MATCHING OBJECT OR RELATION? TRACING ABSTRACT REASONING INSIDE VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Language Models (VLMs) excel on visual benchmarks but fail systematically on tasks requiring abstract reasoning. Existing benchmarks document this failure but cannot say *why* it happens or which cognitive capability is missing. We close this gap by adopting the Relational Match-to-Sample (RMTS) paradigm from comparative and developmental psychology and pairing it with a mechanistic analysis of the model's internals. On a parametrically controlled stimulus set evaluated across frontier API models (GPT, Claude, Gemini) and three open-source families (Qwen3.5, Gemma-4, InternVL3), we identify four levers that shift VLMs toward the relational match—capability tier, model scale, the number of objects per scene, and the absence of per-object stimulus noise—together producing a developmental-like trajectory that mirrors the human *relational shift*. Opening up the model, a per-layer representational similarity analysis and a causal mediation analysis reveal that VLM abstract reasoning is implemented by two competing circuits: an early circuit that organises images by their surface object features, and a late circuit that organises them by their abstract relation. When the model answers directly, its choice reflects which of the two circuits dominates at the output. Extending the analysis to ARC-AGI-1, we find that ablating the relational heads identified on RMTS degrades performance more than ablating random heads, suggesting that the relational circuit is recruited beyond our controlled stimuli. We hope this mechanism-level view serves as a step toward understanding how abstract reasoning is implemented in VLMs, beyond what aggregate benchmark scores can reveal.

1 INTRODUCTION

Vision Language Models (VLMs) have advanced rapidly on visual tasks (Yue et al., 2024; 2025; Li et al., 2025c; Zhang et al., 2024; Li et al., 2025b), yet they continue to fail systematically on a kind of *abstract reasoning* (Mitchell et al., 2023; Campbell et al., 2024; Yiu et al., 2024). Existing benchmarks document this failure but conflate it with object recognition, attention, and configuration tracking; the resulting picture is descriptive rather than mechanistic, leaving open both *why* VLMs underperform and which cognitive capability they lack relative to mature human intelligence. A diagnostic that ties behaviour to internal mechanism has been missing.

To close this gap, we adopt the **Relational Match-to-Sample (RMTS)** paradigm from comparative and developmental psychology (Christie & Gentner, 2014; Shivaram et al., 2023; Gentner & Hoyos, 2017), illustrated in Figure 1. Two sample images are introduced with a novel nonsense label (*e.g.*, "these are both *XXX*"), and the model is asked which of two test images is also a *XXX*. The *relational match* preserves the abstract pattern shared by the samples—here, a `small-large-small` size pattern—with entirely new objects; the *object match* reuses the samples' own objects (a blue circle and a red triangle) but breaks the pattern. The question is not which option is objectively correct, but which kind of generalisation the system reaches for. Put differently, RMTS diagnoses where the model grounds a novel, image-agnostic label: in the surface-level identity of the sample objects, or in the abstract relational structure that links them. Adult humans gradually shift toward the relational match—a developmental trajectory known as the *relational shift* (Gentner et al., 2011; Christie & Gentner, 2014; Goswami, 1991)—while young children and non-human apes begin near the object-match end (Christie et al., 2016; Premack, 1983). RMTS thus offers a calibrated yardstick

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.

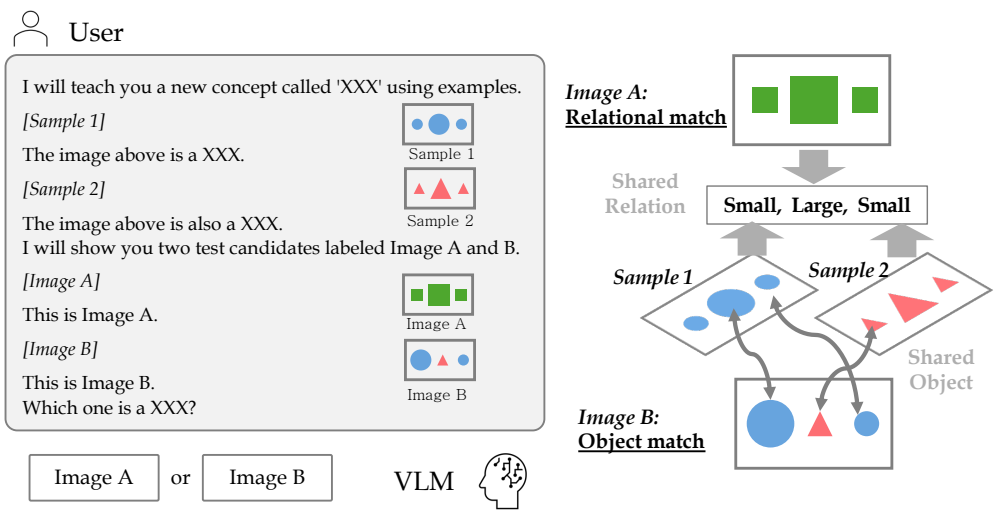


Figure 1: **The Relational Match-to-Sample (RMTS) task as we present it to a VLM. Left:** the prompt as the model sees it. The VLM is taught a novel nonsense label *XXX* from two example images (Sample 1 and Sample 2), and is then asked which of two test candidates—Image A or Image B—is also a *XXX*. The two samples instantiate the same abstract relation—a `small-large-small` size arrangement—but use entirely different object identities (blue circles in Sample 1 vs. red triangles in Sample 2), so the label cannot live in any specific object. **Right:** the two competing solutions the model can latch onto. Image A (top) preserves the `small-large-small` relation with brand-new objects (green squares); the label *XXX* therefore generalises to A only if the model has bound it to the *shared abstract relation* between the samples—the *relational match*. Image B (bottom) reuses the very objects that appeared in the samples (a blue circle and red triangles) but breaks the size pattern; the label generalises to B only if the model has bound it to *shared surface features*—the *object match*. The two readings are mutually exclusive by construction: a relational reasoner picks A, while a system that keys on surface object similarity picks B.

for relational abstraction in any cognitive system, with established human-developmental and cross-species reference points.

We bring this paradigm to VLMs. To eliminate confounds from naturalistic scene content, we design a formally controlled stimulus set whose scenes consist of geometric objects parameterised by `(size, color, shape)`, and evaluate frontier API models (GPT, Claude, Gemini) alongside three open-source families (Qwen3.5, Gemma-4, InternVL3). On this task we make two contributions.

(1) A behavioural map of the object–relational axis in VLMs. (Section 3) We identify four levers that systematically shift VLMs toward the relational match: capability tier (a 58 pp gap inside a single API family between Claude-Opus-4.6 and Claude-Haiku-4.5), within-family scale, the number of objects per scene n, and the absence of per-object stimulus noise. Chain-of-thought (CoT) prompting lifts several non-saturated models by $+20$ to $+57$ pp, although its effect is model-dependent, and the resulting developmental-like trajectory mirrors the human relational shift.

(2) A mechanistic account of how the choice is made. (Section 4) Using a per-layer representational similarity analysis (RSA; Kriegeskorte et al., 2008), we show that early layers organise the four images in line with the object-match solution, while late layers swing toward the relational-match solution, with the crossover arriving earlier in larger models. Together with a causal mediation analysis (CMA; Meng et al., 2022; Wang et al., 2023), this reveals two competing circuits—an early object circuit and a late relational circuit anchored by a few attention heads in Gemma-4, InternVL3 and Qwen2.5-VL—whose balance determines the model's direct answer; CoT instead reaches the relational answer without the relational heads.

We hope that bringing controlled, mechanism-aware experimental designs of this kind to VLMs can serve as a step toward a research programme on abstract reasoning that goes beyond aggregate benchmark scores and engages directly with how that capability is implemented inside the model.

2 THE RELATIONAL MATCH-TO-SAMPLE SETUP

In cognitive science, RMTS is typically conducted with naturalistic scenes containing multiple objects such as animals (Gentner & Hoyos, 2017). For instance, as illustrated in Figure 2, stimuli are constructed by treating entities such as rabbits or penguins as the units over which relational and object matches are defined. In contrast, to probe how VLMs perform abstract reasoning under minimal confounds, we adopt a tightly controlled stimulus design in which the attributes of each object are explicitly parameterised. This design follows a line of work on controlled task construction in deep-learning vision research (Assouel et al., 2025; Park et al., 2024; Okawa et al., 2023). Concrete examples of our stimuli are shown in Figure 1 and, more comprehensively, in Figure 9 in Appendix D.

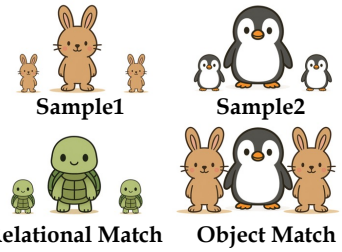


Figure 2: Typical RMTS stimuli in cognitive science.

2.1 TRIAL STRUCTURE: SAMPLES, RELATIONAL MATCH, OBJECT MATCH

We make the structure of an RMTS trial precise. Each scene is built from objects parameterised by three attributes drawn from finite alphabets:

$$\mathcal{C} = \{\text{red}, \text{blue}, \text{green}\}\,, \qquad \mathcal{S} = \{\text{small}, \text{medium}, \text{large}\}\,, \qquad \mathcal{H} = \{\text{circle}, \text{triangle}, \text{square}\}\,,$$

the colour, size, and shape sets respectively. An object is a triple $o = (c, s, h) \in \mathcal{C} \times \mathcal{S} \times \mathcal{H}$.

Definition 1 (Samples)**.** A *sample* of size n is an ordered tuple of n objects that share a common shape[1]:

$$\mathbf{x} = (o_1, \ldots, o_n)\,, \qquad o_i = (c_i, s_i, h) \;\; \forall i\,.$$

The within-scene patterns are the equivalence classes of the colour and size sequences: $P_{\mathcal{C}}(\mathbf{x}) = \text{pat}(c_1, \ldots, c_n)$ and $P_{\mathcal{S}}(\mathbf{x}) = \text{pat}(s_1, \ldots, s_n)$, where for $n = 3$ we use the standard alphabet $\text{pat} \in \{\text{AAA}, \text{AAB}, \text{ABA}, \text{ABB}, \text{ABC}\}$ (e.g. $\text{pat}(\text{small}, \text{large}, \text{small}) = \text{ABA}$). The *relation* of a scene is the pair $R(\mathbf{x}) = \big(P_{\mathcal{C}}(\mathbf{x}),\, P_{\mathcal{S}}(\mathbf{x})\big)$.

A trial is a tuple $T = \big(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}; \mathbf{y}_{\text{rel}}, \mathbf{y}_{\text{obj}};\, \lambda\big)$ consisting of $K \geq 2$ *sample scenes* $\mathbf{x}^{(k)}$ that share the same relation $R^* = R(\mathbf{x}^{(k)})$ but use distinct shapes and distinct attribute-value assignments, two *test scenes* $\mathbf{y}_{\text{rel}}$ and $\mathbf{y}_{\text{obj}}$, and a label λ (a novel three-letter nonce word) attached to each sample. The model is asked which of $\mathbf{y}_{\text{rel}}, \mathbf{y}_{\text{obj}}$ is also a λ.

Definition 2 (Relational match)**.** A scene $\mathbf{y}_{\text{rel}}$ is a *relational match* for samples sharing relation R^* if (i) its relation matches: $R(\mathbf{y}_{\text{rel}}) = R^*$; (ii) its shape is not used by any sample, $h(\mathbf{y}_{\text{rel}}) \notin \{h(\mathbf{x}^{(k)})\}_k$; and (iii) its concrete colour and size values are disjoint from those used by the samples (when the alphabet permits).

Definition 3 (Object match)**.** A scene $\mathbf{y}_{\text{obj}}$ is an *object match* if (i) its relation *differs* from the samples', $R(\mathbf{y}_{\text{obj}}) \neq R^*$; (ii) every object in $\mathbf{y}_{\text{obj}}$ is drawn from the union of objects appearing in the samples, $\{o : o \in \mathbf{y}_{\text{obj}}\} \subseteq \bigcup_k \{o : o \in \mathbf{x}^{(k)}\}$; and (iii) the shapes used in $\mathbf{y}_{\text{obj}}$ are mixed across ≥ 2 of the sample shapes.

The two definitions are mutually antagonistic by construction. Definition 2 guarantees that picking $\mathbf{y}_{\text{rel}}$ requires inferring the abstract relation from the samples and applying it to genuinely new objects (new shape, new colour/size values); Definition 3 guarantees that picking $\mathbf{y}_{\text{obj}}$ requires only matching surface features back to the samples (because every object in $\mathbf{y}_{\text{obj}}$ literally appears in some sample). A system that keys on object similarity therefore deterministically picks $\mathbf{y}_{\text{obj}}$; a system that keys on the relation R^* deterministically picks $\mathbf{y}_{\text{rel}}$. Following the cognitive-science convention (Gentner & Hoyos, 2017; Christie & Gentner, 2014) we report the rate at which the model picks $\mathbf{y}_{\text{rel}}$ (the *relational rate*, Rel%) and treat the contrast as a *bias* rather than an accuracy measurement.

[1] Fixing the shape within a sample preserves the setting illustrated in Figure 2: it lets shape act as a perceptual grouping cue that signals which objects belong to the same scene, while keeping the relation carried entirely by colour and size. Allowing shape to vary within a sample would make the task substantially harder, as the model would also need to identify the grouping itself.

2.2 Prompt and evaluation protocol

Each trial is presented in a single turn containing all four images followed by a short instruction (see concrete prompt in Appendix C). The label is a novel three-letter nonce word drawn from a fixed pool to avoid priming any pre-existing concept; the question is phrased neutrally ("Which one is a XXX?") and constrained with "Final answer: Image " so that the next token is forced to be either A or B. We sweep three open-source families covering 0.8B–38B parameters—Qwen3.5 (Qwen Team, 2026) (0.8B/2B/4B/9B/27B), Gemma-4 (Gemma Team, Google DeepMind, 2026) (E2B/E4B/26B/31B), and InternVL3 (Zhu et al., 2025) (1B/2B/8B/14B/38B)—together with frontier VLM models from OpenAI (GPT-5.4, GPT-5.4-mini) (OpenAI, 2026), Anthropic (Anthropic, 2026) (Claude-Opus-4.6, Claude-Haiku-4.5), and Google (Google DeepMind, 2026) (Gemini-3.1-Pro). For open-source models we read the raw logits of the tokens A and B directly from the next-token distribution, avoiding free-text parsing; for closed-source API models we parse the generated response.

Two sources of bias must be removed for a relational rate to be interpretable. First, the choice of letter labels (A/B) and of nonce word can each induce a constant preference; we cycle the nonce word across a fixed pool of five three-letter strings—XXX, YYY, ZZZ, QQQ, and WWW—assigned uniformly at random per trial, and average the relational rate over labels, integrating out any nonce-specific preference. The pool is restricted to repeated-character strings so that no element carries lexical content that could prime a real concept. Second, the model's prior over the answer position can favour one letter—a hard "A-bias" is observable in some open-source checkpoints. To remove this, every trial is presented in both orderings (relational match as A and as B) and the relational rate is averaged over the two; any constant letter preference therefore cancels. Unless stated otherwise, the relational rate (Rel%) is reported over all trials per model. By default, we use the $n = 3, k = 2$ configuration shown in Figure 1, which yields 729 distinct trials per model; together with the two label orderings used to remove A-bias, this gives 729×2 evaluations per model. Full configuration and model details are deferred to Appendices A and B.

3 Behaviour: When VLMs Pick the Relational Match

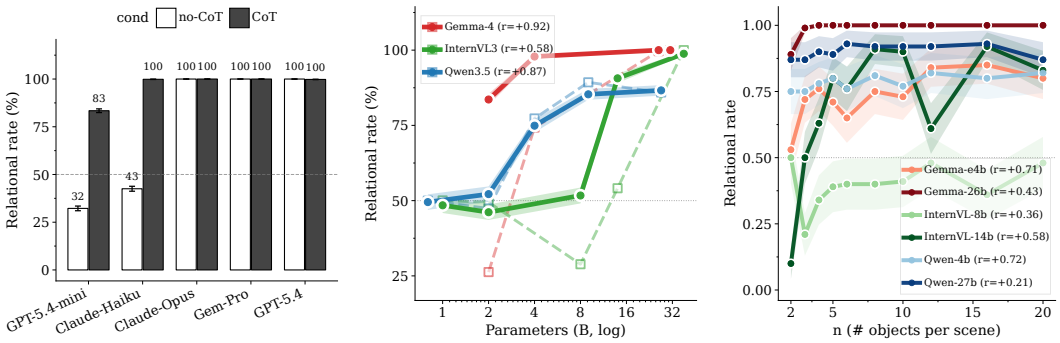
We characterise relational match bias along four levers—(1) capability tier with chain-of-thought (CoT) in Section 3.1, (2) model-size scaling in Section 3.2, (3) number of objects n in Section 3.3, and (4) stimulus noise in Section 3.4. The levers are instances of two factors: the *capability of the model* (tier and scale, with CoT as an inference-time counterpart), the analogue of maturation in the human relational shift; and the *strength of the relational evidence in the stimulus* (n and noise).[2]

3.1 Frontier VLMs show a wide object–relational spread

We begin by measuring RMTS performance across five frontier VLMs—GPT-5.4-mini, Claude-Haiku-4.5, Claude-Opus-4.6, Gemini-3.1-Pro, and GPT-5.4—as shown in Figure 3a. Strikingly, even on this set of frontier models, the no-CoT condition reveals a pronounced spread in relational rate, ranging from 32% for GPT-5.4-mini to 100% for GPT-5.4 (a 68-pp gap). Adding chain-of-thought prompting shifts the two non-saturated models toward the relational match: GPT-5.4-mini and Claude-Haiku-4.5 gain 51 pp ($32 \rightarrow 83\%$) and 57 pp ($43 \rightarrow 99.9\%$) respectively, while the three already-saturated models stay at or near ceiling ($\geq 99.8\%$).[3] Following Kojima et al. (2022), we elicit CoT with a "Let's think step by step" prompt; the exact prompt template is given in Appendix C. Inspecting the generated reasoning traces, we find that the models explicitly extract the abstract relation shared between the two samples and apply it to the test scenes—example traces are provided in Appendix E.

[2] More objects make a shared property overwhelming and each object less individuating; per-object noise degrades exactly the colour/size evidence that defines the relation.

[3] The set is designed as an instrument for analysing *how* a model answers rather than as a benchmark; for closed-source models it does act as one and is at ceiling, but a harder relation (numerical equality) brings them off it (GPT-5.4: 82.5%; Appendix M.1).



(a) Frontier VLMs. (b) Model scale. (c) Number of objects n.

Figure 3: **Capability tier, model scale, and number of objects each promote the relational match.** (a) Capability tier drives a 58 pp spread within Claude (Opus 100% vs. Haiku 42.5%); CoT lifts both non-saturated frontier models (+51 to +57 pp). (b) Within each open-source family, larger checkpoints pick the relational match more reliably (solid lines, filled circles: CoT; dashed lines, open squares: no-CoT). (c) Raising the number of objects per scene from n=2 to n=20 shifts the relational rate upward.

3.2 Model scaling drives the relational match

Within each open-source family, larger models pick the relational match more reliably (Figure 3b; per-model rates in Table 5). Gemma-4 rises monotonically from 26.3% (E2B) to 100% (31B), and Qwen3.5 from 50.0% (0.8B) to 89.3% (9B), with a small but non-negligible dip at 27B (85.9%, −3.4 pp). InternVL3 reaches 100% at 38B; its apparent dip at 8B (28.9%) arises because 1B and 2B sit at chance, committing to neither candidate, whereas 8B commits to the object match before scale flips it.[4] The no-CoT rate correlates with log-model size in Gemma-4 ($r = +0.92$), Qwen3.5 (+0.87) and, more weakly, InternVL3 (+0.58). The effect of CoT is model-dependent: it lifts Gemma-4 and the mid-sized InternVL3 checkpoints substantially (up to +57.3 pp) but slightly lowers several Qwen3.5 and InternVL3 checkpoints (by up to 4.0 pp; Table 5). The aggregate picture is a developmental-like trajectory mirroring the human relational shift (Christie & Gentner, 2014; Gentner & Hoyos, 2017): surface-dominant at small scale, relation-dominant at large scale.

3.3 More objects per scene promote the relational match

Raising the number of objects per scene from n=2 to n=20 shifts the relational rate upward in the object-biased middle tiers of each family (Figure 3c, per-model Pearson $r(n, \text{Rel})$ in the legend; example stimuli in Figure 10). Gemma-4-E4B climbs from 0.53 to 0.80 ($r = +0.71$), InternVL3-14B from 0.10 to 0.83 ($r = +0.58$) and Qwen3.5-4B from 0.75 to 0.82 ($r = +0.72$); InternVL3-8B shows a weaker trend ($r = +0.36$), and the near-ceiling Gemma-4-26B and Qwen3.5-27B show small positive correlations around saturation. As n grows, the abstract relation becomes visually salient, and the surface heuristic that succeeds at n=2 ("the candidate that reuses the samples' objects looks more familiar") no longer suffices. The same trend holds for GPT-5.4-mini, the only non-saturated frontier model in our suite (Figure 13).

3.4 Stimulus noise weakens the relational match

The controlled stimuli used so far are sharp idealisations: every object has a categorical colour (red/blue/green) and a categorical size (small/medium/large), rendered at fixed pixel values. Real-world visual tasks (*e.g.*, natural-image analogues of RMTS) are messier—two "red" objects rarely have identical pixel values, and two "small" objects rarely have identical pixel sizes. To probe how robust the relational behaviour observed above is to this kind of low-level perturbation, we add per-object Gaussian noise to colour and size while leaving the categorical identity of each object

[4] The rank correlation between parameter count and relational rate is positive within every family (Spearman $\rho = 0.80, 1.00, 0.60$ for Qwen3.5, Gemma-4, InternVL3; $\rho = 0.78$ over all fourteen checkpoints), and the scale effect also holds for numerical equality (Appendix M.1).

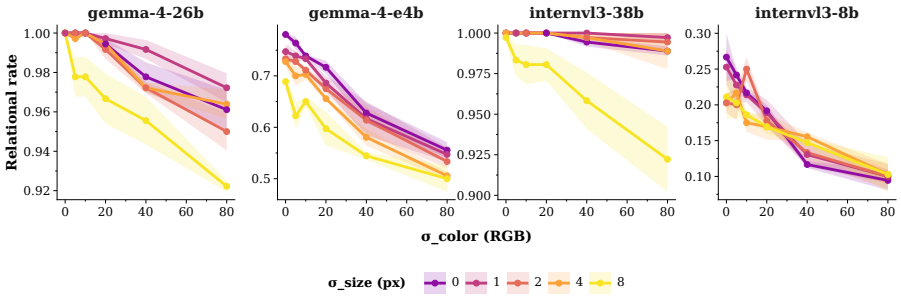


Figure 4: **Per-object stimulus noise reduces the relational rate.** Relational rate as a function of colour-noise scale σ_c (RGB units; x-axis) for four open-source checkpoints, with size-noise scale σ_s (px) encoded by line colour. Both noise channels degrade the relational rate, and within each family the smaller checkpoint degrades more steeply. Bands: ± 1 s.e. over three noise seeds.

unchanged. Concretely, for each object o in a scene we apply

$$\mathrm{rgb}_o \leftarrow \mathrm{rgb}_o + \mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}_3) \quad \text{and} \quad \mathrm{size}_o \leftarrow \mathrm{size}_o + \mathcal{N}(0, \sigma_s^2), \tag{1}$$

clipped to valid pixel and size ranges. We sweep $\sigma_c \in \{0, 5, 10, 20, 40, 80\}$ (RGB units) and $\sigma_s \in \{0, 1, 2, 4, 8\}$ (px) with three independent noise seeds per cell. Examples of the resulting stimuli across the noise sweep are shown in Figure 11.

Figure 4 shows the results for four open-source checkpoints (Gemma-4-26B/E4B, InternVL3-38B/8B). Two patterns are immediately visible. First, both noise channels degrade the relational rate: every panel slopes downward as σ_c grows, and within each panel the curves separate by σ_s (with the largest $\sigma_s = 8$ uniformly the lowest). Second, smaller models within each family are markedly more sensitive to noise. Gemma-4-26B drops only from $\sim 100\%$ (clean) to $\sim 92\%$ under the most extreme perturbation, whereas Gemma-4-E4B falls from $\sim 75\%$ to $\sim 50\%$ over the same sweep—a 25-pp degradation. The InternVL3 family shows the same asymmetry: InternVL3-38B loses only 8–9 pp end-to-end, while InternVL3-8B falls from $\sim 25\%$ to $\sim 10\%$.

Taken together, these results show that relational abstraction is sensitive to low-level visual perturbations even when the categorical identity of every object (red/blue/green; small/medium/large) is preserved. Whether this is because relations are computed over the rendered pixels or because the model's internal categorical representations become noisier remains open; internally, the same perturbation lowers the relational selectivity of late layers (Appendix M.4). See Appendix I for per-attribute differences and results on additional models.

4 Internal Mechanism: Two Competing Circuits

The behavioural results establish *that* VLMs sit at varying positions on the object–relational axis and that several levers shift them, but not *how* the choice is made internally. Our central finding is that it is not made by a single, monolithic "relational module" that grows stronger with scale. Rather, two distinct internal circuits[5]—one keying on object identity, one on the abstract relation—are simultaneously active, and the model's direct answer reflects which circuit dominates. We trace the layer-wise dynamics of the two circuits (Section 4.1), show that the levers that shifted behaviour reshape them in the predicted direction (Section 4.2), and confirm causally that they drive the choice (Section 4.3).

4.1 The two circuits compete across layer depth

For every layer ℓ, we extract the residual-stream activation at the last token of each image, yielding four hidden vectors per trial: $h^{(\ell)}_{s_1}, h^{(\ell)}_{s_2}, h^{(\ell)}_{\mathrm{rel}}, h^{(\ell)}_{\mathrm{obj}} \in \mathbb{R}^d$. We assemble them into a 4×4

[5]We use *circuit* in a functional sense: layers and heads with a distinguishable representational signature (organising the images by object features or by relation) and a measurable causal effect on the choice under patching and ablation—not a complete, reverse-engineered computational subgraph (Elhage et al., 2021; Wang et al., 2023).

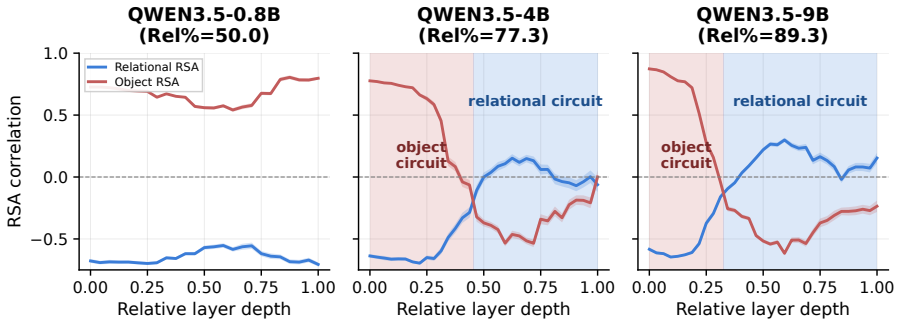


Figure 5: **Layer-wise emergence of the relational circuit.** Per-layer Rel-RSA(ℓ) (blue) and Obj-RSA(ℓ) (red) for three Qwen3.5 checkpoints, plotted against relative layer depth. The smallest model (0.8B; Rel% $=$ 50.0) never produces a relational circuit. The two larger models (4B, 9B; Rel% $=$ 77.3, 89.3) exhibit an early *object circuit* followed by a *relational circuit* that overtakes it at intermediate depth. The crossover shifts earlier with scale. Shaded bands: ± 1 s.e. over trials.

image–image cosine similarity matrix (Kriegeskorte et al., 2008; Assouel et al., 2025), $\mathrm{RSM}_{ij}^{(\ell)} = \cos\big(h_i^{(\ell)}, h_j^{(\ell)}\big)$ $(i, j \in \{s_1, s_2, \mathrm{rel}, \mathrm{obj}\})$, and compare it against two binary 4×4 *target* matrices encoding the two competing solutions. T_{rel} pulls the relational match close to both samples (same abstract pattern); T_{obj} pulls the object match close to both samples (same concrete objects):

$$T_{\mathrm{rel}} = \left(\begin{array}{c|cccc} & s_1 & s_2 & \mathrm{rel} & \mathrm{obj} \\ \hline s_1 & 1 & 0 & 1 & 0 \\ s_2 & 0 & 1 & 1 & 0 \\ \mathrm{rel} & 1 & 1 & 1 & 0 \\ \mathrm{obj} & 0 & 0 & 0 & 1 \end{array}\right), \qquad T_{\mathrm{obj}} = \left(\begin{array}{c|cccc} & s_1 & s_2 & \mathrm{rel} & \mathrm{obj} \\ \hline s_1 & 1 & 0 & 0 & 1 \\ s_2 & 0 & 1 & 0 & 1 \\ \mathrm{rel} & 0 & 0 & 1 & 0 \\ \mathrm{obj} & 1 & 1 & 0 & 1 \end{array}\right).$$

Note that T_{rel} leaves (s_1, s_2) at 0: our hypothesis concerns whether each sample sits close to the relational candidate, not how similar the two samples are to each other, and asserting the latter would reward generic image similarity unrelated to the relation.[6] We score each layer by the Pearson correlation between the upper triangle of $\mathrm{RSM}^{(\ell)}$ and that of each target:

$$\text{Rel-RSA}(\ell) = r\big(\mathrm{RSM}^{(\ell)}, T_{\mathrm{rel}}\big), \qquad \text{Obj-RSA}(\ell) = r\big(\mathrm{RSM}^{(\ell)}, T_{\mathrm{obj}}\big).$$

High Rel-RSA at layer ℓ means the model is organising the four images by their abstract pattern at that depth; **High Obj-RSA** means it is keying on surface features. The two scores are not redundant—a layer can score positively on neither, one, or both—and the layer-wise trajectory of $(\text{Rel-RSA}(\ell), \text{Obj-RSA}(\ell))$ reveals which solution dominates at which depth.

Figure 5 shows the layer-wise trajectories of Rel-RSA(ℓ) and Obj-RSA(ℓ) for three Qwen3.5 checkpoints spanning the behavioural object–relational axis: Qwen3.5-0.8B (Rel% $=$ 50.0), Qwen3.5-4B (77.3), and Qwen3.5-9B (89.3). For Qwen3.5-0.8B, Obj-RSA(ℓ) stays high ($\sim$ 0.6–0.8) and Rel-RSA(ℓ) negative (~ -0.6) at every depth, matching its chance-level Rel%. For Qwen3.5-4B and Qwen3.5-9B, the picture is qualitatively different: Obj-RSA(ℓ) dominates the early layers (relative depth $\lesssim$ 0.3–0.4), and Rel-RSA(ℓ) rises to overtake it at intermediate layers and remains the dominant signal through the rest of the network. We refer to these two regimes as the *object circuit* (early, red shading in Figure 5) and the *relational circuit* (late, blue shading). The crossover depth shifts earlier with scale ($\sim$0.45 in 4B, $\sim$0.30 in 9B), and the asymptotic Rel-RSA grows with scale, mirroring the behavioural ordering (50.0 $\rightarrow$ 77.3 $\rightarrow$ 89.3). Surface-feature matching is thus not absent in models that answer relationally; it is overwritten by the relational circuit at intermediate depth.

4.2 The relational circuit is amplified by the number of objects

In Section 3.3 we showed that increasing the number of objects per scene n shifts the behavioral relational rate upward. We now ask whether this behavioral shift is mirrored in the internal RSA

[6] A variant target that also clusters the two samples together yields layer-wise Rel$-$Obj curves that correlate with ours at $r = 0.91$–1.00 for all fourteen open checkpoints (Appendix M.5).

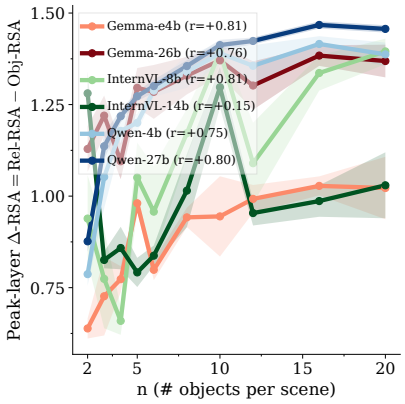


Figure 6: **The relational circuit strengthens as the number of objects per scene grows.** Peak-layer gap Δ-RSA $= \max_\ell[\text{Rel-RSA}(\ell) - \text{Obj-RSA}(\ell)]$ as a function of n for six open-source checkpoints (Pearson r with n in the legend; bands: ± 1 s.e.).

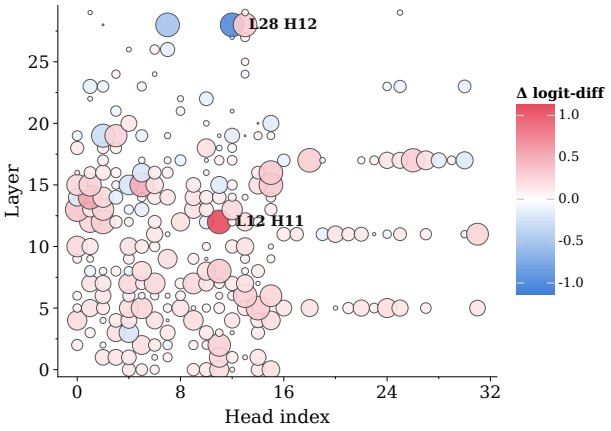


Figure 7: **Per-head activation patching effects in Gemma-4-26B.** $\Delta z = \text{logit}_A - \text{logit}_B$ for each (layer, head) pair: circle size encodes $|\Delta z|$, colour encodes its sign (blue: relational-supporting; red: object-supporting). Two outliers stand out—**L28 H12** (deep, relational) and **L12 H11** (mid, object).

dynamics and whether the two are quantitatively linked. Figure 6 plots, for each of six open-source checkpoints, the peak-layer gap Δ-RSA $= \max_\ell\big[\text{Rel-RSA}(\ell) - \text{Obj-RSA}(\ell)\big]$ as a function of n. The trend is consistent across models: Δ-RSA increases with n, with five of the six checkpoints showing a strong positive Pearson correlation with n (Gemma-4-E4B $r = +0.81$, Gemma-4-26B $r = +0.76$, InternVL3-8B $r = +0.81$, Qwen3.5-4B $r = +0.75$, Qwen3.5-27B $r = +0.80$); InternVL3-14B is the only weak case ($r = +0.15$). The lever that pushed behaviour toward the relational match thus also strengthens the relational circuit in intermediate representations, rather than merely tipping the final answer. Stimulus noise, the lever that weakened behaviour, acts on the same quantity in the opposite direction (Appendix M.4).

4.3 Causal mediation localises the circuits to specific heads

The evidence so far is correlational: Rel-RSA and Obj-RSA identify layers whose representations *align* with the two competing solutions, but they do not establish that those layers are *causally* responsible for the model's choice. To make a causal claim, we follow the mechanistic-interpretability (Bereska & Gavves, 2024; Liu et al., 2025) convention of intervening on activations and measuring the resulting change in behaviour (Meng et al., 2022; Wang et al., 2023; Assouel et al., 2025; Heimersheim & Nanda, 2024; Dumas et al., 2025; Li et al., 2026).

Concretely, we run per-head activation patching on Gemma-4-26B. We form a clean context c_1 (relational match in position A) and a corrupted context c_2 (relational match in position B). For each (layer, head) pair (ℓ, h) we patch the post-o_{proj} activation at the prediction position from c_2 into c_1 and record the resulting shift in the logit margin $\Delta z = \text{logit}_A - \text{logit}_B$. A negative Δz identifies a head that normally carries the relational signal (patching it from a corrupted run flips the model away from the relational answer); a positive Δz identifies a head that carries the object signal.

Figure 7 shows the resulting (layer, head) effect map. The effect is heavy-tailed: most heads have negligible $|\Delta z|$, while a few carry essentially all of the causal signal. The two clear outliers, **L28 H12** (deep, relational-supporting) and **L12 H11** (mid-depth, object-supporting), sit at distinct depths on opposite sides of the sign axis—consistent with the layer-wise RSA in Figure 5, where the object circuit dominates early and the relational circuit takes over late. The two-stage organisation observed in representation space is thus reflected at the level of individual heads. The same organisation appears on Gemma-4-31B (Appendix J) and on five further checkpoints from Gemma-4, InternVL3 and Qwen2.5-VL, where in every case a few heads carry most of the relational effect and the strongest sits in the second half of the network (Appendix M.2).

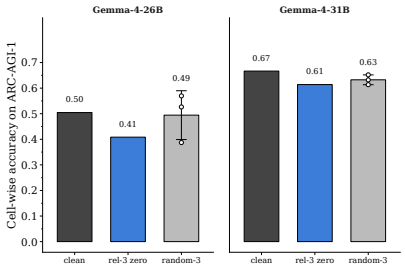


Figure 8: **Ablating relational heads degrades ARC-AGI-1 performance more than random ablation.** Cell-wise accuracy on ARC-AGI-1 for Gemma-4-26B and -31B under *clean*, *rel-3 zero* (top-3 relational heads from CMA), and *random-3* (three random heads) conditions. Targeted ablation drops accuracy more than random in both checkpoints. Error bars on random-3: s.d. across seeds.

CoT bypasses the relational heads. Zeroing the three strongest relational heads of Gemma-4-26B drops its direct answers to chance, yet with CoT the ablated model is back at 100% (Appendix M.3): the competition governs direct answers, while the CoT gains of Section 3.1 run through a separate, language-mediated route.

5 Discussion and Related Work

Summary. Using the RMTS paradigm as a controlled probe of abstract reasoning in VLMs, Section 3 showed how model capability and the strength of relational evidence in the stimulus determine where a model sits on the object–relational axis, and Section 4 showed that the underlying mechanism is not a single "relational module" but two competing circuits—one keying on surface object features, one on the abstract relation—whose relative strength is reflected layer-wise (RSA), tracked by the same levers that shifted behaviour (n and noise), and causally localised to a small number of late-layer attention heads (CMA) in Gemma-4, InternVL3 and Qwen2.5-VL; CoT, in contrast, reaches the relational answer without these heads. Extended related work is in Appendix O.

Mechanistic interpretability of VLMs. Mechanistic interpretability (Bereska & Gavves, 2024; Sharkey et al., 2025; Zhang et al., 2026; Nanda et al., 2023; Olah et al., 2020; Elhage et al., 2021) seeks to reverse-engineer large neural networks at the level of weights, activations, and circuits, originally for LLMs and increasingly for VLMs (Ben Melech Stan et al., 2024; Liu et al., 2025; 2026; Fu et al., 2025; Yang et al., 2025). Recent work has localised several VLM-specific phenomena: information loss already at the vision encoder (Li et al., 2025a), performance gains from strengthened image attention (Chen et al., 2025), and the emergence of binding mechanisms between visual tokens and concept slots (Assouel et al., 2025). What has remained largely uncharted is the mechanism by which a VLM extracts the abstract relation *between* objects, as opposed to the identity of individual objects. Our RSA + CMA pipeline addresses this gap: we show that relation extraction is not a diffuse property of the network but is implemented by a small number of identifiable late-layer heads, sitting downstream of—and competing with—an earlier object-feature circuit. Closest to our work, Lepori et al. (2024) find a perceptual-then-relational decomposition of same/different judgements in vision transformers, mirroring our early object and late relational stages (Appendix O compares the two in detail).

Toward AI capable of abstract reasoning. Recent work has repeatedly identified abstract reasoning as a core weakness of contemporary AI systems, with benchmarks such as ARC-AGI (Chollet, 2019; Chollet et al., 2025) highlighting the gap between human and machine performance. The gap is especially pronounced for VLMs: they underperform on ARC-style tasks (Mitchell et al., 2023), fail tasks that young children solve (Yiu et al., 2024), and struggle with analogical reasoning more broadly (Campbell et al., 2024). By bringing the controlled RMTS paradigm to VLMs and pairing behavioural measurements with internal-state and causal analyses, we provide a more mechanistic account of *which* circuits succeed and fail at relation extraction, and *when*. This account is directly actionable: ablating the top relational heads identified by CMA (Section 4.3) degrades ARC-AGI-1 performance more than random head ablation of the same size (Figure 8; Appendix L). The relational circuit identified on RMTS thus appears to be recruited beyond our controlled stimuli, pending a depth-matched random baseline (Appendix N).

Ethics statement

This work studies publicly released pretrained VLMs and commercial APIs on synthetic geometric stimuli; it involves no human subjects, personal data, or new data collection from people. The work is methodological—a diagnostic for relational reasoning in VLMs—and we see no direct dual-use implications. Identifying specific heads that carry the relational answer could, in principle, be used to induce or suppress particular generalisation patterns at deployment; any such edit should first be evaluated for downstream effects on unrelated tasks. Limitations are discussed in Appendix N.

Reproducibility statement

The stimulus design is fully specified in Section 2 and Appendix A; exact model identifiers, decoding settings, and compute are listed in Appendix B; verbatim prompts are given in Appendix C; the RSA and CMA procedures are defined in Section 4; and the ARC-AGI-1 evaluation protocol, including the head-edit recipe, is described in Appendix L; the protocols of the additional analyses are given in Appendix M. All open-source checkpoints are publicly available, and every experiment uses greedy decoding.

AI use statement

In this work, we used generative AI tools (general-purpose chat LLMs and coding assistants) to help write and edit experiment and analysis code. Additionally, we used them to draft and edit text for readability and to format the manuscript and references. Generative AI tools were not used as a component of the core methodology, and the models evaluated in this paper are the objects of study rather than research assistants. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Anthropic. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, February 2026. Accessed: 2026-05-07.

Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. *arXiv preprint arXiv:2506.15871*, 2025.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. LVLM-Interpret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8182–8187, 2024.

Leonard Bereska and Stratis Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6. Survey Certification, Expert Certification.

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

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas, 2025. URL https://arxiv.org/abs/2503.01773.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Stella Christie and Dedre Gentner. Language helps children succeed on a classic analogy task. *Cognitive Science*, 38(2):383–397, 2014. doi: 10.1111/cogs.12099.

Stella Christie, Dedre Gentner, Josep Call, and Daniel Benjamin Moritz Haun. Sensitivity to relational similarity and object similarity in apes and children. *Current Biology*, 26(4):531–535, 2016. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2015.12.054. URL https://www.sciencedirect.com/science/article/pii/S0960982215015857.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in Transformers. 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. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=qQb1JLrwol.

Gemma Team, Google DeepMind. Gemma 4. https://deepmind.google/models/gemma/gemma-4/, 2026. Accessed: 2026-05-07.

Dedre Gentner and Christian Hoyos. Analogy and abstraction. *Topics in Cognitive Science*, 9(3):672–693, 2017. doi: https://doi.org/10.1111/tops.12278. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12278.

Dedre Gentner, Sarah Brem, Ron Ferguson, and Phillip Wolff. Analogy and creativity in the works of Johannes Kepler. *Creative thought: An investigation of conceptual structures and processes*, January 1997. doi: 10.1037/10227-016.

Dedre Gentner, Florencia K. Anggoro, and Raquel S. Klibanoff. Structure mapping and relational language support children's learning of relational categories. *Child development*, 82 4:1173–88, 2011. URL https://api.semanticscholar.org/CorpusID:440312.

A.K. Goel. Design, analogy, and creativity. *IEEE Expert*, 12(3):62–70, 1997. doi: 10.1109/64.590078.

Google DeepMind. Gemini 3.1 Pro. https://deepmind.google/models/gemini/pro/, 2026. Accessed: 2026-05-07.

Usha Goswami. Analogical reasoning: What develops? a review of research and theory. *Child Development*, 62(1):1–22, 1991. doi: https://doi.org/10.1111/j.1467-8624.1991.tb01511.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8624.1991.tb01511.x.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, November 2008. doi: 10.3389/neuro.06.004.2008.

W. H. Leatherdale. *The Role of Analogy, Model, and Metaphor in Science*. American Elsevier Pub. Co., New York, 1974. URL https://philpapers.org/rec/LEATRO.

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

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulic, and Anders Søgaard. Lost in embeddings: Information loss in vision-language models. *arXiv preprint arXiv:2509.11986*, 2025a.

Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. Visual large language models for generalized and specialized applications, 2025b. URL https://arxiv.org/abs/2501.02765.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1587–1606, 2025c.

Yiming Liu, Yuhui Zhang, and Serena Yeung-Levy. Mechanistic interpretability meets vision language models: Insights and limitations. In *ICLR Blogposts 2025*, 2025. URL https://d2jud02ci9yv69.cloudfront.net/2025-04-28-vlm-understanding-29/blog/vlm-understanding/. https://d2jud02ci9yv69.cloudfront.net/2025-04-28-vlm-understanding-29/blog/vlm-understanding/.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JAI7afWA9e.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=6at6rB3IZm.

Ziming Liu, Eric J Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv:2303.11873*, 2023.

Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Bridging lottery ticket and grokking: Is weight norm sufficient to explain delayed generalization? *arXiv preprint arXiv:2310.19470*, 2023.

Gouki Minegishi, Jingyuan Feng, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Emergent analogical reasoning in Transformers. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.01992.

Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50173–50195. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d0f188c7947eacb0c07f709576824f6-Paper-Conference.pdf.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

OpenAI. GPT-5.4 Thinking System Card. Technical report, OpenAI, March 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf.

Core Francisco Park, Maya Okawa, Andrew Lee, Hidenori Tanaka, and Ekdeep Singh Lubana. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 84698–84729. Curran Associates, Inc., 2024. doi: 10.52202/079017-2689. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/99e6bcf460ea36818cf236da29311e73-Paper-Conference.pdf.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

David Premack. The codes of man and beasts. *Behavioral and Brain Sciences*, 6(1):125–136, 1983. doi: 10.1017/S0140525X00015077.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Apoorva Shivaram, Ruxue Shao, Nina Simms, Susan Hespos, and Dedre Gentner. When do children pass the Relational-Match-To-Sample task? In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*, 2023.

Paul Thagard. Analogy, explanation, and education. *Journal of Research in Science Teaching*, 29 (6):537–544, 1992. doi: 10.1002/tea.3660290603.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

Jiabing Yang, Chenhang Cui, Yiyang Zhou, Yixiang Chen, Peng Xia, Ying Wei, Tao Yu, Yan Huang, and Liang Wang. IKOD: Mitigating visual attention degradation in large vision-language models, 2025. URL https://arxiv.org/abs/2508.03469.

Eunice Yiu, Maan Qraitem, Anisa Noor Majhi, Charlie Wong, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. KiVA: Kid-inspired visual analogies for testing large multimodal models. *arXiv preprint arXiv:2407.17773*, 2024.

Xiang Yue et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Xiang Yue et al. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Qibo Xue, et al. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10317–10362, 2026.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024. URL https://arxiv.org/abs/2302.00923.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A TRIAL COUNTS AND RELATION COVERAGE

A trial is one RMTS instance: two sample images, a relational candidate, and an object candidate. Each image contains n objects. Within an image, the n objects vary along two attributes—colour and size—and each attribute follows one of five abstract patterns: AAA (all the same), AAB (last different), ABA (outer pair shared, middle different), ABB (first different), or ABC (all different). A relation is therefore one pair of patterns $(P_{\text{color}}, P_{\text{size}})$, drawn from the resulting $5 \times 5 = 25$ relation types.

$n = 3$. This is the canonical setting and the basis of every headline number in the paper. Three positions are enough to instantiate all five patterns unambiguously, so a single trial fixes one of the 25 relations. We then enumerate every distinct value assignment that realises each relation— e.g. ABA colour with $A =$ red, $B =$ blue is a different trial from $A =$ blue, $B =$ green—and obtain **729 unique trials** in total. The trials grade naturally by complexity: 9 of them have $P_{\text{color}}{=}P_{\text{size}}{=}$AAA (call this $\text{cpx}{=}0$, both patterns trivial), 144 have one non-trivial attribute ($\text{cpx}{=}1$), and the remaining 576 have both attributes non-trivial ($\text{cpx}{=}2$). Each trial is run twice (relational match in position A and in position B) so the per-model evaluation totals $729 \times 2 = 1458$ datapoints.

Larger n. To probe the effect of visual saliency we re-render the same 25-relation design at $n \in \{2, 3, 4, 5, 6, 8, 10, 12, 16, 20\}$, drawing two value assignments per relation at every n—i.e. $25 \times 2 = 50$ trials per n, and 500 trials in total across the ten n values. At $n{=}3$ a trial in this sweep is structurally identical to a canonical $\text{cpx}{=}2$ trial; at $n \neq 3$ each pattern is extended deterministically: AAA $\rightarrow$ all n positions share one value; AAB $\rightarrow$ only the last position differs; ABB $\rightarrow$ only the first position differs; ABA $\rightarrow$ the two outer positions share one value and all interior positions take a second value; and ABC $\rightarrow$ the n positions are split into three contiguous thirds, each taking a different value. Because each attribute has only three values, "all different" cannot be realised for $n > 3$, so the thirds construction is the natural generalisation that keeps all three values present. The relation pool and the per-relation trial count are held fixed across n, so the only thing that varies along the n-axis is the number of objects per scene—any change in the relational rate therefore reflects scene composition rather than a change in sample size.

B MODEL DETAILS

Decoding. All evaluations use *greedy* decoding (`temperature`=0.0, `do_sample`=False) with `max_tokens`=4096 (no-CoT and CoT). All Hugging Face IDs below are loaded with `trust_remote_code=True` and run in `bfloat16` (vLLM where supported, HF Transformers otherwise).

API checkpoints. The API model identifiers below are used verbatim with the official provider SDKs.

Compute. All open-source experiments were run with NVIDIA H200 nodes (one GPU per job, 80–320 GB host RAM), with the runtime stack containerised in a Singularity image (PyTorch 2.4, Transformers 4.45, vLLM 0.19.1). API-model experiments require no local GPU.

Table 1: **Open-source checkpoints.** HF identifiers as released by the model authors at the time of evaluation.

Tag (used in figures)	Hugging Face identifier
Qwen3.5 (5 sizes)	
qwen3.5-0.8b	Qwen/Qwen3.5-0.8B
qwen3.5-2b	Qwen/Qwen3.5-2B
qwen3.5-4b	Qwen/Qwen3.5-4B
qwen3.5-9b	Qwen/Qwen3.5-9B
qwen3.5-27b	Qwen/Qwen3.5-27B
Gemma-4 (4 sizes)	
gemma-4-e2b	google/gemma-4-E2B-it
gemma-4-e4b	google/gemma-4-E4B-it
gemma-4-26b	google/gemma-4-26B-A4B-it
gemma-4-31b	google/gemma-4-31B-it
InternVL3 (5 sizes)	
internvl3-1b	OpenGVLab/InternVL3-1B-Instruct
internvl3-2b	OpenGVLab/InternVL3-2B-Instruct
internvl3-8b	OpenGVLab/InternVL3-8B-Instruct
internvl3-14b	OpenGVLab/InternVL3-14B-Instruct
internvl3-38b	OpenGVLab/InternVL3-38B-Instruct
Qwen2.5-VL (used only in Appendix M.2)	
qwen2.5-vl-7b	Qwen/Qwen2.5-VL-7B-Instruct

Table 2: **API checkpoints.**

Tag	Provider	Model identifier
gpt-5.4	OpenAI	`gpt-5.4`
gpt-5.4-mini	OpenAI	`gpt-5.4-mini`
claude-opus-4.6	Anthropic	`claude-opus-4-6`
claude-sonnet-4.6	Anthropic	`claude-sonnet-4-6`
claude-haiku-4.5	Anthropic	`claude-haiku-4-5-20251001`
gemini-3.1-pro	Google	`gemini-3.1-pro-preview`
gemini-3.1-flash-lite	Google	`gemini-3.1-flash-lite-preview`

C PROMPT TEMPLATES

No-CoT prompt. The model is explicitly instructed to suppress reasoning and emit only the final answer line.

```
I will teach you a new concept called 'XXX' using examples.

The first image is a XXX.
The second image is also a XXX.

Now look at the third image (Image A) and the fourth image (Image B).
Which one is a XXX?
The answer is equally likely to be Image A or Image B. Do NOT default to
    A
- examine both images carefully. Respond with ONLY one of these two lines
(no reasoning, no explanation):
Final answer: Image A
Final answer: Image B
```

CoT prompt. The model is invited to reason step by step before committing to an answer.

```
I will teach you a new concept called 'XXX' using examples.

The first image is a XXX.
The second image is also a XXX.

Now look at the third image (Image A) and the fourth image (Image B).
Which one is a XXX?
The answer is equally likely to be Image A or Image B. Do NOT default to
    A
- examine both images carefully. Let's think step by step. After your
reasoning, conclude with exactly one of these two lines:
Final answer: Image A
Final answer: Image B
```

D EXAMPLES OF RMTS STIMULI

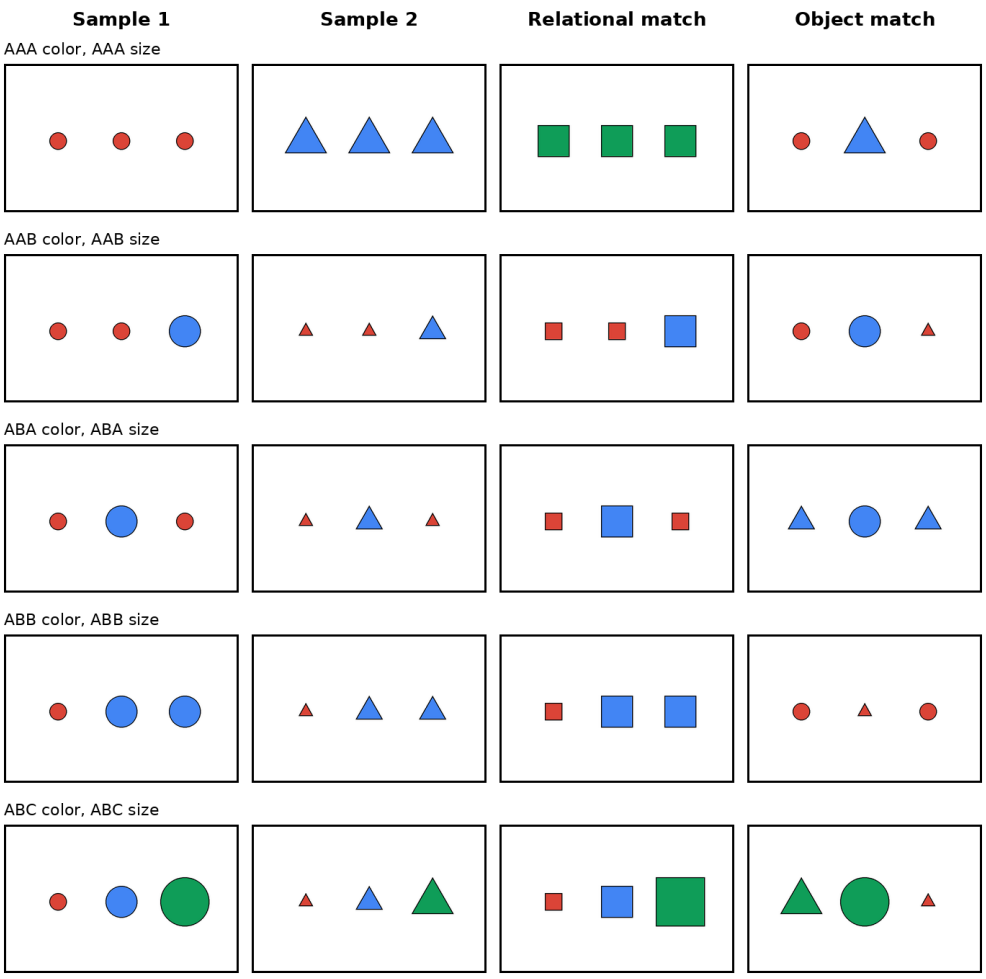


Figure 9: **Canonical stimulus examples** (n=3).

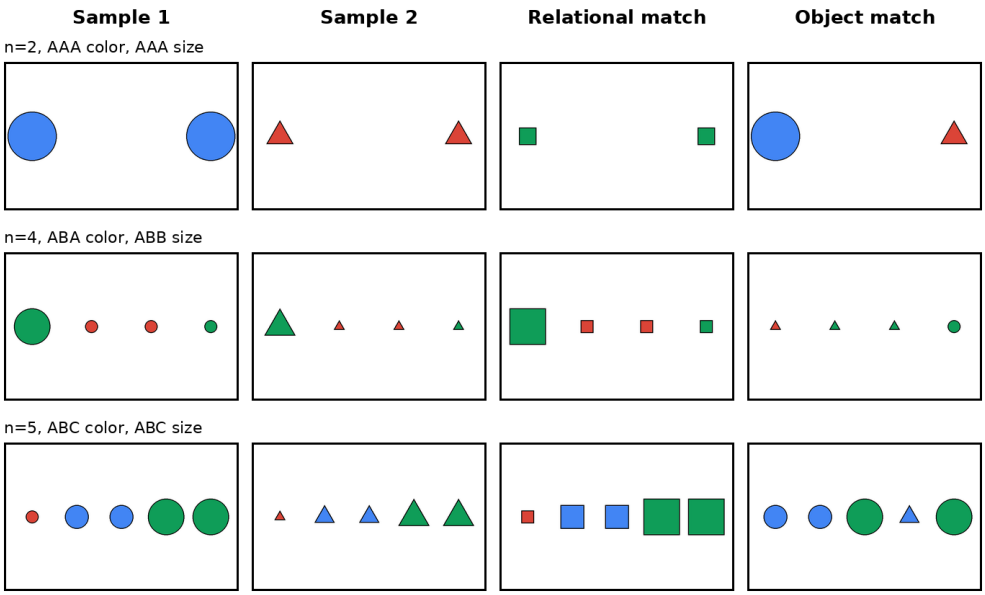


Figure 10: n**-axis stimulus examples.** Same column layout as Figure 9, with the number of objects per scene rising from n=2 (top) to n=5 (bottom); n=3 is omitted because the canonical Figure 9 already shows it.

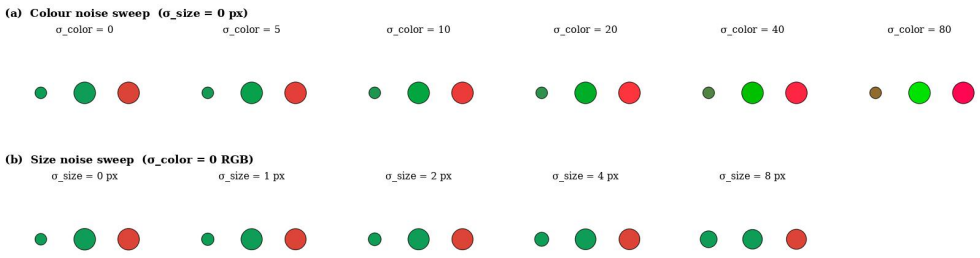


Figure 11: **Per-object noise on a single cpx-2 trial (sample 1 only). (a)** color noise sweep at $\sigma_s = 0$: at $\sigma_c \geq 40$ the categorical color identity becomes visibly ambiguous; at $\sigma_c = 80$ the middle "green" object is closer to lime/yellow. **(b)** size noise sweep at $\sigma_c = 0$: even at $\sigma_s = 8$ the small/medium/large categorical ordering survives, because the categorical sizes are 30/55/85 px and thus at least 25 px apart. This geometric difference is the simplest explanation for the color-vs-size asymmetry reported in Section 3.4.

E CONCRETE COT OUTPUTS

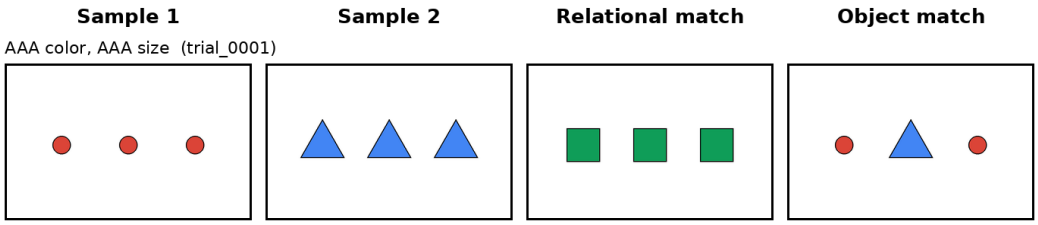


Figure 12: Example RMTS trial used to elicit the chain-of-thought trace shown below. The two top images are the samples sharing the relation R^*; the bottom two are the candidate test scenes (relational match and object match).

Figure 12 shows an example RMTS trial together with the corresponding chain-of-thought trace produced by Claude-Opus-4.6 on this trial. Inspecting the trace, we observe that the model explicitly extracts the pattern shared between the two sample scenes before evaluating the candidates—consistent with the qualitative claim in the main text that CoT prompting surfaces the abstract relation rather than surface-level object features. Because Claude-Opus-4.6 is already at ceiling without CoT, this trace illustrates the form of CoT reasoning rather than the source of the CoT gain.

> "I need to figure out what 'XXX' means from the examples. Example 1: Three red circles arranged in a horizontal row. Example 2: Three blue triangles arranged in a horizontal row. The common pattern: both examples show three identical shapes arranged in a row. The shapes are all the same…" *[selects relational match]*

F COLOR-VS-SIZE ATTRIBUTE ASYMMETRY

The relations in our set are defined over two attributes (color and size). Stratifying by which attribute carries the relation (Table 3) reveals a non-trivial asymmetry under no-CoT. Most open-source families lean *color*: Gemma-4 ($+14.6$ pp), InternVL3 ($+5.6$ pp), Qwen3.5 ($+2.9$ pp). The smaller tiers from Anthropic and OpenAI lean the opposite way: Frontier (Anthropic) -36.1 pp, Frontier (OpenAI) -17.4 pp. The asymmetry is large enough that benchmarks dominated by a single attribute could systematically over- or under-state relational competence; we recommend evaluating on at least two orthogonal attributes whenever possible. Chain-of-thought largely *erases* this asymmetry (right half of Table 3): the smaller Anthropic tier swings from -36.1 pp (no-CoT) to 0.0 pp under CoT, the smaller OpenAI tier from -17.4 to $+4.5$, and Gemma-4 from $+14.6$ to $+0.8$. We read this as a side-effect of the CoT lift documented in Section 3: once CoT pushes every non-saturated model close to ceiling on both attribute strata, the residual asymmetry collapses. A practical implication: benchmarks that report CoT-only relational rates substantially understate the attribute-bias effects we identify under no-CoT.

Table 3: **Color-only vs. size-only relational rate (%), aggregated by family.** "Color-only" = trials with non-trivial color pattern but uniform size; "size-only" is symmetric. For each family the rate is averaged across non-saturated members only (models at 100% on both attributes carry no information about the asymmetry); Frontier (Google) is at ceiling on both for every member under no-CoT, so its Δ collapses to "—". Under no-CoT, families differ in the *direction* of the asymmetry (the smaller Anthropic / OpenAI tiers lean toward size; Gemma-4 leans toward color); under CoT every family's $|\Delta|$ collapses to $\leq$ 7 pp, indicating that explicit reasoning closes the attribute asymmetry.

	No-CoT			CoT		
Family	Color (%)	Size (%)	Δ	Color (%)	Size (%)	Δ
Frontier (OpenAI)	21.5	38.9	-17.4	92.7	88.2	$+4.5$
Frontier (Anthropic)	63.2	99.3	-36.1	98.6	98.6	$+0.0$
Frontier (Google)	100.0	100.0	—	83.1	76.4	$+6.8$
Qwen3.5	72.9	70.0	$+2.9$	82.8	84.1	-1.4
Gemma-4	68.5	53.9	$+14.6$	75.7	74.9	$+0.8$
InternVL3	51.0	45.5	$+5.6$	65.3	66.2	-0.9

G n-AXIS ON FRONTIER API MODELS (COMPANION TO FIGURE 3C)

The main-text n-axis figure (Figure 3c) shows the open-source families to keep the panel readable. The corresponding curves for the frontier API models we evaluated on the n-axis are shown in Figure 13: GPT-5.4 saturates at ceiling for every n, while GPT-5.4-mini exhibits the same competition the rest of the paper documents—chance at n=2 ($\approx$ 0.48), object-biased at n=3 ($\approx$ 0.23, matching its 32% rate on the canonical 729-trial set), and rising back across the chance line as n grows ($\approx$ 0.67 at n=20). The two API curves and the OSS curves in Figure 3c are evaluated on the same n-axis stimuli (Appendix A), which follow the canonical trial design of Section 2.1 at every n (one shape per scene, distinct across scenes; colour and size patterns over the full 5×5 grid; object match drawn from the sample objects), so the n=3 point reproduces the canonical 729-trial rate up to relation-pool sampling. The Anthropic and full Gemini families are omitted from this companion because of API-credit interruptions in the n-axis sweep window; the canonical 729-trial set in Section 3 and the noise sweep in Section 3.4 cover them.

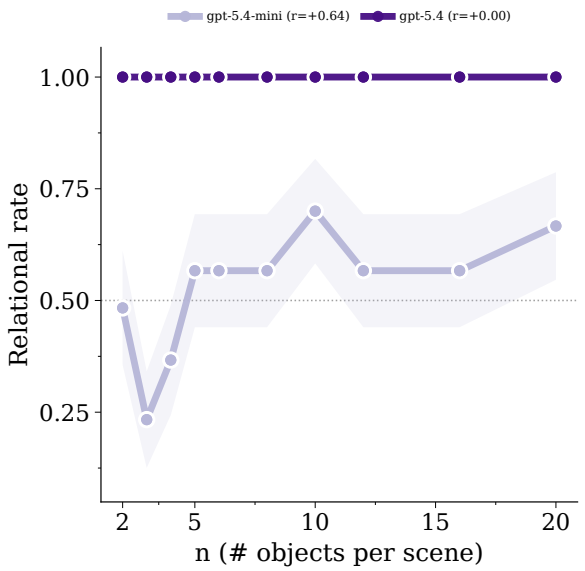


Figure 13: **n-axis on frontier API models (no-CoT, n-axis stimuli of Appendix A).** GPT-5.4 saturates at 1.0 at every n. GPT-5.4-mini drops to $\approx$ 0.23 at n=3—reproducing the 32% canonical rate up to relation-pool sampling—and rises back as n grows. 95% CI bands.

H SHAPE-COHERENCE ABLATION

The confound. The trial structure introduced in Section 2.1 contains a subtle shape-level asymmetry. By definition (Definition 3), the object match mixes shapes from ≥ 2 different samples, while the relational match uses a single fresh shape. The two samples are themselves single-shape (Definition 2). Concretely: the samples are shape-coherent within themselves, the relational match is shape-coherent within itself, but the object match is shape-mixed. A model could in principle exploit this asymmetry alone—"the candidate that is shape-coherent within itself, like the samples, is the answer"—and arrive at the relational match without ever inferring the colour/size relation that defines the trial.

The ablation. To test whether the relational behaviour reported in Section 3 is driven by this shape-coherence cue or by genuine relation extraction, we generated an ablated stimulus set in which the object match is also rebuilt as a single-shape scene, drawn at random per trial from one of the two sample shapes. The relational match is unchanged (it still uses the brand-new third shape). Under this ablation, both candidates are shape-coherent within themselves, so shape coherence no longer disambiguates them; only the colour/size relation does. We refer to this set as *shape-controlled* and to the original set used throughout Section 3 as *canonical*.

Reading the result. If the relational rate on the shape-controlled set matches the rate on the canonical set, the original effect cannot be driven by shape coherence. If instead the rate drops markedly, shape coherence was doing meaningful work and the relational claim should be tempered.

Result. Table 4 summarises the per-model relational rate on the canonical and shape-controlled sets (no-CoT). All ceiling-bound models are essentially unaffected: GPT-5.4 ($100.0 \rightarrow 100.0$), Gemini-3.1-Pro ($100.0 \rightarrow 100.0$), Gemma-4-31B ($100.0 \rightarrow 100.0$), Gemma-4-26B ($99.7 \rightarrow 99.7$, $\Delta = +0.1$ pp), and InternVL3-38B ($100.0 \rightarrow 98.1$, $\Delta = -1.9$ pp) reproduce their canonical-set rates within noise. Qwen3.5-9B, the only non-ceiling model in this comparison, is similarly unaffected ($89.3 \rightarrow 89.6$, $\Delta = +0.3$ pp). All deltas are within ± 2 pp. The relational claim is therefore robust to the shape-coherence confound across the six tested checkpoints.

Table 4: **Shape-coherence ablation: relational rate (no-CoT) on the canonical vs. shape-controlled stimulus set.** On the canonical set, the object match mixes sample shapes; on the shape-controlled set, both candidates are single-shape, so shape coherence cannot disambiguate them. A small $|\Delta|$ means the relational claim is robust to the confound. All six tested checkpoints fall within $|\Delta| \leq 2$ pp.

Model	Canonical	Shape-controlled	Δ
GPT-5.4	100.0	100.0	+0.0
Gemini-3.1-Pro	100.0	100.0	+0.0
Gemma-4-31B	100.0	100.0	+0.0
InternVL3-38B	100.0	98.1	−1.9
Gemma-4-26B	99.7	99.7	+0.1
Qwen3.5-9B	89.3	89.6	+0.3

I NOISE SWEEP: ADDITIONAL MODELS (API AND SMALL QWEN)

The main-text Figure 4 restricts to the four open-source checkpoints that span the clean-rate range (Gemma-4-26B/E4B, InternVL3-38B/8B) so the panel reads cleanly. For completeness we also ran the sweep on two frontier API models (Gemini-3.1-Flash-Lite, GPT-5.4-mini) and the smallest non-A-biased Qwen3.5 (2B); these companion curves are shown in Figure 14. Their qualitative behaviour matches the main figure: clean rate predicts noise robustness, and color perturbation drives a substantially larger drop than size perturbation.

Which attribute matters. Beyond the overall magnitude, *which* attribute the noise sensitivity attaches to also varies across the models in Figure 4. InternVL3-38B, for instance, is almost entirely indifferent to colour noise and to size noise up to $\sigma_s = 4$, only dropping appreciably at $\sigma_s = 8$; its

relational rate stays above 98% across nearly the entire sweep. Other models display the opposite asymmetry, where colour perturbations cost more than comparable size perturbations (cf. Gemma-4-E4B, where the σ_c slope is sharper than the σ_s separation).

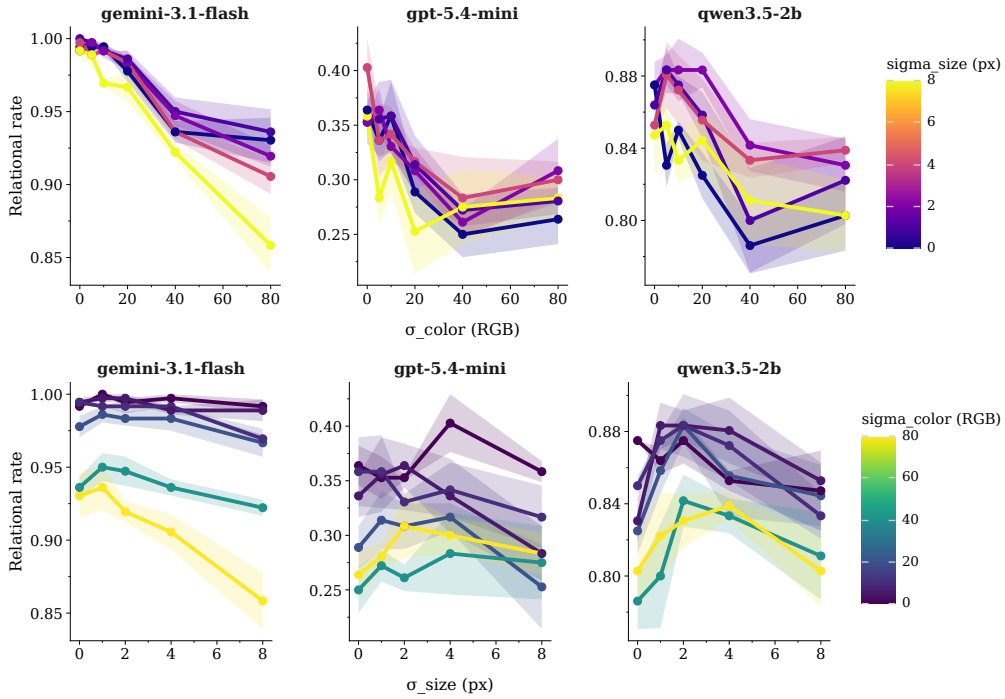


Figure 14: **Noise sweep companions: API frontier models and Qwen3.5-2B.** Top: relational rate vs. σ_c (lines = σ_s). Bottom: same data plotted vs. σ_s (lines = σ_c). SE bands across three seeds. Same σ-grid as Figure 4.

J CMA RESULT ON GEMMA-4-31B

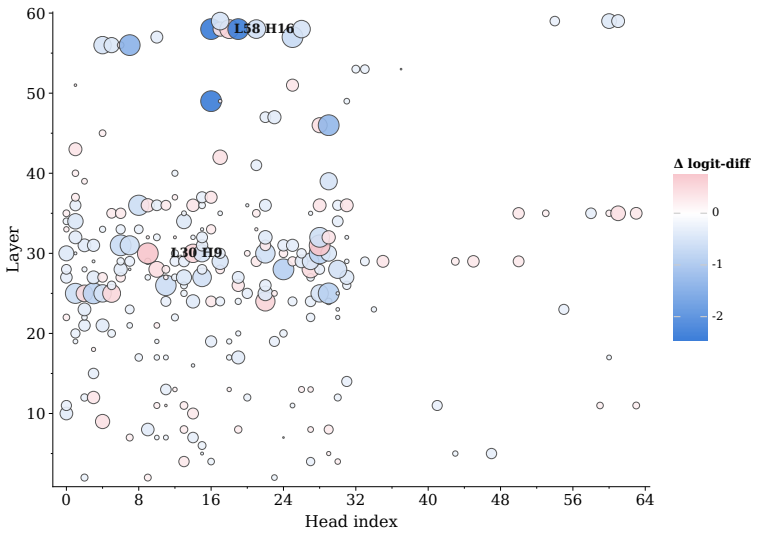


Figure 15: **Per-head activation patching effect on Gemma-4-31B (60 layers, 32/64 heads alternating).** Same encoding as Figure 7: *bubble size* $= |\Delta z|$ capped at the 95th percentile; *color* $=$ signed Δz (blue rel, red object). The dominant rel head sits at **L58 H16** ($\Delta z = -11.7$, an order of magnitude larger than the 26B counterpart and the largest single-head effect in our experiments); dominant object heads cluster at L30–L31. The depth at which the rel head appears scales with total depth ($\sim 95\%$ in both 26B and 31B), so the two-stage organisation is preserved across the family. The wide gap at heads 32–63 on alternating rows again reflects Gemma-4's local/global attention pattern.

K PER-MODEL RELATIONAL RATES

Table 5 reports the per-model no-CoT and CoT rates underlying Figures 3a and 3b. Notable rows: Claude-Haiku-4.5 ($+57.4$ pp; $42.5 \rightarrow 99.9\%$) and GPT-5.4-mini ($+51.2$ pp; $32.2 \rightarrow 83.4\%$) are the two largest gains among the smaller frontier tiers; Gemma-4-E2B ($+57.3$ pp) and Gemma-4-E4B ($+23.6$ pp) show consistent within-family gains for Gemma; Qwen3.5-9B (-4.0 pp), Qwen3.5-4B (-2.4 pp), and InternVL3-2B (-3.3 pp) show a small negative CoT effect.

Table 5: **No-CoT vs. CoT relational rate, full per-model breakdown.** Numbers in percent over parsed (relational+object) trials, excluding parse-failures and UNCLEAR responses; sign of Δ tells whether CoT helps (positive) or hurts (negative). Several Anthropic CoT entries are computed over the parsed subset that survived API failures during the run.

Model	no-CoT (%)	CoT (%)	Δ
GPT-5.4	100.0	99.8	-0.2
GPT-5.4-mini	32.2	83.4	$+51.2$
Claude-Opus-4.6	100.0	100.0	$+0.0$
Claude-Sonnet-4.6	100.0	100.0	$+0.0$
Claude-Haiku-4.5	42.5	99.9	$+57.4$
Gemini-3.1-Pro	100.0	100.0	$+0.0$
Gemini-3.1-Flash-Lite	99.7	99.9	$+0.2$
Qwen3.5-0.8B	50.0	49.5	-0.5
Qwen3.5-2B	47.4	52.2	$+4.8$
Qwen3.5-4B	77.3	74.9	-2.4
Qwen3.5-9B	89.3	85.3	-4.0
Qwen3.5-27B	85.9	86.6	$+0.7$
Gemma-4-E2B	26.3	83.6	$+57.3$
Gemma-4-E4B	74.3	97.9	$+23.6$
Gemma-4-26B	99.7	100.0	$+0.3$
Gemma-4-31B	100.0	100.0	$+0.0$
InternVL3-1B	50.1	48.5	-1.6
InternVL3-2B	49.5	46.1	-3.3
InternVL3-8B	28.9	51.7	$+22.8$
InternVL3-14B	54.1	90.6	$+36.5$
InternVL3-38B	100.0	98.8	-1.2

L ARC-AGI-1 TASK DETAIL

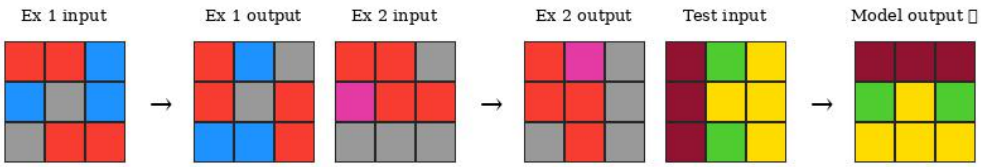


Figure 16: **One ARC-AGI-1 training task as the model sees it.** Two of the four train pairs (input $\rightarrow$ output) and the test input are shown to the model as rendered colour images; the rule here is grid *transposition* (rows become columns).

Figure 8 tests whether the relational heads we identified on RMTS make a measurable contribution on a fully out-of-domain abstract-reasoning benchmark. This appendix gives the exact stimulus, prompt, evaluator, and head-edit recipe.

L.1 DATASET AND TASK SUBSET

We use the public **ARC-AGI-1** corpus (Chollet, 2019) (`training` split; 400 tasks). Each task contains 2–5 train input/output pairs and one test input; the model must emit the test output grid. Two filters are applied before evaluation:

- **Image budget**: the prompt contains $2K+1$ images for K train pairs. We cap at 14 images per prompt (`-limit-mm-per-prompt 14`), which covers $K \leq 6$ pairs.
- **Task cap**: after the image filter we take the first `LIMIT` tasks in canonical filename order (66 for Gemma-4-26B, 80 for Gemma-4-31B; the difference is purely a job-budget choice and has no effect on the comparison since each condition uses the same n within a model).

Identical task subsets are evaluated under every head-edit condition for a given model, so all condition deltas are paired.

L.2 Stimulus rendering

Each $H \times W$ grid is rendered as an RGB image with one 32×32 px cell per integer (0–9) using the canonical ARC palette (0 =black, 1 =blue, 2 =red, 3 =green, 4 =yellow, 5 =grey, 6 =pink, 7 =orange, 8 =light-blue, 9 =maroon), with a 1 px dark border around each cell. The model never sees the integer matrix—it sees the rendered colour image, the same modality used throughout the rest of the paper.

L.3 Prompt format

A single user turn carries an instruction header, the palette key, and an interleaved sequence of (image, shape annotation) blocks for each train pair followed by the test input. Verbatim:

> You are given input/output examples that follow a hidden rule. Each example is shown as a coloured grid; each cell is one integer 0–9. *[palette line]*. Infer the rule from the examples and apply it to the final test input. Be concise and decisive: do NOT second-guess yourself or restate the examples. After your reasoning, output ONLY the final grid for the test input between the literal markers `<grid>` and `</grid>`: one row per line, integers 0–9 separated by spaces, no row numbers, no colour names, no extra text.
> Example 1 input: *[image]* ($H_1 \times W_1$ grid)
> Example 1 output: *[image]* ($H_1' \times W_1'$ grid)
> ...
> Test input: *[image]* ($H_t \times W_t$ grid). Produce the test output grid ... between `<grid>...</grid>`.

We do *not* add the explicit CoT instruction of Appendix C (`-cot` flag off) for the intervention experiments. The prompt nonetheless permits brief reasoning before the grid ("After your reasoning, output ONLY the final grid"), so, unlike the direct-answer RMTS setting, responses may contain short free-text reasoning. We omit the explicit CoT instruction because (a) the rel-3 ablation we want to detect is a *representational* effect that should not require verbal scratchpad to manifest, and (b) CoT lengthens the response, increasing the chance that any head edit becomes visible only via reasoning errors rather than directly via the predicted grid.

L.4 Decoding and parsing

We run the unquantised checkpoint via HuggingFace Transformers (no vLLM in the intervention path, since vLLM does not expose pre-`o_proj` hooks). Decoding is greedy (`do_sample=False`, no temperature) with `max_new_tokens=4096`. The parser:

- First tries the explicit `<grid>...</grid>` block.
- Falls back to the last `Output:` / `Answer:` block.
- For each line, extracts integer tokens; if none are found, tries the colour-name fallback (e.g., "black blue red" $\rightarrow [0, 1, 2]$, including the two-word phrases "light blue" and "dark red").

L.5 Metric: cell-wise accuracy

ARC-AGI-1's official metric is exact-match (whole grid correct or incorrect), but exact-match is too lossy at the capability tier of Gemma-4-26B (clean exact-match ≈ 0.23): every condition crashes to near zero and the rel-vs-random comparison becomes statistically uninformative. We therefore report **cell-wise accuracy**—the fraction of cells in the ground-truth grid matched by the parsed grid, counted as 0 if the parsed shape does not match. This is a natural relaxation that preserves ordering on hard tasks (a one-cell-wrong answer scores $\frac{HW-1}{HW}$, not 0) while still penalising shape errors.

L.6 Head-edit recipe

We use the same per-head intervention machinery as our RMTS CMA scan (Section 4.3): a forward-pre-hook on each layer's `o_proj` reshapes its input to (B, T, n_h, d_h) and multiplies the selected

head slots by the chosen scale. In every condition reported in the main text the scale is 0 (full ablation).

- **Gemma-4-26B (30 layers, 16/32 heads)**: *rel-3 zero*: L28 H12, L28 H7, L19 H2. *obj-3 zero*: L12 H11, L14 H1, L15 H5. *Random-3* (three seeds): each is three (layer, head) pairs sampled reproducibly from the layer-by-head grid excluding the top-5 rel and top-5 object heads, so each random condition zeros out the same number of head slots as the rel/obj conditions.
- **Gemma-4-31B (60 layers, 32/64 heads)**: *rel-3 zero*: L58 H16, L58 H19, L49 H16 (L58 H16 has $\Delta z = -11.7$ in the CMA scan, four-to-five times the runner-up). *obj-3 zero*: L30 H9, L30 H14, L31 H28. Random-3 sampled analogously.

The random-3 baseline controls for the trivial "editing any head hurts" explanation: if rel-3 ablation drops cell-wise accuracy beyond what random-3 ablation does, the rel circuit is contributing causal capacity to ARC, not just disturbing the residual stream.

M ADDITIONAL ANALYSES

This appendix collects analyses that probe the generality of the behavioural and mechanistic findings: a structurally different relation (Appendix M.1), causal mediation on further model families (Appendix M.2), the interaction between CoT and the relational heads (Appendix M.3), the internal effect of stimulus noise (Appendix M.4), and the robustness of the RSA target (Appendix M.5).

M.1 A STRUCTURALLY DIFFERENT RELATION: NUMERICAL EQUALITY

Design. All relations in the main text are ordinal patterns over colour and size. To test whether the findings depend on this relation type, we construct trials in which the two samples share their *object count* and nothing else. With three objects, for example, Sample 1 shows three red squares and Sample 2 three blue triangles, so that count is their only commonality. The relational match shows three green circles—the right count, in a colour and shape used by neither sample—while the object match shows five red squares, Sample 1's objects copied verbatim with the wrong count. A model that matches appearance picks the red squares; only counting reaches the green circles. As in the main experiments, every trial is presented in both answer orderings, so any positional preference cancels (chance $= 50\%$), and no CoT is used.

Scale effect on open-source models. We generate 64 trials (128 judgements per model) and evaluate two checkpoints from each open-source family (Table 6). In every family the relational rate rises with model size, by 46, 30 and 73 pp respectively, so the scale effect of Section 3.2 is robust to changing the relation itself.

Frontier models come off the ceiling. The three frontier models that saturate the ordinal relations are no longer saturated on numerical equality (Table 7). Together with the noise sweep, in which checkpoints that saturate the clean stimuli also fall off the ceiling (Gemma-4-26B $100 \rightarrow 92\%$, InternVL3-38B $100 \rightarrow 92\%$, Gemini-3.1-Flash-Lite $99 \rightarrow 86\%$ under the strongest colour and size jitter; Figure 4 and Appendix I), this indicates that the ceiling on the canonical set partly reflects its simplicity for the strongest models, and that harder relations discriminate among them.

Table 6: **Numerical equality, open-source models.** Relational rate (%, no-CoT), small $\rightarrow$ large checkpoint per family.

Family (small $\rightarrow$ large)	Small	Large
Qwen3.5 (2B $\rightarrow$ 9B)	47.7	93.8
Gemma-4 (E4B $\rightarrow$ 26B)	26.6	57.0
InternVL3 (8B $\rightarrow$ 38B)	25.8	98.4

Table 7: **Numerical equality, frontier models.** Relational rate (%), averaged over $n = 2$–5 and both orderings.

Model	Rel. rate
Gemini-3.1-Pro	99.0
Claude-Opus-4.6	93.0
GPT-5.4	82.5

M.2 CAUSAL MEDIATION ON FURTHER MODEL FAMILIES

To test whether the head-level picture of Section 4.3 is specific to Gemma-4-26B, we run the same per-head activation patching, unchanged, on five further checkpoints covering Gemma-4, InternVL3 and Qwen2.5-VL (30 trials each, both answer orderings). Table 8 reports the strongest relational head per checkpoint. In every family a small number of heads carry most of the relational effect, and in every case the strongest one sits in the second half of the network, as in the Gemma-4-26B result of the main text.

Table 8: **Strongest relational head per checkpoint.** Effect: change in the log-odds of the relational answer when that head alone is patched (negative = relational-supporting, as in Figure 7). Depth: layer index divided by (number of layers -1), so that the last layer has depth 1.

Model	Top relational (layer, head)	Effect	Depth
Gemma-4-26B (main text)	(28, 12)	-1.02	0.97
Gemma-4-E4B	(21, 5)	-1.27	0.51
InternVL3-38B	(55, 23)	-1.36	0.87
InternVL3-14B	(28, 29)	-0.82	0.60
InternVL3-8B	(17, 27)	-0.44	0.63
Qwen2.5-VL-7B	(20, 24)	-0.32	0.74

M.3 COT BYPASSES THE RELATIONAL HEADS

CoT lifts the weaker frontier models and several open-source checkpoints toward the relational match (Section 3.1 and Appendix K). Does it do so by strengthening the relational circuit, or through a different route? On Gemma-4-26B we zero out the relational heads identified by CMA (the single strongest, L28 H12, or the top three, L28 H12, L28 H7 and L19 H2; Appendix L) and run the identical trials twice, once answering directly and once with the CoT prompt of Appendix C. We use 30 trials from the most demanding relation subset in both answer orderings (60 judgements per condition). Table 9 shows that removing the three strongest relational heads reduces direct answers to chance, which also shows that these heads are necessary rather than merely correlated with the behaviour. With CoT, however, the same ablated model is back at 100%. The CoT gain therefore does not run through the circuit we identified.

Table 9: **Relational rate (%) of Gemma-4-26B under head ablation, without and with CoT.**

Heads zeroed	Direct answer	With CoT
None (intact model)	100.0	100.0
Strongest relational head	86.7	100.0
Three strongest relational heads	50.0	100.0

M.4 STIMULUS NOISE AND INTERNAL REPRESENTATIONS

We ask whether stimulus noise, which lowers the relational rate (Section 3.4), acts on the same internal quantity as the number of objects (Section 4.2). We use Gemma-4-E4B, the checkpoint with the largest behavioural noise effect, and apply independent Gaussian jitter to every object's radius, keeping each object inside its original size category so that the categorical description of the scene is unchanged and only the rendered pixels move. We use three noise levels, from no jitter to the strongest jitter of the behavioural sweep, with three noise seeds and 40 trials per seed (120 trials per condition), and measure the relational selectivity $\Delta\text{-RSA} = \text{Rel-RSA} - \text{Obj-RSA}$ used throughout Section 4. Averaged over the late third of the layers (layers 29–42), Δ-RSA falls from $+0.24$ without jitter to $+0.15$ at the strongest jitter (s.e. 0.06), while the relational rate on the same trials falls from 78% to 69%. Noise thus acts on the same internal quantity as n, in the opposite direction, and the internal decline tracks the behavioural one.

M.5 Robustness of the relational RSA target

The relational target T_{rel} (Section 4.1) leaves the sample–sample entry at 0. We also evaluate the alternative in which all relation-sharing images, including the two samples, are clustered together. For each model and each target we compute the layer-wise curve $\text{Rel-RSA}(\ell) - \text{Obj-RSA}(\ell)$ and take the Pearson correlation across layers between the two curves. The two definitions agree at $r = 0.91$–1.00 for all fourteen open checkpoints across the three families, so every layer-wise conclusion in the paper holds under either target.

N Limitations

Shape coherence as a residual cue. Within a trial, every sample scene uses one shape, the relational match introduces a fresh shape, and the object match by construction (Definition 3) mixes ≥ 2 sample shapes; this enforces shape-disjointness with the samples but also means the relational candidate is shape-coherent while the object candidate is shape-mixed. A model could therefore exploit a "which candidate is shape-coherent like the samples" heuristic without inferring the colour/size relation. We address this with a candidate-level ablation in Appendix H: the object match is rebuilt as single-shape (drawn from a sample shape) so both candidates are shape-coherent within themselves, and only the colour/size relation can disambiguate them. Across the six tested checkpoints the relational rate moves by at most $|\Delta| \leq 2$ pp, indicating that the headline effect is not driven by shape coherence.

Stimulus distribution. Our formal geometric stimuli isolate the relational step at the cost of a narrow distribution; we mitigate this by sweeping $n = 2$–20 across the full 5×5 pattern grid (Figure 3c), by adding per-object Gaussian noise (Figure 4), and by testing a structurally different relation, numerical equality (Appendix M.1). These relations still do not exhaust those VLMs will ultimately need; spatial, causal and semantic relations, as well as naturalistic scenes, are an immediate next step.

Ceiling effects. The strongest frontier models are at ceiling on the ordinal colour/size relations. For these closed-source models the stimulus set functions only as a benchmark, and at ceiling it cannot discriminate genuine relational abstraction from simpler strategies; the numerical-equality relation partly addresses this by bringing them off the ceiling (Appendix M.1).

Mechanistic scope. Per-head CMA evidence covers two Gemma-4 checkpoints in full (Section 4.3 and Appendix J) and five further checkpoints from Gemma-4, InternVL3 and Qwen2.5-VL on a smaller trial budget (Appendix M.2). Our use of *circuit* is functional (Section 4); we do not characterise how the identified heads compose. Finally, the circuit account covers direct answers only: under CoT the model reaches the relational answer without the relational heads (Appendix M.3), and the mechanism of that route is left open. Because we patch only at the prediction position, the identified heads may read out or transmit the relational decision rather than perform the comparison between images itself; localising where the relation is computed would require patching at image-token positions.

ARC-AGI-1 baseline. The random-3 baseline (Appendix L) samples heads uniformly across depth, whereas the strongest relational heads sit in the last layers (e.g. L28 of 30 in Gemma-4-26B). Since late-layer heads can have larger effects on the output in general, a depth-matched random baseline is needed to confirm that the ARC-AGI-1 degradation is specific to the relational heads rather than a generic consequence of disrupting late layers. The comparison is also small (66 and 80 tasks, three random-ablation seeds, no significance test) and therefore directional, and because the ARC-AGI-1 prompt permits brief reasoning before the answer (Appendix L), it is not strictly a direct-answer setting, which qualifies its relation to the CoT result of Appendix M.3.

Hidden reasoning in API models. Some frontier API models (e.g. GPT-5.4) may perform internal reasoning that is not exposed even under the no-CoT prompt; for them, "no-CoT" means only that no visible reasoning was requested.

O Extended Related Work

Analogical reasoning. The RMTS task is treated in cognitive science as an instance of *analogical reasoning* (Christie & Gentner, 2014; Gentner & Hoyos, 2017), in which a learner transfers an abstract structural correspondence from a source to a target despite surface-level dissimilarity. Analogy has been studied as a candidate explanation for the remarkable sample efficiency of human learning (Thagard, 1992; Gentner & Hoyos, 2017) and for its role in scientific discovery (Leatherdale, 1974; Goel, 1997; Gentner et al., 1997).

Studies of analogical reasoning in deep neural networks remain comparatively scarce, but several lines of work are directly relevant. Webb et al. (2023) compare large language models to humans on a battery of analogical-reasoning tasks and report broadly human-level behavioural performance. Campbell et al. (2024) document VLM failures on visual analogy benchmarks, in line with the broader picture that vision-grounded analogical inference remains a weakness of current systems. From a more mechanistic angle, Minegishi et al. (2026) formalise analogy in a category-theoretic framework, design a controlled toy task, and analyse how a Transformer trained from scratch on this task implements analogical mappings.

The closest mechanistic study is Lepori et al. (2024), who show that vision transformers decompose a same/different judgement into a perceptual stage followed by a relational stage, close to the early object stage and late relational stage we describe. We differ in three respects: we study generative VLMs on a task where the relational answer must beat a specific object-matching alternative rather than give a binary report; we track that competition across capability tier, scale, CoT, scene size and stimulus noise; and we test the relational stage causally, including on ARC-AGI-1.

These prior efforts are largely either purely behavioural (assessing whether a model gets analogy problems right or wrong) or focused on toy networks trained from scratch on synthetic tasks. To our knowledge, the present work is the first to combine **(i)** a parametrically controlled analogical-reasoning paradigm imported from comparative and developmental psychology, **(ii)** an evaluation across the current frontier of pretrained VLMs (API and open-source families), and **(iii)** a mechanistic dissection that traces how the analogical computation is implemented inside the network at the level of layers and individual attention heads.

Competing internal circuits. The two-circuit picture we observe is reminiscent of similar phenomena documented elsewhere in the mechanistic-interpretability literature. Most strikingly, the grokking phenomenon (Liu et al., 2023; Power et al., 2022) has been explained as a competition between a memorisation circuit and a generalisation circuit, with the latter eventually dominating after extended training (Nanda et al., 2023; Liu et al., 2022; Minegishi et al., 2023; Merrill et al., 2023). Our findings suggest that the same competitive logic surfaces at inference time in pretrained VLMs: a surface-feature (object) circuit and an abstract (relational) circuit are both present, and the model's behavioural output reflects which circuit dominates at the prediction position. We conjecture that large pretrained models harbour many such circuits implementing alternative solutions to the same task, and that the model's response is governed by whichever circuit exerts the largest causal influence on the output logits.