FROM PERCEPTION TO INTEGRATION: REVISITING THE INTERNAL DYNAMICS OF REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can answer simple visual questions, but often struggle when one question requires several visual judgments. We study this gap with controlled tasks for feature binding, numerosity, spatial relations, and amodal completion, together with a Composite task that combines them. Matched counterfactual image pairs isolate changes in the visual evidence needed to answer. Across four models, direct answers, hidden-state readouts, and state interventions show that the individual judgments can be made without explicit reasoning and that intervening on the corresponding states can affect the answer. During reasoning, the Composite answer becomes decodable from hidden states and usable from shortened traces, often before the model stops on its own. We train a small detector to predict this readiness and stop reasoning at that point. On MMStar and RealWorldQA, this reduces mean reasoning tokens by 79.1% and 74.5%, while average accuracy rises by 3.13 and 3.30 percentage points, respectively. These findings connect the internal development of answer readiness to a practical rule for allocating reasoning computation.[1]

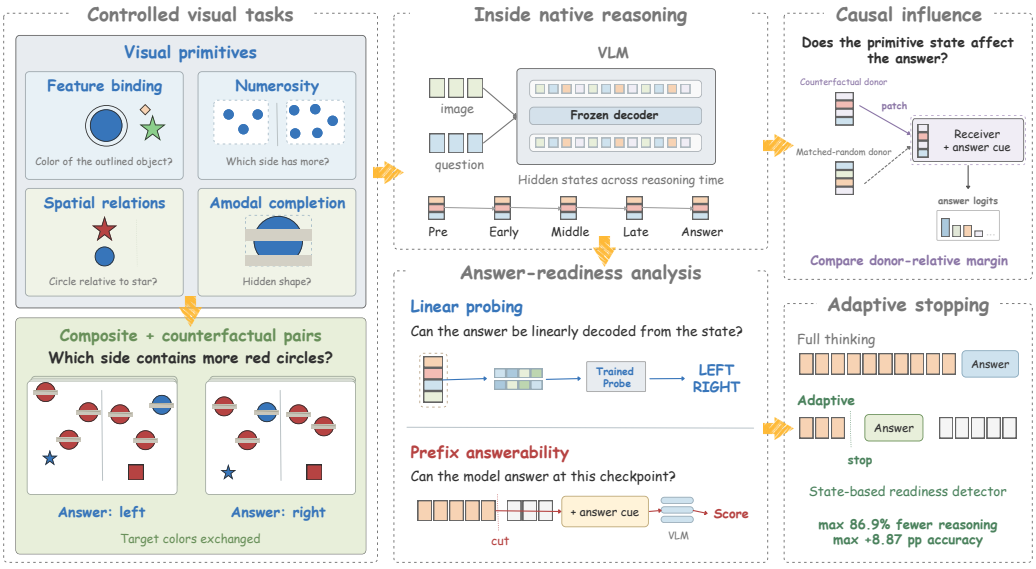

Figure 1: Overview of our study. Controlled visual tasks test individual judgments and their combination in Composite scenes. Linear readouts and counterfactual state interventions examine what visual information is represented and affects the answer. During native reasoning, readouts and answers from shortened traces track when the Composite answer becomes available. A hidden-state readiness detector uses this principle to stop reasoning early on MMStar and RealWorldQA.

[1]Code is available in the supplementary materials.

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.

1 INTRODUCTION

People recognize many objects quickly, but some images require further visual processing. In primates, object information can appear within 150 ms during an initial feedforward pass, while harder images require recurrent processing (Thorpe et al., 1996; DiCarlo et al., 2012; Lamme & Roelfsema, 2000; Kar et al., 2019). VLMs also have a fixed number of layers in each forward pass. Generating more tokens gives them additional steps of computation. This suggests a question: which visual judgments can a VLM make before it starts reasoning, and which improve as reasoning proceeds?

Visual benchmarks show which questions models answer correctly (Johnson et al., 2017; Huang et al., 2025). Mechanistic studies show where visual information appears inside a model (Neo et al., 2024; Zhang et al., 2025b). We ask how the model uses that information to reach an answer (Figure 1). Across four VLMs, accuracy without reasoning is 87.2–91.4% on four separate visual tasks, compared with a 31.3% guess rate. Yet average accuracy falls to 58.6% on a single Composite question with two answer choices, left or right. This question requires the model to combine several visual judgments to select the correct side. High accuracy on the individual tasks therefore does not ensure a correct answer when those judgments must be used together. A correct final answer does not reveal when the model first became able to give it.

We test Feature Binding, Numerosity, Spatial Relations, and Amodal Completion: identifying an object's features, comparing quantities, judging positions, and recognizing a partly hidden shape (Treisman & Gelade, 1980; Feigenson et al., 2004; Logan, 1994; Kellman & Shipley, 1991). Our Composite task combines these demands by asking which side of a partly occluded scene contains more red circles (Figure 2, rightmost column). We also use matched image pairs that change the correct answer while keeping the rest of the scene fixed (Gardner et al., 2020; Bitton et al., 2021). Hidden-state readouts and state interventions test what information the model carries and uses. Answers from shortened reasoning traces show when it can use that information to answer the Composite question.

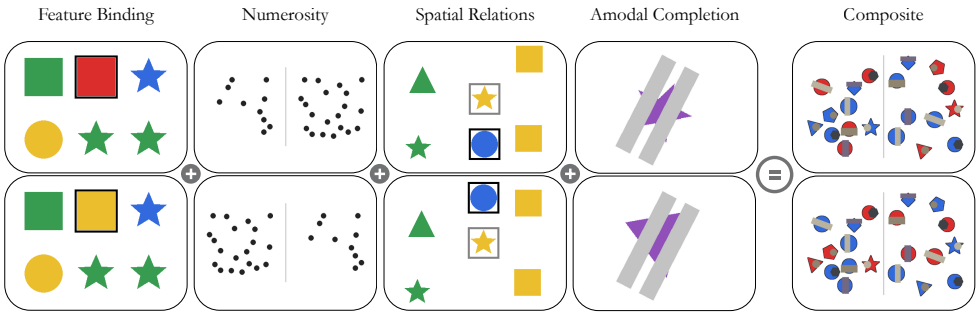

Figure 2: Every column shows a counterfactual pair in which the task-relevant variable and the correct answer change together while the remaining scene content is preserved.

Reasoning raises Composite pair-consistent accuracy by 40.8 points on average. It is also costly: full traces average 1,650 reasoning tokens on MMStar, and 16.6% of Composite runs reach the generation limit without a final answer. We find that a shortened trace can often support a correct answer before the model stops. We call the earliest tested point where this happens *answer readiness* and ask whether the model's state can detect it during generation.

We propose Answer-Readiness-Guided Early Stopping. A linear detector uses the current hidden state and its change since the previous checkpoint to predict whether the model can answer correctly now. If the prediction crosses a threshold selected on validation data, the model stops reasoning and generates its answer. This approach builds on earlier methods that stop when an answer appears correct or sufficient (Zhang et al., 2025a; Yang et al., 2026; Xiang et al., 2026).

Across four VLMs on MMStar and RealWorldQA, stopping shortens mean reasoning by 79.1% and 74.5% while average accuracy rises by 3.13 and 3.30 points.

2 Related Work

Multimodal reasoning and visual evaluation. Controlled benchmarks test how VLMs combine visual information (Johnson et al., 2017; Schiappa et al., 2024; Huang et al., 2025). Multimodal chain-of-thought methods ask models to explain their answers or reason through image regions and objects (Zhang et al., 2023; Shao et al., 2024; Li et al., 2025). Other evaluations find that more reasoning helps some visual questions but not others (Jiang et al., 2025; Jin et al., 2026). These studies measure final answers; they do not show when reasoning first makes an answer possible.

Interpreting visual representations in VLMs. Researchers use causal tracing and token interventions to locate information inside language models and VLMs (Meng et al., 2022; Basu et al., 2024; Neo et al., 2024). Related work studies how visual information moves through direct and text-mediated pathways (Zhang et al., 2025b; Salazar et al., 2026). Layer-wise readouts also separate visual grounding from later processing and answer formatting (Yu & Lee, 2025). These studies focus on a single forward pass. They do not track how the model uses visual information while generating a reasoning trace.

Efficient reasoning and early exit. Hidden-state readouts can predict reasoning success or check intermediate answers (Afzal et al., 2025; Zhang et al., 2025a). Other methods stop when a trial answer has high confidence or appears sufficient (Yang et al., 2026; Xiang et al., 2026). For VLMs, Bi et al. (2026) estimate when reasoning helps and how long it should continue, then intervene on attention heads. These methods do not use a direct measurement of when the visual answer becomes available during reasoning.

Our analysis measures when a Composite answer becomes usable during reasoning. That measurement provides the target for a detector that decides when to stop.

3 Interpreting Composite Visual Reasoning

3.1 Interpretability Analysis Setup

Each task isolates one answer-determining variable while holding the remaining scene content constant.

Feature Binding. Feature Binding asks which color belongs to a marked object. People can notice a color and a shape but sometimes combine them incorrectly (Treisman & Gelade, 1980; Treisman & Schmidt, 1982). Each image contains several colored shapes; a black outline marks the target. The model reports its color.

Numerosity. Numerosity asks which side of a divider has more dots. Even infants can compare such quantities before learning number words (Feigenson et al., 2004; Xu & Spelke, 2000). We vary the relationship between dot count and total dot area, so area alone does not always reveal the answer. The model responds left or right rather than giving an exact count.

Spatial Relations. Spatial Relations asks where a target lies relative to another object. Such judgments require selecting the two relevant objects (Carlson-Radvansky et al., 1999; Logan, 1994). Black and gray outlines mark the target and reference, while their colors and shapes vary. The model chooses left, right, above, or below.

Amodal Completion. Amodal Completion asks what whole shape is partly hidden behind an occluder. People can infer the whole from visible fragments (Kellman & Shipley, 1991; Sekuler & Palmer, 1992). We partly cover a geometric shape and ask the model to choose its complete form. The Composite task needs this judgment for partly hidden objects.

Composite Task. The Composite task asks which side of a divider has more red circles among colored, partly hidden shapes. To answer, the model must recognize each shape, match colors to shapes, assign objects to a side, and compare the two counts. This resembles "challenge images" in

vision research: the parts may be easy to recognize, but combining them can require more processing (Kar et al., 2019; Kar & DiCarlo, 2021).

Counterfactual Pairs. For state interventions, we use matched image pairs that differ in the feature that determines the answer while keeping the rest of the scene fixed (Gardner et al., 2020; Bitton et al., 2021). Feature Binding, Numerosity, and Spatial Relations use separate datasets for direct answers and interventions. Amodal Completion and Composite use paired datasets throughout. Figure 2 shows one pair per task; Appendix A gives the construction rules and splits.

3.2 Visual Judgments Without Reasoning

The four component judgments provide three complementary tests without explicit reasoning: direct-answer accuracy, linear readouts of hidden states, and interventions that measure whether changing those states affects the answer (Belinkov, 2022).

Component judgments are available without reasoning. We freeze each model, disable thinking, and ask it to answer the four tasks directly. All four models perform above the task-specific guess levels (Table 1). They do best on Feature Binding and Spatial Relations and make more errors on Numerosity and Amodal Completion.

Primitive variables are linearly recoverable across depth. Direct answers show that a model can make these judgments, but do not show where it represents them. We therefore average its image-token hidden states at each layer:

$$v_i^{(\ell)} = \frac{1}{|\mathcal{I}_i|} \sum_{t \in \mathcal{I}_i} h_{i,t}^{(\ell)}, \tag{1}$$

where $h_{i,t}^{(\ell)}$ is the representation of sample i at layer ℓ and token position t, and $\mathcal{I}_i$ contains its image-token positions. We standardize these vectors using training-set statistics to obtain $z_i^{(\ell)}$, whose rows form the training matrix Z_ℓ. For each model and task, we fit an independent ridge linear readout at every layer, keeping the VLM frozen and fitting only the readout matrix:

$$W_\ell^* = \arg\min_W \|Z_\ell W - Y\|_F^2 + \lambda \|W\|_F^2, \qquad f_{\text{read}}\big(z_i^{(\ell)}\big) = \arg\max_c \big(z_i^{(\ell)} W_\ell^*\big)_c. \tag{2}$$

Here Y contains the correct task label: color, side with more dots, spatial direction, or complete shape. The parameter λ controls regularization, and f_{read} is the fitted readout. The accuracy of f_{read} measures how well that label can be decoded from the frozen model's states.

We test each readout on held-out samples and keep paired images in the same split. As a control, we fit a separate readout using question tokens without an image. For Amodal Completion, the readout predicts one of six shape labels, whereas direct answering chooses among four options in the question. Appendix A.2 gives the splits and fitting details.

In the upper row of Figure 3, each task label is decodable from several layers. Numerosity and complete-shape accuracy is high; Feature Binding and Spatial Relations vary more across layers. Question-only controls stay near chance for the first three tasks. They score higher for Amodal Completion because the candidate list itself gives some information about the shape. Image-conditioned readouts outperform these controls through most layers. State interventions test whether this decoded information also affects the answer.

State interventions affect answers at some layers. Reading a task label from a state does not prove that the model uses that state to answer. We test this with matched image pairs. One image is the receiver; the other, with a different correct answer, is the donor. At a selected decoder block, we replace the receiver's image-token states with the donor's and continue the forward pass:

$$f_{\text{patch}}(r, d, \ell) \quad \text{assigns} \quad H_{r,\mathcal{I}_r}^{(\ell)} \leftarrow H_{d,\mathcal{I}_d}^{(\ell)}. \tag{3}$$

We swap states in both directions and retain only pairs for which both original answers are correct under the intervention scoring protocol. This leaves a small set whose size varies by model. As a control, we also use a donor from another pair that shares the receiver's label.

Table 1: All four component judgments can be performed above the guess level without explicit reasoning. Answer accuracy (%) on 1,000 images per task and model; guess is uniform random selection over the answer choices (two for Numerosity, four for the others).

Task	Qwen3.5-4B	Qwen3.5-9B	Gemma-4-E4B	Gemma-4-12B	Guess
Feature Binding	98.0	99.8	98.6	99.9	25.0
Numerosity	79.6	85.2	72.4	86.5	50.0
Spatial Relations	99.8	100.0	97.5	89.5	25.0
Amodal Completion	76.8	80.5	81.4	72.9	25.0
Avg	**88.6**	**91.4**	**87.5**	**87.2**	31.3

An intervention can shift answer scores even if the top answer stays the same (Heimersheim & Nanda, 2024). We measure this shift with $m = s(y_d) - s(y_r)$, the score for the donor's answer minus the score for the receiver's answer. We then compare how often the matched donor and the control donor increase this margin:

$$\Delta_{\text{patch}} = 100\,[\Pr(m_{\text{target}} > m_{\text{base}}) - \Pr(m_{\text{random}} > m_{\text{base}})]\,. \tag{4}$$

Here m_{base} is the receiver's score before replacement. A positive Δ_{patch} means that matched donors push scores toward their answers more often than control donors do. It measures how often scores move, not how far they move or how often the final choice flips. Appendix A.2 gives the scoring and matching rules.

In the lower row of Figure 3, matched donors move scores toward their answers more often than control donors in early and middle layers, but less consistently at the final tested block. Some task labels remain decodable in late layers even when matched-donor replacement no longer shows a consistent advantage over control replacement. Thus, decodability and the measured intervention effect can differ across depth. The results suggest that the Composite task is difficult because the model must combine the component judgments, not simply recognize their parts. A readout alone cannot establish that the model will use the information. We then test whether a shortened reasoning trace already supports the correct answer.

3.3 Development of Composite Answer Readiness

Reasoning improves counterfactual consistency. We compare direct answers with the model's normal reasoning on the same held-out Composite pairs. Sample accuracy counts correct images. Pair-consistent accuracy counts a pair only when both images are answered correctly.

Reasoning improves pair-consistent accuracy for all four models (Table 2). Sample accuracy rises for three models but falls for Gemma-4-12B. This model also often reaches the generation limit; these results do not show whether that limit causes its lower sample accuracy. Final answers still leave one question open: when during reasoning could the model first give the correct answer?

Tracing when the Composite answer becomes usable. We examine each model's reasoning trace at five checkpoints: before reasoning (Pre), early, middle, at the last reasoning token (Late), and just before the final choice (Answer). Validation data select one layer per model based on how well it represents the answer before reasoning. We use that same layer at every checkpoint and on all test samples. This analysis includes only pairs with valid checkpoints and parsed answers in both traces, so its sample sizes can differ from the direct-answer test. Appendix A.3 defines the checkpoints and selected layers; Appendix A gives the splits.

At each checkpoint, a separate linear readout predicts the correct side from the selected hidden state. We use the same feature standardization, ridge objective, and pair-preserving splits as in Section 3.2. Test accuracy measures how well the answer is decodable from that state.

We also test whether the model can answer from the reasoning generated so far. At each checkpoint, we truncate the trace after that checkpoint, append the same answer cue, and score the two possible answers with the frozen model:

$$f_{\text{force}}(i,t) = \arg\max_{c \in \{\text{left},\text{right}\}} s_\theta(c \mid x_i, r_{i,\le t}, a), \qquad U(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[f_{\text{force}}(i,t) = y_i\big]. \tag{5}$$

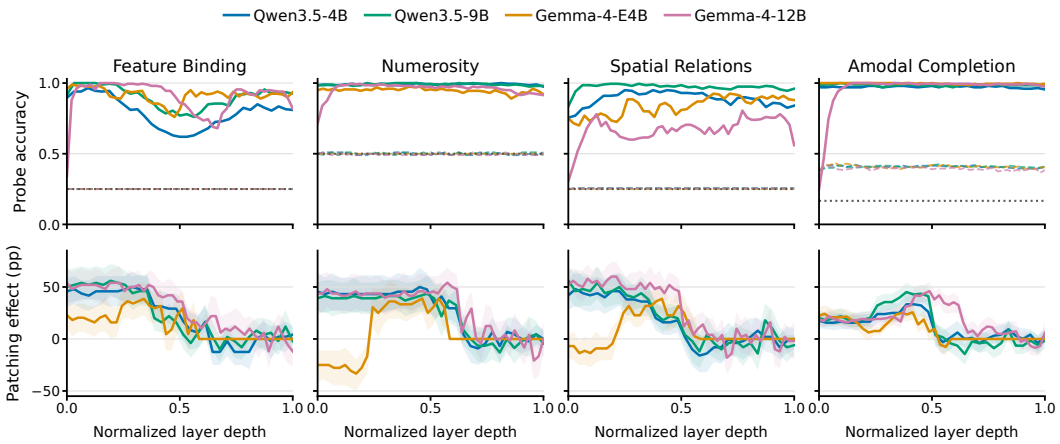

Figure 3: Top: test accuracy of readouts using image states (solid) or question-only states (dashed); gray dotted lines mark chance. Bottom: how often matched donor states move scores toward the donor's answer compared with control donors, in percentage points. Shading shows pair-bootstrap 95% CIs. Layer depth is normalized within each model.

Table 2: Native thinking improves counterfactual consistency. Values are accuracy (%) with pair-bootstrap 95% CIs; Δ is the reasoning-induced change in percentage points. Budget hit counts thinking-mode runs that reach the generation limit, out of 200 per model; the summary row pools all 800 runs.

	Sample Accuracy			**Pair-Consistent Accuracy**			**Budget hit**
Model	Non-thinking	**Thinking**	Δ	Non-thinking	**Thinking**	Δ	Count (%)
Qwen3.5-4B	58.5 [53.0,64.0]	**80.0** [74.0,86.0]	↑ 21.5	26.0 [18.0,35.0]	**67.0** [58.0,76.0]	↑ 41.0	38 (19.0%)
Qwen3.5-9B	66.5 [61.5,71.5]	**94.0** [90.5,97.0]	↑ 27.5	36.0 [27.0,45.0]	**89.0** [83.0,95.0]	↑ 53.0	11 (5.5%)
Gemma-4-E4B	52.0 [49.0,55.0]	**73.5** [67.0,80.0]	↑ 21.5	7.0 [2.0,12.0]	**57.0** [47.0,67.0]	↑ 50.0	0 (0.0%)
Gemma-4-12B	**57.5** [54.0,61.0]	54.5 [47.0,62.0]	↓ 3.0	15.0 [8.0,22.0]	**34.0** [25.0,43.0]	↑ 19.0	84 (42.0%)
Avg	58.6	**75.5**	↑ 16.9	21.0	**61.8**	↑ 40.8	133 (16.6%)

Here x_i is the image and prompt, $r_{i,\leq t}$ is the reasoning kept through checkpoint t, and a is the fixed answer cue, which gives the format but not the answer. The score s_θ compares the two choices. Unlike the readout from one hidden state, f_{force} uses the full retained trace, including any intermediate reasoning in text. Its accuracy measures whether the answer is usable at each checkpoint. This offline test needs the correct answer for scoring; that answer is unavailable when the model is deployed.

Decodability and usability develop during reasoning. Before reasoning, neither the hidden-state readout nor the shortened-trace test is consistently accurate across models (Figure 4). Both improve during reasoning, though at different speeds. Usability rises earlier for the Qwen models and Gemma-4-12B, and later for Gemma-4-E4B. By LATE, both decodability and usability are high for all four models.

Accuracy does not always improve from LATE to ANSWER: it stays high for the Qwen models but falls for both Gemma models. Correct answers are often available at earlier checkpoints, before the model finishes reasoning. Moreover, 16.6% of Composite thinking runs reach the generation limit without a final answer (Table 2). These results motivate predicting usability during generation.

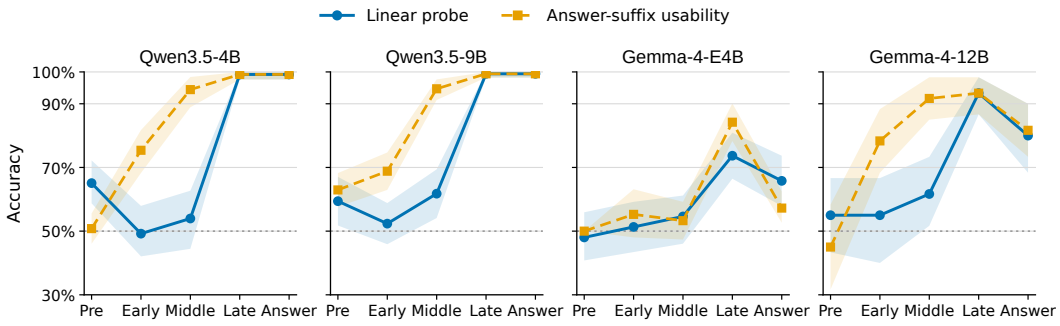

Figure 4: Each panel shows one model at five reasoning checkpoints. Blue: accuracy of a hidden-state readout for the Composite answer. Orange: accuracy when the model answers from a trace cut at that checkpoint. Shading shows pair-bootstrap 95% CIs; the dotted line marks chance.

4 Answer-Readiness-Guided Early Stopping

The shortened-trace test shows that an answer can be usable before reasoning ends. Figure 1 shows how this finding becomes a stopping rule. For each model and dataset, a small detector reads the current hidden state and its change at a layer selected on Composite data. Its predicted readiness score determines when to stop: once it crosses a threshold chosen on validation data, the model generates its answer. This resembles the use of an evidence threshold in studies of human perceptual decisions (O'Connell et al., 2012).

Predicting readiness without the correct answer. The offline test f_{force} uses the correct answer to measure readiness. At inference time, the detector must predict readiness without that answer. For training, we stop recorded traces at fixed checkpoints, ask the frozen VLM to answer, and label each checkpoint by whether that answer is correct. The detector uses the current state, its change since the previous checkpoint, and the elapsed reasoning length:

$$f_{\text{state}}(i,t) = \big[h_{i,t}^{(\ell)},\ d_{i,t}^{(\ell)},\ t/B\big], \qquad f_{\text{ready}}(i,t) = \sigma\big(w^\top \tilde{f}_{\text{state}}(i,t) + b\big), \tag{6}$$

Here t counts generated reasoning tokens, $d_{i,t}^{(\ell)}$ is the change since the previous checkpoint, and B is the generation budget. At the first checkpoint, we use a zero vector as the previous state, so $d_{i,\Delta}^{(\ell)} = h_{i,\Delta}^{(\ell)}$; subsequent checkpoints use $d_{i,t}^{(\ell)} = h_{i,t}^{(\ell)} - h_{i,t-\Delta}^{(\ell)}$. Training and online inference use the same convention. The tilde denotes standardization using training data. Only the detector weights w and bias b are trained.

Detector training. With checkpoint labels $y_{i,t} = \mathbf{1}[\hat{a}_{i,t}^{\text{forced}} = a_i^*]$, we minimize

$$\mathcal{L} = -\frac{1}{M} \sum_{(i,t)\in\mathcal{D}_{\text{train}}} \big[\alpha y_{i,t} \log f_{\text{ready}}(i,t) + (1-y_{i,t}) \log\big(1 - f_{\text{ready}}(i,t)\big)\big], \tag{7}$$

where $\hat{a}_{i,t}^{\text{forced}}$ is the forced answer, a_i^* is its reference, M counts training checkpoints, and α is their negative-to-positive label ratio. Feature normalization and class weights use only training checkpoints.

Stopping at a validated decision bound. We check the detector every $\Delta = 256$ reasoning tokens. At the first checkpoint where $f_{\text{ready}}(i,t) \geq \eta$, we close the reasoning channel and ask the VLM for its final answer. The interval is fixed; validation data select only the threshold η. We choose the threshold that minimizes mean reasoning length while keeping accuracy within a set tolerance of full thinking. If the score never crosses it, generation continues until the model stops naturally or reaches the token budget.

Table 3: Adaptive stopping cuts reasoning cost on both benchmarks while raising average accuracy. All held-out cases per model (1,500 MMStar and 765 RealWorldQA); accuracy in %, reasoning tokens as means, Δ in percentage points, reduction as the token decrease.

		Accuracy (%)			**Reasoning Tokens**		
Benchmark	Model	Full	**Stop**	Δ	Full	**Stop**	Reduction
MMStar	Qwen3.5-4B	64.87	**64.93**	↑ 0.07	2,041	**255**	87.5%
	Qwen3.5-9B	66.87	**69.87**	↑ 3.00	1,951	**255**	86.9%
	Gemma-4-E4B	58.07	**58.67**	↑ 0.60	753	**585**	22.3%
	Gemma-4-12B	56.93	**65.80**	↑ 8.87	1,854	**286**	84.6%
	Avg	61.69	**64.82**	↑ 3.13	1,650	**345**	**79.1%**
RealWorldQA	Qwen3.5-4B	72.81	**75.42**	↑ 2.61	1,696	**285**	83.2%
	Qwen3.5-9B	72.94	**77.91**	↑ 4.97	1,606	**248**	84.6%
	Gemma-4-E4B	**58.17**	57.12	↓ 1.05	442	**346**	21.8%
	Gemma-4-12B	60.13	**66.80**	↑ 6.67	859	**293**	65.9%
	Avg	66.01	**69.31**	↑ 3.30	1,151	**293**	**74.5%**

Table 4: Mean total output tokens and detector-triggered stopping (% of all questions).

	Total output tokens			Stopping (%)	
Model	Full	Stop	Reduction	Any check	First check
MMStar					
Qwen3.5-4B	2,159	326	84.9%	95.1%	94.7%
Qwen3.5-9B	2,066	315	84.8%	94.7%	94.3%
Gemma-4-E4B	944	724	23.3%	24.7%	15.9%
Gemma-4-12B	1,974	351	82.2%	76.3%	68.1%
RealWorldQA					
Qwen3.5-4B	1,699	299	82.4%	79.1%	71.1%
Qwen3.5-9B	1,610	263	83.7%	79.3%	78.4%
Gemma-4-E4B	448	359	19.9%	38.2%	29.5%
Gemma-4-12B	865	314	63.7%	39.9%	33.1%

5 Experiments

5.1 Experimental Setup

We compare each model's adaptive stopping run with its full-thinking run using the same token budget and answer extraction rule. MMStar (Chen et al., 2024) has multiple-choice questions about perception, science, and mathematics. RealWorldQA (x.ai) uses photographs and includes both multiple-choice and short-answer questions. We test all 1,500 MMStar and 765 RealWorldQA questions with two-fold cross-validation. One fold supplies detector training and threshold validation data; the other is held out for testing. All checkpoints from one question, and RealWorldQA questions with the same image, stay in the same fold. Each question therefore receives one held-out prediction per condition. Both conditions use a $B = 6{,}144$-token budget. Appendix B gives the split counts.

Threshold selection and answer transition. Validation selects the threshold that uses the fewest mean reasoning tokens while allowing at most 0.5 percentage points less accuracy than full thinking. When the detector triggers, we close the thinking channel and append `Therefore, the answer is`. Training labels and adaptive inference use this same transition. Each model, dataset, and evaluation direction has its own detector and threshold.

Answer scoring and token accounting. Within each benchmark, both conditions use the same answer extraction rule, which does not depend on the correct answer. MMStar uses the final choice letter. RealWorldQA accepts a choice or short answer, as the question requires. We extract answers

Table 5: Answer completion and changes in correct-answer counts. Missing includes traces without a parsed answer. Open/closed refer to full thinking; direct answers count as closed. Gains are stopping minus full thinking.

	Missing answer		Full status		Net correct gain		
Model	Full	Stop	Unclosed	Closed	Unclosed	Closed	Total
MMStar							
Qwen3.5-4B	297	24	267	1,233	96	−95	1
Qwen3.5-9B	276	44	243	1,257	121	−76	45
Gemma-4-E4B	104	87	32	1,468	20	−11	9
Gemma-4-12B	318	56	207	1,293	100	33	133
RealWorldQA							
Qwen3.5-4B	112	8	112	653	48	−28	20
Qwen3.5-9B	96	12	96	669	53	−15	38
Gemma-4-E4B	3	29	0	765	0	−8	−8
Gemma-4-12B	25	20	16	749	9	42	51

after the thinking channel closes and also accept direct short answers for RealWorldQA. A missing or unparseable answer counts as wrong.

Reasoning length counts tokens before the thinking channel closes. If it never closes, we count all generated tokens as reasoning. A direct answer has zero reasoning tokens. Total output length also includes answer tokens. We average lengths over all held-out questions, including those answered before the first checkpoint.

5.2 Main Results

Shorter reasoning with higher average accuracy. On MMStar, every model uses fewer reasoning tokens with adaptive stopping, and average accuracy rises (Table 3). Gemma-4-12B gains the most accuracy (+8.87 points), while Qwen3.5-4B is nearly unchanged (+0.07). Total output tokens fall by 23.3–84.9% across models (Table 4). The detector stops at the first 256-token check on 15.9–94.7% of questions, depending on the model. Appendix C.1 reports each test fold separately.

Results on short-answer questions. RealWorldQA tests whether the rule also works with short answers (Table 3). All four models use fewer reasoning tokens, and total output falls by 19.9–83.7% (Table 4). Average accuracy rises, although Gemma-4-E4B loses 1.05 points. The rule reduces output length for both answer formats; its accuracy effect varies by model.

Where the accuracy gains come from. Table 5 separates questions by whether full thinking closes its reasoning channel. On MMStar, adaptive stopping reduces missing answers from 297 to 24 for Qwen3.5-4B and from 276 to 44 for Qwen3.5-9B. For Qwen3.5-4B the 96 correct answers gained on open traces are almost exactly offset by 95 lost on closed traces, leaving a net gain of one; Qwen3.5-9B gains 121 on open traces against 76 lost on closed ones. Gemma-4-12B gains on both groups: 100 correct answers on open traces and 33 on closed ones, while Gemma-4-E4B gains 20 on open traces and loses 11 on closed ones. On RealWorldQA, Gemma-4-E4B has no open full-thinking traces to recover; it loses eight correct answers on closed traces.

6 Conclusion

VLMs can make the component visual judgments without explicit reasoning, but combining them is harder. During reasoning, the Composite answer becomes decodable from hidden states and usable from a shortened trace, often before the model stops on its own. A small detector predicts this point and stops reasoning early. Across MMStar and RealWorldQA, it reduces reasoning tokens and improves average accuracy, partly by producing answers from traces that otherwise end without one.

AI USE STATEMENT

In this work, we used generative AI tools only for final-stage language polishing and grammar refinement of the manuscript. We have not used generative AI tools for generating research ideas, designing methods, conducting experiments, implementing code, analyzing data, interpreting results, or producing scientific claims, and other generative AI-assisted tasks requiring disclosure are not applicable to this work. We have reviewed all AI-assisted revisions to ensure their correctness and consistency with the authors' intended content. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work by providing detailed descriptions of the proposed method, experimental settings, implementation details, and evaluation protocols in the main paper and supplementary materials. The supplementary materials include the source code required to reproduce our experiments, and we will publicly release the code after the review process to further facilitate verification and future research. The content of this paper can be completely reproduced.

REFERENCES

Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. Knowing before saying: Llm representations encode information about chain-of-thought success before completion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12791–12806. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.662.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *arXiv preprint arXiv:2406.04236*, 2024. doi: 10.48550/ARXIV.2406.04236.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. ISSN 1530-9312. doi: 10.1162/coli_a_00422.

Jing Bi, Luchuan Song, Dingxin Zhang, Pinxin Liu, Guangyu Sun, Lianggong Bruce Wen, Weidong Cai, Chen Chen, and Chenliang Xu. Does reasoning improve seeing? understanding when vision-language models benefit from thinking. In *Proceedings of the International Conference on Machine Learning 2026*, 2026.

Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of gqa. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 94–105. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.9.

Laura A. Carlson-Radvansky, Eric S. Covey, and Kathleen M. Lattanzi. "what" effects on "where": Functional influences on spatial relations. *Psychological Science*, 10(6):516–521, 1999. ISSN 1467-9280. doi: 10.1111/1467-9280.00198.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. doi: 10.48550/ARXIV.2403.20330.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. ISSN 0896-6273. doi: 10.1016/j.neuron.2012.01.010.

Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. Core systems of number. *Trends in Cognitive Sciences*, 8(7):307–314, 2004. ISSN 1364-6613. doi: 10.1016/j.tics.2004.05.002.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire,

Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. doi: 10.48550/ARXIV.2404.15255.

Jen-Tse Huang, Dasen Dai, Jen-Yuan Huang, Youliang Yuan, Xiaoyuan Liu, Wenxuan Wang, Wenxiang Jiao, Pinjia He, Zhaopeng Tu, and Haodong Duan. Human cognitive benchmarks reveal foundational visual gaps in mllms. *arXiv preprint arXiv:2502.16435*, 2025. doi: 10.48550/ARXIV.2502.16435.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *International Conference on Machine Learning*, pp. 27793–27830, 2025.

Zhuoran Jin, Kejian Zhu, Hongbang Yuan, Yupu Hao, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Look light, think heavy: What multimodal chain-of-thought reasoning can and cannot do. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8567–8608. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.387.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. 2017.

Kohitij Kar and James J. DiCarlo. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1): 164–176.e5, 2021. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.09.035.

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B. Issa, and James J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0392-5.

Philip J Kellman and Thomas F Shipley. A theory of visual interpolation in object perception. *Cognitive Psychology*, 23(2):141–221, 1991. ISSN 0010-0285. doi: 10.1016/0010-0285(91)90009-d.

Victor A.F. Lamme and Pieter R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000. ISSN 0166-2236. doi: 10.1016/s0166-2236(00)01657-x.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3769–3798. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.192.

Gordon D. Logan. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015–1036, 1994. ISSN 0096-1523. doi: 10.1037/0096-1523.20.5.1015.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*, 2022. doi: 10.48550/ARXIV.2202.05262.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. doi: 10.48550/ARXIV.2410.07149.

Redmond G O'Connell, Paul M Dockree, and Simon P Kelly. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, 15(12):1729–1735, 2012. ISSN 1546-1726. doi: 10.1038/nn.3248.

Israfel Salazar, Stella Frank, Dan Oneata, Desmond Elliott, and Constanza Fierro. Pathways of visual information flow in vision-language models. *arXiv preprint arXiv:2607.03358*, 2026. doi: 10.48550/ARXIV.2607.03358.

Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh S. Rawat. Probing conceptual understanding of large visual-language models. 2024.

Allison B. Sekuler and Stephen E. Palmer. Perception of partly occluded objects: A microgenetic analysis. *Journal of Experimental Psychology: General*, 121(1):95–111, 1992. ISSN 0096-3445. doi: 10.1037/0096-3445.121.1.95.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *arXiv preprint arXiv:2403.16999*, 2024. doi: 10.48550/ARXIV.2403.16999.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, 1996. ISSN 1476-4687. doi: 10.1038/381520a0.

Anne Treisman and Hilary Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14(1):107–141, 1982. ISSN 0010-0285. doi: 10.1016/0010-0285(82)90006-8.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. ISSN 0010-0285. doi: 10.1016/0010-0285(80)90005-5.

x.ai. Grok-1.5 vision preview, 2024.

Yang Xiang, Yixin Ji, Ruotao Xu, Dan Qiao, Zheming Yang, Juntao Li, and Min Zhang. When is thinking enough? early exit via sufficiency assessment for efficient reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23541–23556. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1080.

Fei Xu and Elizabeth S. Spelke. Large number discrimination in 6-month-old infants. *Cognition*, 74 (1):B1–B11, 2000. ISSN 0010-0277. doi: 10.1016/s0010-0277(99)00066-9.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weipinng Wang. Dynamic early exit in reasoning models. *International Conference on Learning Representations*, pp. 88170–88210, 2026.

Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. *arXiv preprint arXiv:2508.20279*, 2025. doi: 10.48550/ARXIV.2508.20279.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025a. doi: 10.48550/ARXIV.2504.05419.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025b. doi: 10.1109/cvpr52734.2025.01842.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. doi: 10.48550/ARXIV.2302.00923.

A Datasets and Mechanistic Analysis

Table 6 summarizes the synthetic datasets and their analysis splits.

Table 6: **Synthetic dataset sizes and analysis splits.** Counts are images. Primitive splits are used for linear probing, and the Composite split is used for reasoning-trajectory analysis before trace filtering. Behavioral evaluation uses all 1,000 images per primitive and the 200-image Composite test set. Paired images remain in the same split.

			Analysis Split			
Dataset	Images	Choices	Train	Validation	Test	Split Unit
Feature Binding	1,000	4	800	–	200	Image
Numerosity	1,000	2	800	–	200	Image
Spatial Relations	1,000	4	800	–	200	Image
Amodal Completion	1,000	4	800	–	200	Pair
Composite	1,000	2	640	160	200	Pair

A.1 Primitive Dataset Construction

We generate each dataset so the feature that determines the answer varies while other scene details are controlled. The settings for each task follow.

Feature Binding. Feature Binding uses four target colors (red, blue, green, and yellow), four shapes, 4–8 objects per image, and three layout templates. Target shape, size, and position also vary.

Numerosity. Numerosity uses three ratios between the larger and smaller dot counts (2.0, 1.5, and 1.25). For each ratio, total dot area either agrees with the count difference, is similar on both sides, or points in the opposite direction. Dots do not overlap, and each side is equally often the one with more dots.

Spatial Relations. Spatial Relations uses black and gray outlines for the target and reference, with their colors and shapes sampled independently. Its conditions cross four directions, three target–reference distances (60, 90, and 120 pixels), and three scene sizes (4, 6, and 8 objects).

Amodal Completion. Amodal Completion uses six shapes and six kinds of occluder. Each shape is 25–65% hidden but has at least two visible parts. The two images in a pair have different complete shapes. They share the target's color, center, size, and rotation settings, the occluder, the intended amount of occlusion, and the order of four answer choices. The actual hidden area can differ between the shapes.

Behavioral and probe splits. The VLM stays frozen during direct answering and readout fitting. We balance splits by target color for Feature Binding; by count ratio, area condition, and answer for Numerosity; and by relation, distance, and object count for Spatial Relations. Amodal Completion has 500 matched pairs: 400 for training and 100 for testing, with occluder types balanced. Paired images always stay in the same split.

Counterfactual pairs for primitive interventions. Feature Binding, Numerosity, and Spatial Relations each have a separate intervention set of 125 matched pairs: 100 training and 25 test pairs. Each pair keeps the scene context fixed while changing the feature that determines the answer. Amodal interventions use the 100 held-out Amodal Completion pairs, which change the complete shape under the same occluder. We intervene in both directions and apply the criteria in Section 3.2 to choose pairs for the primary analysis.

A.2 Primitive Readouts and Interventions

This subsection gives the settings for Section 3.2. Direct-answer tests disable thinking, request a short response, and allow at most 40 generated tokens. We accept a normalized choice letter or a

matching word answer. Table 7 summarizes the readouts. For standardized training features X and one-hot labels Y, we solve

$$W = X^\top (XX^\top + 100I)^{-1} Y,$$

and predict the label maximizing zW for a standardized test vector z.

Table 7: **Primitive linear-readout settings.** Dataset splits are given in Table 6.

Setting	Specification
Inputs	No-thinking image and question
Visual features	Mean-pooled image-token states at each tested depth
Question-only control	Mean-pooled question span, including candidates, with no image
Standardization	Per-dimension training mean and standard deviation
Label classes	4 for Feature Binding, 2 for Numerosity, 4 for Spatial Relations, and 6 for Amodal Completion

Patching protocol. The no-thinking prompt ends in an answer cue, and we score the next-token choice. Each choice score is the maximum logit over the last token IDs of uppercase/lowercase and leading-space variants of its letter. Pairs and controls satisfy the following criteria:

- Both intervention directions are available and both unpatched answers are correct.
- The original donor and receiver margins differ by more than 10^{-6} in both directions.
- Matched-random donors come from the same test set, belong to another pair, and share the receiver's semantic label. Amodal controls additionally match question-token length.

The lower row of Figure 3 averages the two directional indicator differences within each pair. Table 8 gives the retained sample sizes.

Table 8: **Retained counterfactual pairs for primitive interventions.**

Task	Qwen3.5-4B	Qwen3.5-9B	Gemma-4-E4B	Gemma-4-12B
Feature Binding	24	25	22	25
Numerosity	22	23	18	24
Spatial Relations	25	25	22	25
Amodal Completion	48	62	72	75

A.3 Composite Data and Reasoning Checkpoints

The Composite analysis uses scenes that require several component judgments. We also define fixed points where we can cut off a reasoning trace and test whether the model can answer.

Composite dataset. The Composite task has 1,000 images in 500 matched pairs. Each side shows 8–12 partly hidden red or blue shapes; the model must choose the side with more red circles. Within a pair, we change which shapes have which colors to reverse the answer. Positions, shapes, sizes, rotations, occluders, and overall color and shape counts remain fixed. Occlusion ranges from 25% to 60%. We reserve 100 pairs (200 images) for direct-answer testing. The other 400 pairs provide 320 training and 80 validation pairs for reasoning-trace analysis, with validation selecting the analysis layer. We retain a pair only if both traces have a parsed answer and usable checkpoints. Sample counts can therefore differ by model. Table 9 defines the five checkpoints.

Table 9: **Reasoning-trajectory checkpoints.** T counts reasoning tokens, and positions are one-based.

Checkpoint	State position
PRE	Last prompt token before reasoning
EARLY	Reasoning token $\lfloor T/3 \rfloor$
MIDDLE	Reasoning token $\lfloor 2T/3 \rfloor$
LATE	Last reasoning token, T
ANSWER	Token immediately preceding the first final-choice token

Validation selects one layer per model: L17 for Qwen3.5-4B, L16 for Qwen3.5-9B, L1 for Gemma-4-E4B, and L20 for Gemma-4-12B. Test samples do not affect this choice. The selected depth differs across models, so one layer cannot represent the Composite answer equally well in every architecture.

MMStar and RealWorldQA use separate early-stopping splits, described in Appendix B.

B EARLY-STOPPING DETAILS

B.1 DATASETS AND CROSS-VALIDATION

We use every MMStar and RealWorldQA question in two-fold cross-validation. Within one fold, we train the detector and select its threshold; we test on the other fold (Table 10). All checkpoints from one question stay together, as do RealWorldQA questions that share an image. Every model uses the same folds.

Table 10: **Early-stopping evaluation splits.** Counts are questions. Each source fold is divided into detector training and threshold validation, and the other fold is held out for testing.

		Source fold		Held-out fold
Dataset	Evaluation	Training	Validation	Test
MMStar	A→B	600	150	750
	B→A	600	150	750
RealWorldQA	A→B	306	77	382
	B→A	305	77	383

Each question contributes one held-out prediction per condition. Pooled accuracy weights questions equally across the two evaluation directions.

C SUPPLEMENTARY EARLY-STOPPING RESULTS

The main text combines the two test folds and analyzes answer completion. Here we report each fold separately.

C.1 PERFORMANCE ACROSS EVALUATION FOLDS

Tables 11 and 12 give accuracy and mean reasoning length for each test fold. Reasoning length falls in every model and dataset, but accuracy changes differ between folds.

Table 11: **Per-direction held-out MMStar results.** Each row evaluates 750 held-out cases. Accuracy values are percentages and reasoning-token counts are means. Δ denotes the change from full thinking to adaptive stopping in percentage points.

		Accuracy (%)			Reasoning Tokens	
Model	Train→Test	Full	**Stop**	Δ	Full	**Stop**
Qwen3.5-4B	A→B	64.13	64.13	0.00	2,101	**256**
	B→A	65.60	**65.73**	↑ 0.13	1,982	**254**
Qwen3.5-9B	A→B	66.93	**69.73**	↑ 2.80	1,979	**256**
	B→A	66.80	**70.00**	↑ 3.20	1,922	**254**
Gemma-4-E4B	A→B	57.47	57.47	0.00	767	**437**
	B→A	58.67	**59.87**	↑ 1.20	739	**733**
Gemma-4-12B	A→B	57.60	**66.67**	↑ 9.07	1,822	**327**
	B→A	56.27	**64.93**	↑ 8.67	1,885	**246**

Table 12: **Per-direction held-out RealWorldQA results.** A→B evaluates 382 questions, and B→A evaluates 383. Accuracy values are percentages and reasoning-token counts are means. Δ denotes the change from full thinking to adaptive stopping in percentage points.

		Accuracy (%)			Reasoning Tokens	
Model	Train→Test	Full	**Stop**	Δ	Full	**Stop**
Qwen3.5-4B	A→B	71.73	**74.87**	↑ 3.14	1,692	**247**
	B→A	73.89	**75.98**	↑ 2.09	1,700	**322**
Qwen3.5-9B	A→B	73.30	**78.27**	↑ 4.97	1,521	**252**
	B→A	72.58	**77.55**	↑ 4.96	1,692	**244**
Gemma-4-E4B	A→B	**57.33**	56.81	↓ 0.52	448	**347**
	B→A	**59.01**	57.44	↓ 1.57	436	**344**
Gemma-4-12B	A→B	59.95	**63.87**	↑ 3.93	781	**217**
	B→A	60.31	**69.71**	↑ 9.40	936	**369**