CERTIFIED INTERFACE ALIASES: EXACT COLLISIONS IN VISION–LANGUAGE PREPROCESSING, AND WHEN THEY EXIST

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language verifiers and routers must distinguish errors repairable by more reasoning from those caused by visual evidence never reaching the language model. This distinction lacks ground truth because annotators see full-resolution images while models receive preprocessed tensors. We introduce ALIASFORGE to create cases where the relevant fact is provably absent from the interface. Fixed-point resampling makes the pre-rounding resize an exact integer linear map that can send nonzero integer perturbations to zero. Hiding a label-flipping perturbation there produces images with opposite step-correctness labels but bit-identical interface states. Every verifier therefore has the same output law on both members, giving pair-balanced accuracy exactly one half and zero gain from language-side repair. We prove that every fixed-point downscaler has such null vectors and bound their smallest size at the most common ratios, which rules out an 8-bit fit whenever the bound exceeds 255. From the resize configuration alone, a lattice criterion supplies realizable collisions and certifies their absence within the specified construction family. It resolves all twenty screened configurations, seventeen as constructible and three as non-constructible. We construct certified pairs across three architectures and certify four additional processors, with zero decision-logit gap on all 18 scored pairs and none of the 18 controls. The pairs also screen routers that waste computation on re-attention or further reasoning. On natural items, per-item routing headroom exists, but no tested interface-only router improves over stopping. Our fiber ceiling bounds the headroom recoverable from the interface.

1 INTRODUCTION

When a vision–language model fails on a visual task, diagnosing the error requires knowing whether the visual evidence ever reached the language model. Nothing the model emits reveals this. Yet practitioners are forced to act without this knowledge. They train perception-aware reward models (Min et al., 2026; Zhang et al., 2026; Diao et al., 2026), decide whether a downsampled view suffices (Yang et al., 2025), and route test-time compute accordingly (Damani et al., 2025; Jeddi et al., 2026). None of these procedures can verify if the visual evidence survived preprocessing, and guessing incorrectly is not free. Consider when a verifier model, tasked with judging a reasoning step against the image, is instructed to re-examine an input whose decisive detail never survived preprocessing. In this scenario, the system wastes compute on a channel that carries no useful signal. Conversely, re-encoding at a higher resolution when the detail was already present wastes compute on an unnecessary intervention. Similarly, a reward model trained on perception labels assigned by a judge model simply inherits the judge's own perceptual blind spots. Ultimately, these mechanisms are making decisions based on missing information. The resulting errors compound quietly because, from the outside, a model that could not see the evidence behaves exactly like one that did not use it (Rahmanzadehgervi et al., 2024; Takishita et al., 2025).

This gap exists because human evaluators establish ground truth from the full-resolution image, whereas the model receives only a tensor. Two inputs whose tensors differ at all, however slightly, can be told apart by some verifier (Proposition 3), so no threshold on tensor similarity can certify

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.

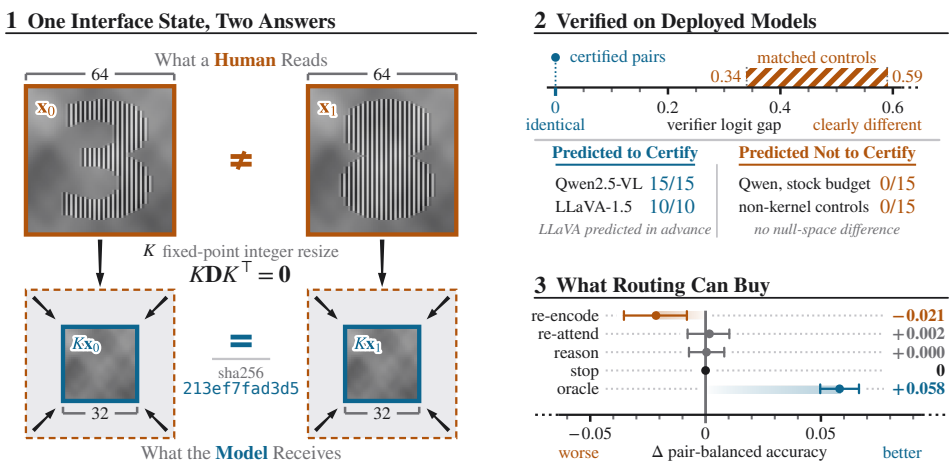

Figure 1: **Certified interface aliases and routing.** (1) Images with different correct answers reach the model as bit-identical tensors because the deployed resize K maps their difference $\mathbf{D}$ to zero (Theorem 2). (2) The criterion predicts which deployed pipelines admit certified pairs, while controls outside the resize null space and identity-resize controls do not certify (Proposition 9). (3) No tested action improves on stopping on natural items despite per-item oracle headroom; error bars are 95% intervals clustered by record, the source problem.

that the visual evidence failed to reach the model. Our certificate comes from one observation. An integer resize has an integer null space, and we hide a label-flipping perturbation in it.

In most deployed preprocessing pipelines the input image is resized by a resampler to the size the model expects. A deployed resampler's filter weights are fixed-point integers, so its downscaling resize is an exact linear map over the integers, and some integer differences between images map to zero. We call two such images a *certified pair*[1] when they have opposite ground-truth labels under one shared question. Because their difference maps to zero, they reach the model as bit-identical tensors, so it provably cannot tell them apart (Figure 1). Our constructor ALIASFORGE hides a label-flipping perturbation in one of those differences (Section 5). On such a pair every failure is a perception failure, and nothing the model does can help unless it changes the image. From the outside, chance accuracy on a certified pair looks the same as chance accuracy on two images that are merely similar. Only on a certified pair is that chance provable rather than observed. Constructed pairs test verifiers and routers on known perceptual failures, and they do not by themselves diagnose a natural error. The mechanism descends from image-scaling attacks (Xiao et al., 2019; Quiring et al., 2020), repurposed from evasion to measurement.

We build ALIASFORGE, the constructor that produces certified pairs, validate it on three architectures, and audit twenty declared configurations with its criterion. We then ask the question that motivated the work: *Can a score read from the preprocessed tensor alone decide when an input deserves more compute?* No score we test can, and the fiber ceiling (Proposition 5) bounds what any such score could recover. Three contributions follow.

- **A certified measurement instrument.** We propose and implement ALIASFORGE, which constructs certified pairs whose members have opposite correct answers but bit-identical preprocessed states. We prove that no verifier or language-side repair can distinguish their members (Theorem 2), and empirically validate the construction across three architectures and four additional processors (Tables 1 and 7).
- **A constructibility criterion from preprocessing alone.** We derive a criterion that uses only the deployed resize operation to test whether our construction can produce a certified pair, without model weights or queries (Theorem 6, Proposition 9). Across all 20 declared configurations, it either produces a pair or certifies that our construction cannot produce one. It also correctly predicts constructibility on a second architecture before any pairs are generated.

[1]This notion differs from certified adversarial robustness, which proves that predictions remain fixed under bounded perturbations. Our certificate proves that inputs with different correct answers reach the model identically.

- **A screen and ceiling for interface-only routing.** We use certified pairs to test whether a router spends compute on re-attention or further reasoning when those actions provably cannot help (Corollary 4). We also derive the fiber ceiling, which bounds the gain of any router that selects actions using only the preprocessed image (Proposition 5). On natural benchmark items, the best action varies across items, but none of the interface-only routers we test improves on always stopping (Tables 2 and 11).

2 Related work

Approximate and exact collisions. Prior work identifies visually distinct images with similar internal representations. The Multimodal Visual Patterns (MMVP) benchmark selects "CLIP-blind" image pairs by thresholding embedding similarity (Tong et al., 2024). Other work uses optimization to make images approach a target embedding (Salman et al., 2024; 2025), while connector reconstruction loss has been used to predict failures in grounded question answering (Li et al., 2025b). These approaches identify approximate similarity or predict failure, but do not certify exact equality of the deployed interface (Proposition 3). Text preprocessing shows the opposite behavior. Visually identical Unicode strings can produce different token sequences (Boucher et al., 2022), while our visually different images produce the same preprocessed representation. The closest constructive precedent is image-scaling attacks, which create inputs whose downscaled content differs from their full-resolution appearance (Xiao et al., 2019), a behavior arising from the interaction between downsampling and convolution (Quiring et al., 2020). These attacks have been demonstrated against deployed multimodal systems (Zeeshan & Satti, 2025), and the preprocessing configuration can be recovered through black-box queries (Hu et al., 2026). Defenses developed for these attacks aim to prevent attacker-chosen content from appearing after downscaling (Quiring et al., 2020), whereas our aliases (Definition 1) leave the downscaled output unchanged while altering the correct answer in the full-resolution image. Randomized resizing can move bicubic preprocessing away from an integer downscaling ratio (Xie et al., 2018), but changing the resampling kernel does not necessarily eliminate aliases. Area scaling has been reported to resist image-scaling attacks (Quiring et al., 2020). However, at the integer resize ratios evaluated in this study, it is equivalent to Pillow's box filter, which admits exact collisions at every ratio we successfully reconstructed.

Representation probing. A representation probe trains a decoder to predict a visual fact from the model's internal representation. Successful decoding provides evidence that the fact is present, but failure does not prove absence because the decoder itself may be inadequate. Control tasks help detect probes that succeed by fitting arbitrary labels (Hewitt & Liang, 2019). Recent theory characterizes information loss through inputs that share a representation (Sevetlidis, 2026a;b), but does not identify such collisions in a particular deployed system. Our certified pairs provide such a witness for a deployed interface.

Perception labels in verifiers and routers. Prior work trains verifiers to identify whether a reasoning error is perceptual (Min et al., 2026; Zhang et al., 2026; Ong et al., 2025; Wang et al., 2026), evaluates this ability at the reasoning-step level (Song et al., 2025), and uses similar signals to decide when additional computation is needed (Damani et al., 2025; Vo et al., 2026; Diao et al., 2026; Jeddi et al., 2026; Yu et al., 2026). Among these methods, Yang et al. (2025) is closest to our setting because it routes compute using an estimate of whether the relevant visual evidence is available. Rahmanzadehgervi et al. (2024) study a complementary failure in which a visual fact is represented but not used, while other studies infer limited visual access from model behavior (Liu et al., 2026; Shahgir et al., 2026; Osaka et al., 2026). Correctly identifying the source of an error matters because one audit assigns 86.9% of a strong model's visual-mathematics errors to perception using a judge model that cannot verify what visual information reached the model (Wu et al., 2026). Our certified pairs establish a known case in which the distinguishing fact is absent from the interface and provide a one-forward-pass test of a router's decision without observing intervention outcomes (Corollary 4). Prior work finds that longer reasoning can disperse visual attention (Peng et al., 2026; Sammani et al., 2026), prompted self-reflection can fail to revisit the image (Shi et al., 2026a), and chains of thought can propagate upstream perceptual errors (Sharif et al., 2026). Our routing evaluation complements these findings by keeping the verifier frozen, separating re-attention (Chu et al., 2025) from re-encoding (Shi et al., 2026b), and measuring regret against a per-item oracle.

3 SETUP

A verifier $f(\mathbf{x}, Q, R) \in [0,1]$ assigns a probability that candidate reasoning step R is correct for image $\mathbf{x}$ and question Q. The image reaches the language model only through the *visual interface* $\mathbf{z} = \Phi(\mathbf{x})$, which contains every image-dependent quantity on which the language model conditions. Formally, the verifier is a Markov kernel from $(\mathbf{z}, Q, R)$ to a verdict in $\{0,1\}$, and $f(\mathbf{x}, Q, R)$ is the probability that it accepts R. This covers deterministic scoring and sampled decoding, and its randomness ω is drawn independently of $(\mathbf{x}, \mathbf{z})$. The deployed processor computes this interface in stages, so $\Phi = \Phi_n \circ \cdots \circ \Phi_1$, and we write $\mathbf{s}_k(\mathbf{x}) = \Phi_k(\cdots \Phi_1(\mathbf{x}))$ for the state after stage k, with $\mathbf{s}_n(\mathbf{x}) = \Phi(\mathbf{x})$. Write $y(\mathbf{x}, Q, R) \in \{0,1\}$ for the ground-truth step-correctness label, determined at full resolution. A *pair* is $(\mathbf{x}_0, \mathbf{x}_1, Q, R)$ with $y(\mathbf{x}_0, Q, R) = 1$ and $y(\mathbf{x}_1, Q, R) = 0$. For a pair $i = (\mathbf{x}_{i,0}, \mathbf{x}_{i,1}, Q_i, R_i)$ and member $b \in \{0,1\}$, write $p_{i,b} = f(\mathbf{x}_{i,b}, Q_i, R_i)$. The *pair-balanced accuracy* is $\mathrm{BA} = \mathbb{E}_i[\frac{1}{2}(p_{i,0} + 1 - p_{i,1})]$.

To study routing, we use the action set $\mathcal{A} = \{\mathsf{stop}, \mathsf{att}, \mathsf{rea}, \mathsf{enc}\}$. Here, stop accepts the current verdict, att (re-attend) instructs the model to re-examine the region named by the question, rea (reason) runs and aggregates further critique passes, and enc (re-encode) recomputes $\mathbf{z}$ from a higher-resolution crop (Appendix F, Table 8). A *language-side* action is anything the system does once $\mathbf{z}$ is fixed. Formally, it is a Markov kernel acting on $(\mathbf{z}, Q, R)$, and we write $\mathcal{L}$ for the set of such actions. $\{\mathsf{stop}, \mathsf{att}, \mathsf{rea}\} \subset \mathcal{L}$, while enc lies outside $\mathcal{L}$ because it recomputes $\mathbf{z}$. The set $\mathcal{L}$ is closed under composition and finite adaptivity, so it also includes procedures outside $\mathcal{A}$, such as self-consistency over any number of critiques (Definition 7). We call the unit scored by the verifier an *item*, either a pair or a single image with its question and candidate step. The verifier's *correctness* on an item is the probability it assigns to the true label, f when $y = 1$ and $1 - f$ when $y = 0$, and on a pair it is the members' average $\frac{1}{2}(p_{i,0} + 1 - p_{i,1})$. We write $\Delta_i(a)$ for the change in the verifier's correctness on item i after action a. Each action has a measured cost $c(a)$ with $c(\mathsf{stop}) = 0$, charged at a price $\lambda \geq 0$, and its gain is $g_i(a) = \Delta_i(a) - \lambda\, c(a)$. A *policy* π chooses one action per item from the information available to it.

4 CERTIFIED INTERFACE ALIASES

Assumption 1 (Interface factorization)**.** Fix a pipeline stage k.

(a) *Factorization.* The conditional law of $\mathbf{z}$ given $\mathbf{s}_k(\mathbf{x})$ does not depend on $\mathbf{x}$. No stage after k reads the raw image or any earlier state except through $\mathbf{s}_k$.
(b) *Shared non-image inputs.* The members of a pair use identical token sequences for Q and R and the same decoding parameters. Any sampling randomness is drawn independently of $(\mathbf{x}, \mathbf{z})$.

Definition 1 (Certified interface alias)**.** A pair $(\mathbf{x}_0, \mathbf{x}_1, Q, R)$ is an *interface alias at stage k* if its members produce exactly the same state at stage k, $\mathbf{s}_k(\mathbf{x}_0) = \mathbf{s}_k(\mathbf{x}_1)$. It is a *certified interface alias*, or a *certified pair* for short, once this equality is verified by comparing cryptographic digests of the full serialized state, including array contents, data type, shape and all non-array metadata.

Theorem 2 (Interface aliasing certificate)**.** *Under Assumption 1, the joint law of $(\mathbf{z}, Q, R, \omega)$ is the same under both members of an interface alias at stage k. Every verifier therefore induces the same output law on both members, and every language-side action preserves that law. Consequently, (i) every decision rule based on that law has $\mathrm{BA} = 1/2$ exactly; (ii) $\Delta_i(a) = 0$ exactly for every $a \in \mathcal{L}$, including att, rea with any number of passes, self-consistency over any number of critiques, and deterministic re-supply of the same image; and (iii) any action, language-side or not, that leaves $\mathbf{s}_k$ shared and preserves Assumption 1 has $\Delta_i(a) = 0$. Therefore, $\Delta_i(a) > 0$ requires an action that gives the members different states at stage k or introduces new image-dependent information.*

Theorem 2 is the data-processing inequality in its degenerate case, where a Markov kernel applied to identical inputs produces identical output laws. Its value is that an exact digest comparison establishes the hypothesis without model weights or queries. Assumption 1(a) is the theorem's architectural requirement. It excludes a parallel high-resolution branch or raw-image side channel unless the collision occurs before that path branches. Certification also requires equality over every image-dependent field emitted by the processor, so we include all enumerated fields in a single digest (Appendix C, Table 4). The certificate assumes that this enumeration is complete. We verify that every enumerated field affects the digest by perturbing each one in isolation.

Three consequences follow. Proofs are in Appendix A.

Proposition 3 (Sharpness). *Fix a pair, write μ_b for the law of $\mathbf{z}$ under member b, and let $\mathrm{BA} = \frac{1}{2}(p_0 + 1 - p_1)$ be that pair's value.* (a) *Every verifier reading only the interface, including after any language-side action, has $\mathrm{BA} = 1/2$ if and only if $\mu_0 = \mu_1$.* (b) *The highest pair-balanced accuracy attainable from the interface is $\frac{1}{2}\big(1 + \mathrm{TV}(\mu_0, \mu_1)\big)$, where the total-variation distance is $\mathrm{TV}(\mu_0, \mu_1) = \sup_A |\mu_0(A) - \mu_1(A)|$. Thus, if $\mu_0 \neq \mu_1$, some verifier outperforms chance. If the interface is deterministic and $\mathbf{z}_0 \neq \mathbf{z}_1$, some verifier attains $\mathrm{BA} = 1$.*

Part (b) is the Neyman–Pearson identity for testing between two laws (Le Cam, 1986; Tsybakov, 2009), so for a deterministic interface only an exact collision certifies chance performance without assumptions about the verifier.

Corollary 4 (Screen and ceiling). *(a) On a certified pair, for every $\lambda > 0$ and any costs satisfying $c(\mathsf{att}), c(\mathsf{rea}) > 0$, $g_i(\mathsf{att}) < g_i(\mathsf{stop})$ and $g_i(\mathsf{rea}) < g_i(\mathsf{stop})$. Thus, att and rea are strictly dominated by stop.*

(b) If a fraction κ of a population of pairs is certified, every verifier, before or after any language-side action, satisfies $\mathrm{BA} \leq 1 - \kappa/2$. Equality holds if the verifier separates every uncertified pair perfectly.

Corollary 4(a) provides a direct screen for routing decisions. On a certified pair, choosing att or rea wastes compute, and applying the screen requires only the router's initial decision, without running an intervention or observing its outcome.

Routing value. The routing value measures the advantage of an oracle that chooses the best action for each item over the best fixed action. Write

$$V = \mathbb{E}_i\big[\max_a g_i(a)\big] - \max_a \mathbb{E}_i\big[g_i(a)\big]. \tag{1}$$

At $\lambda = 0$, $g_i(a) = \Delta_i(a)$. Attaining V requires distinguishing items that benefit from different actions. A router sees only a *signal* $S = \sigma(\mathbf{z}, Q, R)$, some statistic of the interface and the text, and it is *interface-only* when S is a function of $\mathbf{z}$ alone. A router cannot always attain V because it must use the same action distribution for every item in a *fiber* of its signal, defined as the set of items that map to the same signal value. Proposition 5 gives the resulting ceiling, which is the value-of-information inequality in our notation (Blackwell, 1953; Howard, 1966).

Proposition 5 (Fiber ceiling). *Let $\pi(S, \omega)$ be any policy reading a signal S, where ω is auxiliary randomness independent of the item, and define $V_S = \mathbb{E}\big[\max_a \mathbb{E}[g_i(a) \mid S_i]\big] - \max_a \mathbb{E}[g_i(a)]$. Then* (i) *every such policy satisfies $\mathbb{E}[g_i(\pi(S_i, \omega))] - \max_a \mathbb{E}[g_i(a)] \leq V_S$, with equality attained by $\pi_S(S_i) \in \arg\max_a \mathbb{E}[g_i(a) \mid S_i]$;* (ii) *$0 \leq V_S \leq V$, and $V_{S'} \leq V_S$ whenever S' is a function of S; and* (iii) *$V_S = V$ if and only if some policy reading S selects a per-item optimal action almost surely.*

Thus, V_S is the greatest improvement over the best fixed action available to a router reading S, and we write V_Φ for the interface-only case $S = \mathbf{z}$. The gap $V - V_S$ is the routing value hidden by variation within fibers. A certified pair witnesses a fiber with two oppositely labeled items for a particular deployed Φ under every signal, because its members share $\mathbf{z}$, Q and R, so it screens any router.

5 AliasForge: Constructing and Auditing Certified Aliases

AliasForge analyzes a deployed processor and, when possible, constructs two images that have different correct answers but are mapped by its resize to exactly the same output. It then verifies equality across the complete visual interface and records the evidence needed to certify the pair. Algorithm 1 gives the complete audit and construction procedure.

A pixel-offset pattern computed in the real-valued null space generally contains fractional entries, which must be rounded to form a valid image. None of the 72 rounded candidates we tested as a control certified, with roughly 1% of their output values remaining unequal (Appendix E). We therefore construct the offset patterns using the resampler's exact integer arithmetic. The resize implementations selected by the target processors represent their filter weights as fixed-point integers rather

than floating-point numbers and differ only in final rounding. Their pre-rounding computations can therefore be represented exactly by an integer matrix K.

The resize is *separable*, meaning that it applies one-dimensional filtering independently in the horizontal and vertical directions. ALIASFORGE therefore analyzes the horizontal operation. Each column of K describes how the corresponding input-image column contributes to the resized output. Because each output column depends only on nearby input columns, the search can be restricted to W adjacent input columns, which we call a *window*. Let K_W denote the corresponding restriction of K. The *kernel* $\ker K_W$ is the set of offset patterns that K_W maps to zero. ALIASFORGE searches this kernel for a nonzero integer pattern $v \in \ker K_W \cap \mathbb{Z}^W$. Each entry of v specifies how much to change the pixels in one column of the window. Multiplying v by an integer increases its contrast without moving it out of the kernel. Repeating the resulting offsets across selected image rows creates a visible, label-changing pattern. Because the horizontal resize maps every modified row to zero, the subsequent vertical resize also receives zero difference, leaving the final resized output unchanged. We call this pattern the *carrier*. It can be added to any image whose affected pixels have enough range to avoid clipping, including natural images and synthetic renders.

Our construction is not available at every resize ratio. It requires a *realizable* vector, a nonzero integer vector in $\ker K_W$ whose largest absolute entry is at most 255, the largest possible difference between two 8-bit pixel values. Whether such a vector exists depends on both the resize ratio and the resampling filter. These vectors form the integer lattice $L = \ker K_W \cap \mathbb{Z}^W$, a discrete set closed under integer combinations. Because L is determined entirely by the deployed resize configuration, constructibility can be evaluated without model weights or queries. Away from the image border, a resampler at ratio p/q in lowest terms is periodic, so it is described by a $q \times p$ matrix $H(\zeta)$ of integer polynomials in an indeterminate ζ, its *polyphase matrix* (Vaidyanathan, 1993), and an integer offset vector v on interior columns by a polynomial vector $\hat{v}(\zeta) \in \mathbb{Z}[\zeta]^p$ (Appendix A).

Theorem 6 (Existence and size of integer kernel vectors)**.** *Let K be the fixed-point resize operator at ratio $p/q > 1$, and let v be an integer offset vector on interior columns (Remark 11).*

(a) Existence. *$Kv = 0$ if and only if $H(\zeta)\hat{v}(\zeta) = 0$. The solutions form a $\mathbb{Z}[\zeta]$-module of rank at least $p - q \geq 1$, so every wide enough image admits finitely supported kernel vectors $v \neq 0$.*

(b) Principal case. *If $p = q + 1$ and H has full row rank, every integer kernel vector is $\hat{v} = \xi\, w$ for some $\xi \in \mathbb{Z}[\zeta]$, where w is the vector of signed maximal minors of H divided by their greatest common divisor. Every nonzero interior kernel vector of any support width then satisfies $\|v\|_\infty \geq \beta(K) := \max_{s:\, w_s \neq 0} \max\{|\mathrm{lc}(w_s)|, |\mathrm{tc}(w_s)|\}$, where* lc *and* tc *are the highest- and lowest-order nonzero coefficients, and w itself is a kernel vector of sup-norm $\|w\|_\infty$.*

Non-constructibility is therefore always a question of size. The principal case covers $2\times$, $1.5\times$ and $4/3\times$, and the rank condition holds for every operator we reconstruct (Table 3). For bicubic resampling at $2\times$, w encodes the carrier $(1, -3, 3, -1)$ and $\beta(K) = 3 = \|w\|_\infty$, so our carrier is the smallest possible at any interior support width. For Lanczos at $2\times$, $\beta(K) = 63{,}150$, and at $1.5\times$ it exceeds 5×10^9 for bicubic, bilinear and Lanczos, so no realizable interior carrier exists at any width. Outside the principal case, such as $3\times$ and $2.5\times$, we bound window-supported vectors.

We reduce a basis of L to search for a realizable vector. If none is found, the bound in Proposition 9 may certify that no such vector exists within the specified window, because every nonzero $v \in L$ satisfies $\|v\|_\infty \geq \min_i \|b_i^*\|/\sqrt{W}$ for the Gram–Schmidt vectors b_i^* of any basis of L (Appendix A). Otherwise, the configuration remains unresolved. The bound covers the rank-one differences $\mathbf{D} = uv^\top$ that ALIASFORGE produces, where the binary vector u selects the affected rows. Appendix A extends it to differences of any rank, which rules out both declared configurations at $1.5\times$.

Applied to the 20 declared configurations using bicubic resampling, the criterion returns a definitive verdict in every case (Appendix D). Seventeen are constructible and three are ruled out. All constructible configurations use integer resize ratios, while the three exclusions use non-integer ratios (Table 6). These verdicts depend only on K and are obtained before an image is rendered or a model is queried.

We also reconstructed K for 28 resampling-filter and ratio combinations. The reconstruction matched the library bit for bit in 27 cases, and the criterion returned a definitive verdict in 26. No tested filter eliminates the construction at every ratio (Table 5, Appendix D).

Table 1: **Instrument validation.** Targets are Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-1.5-7B (Liu et al., 2024) and VisualPRM-8B (Wang et al., 2025). Predictions matched the observed outcomes in all 192 attempts, and all 18 measured decision-logit gaps between members were exactly zero.

Target or control	Resize configuration	Certified pairs	Member logit gap
Qwen2.5-VL-7B	2× downscale	15 of 15	zero on all 8 scored
LLaVA-1.5-7B	fixed 2× resize	10 of 10	zero on all 10
VisualPRM-8B	3×3 tiles and thumbnail	20 of 20	not available
All negative controls	four control designs	0 of 147	nonzero when available

Instrument validation. Table 1 reports the certification experiment. The constructibility criterion was evaluated before each experiment, so the table compares prospective predictions with observed outcomes. For the second architecture, the criterion predicted constructibility using only the deployed processor, and the construction transferred without modification. The dynamic-tiling target provides a more demanding test because it processes nine 2× tiles and a 6× thumbnail, requiring the pair to collide under both resize operators. All 20 constructed pairs certified across the nine tiles, the thumbnail, and every recorded interface field, while none of the 20 same-support, same-amplitude patterns outside the kernel did. An identity-resize control also did not certify, confirming that the construction requires both exact kernel membership and a lossy resize. Four additional pipelines certified at the processor level (Table 7), and Appendix E gives the full controls.

The construction extends beyond synthetic base images. On 200 benchmark images resized to the eligible 2× regime, selecting carrier rows with sufficient intensity headroom produced 48 certified pairs at amplitude 20, of which 8 keep the marks clearly legible (Figure 2, Appendix D).

6 Interface Accessibility and Routing

To test whether an interface-only router can recognize when relevant visual evidence is unavailable, we need a score computed from the interface alone. We define this score relative to the binary fact associated with each item. On the natural stream of benchmark items, the fact is whether the candidate reasoning step is correct. The question Q_i and step R_i identify the fact being evaluated but are not decoder features. An auxiliary decoder d_θ, implemented as logistic regression on a projection of $\mathbf{z}$ (Appendix G), estimates the probability $\hat{p}_\theta(1 \mid \mathbf{z}_i)$ that the fact is positive. The *interface accessibility score* is $\mathrm{IAS}_i = \mathrm{cal}(\hat{p}_\theta(1 \mid \mathbf{z}_i))$, where cal is an out-of-fold isotonic calibration map. We cross-fit the decoder with folds grouped by source image and, on the natural stream, by *record*, the source problem containing each item's step. The certified set provides an absent-side calibration anchor at $\mathrm{IAS} = 1/2$.

For each action, a linear model fitted on a fixed random half of the records, the development half, estimates its gain $\hat{g}_i(a)$ from IAS_i. The router IAS-Route selects the action with the largest estimated gain. At test time, this routing decision depends only on $\mathbf{z}$. The question and candidate step are passed to the selected action after the decision is made.

Evaluation metrics. A policy's regret is its expected shortfall relative to the per-item oracle $\pi^\star$,

$$\mathrm{regret}(\pi) = \mathbb{E}_i\big[g_i(\pi^\star(i)) - g_i(\pi(i))\big], \qquad \pi^\star(i) = \arg\max_{a \in \mathcal{A}} g_i(a). \tag{2}$$

Order accuracy is the fraction of items on which π selects the same action as $\pi^\star$. The dominated-action rate is the fraction of certified pairs sent to att or rea, which gain nothing there and are strictly dominated for every $\lambda > 0$ (Corollary 4(a)). All headline results use raw gains with $\lambda = 0$. We also evaluate the sweep $\lambda \in \{0, 0.005, 0.01, 0.02, 0.05, 0.1\}$ using the measured action costs in Appendix F.

Experimental design. Because a deployed system does not know which items it answered incorrectly, we evaluate every item rather than only observed errors. To isolate the effect of re-examining the image, gains are measured against a *path-matched* baseline that follows the same critique procedure but omits the instruction to re-examine the image (Appendix J). To avoid selection bias, each

item is measured twice, with per-item maxima and regrets selected on one replicate and evaluated on the other (Appendix I, Table 10). The primary endpoint tests whether a score calibrated on certified pairs can guide routing on natural items. Its pre-registered routing-headroom gate was powered to detect $V \geq 0.03$ using 160 templates of eight items (Appendix H, Table 9). All remaining tests are secondary and Holm-corrected within each family, and intervals are clustered by record.

7 Routing Evaluation

The instrument-validation experiments in Section 5 test four pre-registered claims, all of which pass. The validation experiments show that the criterion correctly predicts which resize configurations admit exact collisions, rounded approximate constructions fail certification, certified pair members produce identical decision logits, and the digest detects changes to every recorded interface field. The primary experiment tests a fifth claim that a label-free interface score calibrated to chance when relevant evidence is provably absent can guide routing on natural items, but the score does not improve over always stopping. Two pre-registered gates make this negative result interpretable. They require measurable headroom for per-item routing and a probe that recovers known-present information without fitting shuffled labels. Both pass, ruling out insufficient routing opportunity and failure on either probe control as simple explanations.

Table 2: **Certified-pair screen and natural-stream routing.** The screen reports dominated-action rates on 80 certified Qwen2.5-VL-7B pairs (Corollary 4(a)). Natural results use 1280 VisualProcess-Bench items from 1018 records with a frozen Qwen2.5-VL-7B verifier. Gain is relative to stopping at $\lambda = 0$, with 95% intervals clustered by record. †Best of six, Holm-adjusted $p = 0.11$.

	Certified pairs	Natural stream		
Routing method	**Dominated action rate ↓**	**Gain over stop ↑** with 95% interval	**Regret ↓**	**Order accuracy ↑**
Fixed policies				
Always stop	0.000	0.000 reference	0.105	0.271
Always re-attend	1.000	+0.002 [−0.008, +0.010]	0.103	0.281
Always reason	1.000	0.000 [−0.007, +0.008]	0.104	0.188
Always re-encode	0.000	–0.021 [−0.035, −0.008]	0.127	0.260
Protocol-specified label-free policies				
Verifier confidence	1.000	–0.001 [−0.009, +0.007]	0.106	0.202
IAS-ROUTE (ours)	1.000	+0.001 [−0.007, +0.009]	0.104	0.188
Median-threshold diagnostic				
Verifier confidence	0.838	+0.015 [+0.008, +0.023]	0.089	0.344
Raw interface distance	0.838	+0.021 [+0.014, +0.028]	0.084	0.369
Random score	0.414	+0.018 [+0.009, +0.026]	0.087	0.368
IAS-ROUTE (ours)	0.838	+0.016 [+0.009, +0.024]	0.089	0.366
Oracle-supervised references				
Learned router on measured Δ, from $\mathbf{z}$	0.063	+0.003 [−0.006, +0.012]	0.102	0.259
Learned router, from $(\mathbf{z}, Q, R)$ †	0.000	+0.010 [+0.002, +0.018]	0.095	0.283
Per-item oracle (split-sample)	0.000	+0.058 [+0.050, +0.067]	0.047	0.659

Routing headroom. The best action varies across items even though no single action improves on stopping on average. The split-sample routing value is $V = 0.056$ over the best fixed action, with a record-clustered 95% interval of $[0.047, 0.062]$, and the per-item oracle gains $+0.058$ over stopping (Table 2). Against the path-matched baseline, re-attention gains $+0.002$ and further reasoning gains 0, both with intervals covering zero, while re-encoding loses 0.021 with an interval of $[-0.035, -0.008]$. Thus, positive routing value reflects item-level variation rather than an action that helps uniformly. By contrast, a split-sample ten-crop oracle gains $+0.035$ over stopping, with a 95% interval of $[+0.022, +0.047]$, while the blind crop loses 0.021 (Zhang et al., 2025). This shows additional headroom from selecting the right interface-changing view, but only with oracle supervision.

Probe controls and the primary endpoint. The probe, our decoder d_θ, behaves as expected at all three anchors. Its AUC is 0.9999 when the fact survives the interface, 0.52 under random la-

bels and exactly $1/2$ on certified pairs, where chance is guaranteed. Yet IAS-ROUTE does not improve over stopping under the protocol-specified rule, with a gain of $+0.001$ and an interval of $[-0.007, +0.009]$, and neither does verifier confidence. The median-threshold diagnostic produces gains of $+0.015$ to $+0.021$ across IAS, confidence, a random score and raw interface distance. Because the random score performs similarly and all gains disappear when the non-stop action is fixed, the improvement comes from replicate-informed action selection rather than the scores. The same pattern holds across all five source benchmarks. Each has positive routing value, but neither IAS-ROUTE nor the learned router significantly beats its best fixed action (Table 12).

Proposition 5 explains how positive routing value can coexist with this result. An interface-only router is limited to V_Φ, which may be smaller than the full per-item value V. Six oracle-supervised routers trained directly on natural-item gains using the vision tower's outputs achieve only -0.003 to $+0.003$ over stopping, with every interval covering zero (Table 11). Thus, poor transfer from certified pairs cannot by itself explain why the routers fail to improve over stopping. These gains only lower-bound V_Φ and do not establish that $V_\Phi = 0$. Six further routers read the verifier's hidden state at the decision position, a signal of $(\mathbf{z}, Q, R)$. The best gains $+0.010$, about one sixth of V, with an interval of $[+0.002, +0.018]$ that does not survive Holm correction across the six. The natural evidence remains behavioral. The verdict is unchanged across three critiques on 45.2% of 2560 measurements and 49.6% of 954 errors. When it changes, repairs and breakages offset one another, contributing $+0.030$ and -0.016 for re-attention (Appendix J, Table 13).

The screen in practice. On a certified pair, Corollary 4(a) proves that att and rea are dominated by stopping for every $\lambda > 0$. A router that recognizes this limit should avoid both actions, yet none of the label-free scores does so. Under the median rule, every score except the random control acts on all 80 certified pairs but only half of the natural stream. The random control acts on half of each. The protocol-specified rule acts on all items and sends every certified pair to reasoning. Routers trained on the vision tower's outputs behave less uniformly (Table 11). One stops on all 80 pairs. Another acts on 70 but selects re-encoding, which the corollary does not rule out, on 65 of them. The remaining four select a dominated action on between 6 and 80 pairs. Five of the six hidden-state routers stop on all 80. The screen therefore distinguishes policies that avoid provably wasted computation from those that do not.

Limitations. All certified witnesses use one resize-based mechanism in one modality, so broader applications of Proposition 5 remain untested. Our exclusions cover only exact-kernel collisions that can be added to any base image (Remark 10). The construction requires a fixed-point resampler whose filter and resize ratio yield an integer kernel vector that fits within the pixel range. Among the bicubic configurations we resolve, only integer resize ratios are constructible, and the construction is unavailable when a deployment forwards images at native resolution (Dehghani et al., 2023; Wang et al., 2024). An adversary can choose an eligible input resolution, but natural images rarely arrive at one. We therefore resized the benchmark images into the eligible regime, where only a small minority supported a legible mark. We do not measure how often natural errors are interface-limited, and identifying such errors would require mining natural failures. Every certificate also assumes that all image-dependent interface fields have been enumerated correctly (Appendix C). Finally, because the primary endpoint is null, we do not ablate the components of IAS.

8 CONCLUSION

Perceptual failures are usually inferred from behavior. When pre-rounding resize is an exact integer map, a certified pair makes one class provable: members with different correct answers have identical interface states, so no verifier or language-side repair can beat chance. From the resize configuration, ALIASFORGE either constructs such a pair or certifies its absence within the specified family. Certified pairs also expose routers that waste compute on re-attention or further reasoning. Natural items exhibit routing headroom, but no tested interface-only router improves over stopping. The fiber ceiling explains how this can occur: a router cannot distinguish inputs within a fiber (Proposition 5). Extending certification beyond fixed-point resampling is the next step.

ETHICS STATEMENT

This work has dual-use implications. Although certified aliases are introduced as a measurement tool, they could also be used to evade moderation or safety verifiers. Related image-scaling attacks and defenses are already public (Xiao et al., 2019; Quiring et al., 2020), and aliases for a known preprocessing pipeline are inexpensive to regenerate. We therefore judge that releasing the constructor with a defense checklist offers greater benefit than withholding it. We document susceptible configurations and practical mitigations involving resampling, input resolution, and preprocessing verification. We use only public datasets under their licenses and collect no new human annotations.

REPRODUCIBILITY STATEMENT

All theoretical claims state their assumptions in Section 4 and are proved in Appendix A. The supplementary material provides machine checks using exhaustive enumeration, exact rational arithmetic, or numerical validation as appropriate. It also contains a self-contained implementation of the deterministic alias construction and certification procedure, along with per-item records and analysis code for reproducing every reported statistic. The complete protocol, cost accounting, fold assignments, and model and dataset versions are documented. All models are frozen and evaluations are inference-only, so reproducing the results requires only modest hardware.

AI USE STATEMENT

Large language models assisted with writing and editing, finding and verifying related work and references, drafting formal statements and proofs, proposing and refining hypotheses, designing experiments, and implementing experiment, analysis, and figure code. They were not used to collect or annotate data or to interpret results. Reported numbers were computed from experimental records by tested analysis code rather than generated by a language model. The authors reviewed all AI-assisted content and checked the formal claims using exhaustive enumeration, exact arithmetic, and numerical tests as appropriate. The authors take full responsibility for the paper and any remaining errors.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, June 1953. ISSN 0003-4851. doi: 10.1214/aoms/1177729032. URL http://dx.doi.org/10.1214/aoms/1177729032.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible NLP attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1987–2004, 2022. doi: 10.1109/SP46214.2022.9833641.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, Dec 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4231-5. URL https://doi.org/10.1007/s11432-024-4231-5.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information, 2025. URL https://arxiv.org/abs/2505.23558.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102783–102802, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ff414825df833edb8b1839e3d5d495e9-Paper-Conference.pdf.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lucic, and Neil Houlsby. Patch n' pack: NaViT, a vision transformer for any aspect ratio and resolution. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 2252–2274. Curran Associates, Inc., 2023. doi: 10.52202/075280-0106. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/06ea400b9b7cfce6428ec27a371632eb-Paper-Conference.pdf.

Xingjian Diao, Zheyuan Liu, Chunhui Zhang, Weiyi Wu, Keyi Kong, Lin Shi, Kaize Ding, Soroush Vosoughi, and Jiang Gui. Addressing overthinking in large vision-language models via gated perception-reasoning optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4393–4410, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.215. URL https://aclanthology.org/2026.findings-acl.215/.

Hao Ding, Zhichuan Yang, Weijie Ge, Ziqin Gao, Chaoyi Lu, and Lei Zhao. TikArt: Stabilizing aperture-guided fine-grained visual reasoning with reinforcement learning, 2026. URL https://arxiv.org/abs/2602.14482.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Kai Hu, Akash Bharadwaj, Weichen Yu, and Matt Fredrikson. Steal the patch size: Adversarially manipulate vision language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=LrMoEItzRJ.

Ahmadreza Jeddi, Minh Ngoc Le, Amirhossein Kazerouni, Hakki Can Karaimer, Hue Nguyen, Iqbal Mohomed, Michael Brudno, Alex Levinshtein, Konstantinos G. Derpanis, Babak Taati, and Radek Grzeszczuk. AVIS: Adaptive test-time scaling for vision-language models, 2026. URL https://arxiv.org/abs/2606.11576.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer New York, 1986. ISBN 9781461249467. doi: 10.1007/978-1-4612-4946-7. URL http://dx.doi.org/10.1007/978-1-4612-4946-7.

A. K. Lenstra, H. W. Lenstra, and L. Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261(4):515–534, Dec 1982. ISSN 1432-1807. doi: 10.1007/BF01457454. URL https://doi.org/10.1007/BF01457454.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulić, and Anders Søgaard. Lost in embeddings: Information loss in vision–language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational*

Linguistics: EMNLP 2025, pp. 22676–22693, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1235. URL https://aclanthology.org/2025.findings-emnlp.1235/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

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

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving vision-language models with perception-centric process reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33099–33109, June 2026.

Brandon Ong, Tej Deep Pala, Vernon Toh, William Chandra Tjhi, and Soujanya Poria. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned, 2025. URL https://arxiv.org/abs/2509.23250.

Hiroto Osaka, Shohei Taniguchi, Gouki Minegishi, Kai Yamashita, Masahiro Suzuki, and Yutaka Matsuo. Visual access boundaries in vision-language model reasoning, 2026. URL https://arxiv.org/abs/2607.12815.

Ruiying Peng, Xueyu Wu, Jing Lei, Lu Hou, Yuanzheng Ma, and Xiao-Hui Li. Deeper thought, weaker aim: Understanding and mitigating perceptual impairment during reasoning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12064–12073, June 2026.

Erwin Quiring, David Klein, Daniel Arp, Martin Johns, and Konrad Rieck. Adversarial preprocessing: Understanding and preventing image-scaling attacks in machine learning. In *Proceedings of the 29th USENIX Security Symposium*, pp. 1363–1380. USENIX Association, August 2020. ISBN 978-1-939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/quiring.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 18–34, December 2024.

Shaeke Salman, Md Montasir Bin Shams, and Xiuwen Liu. Intriguing equivalence structures of the embedding space of vision transformers, 2024. URL https://arxiv.org/abs/2401.15568.

Shaeke Salman, Montasir Shams, Mao Nishino, and Xiuwen Liu. Unaligning everything: Or aligning any text to any image in multimodal models. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2025. doi: 10.1109/IJCNN64981.2025.11228287.

Fawaz Sammani, Tzoulio Chamiti, and Nikos Deligiannis. On test-time scaling for vision-language models. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 168–185, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37271-0.

Vasileios Sevetlidis. Bayes-sufficient representations in supervised learning, 2026a. URL https://arxiv.org/abs/2606.04045.

Vasileios Sevetlidis. A fiber criterion for representation identifiability in supervised learning, 2026b. URL https://arxiv.org/abs/2606.01092.

Haz Sameen Shahgir, Xiaofu Chen, Yu Fu, Erfan Shayegani, Nael Abu-Ghazaleh, Yova Kementchedjhieva, and Yue Dong. VLMs need words: Vision language models ignore visual detail in favor of semantic anchors. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=6DJpzvjWbK.

Omar Sharif, Eftekhar Hossain, Nikhil Singh, and Patrick Ng. Disentangling perception and reasoning in multimodal LLMs via reward design, 2026. URL https://arxiv.org/abs/2601.00215.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? Uncovering the illusion of visual re-examination. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=DdU1o2ZvWi.

Zeru Shi, Kai Mei, Yihao Quan, Dimitris N. Metaxas, and Ruixiang Tang. Improving visual reasoning with iterative evidence refinement, 2026b. URL https://arxiv.org/abs/2603.14117.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1230. URL https://aclanthology.org/2025.acl-long.1230/.

Sho Takishita, Jay Gala, Abdelrahman Mohamed, Kentaro Inui, and Yova Kementchedjhieva. LLMs can compensate for deficiencies in visual representations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15253–15272, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.825. URL https://aclanthology.org/2025.findings-emnlp.825/.

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

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387790527. doi: 10.1007/b13794. URL http://dx.doi.org/10.1007/b13794.

P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, Englewood Cliffs, NJ, 1993. ISBN 978-0-13-605718-5.

Khang Nhat Hoang Vo, Artem Vazhentsev, Artem Shelmanov, Timothy Baldwin, and Yova Kementchedjhieva. Failing to see or failing to know? Attributing errors in vision-language models, 2026. URL https://arxiv.org/abs/2607.04683.

Haozhe Wang, Qixin Xu, Changpeng Wang, Taofeng Xue, Chong Peng, Wenhu Chen, and Fangzhen Lin. Bad seeing or bad thinking? Rewarding perception for multimodal reasoning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=duzdftKtUA.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. VisualPRM: An effective process reward model for multimodal reasoning, 2025. URL https://arxiv.org/abs/2503.10291.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=r7uOjvZdzO.

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

Qixue Xiao, Yufei Chen, Chao Shen, Yu Chen, and Kang Li. Seeing is not believing: Camouflage attacks on image scaling algorithms. In *Proceedings of the 28th USENIX Security Symposium*, pp. 443–460, Santa Clara, CA, USA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL https://www.usenix.org/conference/usenixsecurity19/presentation/xiao.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sk9yuql0Z.

Senqiao Yang, Junyi Li, Xin Lai, Jinming Wu, Wei Li, Zejun Ma, Bei Yu, Hengshuang Zhao, and Jiaya Jia. VisionThink: Smart and efficient vision language model via reinforcement learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 95187–95227. Curran Associates, Inc., 2025. doi: 10.52202/085713-3182. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/88be023075a5a3ff3dc3b5d26623fa22-Paper-Conference.pdf.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning, 2026. URL https://arxiv.org/abs/2602.08236.

M Zeeshan and Saud Satti. Chameleon: Adaptive adversarial agents for scaling-based visual prompt injection in multimodal AI systems, 2025. URL https://arxiv.org/abs/2512.04895.

Chi Zhang, Haibo Qiu, Qiming Zhang, Yufei Xu, Zhixiong Zeng, Siqi Yang, Peng Shi, Lin Ma, and Jing Zhang. Perceptual-evidence anchored reinforced learning for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41111–41120, June 2026.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/aaa0ac4253da75faf9b0dc0dda062612-Paper-Conference.pdf.

A PROOFS

Definition 7 (Language-side class)**.** Fix a question and candidate step (Q, R). A *language-side intervention* is a measurable function $\psi(\mathbf{z}, Q, R, \omega)$, where the randomness ω is drawn independently of $(\mathbf{x}, \mathbf{z})$. A finite adaptive procedure also belongs to this class because its final output remains a measurable function of $(\mathbf{z}, Q, R, \omega)$. Re-supplying the identical image to a deterministic processor is language-side, and enc is not because it recomputes $\mathbf{z}$.

Proposition 3. Let $\nu = \mu_0 + \mu_1$, write q_0 and q_1 for the corresponding densities, and define $A^\star = \{q_0 > q_1\}$. A deterministic verifier whose positive decision region is A has $\mathrm{BA} = \frac{1}{2}\big(1 + \mu_0(A) - \mu_1(A)\big)$, and for every measurable A,

$$\mu_0(A) - \mu_1(A) = \int_A (q_0 - q_1)\, d\nu \le \int_{A^\star} (q_0 - q_1)\, d\nu = \mathrm{TV}(\mu_0, \mu_1),$$

with equality when $A = A^\star$. Randomized rules are mixtures of deterministic rules and cannot exceed this maximum, which proves (b). If the interface is deterministic and $\mathbf{z}_0 \neq \mathbf{z}_1$, then μ_0 and μ_1 are distinct point masses and $\mathrm{TV}(\mu_0, \mu_1) = 1$. For (a), sufficiency holds because a verifier, before or after a language-side action, is a Markov kernel on $(\mathbf{z}, Q, R)$, so $\mu_0 = \mu_1$ gives $p_0 = p_1$, and necessity follows from (b) because $\mathrm{TV}(\mu_0, \mu_1) > 0$ whenever $\mu_0 \neq \mu_1$. □

Corollary 4(b). Partition the population into the certified fraction κ and the remainder. Every verifier attains exactly $\frac{1}{2}$ on the first by Theorem 2(i) and at most 1 on the second, so the aggregate is at most $\kappa \cdot \frac{1}{2} + (1 - \kappa) \cdot 1 = 1 - \kappa/2$, with equality when the remainder is answered perfectly. □

Corollary 4(a). By Theorem 2(ii), $\Delta_i(a) = 0$ for every $a \in \{\mathsf{stop}, \mathsf{att}, \mathsf{rea}\}$ on a certified pair, so $g_i(a) = -\lambda c(a)$ for these actions. Because $c(\mathsf{stop}) = 0$, $c(\mathsf{att}), c(\mathsf{rea}) > 0$ and $\lambda > 0$, we have $g_i(\mathsf{att}), g_i(\mathsf{rea}) < 0 = g_i(\mathsf{stop})$. Within $\mathcal{A}$, stopping is optimal exactly when $\Delta_i(\mathsf{enc}) \le \lambda c(\mathsf{enc})$, which is why the screen treats enc separately. □

Proof of Theorem 2. By Definition 1, both members produce the same state $\mathbf{s}_k$. Assumption 1(a) then gives the same conditional law of $\mathbf{z}$ under both members. Together with Assumption 1(b), this implies that the joint law of $(\mathbf{z}, Q, R, \omega)$ is identical. By Definition 7, every language-side intervention is a measurable function of this tuple, so its push-forward law is also identical.

For any binary rule applied to this common output law, let p be the probability of predicting label 1. Because the two members have labels 1 and 0, $\mathrm{BA} = \frac{1}{2}\big(p + (1 - p)\big) = \frac{1}{2}$, which proves (i). The same calculation holds before and after any language-side action, so $\Delta_i(a) = 0$, proving (ii). For (iii), an action that produces a shared post-action state at stage k and preserves Assumption 1 yields an interface alias at stage k, so the same argument gives $\Delta_i(a) = 0$. □

Remark 8 (Which signal the ceiling refers to). Proposition 5 is stated for a signal $S = \sigma(\mathbf{z}, Q, R)$ because the bound depends on what the router reads. Let $\mathbf{z}$ be constant and Q be uniform on $\{Q', Q''\}$, and suppose action a_1 gains 1 under Q' and 0 under Q'', with the gains reversed for a_2. Then $V_\Phi = 0$, but a policy reading Q gains $1/2$ over the best fixed action. Conversely, in our natural stream the tuple $(\mathbf{z}, Q, R)$ identifies the item, so $V_{(\mathbf{z},Q,R)} = V$ and the ceiling for that signal is vacuous without restricting the router class. IAS, interface distance and the vision-tower routers read $\mathbf{z}$ alone, while verifier confidence and the hidden-state routers read $(\mathbf{z}, Q, R)$.

Proposition 9 (Non-constructibility bound)**.** *Fix a deployed resize with fixed-point integer coefficient matrix K and a window of W input columns. Let K_W be the restriction of K to that window, and let $L = \ker K_W \cap \mathbb{Z}^W$. The construction forms the image difference $\mathbf{D} = uv^\top$ inside the window and sets it to zero elsewhere, where u is a binary vector selecting the affected image rows and $v \in L$ supplies the column offsets. We call v* realizable *when $v \neq 0$ and $\|v\|_\infty \le 255$. For any basis of L with Gram–Schmidt vectors $b_1^*, \ldots, b_n^*$, every nonzero $v \in L$ satisfies $\|v\|_2 \ge \min_i \|b_i^*\|$ and therefore $\|v\|_\infty \ge \min_i \|b_i^*\| / \sqrt{W}$. Consequently, if $\min_i \|b_i^*\| > 255\sqrt{W}$, no realizable perturbation in this rank-one family can be supported on that window.*

The bound follows because the shortest nonzero vector of a lattice is at least as long as the shortest Gram–Schmidt vector of any of its bases (Lenstra et al., 1982). It is basis-dependent, window-local

and one-sided. Exceeding the threshold certifies impossibility, while falling below it establishes only that a search may succeed.

Arbitrary-rank extension of Proposition 9. Assume the same fixed-point operator K_W acts along both image axes, and let $\mathbf{D}$ be an integer difference whose support the window covers along both axes. The deployed resampler applies K_W one axis at a time, so a difference that collides on every unclipped base is annihilated by the first pass and therefore satisfies $K_W \mathbf{D} K_W^\top = 0$. We bound this larger exact-kernel family. The rank-one family $uv^\top$ does not exhaust the solutions. For example, the integer kernel of $G = \left[\begin{smallmatrix} 160 & 96 & 0 \\ 0 & 96 & 160 \end{smallmatrix}\right]$ is spanned by $(3, -5, 3)$, yet $\mathbf{D} = \left[\begin{smallmatrix} 0 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{smallmatrix}\right]$ has rank 2 and sup-norm 1 and satisfies $G\mathbf{D}G^\top = 0$. Define

$$\tau_\infty(L) = \min_{v \in L \setminus \{0\}} \|v\|_\infty \quad \text{and} \quad M = \max_a \|\mathrm{row}_a(K_W)\|_1,$$

and let $Y = K_W \mathbf{D}$. Since $Y K_W^\top = 0$, the transpose of every row of Y lies in L. If $Y = 0$, every column of $\mathbf{D}$ lies in L, so $\|\mathbf{D}\|_\infty \geq \tau_\infty(L)$. If $Y \neq 0$, a nonzero row $Y_a = \mathrm{row}_a(K_W)\mathbf{D}$ satisfies $\tau_\infty(L) \leq \|Y_a\|_\infty \leq \|\mathrm{row}_a(K_W)\|_1 \|\mathbf{D}\|_\infty \leq M\|\mathbf{D}\|_\infty$. Because $M \geq 1$, both cases give $\|\mathbf{D}\|_\infty \geq \tau_\infty(L)/M$.

Substituting the lower bound of Proposition 9 for $\tau_\infty(L)$, the resulting bounds at $1.5\times$ are 3.3×10^5 for bicubic, 2.9×10^5 for bilinear, and 3.1×10^5 for Lanczos, all above 255, so these configurations admit no nonzero window-supported integer collision of any rank in the exact kernel. In the principal case of Theorem 6, every row of Y and, when $Y = 0$, every column of an interior $\mathbf{D}$ is an interior kernel vector, so $\tau_\infty(L)$ can be replaced by $\beta(K)$. At $1.5\times$, this gives 5.7×10^3, 1.2×10^6 and 964 for any interior support (Table 3). The exclusion at $2.5\times$ and the Lanczos exclusions in Table 5 apply only to the rank-one family. □

Remark 10 (Exact kernel and output quantization). A collision need not arise from exact annihilation. The deployed resize rounds and clips its integer accumulator, so distinct pre-quantization values can produce the same 8-bit output, depending on the base image. A perturbation annihilated by the first resize pass instead collides on every base for which it is realizable without input clipping. Our non-constructibility results rule out only exact-kernel perturbations within the stated construction family.

Proof of Proposition 5. For (i), let $p_\pi(a \mid S)$ be the probability that $\pi(S, \omega)$ selects action a. Because ω is independent of the item, $\mathbb{E}[g(\pi) \mid S] = \sum_a p_\pi(a \mid S)\mathbb{E}[g(a) \mid S] \leq \max_a \mathbb{E}[g(a) \mid S]$. Taking expectations and subtracting $\max_a \mathbb{E}[g(a)]$ gives the bound. Because $\mathcal{A}$ is finite, a fixed tie-breaking rule makes $\pi_S(S) \in \arg\max_a \mathbb{E}[g(a) \mid S]$ measurable, and this policy attains equality.

For (ii),

$$\max_a \mathbb{E}[g(a)] \leq \mathbb{E}\Big[\max_a \mathbb{E}[g(a) \mid S]\Big] \leq \mathbb{E}\Big[\max_a g(a)\Big].$$

The first inequality holds because constant policies read any signal, and the second follows from the conditional expectation of the pointwise maximum. If $S' = h(S)$, the tower property and the same inequality applied conditionally on S' give $V_{S'} \leq V_S$, the statement that garbling a signal cannot increase its value (Blackwell, 1953).

For (iii), $V - V_S = \mathbb{E}[\max_a g(a) - g(\pi_S(S))]$ has a nonnegative integrand, so $V_S = V$ exactly when $\pi_S(S)$ selects a per-item optimal action almost surely. □

Proof of Theorem 6. Indexing. Number the output columns so that column $j = qn + \varphi$, with $n \in \mathbb{Z}$ and phase $\varphi \in \{0, \ldots, q-1\}$, reads input columns $pn + o_\varphi + t$ for $t = 0, \ldots, T_\varphi - 1$ with weights $k_\varphi[t]$. Away from the border, the offsets o_φ and the weights depend only on φ, which is the periodicity used in Section 5. Write each input index as $pn + o_\varphi + t = p(n + e) + s$ with residue $s \in \{0, \ldots, p-1\}$, and set $h_{\varphi s}[e] = k_\varphi[pe + s - o_\varphi]$, taken as zero outside the weight range. With $v_s[n] = v[pn + s]$, the equation $(Kv)_{qn+\varphi} = 0$ reads

$$\sum_{s=0}^{p-1} \sum_e h_{\varphi s}[e]\, v_s[n + e] = 0 \qquad \text{for all } n \in \mathbb{Z} \text{ and } \varphi.$$

Define the Laurent polynomials $\tilde{H}_{\varphi s}(\zeta) = \sum_e h_{\varphi s}[e]\zeta^{-e}$ and $\hat{v}_s(\zeta) = \sum_n v_s[n]\zeta^n$. The left-hand side is the coefficient of ζ^n in $\sum_s \tilde{H}_{\varphi s}(\zeta)\hat{v}_s(\zeta)$, so $Kv = 0$ on the interior if and only if $\tilde{H}(\zeta)\hat{v}(\zeta) = 0$. Multiplying row φ by a suitable power of ζ turns $\tilde{H}$ into the polynomial matrix $H \in \mathbb{Z}[\zeta]^{q\times p}$ of Section 5 without changing the solution set, and a solution $\hat{v}$ has finite support if and only if v does.

Existence. Over the field $\mathbb{Q}(\zeta)$, H has rank at most q, so its null space has dimension at least $p-q \geq 1$. Any nonzero rational null vector, multiplied by the least common multiple of its denominators, is a nonzero vector in $\mathbb{Z}[\zeta]^p$ annihilated by H. The set of all such vectors is a $\mathbb{Z}[\zeta]$-submodule of $\mathbb{Z}[\zeta]^p$ that spans the null space, so its rank equals that dimension. Shifting a solution into the interior of the image gives a finitely supported integer kernel vector of K, provided the interior is at least as wide as its support.

Principal case. Let $p = q+1$ and suppose H has full row rank, so the null space over $\mathbb{Q}(\zeta)$ is one-dimensional. Let $r_s = (-1)^s \det H^{(s)}$, where $H^{(s)}$ deletes column s. For each row φ, $\sum_s H_{\varphi s} r_s$ is the Laplace expansion of the determinant of the $(q+1)\times(q+1)$ matrix obtained by prepending row φ of H to H. That matrix has a repeated row, so $Hr = 0$, and $r \neq 0$ because some maximal minor is nonzero. Let $\gamma = \gcd(r_0, \ldots, r_{p-1})$ in $\mathbb{Z}[\zeta]$, including integer content, and $w = r/\gamma$, so that $\gcd(w_0, \ldots, w_{p-1}) = 1$. Now let $\hat{v} \in \mathbb{Z}[\zeta]^p$ be any integer solution. Since the null space is one-dimensional, $\hat{v} = (\xi_1/\xi_2)\, w$ with $\xi_1, \xi_2 \in \mathbb{Z}[\zeta]$ coprime and $\xi_2 \neq 0$. Then ξ_2 divides $\xi_1 w_s$ for every s, so by Gauss's lemma ξ_2 divides every w_s and hence their greatest common divisor 1. Thus $\xi_2 = \pm 1$ and $\hat{v} = \xi w$ with $\xi = \pm\xi_1$. For the bound, fix s with $w_s \neq 0$. The highest-order coefficient of ξw_s is $\mathrm{lc}(\xi)\,\mathrm{lc}(w_s)$ and its lowest-order nonzero coefficient is $\mathrm{tc}(\xi)\,\mathrm{tc}(w_s)$. Both $\mathrm{lc}(\xi)$ and $\mathrm{tc}(\xi)$ are nonzero integers, so $\|\xi w_s\|_\infty \geq \max\{|\mathrm{lc}(w_s)|, |\mathrm{tc}(w_s)|\}$. Taking the maximum over s gives $\|v\|_\infty \geq \beta(K)$. Moving the support multiplies $\hat{v}$ by a power of ζ, which changes neither coefficient. Finally, $\hat{v} = w$ is itself a solution of sup-norm $\|w\|_\infty$. □

Remark 11 (Border columns). Within one filter support of the left or right image edge, the resampler clips and renormalizes its weights, so the periodic description and the bound $\beta(K)$ do not apply there. We call an input column *interior* when every output column that reads it has periodic weights. A carrier that touches the border must be checked with Proposition 9 on that window. AliasForge places carriers on interior columns.

Table 3: **Window-free bounds in the principal case.** $\beta(K)$ bounds every interior integer kernel vector from below at any support width, $\|w\|_\infty$ is the sup-norm of the generator, and $\beta(K)/M$ bounds interior image differences of any rank, where M is the largest absolute row sum of K. Orange entries exceed 255 and exclude a realizable perturbation. Blue entries are attained by the generator. The box operator at $1.5\times$ fails the bit-exact reconstruction check of Table 5.

	2×			**1.5×**			**4/3×**		
Filter	$\beta(K)$	$\|w\|_\infty$	$\beta(K)/M$	$\beta(K)$	$\|w\|_\infty$	$\beta(K)/M$	$\beta(K)$	$\|w\|_\infty$	$\beta(K)/M$
Bicubic	**3**	**3**	< 1	2.8×10^{10}	6.7×10^{12}	5.7×10^{3}	1.2×10^{16}	7.1×10^{18}	2.4×10^{9}
Bilinear	**3**	**3**	< 1	5.2×10^{12}	5.2×10^{12}	1.2×10^{6}	2.3×10^{12}	2.3×10^{12}	5.5×10^{5}
Box	**1**	**1**	< 1		*gate failed*		**1**	**1**	< 1
Lanczos	6.3×10^{4}	1.9×10^{6}	< 1	5.6×10^{9}	6.8×10^{12}	964	1.2×10^{15}	2.3×10^{19}	2.1×10^{8}

Computed bounds. Table 3 evaluates Theorem 6 on the reconstructed operators at the three principal ratios. We compute the maximal minors and their greatest common divisor in exact integer arithmetic and confirm that the operator maps the generator w exactly to zero. In every case the interior columns are exactly periodic and H has full row rank. Where $\beta(K)$ exceeds 255, no realizable interior kernel vector exists at any support width, and where $\beta(K)/M$ also exceeds 255, no interior image difference of any rank lies in the exact kernel.

B The AliasForge procedure

Algorithm 1 separates AliasForge into two procedures. Audit runs once per deployed configuration without model access or image content. It accepts the reconstructed operator K only after

reproducing the deployed resize bit for bit. When a pipeline applies multiple resizes to the same pixels, it stacks their operators so that the resulting kernel is the intersection of their individual kernels, and it searches the integer kernel with LLL basis reduction (Lenstra et al., 1982). CONSTRUCT then builds a pair whose desired fact values are ℓ_0 and ℓ_1. A NOT CONSTRUCTIBLE verdict applies to the rank-one family $uv^\top$.

Algorithm 1 ALIASFORGE configuration audit and pair construction

procedure AUDIT(Φ, W)
  Enumerate the image-dependent fields $\mathcal{F}$ emitted by Φ
  Perturb each field separately and verify that the digest changes
  Reconstruct the integer resize operators from their fixed-point coefficients
  Stack operators applied to the same pixels into K
  if K does not reproduce the deployed resizes bit for bit **then**
    return AUDIT FAILURE
  Compute and LLL-reduce an exact $\mathbb{Z}$-basis B of $L = \ker K_W \cap \mathbb{Z}^W$
  if $L = \{0\}$ **then**
    return NOT CONSTRUCTIBLE
  Compute the Gram–Schmidt vectors $b_1^*, \ldots, b_n^*$ of B
  if $\min_i \|b_i^*\|_2 > 255\sqrt{W}$ **then**
    return NOT CONSTRUCTIBLE ▷ Proposition 9
  Search L for $v \neq 0$ with $\|v\|_\infty \leq 255$
  if no such v is found **then**
    return UNRESOLVED
  return $(\mathcal{F}, K, v)$

procedure CONSTRUCT($\mathbf{x}, \ell_0, \ell_1, \mathcal{F}, K, v$)
  Choose an integer amplitude $m \geq 1$ with $m\|v\|_\infty \leq 255$
  Place copies of v at valid column offsets to form a carrier $\tilde{v}$ satisfying $K\tilde{v} = 0$
  Choose binary row selectors u_0 and u_1 that render fact values ℓ_0 and ℓ_1 without clipping
  Form $\mathbf{x}_b$ by adding $m\tilde{v}$ to $\mathbf{x}$ on the rows selected by u_b, for $b \in \{0, 1\}$
  if $\Phi(\mathbf{x}_0)$ and $\Phi(\mathbf{x}_1)$ have equal digests over $\mathcal{F}$ **then**
    return the certified pair and its certification record
  return FAIL

Construction settings. For the $896 \to 448$ bicubic resize, the shortest integer kernel vector is $v = (1, -3, 3, -1)$, tiled across valid column offsets to form the full-width carrier $\tilde{v}$. Every certified pair in Table 1 uses amplitude $m = 20$, giving a maximum pixel offset of 60. The tiling pipeline uses a vector of sup-norm 7, giving a maximum offset of 140. The fact values are $\ell_0 = 2$ and $\ell_1 = 3$ visible horizontal bands, so $\mathbf{x}_1 - \mathbf{x}_0 = m\tilde{v}$ on the added rows and is zero elsewhere.

Shapes beyond bands. Any image difference $\mathbf{D}$ whose rows lie in $\ker K$ satisfies $K\mathbf{D}K^\top = 0$. Figure 1 applies this row-wise construction to digit glyphs that visibly depict a 3 and an 8. The resulting difference has rank 18 and lies outside the rank-one family bounded by Proposition 9.

C INTERFACE COMPLETENESS CHECKLIST

For each scored prompt, the certificate hashes every field returned by the deployed processor (Table 4). Arrays are serialized with their field name, data type, byte order, shape and raw bytes, while scalars and sequences use a canonical serialization. Fields are ordered by name before being combined into one SHA-256 digest. Downstream image-dependent quantities, including projected visual tokens, position identifiers and expanded image placeholders, are determined by the hashed bundle and the shared non-image inputs.

Validation. Perturbing one element in each enumerated field changed the digest for all 5 Qwen2.5-VL fields and all 3 LLaVA-1.5 fields. Re-hashing five certified pairs per target with the complete scored prompt (Q, R) left every pair identical across all fields.

Table 4: **Fields included in the certificate digest.** Shapes are shown for the audited scored prompt. Image dependence indicates whether a field varies with pixel content, only with image geometry, or not with the image.

		Shape		
Processor field	**Data type**	**Qwen2.5-VL-7B**	**LLaVA-1.5-7B**	**Image dependence**
pixel_values	float32	(1024, 1176)	(1, 3, 336, 336)	content
image_grid_thw	int64	(1, 3)	not emitted	geometry
input_ids	int64	(1, 321)	(1, 617)	none
attention_mask	int64	(1, 321)	(1, 617)	none
mm_token_type_ids	int64	(1, 321)	not emitted	none

D ALIAS CONSTRUCTION IN DETAIL

Windowed criterion. We apply Proposition 9 to 32-column windows at stride 16 and take the minimum bound over the windows. A window of W columns at stride t contains every support of width at most $W - t + 1$, so the 32-column scan certifies supports up to 17 columns wide and the 256-column scan up to 241. Widening the windows to 64, 128 or 256 columns changes no verdict among the 27 filter-and-ratio configurations that pass the bit-exact reconstruction check. Clearing denominators in a basis of the rational kernel may span only a strict subset of L, so we obtain a $\mathbb{Z}$-basis using integer column operations of determinant ± 1, shorten it with LLL and evaluate the Gram–Schmidt bound in exact rational arithmetic.

Resize results. We reconstructed K for 28 resampling-filter and ratio combinations (Table 5). The reconstruction matched the library bit for bit in 27 cases, and the criterion returned a definitive verdict in 26. Unlike bicubic, Lanczos is ruled out through $3.01\times$ but succeeds at $6\times$ and $8\times$. Each positive result is verified by resizing a constructed pair, and Table 7 reports the deployed processors that use other resampling filters. The $4\times$ Lanczos case remains unresolved because the bound falls below 255 but the reduced basis yields no realizable vector, and the box operator at $1.5\times$ fails the bit-exact reconstruction gate.

Table 5: **Constructibility by resampling filter and resize ratio.** Bold numbers give the sup-norm of a constructed integer kernel vector. Entries of the form $\geq B$ give certified lower bounds. Because each such bound exceeds 255, the rank-one construction is excluded. At $1.5\times$ and for Lanczos at $2\times$, Theorem 6 makes the exclusion independent of the window (Table 3). *Unresolved* means that the bound falls below 255 but no realizable vector was found. *Gate failed* means that the reconstructed operator did not reproduce the deployed library bit for bit.

	Resize ratio						
Filter	**1.5×**	**2×**	**3×**	**3.01×**	**4×**	**6×**	**8×**
Bicubic	$\geq 1.6\times10^{12}$	**3**	**13**	≥ 735	**5**	**2**	**2**
Bilinear	$\geq 1.2\times10^{12}$	**3**	**2**	**1**	**1**	**1**	**1**
Box	*gate failed*	**1**	**1**	**1**	**1**	**1**	**1**
Lanczos	$\geq 1.8\times10^{12}$	$\geq 4.8\times10^{5}$	$\geq 1.6\times10^{3}$	$\geq 2.0\times10^{3}$	*unresolved*	**35**	**16**

Configuration screen. Table 6 applies the criterion to 20 declared bicubic-resize configurations, each a released processor's resize at a stated input and output size. Configurations with the same resampling filter and input-output geometry share the same integer operator. Seventeen configurations across four resize ratios are constructible, and three across two ratios are excluded within the rank-one family. Four cases use declared geometries that differ from the processors' default behavior (Table 7), so their verdicts should not be interpreted as certificates for the default pipelines.

Where collisions can be constructed. Whether ALIASFORGE can construct a collision at a preprocessing stage depends on the deployed configuration, which we determine from its implementation. A collision at resizing identifies information lost before the learned encoder, whereas a collision at token selection would identify information lost by the encoder. We exclude post-encoder pruning because global attention allows retained token values to depend on discarded content, violating Assumption 1(a). Across the three target pipelines, only resizing produced certified pairs. Cropping

Table 6: **Screen of 20 declared bicubic-resize configurations.** The bound column gives the lower bound of Proposition 9 on $\|v\|_\infty$, and a value above 255 excludes the construction, and at $1.5\times$ the exclusion holds for every interior support (Theorem 6). For constructible ratios, the witness column gives the smallest $\|v\|_\infty$ of a kernel vector found. A dagger marks a configuration evaluated at a declared geometry that differs from the processor's default, so its verdict does not apply to the default pipeline.

Ratio	Bound	Witness	**Configurations** model and output side in pixels
Constructible integer ratios, 17 configurations			
2×	0.71	3	CLIP 224 · PaliGemma 224 · Qwen2.5-VL 224 · LLaVA-1.5 336 Phi-3.5-V 336† · SigLIP 384 · SmolVLM 384† · InternVL thumbnail 448 Qwen2.5-VL 448 · Pixtral 512 · Cambrian 576
3×	4.17	13	BLIP-2 224 · SigLIP 384 · InternVL tile 448
4×	1.25	5	CLIP 224 · LLaVA-1.5 336
6×	0.59	2	Molmo 336
Excluded non-integer ratios, 3 configurations			
1.5×	1.6×10^{12}	none	Fuyu 300† · Idefics2 378†
2.5×	1.8×10^{4}	none	MiniCPM-V 448

and tiling produced no candidates because the tested pipelines preserve all image regions, and we did not evaluate pre-encoder masking.

Deployment dependence. A resize without downscaling is an identity map and admits no construction. One target resizes only when the image exceeds a pixel budget (Wang et al., 2024), so an aligned image within that budget passes through unchanged, and lowering the budget introduces downscaling. If the pipeline also appends a thumbnail (Chen et al., 2024; Li et al., 2025a), a perturbation must be removed by both resize operations.

Input transport. A certificate applies to the exact 8-bit pixel arrays received by the processor. Lossy re-encoding during transport, such as JPEG compression, invalidates it unless equality is verified again after decoding. The audited Qwen and LLaVA processors resize 8-bit values before rescaling or normalization, so no floating-point transformation precedes the certified collision.

Natural images. On a natural image, every affected pixel must remain within $[0, 255]$ after adding the carrier. A solid band across the image therefore certified none of the 200 benchmark images at any tested amplitude. Because headroom is required only on affected rows, we retain eligible rows and allow the bands, which we call marks there, to vary in height, on benchmark images resized to 896×896 for the deployed $2\times$ regime. None of the 200 images has a native size at which the processor's scale factor is an integer. At amplitude 20, 48 of 200 images certify with all three marks aligned in one column, and 94 certify at amplitude 5. Requiring each mark to occupy a flat, mid-tone region leaves 8 legible certified pairs at amplitude 20. Padding an image onto the required canvas preserves every original pixel and yields 13 certified pairs, of which 5 are legible (Figure 2).

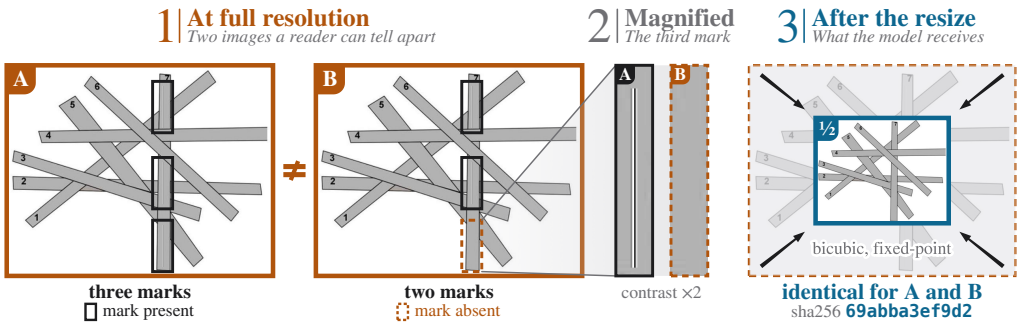

Figure 2: **A certified pair on a MathVision image.** (1) Member A contains three marks and member B contains two, giving different correct counts. (2) The third mark is magnified at twice its original contrast. (3) After $2\times$ resizing, the tensors are bit-identical.

Libraries and numeric modes. We resized five bicubic-carrier and five box-carrier pairs at $2\times$ under the resize modes of Pillow and torchvision and compared the outputs exactly. The bicubic carrier $(1, -3, 3, -1)$ collides bit for bit under antialiased bicubic and bilinear resampling in both libraries, in 8-bit fixed-point and 32-bit floating-point arithmetic alike, because the $2\times$ weights are exactly representable. The box carrier $(1, -1)$ collides under box, area and non-antialiased bilinear resampling. Neither collides under any other tested mode, including Lanczos. A pair is therefore specific to one operator, whatever the library or number format, while Theorem 6 gives every such downscaler kernel vectors of its own. These are the two image backends the audited processors offer. No audited pipeline uses OpenCV or TensorFlow, which we did not evaluate.

Breadth across deployed processors. We applied the same construction to every pipeline whose processor was available (Table 7). Across four additional pipelines using a single fixed-point resize at an eligible integer ratio, all six attempts per pipeline certified, for $24/24$ overall. The current procedure does not resolve tiled policies, patch-based geometries or floating-point resizing. One tiled pipeline first resizes at $4/3\times$ with a Lanczos filter. Under the fixed-point resampler, $\beta(K) = 1.2 \times 10^{15}$ there, so that stage admits no realizable exact-kernel perturbation on interior support at any rank (Table 3).

Table 7: **Breadth across deployed processors.** All six attempts certified on each pipeline with an eligible fixed-point resize. The remaining pipelines fall outside the current procedure's scope.

	Observed resize			
Pipeline	**Input side**	**Output side**	**Ratio**	**Result**
Eligible fixed-point resizes				
SigLIP-B/16-384	768	384	$2\times$	$6/6$ certified
SigLIP-SO400M/14-384	768	384	$2\times$	$6/6$ certified
BLIP-2 OPT-2.7B	448	224	$2\times$	$6/6$ certified
Pixtral-12B	2048	1024	$2\times$	$6/6$ certified
Outside the current procedure's scope				
SmolVLM-Instruct	2048	1536	$1.33\times$	tiled policy
Idefics2-8B	2048	1960	$1.04\times$	tiled policy
Qwen2-VL-7B-Instruct	512	504	$1.02\times$	near-identity policy
Phi-3.5-Vision-Instruct	1344	672	$2\times$	floating-point resize
Fuyu-8B		no resize		patch geometry

E INSTRUMENT VALIDATION

Table 1 summarizes the end-to-end validation. On Qwen2.5-VL-7B, all 15 constructed pairs certified under a $2\times$ downscale, while none of the 15 matched non-kernel controls, the 15 identity-resize controls or the 15 non-kernel controls under identity resize did. The identity-resize controls keep the stock pixel budget, so nothing is downscaled. For the 8 pairs with available logits, member differences at the decision position were exactly zero and all matched controls were nonzero. Without adapting the construction, all 10 LLaVA-1.5-7B pairs certified with zero logit difference and all 10 controls separated. VisualPRM-8B processes a tiled view and a thumbnail, so the construction uses the intersection of the two resize kernels. All 20 pairs collided across the complete interface and all 20 controls separated. We omit its logit gap because the checkpoint returned non-finite language-model logits under every tested hardware and precision configuration. Across 72 rounded real-valued null-space candidates spanning downscale factors from $1.41\times$ to $8\times$, roughly 1% of the 602,112 resized values remained unequal, about 6000 mismatches per tensor, and no candidate certified. In total, the 147 negative controls comprise 45 non-kernel patterns across the three targets, 30 identity-resize attempts and these 72 rounded candidates.

On all 80 certified pairs of the routing screen, a fixed deterministic critique produced pair-balanced accuracy of exactly $1/2$. Re-attention and further reasoning use independently sampled critiques, so their finite-sample outcomes can vary around $1/2$ even though Theorem 2 guarantees identical output laws. Re-encoding lies outside the language-side class, and its gain was positive on 13 pairs, zero on 60 and negative on 7. Near-chance accuracy alone is not evidence of a collision. The identity-resize controls have pair-balanced accuracies from 0.4975 to 0.5028 but nonzero member logit differences.

Table 8: **Action set and diagnostic controls.** The policy actions use the path-matched costs shown below. The placebo critique is the path-matched baseline, the ten-crop oracle bounds crop-based re-encoding, and the same-interface re-encode isolates re-encoding overhead.

Action	Mechanism	Changes $\mathbf{z}$	Cost accounting
Policy actions			
stop	accept the current verdict	no	0
att re-attend	instruct the model to re-examine the region named by Q	no	prompt + one critique
rea reason	run and aggregate t additional critique passes	no	$t\times$ critique
enc re-encode	recompute $\mathbf{z}$ from a higher-resolution center crop	yes	re-prefill + one critique
Diagnostic controls			
Placebo critique	the att path without the instruction to re-examine	no	prompt + one critique
Ten-crop oracle	re-encode the best of ten half-frame crops	yes	matched to enc
Same-interface re-encode	re-encode while reproducing the original interface state	no	matched to enc

F ACTIONS AND COST ACCOUNTING

Action costs. Table 8 specifies the four actions. Normalized to one critique and net of stopping, the decode-token and prefill-token cost vectors are $(0,\ 1.00,\ 2.75,\ 1.01)$ and $(0,\ 1.00,\ 2.93,\ 1.02)$ for $(\mathsf{stop}, \mathsf{att}, \mathsf{rea}, \mathsf{enc})$, and wall-clock time gives $(0,\ 1.00,\ 3.51,\ 1.00)$. The λ sweep of Section 6 uses decode-token cost. The enc action uses a question-independent crop from the central half of the image. A same-interface re-encode produces an identical digest, hence an identical output law and exactly zero gain on all 1280 items. Question-conditioned acquisition policies (Wu & Xie, 2024; Ding et al., 2026) could perform better, so the measured $\Delta(\mathsf{enc})$ does not upper-bound re-encoding in general.

Labels. Natural-item labels are the step-correctness annotations released with Wang et al. (2025). Certified-pair labels need no adjudication because the construction writes the two fact values into known image regions.

G EXPERIMENTAL PROTOCOL

Routing rules and controls. At $\lambda = 0$, the specified IAS rule selects reasoning on 1266 of 1280 natural items and re-encoding on the remaining 14. At $\lambda = 0.01$, it stops on 1184 natural items and every certified pair. Across the full λ sweep, every protocol-specified policy remains within 0.002 regret of stopping, and all adaptive rules reduce to stopping from $\lambda = 0.05$ onward under each cost measure. The specified rule requires no labels or measured gains from test items. The median-threshold diagnostic acts on the items whose score is at or below its median and stops on the rest. It takes the non-stop action with the largest gain on the choosing replicate, and the random score controls for this outcome-informed selection. Raw interface distance is the root-mean-square magnitude of the pixel tensor. Besides the routers of Table 11, two interface-only routers use a 1024-dimensional size-invariant projection of the pixel tensor, and neither improves over stopping.

Statistical plan. All bootstrap intervals are clustered over the 1018 source records containing the 1280 items. Each item is measured in two replicates. Across replicates, verdict probabilities correlate between 0.84 and 0.94, path-matched gains between 0.47 and 0.72, and the per-item optimal action agrees on 65.9% of items. Reversing the choosing and scoring replicates changes every regret in Table 2 by at most 0.004.

Decoder. The decoder d_θ is an ℓ_2-regularized logistic regression on a 128-dimensional Gaussian random projection of the flattened interface tensor. No hyperparameters are tuned. We use five-fold cross-fitting, with folds grouped by source image for rendered items and by record for the natural stream to prevent leakage between related items. Confidence is calibrated out of fold using monotone isotonic regression. The question and candidate step define the fact being decoded but are not decoder features.

H POWER CALCULATION FOR THE ROUTING GATE

We estimate power by simulation because V maximizes over actions and items are clustered within templates, the planned groups of eight related items. Each simulated template receives a best action. An item inherits it with probability ρ and receives an independently drawn action otherwise (Table 9). No false positives occurred in the null simulations over this grid.

Table 9: **Power of the routing-value gate.** Each cell gives the simulated probability that the 95% lower bound on the split-sample routing value V exceeds zero. The columns give the number of eight-item templates, and ρ controls dependence among items from the same template. Bold values mark the smallest template count achieving at least 0.90 power for the pre-registered target $V = 0.03$.

Simulation setting		Number of eight-item templates					
V	ρ	10	20	40	80	160	320
0.01	0.5	0.03	0.07	0.12	0.27	0.45	0.71
0.02	0.5	0.08	0.16	0.35	0.68	0.94	0.99
0.03	0.5	0.13	0.33	0.68	**0.93**	1.00	1.00
0.03	0.9	0.03	0.15	0.52	0.86	**0.99**	1.00
0.05	0.5	0.34	0.77	0.96	1.00	1.00	1.00

We power the gate for $V = 0.03$, the smallest value that remains positive under the cost accounting. At this effect size, 160 eight-item templates give 0.99 power at $\rho = 0.9$, so the design uses 1280 items. The realized stream has 1018 records and an estimated within-record correlation of 0.25, which gives power 1 at $V = 0.03$ and 0.92–0.98 at $V = 0.02$.

I ESTIMATING THE ROUTING VALUE

Bias of the plug-in estimate. A plug-in estimator replaces true gains with measured gains. Taking the maximum within each item then favors actions with positive errors, and more items do not remove this bias. In a simulation with no item-specific signal, the plug-in estimate was 0.125 with 80 items and 0.128 with 1200, and its bootstrap interval excluded zero in every simulation run. The simulated noise scale was not fitted to the observed data, so these values should not be compared with the 0.101 in Table 10. We therefore select per-item maxima on one replicate and evaluate them on the other (Table 10). Sampling is needed because rea aggregates multiple critiques, and verdicts read directly from decision-position logits would make the replicates identical.

Table 10: **From estimated routing headroom to realized gain** on the natural stream. Every entry is an estimate.

Symbol	Quantity	Value	Interpretation
$\hat{V}_{\text{plug}}$	plug-in per-item maximum	0.101	inflated by maximizing noisy gains
$\hat{V}$	split-sample routing value	0.056	estimated headroom after bias control
$\hat{V}_{\text{fit}}$	best of eight fitted interface routers	≈ 0	no fitted router recovers the headroom
$\hat{V}_{\text{IAS}}$	IAS-ROUTE as realized	0	matches the best fixed action

The selected action and its regret are unchanged when a per-item constant is added to every estimated gain, so we report action-ordering accuracy and regret.

J EXTENDED RESULTS

Unprompted baseline and the crop arm. Relative to the model's unprompted verdict, re-attention, reasoning and re-encoding have gains of -0.069, -0.070 and -0.092. Their difference from the path-matched results in Section 7 is the prompt path itself, which has a gain of -0.070. This per-item offset is identical across replicates, so it leaves action ordering and regret unchanged. A ten-crop oracle over the half-frame center crop and a 3×3 grid of half-frame crops upper-bounds any policy restricted to those crops (Zhang et al., 2025). When the crop is selected on one replicate

and scored on the other, the oracle gains $+0.035$ over the path-matched stopping baseline, with a 95% interval of $[+0.022, +0.047]$, but has a gain of -0.035 relative to the unprompted verdict.

Table 11: **Oracle-supervised routers by signal.** The vision-tower routers read $\mathbf{z}$ alone, and the hidden-state routers read $(\mathbf{z}, Q, R)$. Gain intervals are clustered by record, and *Action rate* is the fraction of natural items sent to a non-stop action. On certified pairs, *Any action* counts non-stop choices, while *Dominated* counts choices of re-attention or further reasoning.

		Natural stream			Certified pairs	
Features	**Learner**	**Gain over stop, 95% CI**	**Order acc.**	**Action rate**	**Any action**	**Dominated**
Vision-tower outputs, a signal of $\mathbf{z}$						
projected tokens	ridge	-0.003 $[-0.011, +0.006]$	0.244	0.77	0/80	0/80
projected tokens	boosting	-0.003 $[-0.012, +0.007]$	0.254	0.78	71/80	37/80
patch features	ridge	$+0.002$ $[-0.006, +0.010]$	0.245	0.77	80/80	80/80
patch features	boosting	$+0.003$ $[-0.006, +0.012]$	0.259	0.78	70/80	5/80
both	ridge	$+0.000$ $[-0.008, +0.009]$	0.247	0.76	6/80	6/80
both	boosting	-0.002 $[-0.011, +0.007]$	0.248	0.80	62/80	62/80
Verifier hidden state at the decision position, a signal of $(\mathbf{z}, Q, R)$						
final layer	ridge	$+0.008$ $[+0.000, +0.016]$	0.270	0.79	0/80	0/80
final layer	boosting	$+0.002$ $[-0.007, +0.011]$	0.258	0.80	0/80	0/80
middle layer	ridge	$+0.008$ $[-0.001, +0.016]$	0.282	0.80	0/80	0/80
middle layer	boosting	$+0.005$ $[-0.004, +0.013]$	0.268	0.80	19/80	19/80
both	ridge	$+0.010$ $[+0.002, +0.018]$	0.283	0.79	0/80	0/80
both	boosting	$+0.004$ $[-0.005, +0.012]$	0.260	0.81	0/80	0/80

Oracle-supervised vision-tower routers. The first block of Table 11 reports six routers trained directly on measured action gains with ridge regression or gradient-boosted decision trees. Projected visual tokens delivered to the language model are summarized by their mean and maximum (7,168 dimensions), and pre-projection patch features by their mean and standard deviation (2,560 dimensions). The routers are cross-fitted by record and scored on the held-out replicate. For the certified-pair screen, each is refit on the full natural stream, which also supplies the reported natural action rate. Table 2 reports the lowest-regret router of each block, boosting on patch features and ridge on both hidden layers.

Oracle-supervised hidden-state routers. The second block repeats the experiment with the richest signal the frozen verifier computes from $(\mathbf{z}, Q, R)$, its language-model hidden state at the decision position, taken at the final layer, a middle layer or both (3,584 dimensions each). These states are bit-identical across all 80 certified pairs. The best interval, $[+0.002, +0.018]$, excludes zero with an unadjusted bootstrap $p = 0.019$, which Holm correction across the six routers raises to 0.11. Reversing the choosing and scoring replicates gives the same gain and interval. After correction the gain is not significant, so we do not claim that the text-conditioned signal improves routing.

Table 12: **Results by source benchmark.** Policies are fit once on the full stream and evaluated within each source. Gains are relative to the source-specific best fixed action. $\hat{V}$ is the split-sample routing value, and intervals are 95% confidence intervals clustered by record. The learned router is boosting on patch features. Median-rule entries report IAS and random-score gains, respectively.

Source	**Best fixed**	$\hat{V}$	**IAS-ROUTE specified**	**Median IAS-ROUTE / random**	**Learned router**
MathVerse (454)	att	$+0.067$ $[+0.051, +0.078]$	$+0.002$ $[-0.012, +0.015]$	$+0.020$ / $+0.023$	-0.004 $[-0.018, +0.011]$
MathVision (331)	stop	$+0.031$ $[+0.013, +0.042]$	-0.012 $[-0.028, +0.003]$	$+0.000$ / $+0.002$	-0.007 $[-0.021, +0.006]$
DynaMath (260)	stop	$+0.065$ $[+0.044, +0.079]$	-0.004 $[-0.022, +0.017]$	$+0.019$ / $+0.018$	$+0.009$ $[-0.013, +0.031]$
WeMath (126)	rea	$+0.044$ $[+0.018, +0.055]$	$+0.001$ $[-0.001, +0.005]$	$+0.019$ / $+0.025$	$+0.006$ $[-0.012, +0.024]$
MMMU (109)	rea	$+0.034$ $[+0.011, +0.055]$	-0.001 $[-0.003, +0.001]$	-0.006 / -0.009	-0.004 $[-0.022, +0.012]$
All five (1280)	rea	$+0.056$ $[+0.047, +0.062]$	$+0.000$ $[+0.000, +0.001]$	$+0.016$ / $+0.018$	$+0.002$ $[-0.005, +0.010]$

Source benchmarks. In Table 12, the routing-value interval excludes zero in all five source benchmarks, while the intervals for the specified IAS rule and the learned router contain zero throughout. Under the median rule, the gains from IAS and a random score differ by at most 0.006.

Natural verifier errors under each action. Table 13 divides the 2560 item-by-replicate measurements into 954 initially incorrect and 1606 initially correct stopping verdicts. The placebo critique, the path-matched baseline of Section 6, follows the same prompt path without instructing the model to re-examine the image. *Inert* means that all three sampled critiques produce the same verdict.

Correct after is the fraction of verdicts that are correct after an action, and *Gain* is the mean path-matched change relative to the placebo. About half of the initial errors are inert, which does not establish that interface states collide (Proposition 3). Where verdicts change, each action repairs more errors than the placebo but also breaks more initially correct verdicts. These effects cancel over the stream for re-attention and reasoning and are net negative for re-encoding.

Table 13: **Natural verdicts under each action.** Columns split the 2560 measurements by the initial stopping verdict. Proportions use Wilson intervals, and gain intervals are bootstrapped by record.

		Initial **stop** verdict	
Quantity	**All**	**Wrong**	**Correct**
Measurements, n	2560	954	1606
Inert across three critiques			
Rate	0.452 [0.433, 0.472]	0.496 [0.464, 0.527]	0.426 [0.403, 0.451]
Correct after action			
Placebo	0.525 [0.506, 0.544]	0.126 [0.106, 0.148]	0.762 [0.741, 0.782]
att	0.528 [0.509, 0.547]	0.187 [0.163, 0.213]	0.731 [0.709, 0.752]
rea	0.523 [0.504, 0.542]	0.152 [0.131, 0.176]	0.744 [0.722, 0.764]
enc	0.500 [0.480, 0.519]	0.210 [0.185, 0.237]	0.672 [0.649, 0.694]
Mean gain relative to placebo			
att	+0.001 [−0.007, +0.009]	+0.030 [+0.018, +0.041]	−0.016 [−0.026, −0.006]
rea	+0.003 [−0.004, +0.010]	+0.014 [+0.004, +0.024]	−0.003 [−0.013, +0.006]
enc	−0.021 [−0.034, −0.009]	+0.042 [+0.026, +0.059]	−0.059 [−0.075, −0.043]