LOCALIZING AND DISRUPTING ENTITY IDENTIFICATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Shown a photograph of a famous person, brand, or landmark, a vision–language model (VLM) can report facts that no pixel contains: it first *recognizes* the entity and then retrieves what it has memorized about it. The same step lets memory override the image. Which components of a VLM support recognition of familiar entities? We answer this question with two tools: *familiarity tiers*, derived from the model's own behavior, which separate entities it knows robustly, knows weakly, or does not know; and an exhaustive sweep of attention-layer knockouts, scored against perceptual control questions on the same images. Across four open VLMs (Qwen2.5-VL, Qwen3-VL, Gemma-4, and LLaVA-1.5) recognition loss exhibits prominent peaks in mid-decoder attention layers, with selectivity relative to perceptual controls varying across models. In Qwen2.5-VL and LLaVA-1.5, where the band is one or two layers wide, ablating it substantially lowers recognition accuracy while having smaller effects on answers to the perceptual control questions: an *entity-recognition gate*. In Gemma-4 and Qwen3-VL the band spans five layers, and ablating it also changes a substantial share of the control answers. An in-depth analysis of Qwen2.5-VL shows that its gate generalizes across faces, logos, landmarks, and artworks; that ablating it lowers identification of highly-known-tier edited faces from 58% to 15% while reports of the same images' attributes are unchanged and general VQA accuracy drops by 1–3 points; and that ablating five of its attention heads lowers logo recognition from 96% to 3%. Finally, the gate influences behavior when memory and image disagree: ablating it reduces memory-driven identification errors on VLMBIAS by up to 24 points, and on people edited to wear a doctor's coat it shifts the model's answer from their real occupation toward "doctor".

1 INTRODUCTION

Show a vision–language model (VLM) a photograph of Lionel Messi and ask what he does for a living, and it will answer "footballer", a fact that is nowhere in the pixels. To answer, the model must first *recognize* the person or entity and then retrieve what it has memorized about them. Recognition actually entangles perception with memory, to the extent that in some cases prior knowledge wins over perceived details: VLMs report three stripes on an Adidas logo edited to have four (Vo et al., 2026) and show conflicts between visual evidence and prior knowledge (Golovanevsky et al., 2025; Ortu et al., 2026)..

Unlike text-only language models, for which much is known about how memorized knowledge is recalled (Geva et al., 2021; Meng et al., 2022; Haviv et al., 2023; Ferrando et al., 2025), VLMs have seen less progress (Kaduri et al., 2025; Cohen et al., 2025; Zhang et al., 2025), leaving many important questions unanswered. In this work, we investigate the phenomenon of entity recognition in VLMs. We ask what is recognized, where in the network an entity is recognized, and employ the drawn conclusions to improve bias scores through a simple application.

In more detail, we define *familiarity tiers* from the model's own behavior: we find that for some entities the model can name them even when half the image is masked or cropped, for others naming occurs only on the clean image. We define the first group as *highly known*, and second as *known*, and rest as *unknown* (§3.1). A recognition mechanism should act on the first two tiers and have nothing

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.

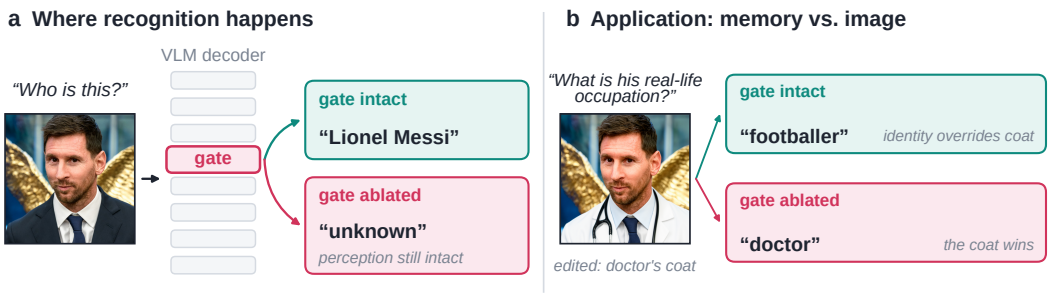

Figure 1: **Entity familiarity is gated by a mid-network band of attention layers.** (a) Asked who is shown, Qwen2.5-VL-7B names Lionel Messi; mean-ablating its gate (layers 16–17) makes it answer "unknown". (b) Asked for the real-life occupation of the same person edited to wear a doctor's coat, the model answers "footballer", letting memorized identity override the visual cue; with the gate ablated, it answers "doctor".

to act on in the third. We then mean-ablate the attention heads in windows of up to five consecutive layers and measure the effect on the various identity related answers (§2).

We apply this protocol in four domains of entities (faces, logos, landmarks, and artworks) and four open VLMs, with one of them, Qwen2.5-VL-7B, analyzed in more depth (since it is widely used, but with an affordable size for sweep analysis). We find that (1) performance on the recognition task of entities in the known tiers is **gated by a narrow band of attention layers** located between 35% and 60% of the decoder's depth in every model; (2) the gate is heavily **carried by a few heads**: removing five of the 56 heads in Qwen2.5-VL-7B's gate band lowers logo recognition from 96% to 3%.(3) as expected, the gate is **tier-graded**: the ablation lowers recognition of highly-known-tier logos by dozens of points, and of unknown-tier logos by none; and (4) **performance on perceptual tasks is nearly unchanged**: on the same images, attribute reports and general VQA accuracy stay within a few points.

To demonstrate the found gate is effective, we simply knock out the gate layers (§5). For portraits edited to show a doctor's coat, where the image supports "doctor" and the memorized occupation does not, removing the gate shifts answers from the memorized occupation to the image-supported one: medical answers rise from 42% to 69% in the highly-known tier. On the VLMBIAS dataset, the knockout improves image-grounded identification by up to 24 points.

2 Problem Formulation

We want to find the components at which an is recognized by a VLM, and to test whether removing them removes recognition without degrading other answers about the same image. This section defines the three ingredients: a way to *observe* recognition in the model's answers, an *intervention* that removes a chosen set of attention layers and nothing else, and *measures* that separate its effect on recognition from its effect on perceptual answers, so that a window that merely breaks the model is not mistaken for one that implements recognition.

A VLM M receives an image I and a prompt q; a vision encoder turns the image into visual tokens, which are interleaved with the text tokens into an input sequence of N tokens and processed by a decoder of L transformer layers. At each layer ℓ, multi-head attention adds to the hidden state of every token i a vector $\boldsymbol{a}_i^\ell$ that is the sum of its H heads' contributions,

$$\boldsymbol{a}_i^\ell = \sum_{j=1}^{H} \boldsymbol{W}_O^{\ell,j}\, \boldsymbol{a}_i^{\ell,j}, \tag{1}$$

where $\boldsymbol{a}_i^{\ell,j}$ is the output of head j (its attention-weighted sum of value vectors over positions $s \le i$) and $\boldsymbol{W}_O^{\ell,j}$ is the part of the output projection that reads head j. Answers are decoded greedily. Equation 1 is all our intervention relies on, and it holds for every model we study; the models differ in normalization, positional encoding, key–value sharing, and how visual tokens enter the decoder, none of which the intervention touches (Appendix B).

Recognition. Let e be an entity with canonical name $n(e)$, shown in an image. An *identity probe* asks whether the model knows the depicted entity and, if so, to name it, with the output contract `known,<name>` or `unknown,<n>`. The model *recognizes* e when it answers `known` with a name that matches $n(e)$ after normalization (one string contains the other). Each domain also has *control probes* about the same image whose answers are fixed by the pixels (for LOGOS, "what is the the dominant color" and whether the image contains text or a circle) . In turn, we look for components whose removal stops the model from recognizing entities it knows, while leaving the control answers unchanged.

Attention knockout. Let $S \subseteq \{0, \ldots, L-1\}$ be a set of consecutive layers, or a *window*(e.g. $S = \{16, 17\}$). To test how the recognition depends on the layers , we replace the output of every attention head in those layers, at every input position, by that head's mean over the input sequence:

$$\tilde{\boldsymbol{a}}_i^{\ell,j} = \frac{1}{N} \sum_{s=1}^{N} \boldsymbol{a}_s^{\ell,j} \qquad \text{for all } \ell \in S,\ j \in [H],\ i \in [N]. \tag{2}$$

We notate by $\mathrm{KO}(S)$ the network obtained by knocking out S. Because Eq. 1 is linear in the head outputs, $\mathrm{KO}(S)$ adds the *same* vector to every position, so the ablated heads no longer move position-specific information; Unlike zero-ablation, this intervention replaces each head's outputs with their mean over the current input sequence, removing variation across positions while retaining an input-dependent contribution. The intervention acts on the forward pass over the input, from which the first answer token is predicted (Appendix B). We apply $\mathrm{KO}(S)$ to every window of $w \leq W$ consecutive layers, $S = \{\ell, \ldots, \ell{+}w{-}1\}$ with $0 \leq \ell \leq L - w$, where W is the largest width swept ($W{=}4$ for Qwen2.5-VL-7B, 106 windows; $W{=}5$ for the deeper Qwen3-VL and Gemma-4; $W{=}2$ for LLaVA-1.5). The width of the gate itself is a result (§4.1). For head-level analyses, Eq. 2 is applied to a chosen set of heads only.

Measuring the effect of a knockout. Let $\mathcal{E}$ be a set of entities, each with one image. Write $\mathrm{R}(e) \in \{0, 1\}$ for whether the entity is recognized by the intact model e from its image, and $\mathrm{R}_S(e)$ for the same under $\mathrm{KO}(S)$. The *recognition loss* caused by ablating the layers in S is the fraction of entities that the intact model recognizes and the ablated model does not,

$$\mathrm{RL}(S) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathrm{R}(e) \big(1 - \mathrm{R}_S(e)\big), \tag{3}$$

so an entity the model never recognized contributes nothing. Recognition loss alone cannot tell a recognition mechanism from layers the whole network depends on. Ablating the first decoder layer, for instance, abolishes recognition in every model but also every other answer, consistent with the known fragility of the earliest layers (Gromov et al., 2025; Sun et al., 2024; Xiao et al., 2024); we therefore exclude windows containing it. To rule out subtler cases we record, for every window, the *control change*: the fraction of answers to the control probes that change under $\mathrm{KO}(S)$. General damage raises both numbers; a recognition-specific component raises only recognition loss. We call their difference the *familiarity selectivity* of S, and call the narrowest window at which selectivity peaks, with control change no higher than elsewhere, the model's *entity-recognition gate* $\mathcal{G}_M$. Section 3.1 adds one more requirement: that recognition loss track how well an entity is stored. To compare models of different depth we report the relative depth of a window's centre, $\delta(S) = (\min S + \max S)/\big(2(L-1)\big)$.

3 MODELS, DATA, AND FAMILIARITY TIERS

Models. We study four open VLMs that differ in vision encoder, language backbone, and depth: Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), Gemma-4-E4B (Team et al., 2026), and LLaVA-1.5-7B (Liu et al., 2024), with 28 to 42 decoder layers (Appendix Table 5). The reported windows lie in broadly middle decoder regions, with centers at relative depths of 0.48–0.61 (§4.1; Appendix Table 5). The full protocol is expensive, so the in-depth analysis is carried out on one model, Qwen2.5-VL-7B ($L{=}28$, $H{=}28$): it is widely used, and its size keeps the complete sweep and the head-level analyses affordable.

Table 1: Datasets. Entity datasets localize the gate in all four models; the other groups support the in-depth analysis of Qwen2.5-VL-7B only.

	Dataset	Items	What we ask
Entity	LOGOS	352 logos	identity; color, text, circle
	FACES	600 people	identity; color, text, circle
	LANDMARKS	600 landmarks	identity; color, text, circle
	ARTWORK	600 paintings	artist, title; color, tree, person
Ability	VQAV2	214,354 questions	general VQA
	GQA	12,578 questions	compositional VQA
	TEXTVQA	5,000 questions	text reading
	POPE	3,000 questions	object hallucination
Memory	VLMBIAS	9,350 images	identification; counting
	OCCUPATION	342 people $\times$ 2	real-life occupation

Entity datasets. To localize the gate we need real entities that span the range from globally famous to obscure, in several domains. For FACES, LANDMARKS, and ARTWORK we build the datasets from Wikidata (Vrandečić & Krötzsch, 2014) and Wikimedia Commons, using the number of Wikipedia language editions with an article on an entity (its *sitelinks*) as a popularity signal: half of each dataset is sampled from entities with many sitelinks and half from entities with one or two, so that both famous and obscure entities are real and drawn from the same classes. FACES are people with a portrait image, sampled by occupation (politicians, actors, singers, footballers, musicians) with at least 40 sitelinks for the famous half and a single sitelink for the obscure half; ARTWORK are paintings with at least 15 sitelinks or at most 2; LANDMARKS are built the same way from Wikidata landmark classes (monuments, bridges, tourist attractions). Each dataset has 600 entities. LOGOS are 352 brand logos from LogoDet-3K (Wang et al., 2022), a collection that mixes global brands with many obscure ones. Every domain has an identity probe and three control probes (Table 1; examples in Figure 2; queries, thresholds, and filters in Appendix C).

Evaluation datasets. Two further groups of data test what the gate does and does not control. *(i) General ability*: to check that ablating the gate does not damage the model, we evaluate general VQA on VQAV2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), and TEXTVQA (Singh et al., 2019), and hallucination on POPE (Li et al., 2023); none of these requires recognizing a specific entity. *(ii) Memory versus image*: VLMBIAS (Vo et al., 2026) shows familiar subjects with a counterfactual modification (e.g., a four-stripe Adidas-like logo) and asks identification and counting questions that each have an image-grounded and a memory-grounded answer; and OCCUPATION edits the FACES images so that the person wears a doctor's coat, a visual cue that competes with the person's real occupation, together with a *control variant*, the same person edited to wear a plain T-shirt, which shows that the edit itself does not change the answer (both generated with FLUX.2 (Labs, 2025)).

3.1 FAMILIARITY TIERS

Fame is only a proxy for what a particular model has memorized, so we assign tiers from each model's own behavior. The intuition comes from human face perception: familiar faces are recognized from severely degraded images, unfamiliar faces are not (Burton et al., 1999; Hancock et al., 2000). An identity that is stored robustly should likewise survive the loss of part of the visual evidence.

We show each image in five perturbed versions: the right half masked in black, the right half masked in white, only the left half, only the right half, and a 32×32 downsampled version. Let $\mathrm{R}(e)$ be recognition from the clean image (§2) and $\kappa(e)$ the number of perturbed versions on which the model still gives the same name. Then

$$\tau_M(e) = \begin{cases} \text{HK (highly known)} & \text{if } \mathrm{R}(e) = 1 \text{ and } \kappa(e) \geq k, \\ \text{K (known)} & \text{if } \mathrm{R}(e) = 1 \text{ and } \kappa(e) < k, \\ \text{U (unknown)} & \text{if } \mathrm{R}(e) = 0. \end{cases} \tag{4}$$

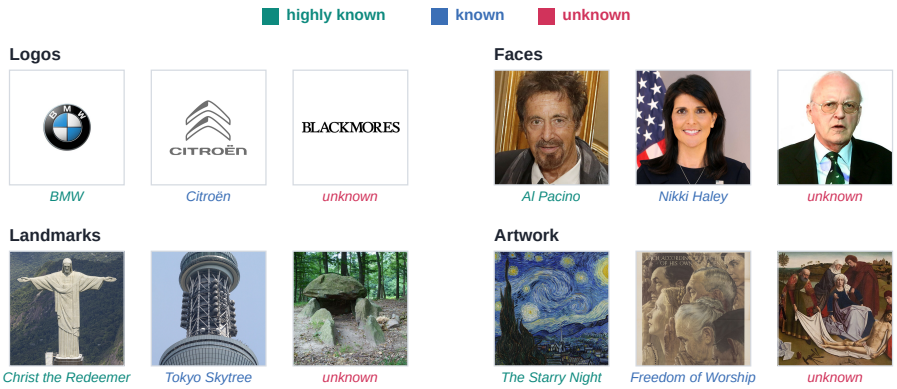

Figure 2: **Tier examples** (Qwen2.5-VL-7B): a highly known, a known, and an unknown entity per domain.

Table 2: Tier sizes (highly known / known / unknown) per model and dataset.

	Logos	Faces	Landmarks	Artwork
Qwen2.5-VL	36 / 218 / 98	44 / 100 / 456	70 / 98 / 432	158 / 226 / 216
Qwen3-VL	63 / 234 / 55	92 / 354 / 154	137 / 352 / 111	301 / 269 / 30
Gemma-4	7 / 249 / 96	30 / 313 / 257	68 / 239 / 293	92 / 267 / 241
LLaVA-1.5	13 / 122 / 217	147 / 280 / 173	10 / 46 / 544	34 / 188 / 378

The threshold k depends on how destructive the perturbations are in each domain. For logos, a highly known brand must be named correctly under all four masking and cropping perturbations (k=4; downsampling is not used). For faces, landmarks, and artworks, where masking half of the image removes much of what identifies it, a highly known entity must receive the same name as on the clean image under at least two of the five perturbations (k=2).

Tiers let us test whether intervention effects differ with baseline recognition robustness. The informative comparison is between the highly known and known tiers, both recognized on the clean image. Recognition loss is zero for the unknown tier by construction, rather than independent evidence of selectivity. These behavioral tiers may also reflect image composition and the visibility of identifying features. Tiers are model-relative and need not agree with fame: Qwen2.5-VL-7B robustly recognizes the Adidas, Starbucks, and Lego logos (HK), recognizes 3M and Accenture only on the clean image (K), and does not identify Airhawk or Aluratek (U; more examples in Appendix Table 6). Table 2 gives the tier sizes for every model.

4 Localizing the Entity-Recognition Gate

4.1 The gate across models

Figure 3 plots the Recognition loss and control change for every window of up to five consecutive layers, per model, tier, and relative depth.

Recognition loss peaks in the middle of every model. In In all four models, recognition loss on the known tier is low across most of the network and rises within a band between roughly 40% and 65% of the decoder's depth: L16–17 in Qwen2.5-VL-7B (δ=0.61; recognition loss of 94–100% on the known tier in every domain); layer 15 in LLaVA-1.5 (δ=0.48; 80–100% except logos, which respond only to layer 19); layers 18–22 in Gemma-4 (δ=0.49; 48–99%); and layers 15–19 in Qwen3-VL (δ=0.49; 38–92%).

Tiers differ in recognition, not in perception. The control curves of the two tiers in Figure 3 lie on top of each other: at every model's gate, control change for entities in the known and highly-known tiers differs by at most three points (Qwen2.5-VL-7B faces 18.2% vs. 15.7%). The recognition

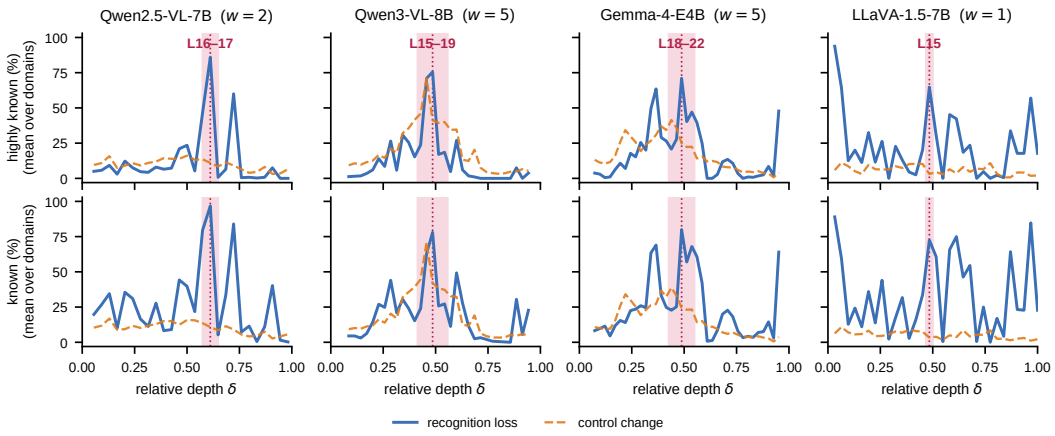

Figure 3: **Recognition loss and control change across models.** For every window of w consecutive layers, plotted at the window's relative depth: recognition loss (blue) and control change (orange, dashed), each averaged over the four entity domains, on highly known (top) and known (bottom) entities. Shaded band: each model's gate window; dotted line: its center. Recognition loss peaks in the middle of every model and is graded by tier; control change is the same for both tiers and stays low at the gate only where the gate is narrow (Qwen2.5-VL-7B, LLaVA-1.5). Per-domain curves are in Appendix Figure 7.

curves do not: known entities lose recognition at least as often as highly known ones and often more (Qwen2.5-VL-7B faces 94% vs. 77%; Qwen3-VL faces 45% vs. 3%), and across more windows. Recognition loss depends on the tier; control change does not, which is the signature required in §3.1.

Only narrow windows are selective. Qwen2.5-VL-7B and LLaVA-1.5, whose gates are one or two layers wide, satisfy the criterion of §2: control change at the gate is 3–18% and 1–7%, against recognition loss of 77–100%. Gemma-4 and Qwen3-VL reach comparable recognition loss only with five-layer windows, and at that width performance on the control probes degrades as well: control answers change on 13–31% of images at Gemma-4's gate and on 35–46% at Qwen3-VL's recognition-loss peak.

The changed control answers do not become random: each yes/no probe flips in a single direction ("circle" no→yes on 219 of 219 changed landmarks for Qwen3-VL at layers 15–19, "tree" no→yes on 319 of 319 artworks, "text" yes→no on most landmarks) and colors drift to a few defaults, i.e., under these windows the control answers become the same across images. A plausible reading, which we do not test directly, is that a five-layer window also removes image information that the control questions depend on, so that it removes more than the entity. The failure mode also differs by model: Qwen2.5-VL-7B and LLaVA-1.5 answer `unknown` on about 90% of the known tiers and rarely name another entity ($\leq$8%), whereas Gemma-4 often substitutes another entity (64% of known-tier faces) and Qwen3-VL answers `known` without producing a name (Appendix D).

We therefore report a gate with largely unchanged control performance for Qwen2.5-VL-7B and LLaVA-1.5, and a wider, partly selective band for Gemma-4 and Qwen3-VL.

4.2 Anatomy of the gate in Qwen2.5-VL-7B

We now characterize one model's gate in depth: which entities it affects, what it leaves intact, and which heads implement it.

Which tiers it affects. A deeper look at this model reveals that for highly-known-tier logos, recognition loss is essentially zero everywhere *except* at the gate (97%); the known tier peaks as high over a wider band; the unknown tier, which has no recognition to lose, stays at zero everywhere. The same holds for the other domains (Appendix E); for landmarks and artworks a window extending to layer 19 is stronger still, so the gate is a narrow band centered on L16–17.

Table 3: General VQA accuracy (%) in Qwen2.5-VL-7B, intact and under KO($\{16, 17\}$), with bootstrap 95% intervals on the paired difference (official splits).

Benchmark	n	Intact	KO($\{16, 17\}$)	Δ [95% CI]
VQAv2 (all)	214,354	79.5	78.2	-1.31 [$-1.40, -1.22$]
yes/no	80,541	94.7	93.8	-0.9
number	28,134	73.8	69.8	-4.0
other	105,679	69.5	68.6	-0.9
GQA	12,578	58.2	55.1	-3.13 [$-3.63, -2.61$]
TextVQA	5,000	74.6	72.2	-2.37 [$-3.03, -1.71$]
POPE	3,000	86.2	83.1	-3.10 [$-4.00, -2.27$]

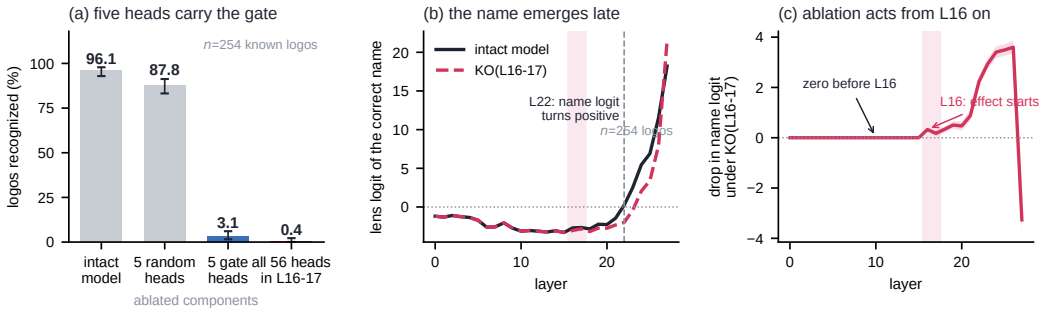

Figure 4: **Which heads carry the gate, and when the name is written** (Qwen2.5-VL-7B, Logos, n=254 logos in the known and highly-known tiers). (a) Recognition with different components ablated: the five heads of L16–17 whose individual ablation reduces recognition most remove it almost as completely as all 56 heads of those layers; five random heads from the same layers do not (bars: Wilson 95% intervals). (b) Logit-lens logit of the correct name at the last position, by layer: in the intact model it turns positive only at layer 22. (c) Drop in that logit under KO($\{16, 17\}$): zero through layer 15 and growing from layer 16 on (band: 95% interval across logos).

Performance on perceptual probes. A knockout that lowered recognition by degrading the visual input would also change the answers to perceptual questions. To test this on the same images, we edit the hair or eye color of each Faces portrait and ask both who the person is and what color the edited attribute is (Appendix F). Under KO($\{16, 17\}$) the share of highly-known-tier people still named as on the unedited photo falls from 58.3% to 14.9%, while the reported color of the edited attribute is unchanged on 99.0% of the same images.

Performance on general benchmarks. KO($\{16, 17\}$) lowers accuracy by 1.3 to 3.1 points on four general VQA and hallucination benchmarks (Table 3), against 46 points of recognition; while the VQAv2 errors are mostly in *number* questions (-4.0; see §5.1).

Which heads implement it, and when. Two analyses on Logos look inside the gate (Figure 4; Appendix G). First, we ablate the five heads of L16–17 whose individual ablation reduces recognition most (four in layer 16, one in layer 17 - heads 1, 14, 17, and 20 of layer 16 and head 21 of layer 17). Re-measured in this run (Appendix G), the intact model recognizes 96.1% of the 254 logos in the known and highly-known tiers; recognition falls to 3.1% with these five heads removed, close to the 0.4% obtained by removing all 56 heads of the two layers, whereas five random heads from the same layers leave recognition at 87.8%. Second, a logit lens (nostalgebraist, 2020) at the last position tracks the layerwise readout of the correct name: in the intact model the logit of the correct name stays negative until layer 22 and then rises sharply, well after the gate. Under KO($\{16, 17\}$) the curve is identical up to layer 15 and falls below the intact curve from layer 16 on, the gap growing to 3.5 logits by layer 25.

Table 4: VLMBIAS identification questions: image-grounded accuracy (%) in Qwen2.5-VL-7B, intact and under KO($\{16, 17\}$). Counting results in Appendix Table 8.

	Chess pieces	Animals	Logos	Patterned grids	Flags	Game boards	Optical ill.
Intact	71.4	5.6	10.8	91.2	38.9	70.1	27.6
KO($\{16, 17\}$)	95.2	29.1	27.1	100.0	46.1	72.8	29.9
Δ	+23.8	+23.5	+16.2	+8.8	+7.2	+2.7	+2.3



Figure 5: **OCCUPATION examples** Original portraits with their T-shirt (control) and doctor's-coat edits; identity, pose, and background are preserved.

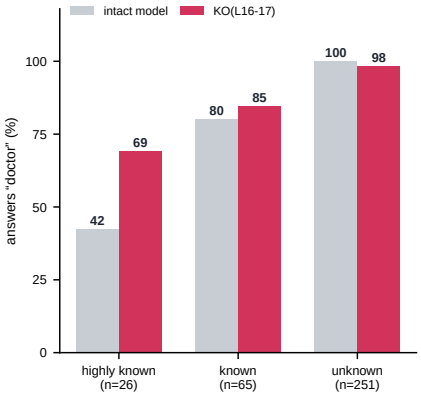

Figure 6: **OCCUPATION results** (Qwen2.5-VL-7B): share answered "doctor" on the doctor's-coat image, by tier.

5 APPLICATIONS: WHEN MEMORY AND IMAGE DISAGREE

If the gate decides whether a VLM consults its memory, removing it should change behavior whenever memory and image disagree. We test this in Qwen2.5-VL-7B on two tasks.

5.1 MITIGATING MEMORY-DRIVEN ERRORS

VLMBIAS shows familiar subjects with a counterfactual modification and asks which subject it is or how many parts it has. Under KO($\{16, 17\}$), image-grounded identification improves in all seven topics, significantly in five, by up to +23.8 points (Table 4); counting shows no consistent improvement, since a biased identification *is* a retrieved entity label while a count is not (Appendix I). Raising image resolution from 384 to 1152 pixels leaves baseline logo accuracy at 12–13%, so the gains are not from seeing better.

5.2 MEMORIZED VERSUS IMAGE-SUPPORTED ANSWERS

VLMBIAS shows the gate removing a memorized answer; the second application asks whether the gate also determines which answer is given when the image supports one answer and the memorized occupation another. We built OCCUPATION dataset for this purpose: a sample of FACES images is edited so that the person wears a doctor's coat (examples in Figure 5), and the model is asked "What is this person's real-life occupation?". The image supports "doctor"; the memorized occupation is a different one, as in Figure 1b, where Lionel Messi is answered "footballer" by the intact model and "doctor" under the ablation. A matched T-shirt edit serves as control and never elicits a medical answer (0%), so the coat is the only cue. Figure 6 shows the share of "doctor" answers by tier: intact, it is 42% in the highly-known tier, 80% in the known tier, and under KO($\{16, 17\}$) it rises to 69%, 85%. Ablating the gate thus moves answers from the memorized occupation to the image-supported one, most in the tier where the memorized answer was given most often. This shift measures increased reliance on clothing cues, not improved accuracy about real-life occupation, which the edit does not change.

6 Related Work

Memory versus the image in VLMs. Reliance on language priors is well documented in VQA (Goyal et al., 2017; Agrawal et al., 2018). Related work examines object hallucination, visual shortcomings, and counterfactual reasoning (Li et al., 2023; Tong et al., 2024; Bitton-Guetta et al., 2023; Lee et al., 2025).Other studies examine conflicts across modalities and between parametric knowledge and visual evidence (Zhu et al., 2024; Liu et al., 2025a; Vo et al., 2026). Closest to our work, predictions move from priors to visual evidence in mid-to-late layers (Golovanevsky et al., 2025) and a few attention heads mediate such conflicts (Ortu et al., 2026; Lietzow et al., 2026). These works study visual-attribute and commonsense conflicts, we study identification of *specific entities*, using familiarity tiers and matched controls to compare recognition loss with perceptual changes.

Recall and entity knowledge in language models. Factual recall in text-only models has been traced to feed-forward memories (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022) and to staged extraction by attention (Geva et al., 2023); whether a model knows an entity can be read from its representations and controls whether it answers or refuses (Gottesman & Geva, 2024; Ferrando et al., 2025; Lindsey et al., 2025). In text, memory-context conflicts have been studied behaviorally (Longpre et al., 2021; Xie et al., 2024) and linked to competing attention heads (Jin et al., 2024). Whether a model can report the limits of its own knowledge has also been studied in text: models show imperfect awareness of what they know (Kadavath et al., 2022; Yin et al., 2023), and the truth of factual statements can be linearly decoded from activations (Marks & Tegmark, 2024). Our tiers measure recognition robustness behaviorally, and the gate causally influences the model's identity reports.

Localization and its limits. Circuit discovery and activation patching identify components implicated in model behavior (Conmy et al., 2023; Zhang & Nanda, 2024), while open-vocabulary probes interpret hidden representations (Ghandeharioun et al., 2024). Localization methods have been evaluated using injection and deletion benchmarks (Chang et al., 2024), but localization does not by itself identify the best editing site (Hase et al., 2023). We therefore report ablation effects and leave transplantation to future work.

Removing entity knowledge from VLMs. Knowledge can be removed or modified by editing weights, and multimodal editing benchmarks and methods exist (Cheng et al., 2023; Basu et al., 2024). Machine unlearning for VLMs removes the recognition of concepts or identities by fine-tuning (Li et al., 2024), with benchmarks built on fictitious and real profiles (Liu et al., 2025b; Ma et al., 2025; Dontsov et al., 2025), and with pruning of modality-specific neurons (Liu et al., 2025c). These methods change the weights and are evaluated by what the model forgets. Our intervention is inference-time and reversible, and we use it to localize computations involved in entity identification, recognition removed at the gate leaves perception and general ability largely intact.

7 Discussion and Conclusion

We asked where entity recognition happens in a VLM, and found an entity-recognition gate: a narrow band of mid-network attention layers whose ablation removes recognition of entities in the known tiers while, where the band is narrow, performance on perceptual probes is nearly unchanged; recognition loss peaks at the same relative depth in all four models. Ablating the gate turns identity answers into "unknown" while answers to perceptual questions about the same image stay the same, and it raises image-grounded accuracy when the image and the memorized answer differ, at a cost of 1–3 points on general benchmarks.

In terms of limitations, we localize the gate with layer windows, the simplest intervention that exposes it; in the two deeper models this resolution is too coarse to separate recognition from the control answers, and head-level interventions of the kind we apply to Qwen2.5-VL-7B would be needed. Finally, knockouts show that the gate is necessary for recognition, not that it is sufficient; whether transplanting its output transfers identity between entities, and how the gate forms during training, are natural next steps. The interventions do not distinguish disruption of identity-relevant visual processing, retrieval, confidence, or answer production.

AI USE STATEMENT

We used generative AI tools as coding and writing assistants. Specifically, we used them to help implement experiment scripts (model-loading adapters, attention-head ablation hooks, evaluation and plotting code) and to assist with drafting and editing the manuscript text. We used generative AI tools to obtain feedback on the final paper (simulating a review process), and to assist with literature lookup and submission checks. We used a generative image editor, FLUX.2, to produce the edited images of the FACE EDITS and OCCUPATION datasets (changed hair or eye color; doctor's coat and T-shirt); the source portraits, edit instructions, and the selection and scoring of the edited images are the authors' own. We did not use generative AI tools to generate research ideas, to design the core hypotheses or experimental protocol, or to produce any experimental results, or numbers; all reported quantitative results were produced by running our own code on the stated models and datasets. AI-assisted code was reviewed and tested for correctness by the authors, including verification against known baselines (e.g. reproducing the localization result before extending it), and all statistical analyses were independently checked. We reviewed all AI-assisted text for accuracy and revised it to reflect our own claims. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work studies how vision–language models recognize and recall knowledge about real, named entities (public figures, brands, landmarks, and artworks) using publicly available images. Because the analysis involves images of identifiable public figures, we restrict our study to well-known public figures in contexts relevant to their public role, use images only for scientific analysis of model behavior, and do not attempt to identify private individuals or infer sensitive personal attributes. Our counterfactual stimuli are digital clothing edits (e.g. a doctor's coat) applied to study occupation recall; these are used solely to probe model behavior and are not presented as genuine depictions.

A central contribution, identifying and removing the mechanism by which a VLM commits an entity identity, has a dual-use character. On the beneficial side, the same intervention mitigates identification bias on VLMBIAS, reducing a model's tendency to answer from memorized priors rather than from the image, and it offers a lightweight, prompt-based control over prior-versus-visual conflicts. On the risk side, a method that selectively erases an entity's stored identity could in principle be misused to degrade or manipulate a model's recognition of specific entities. We mitigate this by intervening only at inference time (a reversible activation hook, with no weight edits or persistent changes), by evaluating on public, non-sensitive entity categories, and by framing the intervention as an interpretability and bias-mitigation tool operating on already-public models and images. We declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The problem setup, intervention (mean-ablation of attention heads via a reversible forward hook), and metrics (familiarity selectivity, recognition status, tier grading) are defined in Section 2, including the exact model, layer/head counts, and default gate configuration. The localization procedure (exhaustive single-layer and adjacent-pair sweeps against matched perceptual controls) is described in Section 4, and the application protocols for VLMBIAS (Section 5.1) and the counterfactual occupation task (Section 5.2) specify the prompts, image variants, and paired comparisons. Implementation details—hook targets, answer-scoring rules (the medical-term matcher and the concrete-vs-unknown status detector), familiarity-tier derivation, and statistical tests are given in Appendix B. We use publicly available models and datasets throughout, and describe all data processing (perturbation variants, clothing edits, tier labels) in Section 3. We will publicly release all code and per-experiment configurations.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980. IEEE, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Ke qin Chen, Xue-Jing Liu, Jia-Lin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jian-Qiang Wan, Peng-Fei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025b.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. volume 37, pp. 7400–7426, 2024.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2616–2627. IEEE, 2023.

A Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, 1999.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. Do localization methods actually localize memorized data in llms? a tale of two benchmarks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3190–3211, 2024.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 13877–13888, 2023.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. Performance gap in entity knowledge extraction across modalities in vision language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29095–29108, 2025.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. pp. 20582–20603, 2025.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, volume 2025, pp. 77965–78000, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5484–5495, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024.

Michal Golovanevsky, William Rudman, Michael A Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24848–24863, 2025.

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers. 2025:81906–81920, 2025.

Peter JB Hancock, Vicki Bruce, and A Mike Burton. Recognition of unfamiliar faces. *Trends in cognitive sciences*, 4(9):330–337, 2000.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. volume 36, pp. 17643–17668, 2023.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 248–264, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. volume 37, pp. 35414–35453, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Niclas Lietzow, Danielle Bitterman, Carsten Eickhoff, William Rudman, and Michal Golovanevsky. Vision-default, prior-override: Causal mechanisms of perception-knowledge conflict in vision-language models. *arXiv preprint arXiv:2606.28273*, 2026.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17825–17846, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4105–4135, 2025b.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5913–5933, 2025c.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7052–7063, 2021.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, et al. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *International Conference on Learning Representations*, volume 2025, pp. 87806–87827, 2025.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. volume 35, pp. 17359–17372, 2022.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130, 2026.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

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

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, THI TUONG VY DANG, Anh Nguyen, and Daeyoung Kim. Vision language models are biased. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 3689–3774, 2026.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(1), January 2022. ISSN 1551-6857. doi: 10.1145/3466780.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 35623–35646, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025. doi: 10.1109/CVPR52734.2025.01842.

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024.

A Appendix

B Implementation Details

Architectures. All four models are decoder-only transformers in which each layer adds an attention output and an MLP output to the residual stream, $\boldsymbol{x}_i^\ell = \boldsymbol{x}_i^{\ell-1} + \boldsymbol{a}_i^\ell + \boldsymbol{m}_i^\ell$, and in which the attention output is a linear projection $\boldsymbol{W}_O$ of the concatenated head outputs (Eq. 1). Our hook intercepts the input to $\boldsymbol{W}_O$, so it applies unchanged to all of them. The models differ in details the intervention does not touch (normalization placement, positional encoding, key–value sharing, and how visual tokens enter the decoder): Qwen2.5-VL-7B (L=28, H=28, d_h=128, four key–value groups) uses a native-resolution ViT with 2D rotary positions and grouped-query attention (Bai et al., 2025b); Qwen3-VL-8B (L=36) injects visual features into several early decoder layers (Bai et al., 2025a); Gemma-4-E4B (L=42) is the 4B-effective model of the Gemma-4 family (Team et al., 2026); LLaVA-1.5-7B (L=32, H=32) connects a CLIP ViT-L/14 encoder to a Vicuna-7B backbone through a two-layer MLP (Liu et al., 2024). Table 5 lists depth and gate per model.

Table 5: Models, depth, and the gate window found in §4.1; δ is the relative depth of the window's center and "sel." its selectivity (recognition loss minus control change, mean over domains).

Model	Layers L	Gate window	Width	δ	Sel.
Qwen2.5-VL-7B	28	16–17	2	0.61	82
LLaVA-1.5-7B	32	15	1	0.48	67
Gemma-4-E4B	42	18–22	5	0.49	55
Qwen3-VL-8B	36	15–19	5	0.49	27

Hooks and scope. $\mathrm{KO}(S)$ is a forward pre-hook on the attention output projection (`o_proj`) of each layer in S: it receives the concatenated head outputs of shape (B, N, Hd_h), reshapes them to (B, N, H, d_h), replaces every position by the mean over the sequence axis, and reshapes back (Eq. 2). Head-level knockouts apply the same average to selected heads only. Hooks are registered in a context manager for one generation and removed afterwards; the intact condition runs through the same code path with $S = \emptyset$. The hook acts on the forward pass over the input tokens $t_1, \ldots, t_N$, from which the first answer token is predicted, so every answer is affected; subsequent tokens are decoded with a key–value cache and attend to the cached keys and values of the input, which in layers above $\min S$ were computed from ablated hidden states.

Decoding and scoring. Answers are generated greedily with a per-probe token budget (identity probes 60 tokens; control probes 5–8). The identity parser lowercases the answer, strips quotes and line breaks, and splits on the first comma; answers not beginning with `known` count as unknown, and a name matches if one normalized string contains the other. Color answers are mapped to a fixed list of eleven color names, yes/no answers to two classes. In Occupation, medical answers are detected by case-insensitive word-boundary matching against a fixed list of medical terms.

C Dataset and Tier Details

Entity datasets. Faces, Landmarks, and Artwork are built from Wikidata (Vrandečić & Krötzsch, 2014) and Wikimedia Commons by SPARQL queries anchored on a class, with the number of Wikipedia language editions linking to an entity (*sitelinks*) as the popularity signal; half of each dataset is sampled from the high-sitelink and half from the low-sitelink range of the same class. Faces: humans (P31 = Q5) with a portrait (P18), queried by occupation (P106: politician, actor, singer, footballer, musician) and filtered with a face detector; famous $\geq$40 sitelinks, obscure =1. Artwork: paintings (P31 = Q3305213) with artist and title; famous $\geq$15 sitelinks, obscure $\leq$2. Landmarks: Wikidata landmark classes (monuments, bridges, tourist attractions), sampled the same way. Each has 600 entities. Logos: 352 logos from LogoDet-3K (Wang et al., 2022), including the text-only version of wordmark brands.

Evaluation datasets. Face edits: the Faces images with one natural attribute edited with FLUX.2 (Labs, 2025), four edits per person (gray, brown, or blonde hair; green eyes; 1,012 images

over 253 people). OCCUPATION: the FACES images edited with FLUX.2 (Labs, 2025) so that the person wears a white doctor's coat with a stethoscope, plus a matched plain-T-shirt edit (Figure 5); 342 people. VLMBIAS (Vo et al., 2026): 9,350 counterfactual images of familiar subjects in seven topics with identification and counting questions, each with an image-grounded and a memory-grounded answer. General VQA: the official splits of VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), TEXTVQA (Singh et al., 2019), and POPE (Li et al., 2023).

Tiers. The five perturbations used for tiering are right-half black mask, right-half white mask, left-half crop, right-half crop, and 32×32 downsampling (rotation is generated but not used). For logos, highly known requires the correct brand under the four masking and cropping perturbations; for the other domains, the same name as on the clean image under at least two of the five. For the occupation experiments, unknown entities are further split into *slightly known* (named when asked directly, "Identify this celebrity or public figure by name") and *never known*; §5.2 reports them merged. Table 6 lists examples.

Table 6: Examples of familiarity tiers assigned by Qwen2.5-VL-7B.

Domain	Highly known	Known	Unknown
LOGOS	Adidas, Starbucks, BMW	3M, Accenture, Citroën	Airhawk, Aluratek, Blackmores
FACES	Al Pacino, Barack Obama	Nikki Haley, Britney Spears	Roman Herzog, Mike Oldfield
LANDMARKS	Christ the Redeemer, Gateway of India	Tokyo Skytree, Column of Marcus Aurelius	Albersdorf dolmen, a kurgan near Bratske
ARTWORK	*The Starry Night*, *Campbell's Soup Cans*	*Freedom of Worship*	*Lamentation of Christ* (Q3209944)

D GATE ACROSS MODELS

Figure 7 gives the per-domain version of Figure 3, and Table 7 the best interior windows per model by selectivity.

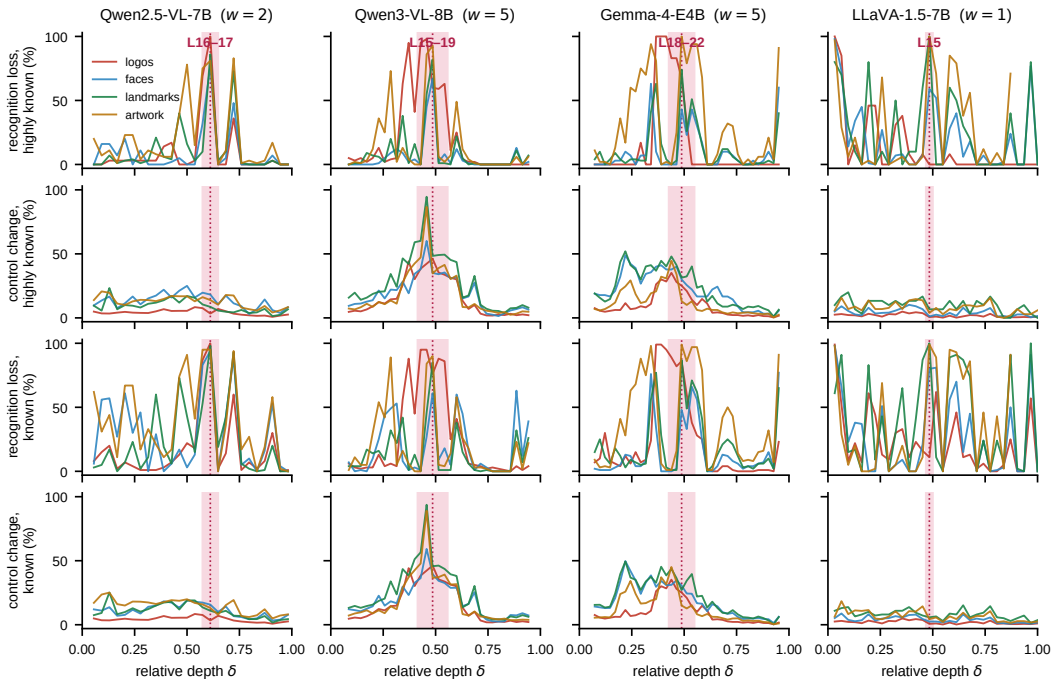

Figure 7: **Per-domain version of Figure 3.** Recognition loss (rows 1 and 3) and control change (rows 2 and 4) for every window of w consecutive layers, on the highly-known and known tiers, per entity domain. Shaded band: each model's gate window; dotted line: its center. Logo control curves are computed over all recognized logos in both tier rows, as the logo files of the other models do not carry per-entity tier labels.

Table 7: Best interior windows (first and last two layers excluded) by selectivity, with recognition loss and control change (%, mean over domains, highly-known and known tiers). Per-domain losses in parentheses (logos / faces / landmarks / artworks).

Model	Window	Sel.	Loss	Ctl.	Loss per domain
Qwen2.5-VL-7B	16–17	82	93	11	99 / 89 / 93 / 89
	19–20	68	77	9	57 / 76 / 85 / 89
LLaVA-1.5	15	67	70	3	8 / 73 / 100 / 100
	19	67	70	4	47 / 59 / 84 / 91
Gemma-4	18–22	55	80	25	87 / 48 / 84 / 99
	21–25	43	58	14	27 / 53 / 56 / 95
	20–23	35	54	18	36 / 45 / 47 / 87
Qwen3-VL	15–19	27	69	42	38 / 63 / 84 / 92
	20–24	10	23	13	24 / 37 / 12 / 18

Additional detail on the control and identity answers. The default colors differ by model: Qwen3-VL drifts to blue, green, or brown and Gemma-4 to gray or brown. Qwen2.5-VL-7B shows the same one-directional pattern at a much smaller scale at its own gate ("circle" no→yes on 46 of 144 faces). For identity, Gemma-4's substitutions are mostly other famous entities of the same type (e.g., Admiralty Arch → "Eiffel Tower"), and Qwen3-VL's bare `known` answers occur on 84–92% of landmarks and artworks but only 22% of faces, where it more often names a different person (36%).

E Layer Sweep in Qwen2.5-VL-7B

Figure 8 shows the complete single-layer and adjacent-pair sweep of Qwen2.5-VL-7B on Logos with the two measures of §2; Figure 3 in the main text shows their average over domains, and Figure 7 the other domains. Two things are visible here. First, recognition loss on the known tiers peaks sharply at L16–17 (99%) with control change of 3%, and no other adjacent pair comes close except a secondary peak at layers 19–20 (57%). Second, single layers 16 and 17 on their own remove far less (left), so the two layers act together. The same pattern holds in the other domains (per-domain losses at L16–17: 89% faces, 93% landmarks, 89% artworks), where a window extending to layer 19 is sometimes stronger still, which is why §4.2 describes the gate as a band centered on L16–17.

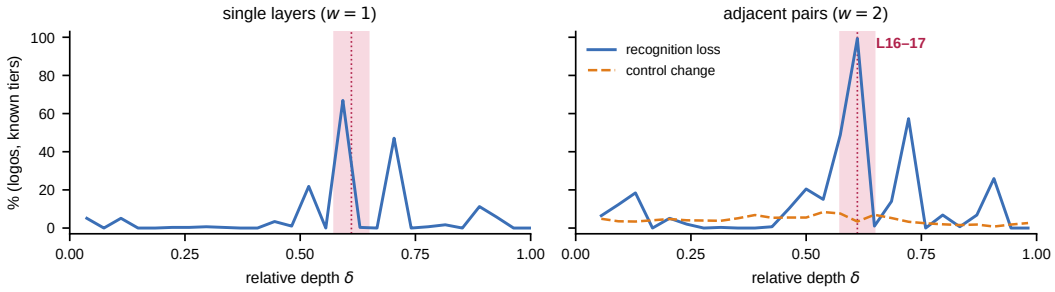

Figure 8: **Exhaustive sweep in Qwen2.5-VL-7B on Logos.** Recognition loss on the known tiers (blue) for every single layer (left) and every adjacent pair (right), at the window's relative depth, with control change (orange, dashed; recorded for adjacent pairs). Shaded band: the gate L16–17; dotted line: its center.

F Recognition and Perception on the Same Image

The control probes of the sweep ask generic questions, about a dominant color or the presence of text, that a model could answer while ignoring the entity. A stronger test of the dissociation asks two questions about the same image: one that requires recognizing the person and one that can only be answered by looking, because its answer was changed in the picture and has no memorized

counterpart. If the gate removes recognition without degrading perception, the first answer should change under KO($\{16, 17\}$) and the second should not.

To construct such pairs we took the 253 people of the face run (42 highly known, 111 known, and 100 unknown by that run's tiering) and used FLUX.2 (Labs, 2025), the same editor as for the OCCUPATION edits, to change one natural attribute of each portrait, producing four edits per person: gray, brown, or blonde hair, and green eyes (Figure 10; 1,012 images). Two further edits, hazel eyes and auburn hair, were generated but are not scored, because the model never uses either word in a one-word answer (it says "brown" for 67% of hazel edits and "red" for 98% of auburn ones). For every edited image we ask the identity and attribute questions of Appendix J and record whether the model returns the same name as on the unedited photo and whether its attribute answer matches the edit.

The two answers come apart (Figure 9). Under KO($\{16, 17\}$), the share of edited images on which the model still returns the same name falls from 58.3% to 14.9% in the highly-known tier (n=168 images; 73 losses and 0 gains), from 37.8% to 8.1% in the known tier (n=444), and from 1.8% to 0.0% in the unknown tier (n=400). On the same images, the attribute answer matches the edit for 97.0% before the knockout and 97.6% after, and is literally unchanged on 99.0%. Three conclusions follow. The knockout does not act by degrading the visual input, since an answer that can only be read from the pixels is untouched. What it removes is graded by tier, as recognition was in the sweep. And the effect is specific to the gate rather than to any nearby layer: ablating layer 19 alone lowers identity preservation in the highly-known tier only from 58.3% to 44.0%, and layer 20 alone leaves it at 58.3%.

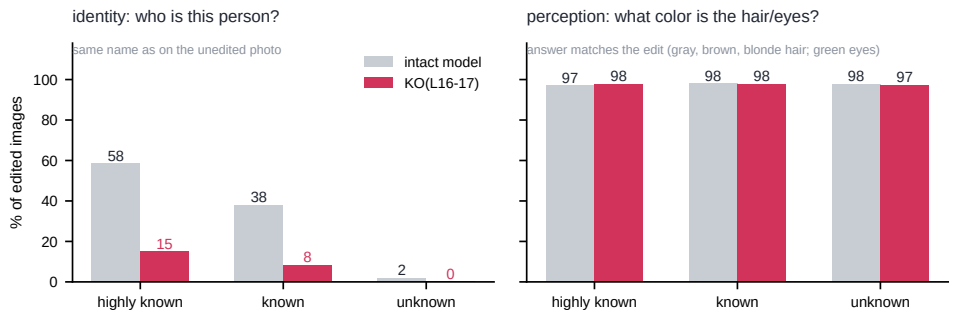

Figure 9: **Recognition loss without change in the perceptual probe** (FACE EDITS, Qwen2.5-VL-7B). *Left:* identity is preserved for far fewer images under KO($\{16, 17\}$), graded by tier. *Right:* on the same images, the reported hair or eye color still matches the edit at the same rate.

G HEAD-LEVEL AND LOGIT-LENS DETAILS

Prompt. These runs use the logo identity probe of Appendix J, re-run from scratch; the intact model recognizes 96.1% of the 254 logos that the sweep had placed in the known and highly known tiers (10 logos differ between the two runs).

Heads. Each of the 56 heads of L16–17 is ablated individually and scored by the recognition it removes; the five most selective are heads 14, 1, 20, and 17 of layer 16 and head 21 of layer 17. Ablating these five leaves 3.1% of logos recognized, all 56 heads 0.4%, and five random heads from the same layers 87.8% (Figure 4a).

Logit lens. At the last position we project the residual stream after each layer through the final normalization and unembedding (nostalgebraist, 2020) and record the logit of the first token of the correct name. In the intact model it stays negative until layer 22 and then rises sharply; under KO($\{16, 17\}$) it is identical up to layer 15 and falls below the intact curve from layer 16 on, with the gap reaching 3.5 logits by layer 25.

A familiarity direction is weaker than the heads. The difference of mean residual streams between the known and unknown tiers of faces at layer 16 defines a direction. Projecting it out



Figure 10: **FACE EDITS examples.** Original portraits and their four edits (gray, brown, and blonde hair; green eyes).

reduces recognition in the highly-known tier from 92% to 84% and in the known tier from 84% to 67%, with no effect on the unknown tier and none for a random direction, but does not reduce errors on VLMBIAS (29%→30%): the gate is better captured by its heads than by a single direction.

H NEGATIVE CONTROLS: ANSWERS THE GATE DOES NOT CHANGE

Occupation answers change, hair-color answers do not. On unedited faces we ask separately for the person's occupation (identity-dependent) and hair color (perceptual). Under $\mathrm{KO}(\{16, 17\})$ the occupation answer changes for 50% of the highly-known and 77% of the known tier, while hair color changes for 2–5%; explicit "unknown" answers rise from 2% to 62% (highly known) and from 8% to 85% (known), so the change is a refusal, not a substitution.

Color answers on grayscale images do not depend on the gate. On desaturated face images the model still names hair and eye colors, agreeing with its color-image answer for ∼80% of highly-known-tier faces and equally (∼77%) for unknown-tier faces; the answers follow luminance, and $\mathrm{KO}(\{16, 17\})$ leaves them unchanged while recognition on the same images falls from 83.4% to 19.2%.

Counting errors are not recovered by patching any single site. Patching clean residual streams into counterfactual runs on 546 VLMBIAS animal pairs, at every layer and for image, text, or final positions, recovers essentially none of the true-count logit.

I FULL VLMBIAS RESULTS

Table 8 gives every VLMBIAS question type. Identification accuracy rises in all seven topics under $\mathrm{KO}(\{16, 17\})$, most where the intact model's answer is a memorized label (chess pieces, animals,

logos). Counting accuracy never rises and falls most on FLAGS (-12.7 and -10.5 points on the two counting questions), where the count asked for (e.g., the number of stars) is one the intact model answers from the recognized flag rather than by counting; once the flag is no longer recognized, the model has to count, and does so less accurately. The other counting drops are within a few points. Increasing image resolution from 384 to 1152 pixels leaves baseline logo identification at 12–13%, so the identification gains are not due to seeing better.

Table 8: VLMBIAS image-grounded accuracy (%) in Qwen2.5-VL-7B, intact and under KO($\{16, 17\}$), per question type and topic (counting rows with a change of at least one point).

Question	Topic	n	Intact	KO($\{16, 17\}$)	Δ
Identification	Chess pieces	252	71.4	95.2	$+23.8$
	Animals	409	5.6	29.1	$+23.5$
	Logos	314	10.8	27.1	$+16.3$
	Patterned grids	294	91.2	100.0	$+8.8$
	Flags	180	38.9	46.1	$+7.2$
	Game boards	147	70.1	72.8	$+2.7$
	Optical illusions	693	27.6	29.9	$+2.3$
Counting	Flags (Q2)	252	31.0	18.3	-12.7
	Flags (Q1)	285	27.7	17.2	-10.5
	Optical ill. (Q2)	792	55.7	51.5	-4.2
	Chess (Q1)	432	3.9	0.7	-3.2
	Patterned grids (Q1)	504	19.2	16.3	-2.9
	Game boards (Q1)	240	17.5	15.0	-2.5
	Optical ill. (Q1)	1,188	55.6	53.4	-2.2
	Logos (Q2)	483	12.6	14.5	$+1.9$

J PROMPTS

All prompts are given verbatim. Identity probes share one output contract; the value after `unknown` (a count of visible elements) is never used and only keeps the two branches of the contract symmetric.

Identity probes. FACES:

```
Return exactly: "<status>,<value>"
Rules:
- <status> must be either "known" or "unknown".
- Do you know who the person is in the image? Answer with
"known" if yes and "unknown" otherwise.
Output:
- If status is "known", <value> must be the person's full
real-world name in lowercase (e.g. "barack obama").
- If status is "unknown", <value> must be the number of
distinct visible elements in the image as digits only.
Examples: known,barack obama / known,serena williams /
known,roger federer / unknown,4
Do not explain your reasoning. Output only the answer.
```

LANDMARKS (identity):

```
Return exactly: "<status>,<value>"
Rules:
- <status> must be either "known" or "unknown".
- Do you know which specific real-world landmark is shown
in the image? Answer "known" if you recognise it (a famous
building, monument, natural wonder, bridge, temple, tower,
statue, square, etc.), and "unknown" otherwise.
Output:
- If status is "known", <value> must be the landmark's full
real-world name in lowercase (e.g. "eiffel tower").
```

```
- If status is "unknown", <value> must be the number of
distinct visible elements in the image as digits only.
Examples:  known,eiffel tower / known,statue of liberty /
known,taj mahal / unknown,5
Do not explain your reasoning.  Output only the answer.
```

LANDMARKS (country; a second familiarity probe, not counted as a control):

```
Return exactly:  "<status>,<value>"
Rules:
- <status> must be either "known" or "unknown".
- Do you know which country the landmark in the image is
located in?  Answer "known" if you can name the country, and
"unknown" otherwise.
Output:
- If status is "known", <value> must be the country name in
lowercase (e.g.  "france", "united states", "india").
- If status is "unknown", <value> must be the number of
distinct visible elements in the image as digits only.
Examples:  known,france / known,united states / known,india
/ unknown,4
Do not explain your reasoning.  Output only the answer.
```

ARTWORK (artist):

```
Return exactly:  "<status>,<value>"
Rules:
- <status> must be either "known" or "unknown".
- Do you know which artist painted the artwork in the image?
Answer "known" if you recognise the artist, and "unknown"
otherwise.
Output:
- If status is "known", <value> must be the artist's full
real-world name in lowercase (e.g.  "vincent van gogh").
- If status is "unknown", <value> must be the number of
distinct visible elements in the image as digits only.
Examples:  known,vincent van gogh / known,leonardo da vinci
/ known,claude monet / unknown,5
Do not explain your reasoning.  Output only the answer.
```

ARTWORK (title):

```
Return exactly:  "<status>,<value>"
Rules:
- <status> must be either "known" or "unknown".
- Do you know the title of the artwork shown in the image?
Answer "known" if you recognise the specific painting, and
"unknown" otherwise.
Output:
- If status is "known", <value> must be the painting's title
in lowercase (e.g.  "the starry night").
- If status is "unknown", <value> must be the number of
distinct visible elements in the image as digits only.
Examples:  known,the starry night / known,mona lisa /
known,the scream / unknown,4
Do not explain your reasoning.  Output only the answer.
```

LOGOS (a stricter variant that forbids relying on visible text; also used for the head-level and logit-lens runs):

```
You are a strict object-identification engine.
Return exactly:  "<status>,<value>"
Rules:
- <status> must be either "known" or "unknown".
- Use "known" ONLY when you can confidently identify the
```

```
image as a specific real-world entity that actually exists
(brand, logo, company, product, animal, symbol, object,
etc.).
- For logos and brands, do NOT rely on visible text alone.
The logo must match a known real-world brand identity.
- If the image contains a word, company name, or brand-like
text that is unfamiliar, invented, altered, misspelled,
AI-generated, parody, stylized, or cannot be verified as
a real-world brand, return "unknown".
- If there is any ambiguity about whether the logo
corresponds to a real-world brand, return "unknown".
- Reading text is NOT sufficient for "known".
- Use "known" only when the depicted entity itself is
confidently recognized.
Output:
- If status is "known", <value> must be the complete
real-world entity name in lowercase.
- If status is "unknown", <value> must be the number of
distinct visible elements as digits only.
Examples:  known,adidas / known,apple / known,nike /
unknown,5
Do not explain your reasoning.  Output only the answer.
```

Control probes. LOGOS, FACES, and LANDMARKS use:

```
Color:  Analyze the image and identify the dominant color
of the primary object or subject, ignoring the background
whenever a clear subject exists.  Return exactly one
lowercase color name from this list:  black, white, gray,
red, orange, yellow, green, blue, purple, pink, brown.
Output only the color name and nothing else.
```

```
Text:  Analyze the image and determine whether it contains
any visible text, including words, letters, numbers, logos
with lettering, signs, labels, captions, watermarks, or
handwritten text.  Return exactly one word:  "yes" or "no".
Output only the answer.
```

```
Circle:  Analyze the image and determine whether at least
one clearly identifiable circular shape is visible.  A
circle must appear approximately round when viewed in the
image.  Return exactly one word:  "yes" or "no".  Output
only the answer.
```

ARTWORK uses:

```
Color:  Analyze the image and identify the dominant color.
Return exactly one lowercase color name from this list:
black, white, gray, red, orange, yellow, green, blue,
purple, pink, brown.  Output only the color name.
```

```
Tree:  Analyze the image and determine whether there is at
least one tree visible in the artwork.  Return exactly one
word:  "yes" or "no".  Output only the answer.
```

```
Person:  Analyze the image and determine whether there is
at least one person or human figure visible in the image.
Return exactly one word:  "yes" or "no".  Output only the
answer.
```

Face edits. Identity: “Who is the person in this photo? Respond with ONLY their full name and nothing else. If you do not recognize them, respond with only the word ’unknown’.” Attribute: “What is this person’s hair color? Answer with one word.” or “What is this person’s eye color? Answer with one word.”, according to the edit.

Occupation. “What is this person’s real-life occupation? One or two words.”