WHERE CROSS-MODAL REASONING DISTILLATION HURTS, AND HOW TO UNDO IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Distilling chain-of-thought traces from a text-only language model into a vision-language model is a standard route to multimodal reasoning, and it carries a known side effect: the student sees worse and hallucinates more. The field prices this as the cost of the reasoning gained. But the two instruments behind that price are bent: the answer extractor misses a concise model's answers, in favor of the model that writes more, and the hallucination matcher mistakes similarity for identity and under-counts hallucination. Repaired, they show no measurable gain for the student over the base it started from on any of five benchmarks, while the cost is paid in full. That cost is not one thing but three, and each has its own remedy. (i) *Answers the student never states*: forcing a conclusion at inference brings them back, and no weight changes. (ii) *Hallucination*: it is mostly the cost of length, not of reasoning, and a teacher with shorter traces removes it. (iii) *Perception cost*: it has an address, resolvable only on questions that cannot be answered without the picture — a blind teacher takes the sight its student already had. Averaging the distilled weights back toward the original ones, with no training, recovers most of the lost accuracy while keeping most of the reasoning habit. We call it *behavior-preserving interpolation*.

1 INTRODUCTION

Distilling the chain-of-thought traces of a text-only language model into a vision-language model is a standard route to multimodal reasoning. It has two known side effects: the student sees worse, and it hallucinates more. Both have been reported on the 3B, 7B, and 72B Qwen2.5-VL, on Qwen3-VL, on InternVL3, and even on GPT-4o (Bousselham et al., 2025; Tian et al., 2025; Kancheti et al., 2026; Wu et al., 2026; Liu et al., 2025a). The two are not accidents; they are accepted: the field prices them as the cost of the reasoning ability the student gains in exchange, so every model trained along this route pays it. Gain and cost are each reported only as a single average over a whole benchmark. But the average hides two things: whether that gain is real, and on whom that cost falls.

It is not only what the average hides; the two numbers themselves are wrong—the two instruments the field now uses to measure them are both bent. The answer extractor is tuned to the reasoning template and misses a concise model's answers, favoring the model that writes more, and the model that writes more is precisely the distilled one: this alone creates, out of nothing, a 10-point gain for the student on WeMath. The hallucination matcher judges objects by word-vector similarity and mistakes similarity for identity, reporting every model's hallucination about four points too low. The matcher's defect is in AMBER's reference implementation itself. Once they are repaired, this pricing no longer holds. We reproduced the cold-start stage following the original authors' recipe, and no gain is measured on any of five benchmarks; three of these five are exactly the three on which the original paper reports its largest gains. The field does not deny this point (§4.1), but the real question is where the cost arises, and on that the field has had no definite conclusion.

We use the teacher itself as a probe: questions it answers correctly without the image are text-solvable, the rest image-requiring. On the image-requiring questions the student drops below the base by 12.55 points on MathVista and 7.05 on MMStar, and no loss is measured on the text-solvable ones. A teacher that cannot see took away the sight the student already had. On MathVista the student's median answer is 1,107 words and the base's 3—the student did learn the teacher's habit. What it learned is the form, not the ability. No measurable gain, yet the cost is paid in full.

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.

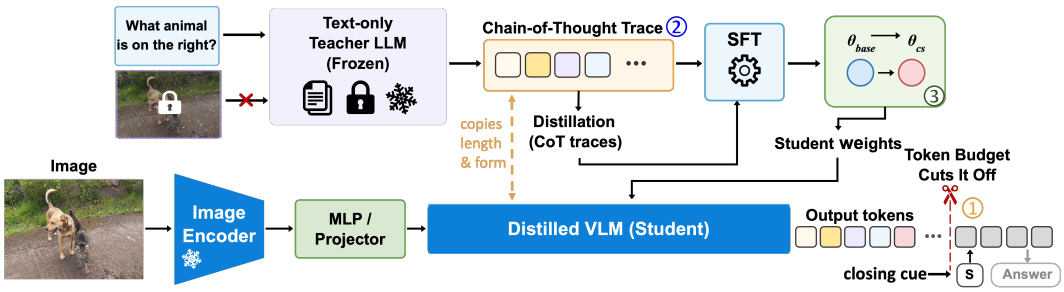


Figure 1: The distillation pipeline. ①–③: where the three parts of the cost arise (Figure 3).

The cost is not one thing but three (Figures 1 and 3), and so need not be accepted as a whole: each can be removed on its own. Our contributions are four: one for each of the three, plus repairing the instruments. **Answers the student never states.** Before a conclusion is forced, the student drops 19.00 points on MathVista, and half of this closes once it is made to state an answer at the budget: forcing a conclusion at inference recovers +10.00 points, without a single weight changed. **Hallucination: it is mostly the cost of length, not of reasoning.** The teacher sets how long the student writes: distilled again from a reasoning teacher with shorter traces, the student's 8.26 points of extra hallucination fall to −0.36, no longer resolvable from the base. **Perception cost: the one neither of the first two can cure.** One scalar that averages the distilled weights back toward the original ones separates it from the reasoning habit, with no training. On a test set of 400 MathVista questions used for no choice, the ratio a fixed rule picks on a separate selection set scores 8.00 points above the student with no resolvable perception cost, and a ratio that keeps 84% of the reasoning habit scores 5.00 above it, again with none. We call this behavior-preserving interpolation (BPI). **Three instruments.** We replaced all three: the extractor by one set of rules for every model, the matcher by exact object matching, and the benchmark average by blind-teacher stratification, with which any comparison asks "where does the loss fall".

2 Setup

We rebuilt the two-stage recipe of VOLD (Bousselham et al., 2025) from its description: first fine-tune the vision-language model on the reasoning traces of a text-only teacher (cold start), then run on-policy reinforcement learning together with on-policy distillation from the teacher (the student generates, and the teacher's distribution at each token pulls it toward the teacher). The model being distilled is Qwen2.5-VL-3B-Instruct (Bai et al., 2025), the teacher is Qwen3-8B (Yang et al., 2025), and the vision tower is frozen throughout, so every point of perception loss below occurs with the vision encoder unchanged. Below, *base* is the model before distillation and *student* the checkpoint after cold start; every claim about capability concerns only the cold-start step, which was trained only once. This recipe has no public implementation (configuration in Appendix A.1), so we do not pursue matching VOLD's absolute scores: every comparison is between models of one pipeline. Accuracy is measured on five benchmarks: MathVista (Lu et al., 2023) and MMStar (Chen et al., 2024) (perception-heavy), plus LogicVista (Xiao et al., 2024), WeMath (Qiao et al., 2024), and MathVision (Wang et al., 2024), each on its first 300 questions, all with a conclusion forced at the budget by Eq. (4), the base included. Hallucination is CHAIR (Rohrbach et al., 2018) on all 1,004 AMBER (Wang et al., 2023) images, lower is better (Eq. (1)). Every comparison is item- or image-paired, with a 95% bootstrap interval; a difference whose interval excludes zero is called resolvable, and otherwise unresolvable—and we never read an unresolvable difference as equality.

3 None of the Three Instruments Is Neutral

Every comparison in the original paper passes through three instruments, all used in the field as standard, and none of them is neutral. Any work comparing a long-reasoning model with a concise base inherits all three errors at once. The extractor is not bent at random: it creates a gain that does not exist in one place and hides a cost that does in another, both pushing the numbers toward "the bargain is worth it". The matcher is bent evenly, reporting every model's hallucination about four

points too low. The third gives only a total, like a bill with no itemization: this cost looks as if the model got worse, when only one ability was taken away (§4).

3.1 THE ANSWER EXTRACTOR

Long-reasoning models end with a template (`\boxed{}`, "the answer is (B)"); models that answer directly do not. An extractor tuned to the template can read the former but not the latter. This defect appeared four times in our reproduction, and each time it overestimated the side that writes longer. The fourth time went like this: on WeMath, 105 of the base's responses cannot be read, most often written as "the correct answer is: D". The extractor does not recognize these forms, so these 105 are recorded as unanswered and all counted wrong, which puts the base's score at 41.67%, while its true score is 52.00%. The student ends with `\boxed{}`, which the extractor recognizes: 52.00%—not a hair different from the base's true score. The student appears to earn 10 points and in fact earns none. VOLD reports on HallusionBench that the student has no hallucination cost, but scored under the convention it was published with, the gap is in fact -1.68, all of it from the student's unreadable responses (Appendix B.1). The same extractor defect creates a gain out of nothing in one place and hides a cost in another. These errors point to one thing: the extraction rules grew out of the output format of one particular model. We designed a set of rules that does not look at which model it is reading, called **symmetric extraction**; it consists of two principles. *First, the set of forms is fixed.* The accepted answer forms are listed in advance (Appendix B.1), the same set is used for all models, and nothing may be added or removed because of some model's output format. *Second, the answer must be stated, not merely present.* We take the last answer stated in the response and never match by substring containment—a "B" appearing in a piece of reasoning does not mean the answer is B. These rules carry over directly to any evaluation that compares models of different lengths; with them in place, the base's suppressed score comes back while the student, which writes long reasoning, barely moves—the defect only ever pressed on one side (Appendix B.1).

3.2 THE HALLUCINATION MATCHER

AMBER judges whether an object is a hallucination by whether it is on the image's annotation list. Write O_r for the objects that response r mentions, G_r for the objects annotated in the image, and A_r for the hallucination-prone objects the annotation lists; CHAIR is pooled over all responses:

$$\text{CHAIR} = 100 \cdot \frac{\sum_r \left|\{\, o \in O_r : h(o) = 1 \,\}\right|}{\sum_r |O_r|}. \tag{1}$$

The two implementations differ only in how h decides. The reference implementation asks about similarity first: as soon as the spaCy word-vector similarity between o and some object in G_r exceeds the threshold (0.8 by default), it sets $h(o) = 0$, before A_r is ever consulted. But similar is not the same: the model says "dog", the list has only "cat", the two words' similarity is 0.802, and the matcher decides that "dog is on the list"—so this hallucination does not count. On the base's 1,004 descriptions, this rule reports 203 fewer hallucinated objects than exact matching—real hallucinations, let through only because they are close to some word on the list (Appendix B.2); CHAIR therefore shrinks from 15.38 to 11.13. The CHAIR of six models spans 12.27 to 24.34, while the understatement stays between 3.64 and 4.25. Any AMBER CHAIR reported with the reference implementation can be read directly as about 4 points too low, whichever model is measured. Exact matching is not our invention: the definition of CHAIR (Rohrbach et al., 2018) requires identity, and item-by-item comparison is how it was done in the first place. We only change it back, letting identity decide first: o is a hallucination if it appears verbatim in A_r and faithful if it appears verbatim in G_r, and similarity is consulted only when it is on neither list. All our CHAIR values use it.

3.3 AN AVERAGE THAT MEASURES SIZE, NOT LOCATION

The third instrument favors no one; it is only too coarse. A benchmark average is one number computed by mixing "questions that can be answered only by looking at the image" with "questions that can be answered without looking at it", and the accuracies measured on these two kinds can move in opposite directions. The reported number neither correctly represents the size of the loss nor points to its location. To answer "where does it fall", we first sort the questions by **blind-teacher stratification**. We let the text-only teacher T itself answer every question q once without

Table 1: Base, student, $\lambda = 0.75$ (§6). Paired differences, 95% bootstrap CIs; $\star$: excludes zero (no λ−Base interval does). Length kept: λ's median length over the student's. †114 free-form items. Questions 0–299, the set with the largest cost of the three; all 1,000 in Appendix E.1.

					$\lambda = 0.75$ (§6)			
Benchmark	n	Base	Student	Student − Base	Acc.	− Student	− Base	Length kept
MathVista	300	62.33	52.67	−9.67 [−15.67, −3.67]*	61.33	+8.67 [+3.00, +14.33]*	−1.00	83%
MMStar	300	58.67	55.33	−3.33 [−8.33, +2.00]	58.00	+2.67 [−2.33, +8.00]	−0.67	75%
LogicVista	300	39.33	35.33	−4.00 [−11.00, +2.33]	45.33	+10.00 [+4.00, +16.00]*	+6.00	96%
WeMath	300	52.00	52.00	+0.00 [−6.33, +6.33]	58.00	+6.00 [+0.67, +11.67]*	+6.00	96%
MathVision†	114	15.79	20.18	+4.39 [−3.51, +12.28]	20.18	+0.00 [−7.02, +7.02]	+4.39	94%

the image; the ones it answers wrong count as image-requiring, and the ones it answers correctly as text-solvable: $S_{\text{img}} = \{q : T(q) \neq y_q\}$ and $S_{\text{txt}} = Q \setminus S_{\text{img}}$. The accuracy difference between any two models A and B over a whole benchmark is then exactly a weighted sum over the two strata:

$$\Delta(Q) = \frac{|S_{\text{img}}|}{|Q|}\Delta(S_{\text{img}}) + \frac{|S_{\text{txt}}|}{|Q|}\Delta(S_{\text{txt}}), \qquad \Delta(S) = \frac{100}{|S|}\sum_{q\in S}\big(c_A(q) - c_B(q)\big), \tag{2}$$

where $c_M(q)$ is whether model M answers q correctly. When the two terms have opposite signs, the average hedges the loss away. MMStar is exactly this case: the student is 3.33 points below the base, which splits as $-3.33 = \frac{241}{300}(-7.05) + \frac{59}{300}(+11.86)$—7.05 points lower on the 241 image-requiring questions, while on the other 59 text-solvable questions its point estimate is instead higher, so a seven-point loss is hedged down to three (MMStar's text-solvable stratum is mostly items the teacher guessed right; Appendix B.3). The split need not be perfect, only fixed in advance and the same for all models—a question sorted wrongly is sorted wrongly for both sides. With this split, "where does the cost of one training intervention fall" becomes a question that can be answered. The labels come from one greedy pass capped at 1,024 tokens, where most answers are cut off, so "image-requiring" means "not answered correctly within that budget". Relabeling three other ways moves up to a third of the items; the MathVista result holds under all of them (Appendix B.3).

Takeaway. A reported number is not a property of the model; it is the joint product of the model and the instruments that measure it. And the choice of instrument is not an error term: the amount it moves the number is as large as the effect being measured, and sometimes it is the effect itself. All three instruments are cheap to repair: the first two need only a rerun over responses already generated, and the third only one pass of the teacher over the questions without the image.

4 What the Cost Is

On none of the five benchmarks is the cold-start student resolvably more accurate than the base; the cost is resolvable only on the image-requiring questions, and is concentrated on a set of questions the base gets at a glance—which the student, after reasoning at length, instead answers wrong.

4.1 No Gain on Five Benchmarks

Table 1 compares base and student on the same questions and extractor, with forced conclusions (Eq. (4)). Of the five benchmarks, four show no resolvable difference, and the fifth, MathVista, shows a 9.67-point loss; without forcing, 36 of the student's responses yield no answer, and the loss is 19.00 (§5.1). And three of these five—MathVision, LogicVista, and WeMath—are exactly the three that the recipe reports as rising most after its whole pipeline has run.

VOLD itself reports that its SFT-only checkpoint is below the base on all eight of its benchmarks, and credits the whole gain to RL. We ran the RL stage (seed 1 on four benchmarks; three seeds on MathVision): the only resolvable gain over the base is on MathVision's free-form questions, only 3.80 points above the cold-start student, too little to credit to RL (Appendix C.8).

What transfers is length and form, and they do not buy accuracy. The student copies the teacher's length (Table 2) and form, down to the cut in truncated traces (Appendix C.2): what it copies is not the content but the template of "how long to write and where to stop". This template does not

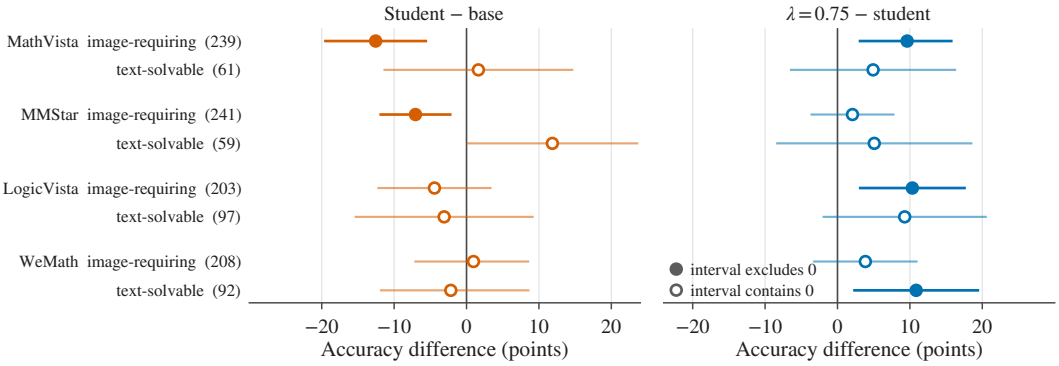


Figure 2: Blind-teacher strata: item-paired accuracy differences, 95% bootstrap intervals (filled: excludes zero). Left: the student against the base. Right: $\lambda = 0.75$ against the student (§6.2).

buy accuracy: across a 24-fold span of MathVista decoding budgets, accuracy moves by only 2.33 points; yet between the two ends forty percent of the questions change their answer, 65 of them between right and wrong—win a batch, lose a batch, and the net comes to zero (Appendix C.1).

Takeaway. The gain is either in the RL stage or cannot be placed in any stage; the cost is paid in the cold-start stage. The two are never in the same place. The transferred template is inert for the total score but not for individual questions: 52.67 does not correspond to a definite set of questions.

4.2 The Perception Cost Falls on the Image-Requiring Questions

The stratification of §3.3 is applied here (Figure 2, left), with forcing on as in Table 1. On MathVista, the student is 12.55 points below the base on the image-requiring questions, and almost no change is measured on the text-solvable ones; MMStar is the example of Eq. (2). The text-solvable stratum is small, with intervals over twenty points wide, so we claim only that no loss is resolvable there, not a gain. The difference between strata is resolvable on MMStar and marginal on MathVista under these labels; neither survives every relabeling (Appendix B.3). That is 30 and 17 questions.

Are the broken questions always the same set? We trained five students, differing in teacher and in trace truncation length, and then compared question by question what each of them got wrong. On the set C of 187 MathVista questions the base answers correctly, if the five students erred independently, the expected number of questions answered wrong by at least k of them would be

$$\mathbb{E}[N_{\geq k}] = |C| \cdot \Pr\Big(\sum_{a=1}^{5} X_a \geq k\Big), \qquad X_a \sim \text{Bernoulli}(e_a) \text{ independent}, \tag{3}$$

where e_a is student a's error rate on C (0.28 to 0.38). In fact the overlap is extremely large: 28 questions are answered wrong by four or more, against only 6.8 expected; 14 by all five, against 0.57 expected, an enrichment of 25 times. MMStar is the same. The teacher changed and the truncation length changed, yet the questions broken are the same set: this is a property of the recipe, not an accident of one training run. These questions (the damage set below) are all very simple: the base answers them with a median length of just 1 word (Appendix C.3). *What causes it?* Of eight candidate explanations, each tested by one direct measurement, none holds (Table 11). Two stand out. The first is "there are no images in the training data": another student, whose 997 training examples each carry an image, likewise drops 7.05 points on MMStar's image-requiring questions. The second is "the student simply stops looking at the image": if so, deleting the image should affect it less; the opposite holds: on AMBER descriptions, deleting the image perturbs the student's output distribution more than the base's (mean KL rises from 0.566 to 0.686). The student does not fail to look at the image; it looks and then cannot use it. We call this the **perception cost**.

Are the perception cost and hallucination the same thing? Table 2 lists the ten models next to the base. Among the seven students, the longer they write, the more they hallucinate ($\rho = +0.893$); yet across the ten models, the longest-writing of which writes 17 times more than the shortest, whether a model writes long shows no resolvable relation to how many points it drops on the image-requiring questions ($\rho = -0.127$, $p = 0.73$). The student taught by Qwen2.5-VL-7B writes only 140 words

Table 2: Eleven models, one pipeline; the seven students differ only in teacher or trace cap. Teacher trace: median words. Cost: accuracy − base on the 239 image-requiring MathVista items.

		MathVista			AMBER	
Model	Teacher trace (words)	Acc.	Cost	Words	Words	CHAIR
Base (undistilled)	—	62.33	—	3	79	15.38
Teacher Qwen3-8B (ours)	2,047	52.67	−12.55	1,107	355	23.64
Teacher R1-Distill-Llama-8B	1,328	47.67	−17.15	401	196	15.02
Teacher Qwen2.5-7B (text-only)	273	55.00	−7.95	125	91	17.10
Teacher Qwen2.5-VL-7B (vis. trained)	260	52.00	−12.97	140	83	15.56
Qwen3-8B traces cut at 2,048 tok.	1,394	53.33	−11.72	1,170	381	26.96
cut at 1,024 tok.	753	36.00	−29.71	764	358	26.02
cut at 512 tok.	387	22.00	−42.26	343	397	27.83
Interpolated, $\lambda = 0.25$	—	59.33	−5.02	66	87	15.20
$\lambda = 0.50$	—	59.67	−3.35	206	108	18.39
$\lambda = 0.75$	—	61.33	−2.93	922	184	22.43

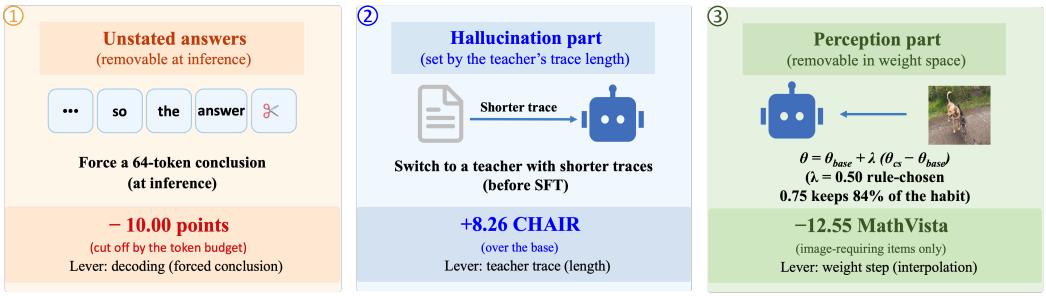


Figure 3: Three parts of the cost and their levers. Bottom: ① MathVista points lost to the budget cut; ② extra AMBER CHAIR over the base; ③ MathVista points lost on image-requiring questions 0–299.

yet drops 12.97 points; the model whose weights were moved back toward the base (§6) writes 922 words and drops only 2.93. Writing more does not mean seeing worse.

Takeaway. The cost has a location: it falls on the image-requiring questions, and on one set that the base gets at a glance; the student does not fail to see the image but cannot use it; talking at length brings hallucination, but the perception cost shows no resolvable relation to how long it writes.

5 Levers One and Two: The First Two Costs Both Come from Length

Each lever mainly moves its own part of the cost (Figure 3). The first two parts both come from the student writing too long, and this section covers them; the third is left to §6.

5.1 Half the Gap Is Locked in Answers Never Stated

Of the three, the first is the cheapest. When the student hits the budget limit, the whole reasoning is often there, missing only the last sentence of conclusion. Write r for the response a model produces within the budget B and E for symmetric extraction; forcing takes effect only when r hits the budget:

$$\hat{y} = \begin{cases} E(r) & \text{if } r \text{ ends within } B, \\ E\big(r \oplus s \oplus g_{64}(r \oplus s)\big) & \text{otherwise,} \end{cases} \tag{4}$$

where s is one fixed closing phrase shared by all five benchmarks (Appendix A.1), and g_{64} is the at most 64 tokens the model continues after it. No retraining, no added data, not a single weight touched. On MathVista, this brings the student from 42.67% back to 52.67%, +10.00 points item-paired (interval $[+4.33, +15.67]$), and not one response is left without an extractable answer. Of the 55 questions it turns right, 16 had no answer at all before; the rest changed their answer when forced

to conclude at the budget (Appendix D.1). The forced run also has a smaller budget (1,536 against 2,048), which alone does not move accuracy (§4.1). Taking the base as reference, the gap shrinks from 19.00 points to 9.67 (both sides unforced, then both forced)—half of the gap closes with no change in ability, once an answer is forced at the budget. The lever is picky, and that is its value: on a model that finishes speaking, it moves MathVista by $+0.67$ and MMStar by -0.33. So it can stay on: where there is something to recover it recovers half, and elsewhere it barely moves.

5.2 Hallucination Follows Word Count, and Word Count Is Given by the Teacher

The student hallucinates more than the base on AMBER: CHAIR rises from 15.38 to 23.64, a paired difference of $+8.26$ points (interval $[+7.13, +9.38]$), and its responses are much longer. More hallucination and longer responses are directly linked. Looking across, the seven students of Table 2 hallucinate more the longer they write ($\rho = +0.893$); looking along, the same model hallucinates less as soon as it writes a few words fewer (for all six models of four families; Appendix D.2); and the student–base pair is the same: with both generation budgets pressed down to 110 tokens, the student's extra hallucination is only $+2.48$ points ($[+1.36, +3.59]$), and the rest is brought by talking more. Who gives this length? Word count is given by the teacher. The first four students of Table 2 differ only in the teacher: the four teachers' trace lengths and the students' AMBER word counts are in exactly the same order ($\rho = +1.000$, $n = 4$), and changing the teacher moves the student's CHAIR by 8.62 points, more than the whole cost. For the student brought up by the reasoning teacher with the shortest traces (R1-Distill-Llama-8B (DeepSeek-AI et al., 2025)), the cost falls from 8.26 points to -0.36, with the interval $[-1.55, +0.77]$ enclosing zero; on the image-requiring questions it drops 4.60 points more than our student, which is not resolvable (Table 22).

Is it the teacher's visual training? No: for two teachers that differ only in visual training, the gap between their students cannot be separated from zero once both are cut to 110 tokens (Appendix D.2). The advantage the visually trained teacher enjoys is mostly that it talks less, not its visual training. Nor can cutting traces replace it: students of truncated traces hallucinate more (Table 2).

Takeaway. This cost is not an inherent property of distillation; it is chosen along with the teacher. Removing it takes no new architecture or data, only the same SFT run with a teacher that talks less.

6 Lever Three: One Scalar Removes the Perception Cost

Neither of the first two levers reaches the perception cost: it shows no resolvable relation to length (§4.2). But this cost was caused by the cold start's weight update, so we go back into the weights to find it: mixing the weights before and after distillation by one scalar, with no retraining. Moving the student's weights back toward the base by one scalar is what we call behavior-preserving interpolation (BPI). Interpolation itself is an existing tool, and reasoning is known to change in stages along it (Wu et al., 2025); what we add is picking the point by where the cost lands.

6.1 Along the Line, the Perception Cost Is Recovered First and the Reasoning Habit Returns Later

Every parameter, the vision tower included, is averaged with the same ratio λ, the weight on the student's side ($\lambda = 0$ is the base, $\lambda = 1$ the student):

$$\theta_\lambda = \theta_{\text{base}} + \lambda\,(\theta_{\text{cs}} - \theta_{\text{base}}) \tag{5}$$

The reasoning habit below is response length; it buys no accuracy (§4.1), but it is where the subsequent on-policy distillation starts (Bousselham et al., 2025). We take one model each at $\lambda = 0.25$, 0.50, and 0.75 and measure them under the protocol of Table 1; the line is drawn in Figure 11 (Appendix), with its numbers in Table 2. *The perception cost is recovered early.* On MathVista, every one of the three interpolated models is more accurate than the student, and between them and the base, all three show no resolvable difference. This is already so at $\lambda = 0.25$. *The reasoning habit comes much later.* The median response at $\lambda = 0.25$ is only 66 words: its accuracy is already back at the base's level, but its way of writing is still almost the base's. Stopping here, the perception cost is recovered, but the reasoning habit learned by distillation is lost along with it; only $\lambda = 0.75$ comes

back to 922 words. *So the rule for picking the point is not "which recovers the most accuracy"*: all three recover it. The rule takes the point closest to the student that is resolvably above the student and not resolvably below the base on image-requiring questions; here that is $\lambda = 0.75$, which also keeps the habit. On a selection set of 300 further MathVista questions, however, $\lambda = 0.75$ is not above the student (-1.67) and the rule picks $\lambda = 0.50$. We evaluated both on a test set of 400 questions used for no choice (Appendix E.1): $\lambda = 0.50$ is 8.00 points above the student ($[+3.00, +13.25]$) with no resolvable perception cost, but keeps only 18% of the habit (194 words); $\lambda = 0.75$ keeps 84% and is 5.00 above the student ($[+0.50, +9.50]$), also with no resolvable cost. We call the first the rule-chosen point and the second the behavior-preserving point; §6.2–6.3 analyse the latter.

Takeaway. The field has always tied these two together: if the student is to learn the teacher's reasoning habit, the perception cost must be accepted. Were they tied, both would retreat together toward the base; in fact the perception cost is recovered first while the reasoning habit remains. **The two can be separated, and separating them takes only adjusting one number, with no retraining.**

6.2 The Recovered Points Land on the Broken Questions

What exactly did $\lambda = 0.75$ do? First read it the way the field reads now (the last four columns of Table 1): on the average over a whole benchmark, it is above the student on MathVista, LogicVista, and WeMath, all three resolvable; against the base, all five benchmarks show no resolvable difference: on MathVista it is only 1.00 point lower, while the student is 9.67 lower. By this instrument, only one conclusion can be drawn—it undid the cold-start step. Yet it keeps 75% to 96% of the student's response length, and a model whose training had been undone would not look like this. The average cannot read what it actually moved. *On which questions did it return the points?* Apply the stratification of §4.2 unchanged to $\lambda = 0.75$ (Figure 2, right). On MathVista's 239 image-requiring questions, $\lambda = 0.75$ is 9.62 points above the student; on the 61 text-solvable ones no difference is measured. And where the student is 12.55 points below the base on the image-requiring questions, $\lambda = 0.75$ is only 2.93 below ($[-9.21, +3.35]$, no resolvable difference from the base), less than a quarter of the original remaining. On MathVista, the recovered points are resolvable only on the image-requiring stratum, though the difference between strata is not; on other benchmarks averaging brings back more than the cost, and on MMStar the recovery is not resolvable (Appendix E). Where it lands shows at the level of single questions. *Could it just be a little better everywhere?* The 28 broken questions of §4.2 allow a direct check. The base is at 100% by definition, the student at 7%, $\lambda = 0.75$ at 46%, $\lambda = 0.50$ at 50%, $\lambda = 0.25$ at 71%—the further toward the base, the more comes back; while the benchmark average cannot tell these three points from the base, hiding the cost that remains here. As a control, take the 67 questions that no student broke: $\lambda = 0.75$, 0.50, and 0.25 are at 97%, 91%, and 93%. One side rises from 7% to 46%, the other falls only three points from 100%: what this scalar restores is the set of broken questions, not all questions lifted together. The perception cost has an address, and the way to remove it recognizes this address too.

Takeaway. The same model and questions, different instruments: the benchmark average reads "it did nothing", the repaired instruments "it gave back the questions the cold start broke".

6.3 The Perception Cost Leaves, and Most of the Hallucination Stays

If the perception cost and hallucination were the same thing, the two should improve together toward the base. On AMBER, only one of the two came back: the CHAIR of $\lambda = 0.75$ is 7.05 points above the base ($[+5.97, +8.12]$), against 8.26 for the student—at this point it already shows no resolvable difference from the base on the image-requiring questions, yet 85% of the hallucination increase remains; it comes down only at $\lambda = 0.50$, and at $\lambda = 0.25$ it shows no resolvable difference from the base. Length falls with it (Table 2). $\lambda = 0.75$ is the one picked to keep the reasoning habit: keep the length, and the hallucination is kept. The correlation of §4.2 is observation; this is intervention—the same scalar takes the perception cost away and leaves 85% of the hallucination. Moving hallucination takes a change of teacher (§5.2); at the point that keeps the reasoning habit, changing the coefficient removes only 15% of it. Three levers, each mainly moving one thing, the third at the behavior-preserving point (Table 22): the whole evidence for "the cost is three things".

7 Related Work

Reasoning length and accuracy. On text-only tasks, longer reasoning can lower accuracy, and models abandon answers they had right (Gema et al., 2025; Liu et al., 2024; Zhou et al., 2026); §4.1 sees the abandoning happen on multimodal tasks, offset by as many wrong answers turning right. Kaiser et al. (2026) split "whether the answer is completed" out of accuracy on text tasks, and our first cost is this term in multimodal distillation; removing it uses the budget forcing of s1 (Muennighoff et al., 2025), only its cutting end: not to compute more, but to let the student say the answer it never stated. **Hallucination.** The longer the reasoning chain, the more the hallucination (Liu et al., 2025a; Zheng et al., 2025); overly detailed training targets keep a model writing past what it sees (Yue et al., 2024); long teacher traces teach talkative students (Griot et al., 2026), and small students learn better from shorter ones (Li et al., 2025). §5.2 joins these into one chain. Evaluation usually controls length away (Yan et al., 2024; Wei et al., 2024; Zhai et al., 2023a): right for comparing models' tendencies, but for what a recipe costs, it controls away what is measured.

Perception cost. Text-only cold starts are also reported to help multimodal reasoning (Chen et al., 2025b; Liu et al., 2025b); with the instruments repaired, VOLD's cold start shows no measurable gain. That the student sees worse is known: VOLD (Bousselham et al., 2025) itself reports its MMStar perception score falling from 58.4 to 54.4, a "perception-reasoning trade-off" long documented; even with the vision encoder frozen, fine-tuning a multimodal model loses visual ability (Zhai et al., 2023b), and on-policy distillation loses visual grounding, restored by gradient steering with retraining (Yoon et al., 2026). The usual explanation is that a long reasoning chain stops looking at the image (Yu et al., 2026; Tian et al., 2025; Sun et al., 2025; Jin et al., 2026), but the perception cost shows no resolvable relation to how long the student writes (§4.2). These works report only its size, not where it falls, and their remedies all retrain (Yu et al., 2026; Tian et al., 2025; Wu et al., 2026). Answering without the image is not new: MMStar and MMMU-Pro (Chen et al., 2024; Yue et al., 2025) use it to build vision-dependent benchmarks, MathVerse (Zhang et al., 2024) and Kancheti et al. (2026) to expose reliance on text, describing a model or benchmark; we use it for attribution, giving the cost an address (image-requiring questions) and a retraining-free remedy (one scalar, §6).

Weight interpolation. Averaging fine-tuned weights back toward the start is a known way to recover what fine-tuning lost (Wortsman et al., 2022; Ramé et al., 2024; Lin et al., 2024; Farn et al., 2025), or scaling task vectors (Ilharco et al., 2023; Wang et al., 2025); interpolating between an Instruct and a Thinking model, reasoning likewise appears in stages along the coefficient (Wu et al., 2025). Closest to ours is Chen et al. (2025a): no distillation, directly merging a reasoning LLM's weights into a VLM. We use the simplest, across modalities, choosing the point on a selection set (§6).

8 Discussion and Conclusion

Whether these three levers carry over to other recipes. It depends on whether the cause each acts on is specific to this student. The first two are not. Where a model stops is set by the decoding budget, which every reason-then-answer evaluation runs into; how long the student writes is set by the teacher's traces, which every distillation on ready-made traces inherits; and both numbers are in the recipe author's hands. The third is different: it works only if this training did not go far in weight space, and this distance can be measured. The student's language part comes from a text-only model that could not see images; its ability to see was added by one transformation, which moved the weights by 38%, 35 times as far as instruction tuning of the same backbone; our cold start moved them only 3.5%, so the precondition is met. To use BPI on another recipe, measure this number first.

What else this split implies. Two untested things should follow. *One*, BPI can be applied unchanged to the RL checkpoint: RL adds no resolvable cost to the cold start and moves less than a twentieth as far, orthogonally (Appendix C.5). *Two*, with a shorter-trace teacher and the weights averaged back, both costs can go unpaid together: no coupling between the two is resolvable (§4.2).

Under repaired instruments, distilling a text-only teacher into a VLM buys no measurable accuracy at the cold start, and its cost falls on image-requiring questions. This cost has been taken as the price of reasoning, yet each of its three parts comes off on its own—a forced conclusion, a shorter-trace teacher, one weight-averaging ratio: none of them is what reasoning itself costs.

AI USE STATEMENT

We conceived the research questions, designed the experiments, and formed the hypotheses, claims, and interpretations in this paper. We made the methodological decisions, drew Figures 1 and 3, and take full responsibility for the content. We used generative AI tools as assistants in two places. In the experiments, we used an AI coding assistant to help write evaluation, analysis, and plotting scripts, which we reviewed and tested, and we recomputed every reported number from raw model outputs. In writing, we laid out the claims, the evidence for each, and the structure in a written plan, used an AI assistant to help turn that plan into prose, and then reviewed and revised every sentence ourselves.

REPRODUCIBILITY STATEMENT

§2 states the models, benchmarks, and statistical conventions that govern every comparison, and §3 the three instruments; Appendix A.1 gives the training and evaluation configuration and Appendix A.2 the statistical details. Code, the evaluation protocol, all raw predictions, and the interpolated checkpoints will be released.

ETHICS STATEMENT

This work evaluates object hallucination and reasoning accuracy in vision-language models on public benchmarks (AMBER, POPE, MathVista, MMStar, LogicVista, WeMath, MathVision) and involves no human subjects and no new data collection. Our findings concern measurement practice: commonly reported hallucination and accuracy numbers depend on response length and on the answer extractor, and uncorrected they can overstate or understate the reliability of deployed systems.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Arnav Bendre, Guneesh Gupta, Kavish Grover, Chayan Aggarwal, and Shreyansh Modi. Rethinking CD: A Reproducibility Study and Extension on the Ineffectiveness of Contrastive Decoding at Mitigating Object Hallucinations in MLLMs. *arXiv preprint arXiv:2607.25196*, 2026.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. VOLD: Reasoning Transfer from LLMs to Vision-Language Models via On-Policy Distillation. *arXiv preprint arXiv:2510.23497*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, et al. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*, 2024.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring Reason to Vision: Understanding Perception and Reasoning through Model Merging. In *International Conference on Machine Learning (ICML)*, 2025a. arXiv:2505.05464.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, et al. Advancing Multimodal Reasoning: From Optimized Cold Start to Staged Reinforcement Learning. *arXiv preprint arXiv:2506.04207*, 2025b.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hua Farn, Hsuan Su, Shachi H Kumar, Saurav Sahay, Shang-Tse Chen, and Hung-yi Lee. Safeguard Fine-Tuned LLMs Through Pre- and Post-Tuning Model Merging. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2412.19512.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, et al. Inverse Scaling in Test-Time Compute. *arXiv preprint arXiv:2507.14417*, 2025.

Maxime Griot, Paul Steven Scotti, and Tanishq Mathew Abraham. Compress-Distill: Reasoning Trace Compression for Efficient Knowledge Distillation. *arXiv preprint arXiv:2606.05988*, 2026.

Joseph Hoche, David Brellmann, and Gianni Franchi. Leveraging Visual Signals for Robust Token-Level Uncertainty in Vision-Language Generation. *arXiv preprint arXiv:2605.27136*, 2026.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

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 (ACL)*, 2026. arXiv:2606.22565.

Daniel Kaiser, Arnoldo Frigessi, Ali Ramezani-Kebrya, and Benjamin Ricaud. Beyond Accuracy: Decomposing the Reasoning Efficiency of LLMs. *arXiv preprint arXiv:2602.09805*, 2026.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian, and Tanuja Ganu. Chain-of-Thought Degrades Visual Spatial Reasoning Capabilities of Multimodal LLMs. *arXiv preprint arXiv:2604.16060*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2305.10355*, 2023.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small Models Struggle to Learn from Strong Reasoners. *arXiv preprint arXiv:2502.12143*, 2025.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the Alignment Tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2309.06256.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More Thinking, Less Seeing? Assessing Amplified Hallucination in Multimodal Reasoning Models. *arXiv preprint arXiv:2505.21523*, 2025a.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, et al. X-Reasoner: Towards Generalizable Reasoning Across Modalities and Domains. *arXiv preprint arXiv:2505.03981*, 2025b.

Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse. *arXiv preprint arXiv:2410.21333*, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, et al. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, et al. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. ALOHa: A New Measure for Hallucination in Captioning Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. arXiv:2404.02904.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. WARP: On the benefits of weight averaged rewarded policies. *arXiv preprint arXiv:2406.16768*, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. arXiv:1809.02156.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating Visual Forgetting via Take-along Visual Conditioning for Multi-modal Long CoT Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2503.13360.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More Thought, Less Accuracy? On the Dual Nature of Reasoning in Vision-Language Models. *arXiv preprint arXiv:2509.25848*, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, et al. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. *arXiv preprint arXiv:2402.14804*, 2024.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training Layer Scaling Prevents Forgetting and Enhances Model Merging. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.17146.

Hongliang Wei, Xingtao Wang, Xianqi Zhang, Xiaopeng Fan, and Debin Zhao. Toward a Stable, Fair, and Comprehensive Evaluation of Object Hallucination in Large Vision-Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust Fine-Tuning of Zero-Shot Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2109.01903.

Juncheng Wu, Hardy Chen, Haoqin Tu, Xianfeng Tang, Freda Shi, Hui Liu, Hanqing Lu, Cihang Xie, and Yuyin Zhou. From Seeing to Thinking: Decoupling Perception and Reasoning Improves Post-Training of Vision-Language Models. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.20177.

Taiqiang Wu, Runming Yang, Tao Liu, Jiahao Wang, and Ngai Wong. Revisiting Model Interpolation for Efficient Reasoning. *arXiv preprint arXiv:2510.10977*, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Bei Yan, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. Measuring the measurers: Quality evaluation of hallucination benchmarks for large vision-language models. *arXiv preprint arXiv:2406.17115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, et al. Decomposed On-Policy Distillation for Vision-Language Reasoning: Steering Gradients for Visual Grounding. *arXiv preprint arXiv:2606.00564*, 2026.

Seonghoon Yu, Dongjun Nam, Byung-Kwan Lee, and Jeany Son. Hide to See: Reasoning-prefix Masking for Visual-anchored Thinking in VLM Distillation. *arXiv preprint arXiv:2605.11651*, 2026.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2409.02813.

Zihao Yue, Liang Zhang, and Qin Jin. Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.14545.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. HallE-Control: Controlling Object Hallucination in Large Multimodal Models. *arXiv preprint arXiv:2310.01779*, 2023a.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the Catastrophic Forgetting in Multimodal Large Language Models. *arXiv preprint arXiv:2309.10313*, 2023b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2403.14624.

Ruipeng Zhang, Zhihao Li, C. L. Philip Chen, and Tong Zhang. Steer Where It Matters: Token-Level Visual-Sensitivity Steering for LVLMs Hallucination Mitigation. *arXiv preprint arXiv:2606.07647*, 2026.

Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibei Yang. Why LVLMs Are More Prone to Hallucinations in Longer Responses: The Role of Context. *arXiv preprint arXiv:2510.20229*, 2025.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When More Thinking Hurts: Overthinking in LLM Test-Time Compute Scaling. *arXiv preprint arXiv:2604.10739*, 2026.

A Setup Details

A.1 Configuration

Table 3 lists every setting the main text relies on. All capability claims concern the cold-start step, which was trained once.

A.2 Statistical Conventions

Every comparison is paired: two models are compared on the same items (accuracy) or on the same images (AMBER). Intervals are 95% percentile bootstrap intervals that resample items (3,000 resamples) or images (2,000). A difference is *resolvable* when its interval excludes zero, and we do not read an unresolvable difference as zero.

A CHAIR difference is the difference of two pooled CHAIR values, hallucinated objects over generated objects, each summed over the paired images. Every CHAIR difference in the paper is therefore the difference of two CHAIR values that appear in the tables. The three interpolation intervals of §6.3 resample 20 fixed blocks of images instead of single images, because those runs were scored block by block.

Permutation tests shuffle the group label 10,000 times and report the share of shuffles at least as extreme as the observed statistic. They are used for the response-length comparison of the damage set and for the ten-model rank correlation of §4.2.

Table 3: Training and evaluation configuration.

Models	
Base (student initialization)	Qwen2.5-VL-3B-Instruct
Teacher	Qwen3-8B, prompted with text only
Cold start (supervised fine-tuning)	
Prompts	open-r1/Mixture-of-Thoughts, text only
Teacher traces	52,028; median 2,997 tokens (2,047 words)
Trainer	verl 0.10.0.dev SFT trainer, full fine-tuning
Learning rate / global batch / max length	5×10^{-5} / 256 / 8,192 tokens
Epochs / steps / seed / runs	3 / 612 / 1 / 1
Vision tower	0 of its 390 parameter tensors differ from the base
Reinforcement learning (from the student)	
Prompts	ORZ-Math-57k: 56,878 text-only math problems
Algorithm / reward	GRPO / rule-based answer correctness
On-policy distillation	From the teacher, reverse KL on its top-64 tokens, coefficient 10^{-3}, conditioned on the task reward
Steps / prompts per step / rollouts per prompt	60 / 64 / 5
Actor learning rate / temperature	6×10^{-6} / 1.0
Clip ratio (low, high) / KL loss / entropy bonus	0.2, 0.3 / none / none
Max prompt / response length	1,024 / 7,168 tokens
Seeds	3
Vision tower	0 of 390 parameter tensors differ from the base, in all three checkpoints
Evaluation	
Decoding	Greedy
Accuracy benchmarks	MathVista, MMStar, LogicVista, WeMath, MathVision; the first 300 items of each (on MathVision, the 114 free-form items among them)
Generation budget	1,536 new tokens (the unforced student run of §4: 2,048)
Forced conclusion	If the budget is reached, append "`Wait, I have spent enough time. The final answer is \boxed{`" and generate up to 64 more tokens
Hallucination	AMBER generative split: 1,004 images, prompt "Describe this image.", 512 new tokens (110 in the budget-matched runs), not forced; descriptions within two tokens of the cap: base 4 of 1,004, student 394, R1-Distill-Llama-8B teacher 16, Qwen2.5-7B and Qwen2.5-VL-7B teachers 0, traces cut at 2,048 / 1,024 / 512 tokens 483 / 419 / 348, $\lambda = 0.25$ / 0.50 / 0.75 1 / 11 / 95; CHAIR with exact object matching
Hardware	2 × NVIDIA H200

B The Three Instruments

B.1 The Extractor

§3.1 reports that an extractor tuned to reasoning templates misread the base four times, each time in favor of the student. Table 4 lists the four cases. In each, the extractor read the student's `\boxed{}` answers but left a share of the base's direct answers unread, and an unread answer counts as wrong. Symmetric extraction reads both models with one fixed set of answer forms.

POPE (Li et al., 2023) shows how far such choices can move a result (Table 5). The base's 173 template responses split 87 *yes* and 86 *no* in the ground truth, so the template carries no answer; the student never produces it. The repaired extractor also reads `\boxed{Yes}` and a sentence-final *yes* or *no*, forms only the student writes. Depending on which choices a scorer makes, the same responses put the student anywhere from 8.33 points ahead of the base to 6.42 behind. POPE therefore carries no claim in this paper.

The second principle of symmetric extraction, that an answer must be stated rather than merely present, comes from an earlier MathVista scorer. It credited an answer whenever the normalized gold

Table 4: The same extraction defect, four times. Unread counts are out of 300 items (600 on POPE) under the extractor in use before the repair.

Benchmark	Base answers the extractor did not read	Unread (base / student)	Effect of symmetric extraction
MathVista	Chinese option markers; spelled-out numbers (*Zero*, *Two*); "answer is: (B)"	—	Base (unforced) 58.67 → 61.67; student (forced) 52.67 → 52.67
POPE	One refusal template, "there is no existence of X in the image description"	173 / 69	Student − base: −6.42 → −2.76 (Table 5)
LogicVista	Direct answers in forms not on the extractor's list	57 / 3	Base 33.00 → 39.33; student 35.33 → 35.33. A +2.33 gain becomes −4.00
WeMath	The same; typically a direct answer followed by "the correct answer is: D"	105 / 1	Base 41.67 → 52.00; student 52.00 → 52.00. A +10.33 gain becomes 0.00

Table 5: POPE, the same two sets of responses scored six ways (600 items). *Parsed only* drops unread answers from the denominator; *all items* counts them as wrong. Each row adds one change to the row above.

	Parsed only			All items		
Reading	Base	Student	Δ	Base	Student	Δ
Extractor before repair	59.72	53.30	−6.42	42.50	47.17	+4.67
+ repaired extractor	59.72	54.08	−5.64	42.50	50.83	+8.33
+ template read as *no*	56.83	54.08	−2.76	56.83	50.83	−6.00

string appeared anywhere in the response, so a gold answer of "1" matched any "1" in a reasoning trace. On the base, 13.5 of the 40.9 accuracy points it received under that scorer, a third, came from this containment alone.

The fixed set of accepted answer forms is `\boxed{}`, option letters, Chinese option markers, spelled-out numbers (*Zero*, *Two*), "answer is (B)", and "the correct answer is: D". The last answer stated in the response is taken, and substring containment is never used.

VOLD's conclusion that the student has no hallucination cost on HallusionBench holds only when "counting only the responses that can be read" (−0.12); scored under the convention it was published with—unreadable counts as wrong—the gap is −1.68, and these 1.68 points come entirely from the student's unreadable responses. The student writes long, is truncated more, and has more unreadable responses.

B.2 The Matcher

AMBER's reference implementation decides whether a generated word names an annotated object by word-vector similarity with a threshold of 0.8 (its default). Under that rule a hallucinated *dog* is excused whenever the image contains a *cat*, since the two words have a similarity of 0.802. Our rule lets an exact match decide first and falls back to similarity (at 0.75) only for words on neither list. The fallback decides at most 16 of the 3,644 to 7,134 object words a model generates; setting it to 0.8 as well moves no model's CHAIR by more than 0.18, and the student's excess over the base from 8.26 to 8.29. What separates the two rules is the order of the checks, not the threshold. Open-vocabulary metrics such as ALOHa (Petryk et al., 2024) likewise match objects by semantic similarity. Table 6 scores the same responses under both rules.

B.3 Blind-Teacher Stratification

Figure 4 sketches the procedure. The blind teacher answers 20.33% of MathVista, 32.33% of LogicVista and 30.67% of WeMath correctly, above the chance rate of guessing among the offered

Table 6: CHAIR under AMBER's similarity matching and under exact matching, on the same responses to all 1,004 AMBER images. Similarity matching lowers CHAIR by 3.64 to 4.25 points on every model, while CHAIR itself ranges from 12.27 to 24.34, so the offset is close to a constant: the student's excess over the base is 8.48 under similarity matching and 8.26 under exact matching. Both rules see the same generated objects, so coverage does not change.

	Objects	Hallucinated objects		CHAIR	
Model	generated	similarity	exact	similarity	exact
Qwen2.5-VL-3B (base)	4,778	532	735	11.13	15.38
Student	6,209	1,218	1,468	19.62	23.64
Qwen2.5-VL-7B	4,572	384	561	8.40	12.27
LLaVA-1.5-7B	3,644	473	620	12.98	17.01
SmolVLM	5,627	1,061	1,266	18.86	22.50
InternVL2.5-8B	6,969	1,423	1,696	20.42	24.34

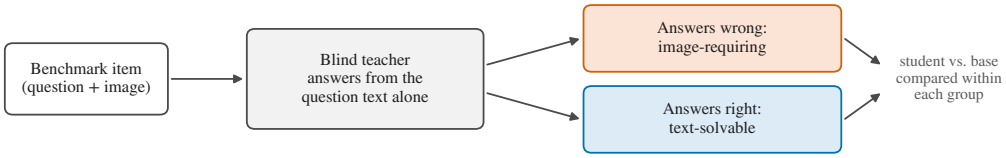


Figure 4: Blind-teacher stratification. The teacher answers every question from its text alone. Items it answers wrong are image-requiring; items it answers right are text-solvable. The two groups are compared separately (§4.2). Splits: MathVista 239/61, MMStar 241/59, LogicVista 203/97, WeMath 208/92, fixed before the comparisons of the main text and shared by every model.

options (17.67%, 23.21% and 22.23%; free-form items count as zero). On MMStar, a benchmark built so that text alone cannot answer it, it answers 19.67%, below the 25% chance of four options: there the text-solvable stratum is mostly items the teacher guessed right. Table 7 tests the difference between the two strata directly, shuffling the stratum labels 10,000 times. Under these labels, the student's cost is resolvably larger on image-requiring items on MMStar and marginal on MathVista; on LogicVista and WeMath, where the student has no resolvable cost in either stratum, there is no difference. The recovery at $\lambda = 0.75$ is not resolvably concentrated by stratum on any benchmark; its concentration shows at the level of single questions (§6.2).

Table 7: Blind-teacher strata: the teacher's blind accuracy against chance, and a permutation test of the difference between strata (image-requiring minus text-solvable, in points).

	Blind teacher			Student − base		$\lambda = 0.75$ − student	
Benchmark	Accuracy	Chance	Img / txt	Difference	p	Difference	p
MathVista	20.33	17.67	239 / 61	−14.19	0.09	+4.71	0.56
MMStar	19.67	25.00	241 / 59	−18.92	0.004	−3.01	0.74
LogicVista	32.33	23.21	203 / 97	−1.34	0.92	+1.07	0.91
WeMath	30.67	22.23	208 / 92	+3.14	0.73	−7.02	0.32

Relabeling with a larger budget. The blind teacher answers from a plain prompt, greedily, with at most 1,024 new tokens, well below the length of its own traces (Table 3). Rerun with these settings, 274 to 280 of the 300 answers on each benchmark's questions 0–299 reach that budget, so an image-requiring item is, in practice, one the teacher did not answer correctly within it. We therefore relabel every item with the teacher's own chat template, thinking enabled, and 8,192 tokens, where at most 34 answers per benchmark reach the budget. Between 18% and 30% of items change stratum (Table 8). The MathVista result is unchanged: the cost stays on image-requiring items and $\lambda = 0.75$ recovers it there. On LogicVista the same hedging appears as on MMStar, a resolvable cost on image-requiring items offset on text-solvable ones. On MMStar the cost on image-requiring

items remains resolvable, but the difference between strata becomes marginal. We chose this setting in advance because it is how the teacher reasons when it writes the distillation traces. Two further settings bracket it: the original settings rerun (greedy decoding reproduces 85–92% of the original labels, since batching changes the arithmetic) and the chat template without thinking at 4,096 tokens. Under the latter the MMStar cost on image-requiring items is not resolvable (-4.25 $[-9.27, +0.77]$) and LogicVista shows no hedging (between strata $p = 0.58$). Only the MathVista result holds under all four labelings: a resolvable cost on image-requiring items (-12.55 to -13.43), none on text-solvable ones, and a resolvable recovery by $\lambda = 0.75$ there ($+8.44$ to $+11.64$); the difference between its strata is never resolvable (p from 0.09 to 0.14).

Table 8: Blind-teacher strata before and after relabeling (teacher's chat template, thinking enabled, 8,192 tokens), questions 0–299 of each benchmark. Changed: items that move between strata. p: permutation test of the difference between strata. $\star$: interval excludes zero.

	Img / txt		Student − base, image-requiring		p	$\lambda = 0.75$ − student,
Benchmark	(relabeled)	Changed	before	after	before → after	image-req. (after)
MathVista	189 / 111	80	$-12.55^\star$	-13.23 $[-20.63, -5.29]^\star$	$0.09 \to 0.14$	$+11.64$ $[+4.23, +19.05]^\star$
MMStar	236 / 64	53	$-7.05^\star$	-5.93 $[-11.86, -0.42]^\star$	$0.004 \to 0.05$	$+3.81$ $[-1.69, +9.32]$
LogicVista	193 / 107	90	-4.43	-11.92 $[-20.21, -4.15]^\star$	$0.92 \to 0.003$	$+9.33$ $[+1.55, +17.10]^\star$
WeMath	173 / 127	87	$+0.96$	-0.58 $[-9.25, +8.09]$	$0.73 \to 0.91$	$+4.62$ $[-2.89, +12.14]$

C Where the Cost Is

C.1 Budget Sweep

§4.1 varies the student's generation budget on MathVista from 64 to 1,536 tokens, forcing a conclusion at each (Table 9). The last column compares the student at each budget with the base on image-requiring items. The base is evaluated once, at 1,536 tokens; its answers are short, and 3 of its 300 MathVista responses reach the budget.

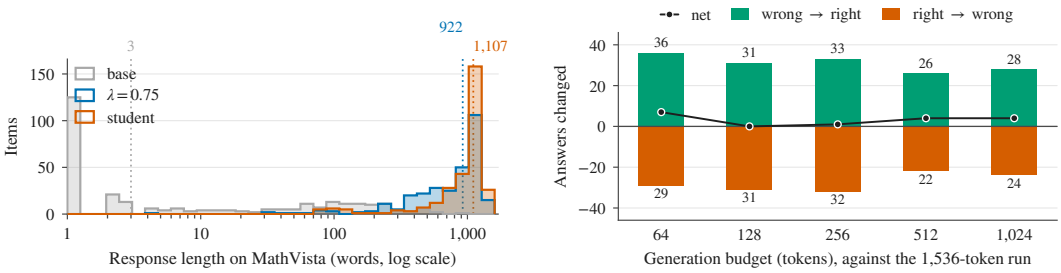
The cost on image-requiring items is resolvable at every budget, between -10.04 and -12.55 points: shortening the student's responses tenfold does not remove it. Between the 64- and 1,536-token levels only sixty percent of the questions keep the same answer: 36 turn right and 29 turn wrong. Figures 5a and 5b plot the response lengths and the answers that change with the budget.

Table 9: MathVista, student at six generation budgets ($n = 300$; 239 image-requiring items). Δ is item-paired against the 1,536-token run. Last column: student minus base on image-requiring items. $\star$: interval excludes zero.

Budget	Acc.	Median words	Δ vs. 1,536	Turned right / wrong	Cost, image-requiring
64	55.00	106	$+2.33$ $[-3.00, +7.33]$	36 / 29	-10.04 $[-16.32, -3.77]^\star$
128	52.67	154	$+0.00$ $[-5.00, +5.33]$	31 / 31	-11.72 $[-17.99, -5.44]^\star$
256	53.00	251	$+0.33$ $[-4.67, +5.67]$	33 / 32	-10.88 $[-17.57, -4.18]^\star$
512	54.00	439	$+1.33$ $[-3.33, +5.67]$	26 / 22	-11.30 $[-17.57, -5.02]^\star$
1,024	54.00	787	$+1.33$ $[-3.33, +6.00]$	28 / 24	-11.72 $[-18.41, -5.44]^\star$
1,536	52.67	1,107	—	—	-12.55 (Figure 2)

C.2 Copying the Template

We truncated the teacher's traces at three different lengths and trained one student on each; the truncated traces stop in the middle of a sentence, with no closing tag and no answer (Table 10). The base never opens `<think>`, its median response is 3 words, and it is the row with the highest accuracy in the table. The four distilled students open `<think>` in all 300 responses; the opening format is copied one hundred percent. After that they part ways: with the teacher's traces cut at 512 tokens, the student closes the tag in only 4 responses and gives a `\boxed{}` answer in none, with accuracy 22.00%; cut at 2048, 296 give an answer, almost the same as the 297 with no truncation at all. The cutting of the teacher's traces was originally an accident of the training budget, yet the student took the cut as part of the format.



(a) Response length.

(b) Answers that change with the budget.

Figure 5: The reasoning habit, and what it does to answers. **(a)** On MathVista the base answers in a median of 3 words, the student writes 1,107, and $\lambda = 0.75$ writes 922 (§4.1). **(b)** Against the 1,536-token run, each smaller budget turns between 26 and 36 of the student's answers from wrong to right and between 22 and 32 from right to wrong. The accuracy moves by at most 2.33 points while the individual answers do not hold still (§4.1).

Table 10: Students trained on teacher traces truncated at different lengths (MathVista, 300 questions).

Model	Median words	Accuracy	Opens `<think>`	Closes `</think>`	Gives `\boxed{}`
Base	3	62.33	0	0	19
Truncated at 512	343	22.00	300	4	0
Truncated at 1024	764	36.00	300	43	28
Truncated at 2048	1,170	53.33	300	133	296
Not truncated (our student)	1,107	52.67	300	134	297

C.3 Two Items from the Damage Set

Figure 6 places every MathVista item for every model, and Figure 7 shows two items of the damage set of §4.2 with the base's and the student's answers verbatim.

C.4 Eight Explanations, Each Excluded

Table 11 lists the eight explanations of §4.2 and the evidence against each. The paragraphs below describe the three tests not reported elsewhere; parameter drift and position effects follow in Appendices C.5 and C.6.

Images in the training data. We built a second cold-start set in which every example carries an image: the student answered Geometry3K questions, and 997 answers that were correct and followed the teacher's structure were kept. A model trained on this set from the base never receives a text-only gradient. On image-requiring MMStar items it loses 7.05 points ($[-13.69, -0.41]$), the same net loss of 17 items as the text-only student, reached through different items: the two disagree on 72 of the 241. Its MathVista accuracy is 28.67, so the set fails as a recipe and serves only as a control.

Attention to image tokens. Forward hooks on all 36 layers record the share of attention that generated tokens place on image tokens, over 200 items. The share rises from 4.81% in the base to 5.05% in the student. It moves between depths: up 33% in the middle third of the layers, down 21% in the last third (Figure 8).

Deleting the image. Attention is correlational, so we also test dependence causally. Each model describes each AMBER image (the first 128 generated tokens); we then score that same text with and without the image and measure how far the next-token distributions move (mean KL divergence) and how much the log-probability of the generated tokens drops. Both move more for the student than for the base: KL $0.566 \to 0.686$ (difference $[+0.11, +0.13]$) and log-probability drop $0.608 \to 0.733$ (difference $[+0.112, +0.137]$), paired over 1,004 images. The difference is positive in each of four position bands (tokens 0–24, 25–49, 50–99, 100–127). The student uses the image more, not less.

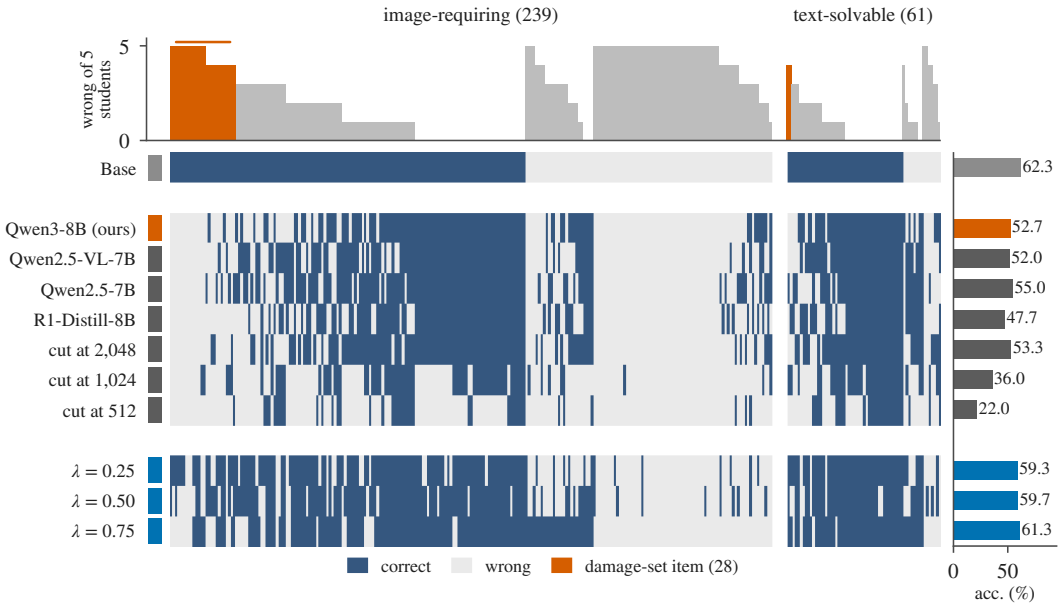
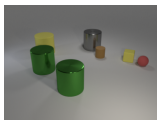


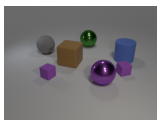
Figure 6: Every MathVista item (columns) for every model (rows); dark cells are correct answers. Columns are split into image-requiring and text-solvable items. Within each block, items the base answers correctly come first, ordered by how many of the five students that define the damage set get them wrong (top bars; these five are all students except the two whose teacher traces were cut at 512 and 1,024 tokens), followed by items the base gets wrong. The 28 damage-set items (vermilion; 26 image-requiring, 2 text-solvable) sit at the left edge: the base answers every one, the students fail most of them, and the interpolated models answer more of them the closer they are to the base (71%, 50% and 46% at $\lambda = 0.25$, 0.50 and 0.75). Right: accuracy of each row.



Subtract all yellow matte blocks. Subtract all tiny brown cylinders. How many objects are left? (gold: **5**)

base **5** ✓ 1 word

student Okay, let's see. The question is asking me to subtract all yellow matte blocks and all tiny brown cylinders … This results in **three objects** remaining. $\boxed{3}$ × **913 words**



Subtract all big red metallic spheres. Subtract all big brown matte things. How many objects are left? (gold: **6**)

base **6** ✓ 1 word

student Okay, let's see. The question is asking me to subtract all the big red metallic spheres … Wait, I have spent enough time. The final answer is $\boxed{5}$? Wait, but how can I be sure? Let me try to think again. … Then subtracting all big × **1,277 words**

Figure 7: Two of the 28 MathVista items that at least four of the five independently trained students answer wrong while the base answers them correctly (§4.2). Both are counting questions the base settles in one word. The student reasons for about a thousand words and settles on the wrong count. In the second item the student reaches the budget; the forced conclusion (§5.1; cue in grey) yields 5, and the student starts doubting it before its 64 tokens run out. Text is verbatim, elided at …

Table 11: Eight explanations for the perception cost, each tested against one direct measurement and not supported by it (§4.2).

Explanation	Evidence against
The training data contain no images	A student trained on 997 examples that all carry an image loses 7.05 points on image-requiring MMStar items ($[-13.69, -0.41]$), the same net loss of 17 items as the text-only student
The backbone reverts toward the text model	The cold start moves the language backbone by 0.0348, and its distance to the text model stays at 0.38 (0.3804 → 0.3812; Appendix C.5)
The vision encoder is damaged	None of its 390 parameter tensors changes, in the student or in any RL checkpoint; 433 of the 434 language tensors do
Attention to the image fades with generated length	The pooled rise of hallucination with position (4.56% to 15.87%) mostly disappears within groups of equal response length; image dependence fades by 74 to 87% from the first to the last position band in all five models measured, base included (Appendix C.6)
Long reasoning traces cause it	Across ten models, response length and the cost on image-requiring items show no resolvable relation ($\rho = -0.127$; Appendix C.7)
The model stops attending to the image	The share of attention on image tokens rises from 4.81% to 5.05% (36 layers, 200 items)
The model depends less on the image	Deleting the image moves the student's token distributions more than the base's (mean KL 0.566 → 0.686, log-probability drop 0.608 → 0.733; 1,004 AMBER images, both intervals excluding zero)
The teacher's lack of visual training transfers	The student of a visually trained teacher loses 12.97 points on image-requiring MathVista items, more than the 7.95 lost by the student of its text-only counterpart (Appendix C.7)

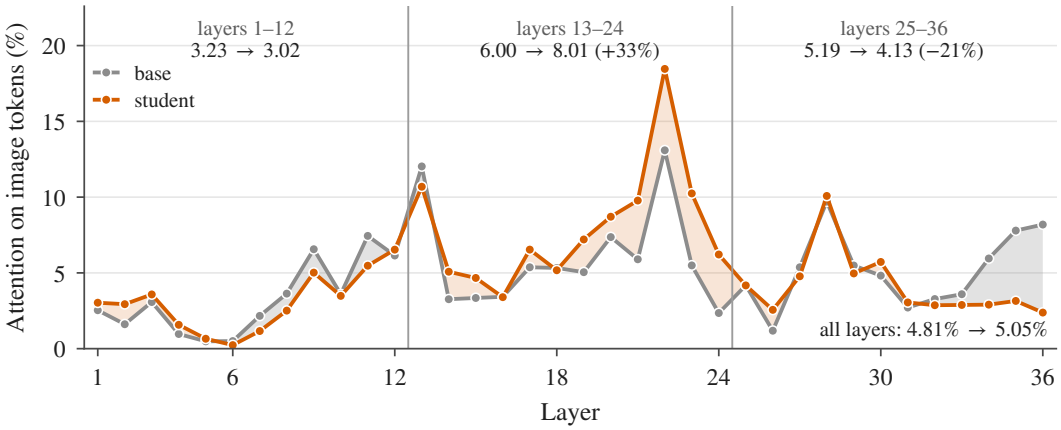


Figure 8: Share of attention that generated tokens place on image tokens, layer by layer, over 200 items. The student attends to the image more than the base in the middle layers and less in the last ones; over all layers the share rises from 4.81% to 5.05%.

C.5 Parameter Drift

We compare the 434 language-backbone parameter tensors that Qwen2.5-3B, Qwen2.5-3B-Instruct, Qwen2.5-VL-3B-Instruct and the student share, with the same names and shapes. For each tensor we take the relative distance $\|B-A\|_F/\|A\|_F$ and report the median over tensors (Table 12).

Multimodal alignment moved the backbone 35 times as far as instruction tuning. The cold start moves it 3.5%, under a tenth of that alignment step, and does not move it back toward the text model: the distance to Qwen2.5-3B-Instruct stays at 0.38. Per tensor, the RL step has a median norm 4.0% of the cold-start update's (the ratio of the two relative drifts in Table 12, 0.0017 to 0.0348, is 4.9%) and is orthogonal to it (median cosine over tensors −0.0004). During the cold start 433 of the 434 language tensors change, and none of the 390 vision-tower tensors do; the same holds for all three RL checkpoints, whose prompts contain no images. Figure 9 breaks the distance down by layer and weight matrix.

Table 12: How far each training step moves the language backbone: median relative distance over the 434 shared parameter tensors.

Step	From → to	Median distance
Instruction tuning	Qwen2.5-3B → Qwen2.5-3B-Instruct	0.0108
Multimodal alignment	Qwen2.5-3B-Instruct → Qwen2.5-VL-3B-Instruct	0.3804
Cold start (ours)	Qwen2.5-VL-3B-Instruct → student	0.0348
RL (ours, 3 seeds)	student → RL checkpoint	0.0017 / 0.0017 / 0.0017
Distance to the text model (Qwen2.5-3B-Instruct): base / student		0.3804 / 0.3812

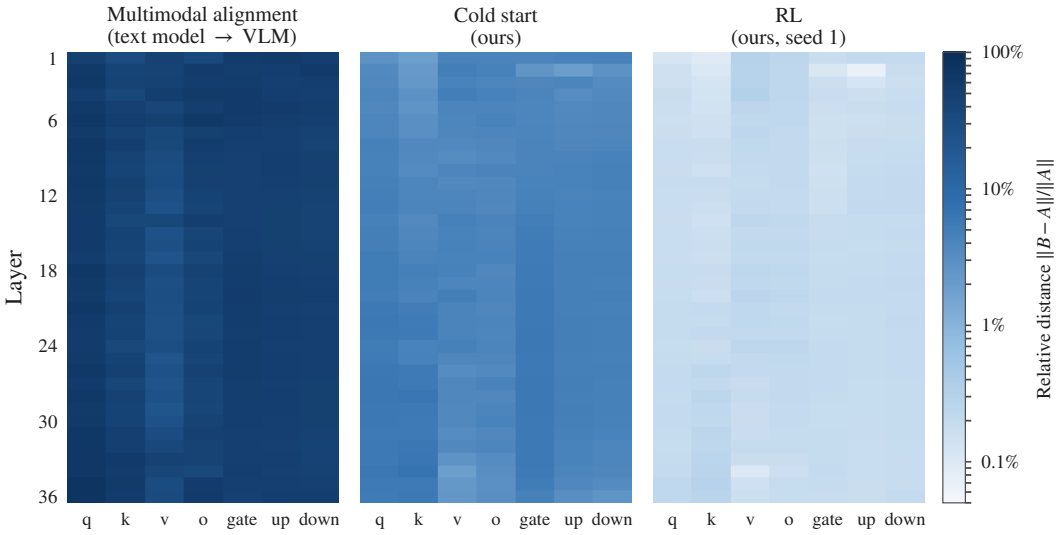


Figure 9: How far each training step moves each weight matrix of the language backbone, layer by layer (relative distance, log color scale; q, k, v, o: attention projections; gate, up, down: MLP). Multimodal alignment moves every matrix by 22 to 79%; the cold start moves each by 1.8 to 6.8%; RL moves each by less than a third of a percent. Medians over all 434 tensors are in Table 12.

C.6 Position Effects

Token position is a strong free signal for hallucinated object words, about as strong as the image-contrast signals (Appendix F), and it is tempting to read this as grounding fading with distance from the image. The data do not support that reading.

Pooled rates mix response lengths. On Qwen2.5-VL-3B the hallucination rate of object words rises with absolute token position: 4.56% at positions 0–39, 5.72% at 40–79, 9.45% at 80–119, 15.87% at 120–159. High positions exist only in long responses, and long responses hallucinate more throughout. Grouping responses by their number of object words removes most of the rise (Table 13). Within each group the trend is flat or non-monotonic, and the group of long responses is highest at the earliest positions, the opposite of what fading grounding predicts.

Table 13: Hallucination rate of object words by absolute token position, grouped by the response's number of object words (Qwen2.5-VL-3B, AMBER).

Position	Short (≤ 7)	Medium (8–12)	Long (≥ 13)
0–19	4.46%	4.52%	7.83%
20–39	4.97%	3.77%	3.85%
40–59	5.96%	4.51%	6.52%

Within a response. Splitting each response's object words into a first and a second half, the second half hallucinates more in three of four models; on Qwen2.5-VL-7B the interval contains zero (Table 14).

Table 14: Hallucination rate in the first and second half of each response's object words, averaged over responses with at least two object words (with an odd count, the middle word goes to the second half). Paired over responses.

Model	Responses	First half	Second half	Difference
Qwen2.5-VL-3B	1,002	4.94	6.75	$+1.80$ $[+0.80, +2.77]$
Qwen2.5-VL-7B	1,000	5.11	5.84	$+0.73$ $[-0.18, +1.71]$
LLaVA-1.5-7B	998	4.50	8.56	$+4.06$ $[+3.10, +5.00]$
SmolVLM	986	4.81	6.09	$+1.28$ $[+0.23, +2.33]$

Image dependence fades in every model. Per-token image dependence, the KL divergence between the next-token distributions with and without the image, falls from the first position band (tokens 0–24) to the last (100–127) by 80% in Qwen2.5-VL-3B, 82% in Qwen2.5-VL-7B, 87% in LLaVA-1.5-7B, 74% in SmolVLM, and 78% in the student. The fade is a property of these models that distillation leaves in place, not a change it introduces.

C.7 The Ten Models Behind §4.2

§4.2 separates the perception cost from hallucination using the ten models of Table 2: seven students trained under different teachers or trace caps, and the three interpolated models of §6. Across the ten, response length and the cost on image-requiring items show no resolvable relation (Spearman $\rho = -0.127$, permutation $p = 0.73$). Across the seven students, response length and CHAIR on AMBER rise together ($\rho = +0.893$; Figure 10).

The 512-token cap cuts many of the long-writing students' AMBER descriptions (Table 3), which undercounts their hallucination, since the cut tails hallucinate most. On the 607 images where neither the base nor the student reaches the cap, the student's excess CHAIR is $+5.96$ $[+4.46, +7.43]$, against $+8.26$ on all images; this subset in turn favors the student's shorter descriptions, so neither number is unbiased, and both are resolvable.

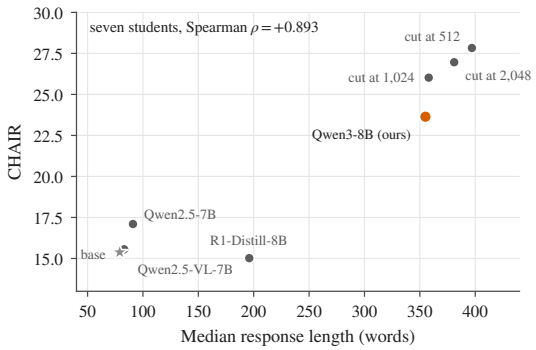


Figure 10: AMBER median response length against CHAIR for the seven students of Table 2 ($\rho = +0.893$).

C.8 The RL Stage

§4 runs the RL stage from the student with three seeds. On MathVista, seed 1 scores 53.00 after forcing, against the student's 52.67. On the 114 free-form MathVision items the seeds score 26.32, 24.56 and 21.05, against 15.79 for the base, 20.18 for the student and 20.18 for $\lambda = 0.75$. Averaged over seeds, the RL checkpoint is 8.19 points above the base ($[+1.75, +14.62]$), 3.80 above the student ($[-3.22, +10.53]$) and 3.80 above $\lambda = 0.75$ ($[-1.75, +9.36]$). Seed by seed, the differences from the base are $+10.53$, $+8.77$ and $+5.26$; the first two intervals exclude zero. Table 15 gives

seed 1 on the four benchmarks evaluated after RL: against the base, the only resolvable gain is on MathVision; against the student, RL also recovers LogicVista (+7.00).

Table 15: The RL stage against the base and the student, item-paired. MMStar was not evaluated after RL. †The 114 free-form items. ⋆: interval excludes zero.

Benchmark	Seed	Base	Student	RL	RL − base	RL − student
MathVista	1	62.33	52.67	53.00	−9.33 [−15.00, −3.33]⋆	+0.33 [−4.67, +5.33]
LogicVista	1	39.33	35.33	42.33	+3.00 [−4.33, +10.00]	+7.00 [+1.33, +12.67]⋆
WeMath	1	52.00	52.00	54.00	+2.00 [−4.00, +7.67]	+2.00 [−3.33, +7.33]
MathVision†	1	15.79	20.18	26.32	+10.53 [+2.63, +18.42]⋆	+6.14 [−1.75, +14.04]
MathVision†	2	15.79	20.18	24.56	+8.77 [+0.88, +16.67]⋆	+4.39 [−4.39, +13.16]
MathVision†	3	15.79	20.18	21.05	+5.26 [−1.75, +11.40]	+0.88 [−7.02, +8.77]

D Levers One and Two

D.1 Forcing a Conclusion

Against the unforced run at 2,048 tokens, forcing at 1,536 turns 55 MathVista questions right and 25 wrong (net +30, +10.00 points). Of the 55, 16 had no extractable answer before; the other 39, and all 25 that turned wrong, had an answer, which changed when the response was cut at 1,536 tokens and forced to conclude. At 1,536 tokens, 226 of the student's 300 responses reach the budget and are forced; the other 74 end on their own, each with an answer. Forcing also raises accuracy on the image-requiring items by +10.46 [+4.18, +16.74]: the perception cost of §4.2 is measured with forcing on. On the RL checkpoint, forcing is worth +9.33 [+4.00, +14.67].

D.2 Truncating Responses

§5.2 cuts each model's own AMBER descriptions to their first 50, 100 and 200 words and rescores them. CHAIR rises with the number of words kept in all six models, from four families (Table 16). The steepest rise belongs to a model we did not train: InternVL2.5-8B goes from 9.93 at 50 words to 24.34 in full.

Table 16: CHAIR after truncating each model's AMBER descriptions to their first K words (all 1,004 images).

Model	$K = 50$	$K = 100$	$K = 200$	Full	Median words
Qwen2.5-VL-3B (base)	10.51	13.46	15.26	15.38	79
Student	12.28	16.98	21.72	23.64	355
Qwen2.5-VL-7B	10.18	12.10	12.27	12.27	84
LLaVA-1.5-7B	13.29	17.00	17.01	17.01	66
SmolVLM	11.30	14.41	19.41	22.50	300
InternVL2.5-8B	9.93	14.45	20.12	24.34	382

The same cut separates the two teachers of §5.2 that differ only in visual training (Table 17). Their students differ by 1.54 CHAIR points in full. Cut to the same length with the student's own tokenizer, the gap is not resolvable up to 110 tokens and returns from 160 on. The median responses are 106 and 96 tokens long.

Table 17: CHAIR of the text-only teacher's student (Qwen2.5-7B) minus that of the visually trained teacher's student (Qwen2.5-VL-7B), after cutting both students' AMBER descriptions to their first K tokens. Paired over 1,004 images. ⋆: interval excludes zero.

K (tokens)	50	80	110	160	256	Full
Difference	+0.21	+0.33	+0.80	+1.44⋆	+1.52⋆	+1.54⋆
95% CI	[−0.87, +1.31]	[−0.72, +1.39]	[−0.28, +1.86]	[+0.32, +2.49]	[+0.42, +2.60]	[+0.44, +2.62]

E LEVER THREE

Figure 11 and Table 18 give the full interpolation line on MathVista and MMStar; Table 19 gives $\lambda = 0.75$ on all five accuracy benchmarks; Table 20 splits the result by blind-teacher stratum.

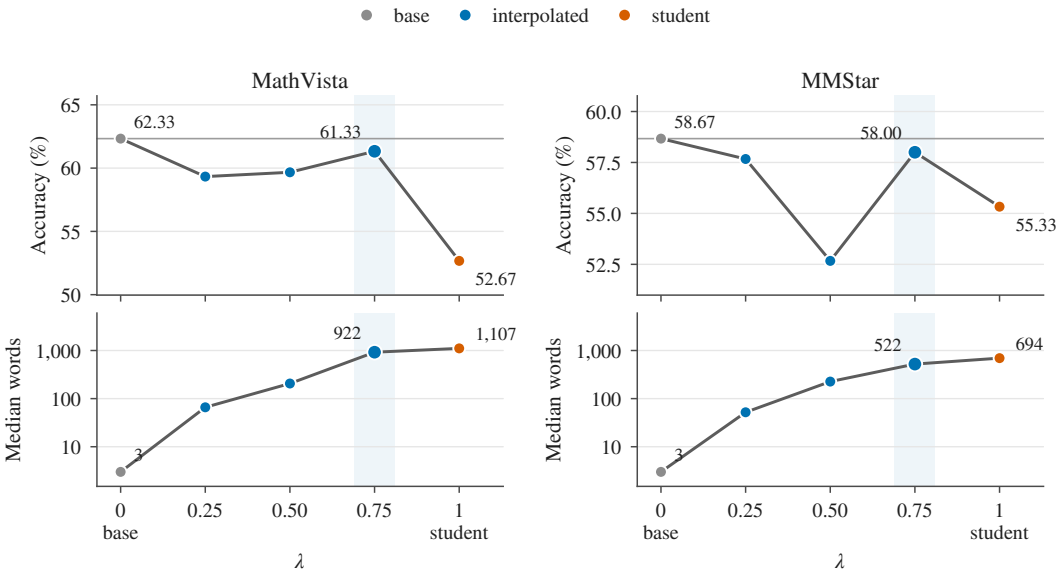


Figure 11: The interpolation line from the base ($\lambda = 0$) to the student ($\lambda = 1$). Left axis: accuracy. Right axis: median response length, log scale. On MathVista every interpolated model is more accurate than the student, already at $\lambda = 0.25$, while response length approaches the student's only at $\lambda = 0.75$ (§6.1). Numbers and intervals in Table 18.

Table 18: The full interpolation sweep on MathVista and MMStar ($n = 300$ each), item-paired differences. Same budget, forcing and extractor as Table 1; the $\lambda = 0.75$ row shares its resamples with Tables 1 and 19. $\star$: interval excludes zero.

	MathVista			MMStar		
λ	Acc.	− Base	− Student	Acc.	− Base	− Student
0 (base)	62.33	—	+9.67 [+3.67, +15.67]*	58.67	—	+3.33 [−2.00, +8.33]
0.25	59.33	−3.00 [−7.33, +1.33]	+6.67 [+0.67, +12.67]*	57.67	−1.00 [−5.00, +3.00]	+2.33 [−3.33, +7.67]
0.50	59.67	−2.67 [−8.33, +2.67]	+7.00 [+1.33, +13.00]*	52.67	−6.00 [−11.67, −0.33]*	−2.67 [−8.67, +3.00]
0.75	**61.33**	−1.00 [−6.33, +4.67]	+**8.67** [+3.00, +14.33]*	**58.00**	−0.67 [−6.00, +5.00]	+2.67 [−2.33, +8.00]
1 (student)	52.67	−9.67 [−15.67, −3.67]*	—	55.33	−3.33 [−8.33, +2.00]	—

Table 19: $\lambda = 0.75$ on all five accuracy benchmarks, item-paired against the student and the base. Last column: the median response length of $\lambda = 0.75$ as a share of the student's. $\star$: interval excludes zero.

Benchmark	Base	Student	$\lambda = 0.75$	λ − Student	λ − Base	Length kept
MathVista	62.33	52.67	61.33	+8.67 [+3.00, +14.33]*	−1.00 [−6.33, +4.67]	83%
MMStar	58.67	55.33	58.00	+2.67 [−2.33, +8.00]	−0.67 [−6.00, +5.00]	75%
LogicVista	39.33	35.33	45.33	+10.00 [+4.00, +16.00]*	+6.00 [−1.00, +13.00]	96%
WeMath	52.00	52.00	58.00	+6.00 [+0.67, +11.67]*	+6.00 [−0.67, +12.67]	96%
MathVision (free-form)	15.79	20.18	20.18	+0.00 [−7.02, +7.02]	+4.39 [−2.63, +12.28]	94%

Where the recovery lands, by benchmark. Table 20 applies blind-teacher stratification to four benchmarks. MathVista and MMStar carry a resolvable cost, both on image-requiring items. On MathVista, $\lambda = 0.75$ returns 9.62 of the 12.55 points lost there, a resolvable gain (§6.2); on MMStar its gain of 2.07 is not resolvable. On LogicVista and WeMath the student's cost is not resolvable in either stratum, and $\lambda = 0.75$ still improves on the student in one stratum each: by 10.34 on

image-requiring LogicVista items and by 10.87 on text-solvable WeMath items. Figure 12 follows the individual answers on the image-requiring MathVista items. Against the base, $\lambda = 0.75$ leaves no resolvable cost on the image-requiring items of MathVista (-2.93 $[-9.21, +3.35]$) or MMStar (-4.98 $[-11.20, +0.83]$).

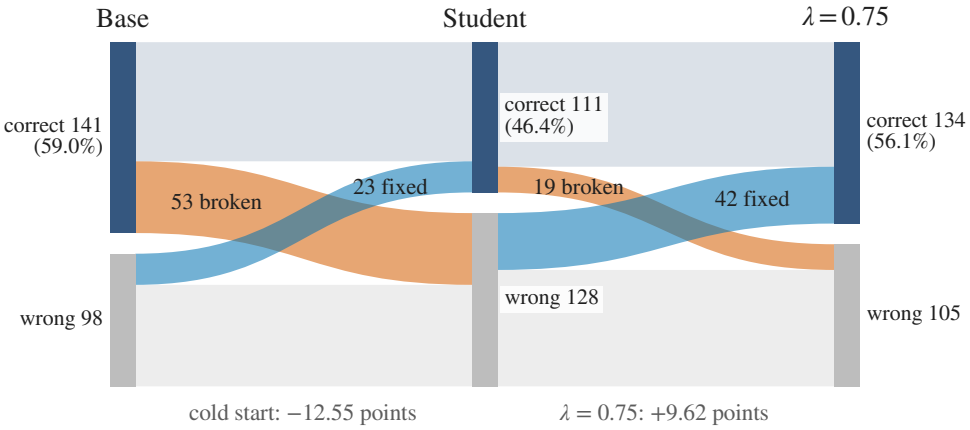


Figure 12: Where the answers go on the 239 image-requiring MathVista items. Bars: items each model answers correctly (dark) or wrongly (grey). Ribbons: items whose answer changes from one model to the next. The cold start breaks 53 answers and fixes 23 (-12.55 points); $\lambda = 0.75$ fixes 42 of the student's wrong answers and breaks 19 ($+9.62$ points).

Table 20: Blind-teacher strata on four benchmarks: the student's cost against the base, and the gain of $\lambda = 0.75$ over the student. $\star$: interval excludes zero.

Benchmark	Stratum	Items	Student − base	$\lambda = 0.75$ − student
MathVista	image-requiring	239	-12.55 $[-19.67, -5.44]^\star$	$+9.62$ $[+2.93, +15.90]^\star$
	text-solvable	61	$+1.64$ $[-11.48, +14.75]$	$+4.92$ $[-6.56, +16.39]$
MMStar	image-requiring	241	-7.05 $[-12.03, -2.07]^\star$	$+2.07$ $[-3.73, +7.88]$
	text-solvable	59	$+11.86$ $[+0.00, +23.73]$	$+5.08$ $[-8.47, +18.64]$
LogicVista	image-requiring	203	-4.43 $[-12.32, +3.45]$	$+10.34$ $[+2.96, +17.73]^\star$
	text-solvable	97	-3.09 $[-15.46, +9.28]$	$+9.28$ $[-2.06, +20.62]$
WeMath	image-requiring	208	$+0.96$ $[-7.21, +8.65]$	$+3.85$ $[-3.37, +11.06]$
	text-solvable	92	-2.17 $[-11.96, +8.70]$	$+10.87$ $[+2.17, +19.57]^\star$

E.1 Choosing λ on a Selection Set

The selection rule is the one stated in §6.1: among $\lambda \in \{0.25, 0.50, 0.75\}$, take the largest whose accuracy is resolvably above the student's and whose image-requiring accuracy is not resolvably below the base's. We used a selection set and a test set. For the selection set, MathVista testmini 300–599, our first written note recorded only the second condition, which every λ passes there; the full rule, applied afterwards, picks $\lambda = 0.50$, since $\lambda = 0.75$ is not above the student on this set. We then wrote the full rule down before running questions 600–999, which were used for no choice, and report all three λ there. Only this last step is prospective. All runs follow the protocol of Table 1, with strata from the same blind teacher. Table 21 gives each set and all 1,000 questions. On 300–599 the student's cost is small and $\lambda = 0.75$ is not above the student, so the rule picks $\lambda = 0.50$. On 600–999, $\lambda = 0.50$ is 8.00 points above the student and 2.75 above the base; $\lambda = 0.75$ is also resolvably above the student there and not resolvably below the base. Over all 1,000 questions, $\lambda = 0.75$ recovers two-thirds of the student's cost and keeps 81% of its median length.

E.2 Three Levers, Three Costs

Table 22 applies each lever to the student and measures all three parts of the cost. Each lever moves its own part resolvably. Off the diagonal, forcing also raises image-requiring accuracy, because the answers it recovers include image-requiring ones; the teacher change leaves 16 responses without

Table 21: MathVista testmini in three sets and in full: the 300 questions of the main text, the 300 used to choose λ, and 400 used for no choice. Item-paired differences with 95% bootstrap intervals. Length kept: median length over the student's. $\star$: interval excludes zero.

	0–299 (main text)	300–599 (selection)	600–999 (test)	All 1,000
Items (image-requiring)	300 (239)	300 (221)	400 (299)	1,000 (759)
Student − base	-9.67 $[-15.67, -3.67]^\star$	-4.00 $[-10.00, +2.00]$	-5.25 $[-10.75, +0.00]$	-6.20 $[-9.60, -2.90]^\star$
image-requiring	-12.55 $[-19.67, -5.44]^\star$	-3.62 $[-11.31, +3.62]$	-6.35 $[-12.71, -0.33]^\star$	-7.51 $[-11.33, -3.56]^\star$
$\lambda = 0.50$ − student	$+7.00$ $[+1.33, +13.00]^\star$	$+5.67$ $[+0.33, +11.33]^\star$	$+8.00$ $[+3.00, +13.25]^\star$	$+7.00$ $[+3.90, +10.10]^\star$
$\lambda = 0.50$ − base, image-req.	-3.35 $[-9.21, +2.51]$	$+0.45$ $[-5.43, +6.33]$	$+1.67$ $[-3.68, +6.69]$	-0.26 $[-3.43, +3.03]$
$\lambda = 0.75$ − student	$+8.67$ $[+3.00, +14.33]^\star$	-1.67 $[-7.00, +3.67]$	$+5.00$ $[+0.50, +9.50]^\star$	$+4.10$ $[+1.20, +7.10]^\star$
$\lambda = 0.75$ − base, image-req.	-2.93 $[-9.21, +3.35]$	-6.33 $[-13.57, +1.36]$	$+0.67$ $[-5.02, +6.69]$	-2.50 $[-6.19, +1.19]$
Length kept, $\lambda = 0.50$ / 0.75	19% / 83%	17% / 77%	18% / 84%	18% / 81%

an answer and moves the perception cost by -4.60, which is not resolvable; and $\lambda = 0.75$ removes 1.21 of the 8.26 points of extra hallucination, which is resolvable but small.

Table 22: Each lever against each part of the cost: the student after the lever minus the student before it. No answer: MathVista responses with no extractable answer (for forcing, against the unforced run; otherwise after forcing). Hallucination: AMBER CHAIR, bootstrap over images. Perception: accuracy on the 239 image-requiring MathVista items. [a]The AMBER runs are not forced: the forcing cue asks for an answer, which a description does not have; how many descriptions reach the budget is in Table 3. [b]The answers forcing recovers include image-requiring ones, which is why the perception cost is measured with forcing on (Appendix D.1). $\star$: interval excludes zero.

Lever	Acts	No answer	Hallucination (CHAIR)	Perception (accuracy)
Force a conclusion	inference	$36 \to 0$	n/a[a]	$+10.46$ $[+4.18, +16.74]^{\star b}$
Teacher R1-Distill-Llama-8B	training	$0 \to 16$	-8.62 $[-9.68, -7.56]^\star$	-4.60 $[-11.30, +1.67]$
Interpolation, $\lambda = 0.75$	after training	$0 \to 0$	-1.21 $[-2.14, -0.26]^\star$	$+9.62$ $[+2.93, +15.90]^\star$

F Signal Audit

A line of work locates ungrounded tokens by comparing the next-token distribution with and without the image (Leng et al., 2023; Zhang et al., 2026; Yoon et al., 2026; Hoche et al., 2026), at the price of a second, image-free forward pass, and Bendre et al. (2026) find that the end-task gains of contrastive decoding built on such signals are largely spurious. We ran the same audit on four models (Table 23, Figure 13). Two signals that cost nothing, token entropy and relative position in the response, separate hallucinated from faithful object words as well as the three image-contrast signals do, and adding the three to the two changes AUC by -0.001 to $+0.020$. Four contrastive-decoding variants built on these signals do not lower CHAIR (Table 24).

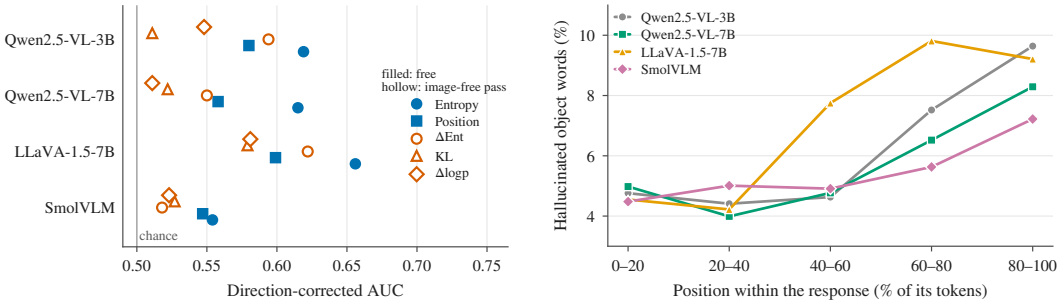
AUCs are direction-corrected, $|\text{AUC} - 0.5| + 0.5$, so that a signal that is lower on hallucinated words also counts as discriminative. Combined AUCs come from logistic regression under 5-fold cross-validation grouped by image, so the tokens of one description never fall on both sides of a split.

Table 23: Left: how well each signal alone separates hallucinated from faithful object words. Right: the two free signals (entropy and relative position) alone, and with the three image-contrast signals added. AMBER, all 1,004 images.

	Single signal					Combined		
Model	Entropy	Position	ΔEnt	KL	Δlogp	Free	+Image	Gain
Qwen2.5-VL-3B	0.619	0.580	0.594	0.511	0.548	0.626	0.632	+0.005
Qwen2.5-VL-7B	0.615	0.558	0.550	0.522	0.511	0.600	0.599	−0.001
LLaVA-1.5-7B	0.656	0.599	0.622	0.579	0.581	0.674	0.683	+0.009
SmolVLM	0.554	0.547	0.518	0.527	0.523	0.551	0.571	+0.020

Table 24: Four contrastive-decoding variants against no steering, Qwen2.5-VL-3B on all 1,004 AMBER images, paired by image (CHAIR without steering: 14.51; all five variants, this unsteered reference included, come from one separate generation run, so the reference differs from the base's 15.38 in the main text). None lowers CHAIR; the confidence-weighted variant raises it. $\star$: interval excludes zero.

Variant	ΔCHAIR	95% CI
constant	$+0.82$	$[-0.19, +1.80]$
visual	-0.49	$[-1.40, +0.44]$
confidence	$+1.14^{\star}$	$[+0.08, +2.19]$
both	$+0.72$	$[-0.20, +1.59]$



(a) Direction-corrected AUC of each signal alone.

(b) Hallucination rate by position within the response.

Figure 13: Signal audit (Table 23). Hallucinated object words concentrate toward the end of a response, which is why position alone competes with signals that need an extra forward pass.