{i,j}^{\ell,h}$ is the prefill softmax attention from token i to j. We measure the attention each token receives from the question and each image token's share of in-image attention:

$$a_j^\ell = \frac{1}{H|\mathcal{Q}|}\sum_h \sum_{i\in\mathcal{Q}} \alpha_{i,j}^{\ell,h}, \qquad r_j^\ell = \frac{a_j^\ell}{\sum_{j'\in\mathcal{I}} a_{j'}^\ell} \;\; (j \in \mathcal{I}). \tag{1}$$

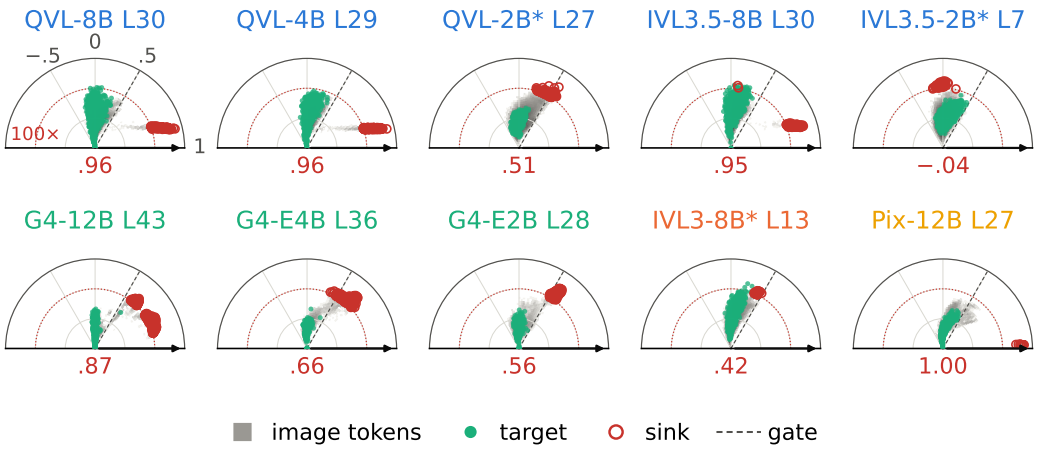


Figure 2: Image tokens on V* at each model's peak layer. Angle: centered cosine of a token's key with the text-sink key (Equation 2). Radius: attention received relative to uniform, log scale. Sinks: tokens above $100\times$ uniform (dotted ring) outside the target box. Dashed ray: gate threshold .5. Red numbers: median cosine of the sinks; *: sink-free

Uniform attention gives $r_j^\ell = 1/n$, so we report a token's attention relative to uniform ($n\, r_j^\ell$).

Every model except Gemma-4 is full-attention. Most Gemma-4 layers instead use a sliding window.

Sinks. The text-sink $t^\star$ is the text position that receives the most attention from the question, $t^\star = \arg\max_{j \notin \mathcal{I}} a_j^\ell$, at the middle full-attention layer. A visual attention sink at layer ℓ is an image token that receives at least $100\times$ its uniform share, $n\, r_j^\ell \geq 100$, and lies outside the target bounding box $\mathcal{B}$ of the queried object.

Key directions. The key of token j at layer ℓ and key–value head g is $k_j^{\ell,g}$, the mean image key is $\bar{k}^{\ell,g} = \frac{1}{n}\sum_{j\in\mathcal{I}} k_j^{\ell,g}$, and the direction of the text-sink key is $\hat{u}^{\ell,g} = k_{t^\star}^{\ell,g}/\|k_{t^\star}^{\ell,g}\|$.

Sink-bearing and free models. The visual attention sink share $s^\ell = \sum_{j\in\mathcal{V}^\ell} r_j^\ell$ of layer ℓ is the in-image attention held by its visual attention sinks $\mathcal{V}^\ell$, averaged over images. A model's peak layer $\ell^\star$ is the layer with the largest s^ℓ in the last three quarters of its language model.[1] A model is sink-bearing if $s^{\ell^\star} > 0.15$ and sink-free otherwise. We define this threshold only for analysis as a sink-free model under this definition may still have a few weak visual attention sinks. They key edit experiments in § 3.4 treats every model alike, removing little from sink-free ones.

3.2 Visual attention sinks come from the language model

In Table 1, we report the number of visual attention sinks per image, their attention relative to uniform, and their share of the in-image attention at each model's peak layer.[2] We find that seven of the ten models are sink-bearing, with sinks holding a quarter to four fifths of the in-image attention, and three are sink-free, with at most 12%.

Table 1: Visual attention sinks in VLMs on V* at each model's peak layer. key1: sinks per image; key 2: their attention relative to uniform; key 3: their share of the in-image attention.

Model	LM	per image	×uniform	share
QVL-8B	Qwen3	3.9	1259	.66
QVL-2B	Qwen3	2.7	163	.12
IVL3.5-8B	Qwen3	1.5	342	.24
IVL3.5-2B	Qwen3	1.0	151	.08
IVL3-8B	Qwen2.5	0.4	93	.02
Pix-12B	Mistral	1.0	2093	.76
G4-12B	Gemma-4	4.4	205	.63
G4-E2B	Gemma-4	2.6	488	.81

The language model largely decides which group a model falls in. The Qwen3, Mistral, and Gemma-4 backbones produce sinks, while the Qwen2.5 backbone of InternVL3-8B does not. The clearest case is InternVL3-8B and InternVL3.5-8B, which share a vision encoder but use Qwen2.5 and Qwen3 respectively. With these models visual attention sinks take 2% of the in-image attention in the first and 24% in the second. Within a family, smaller models have weaker sinks. Visual attention sinks take 66% and 73% of the in-image attention in Qwen3-VL-8B and 4B but only 12% in Qwen3-VL-2B, and 24% in InternVL3.5-8B but 8% in InternVL3.5-2B.

Gemma-4 is sink-bearing for a different reason. Most of its layers attend only within a sliding window and the question sits too far after the text-sink to reach it. In these layers a few image tokens take 63% to 81% of the in-image attention. In its full-attention layers, which do reach the text-sink, no image token takes a large share.

3.3 A sink is an image key aligned with the text-sink key

Prior work identifies sinks by the size of their activations. We test three such features, the hidden-state norm (Kang et al., 2025b) and the value and key norms (Sun et al., 2024), as detectors of visual attention sinks at each model's peak layer. We propose a detector based instead on the direction of a token's key. The text-sink acts as a bias that attracts every query regardless of content (Gu et al., 2025), so if image tokens become sinks by the same mechanism, their keys should point along the text-sink key. We measure this with the centered cosine similarity between an image token's key and

[1] The first quarter is excluded because target localization is still weak there (Appendix A).

[2] Table 5 gives all ten models and adds the text-sink and localization columns.

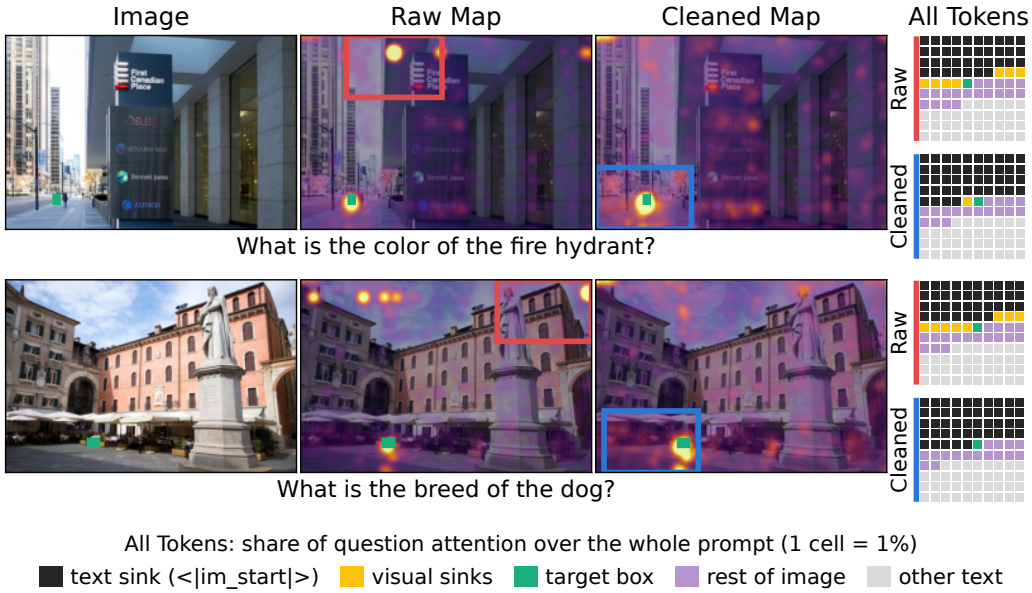


Figure 3: Crop selection in Qwen3-VL-8B-Thinking on two V* images. Green: target box; red/blue: raw/edited window. Right: question attention over the whole prompt (one cell per 1%). The edit moves the visual attention sinks' share to the text-sink.

the text-sink key, which subtracts the mean image key $\bar{k}$ to remove the offset all image keys share

$$\tilde{c}_j = \frac{\langle k_j - \bar{k},\ k_{t^\star} - \bar{k} \rangle}{\|k_j - \bar{k}\|\,\|k_{t^\star} - \bar{k}\|}, \qquad \tilde{c}_j^{\,\ell} = \frac{1}{H_{\mathrm{kv}}} \sum_g \tilde{c}_j^{\,\ell,g}, \tag{2}$$

where the first expression is computed per layer and the key–value head, and the second averages it over the H_{kv} key–value heads. We also test two simpler versions of the same idea, the raw cosine without centering and the logit spread, the standard deviation of $\langle q_i, k_j \rangle / \sqrt{d}$ over question tokens i.

Detecting sinks. In Table 3, we report each feature's AUC at separating visual attention sinks from other image tokens, as the mean and minimum over the seven sink-bearing models. We find that only the centered cosine separates sinks in every model, with a minimum AUC of 99.0, while the features from prior work are family-specific. Hidden-state and value norms separate sinks in the Qwen3 backbones and Pixtral but not reliably in Gemma-4, and key norm works only with its sign reversed, since sinks have unusually small keys, and still fails on Gemma-4. Of the other two new features, logit spread works only on Pixtral, and the raw cosine matches the centered one except on Gemma-4-E2B, where it drops to 94.0. Figure 2 shows this separation as a compass. In every sink-bearing model the sinks are the only tokens that both point toward the text-sink key and receive high attention: their median cosine is .95 to 1.00 in the full-attention models and .56 to .87 in Gemma-4, while the target and the other image tokens stay near cosine 0 and receive far less attention.

Editing keys. To test whether the alignment causes sink attention, we edit keys at the peak layer of the four sink-bearing full-attention models on 20 V* images.[3] First, we remove the direction. Projecting $\hat{u}$ out of the sink keys *drops their share of in-image attention to zero*, while projecting out a random direction leaves it unchanged. Second, we write the direction into other tokens. Setting the keys of eight random image tokens to the text-sink key *gives them 91% to 97% of the in-image attention*, while a random key of the same norm gives them at most 9%.

Alignment with the text-sink key is thus the signature of a visual attention sink. It separates sinks in every sink-bearing model, and in the models we edit it is both necessary and sufficient for sink attention. Because this direction belongs to the language model, the signature is also consistent with the backbone deciding which models have sinks (§ 3.2).

[3]Full results are in Table 7. For Pixtral, whose image keys as a whole sit at an offset along $\hat{u}$, we move the sink keys' component to the median image key instead of zero (§ A.4).

Table 2: Effects of the key edit on V*. L: number of language-model layers. text-sink: share of question attention on the text-sink, averaged over the probed layers

Model	Map $\max_j r_j$	Map top-1 $\in \mathcal{B}$	Map text-sink	Model accuracy	Model agreement	Model KL	Model caption
Qwen3-VL-8B	.25 → .04	.00 → .47	.57 → .67	.73 → .73	.99	.004	.83
Qwen3-VL-4B	.24 → .05	.00 → .41	.54 → .68	.73 → .73	.98	.003	.82
Qwen3-VL-2B	.04 → .03	.28 → .23	.63 → .63	.68 → .61	.81	.039	.57
InternVL3.5-8B	.34 → .05	.00 → .72	.44 → .65	.71 → .68	.94	.001	.85
InternVL3.5-2B	.04 → .04	.60 → .59	.65 → .65	.64 → .61	.93	.002	.80
InternVL3-8B	gate off	.59	.50 → .50	.70 → .71	.97	.001	.87
Pixtral-12B	.57 → .02	.00 → .16	.40 → .51	.50 → .50	.97	.002	.73
Gemma-4-12B	.15 → .13	.01 → .00	–	.68 → .65	.85	.531	.73
Gemma-4-E4B	.32 → .30	.00 → .00	–	.44 → .44	.96	.031	.70
Gemma-4-E2B	.41 → .39	.00 → .00	–	.39 → .36	.78	.132	.66

Gemma-4. Gemma-4's sinks carry the same signature but play a different role. Shrinking the image until the whole prompt fits inside the sliding window moves a third to half of the question's attention in these layers back to the text, and attention on the strongest image token falls by two thirds or more (§ A.5). Its image sinks therefore stand in for a text-sink the model cannot reach. Unlike the sinks of the full-attention models, they carry values of ordinary size and dominate the attention output of their layers, so the model does read them (§ A.7).

Table 3: AUC of each feature at separating visual attention sinks from other image tokens, at the peak layer on V*. Mean and min over sink-bearing models.

Feature	mean	min
hidden norm	80.1	25.0
$-\|v\|$	72.0	8.0
$-\|k\|$	77.4	49.0
−logit std	78.4	65.0
raw key cos	99.0	94.0
centered key cos	99.9	99.0

3.4 Removing the sinks leaves the model unchanged

Removing visual attention sinks. Since the signature is a single direction, a sink can be removed by editing its key along that direction. At each layer and key–value head, an image token passes the gate when its centered cosine exceeds a threshold τ_c (Equation 2). If more than a fraction ρ of the layer's image tokens pass, the alignment belongs to the image as a whole rather than to a few sinks, and we leave the layer unedited. For each token that passes, we move its component along $\hat{u}$ to the median over image keys,

$$k_j \leftarrow k_j - \big(\langle k_j, \hat{u}\rangle - \mathrm{median}_{j' \in \mathcal{I}} \langle k_{j'}, \hat{u}\rangle\big)\, \hat{u}, \tag{3}$$

with $\tau_c = 0.5$ and $\rho = 0.02$ for all models. In the full-attention models, any τ_c from 0.4 to 0.9 gives the same edit (§ A.6). The target is the median rather than zero because the image keys as a whole can sit at an offset along $\hat{u}$, as in Pixtral, where a key with zero component would still stand out (§ A.4). Queries and values are untouched, so a gated token keeps its content and loses only its pull through the sink direction.

Where the attention goes. In Table 2, we report the effect of removing the sinks on the attention map and on the output for all ten models.[4] We find that the attention the sinks held is borrowed from the text-sink. In the four sink-bearing full-attention models, the edit raises the text-sink's share of question attention by 0.10 to 0.21, while attention on the target box stays where it was. Forcing the freed attention onto the image instead raises target attention by up to $2.6\times$ but leaves accuracy unchanged, and it disturbs the output more than the key edit does (§ A.9).

Effect on the output. The model does not see the sinks. In the same four models, accuracy shifts by at most three points and first-token KL from the unedited model is of order 10^{-3} nats. The reason is that sinks pass on almost nothing. At the peak layer they hold 29% to 77% of the in-image attention but supply only 3% to 16% of the attention output norm, since their values are small and copy the text-sink's value direction (§ A.7). Gemma-4 is the exception. As § 3.3 showed, the model

[4] HR-Bench gives the same results (Table 12).

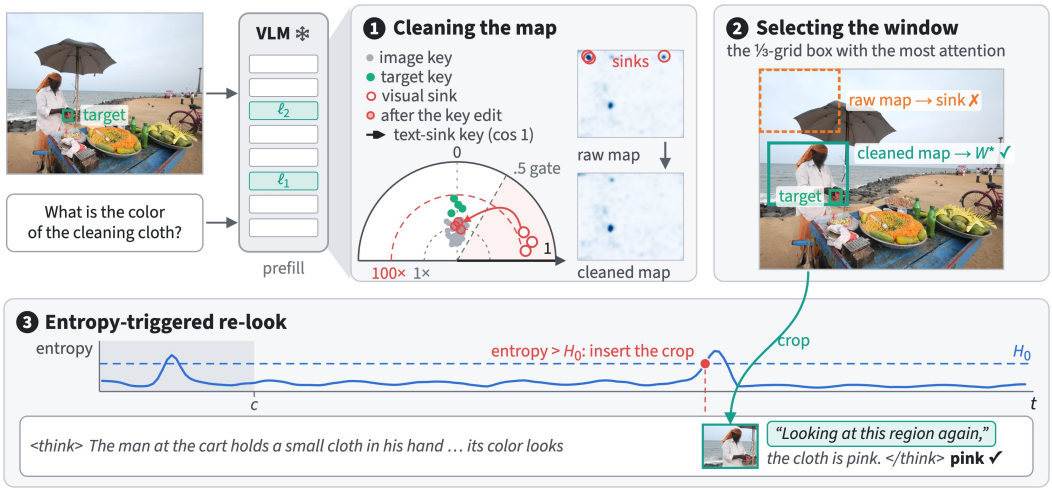


Figure 4: Overview of the entropy-triggered re-look. (1) The key edit removes the visual attention sinks from the prefill attention map. (2) The selector keeps the $\frac{1}{3}$-grid window with the most attention on the cleaned map. (3) When the next-token entropy first exceeds H_0 after c tokens of thought, the crop of that window is inserted into the chain of thought.

reads its image sinks, and removing them changes the output (KL up to 0.53) without clearing the map, so we do not apply the edit to Gemma-4. On the sink-free InternVL3 the gate admits under one token per image and the edit is an identity. On the two 2B sink-free models, the gate admits weakly aligned tokens and costs up to seven points, and restricting the edit to the upper three quarters of layers restores baseline accuracy.

Effect on the map. The attention map changes sharply. In the four sink-bearing full-attention models, the largest single-token share of in-image attention falls by an order of magnitude, and the most-attended image token, which never lay in the target box before the edit, lands in it on 16% to 72% of images afterwards (Figure 3). The target gains no attention. It becomes the most-attended image token because the sinks no longer outrank it.

4 Method

Section 3 showed that removing visual attention sinks leaves the model's answers unchanged but changes its attention map. The harm sinks do is therefore to anything that reads the map from outside the model. We study this harm through a task where reading the map correctly matters. On fine-detail questions, the evidence is often a small region of a large image, which the model may see only at reduced resolution. As the model reasons toward an answer, it reaches points of high uncertainty (Wang et al., 2025a), and zooming into the relevant region is known to improve accuracy (Zhang et al., 2025a; Shen et al., 2025; Zheng et al., 2026). Finding that region is a map-reading problem, and § 3.4 showed that the cleaned map places its most-attended token on the target far more often than the raw map.

We combine the two in a training-free, entropy-triggered re-look that zooms into the region the cleaned map selects at the moment the model becomes uncertain (Figure 4). First, the key edit of § 3.4 removes the visual attention sinks from the attention map. The edit is applied only in a separate prefill that reads the map, so the model that writes the answer is unedited. Second, a selector reads the cleaned map and picks the image window with the most attention (§ 4.1). Third, during chain-of-thought decoding, a trigger waits for the first step where the next-token entropy exceeds a threshold, then crops the selected window from the original image and inserts it into the trace (§ 4.2). Running the same re-look on the raw map isolates the effect of the sinks.

4.1 SELECTING THE WINDOW

The selector reads the question-to-image attention a_j^ℓ of Equation 1 from the edited prefill, averaged over two layers near half depth where the map responds most to the question, $\bar{a}_j = (a_j^{\ell_1} + a_j^{\ell_2})/2$. Because Gemma-4 reads its image sinks, its selector instead uses two full-attention layers, which reach the text-sink and carry almost no image sinks, and skips the edit.

The window's length is one-third of the side of the image-token grid. For tiled inputs, the tiles are first stitched back together into one grid. From the set of all same-sized windows $\mathcal{W}$ that lie within the grid, the selector chooses the window with the highest aggregate attention by applying a softmax over the region: $W^\star = \arg\max_{W \in \mathcal{W}} \sum_{j \in W} \bar{a}_j$. The window is mapped back to pixel coordinates, padded by 20 pixels on each side, and kept fixed for the rest of the question.

4.2 TRIGGERING AND INSERTING THE CROP

The trigger decides when to insert the crop into the chain of thought. At every decoding step, it computes the entropy of the next-token distribution over the full vocabulary, after temperature scaling and before any top-k or top-p truncation. The crop is inserted at the first step where this entropy exceeds H_0, provided at least c tokens of thought have been generated and decoding is still inside the thought channel. We use $H_0 = 1$ nat and $c = 32$. The trigger fires at most once per question, and the token sampled at that step is kept.

At that step, we cut the selected window from the original pixels, so the crop is a zoom rather than a thumbnail, and encode it with the model's own image processor under the same budget cap as the first image. We append it to the reasoning trace as an ordinary second image, with no special positions or embeddings, followed by the text "Looking at this region again," and decoding continues with the same sampling state. An untriggered question decodes exactly as the baseline.

Every question costs one extra prefill of the original prompt for the selector, and each triggered question adds one image encoding and one image-token block to the context. The weights are unchanged, and the crop and its transition text are the only additions to the prompt.

5 EXPERIMENTS

5.1 SETUP

We run the re-look on the ten models of section 3 with frozen weights. Eight reason in a native thinking mode, and we prompt the other two, the instruction models InternVL3-8B and Pixtral-12B, to produce a chain of thought. We evaluate on V* and HR-Bench 4K and 8K, where the evidence is a small patch of a large image, and on MME-RealWorld-Lite (Zhang et al., 2025c), which has 1,919 five-choice questions on high-resolution photographs, remote-sensing and monitoring imagery, diagrams, and driving scenes. Baseline and re-look runs share the sampling seed, so their traces are identical up to the trigger and every difference after it comes from the crop. To isolate the effect of the sinks, we repeat each run with the window selected on the raw map. We test each (model, benchmark) pair with an exact McNemar test. Prompts and sampling details are in Appendix C.

5.2 RESULTS

Overall. In Table 4, we report the baseline accuracy and the paired gain from one re-look for all 40 (model, benchmark) pairs. The re-look improves accuracy on all but one pair, by 4.3 points on average. The gains concentrate on the fine-detail benchmarks, where the evidence is a small patch of a large image, averaging 5.0 points on V* and HR-Bench against 2.1 on MME-RealWorld-Lite, whose questions cover a wider range of tasks.

Raw versus cleaned map. The Raw columns of Table 4 repeat each run with the window selected on the raw map. Gemma-4 has no raw run, since its selector already skips the edit (§ 4.1). Over the 28 pairs with both runs, the cleaned map gains 4.1 points on average and the raw map 2.9, and the whole gap comes from the models with sinks. In Qwen3-VL-8B, Qwen3-VL-4B, and Pixtral-12B, the raw map gives back one to five points per pair, five pairs become significantly worse, and two,

Table 4: Base accuracy (%) and paired re-look gain. *Clean* Δ: crop selected on the sink-cleaned map. *Raw* Δ: on the raw attention map. Shading scales with gain (green) or loss (red); †: $p < .05$ (exact McNemar, clean)

	V*			HR-4K			HR-8K			MME-RW-Lite		
Model	Base	Raw Δ	Clean Δ	Base	Raw Δ	Clean Δ	Base	Raw Δ	Clean Δ	Base	Raw Δ	Clean Δ
Qwen3-VL-Thinking												
8B	79.1	+2.1	+3.7	72.6	+0.8	+4.1†	68.9	+2.0	+4.4†	46.5	+1.0	+2.0†
4B	82.2	−2.1	+2.6	74.1	−0.5	+0.9	67.4	+1.5	+5.1†	45.1	+0.8	+3.4†
2B	70.2	+3.1	+1.6	69.5	+3.0	+3.0†	63.9	+4.1	+3.1†	47.2	+2.1	+2.0†
InternVL3.5 (thinking)												
8B	69.1	+7.3	+8.9†	64.5	+6.2	+6.1†	59.0	+7.8	+7.0†	48.7	+2.1	+3.1†
2B	58.6	+9.9	+10.5†	62.4	+4.6	+4.2†	51.0	+9.1	+9.5†	44.2	+0.1	+0.4
Gemma-4 (thinking; selector reads the raw map)												
12B	75.9	–	+6.8†	76.4	–	+4.6†	72.0	–	+9.9†	52.5	–	+2.8†
E4B	55.0	–	+11.0†	68.1	–	+1.9	62.4	–	+5.0†	46.4	–	+3.6†
E2B	50.8	–	+5.8	66.3	–	+1.3	57.9	–	+6.6†	44.8	–	−0.5
Chain-of-thought prompted												
InternVL3-8B	70.7	+5.8	+5.8	70.9	+0.2	+0.4	64.2	+5.4	+5.4†	48.8	+0.9	+0.9
Pixtral-12B	48.7	+4.2	+7.3	55.1	−2.1	+1.1	47.0	+0.6	+3.1	41.0	+1.4	+3.7†
Mean (all)	66.0	–	+6.4	68.0	–	+2.8	61.4	–	+5.9	46.5	–	+2.1
Mean (with raw run)	68.4	+4.3	+5.8	67.0	+1.7	+2.8	60.2	+4.4	+5.4	45.9	+1.2	+2.2

Qwen3-VL-4B on V* and Pixtral-12B on HR-4K, turn from gains into losses. The difference is where the crop lands. On Qwen3-VL-8B's V* questions, the raw crop misses the target in 78% of triggers and the cleaned crop in 10%. InternVL3.5-8B, whose sinks are the weakest of the sink-bearing models, sits in between. On the sink-free models, the two maps pick nearly the same crops and the gains match, with InternVL3-8B differing on at most four questions per set. The same split holds on MME-RealWorld-Lite alone, at 2.2 points cleaned against 1.2 raw on average. Cleaning matters where there is a sink to clean.

Image resolution. The re-look helps most when the first pass could not see the evidence clearly. On V* and HR-Bench, Qwen3-VL, which has the largest image budget of the four families, gains about 3 points on average, against about 7 for InternVL3.5 and Gemma-4-12B. Within Qwen3-VL, all three sizes gain significantly on HR-8K, where the first pass sees about a third of the pixels, and none does on V*, where the image already fits the budget. On a large image, the crop recovers pixels the first pass discarded. On an image the model already saw in full, it offers only a second look at the region, which is worth a few points at most.

Trigger coverage. An untriggered question decodes identically in both runs, so every gain comes from the triggered subset. On V* and HR-Bench, the trigger fires on 60% to 78% of questions for Qwen3-VL, 79% or more for InternVL3.5, 91% to 95% for Pixtral-12B, and 96% or more for InternVL3-8B and Gemma-4. On MME-RealWorld-Lite, it fires on 91% or more for every model.

6 Conclusion

A visual attention sink is an image token whose key is aligned with the key of the language model's text-sink. This is the one feature that separates sinks from other image tokens in every family we test. Removing the direction from an image key removes the sink and writing it onto an image key creates one. The sink is inherited from the text-sink of the language model. When we remove it the attention flows back to the text-sink and the model's output does not change. Thus cleaning the attention map is a single cheap key edit. Building on this we design an entropy-triggered re-look. When a high-entropy token appears in the chain of thought the method crops the most-attended region of the cleaned map and inserts it into the trace. This improves accuracy where the same crop method from the raw map does not. Since cleaning the map costs almost nothing the method can also be used inside training pipelines to provide a grounding signal at no additional cost.

AI DISCLOSURE

In this work, we used generative AI tools for three purposes: (1) polishing the writing, (2) retrieval and discovery of related work. We did not use generative AI tools for research ideation or research execution, and the remaining disclosure categories are not applicable to this work.

We use AIs for

- Improve the grammar, wording, and clarity of author-written text. All content, claims, and technical details were written by the authors and every AI-edited passage was reviewed to ensure that the original meaning was preserved.
- AI-assisted search tools were used to supplement our manual literature survey and reduce the risk of overlooking relevant prior work. Every reference surfaced this way was independently located, read, and verified by the authors before being cited.

We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Jing Bi, Filippos Bellos, Junjia Guo, Yayuan Li, Chao Huang, Yunlong Tang, Luchuan Song, Susan Liang, Zhongfei Zhang, Jason J. Corso, and Chenliang Xu. When to think and when to look: Uncertainty-guided lookback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5104–5113, June 2026.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21635–21644. IEEE, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19520–19529. IEEE, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962. IEEE, 2023.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image*f* a deep-dive into the vision of vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558. IEEE, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350. IEEE, 2025a.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025b.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, volume 2026, pp. 76722–76746, 2026.

Shiyu Li, Zi-Yuan Hu, Shijia Huang, Yanyang Li, Yiwu Zhong, and Liwei Wang. Sinkpruner: Sink-free visual token pruning for multimodal large language models. *arXiv preprint arXiv:2609.01004*, 2026.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/76818d8d85e05e45ce3a16a8468619d1-Paper-Conference.pdf.

Jiayun Luo, Wan-Cyuan Chris Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 93626–93660, 2026.

Marco Morini, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Look twice: Training-free evidence highlighting for knowledge-based visual question answering, 2026. URL https://arxiv.org/abs/2604.01280.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.335. URL https://aclanthology.org/2025.emnlp-main.335/.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are vlms seeing or just saying? uncovering the illusion of visual re-examination. *arXiv preprint arXiv:2605.15864*, 2026.

Tripti Shukla and Zsolt Kira. Sage: Sink-aware grounded decoding for multimodal hallucination mitigation. *arXiv preprint arXiv:2603.27898*, 2026.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 8222–8251. Curran Associates, Inc., 2025. doi: 10.52202/085713-0280. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0c38f54740062529aa4117a04b583f3c-Paper-Conference.pdf.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, Jing Xiong, Hui Shen, Keyu Fan, Weihao Ye, Chaofan Tao, Taiqiang Wu, Zhongwei Wan, Tiantian Zhang, Bowen Yan, Zhen Li, Yiming Zhang, Congkai Xie, Yulei Qian, Yuchen Xie, Yik-Chung Wu, Hongxia Yang, and Ngai Wong. Attention sink in transformers: A survey on utilization, interpretation, and mitigation, 2026. URL https://arxiv.org/abs/2604.10098.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5158–5171, 2025.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11854–11864. IEEE, 2023.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana

Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 115452–115486. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3850. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Paper-Conference.pdf.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025c.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=DgaY5mDdmT.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision–language models. *Information Processing & Management*, 63(2):104431, 2026.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms, 2025b. URL https://arxiv.org/abs/2505.15436.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/df29d63af05cb91d705cf06ba5945b9d-Paper-Conference.pdf.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference, 2025d. URL https://arxiv.org/abs/2410.04417.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cdb347d7516d52a9280ea9d6708911f6-Paper-Conference.pdf.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Sink Measurements on V*

A.1 Setup and the ten models

We measure every softmax attention layer and head on V*. Its target boxes cover 0.3 to 0.4% of the image tokens, depending on the model's tokenization. Prompts and image budgets are those of Table 4. Under those budgets a V* image is about 3,300 image tokens for Qwen3-VL and an HR-Bench image about 12,400. InternVL3.5 runs under its R1 system prompt and Gemma-4 under its thinking template. Every number below depends on that prompt. The accuracies in Table 2 and Table 12 are greedy and letter-only with an empty thinking block, so they are not comparable with the chain-of-thought accuracies of Table 4. Table 5 extends Table 1 of the main text with the text-sink position and share and with the localization columns.

Table 5: Text and visual sinks in ten models on V*, extending Table 1. LM is the language-model backbone. Text sink: the token index of the text sink and its share of the question's attention. Visual sink (peak layer): attention relative to uniform, count per image, share of the in-image attention and the peak layer out of the model's layer count. Target top-1$\in \mathcal{B}$: the fraction of images whose most-attended image token lies in the target box, at the peak layer and at the best layer.

		Text sink		Visual sink (peak layer)				Target top-1$\in \mathcal{B}$	
Model	LM	position	share	$\times$uniform	per image	share	layer	at peak	best layer
Qwen3-VL-8B	Qwen3	0	.74	1259	3.9	.66	30/36	.00	.80
Qwen3-VL-4B	Qwen3	0	.69	1049	4.9	.73	29/36	.00	.82
Qwen3-VL-2B	Qwen3	0	.75	163	2.7	.12	27/28	.00	.75
InternVL3.5-8B	Qwen3	0	.65	342	1.5	.24	30/36	.14	.62
InternVL3.5-2B	Qwen3	0	.77	151	1.0	.08	7/28	.00	.69
Gemma-4-12B	Gemma-4	2	.10	205	4.4	.63	43/48	.00	.66
Gemma-4-E4B	Gemma-4	2	.21	444	2.7	.66	36/42	.00	.53
Gemma-4-E2B	Gemma-4	1	.22	488	2.6	.81	28/35	.00	.24
InternVL3-8B	Qwen2.5	2	.62	93	0.4	.02	13/28	.05	.76
Pixtral-12B	Mistral	0	.49	2093	1.0	.76	27/40	.00	.00

Text-sink position. The text-sink $t^\star$ of section 3 is fixed per model. In the full-attention models it is also the text position with the largest share of the question's attention at every layer from about a third of the depth onward. Which position it is depends only on the backbone. Qwen3 (in Qwen3-VL and InternVL3.5) and Mistral (in Pixtral-12B) use the first token, and Qwen2.5 (in InternVL3-8B) uses a newline in the system prompt. The share is .62 to .77 in the full-attention models, except Pixtral-12B at .49. Gemma-4 has only seven or eight full-attention layers. In those layers the sink is the turn marker just after <bos>, at position 2 in Gemma-4-12B and Gemma-4-E4B and at position 1 in Gemma-4-E2B (Table 5). It takes .10 to .22 of the question's attention, a sixth to a third of what the other backbones give their sink.

Which models have visual sinks. At their peak layer Qwen3-VL-8B, Qwen3-VL-4B, InternVL3.5-8B, Pixtral-12B and the three Gemma-4 sizes have image tokens at 200 to $2{,}100\times$ uniform attention. One to five such tokens take a quarter to four fifths of the in-image attention. InternVL3-8B has only a faint version ($93\times$ uniform, 2% share). Qwen3-VL-2B's tokens sit just above $100\times$ uniform and hold only a 12% share. In InternVL3.5-2B a single such token appears at

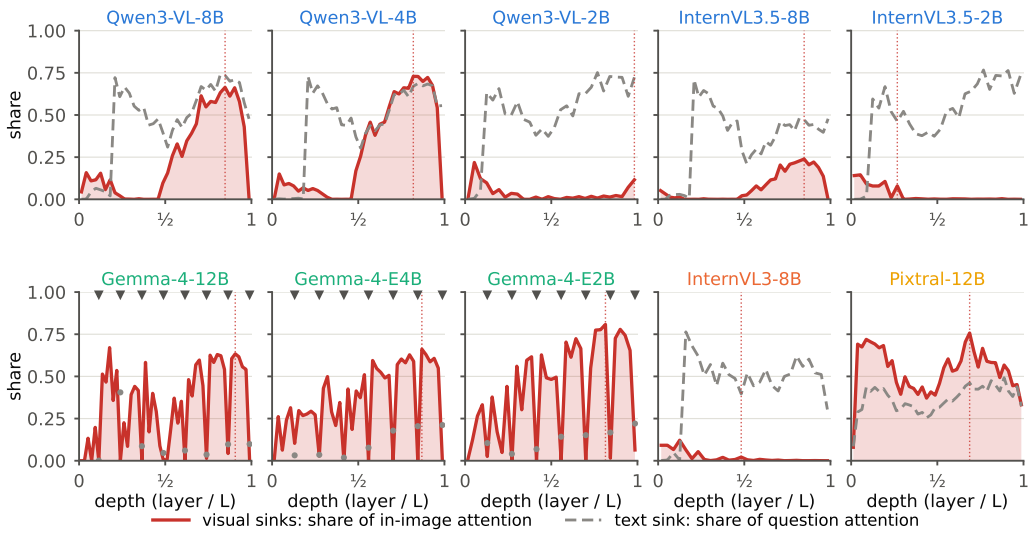


Figure 5: Sink attention across depth on V* for all ten models. Red: visual attention sink share of in-image attention. Gray dashed: text-sink share of question attention. Dotted: peak layer. Gemma-4 reaches the text sink only in its full-attention layers (triangles). Title color marks the language-model backbone (Qwen3, Gemma-4, Qwen2.5 and Mistral).

layer 7. Its key lacks the signature of **??** (centered cosine $-.03$, AUC .42). It holds the same share on every dataset and is gone from layer 8 on. Layer 7 is the first layer the peak-layer rule admits. By the 15% cut of section 3 this makes Qwen3-VL-8B, Qwen3-VL-4B, InternVL3.5-8B, Pixtral-12B and the three Gemma-4 sizes sink-bearing and Qwen3-VL-2B, InternVL3.5-2B and InternVL3-8B sink-free.

Query independence. Does the sink set depend on the question? We replace the V* question by a generic prompt and recompute the late-layer sink set. It is almost unchanged (Jaccard overlap .73 to .92).

Sink depth. Figure 5 shows at which layers the visual attention sinks appear in each of the ten models. Qwen3 sinks switch on near half depth and persist to the penultimate layer. Pixtral-12B has one sink at a fixed grid position at every layer but the first few, and we call it the register. Gemma-4's sinks appear only in its sliding-window layers, where the text-sink is out of reach. In every model but Pixtral-12B the sinks collapse in the last layer. Pixtral-12B's register keeps a third of the in-image attention there.

Localization depth. The raw-map curves of Figure 8 track where the most-attended image token lands at each layer. In Qwen3-VL-8B localization peaks at mid depth and drops to zero once the sinks take over. InternVL3.5-8B keeps localizing while its sinks grow. On the raw map Pixtral-12B never localizes, and InternVL3-8B does from mid depth on. Gemma-4 localizes in its full-attention layers, the only Gemma-4 layers the figure plots. At its sliding-window peak layer the most-attended token is never in the box (Table 5). At the peak layer localization is .00 in six of the seven sink-bearing models and .14 in InternVL3.5-8B (Table 5). At the best layer five of the seven sink-bearing models reach .53 to .82 and Gemma-4-E2B reaches .24, while Pixtral-12B's register stays on top at every layer. This is why the selector reads mid-depth layers (Appendix B).

A.2 The first quarter of depth

Why does the peak-layer rule of section 3 skip the first quarter of each model's depth? Localization is weak there (the gray band of Figure 8). In the first quarter the most-attended image token lies in the target box on 0 to 9% of V* images in every model. The rate first reaches 20% between layers 9

Table 6: Token features as detectors of visual sinks at each model's peak layer on V*. Cells: precision at k / top-1 / AUC, where k is the number of sinks in the image. The three sink-free models are below the rule. There the few tokens above $100\times$ uniform (0.4 to 2.7 per image) stand in for the sinks. The first two columns are the centered cosine between the token's key and the text-sink key (Equation 2) and the same cosine without centering. The next five are the negated value norm, the hidden-state norm, the key norm with both signs and the negated standard deviation of the token's attention logits. Apart from the key norm, every feature is signed so that a larger value is more sink-like. Hidden norm is the feature of Kang et al. (2025b).

Model	centered key cos	raw key cos	$-\|v\|$	hidden norm	$\|k\|$	$-\|k\|$	$-$logit std
Qwen3-VL-8B	.99/.98/1.00	.99/.96/1.00	.99/.94/1.00	.99/.93/1.00	.00/.00/.01	.23/.00/.99	.00/.00/.69
Qwen3-VL-4B	.99/.95/1.00	.99/.93/1.00	.99/.68/1.00	.98/.80/1.00	.00/.00/.28	.01/.00/.72	.00/.00/.71
InternVL3.5-8B	.95/.70/1.00	.95/.66/1.00	.95/.73/1.00	.95/.61/.99	.00/.00/.01	.74/.28/.99	.00/.00/.90
Gemma-4-12B	.61/.00/1.00	.61/.00/1.00	.00/.00/.40	.00/.00/.93	.01/.00/.33	.18/.21/.68	.02/.00/.85
Gemma-4-E4B	.18/.00/1.00	.04/.00/.99	.00/.00/.56	.00/.00/.44	.00/.00/.50	.00/.00/.49	.00/.00/.65
Gemma-4-E2B	.07/.00/.99	.00/.00/.94	.00/.00/.08	.00/.00/.25	.06/.00/.46	.15/.12/.55	.00/.00/.69
Pixtral-12B	1.00/1.00/1.00	1.00/1.00/1.00	1.00/1.00/1.00	.00/.00/1.00	.00/.00/.00	1.00/1.00/1.00	1.00/1.00/1.00
sink-free by the 15% cut (section 3); the tokens above $100\times$ uniform stand in for the sinks							
Qwen3-VL-2B	.51/.30/.99	.78/.73/1.00	.00/.00/.12	.00/.00/.21	.00/.00/.05	.00/.00/.95	.00/.00/.10
InternVL3.5-2B	.00/.00/.42	.01/.00/.91	.00/.00/.75	.00/.00/.06	.00/.00/.87	.00/.00/.13	.00/.00/.48
InternVL3-8B	.13/.09/1.00	.50/.35/1.00	.53/.45/1.00	.00/.00/.30	.00/.00/.24	.00/.00/.76	.00/.00/.30

and 17, and never in Pixtral-12B. Tokens above $100\times$ uniform do appear in the first quarter of every model, with or without later sinks.

Without the rule the peak layer would fall in the first quarter for four models. On V* Qwen3-VL-2B and InternVL3.5-2B would peak at layer 1, InternVL3-8B at layer 3 and Gemma-4-12B at layer 8. The same four models move on both HR-Bench sets. Gemma-4-12B is the one sink-bearing model among them. Its sliding-window layers carry sinks from layer 4 on, with a share comparable to the late layers. The gate of § 3.4 acts per layer, and the selector reads mid-depth layers (§ 4.1). The method is therefore unaffected either way.

The centered cosine of the sliding-window sinks stays below $.4$ in the first quarter of Gemma-4-12B, Gemma-4-E4B and Gemma-4-E2B. From mid depth on it settles at $.5$ to $.8$, the alignment reported in § 3.3. Gemma-4's early sinks are therefore not yet aligned with the text-sink key. We have not tested whether the direction is carried up the residual stream from the full-attention layers. On Gemma-4-12B, the one size whose late-layer keys are its own (§ A.3), the gate does not flag these early sinks. The selector never reads sliding-window layers on Gemma-4, so it never sees them.

A.3 Detectors

Table 6 scores six token features as detectors of the attention-defined sinks at each model's peak layer. Each cell holds three numbers. Here k is the number of sinks in the image. Precision at k is the fraction of the feature's top-k tokens that are sinks. The top-1 entry asks whether the feature's top token is the most-attended one. AUC measures the separation of sinks from all other tokens, and it can round to 1.00 while a sink still ranks only in the top 1% of three thousand tokens. We call a feature a detector when its AUC is at least .99 in every sink-bearing model. Precision at k then asks whether it also names the individual sinks.

The centered key cosine has an AUC of at least .99 in every sink-bearing model, and in the full-attention families it also names the individual sinks. It is the one feature that meets the criterion. In Gemma-4 the AUC stays high but precision is low. Other tokens there sit even closer to the weak text-sink key and outrank the sinks. The raw key cosine matches the centered one in the full-attention families and on Gemma-4-12B. On Gemma-4-E4B and Gemma-4-E2B its precision falls to $.04$ and $.00$, and on Gemma-4-E2B the AUC falls to $.94$.

A low value norm finds the sinks in the Qwen3 backbones and Pixtral-12B. It does not find them in Gemma-4, whose sinks have values of ordinary size. The hidden-state norm is the massive-activation signature of Kang et al. (2025b) and Sun et al. (2024). By the AUC criterion it passes in the Qwen3 backbones and Pixtral-12B and fails in Gemma-4. In the Qwen3 backbones it also names the individual sinks. In Pixtral-12B it does not, because the register is a massive activation

but never the largest in the image and some border tokens outrank it (precision .00), just as other tokens outrank the Gemma-4 sinks under the centered cosine. In hidden norm the Gemma-4 sinks stay within a factor of two of the median at every depth. A massive activation is therefore one route to the sink direction. It is not the only one. The key norm itself has AUC near 0 in Qwen3-VL-8B, InternVL3.5-8B and Pixtral-12B, so the sinks have the smallest keys there and the negated key norm separates them (AUC .99 to 1.00). It names them in Pixtral-12B and partly in InternVL3.5-8B (precision .74) but hardly in Qwen3-VL-8B (.23). It fails in Qwen3-VL-4B (AUC .72) and in Gemma-4. The logit spread separates the sinks only in Pixtral-12B.

The rows of the three sink-free models describe 0.4 to 2.7 tokens per image above $100\times$ uniform. The centered cosine separates them in Qwen3-VL-2B and InternVL3-8B (AUC .99 and 1.00) but not in InternVL3.5-2B (.42), whose single layer-7 token lacks the key signature (subsection A.1).

Gemma-4-E2B and Gemma-4-E4B share keys and values across their last 20 and 18 layers (from layers 15 and 24), so their peak-layer keys are those of an earlier layer. Only Gemma-4-12B has its own keys at its peak layer. The key and value features in the Gemma-4-E4B and Gemma-4-E2B rows of Table 6 are therefore those of layers 24 and 15, scored against the attention of layers 36 and 28.

A.4 KEY INTERVENTIONS

Table 7 gives the numbers behind the causal claim of **??**. We run two interventions on Qwen3-VL-8B, Qwen3-VL-4B, InternVL3.5-8B and Pixtral-12B at each model's peak layer on 20 V* images. *Remove* projects $\hat{u}$ out of the key of every image token whose cosine to $\hat{u}$ exceeds .5. This cosine is uncentered, unlike the gate of subsection 3.4, which is why the counts below differ from those of subsection A.6. The table lists those tokens' in-image share before and after. The control projects out a random unit direction instead. On these 20 images the gate admits five tokens per image on Qwen3-VL-8B, seven on Qwen3-VL-4B and one on Pixtral-12B. Fifteen of the 20 images have such a token on InternVL3.5-8B. *Align* takes $k = 8$ random image tokens and sets their keys to the text-sink key. The Align columns list their share before and after. The control sets their keys to a random key of the same norm.

Table 7: Key interventions at the peak layer on 20 V* images. *Remove*: in-image share of the gated tokens before and after projecting out $\hat{u}$, with a random unit direction as the control. *Align*: share of k=8 random image tokens before and after setting their keys to $k_{t^\star}$, with a random key of the same norm as the control. *Median target*: the $\hat{u}$-component of the edited keys, in the Remove edit and in the Align random-key control, is moved to the population median instead of zero.

		Remove			Align		
Model	Layer	before	$-\hat{u}$	−random	before	$:= k_{t^\star}$	random key
Qwen3-VL-8B	30	.68	.00	.66	.00	.97	.04
Qwen3-VL-4B	29	.75	.00	.74	.00	.95	.00
InternVL3.5-8B	30	.36	.00	.35	.00	.96	.09
Pixtral-12B	27	.49	.25	.49	.00	.91	.64
median target	27	.49	.00	.49	.00	.91	.01

On the three Qwen3 backbones projecting out $\hat{u}$ removes the sinks' share entirely and the random direction leaves it alone (Table 7). Random keys aligned with the text-sink key capture nearly all the in-image attention. The full key edit of subsection 3.4 behaves the same way at every layer. The removed mass returns to the text-sink (text-sink column of Table 2).

Pixtral-12B's image keys point away from the text-sink (centered cosine $-.59$ to $\hat{u}$ for the population). A key with zero component along $\hat{u}$ therefore sits far above the population. With the projection to zero, the register's share only halves (.49 to .25). The same projection applied to random non-sink tokens on the same images lifts their share from .00 to .56. It does nothing on Qwen3-VL-8B, whose image keys sit near zero along $\hat{u}$. Moving the register's component to the population median instead removes it (.49 to .00), and the alignment control separates cleanly (.91 versus .01). The key edit of **??** therefore takes the median as its target.

A.5 GEMMA-4: WHAT THE SLIDING WINDOW CANNOT REACH

Are Gemma-4's image sinks a substitute for a text-sink that the sliding window cannot reach? The released budget is 1,120 soft tokens, of which about 1,080 are image tokens on V*. At that budget the text before the image lies more than one window back, so the question tokens cannot attend to it in the sliding-window layers. Table 8 repeats the sink measurements with the budget cut to 280 soft tokens (261 image tokens). This puts the whole prompt inside the window. The run uses the same 100 V* images and the same prompt as the released-budget run.

At the released budget the image takes .51 to .60 of the question's attention in these layers. With the text in reach it takes .14 to .19, and .37 to .49 goes to the text before the image. The strongest image token's absolute attention drops three- to five-fold. Its share within the image changes little in two of the three models (.37 to .35 in Gemma-4-E2B and .17 to .19 in Gemma-4-12B). In Gemma-4-E4B it falls from .30 to .18. The attention spreads over the text before the image. The text-sink position itself takes little of it. The image tokens of these layers therefore hold attention that the text would absorb if it were reachable.

The count of tokens above $100\times$ uniform also falls. That threshold is not comparable across budgets, because $100\times$ uniform is a much larger share of the image at 261 tokens than at 1,080. The smaller budget coarsens the image as well, so two things change at once.

Table 8: Gemma-4 sliding-window layers at 1,120 versus 280 soft tokens on 100 V* images, averaged over heads, layers and images. Each cell reads released budget → reduced budget. The *text before image* and *image* columns are shares of the question's attention. The strongest image token's share of the question's attention is *top-1, absolute*, and its share of the in-image attention is *top-1, share of image*.

Model	text before image	image	top-1, absolute	top-1, share of image
Gemma-4-E2B	.00 → .38	.55 → .19	.21 → .07	.37 → .35
Gemma-4-E4B	.00 → .37	.51 → .16	.16 → .03	.30 → .18
Gemma-4-12B	.00 → .49	.60 → .14	.10 → .03	.17 → .19

A.6 HEADS AND THE GATE THRESHOLD

Does a head that sinks also localize? Figure 6a plots each head's sink share against its attention on the target at the peak layer. In every sink-bearing model the two are negatively correlated. This is not mass conservation. The target box covers 0.3% of a V* image. A head that gives .8 of its image attention to the sinks could still put its remaining .2 on the target at $50\times$ uniform. Such heads sit at $0.1\times$ instead, below uniform. Heads that sink under half their image attention reach 10 to $30\times$. Different heads sink and localize, as different layers do.

How much does the edit depend on τ_c? Figure 6b plots the precision and recall of the gate against the attention-defined sinks (image tokens outside the target box above $100\times$ uniform, section 3) as τ_c varies. In the full-attention families recall is 1.0 at every threshold from 0.4 to 0.9, so the gate never misses a sink. Precision is flat over the same range, at .6 on Qwen3-VL and .8 on InternVL3.5. Any threshold in the band gives the same edit. We use 0.5. Below 0.4 the gate admits ordinary tokens and precision collapses. On Qwen3-VL-8B it admits 17 per image at 0.4 and 84 at 0.3.

Which tokens does the gate admit inside the band? At 0.5 the gate of Figure 6b admits about seven tokens per image on Qwen3-VL-8B (five on the 20-image subset of Table 7). Four of them are sinks by the $100\times$ definition. The other three sit at 10 to $100\times$ uniform with keys aligned just as tightly. They are sinks below an arbitrary attention cut, and we want them removed too. Precision below 1 inside the band is therefore not a false-positive problem.

Pixtral-12B and Gemma-4 fall outside this picture. In Pixtral-12B the band admits about ten tokens around its single register. Only $\tau_c > 0.8$ isolates the register. The extra tokens do no harm, since the cleaned map still beats the raw one on Pixtral-12B (Table 4). Gemma-4 stays below .5 precision at every threshold and loses recall beyond 0.7, which is a second reason, besides the output shift in Table 2, that the re-look selector skips the edit on Gemma-4 (§ 4.1).

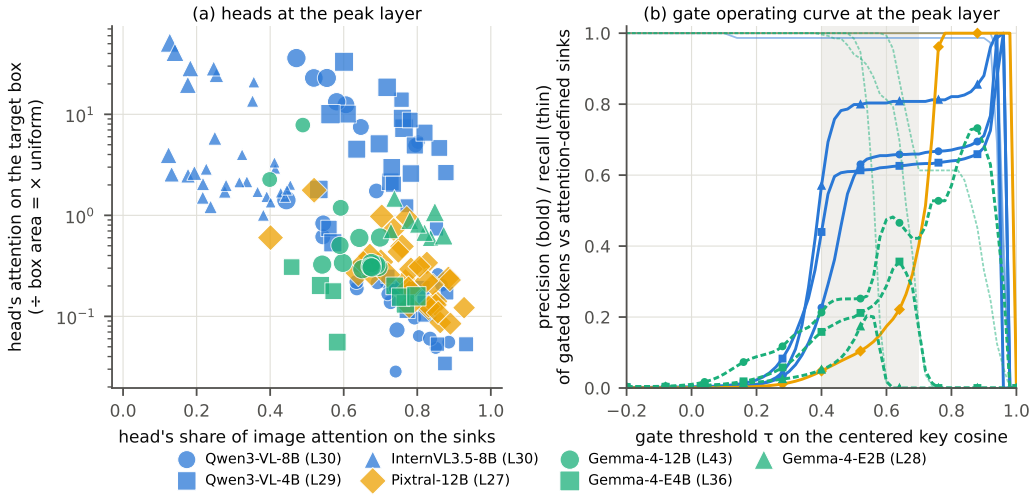


Figure 6: (a) Sink share vs. target attention per head at the peak layer on V*. Marker size is the head's share of image attention. (b) Precision and recall of the centered-cosine gate at threshold τ_c against the attention-defined sinks, one curve per sink-bearing model on V*. Shaded band: thresholds that give the same edit. The paper uses 0.5.

The cap $\rho = 0.02$ is set by hand and not swept. It matters only below the band. At $\tau_c = 0.3$ the 84 admitted tokens are 2.5% of a 3,300-token image. That is above the cap, so the layer is left alone.

A.7 What the sinks contribute to the output

Does the attention a sink receives reach the output? A large attention weight is not a large contribution. The output for query i is $\sum_j \alpha_{ij} v_j$, so a token with a near-zero value adds little however much attention it gets. A value aligned with the text-sink's adds in a direction the text-sink already supplies. In the full-attention families a sink's value has both properties. Its norm is under a tenth of the image median and close to the text-sink's own value norm. The cosine between its direction and the text-sink's value direction is .7 to .8, against a population median of $-.1$. At each model's peak layer and in each head we measure the norm of the visual attention sinks' contribution relative to the full output. We also record the cosine between the output with and without that contribution (Table 9).

In the full-attention families the sinks take .29 to .77 of the in-image attention but supply .03 to .16 of the output norm. Removing their contribution barely rotates the output (cosine at least .98). HR-Bench gives the same picture (§ A.12). On the three sink-free models the tokens above $100\times$ uniform take .02 to .16 of the in-image attention and .01 to .03 of the output norm. A FastV-style pruner (Chen et al., 2024) ranks image tokens by attention at layer 20 and keeps the top half or quarter from layer 21 on. Ranked on the raw map it keeps the sinks, and ranked on the cleaned map at one quarter it drops them. Qwen3-VL-8B's V* accuracy is .73 in every setting with perfect answer agreement, so dropping the sinks outright changes nothing either.

Gemma-4's sliding-window layers cannot reach the text-sink, and its visual attention sinks have values of ordinary size. Their readout is 1.00 to 1.20, so they supply the dominant share of the output. Removing them rotates the output (cosine .02 to .67). The same attention pattern is inert in the full-attention families and load-bearing in Gemma-4.

A.8 Masking instead of editing

If the sinks are just the most-attended image tokens, can a reader of the map drop them without touching the model? Table 10 tests this on the raw maps. At each upper-half layer we zero every image token above $100\times$ uniform (or above $10\times$) without knowledge of the target box. We then recompute top-1 localization on 100 V* images. In the Qwen3 backbones the mask recovers less

Table 9: Visual-sink contributions to the attention output at the peak layer on V*. The *attention* column is the sinks' share of the head's in-image attention at the peak layer, averaged over heads. It therefore differs slightly from the share in Table 5, which is read from the head-averaged map. The *readout* columns divide the contribution norm by the full output norm, for the visual sinks and for the text sink $t^\star$. A readout above 1 means the other contributions partly cancel the sinks'. The *cos* column is the cosine between the output with and without the sink contribution. The three sink-free models are below the rule, with their tokens above $100\times$ uniform in place of the sinks.

Model	attention	readout: sinks	readout: $t^\star$	cos
Qwen3-VL-8B	.70	.11	.66	.994
Qwen3-VL-4B	.75	.13	.52	.991
InternVL3.5-8B	.29	.03	.38	.999
Pixtral-12B	.77	.16	.49	.983
Gemma-4-12B	.62	1.20	.00	.67
Gemma-4-E4B	.65	1.00	.00	.41
Gemma-4-E2B	.80	1.07	.00	.02
Qwen3-VL-2B	.10	.03	.55	.999
InternVL3.5-2B	.16	.03	.43	.999
InternVL3-8B	.02	.01	.41	1.000

than half of the key edit's gain. The sink-free models lose localization under the mask. There 25% to 86% of the tokens above the threshold lie inside the target box (in-box column), so masking them masks the target. The key edit moves the sink-free models by at most .05. The $10\times$ threshold is worse on every model but InternVL3.5-8B, since target tokens sit at 10 to $100\times$ uniform. On Gemma-4 every column is near zero because most upper-half layers use a sliding window and never localize. Over its full-attention layers alone the raw map localizes on .08 to .17 of images. The $100\times$ mask changes this by at most .05, the $10\times$ mask lowers it to .04 to .08, and the key edit raises it to .17 to .31. The re-look does not edit Gemma-4's keys (§ 4.1).

Table 10: Share of V* images whose most-attended image token lies in the target box. The columns give the raw map, the map after masking (every token above $100\times$ or above $10\times$ uniform zeroed) and the map after the key edit. *In box* is the share of the $> 100\times$ tokens inside the target. Numbers average the upper-half layers on 100 images, so the key-edit column is not the value of Table 2. The Gemma-4 rows average all upper-half layers, sliding-window layers included.

Model	raw	masked $> 100\times$	masked $> 10\times$	key edit	in box
Qwen3-VL-8B	.00	.20	.10	.45	5%
Qwen3-VL-4B	.00	.17	.10	.39	3%
InternVL3.5-8B	.00	.18	.28	.66	2%
Pixtral-12B	.00	.09	.03	.14	0%
Qwen3-VL-2B	.22	.18	.05	.17	25%
InternVL3.5-2B	.57	.47	.05	.56	86%
InternVL3-8B	.62	.50	.05	.61	85%
Gemma-4-12B	.03	.04	.10	.03	1%
Gemma-4-E4B	.03	.03	.02	.04	0%
Gemma-4-E2B	.02	.02	.02	.07	0%

A.9 Redistribution instead of removal

Does forcing the sinks' attention onto the image help the answer? The redistribution of Kang et al. (2025b) keeps the sinks and moves their attention. At every language-model layer during prefill the attention on gated tokens is spread over the remaining image tokens in proportion to their own attention. We run two variants, one in all heads and one only in heads that put more than half their attention on the image. On Qwen3-VL-8B this raises absolute target attention by up to $2.6\times$ and restores localization at the layers that hold the sinks. InternVL3.5-8B's target attention doubles or

triples. Accuracy stays flat on both models for both variants. Agreement (.89 to .97) and KL (.01 to .09) show that redistribution disturbs the output more than the key edit.

A.10 The edit along a chain of thought

Does the first-token KL of Table 2 stay small along a full trace? We generate greedy chains of thought of up to 192 tokens from InternVL3.5-8B and Pixtral-12B on 60 V* questions with the edit applied at every step. Each trace is scored against the base model's trace. The control is the unedited model rerun with SDPA in place of the eager attention kernel. Per position the teacher-forced KL stays at a few 10^{-3} nats along the whole trace (Figure 7). This is an order of magnitude above the kernel control and shows no upward trend.

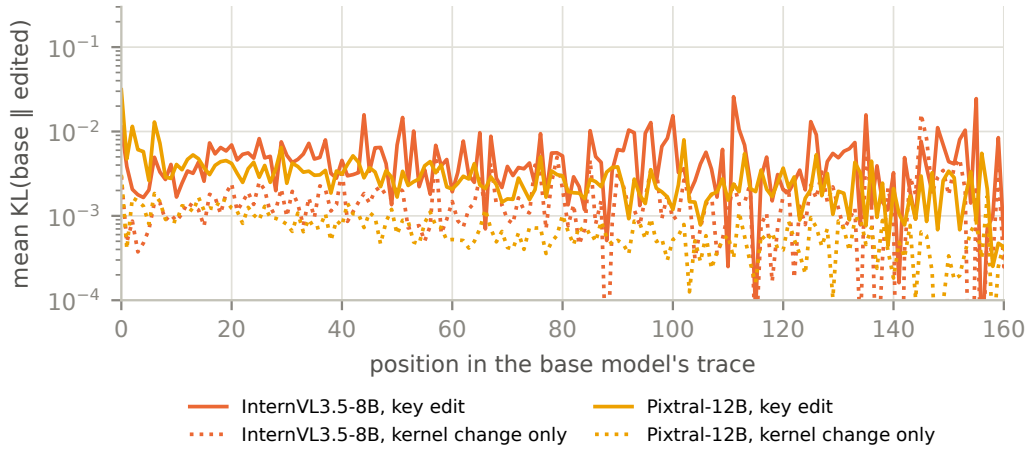


Figure 7: Per-position KL along the base model's chain-of-thought trace with the key edit applied at every step (solid) against a kernel-only control (dotted). InternVL3.5-8B and Pixtral-12B on 60 V* questions, greedy chains of up to 192 tokens.

A.11 The 2B models

The key edit costs Qwen3-VL-2B 7 points on V* (.68 to .61, Table 2) and 13 to 14 points on HR-Bench (Table 12). InternVL3.5-2B loses 3 points on V*. Why does the edit hurt them? The sparsity cap and median target of § 3.4 guard against layers where the whole image-key population carries a small component along $\hat{u}$. Shifting that component can turn repulsion into attraction and create a new sink. Qwen3-VL-2B and InternVL3.5-2B sit at the edge of that guard. In InternVL3.5-2B the gate admits weakly aligned non-sink tokens. Qwen3-VL-2B is hurt more. At layer 2 its gate admits tokens with a small $\hat{u}$ component, and moving them to the median creates a new concentration at nearly 1,000× uniform. Restricting the edit on Qwen3-VL-2B to the upper three quarters of depth restores accuracy to .65. A second unedited run scores .66, so this is within run-to-run variation. Agreement is .98 and KL is .008. For small models the edit should be depth-restricted.

In the re-look the key edit is applied only in the prefill that selects the crop, and the generation passes run on the unedited model. The accuracy loss above therefore does not enter Table 4. On Qwen3-VL-2B the edit can only move the window.

A.12 The same measurements on HR-Bench

Do the V* measurements hold on HR-Bench? We repeat the measurements of section 3 on the first 50 distinct images of HR-4K and of HR-8K with the same budgets and prompts. The 800 questions of each set share 151 images. HR-Bench has no target boxes, so the box columns are absent and a sink is defined by the 100× threshold alone. Table 11 gives the sink measurements and Table 12 the key edit. Detector AUC and value readout match V* to within the rounding of Table 6 and Table 9. We do not reproduce those two tables for HR-Bench. On the four sink-bearing full-attention models

the key edit acts as it does on V*. Gemma-4-12B and Qwen3-VL-2B are the exceptions, and the removal paragraph below says how they differ.

Table 11: Text and visual sinks on the first 50 distinct images of HR-4K and of HR-8K. Columns as in Table 5 without the LM, position, layer and target columns. The visual-sink columns are read at the peak layer. n is the number of image tokens per prompt.

	HR-4K					HR-8K				
Model	n	$t^\star$ share	×unif.	per image	share	n	$t^\star$ share	×unif.	per image	share
Qwen3-VL-8B	12194	.72	4626	6.1	.69	12416	.71	4659	6.1	.71
Qwen3-VL-4B	12194	.67	3468	7.6	.76	12416	.67	3675	8.1	.78
Qwen3-VL-2B	12194	.74	1096	5.5	.17	12416	.75	1366	6.4	.20
InternVL3.5-8B	2555	.64	652	2.5	.41	1987	.65	314	1.8	.28
InternVL3.5-2B	2555	.76	143	1.6	.10	1987	.77	110	0.9	.07
Gemma-4-12B	1088	.10	192	4.4	.62	1090	.10	192	4.5	.62
Gemma-4-E4B	1088	.21	417	2.7	.66	1090	.21	397	2.7	.67
Gemma-4-E2B	1088	.22	525	2.7	.83	1090	.22	481	2.8	.82
InternVL3-8B	2555	.60	126	1.1	.07	1987	.61	89	0.6	.04
Pixtral-12B	3919	.48	3063	1.7	.79	2742	.49	2093	1.0	.78

The sink-bearing split of section 3 is made on V*. On HR-Bench Qwen3-VL-2B's peak share rises to .17 and .20 and crosses the 15% cut. This goes with the larger accuracy loss of the all-layer edit reported below.

Detectors. At the peak layer the centered key cosine keeps near-perfect AUC on every sink-bearing model on both sets. The family-level pattern of precision and of hidden-norm AUC matches Table 6.

Value side. The sinks' share of the output matches the V* values of Table 9 to within .01 on every model and dataset. This includes the large Gemma-4 contribution.

Table 12: The key edit on the first 50 distinct images of HR-4K and of HR-8K. Columns as in Table 2 without the box and caption columns. *Gate off* means the gate admits under one token per image, so the edit is an identity.

	HR-4K					HR-8K				
Model	$\max_j r_j$	text sink	accuracy	agreement	KL	$\max_j r_j$	text sink	accuracy	agreement	KL
Qwen3-VL-8B	.25 → .02	.54 → .64	.72 → .72	.99	.001	.24 → .02	.54 → .65	.68 → .67	.99	.001
Qwen3-VL-4B	.21 → .03	.51 → .65	.74 → .74	.97	.000	.22 → .03	.51 → .65	.69 → .69	.99	.000
Qwen3-VL-2B	.02 → .02	.60 → .59	.71 → .57	.74	.017	.02 → .02	.61 → .59	.72 → .59	.78	.024
InternVL3.5-8B	.31 → .04	.42 → .63	.73 → .72	.98	.000	.33 → .04	.44 → .64	.69 → .70	.99	.000
InternVL3.5-2B	.04 → .04	.62 → .61	.76 → .76	.87	.002	.04 → .04	.63 → .63	.62 → .63	.90	.008
InternVL3-8B	gate off	.46 → .46	.78 → .77	.99	.000	gate off	.47 → .47	.73 → .74	.98	.001
Pixtral-12B	.56 → .02	.38 → .49	.61 → .60	.95	.003	.59 → .02	.39 → .50	.53 → .52	.96	.002
Gemma-4-12B	.15 → .12	–	.67 → .68	.89	.002	.15 → .14	–	.54 → .57	.77	.085
Gemma-4-E4B	.30 → .27	–	.66 → .62	.94	.001	.30 → .27	–	.58 → .56	.95	.001
Gemma-4-E2B	.40 → .38	–	.60 → .59	.93	.000	.38 → .36	–	.54 → .51	.93	.001

Removal. On Qwen3-VL-8B, Qwen3-VL-4B, InternVL3.5-8B and Pixtral-12B the edit clears the map and leaves the answers alone (Table 12). The freed mass goes to the text-sink, whose share rises by .10 to .21 on those four models. On InternVL3-8B the gate admits under one token per image. The map does not move on Gemma-4. Its output is close to unchanged on HR-4K (KL at most .002, though Gemma-4-E4B loses four points of accuracy) and shifts on HR-8K for Gemma-4-12B (KL .085, agreement .77). That value is far below the V* KL of .531 for the same model. We have not traced this gap. The re-look does not edit Gemma-4's keys (§ 4.1), so it does not enter Table 4. On Qwen3-VL-2B the degradation grows with input length. The longer sequences admit many more tokens through the gate, and the all-layer edit costs over ten points while the map does not change. The depth-restricted edit of § A.11 is the form to use on this model.

B SELECTOR LAYERS

Which layers carry a question-dependent map? For Qwen3-VL-8B we sweep all 36 layers on 150 images. The images come 30 each from Geo3K, RealWorldQA, V*, VMCBench and MMVP. At each layer we compute the question-to-image attention map under the real question and again under an unrelated prompt. The statistic is the overlap of the eight most-attended image tokens between the two maps. A low overlap means the map depends on the question. It is lowest at layers 18 to 21 (we do not plot the sweep). We take the two ends of this band, layers 18 and 21. The choice predates the main experiments and is not tuned per benchmark. Thirty of the 150 sweep images are V* images, so the layer choice has seen part of the V* set. The sweep statistic uses neither the target boxes nor the answers.

The other full-attention models use the same relative depths of about .5 and .58 and are not searched separately. Table 13 lists the resulting layers for every model.

Table 13: Selector layers per model. The selector averages the head-averaged question-to-image attention of these two layers (§ 4.1). The number after the slash is the count of language-model layers. The other full-attention models read the cleaned map at the same relative depths as Qwen3-VL-8B. The Gemma-4 rows list full-attention layers, which the selector reads on the raw map.

Model	layers / total
Qwen3-VL-8B	18, 21 / 36
Qwen3-VL-4B	18, 21 / 36
InternVL3.5-8B	18, 21 / 36
Qwen3-VL-2B	14, 16 / 28
InternVL3.5-2B	14, 16 / 28
InternVL3-8B	14, 16 / 28
Pixtral-12B	20, 23 / 40
Gemma-4: full-attention layers, raw map	
Gemma-4-E2B	19, 24
Gemma-4-E4B	17, 23
Gemma-4-12B	23, 29

On Gemma-4 the selector reads only full-attention layers (§ 4.1). These lie five or six layers apart, so relative depth does not carry over. We take the adjacent pair of full-attention layers whose crop hits the target most often on 60 V* images. These 60 images are part of the V* evaluation set, so Gemma-4's V* column in Table 4 is not fully held out. Its HR-Bench and MME-RealWorld-Lite columns are.

Does the cleaned map localize at the chosen layers? Figure 8 plots top-1 localization per layer on V* for the raw and the cleaned map, with the selector layers marked. Every full-attention model but Pixtral-12B has a selector layer at which the cleaned map puts its most-attended token in the target box on .68 to .84 of V* images. This is within .05 of the model's best layer past the first quarter of depth. Pixtral-12B reaches .37 at layer 20 and .45 at its best layer, layer 19. These are per-layer values read from Figure 8. Table 2 averages over the probed layers, which is why its box column is lower.

After the key edit the most-attended token still lands on the outermost row or column of the grid several times more often than chance in the late layers of every model. The effect is strongest in Pixtral-12B, where the tokens next to `[IMG_BREAK]` keep an alignment just under the gate threshold at every layer past mid depth. At the Qwen3-VL selector layers the border rate is at chance. No run in this paper corrects for this effect.

C EXPERIMENTAL DETAILS

Models and prompts. Every model receives the image followed by the question and its lettered options, verbatim from the benchmark. One answer-format sentence closes the prompt and depends on how the model reasons. The native thinking models are asked only for the letter. They reason in their own thinking channel. Qwen3-VL-Thinking uses its default chat template with no system

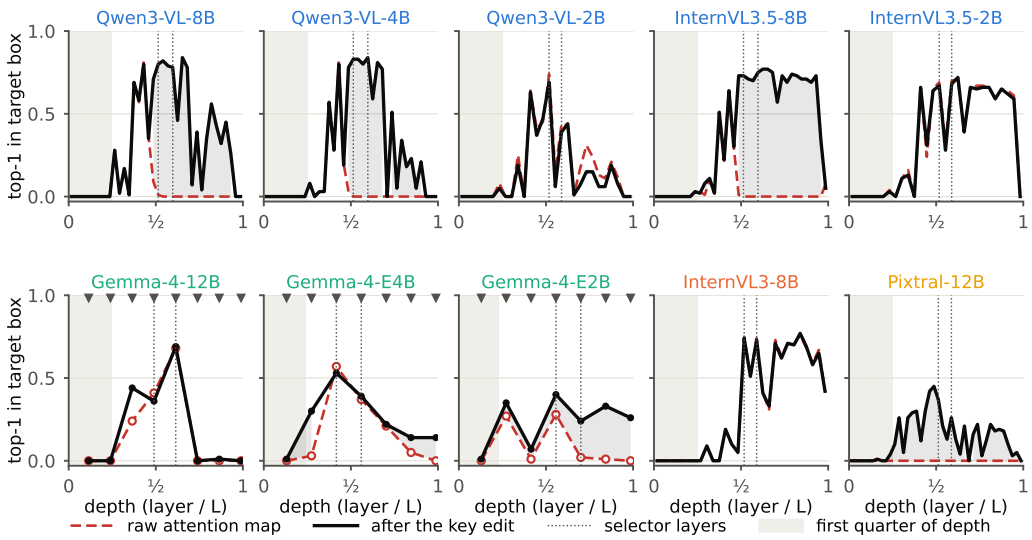


Figure 8: Per-layer top-1 localization on V* for the raw map (dashed) and the cleaned map (solid). Here top-1 localization is the fraction of images whose most-attended image token lies in the target box. The shaded area between the curves is the gain from the key edit. Dotted lines mark the selector layers. The gray band is the first quarter of depth, which the peak-layer rule skips. Gemma-4 shows full-attention layers only (triangles).

Thinking models: `{image} {question}\nPut only the letter of the correct option inside \boxed{}.`

Chain-of-thought prompted models: `{image} {question}\nThink step by step, then put only the letter of the correct option inside \boxed{}.`

Transition after the spliced crop: `{crop} Looking at this region again,`

Figure 9: Prompts used in every run. The native thinking models receive the first line. Qwen3-VL-Thinking (8B, 4B and 2B) uses its default chat template with no system prompt. For InternVL3.5 (8B and 2B) the R1 system prompt of its model card switches on the thinking channel. Gemma-4 (12B, E4B and E2B) runs its default chat template with `enable_thinking=True`. InternVL3-8B and Pixtral-12B have no thinking channel and receive the second line. The third line follows the spliced crop.

prompt. For InternVL3.5 the R1 system prompt of its model card switches on the thinking channel. Gemma-4 runs through its template with thinking enabled. InternVL3-8B and Pixtral-12B have no thinking channel and are asked to think step by step in the answer. MME-RealWorld-Lite has five options (A to E) and the same wording. Figure 9 gives the exact strings.

Image budgets. Qwen3-VL-8B, Qwen3-VL-4B and Qwen3-VL-2B see at most 16,384 patches of 28^2 pixels (12.8 Mpix). InternVL3.5-8B, InternVL3.5-2B and InternVL3-8B keep their default dynamic tiling of up to 12 tiles of 448^2 pixels. Pixtral-12B uses its processor default of a 1,024-pixel longest edge. Gemma-4 runs at its maximum of 1,120 soft tokens. HR-4K and HR-8K images have medians of 16 and 39 Mpix, against 3.4 Mpix on V*.

Sampling and trigger. Sampling uses temperature 0.6 with top-p 0.95, top-k 30 and min-p 0.001. Generation is capped at 4,096 new tokens. The seed is the question index. The re-look is triggered at entropy $H_0 = 1$ nat after at least $c = 32$ tokens of thought (§ 4.2). The selector reads two mid-depth layers per model (Appendix B) with the key edit at $\tau_c = 0.5$ and $\rho = 0.02$ (§ 3.4). The edit is applied only in the prefill that selects the crop. On Gemma-4 the selector reads two full-attention layers without the edit. The window is a third of the grid on each side, padded by 20 pixels.

Trigger coverage. Table 14 gives the per-model trigger rates. For six of the ten models the trigger fires on more than 90% of questions on every benchmark. On those models the method is close to an unconditional re-look. We do not report an always-trigger or random-step control.

Table 14: Fraction of questions (%) on which the re-look is triggered, per model and benchmark. V* has 191 questions, HR-4K and HR-8K 800 each, MME-RealWorld-Lite 1,919. Same runs as Table 4.

Model	V*	HR-4K	HR-8K	MME-RW-Lite
Qwen3-VL-Thinking				
8B	59.7	60.1	73.1	90.9
4B	65.4	65.4	74.9	94.2
2B	72.3	71.8	78.2	92.5
InternVL3.5 (thinking)				
8B	79.6	79.9	83.9	91.0
2B	100.0	100.0	100.0	100.0
Gemma-4 (thinking)				
12B	99.5	98.2	99.1	99.6
E4B	99.5	99.1	99.8	99.3
E2B	97.9	98.4	98.6	99.0
Chain-of-thought prompted				
InternVL3-8B	96.3	97.9	97.5	98.3
Pixtral-12B	95.3	91.4	93.0	96.5

Paired comparisons. Table 15 gives the per-question counts behind the gains of Table 4. Each cell counts the discordant questions between the re-look on the cleaned map and a reference run. Against the base run the re-look wins 1,997 discordant questions and loses 1,181 over the 30 (model, benchmark) pairs on V* and HR-Bench, a ratio of 1.7. Against the raw map the cleaned map is significantly better in seven of the 28 pairs that have a raw run and never significantly worse. Five of the seven are on V* and HR-Bench. The last two columns show where the crops land on V*. A crop counts as a hit when it contains the center of a target box. On Qwen3-VL-8B the cleaned map hits on .90 of triggered questions and the raw map on .22. On the three sink-free models the raw map hits as often as the cleaned map or slightly more often (.82 against .76 on Qwen3-VL-2B).

Table 15: Paired comparisons behind Table 4. For every model and benchmark, wins/losses counts the questions the re-look with the cleaned map answers correctly and the reference answers wrongly, against the reverse. The reference is the base run (*base*) or the same re-look with the crop selected on the raw map (*raw*). † marks $p < .05$ under the exact McNemar test. The last two columns give the V* crop-hit rate on triggered questions for the cleaned and for the raw map. A crop hits when it contains the center of a target box. Gemma-4's selector reads the raw map, so its one hit rate is listed under *cleaned* and it has no raw columns. V* has 191 questions, HR-4K and HR-8K 800 each, MME-RealWorld-Lite 1,919. Same runs as Table 4.

	V*		**HR-4K**		**HR-8K**		**MME-RW-Lite**		**V* crop hit**	
Model	vs. base	vs. raw	vs. base	vs. raw	vs. base	vs. raw	vs. base	vs. raw	cleaned	raw
Qwen3-VL-Thinking										
8B	9/2	7/4	51/18†	40/13†	65/30†	41/22†	114/75†	83/64	.90	.22
4B	11/6	13/4†	30/23	32/21	62/21†	59/30†	144/79†	118/68†	.89	.08
2B	14/11	4/7	58/34†	12/12	71/46†	16/24	159/120†	38/40	.76	.82
InternVL3.5 (thinking)										
8B	23/6†	5/2	65/16†	17/18	86/30†	10/16	172/113†	95/77	.88	.80
2B	34/14†	4/3	108/74†	8/11	142/66†	12/9	189/181	57/51	.81	.81
Gemma-4 (thinking)										
12B	22/9†	–	88/51†	–	122/43†	–	200/147†	–	.84	–
E4B	38/17†	–	96/81	–	108/68†	–	216/146†	–	.75	–
E2B	33/22	–	83/73	–	127/74†	–	182/191	–	.50	–
Chain-of-thought prompted										
InternVL3-8B	21/10	0/0	85/82	1/0	121/78†	2/2	226/208	9/9	.91	.91
Pixtral-12B	30/16	13/7	83/74	60/34†	111/86	58/38	253/182†	152/107†	.38	.13