VISUAL EVIDENCE UNDER CROSS-EXAMINATION: EVALUATING AND CONTROLLING DECISION-LEVEL EVIDENCE USE IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models increasingly reason through crops, regions, and tool-produced observations. Yet an observation can influence the answer without benefiting the candidate it supports. We study *candidate-bound visual contribution*: valid evidence should help, invalidating its supporting relation should remove its additional effect, and valid rebinding should redirect that effect to the newly supported candidate. We introduce **CROSS-Bench**, a benchmark of 28,000 decision problems with matched invalidation and rebinding tests. Our **RIVET** interface preserves evidence identity and uncertainty, composes a candidate-conditioned response, and separately controls its strength. Shared-evidence experiments show that task accuracy and evidence ownership can diverge. Under matched capacity and training, RIVET increases normalized effect transfer from 0.512 to 0.651 where clean evidence has a positive effect. The advantage persists on common evaluation examples and across repeated decision-layer fits. With evidence predicted from raw inputs, RIVET improves CROSS-Bench accuracy by an average of 5.70 pp across four frozen backbones, relative to the same models without auxiliary evidence. These results separate the utility of visual evidence from the candidate-specific destination of its effect.

1 INTRODUCTION

Regions, crops, marked images, visual memories, and tool-produced observations expose intermediate visual evidence during vision-language reasoning (Shao et al., 2024; Wu & Xie, 2024; Yang et al., 2023; Zheng et al., 2025; Jiang et al., 2025). These interfaces make visual information inspectable, but do not by themselves ensure that its effect reaches the supported candidate. A model may localize the right object and read the right state yet assign the resulting update to an unsupported candidate. Evaluating visual evidence therefore requires examining both its content and the candidate-level destination of its effect.

Consider the scarf example in Fig. 1. For a request about one person's scarf, another person's red scarf cannot support the answer "red." Rebinding the request to that other person makes the observation relevant, so its effect should reach the red answer candidate.

Existing work improves visual access, grounds intermediate claims, and diagnoses whether perceived information enters the decision (Liu et al., 2025b; Chen et al., 2026; Wang et al., 2025c). Counterfactual objectives encourage visual dependence and evidence–answer consistency (Wang et al., 2026b; Xu et al., 2025). A complementary question is *which candidate receives the observation's effect?* Correctness may follow from priors or other evidence. Corruption sensitivity reveals dependence on the perturbed input, but leaves the destination of that effect open. We call this property *decision ownership*: whether an established evidence effect follows the answer candidate supported by the current visual relation.

We formalize the broader criterion as *candidate-bound visual contribution*. **Utility** measures the benefit of valid evidence; **specificity** tests whether invalidating its supporting relation removes the additional effect; and **ownership** asks whether an established effect follows a valid change of owner. Suppressing every observation can pass invalidation without using evidence. Conversely, rewarding a fixed candidate can improve accuracy without following the visual relation. Valid rebinding supplies

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.

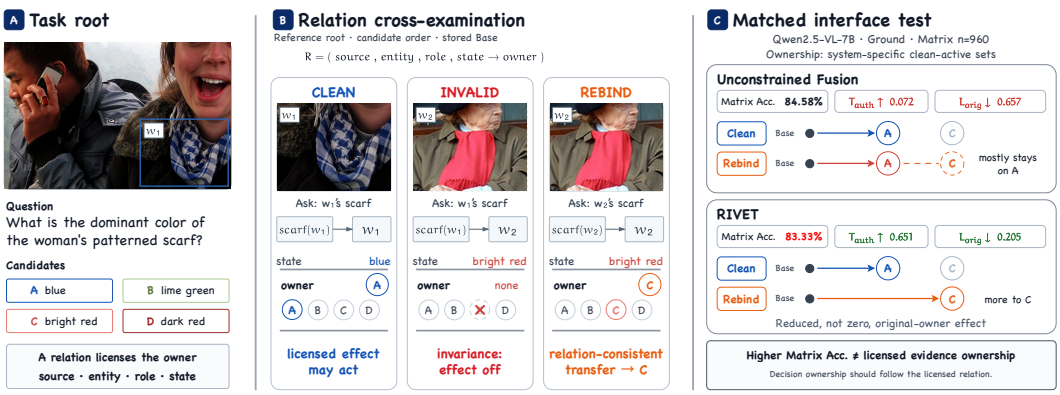


Figure 1: **Candidate-bound visual contribution.** (A) The observation-to-candidate relation; (B) clean, invalid, and rebind request–evidence pairs; (C) utility–ownership dissociation under matched evidence on the 960-root matrix (Qwen2.5-VL-7B, predicted evidence). Ownership uses method-specific clean-active sets. Arrows denote schematic effect destinations, not magnitudes.

the positive requirement missing from either behavior: the established effect should redirect to the newly supported candidate.

CROSS-Bench operationalizes this criterion through matched clean, invalid, and rebinding conditions. The criterion also motivates separating *which candidates the evidence supports* from *how strongly it should act*. An observation's relevance depends on its source and semantic role, which a confidence score alone does not specify. These relations must remain available when composing multiple observations for each candidate. **RIVET** preserves evidence identity, bindings, candidate alignment, and uncertainty across specialized visual readers, composes evidence around candidate queries, and then controls response strength with a scalar.

Our contributions are threefold:

- We introduce *candidate-bound visual contribution* and **CROSS-Bench**, a 28,000-root benchmark that evaluates whether visual evidence helps when valid, loses its additional effect when invalidated, and transfers that effect under valid rebinding.
- We develop **RIVET**, a shared evidence-to-decision interface that preserves identity, candidate alignment, and uncertainty across specialized visual readers, composes evidence through candidate-conditioned representations, and separately controls response strength with a scalar posterior controller.
- We show that task accuracy and evidence ownership can diverge under matched evidence. Relative to a capacity-matched fusion control, RIVET improves effect transfer. With evidence predicted from raw inputs, it also improves task accuracy over Base across four frozen backbones.

2 Related Work

Visual evidence acquisition and grounding. Visual reasoning interfaces expose local content through crops, marks, and region references (Shao et al., 2024; Yang et al., 2023). Tool-using models acquire additional views during reasoning (Zheng et al., 2025; Jiang et al., 2025), while localization supervision and grounded chains of thought connect intermediate claims to visible objects and relations (Wu et al., 2026; Wang et al., 2026a; Sun et al., 2026). These approaches improve visual access and the traceability of reasoning.

Evidence utilization and counterfactual dependence. Multimodal benchmarks assess visual necessity and fine-grained perception (Chen et al., 2024a; Zhang et al., 2025). Diagnostic studies examine perception–reasoning gaps (Liu et al., 2025b; Chen et al., 2026; Wang et al., 2025c; Lv et al., 2026). Counterfactual objectives probe visual dependence (Wang et al., 2026b), including evidence–answer consistency under image masking (Xu et al., 2025). Faithfulness studies test explanations against decisions (Turpin et al., 2023; Chuang et al., 2026), and Self-Critical Reasoning regularizes correct and competing answers' visual gradient sensitivities (Wu & Mooney, 2019). Visual

evidence selection evaluates output-distribution utility (Luo et al., 2026), while receiver-dependent integration models candidate-specific evidence relative to a prior (Kawada & Kellis, 2026). Our paired invalidation and valid rebinding tests examine whether a Base-relative candidate effect follows its annotated visual support relation.

Visual binding and compositionality. Attribute and relation benchmarks test whether models distinguish compositions built from the same objects and concepts (Yuksekgonul et al., 2023; Thrush et al., 2022), with careful negative construction needed to avoid linguistic shortcuts (Hsieh et al., 2023). Entity-grounded composition and reusable visual programs structure visual reasoning (Li et al., 2024; Chen et al., 2024b), while mechanistic analyses examine internal feature binding (Assouel et al., 2026). We measure the decision effect associated with a visual relation, complementing tests of compositional correctness and internal binding.

3 Candidate-Bound Visual Contribution

We compare visual evidence effects over a finite, ordered candidate set $\mathcal{C}_i$. Each decision problem, or root, $r_i = (I_i, q_i, \mathcal{C}_i, y_i)$ contains an image, request, candidates, and clean answer. The Base posterior $\mathbf{p}_i^0$ uses the same image, request, candidates, and scoring procedure, without auxiliary observations. For each evidence condition ι, the evaluator records the support relation

$$\mathcal{R}_i^\iota : (\mathrm{src}, \mathrm{ent}, \mathrm{role}, \mathrm{state})_i^\iota \mapsto o_i^\iota, \tag{1}$$

where $o_i^\iota \in \mathcal{C}_i \cup \{\varnothing\}$ is the candidate supported by this relation, with $\varnothing$ denoting no valid owner. Each trial supplies an observable evidence–request pair u_i^ι. The request specifies the relevant source, entity, and roles without revealing the answer. Ground-truth states, owners, and intervention labels remain evaluator-side.

The same candidate interface answers each condition's supplied request. Clean trials address q_i; valid rebinding addresses q_i' with matching evidence. The primary rebinding test retains the original root, candidate order, and stored $\mathbf{p}_i^0$ as references. It measures the new owner's effect relative to the original Base when the request and supporting relation change together. A complementary test fixes q_i', recomputes Base and evidence-conditioned responses, and varies evidence relevance. This measures within-request selectivity, complementing the cross-request transfer test. Both tests apply to any system exposing matched candidate posteriors, without requiring RIVET's typed representation.

4 CROSS-Bench: Evaluating Candidate-Bound Evidence Use

CROSS-Bench operationalizes this criterion through paired evidence interventions. It contains 28,000 decision roots constructed from public annotations, including Visual Genome, GQA, and TextOCR (Krishna et al., 2017; Hudson & Manning, 2019; Singh et al., 2021), and deterministic renderings, with a frozen test set of 5,600 roots. Each root pairs an image and question with $N_i = |\mathcal{C}_i| \in \{2, 3, 4, 5\}$ candidates and a source–entity–role–state relation. Accuracy uses the clean question; primary relation probes vary the evidence relation against the same Base reference (Fig. 1).

A prediction-blind subset of 960 test roots supports complete relation interventions, with 707 admitting valid rebinding. This subset diagnoses evidence use; all 5,600 roots measure full-suite accuracy. Train/dev/test partitions are source-cluster-disjoint. Across 1,200 reviewed roots, both reviewers confirm unique valid targets on 98.8% and answer neutrality on 99.3%; they also assess alternative-owner validity (App. A).

Two evidence-access routes. Controlled supplies the required-fact specification, source identity, and ordered roles; **Ground** predicts them from raw inputs. Both infer typed state without owner or intervention labels. Crops in Fig. 2(b) illustrate the carriers; they are not extra Ground inputs. Ground predicts evidence access before intervention replay; the replay then changes only the designated evidence or relation input at the specified processing stage (App. A.1).

For each matrix root, **clean** preserves the relation and **canonical null** removes auxiliary contribution. Wrong-source, wrong-entity, wrong-role, and wrong-state invalidate individual coordinates; unauthorized **candidate rotation** breaks evidence–candidate association without reordering candidates. These invalid conditions should remove the additional effect.

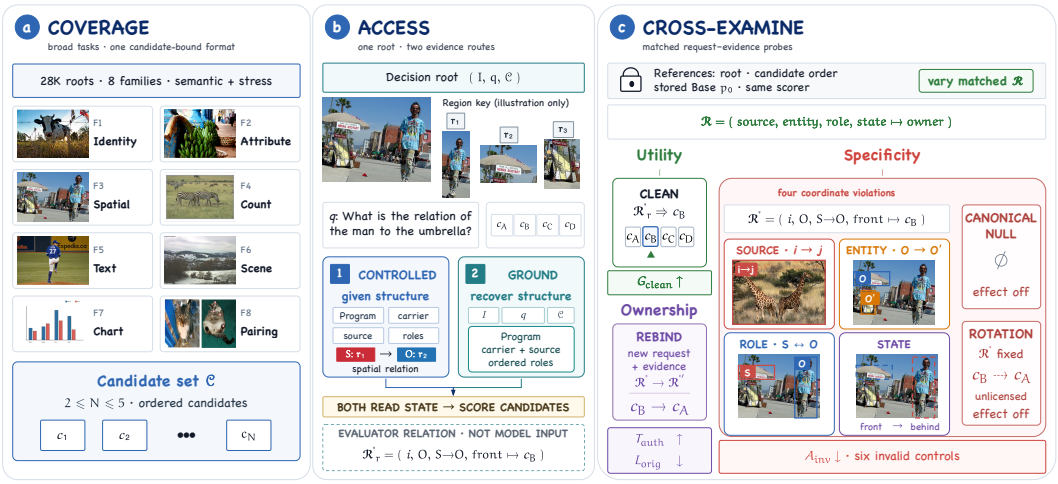


Figure 2: **CROSS-Bench evaluation protocol.** Controlled supplies evidence-access structure and Ground predicts it; both estimate visual state. Clean, invalid, and rebind conditions test utility, specificity, and ownership over 2–5 candidates.

Valid rebinding changes the request and supporting relation together, so the established effect should follow the newly supported candidate. In the scarf example, selecting the other person's scarf and asking about that person makes the new color relevant. Requests identify the person and attribute without stating the color. A companion control fixes the new request and varies its evidence (App. C.2).

We summarize utility by clean gain G_{clean}, specificity by invalid activity A_{inv}, and ownership by authorized transfer T_{auth} and original-owner leakage L_{orig}. Let $\mathbf{p}_i^{\iota}$ denote the posterior under condition ι, let $\bar{p}_{i,c}^{\iota} = \max(p_{i,c}^{\iota}, \epsilon)$ for a numerical floor $\epsilon > 0$, and let $\mathcal{I}_{\text{inv}}$ contain canonical null, single-coordinate relation violations, and unauthorized rotation:

$$G_{\text{clean}} = \mathbb{E}_i \left[\log \frac{\bar{p}_{i,y_i}^{\text{clean}}}{\bar{p}_{i,y_i}^{0}} \right], \qquad A_{\text{inv}} = \mathbb{E}_i \left[\max_{\iota \in \mathcal{I}_{\text{inv}}} \text{TV}\left(\mathbf{p}_i^{\iota}, \mathbf{p}_i^{0}\right) \right]. \tag{2}$$

G_{clean} measures the correct candidate's log-score change, including improvements that leave the argmax unchanged. With $\text{TV}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$, A_{inv} captures each root's largest posterior movement across invalid relations, rather than averaging across invalid conditions.

Ownership tests an effect that clean evidence has already established. We therefore evaluate it on scoreable rebinding-eligible roots where clean evidence increases the original owner's probability; this clean-active set is method-specific. Under clean evidence, $o_i = y_i$; valid rebinding introduces another valid relation supporting a new owner $n_i \neq o_i$. With $[x]_+ = \max(x, 0)$, define $\Delta_{i,c}^{\iota} = p_{i,c}^{\iota} - p_{i,c}^{0}$, $\mathcal{D}^+ = \{i : \Delta_{i,o_i}^{\text{clean}} > \delta_\Delta\}$ for clean-effect threshold δ_Δ, and $b_i = [\Delta_{i,o_i}^{\text{clean}}]_+ + \epsilon$:

$$T_{\text{auth}} = \mathbb{E}_{i \in \mathcal{D}^+} \left[\min\left(1, \frac{[\Delta_{i,n_i}^{\text{rebind}}]_+}{b_i}\right) \right], \qquad L_{\text{orig}} = \mathbb{E}_{i \in \mathcal{D}^+} \left[\frac{[\Delta_{i,o_i}^{\text{rebind}}]_+}{b_i} \right]. \tag{3}$$

Relative to the original Base, T_{auth} measures the positive new-owner increment and L_{orig} the positive original-owner residual, both normalized by the clean effect. These quantities are not success rates. Unnormalized candidate changes provide the complementary effect scale. Transfer is capped; leakage remains uncapped to expose amplification. Fixed-request controls instead use each request's own Base to measure evidence selectivity (App. C.4). Estimation and sensitivity analyses appear in Apps. A.2 and C.1.

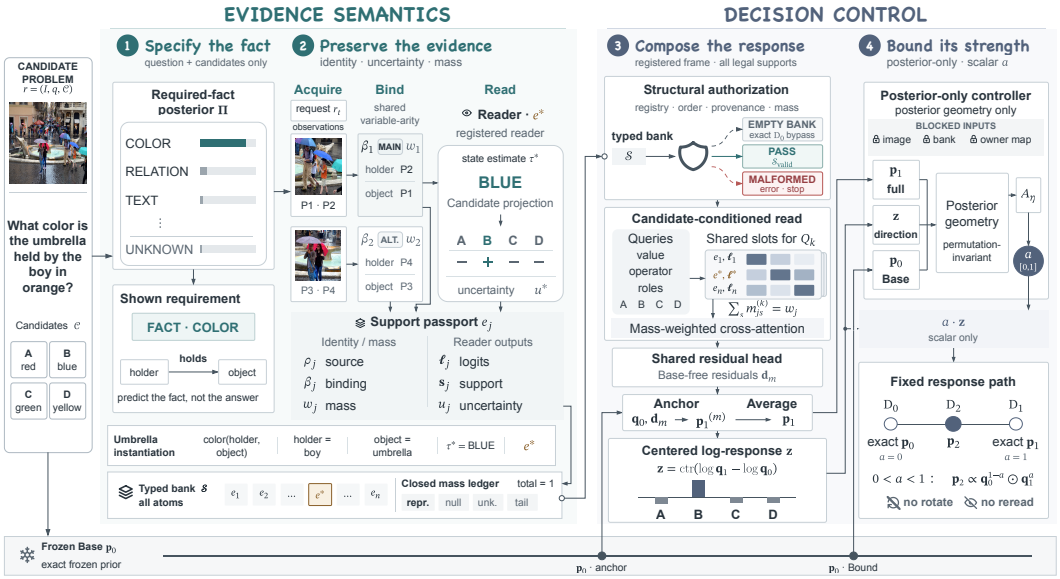


Figure 3: **Overview of RIVET.** SPECIFY plans visual requirements; PRESERVE acquires, binds, and reads observations through a shared typed interface. COMPOSE uses candidate queries to organize and read the complete evidence set; BOUND scales the resulting posterior response. Highlighted e^* follows one illustrative support through the pipeline. Here $\mathbf{q}_t = \max(\mathbf{p}_t, \epsilon)$ for $t \in \{0, 1\}$.

5 RIVET: A CANDIDATE-BOUND EVIDENCE-TO-DECISION INTERFACE

RIVET preserves evidence identity, bindings, and candidate meaning through composition, then separately scales the response (Fig. 3). SPECIFY identifies required facts; PRESERVE binds and reads observations; COMPOSE organizes evidence around each candidate; and BOUND controls response strength.

For ordinary task inference with candidate set $\mathcal{C}$, we omit the root index and use subscripts $0, 1, 2$ for decision endpoints:

$$\begin{aligned} \mathbf{\Pi} &= \mathrm{Specify}(q, \mathcal{C}), & \mathcal{S} &= \mathrm{Preserve}(I, \mathbf{\Pi}, \mathcal{C}), \\ \mathbf{p}_1 &= \mathrm{Compose}_\theta(\mathbf{p}_0, \mathcal{S}, \mathcal{C}), & \mathbf{p}_2 &= \mathrm{Bound}(\mathbf{p}_0, \mathbf{p}_1, \mathbf{z}). \end{aligned} \tag{4}$$

The three endpoints are anchored to the same Base posterior, with $D_0 = \mathbf{p}_0$. $D_1 = \mathbf{p}_1$ incorporates the full evidence response, whose Base-relative log change defines $\mathbf{z}$ in Eq. 6; $D_2 = \mathbf{p}_2$ controls its strength. Primary relation probes supply u_i^ι (Sec. 3) at the stage receiving the changed evidence or request, then rerun subsequent steps against the same Base.

5.1 SPECIFY AND PRESERVE

Required-fact planning. The question and complete candidate set determine which visual distinctions matter: color alternatives require an attribute, whereas spatial alternatives require an ordered relation. SPECIFY predicts these requirements from $(q, \mathcal{C})$ without image or Base access. Its output $\mathbf{\Pi}$ retains uncertainty over registered requirements, specifying what to observe rather than a preferred answer.

PRESERVE executes the active requirements. Acquisition obtains observations; a shared Binder retains weighted assignments to semantic roles. Specialized Readers estimate attributes, counts, spatial relations (direction, relative depth, and proximity), or text/chart/graph states for these bindings. Their distinct state vocabularies map into one candidate-aligned interface for the shared Composer. One observation may support multiple requirements with separate identities and weights: shared pixels need not imply the same semantic role.

Candidate-bound support. Reading "red" is insufficient without knowing whose scarf was observed. Each atom joins provenance ρ_j and ordered binding β_j with candidate-aligned Reader logits $\boldsymbol{\ell}_j$, their

centered form $\mathbf{s}_j$, uncertainty u_j, and represented mass w_j. Here u_j summarizes evidence uncertainty, while w_j is the share of upstream hypothesis probability represented by this atom. With $N = |\mathcal{C}|$ and $\mathrm{ctr}(\mathbf{v}) = \mathbf{v} - N^{-1}(\mathbf{1}^\top \mathbf{v})\mathbf{1}$,

$$e_j = (\rho_j, \beta_j, \boldsymbol{\ell}_j, \mathbf{s}_j, u_j, w_j)\,, \quad \mathbf{s}_j = \mathrm{ctr}(\boldsymbol{\ell}_j), \qquad \mathbf{1}^\top \mathbf{s}_j = 0. \tag{5}$$

Binding identifies whose state is read; a registered deterministic map aligns the Reader state to candidate values in $\boldsymbol{\ell}_j$. Subtracting the candidate mean yields centered relative evidence $\mathbf{s}_j$. Unrepresented mass remains in null, unknown, or tail channels, so surviving observations are not promoted to complete support by renormalization (App. B).

5.2 COMPOSE: ONE FULL-EVIDENCE RESPONSE

COMPOSE turns the aligned observations into a joint candidate response. A deterministic structural check verifies the operator, candidate order, positive mass, and provenance, yielding $\mathcal{S}_{\mathrm{valid}}$. Reading or binding errors with valid schemas can pass these checks. All authorized atoms participate; an empty evidence set leaves Base unchanged. The full response and its centered log-posterior change are

$$\begin{aligned} \mathbf{p}_1 &= \mathrm{Compose}_\theta\left(\mathbf{p}_0, \mathcal{S}_{\mathrm{valid}}, \mathcal{C}\right), \\ \mathbf{q}_t &= \max(\mathbf{p}_t, \epsilon), \qquad t \in \{0, 1\}, \\ \mathbf{z} &= \mathrm{ctr}\left(\log \mathbf{q}_1 - \log \mathbf{q}_0\right). \end{aligned} \tag{6}$$

Candidate query $\mathbf{x}_k$ encodes its value, operator, and roles. With atom embeddings and aligned Reader logits, it determines assignment scores a_{jkv} from atom j to slot v:

$$\omega_{jkv} = w_j \frac{\exp a_{jkv}}{\sum_u \exp a_{jku}}, \qquad \sum_v \omega_{jkv} = w_j.$$

Each candidate can organize the atoms differently, but the assignments preserve every atom's total mass w_j for that candidate. Candidate conditioning thus acts inside aggregation, while unresolved mass stays separate. Mass-weighted cross-attention reads the slots, and a shared head produces Base-free residuals anchored to $\mathbf{p}_0$ afterward.

Separately fitted Composer members define $\mathbf{p}_1$ by posterior averaging, and hence the canonical response $\mathbf{z}$ in Eq. 6. Residual transformation, Base anchoring, and ensemble details are given in App. B.2.

5.3 BOUND: STRENGTH-ONLY CONTROL

Once COMPOSE forms the response, BOUND controls its strength. Geo_N collects permutation-invariant summaries of posterior uncertainty, disagreement, and the centered response. These determine one scalar a, not a candidate-specific correction. When structurally valid support exists,

$$\begin{aligned} a &= \sigma[A_\eta(\mathrm{Geo}_N(\mathbf{p}_0, \mathbf{p}_1, \mathbf{z}))]\,, \\ \mathbf{p}_2 &= \begin{cases} \mathbf{p}_0, & a = 0, \\ \mathrm{Normalize}\left(\mathbf{q}_0^{1-a} \odot \mathbf{q}_1^{a}\right), & 0 < a < 1, \\ \mathbf{p}_1, & a = 1. \end{cases} \end{aligned} \tag{7}$$

The learned branch yields $0 < a < 1$; $a = 0$ and $a = 1$ define exact Base and full-response control endpoints. For any candidates c, k, the interior branch satisfies

$$\log \frac{p_{2,c}}{p_{2,k}} - \log \frac{q_{0,c}}{q_{0,k}} = a \left[\log \frac{q_{1,c}}{q_{1,k}} - \log \frac{q_{0,c}}{q_{0,k}}\right]. \tag{8}$$

Equation (8) scales every evidence-induced pairwise log-odds change by the same nonnegative scalar, preserving its sign. BOUND attenuates the existing response; candidate-wise control would additionally allow its effect to be redirected across candidates.

Training. The SPECIFY planner and Binder use separate program and role supervision. Composer learns task likelihood and calibration on CROSS-Bench train. BOUND minimizes NLL on fixed

posterior pairs produced by Composer members without gradient updates on their roots; fitting and checkpoint selection use disjoint roots. Acquisition and Readers remain frozen; projection and authorization are deterministic. Owner maps and evaluation-condition labels supervise no component (App. B.3).

6 EXPERIMENTS

We ask four questions: whether task accuracy and evidence ownership can diverge; whether candidate-conditioned composition matters beyond aligned evidence; whether these effects persist across backbones and generalize to new tasks and independently authored relations; and which evidence and control mechanisms account for them.

Systems. We diagnose a Qwen2.5-VL-7B Base and four released reasoners: VLM-R^3-7B, Pixel-Reasoner-7B, DeepEyes-7B, and TreeVGR-7B (Bai et al., 2025b; Jiang et al., 2025; Wang et al., 2025a; Zheng et al., 2025; Wang et al., 2026a). Shared-evidence comparisons test prompting, pooled fusion, Unconstrained Residual Fusion (Unconstrained Fusion), calibrated aligned summation, and Capacity-matched Fusion. RIVET uses four frozen backbones: Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL3.5-8B, and LLaVA-OneVision-2-8B (Bai et al., 2025b;a; Wang et al., 2025b; An et al., 2026).

Metrics and protocol. CROSS-Bench accuracy weights roots equally; ΔAcc. is the paired change in percentage points (pp). Accuracy gains use all 5,600 test roots; Table 1 diagnoses the nested 960-root intervention matrix. Ownership retains the original Base; fixed-request selectivity uses the request's own Base. Scoring and uncertainty estimation appear in App. A.2, with paired comparisons in App. C.1. RIVET uses Ground unless noted; CV-Bench uses official source-balanced aggregation.

6.1 MAIN RESULTS: ACCURACY AND OWNERSHIP DIVERGE

Higher accuracy does not imply stronger ownership. Table 1(a) diagnoses released systems; panel (b) compares interfaces using shared evidence. Released reasoners benefit from clean evidence but still respond to invalid evidence and transfer the clean effect incompletely. Unconstrained Fusion has the highest observed accuracy and clean utility in the shared-evidence comparison, yet transfers little of its established effect and retains substantial original-owner leakage. With the same predicted evidence, RIVET raises transfer from 0.072 to 0.651 and lowers mean leakage from 0.657 to 0.205. The ownership ordering persists on common clean-active roots (App. C.1). Thus, task utility and evidence ownership can diverge.

Candidate conditioning matters beyond alignment. Calibrated Aligned Sum achieves transfer of 0.603; RIVET combines higher transfer with higher matrix accuracy (83.33% versus 79.79%).

Table 1: **CROSS-Bench results on the 960-root intervention matrix.** (a) Released-system diagnosis; posterior metrics use each checkpoint's no-auxiliary reference. (b) Shared predicted evidence and Qwen2.5-VL-7B Base. Clean-active counts give scoreable positive-effect roots out of 707 rebinding-eligible roots; ownership uses these method-specific sets ("–": undefined).

Method	Acc. (%)↑	G_{clean} ↑	A_{inv} ↓	Clean-active	T_{auth} ↑ / L_{orig} ↓
(a) Released-system diagnosis					
Qwen2.5-VL-7B (Base)	64.90	0.000	0.000	0/707	–
VLM-R^3-7B	72.50	0.153	0.162	355/707	0.413 / 0.274
Pixel-Reasoner-7B	71.04	0.128	0.143	332/707	0.427 / 0.253
DeepEyes-7B	75.10	0.184	0.171	386/707	0.462 / 0.283
TreeVGR-7B	73.96	0.161	0.187	367/707	0.497 / 0.308
(b) Shared-evidence interface comparison					
Direct Evidence Prompt	70.73	-0.241	0.310	248/707	0.107 / 0.899
Pooled Support Fusion	80.94	0.218	0.477	574/707	0.190 / 0.362
Calibrated Aligned Sum	79.79	0.221	0.000	487/707	0.603 / 0.237
Unconstrained Fusion	**84.58**	**0.296**	0.346	612/707	0.072 / 0.657
Capacity-matched Fusion	84.27	0.289	0.000	547/707	0.512 / 0.286
RIVET	83.33	0.281	0.000	521/707	**0.651 / 0.205**

Capacity-matched Fusion matches evidence, capacity, training, output queries, scalar control, and structural checks, but pools evidence independently of the queried candidate. RIVET raises transfer from 0.512 to 0.651; paired source-cluster bootstrap supports this increase (App. C.1). The advantage persists on common clean-active roots and across three paired fits. On common roots, RIVET also yields a larger unnormalized positive new-owner increment over the original Base (Apps. C.1 and D.4).

Selective responses to semantic errors and evidence mismatch. With shared structural checks, RIVET responds less to schema-valid semantic errors than Capacity-matched Fusion. On 400 clean/error pairs with fixed requests, identifiers, candidate order, and Base, mean invalid TV is 0.022 versus 0.058; both retain positive clean-evidence gains (App. C.2).

Fixing the request and its Base, we compare matched evidence about the requested entity and role with schema-valid crossed evidence about the wrong entity or role. On 638 common roots under the new request, RIVET has similar mean matched log gain to Capacity-matched Fusion (0.229 versus 0.225) but lower crossed gain (0.022 versus 0.063) at the supported candidate. A separate candidate-probability analysis under this Base leaves between-method differences inconclusive (App. C.2). The original request shows the same log-gain pattern (App. C.4).

6.2 Backbone Reuse and Task Generalization

Task gains and ownership patterns persist across Base models. Ground improves all four backbones, averaging +5.70 pp on the 5,600 CROSS-Bench test roots, while Ground ownership remains near 0.65 transfer and 0.21 leakage (Table 2a). Controlled adds 1.84 pp on average with the same Readers, Composer, and Bound, indicating remaining headroom in fact planning, source selection, and role assignment. Composer is trained jointly with Base posteriors from all four backbones; the fitted interface is reused without backbone-specific refitting.

Frozen Ground also improves accuracy over Base in all 16 backbone–benchmark pairs on V⋆Bench, MMStar, CV-Bench, and MME-RealWorld-Lite (MME-RW-L) (Wu & Xie, 2024; Chen et al., 2024a; Tong et al., 2024; Zhang et al., 2025), averaging +2.95 to +7.03 pp by benchmark (Table 2b; App. D.1). Shared evidence also benefits Capacity-matched Fusion; accuracy rankings vary across tasks.

Independent relations show the same directional pattern. On 320 independently authored attribute and spatial roots, RIVET has transfer of 0.563 versus 0.447 for Capacity-matched Fusion. The ordering persists on common clean-active roots, although the paired comparison remains statistically inconclusive (App. D.2). A separate 900-root collection spans four relation types. On its 830 common scoreable roots, RIVET again has similar mean matched gain but lower crossed gain at the request-supported candidate (App. D.3).

Table 2: **Backbone reuse and task generalization.** (a) CROSS-Bench gains on 5,600 roots; Ground ownership on the 960-root matrix with method-specific active sets. (b) Frozen Ground task generalization averaged over four backbones with shared predicted evidence.

(a) CROSS-Bench: backbone and route replication

Backbone	Ctrl. ΔAcc. (pp)	Ground ΔAcc. (pp)	Ground $T_{\text{auth}} \uparrow$ / $L_{\text{orig}} \downarrow$
Qwen2.5-VL-7B	+6.84	+4.96	0.651 / 0.205
Qwen3-VL-8B	+7.93	+5.96	0.638 / 0.216
InternVL3.5-8B	+8.27	+6.46	0.659 / 0.225
LLaVA-OneVision-2-8B	+7.11	+5.39	0.647 / 0.196
Mean over backbones	**+7.54**	**+5.70**	**0.649 / 0.211**

(b) Frozen Ground task generalization across external benchmarks

Endpoint / contrast	V⋆Bench	MMStar	CV-Bench	MME-RW-L
Mean Base Acc. (%)	77.49	61.77	81.62	49.19
Mean Capacity-matched Acc. (%)	82.85	64.92	84.85	55.90
Mean RIVET Acc. (%)	83.38	64.72	85.14	56.23
RIVET − Base (pp)	+5.89	+2.95	+3.52	+7.03

Table 3: **State-only Oracle and architectural controls.** (a) State replacement with predicted bindings and hidden owner labels. (b) Accuracy gains on 5,600 roots; relation metrics on the 960-root matrix, with ownership fixed to Full RIVET's 521 active roots and clean-effect normalizers.

(a) State-only Oracle

Decision interface	Predicted state $T_{\text{auth}}/L_{\text{orig}}$	Oracle state $T_{\text{auth}}/L_{\text{orig}}$
Unconstrained Fusion	0.072 / 0.657	0.101 / 0.612
Capacity-matched Fusion	0.512 / 0.286	0.564 / 0.253
RIVET	0.651 / 0.205	0.685 / 0.171

(b) Architectural controls

Variant	Full ΔAcc. (pp)↑	$G_{\text{clean}}\uparrow$	$A_{\text{inv}}\downarrow$	$T_{\text{auth}}\uparrow$ / $L_{\text{orig}}\downarrow$
Full RIVET	+4.96	0.281	**0.000**	0.651 / 0.205
Hard program selection	+4.18	0.179	0.004	0.083 / 0.056
Shuffled evidence bindings	+4.63	0.264	0.006	0.296 / 0.402
Renormalize represented mass	+4.41	0.221	0.014	0.421 / 0.344
Pre-Compose evidence selection	+4.29	0.234	0.007	0.507 / 0.276
No scalar Bound (D_1)	+4.64	0.265	**0.000**	0.676 / 0.194
Candidate-wise Bound	**+5.21**	**0.298**	**0.000**	0.468 / 0.346

6.3 Analysis of Evidence Alignment and Response Control

Better state estimates do not close the ownership gap. Oracle states raise transfer from 0.512 to 0.564 for Capacity-matched Fusion and from 0.651 to 0.685 for RIVET, lowering leakage in both (Table 3a). Acquisition, bindings, candidate mapping, and predicted-state ownership references are fixed; owner labels stay hidden. State correction thus improves transfer while leaving the ordering between interfaces intact. Better readings alone do not resolve how their decision effects are assigned to candidates.

Identity and complete support matter beyond accuracy. In frozen replays with Full RIVET's ownership reference fixed, binding shuffles retain most accuracy gain but more than halve transfer and increase leakage (Table 3b). Hard program selection weakens clean utility and nearly removes transfer despite low leakage; low leakage alone therefore does not imply successful transfer. Renormalizing represented mass while setting unresolved-mass channels to zero reduces transfer and increases leakage in frozen replay; pre-Compose selection shows the same pattern (App. C.3).

Bound calibrates an existing evidence response. Bound adds 0.32 pp in accuracy and improves clean utility over D_1, whose mean ownership scores are slightly better (Table 3b). A fitted global scalar has intermediate utility, while refitted candidate-wise control increases task gain but worsens ownership (App. C.3). With supports and composition fixed, these comparisons support separating response formation from strength calibration. A shared scalar rescales the composed log-odds response, whereas candidate-wise control can change its direction.

7 Conclusion and Limitations

CROSS-Bench shows that useful visual evidence need not direct its effect to the candidate supported by the underlying relation. RIVET addresses this gap through candidate-conditioned composition and separate strength control, improving transfer over Capacity-matched Fusion while retaining task gains over Base across frozen backbones and external benchmarks. These findings motivate evaluating evidence utility and effect destination together. The criterion focuses on positive contributions in finite candidate sets with explicit, annotatable relations; RIVET relies on registered typed state vocabularies and remains susceptible to schema-valid reading errors. Extending the analysis to contradictory or temporal evidence and open-ended generation, and relating decision-level effects to hidden-state causal mediation, are directions for future work.

AI USE STATEMENT

We used generative AI tools for manuscript development, language editing, literature assistance, and feedback on experimental methodology and interpretation. Based on author-provided outlines and technical content, the tools helped draft and revise passages, improve organization and clarity, and check terminology and consistency between the main paper and appendices. Literature assistance covered search queries, relevant studies, and reference organization. We also used AI to discuss control experiments; clarify evaluation protocols, quantities, sets, and reference distributions; interpret evidence utility, invalid-evidence responses, and ownership; and distinguish point-estimate patterns from statistically supported comparisons and assess the scope of conclusions. The authors reviewed and revised all retained AI-assisted material, checked methodological descriptions and reported values, and verified bibliographic details and citation relevance against primary sources. They made the final decisions on analyses and conclusions and take responsibility for the manuscript, references, experimental results, and scientific claims.

Reproducibility statement. Appendix A documents data sources, splits, interventions, and statistical estimation. Appendix B specifies model dependencies, support representation, aggregation, training, and optimization. Apps. C and D detail evaluation sets, controls, paired fits, and failure accounting.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019. doi: 10.1609/aaai.v33i01.33018076. URL https://ojs.aaai.org/index.php/AAAI/article/view/4815.

Xiang An, Yin Xie, Feilong Tang, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Rim Assouel, Declan Iain Campbell, Yoshua Bengio, and Taylor Whittington Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3RQ863cRbx.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024a. URL https://arxiv.org/abs/2403.20330.

Shuhang Chen, Yunqiu Xu, Junjie Xie, Aojun Lu, Tao Feng, Zeying Huang, Ning Zhang, Yi Sun, Yi Yang, and Hangjie Yuan. CogFlow: Bridging perception and reasoning through knowledge internalization for visual mathematical problem solving. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.01874.

Zhenfang Chen, Rui Sun, Wenjun Liu, Yining Hong, and Chuang Gan. GENOME: Generative neuro-symbolic visual reasoning by growing and reusing modules. In *International Conference on Learning Representations*, pp. 2012–2039, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/088d99765bc121c6df215da7d45bc4e9-Abstract-Conference.html.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Shaochen Zhong, Fan Yang, Andrew Wen, Mengnan Du, Xuanting Cai, Vladimir Braverman, and Xia Hu. FaithLM: Towards faithful explanations for large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pp. 3802–3824, 2026. doi: 10.18653/v1/2026.eacl-long.177. URL https://aclanthology.org/2026.eacl-long.177/.

Julian Eisenschlos, Hernán Maina, Guido Ivetta, and Luciana Benotti. Selectively answering visual questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4219–4229, 2024. doi: 10.18653/v1/2024.findings-acl.250. URL https://aclanthology.org/2024.findings-acl.250/.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. VLM-R^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought, 2025. URL https://arxiv.org/abs/2505.16192.

Sebastien Kawada and Manolis Kellis. Evidence integration in large language models, 2026. URL https://arxiv.org/abs/2609.04290.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics*, 2014. doi: 10.1145/2601097.2601101. URL https://doi.org/10.1145/2601097.2601101.

Junyan Li, Delin Chen, Yining Hong, Zhenfang Chen, Peihao Chen, Yikang Shen, and Chuang Gan. CoVLM: Composing visual entities and relationships in large language models via communicative decoding. In *International Conference on Learning Representations*, pp. 16533–16550, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47561f5e1dc53c7d119185e217b523d0-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pp. 38–55, 2025a. doi: 10.1007/978-3-031-72970-6_3. URL https://doi.org/10.1007/978-3-031-72970-6_3.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs, 2025b. URL https://arxiv.org/abs/2510.17771.

Weiqing Luo, Zongye Hu, Xiao Wang, Zhiyuan Yu, Haofeng Zhang, and Ziyi Huang. Utility-oriented visual evidence selection for multimodal retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35091–35124, 2026. doi: 10.18653/v1/2026.acl-long.1620. URL https://aclanthology.org/2026.acl-long.1620/.

Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen, and Bo Chen. SPD-Faith bench: Diagnosing and improving faithfulness in chain-of-thought for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19875–19927, 2026. doi: 10.18653/v1/2026.findings-acl.995. URL https://aclanthology.org/2026.findings-acl.995/.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence, 2019. URL https://arxiv.org/abs/1908.10543.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72983–73007, 2023. doi: 10.52202/075280-3191. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13018–13028, 2021. doi: 10.1109/CVPR46437.2021.01282. URL https://openaccess.thecvf.com/content/CVPR2021/html/Pham_Learning_To_Predict_Visual_Attributes_in_the_Wild_CVPR_2021_paper.html.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2015. URL https://arxiv.org/abs/1505.04870.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024. URL https://arxiv.org/abs/2403.16999.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8802–8812, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Singh_TextOCR_Towards_Large-Scale_End-to-End_Reasoning_for_Arbitrary-Shaped_Scene_Text_CVPR_2021_paper.html.

Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, and Cees G. M. Snoek. RegionReasoner: Region-grounded multi-round visual reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.03733.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-CLIP: A CLIP model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13019–13029, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Sun_Alpha-CLIP_A_CLIP_Model_Focusing_on_Wherever_You_Want_CVPR_2024_paper.html.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2204.03162.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2771. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ee3a664ccfeabc0da16ac6f1f1cfe59-Abstract-Conference.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2507.07999.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning, 2025a. URL https://arxiv.org/abs/2505.15966.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Yucheng Wang, Yifan Hou, Aydin Javadov, Mubashara Akhtar, and Mrinmaya Sachan. Compose and fuse: Revisiting the foundational bottlenecks in multimodal reasoning, 2025c. URL https://arxiv.org/abs/2509.23744.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=izbBqTL8vb.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 10216–10225, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wu_PhraseCut_Language-Based_Image_Segmentation_in_the_Wild_CVPR_2020_paper.html.

Jialin Wu and Raymond J. Mooney. Self-critical reasoning for robust visual question answering. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/33b879e7ab79f56af1e88359f9314a10-Abstract.html.

Penghao Wu and Saining Xie. V^*: Guided visual search as a core mechanism in multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.01243. URL https://arxiv.org/abs/2312.14135.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33577–33587, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_Grounded_Chain-of-Thought_for_Multimodal_Large_Language_Models_CVPR_2026_paper.html.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. DeFacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning, 2025. URL https://arxiv.org/abs/2509.20912.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark prompting unleashes extraordinary visual grounding in GPT-4V, 2023. URL https://arxiv.org/abs/2310.11441.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *IEEE International Conference on Computer Vision*, 2011. URL https://people.csail.mit.edu/khosla/papers/iccv2011_yao_full.pdf.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, pp. 89655–89701, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning, 2025. URL https://arxiv.org/abs/2505.14362.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. doi: 10.1109/TPAMI.2017.2723009. URL https://doi.org/10.1109/TPAMI.2017.2723009.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287, 2021. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

Appendix at a Glance

A Benchmark Construction and Evaluation p. 15
A.1 Data and interventions; A.2 Statistics and failures.
B RIVET Implementation and Training p. 19
B.1 Readers; B.2 Posterior computation; B.3 Training (p. 21).
C Additional Experimental Results p. 23
C.1 Paired comparisons; C.2 Specificity; C.3 Mechanisms; C.4 Crossed request–evidence analysis.
D Transfer, Robustness, and Evaluation Coverage p. 31
D.1 Task transfer; D.2 Ownership probes; D.3 Independent selectivity; D.4 Training variation; D.5 Coverage.
E Mechanism Illustrations and Limitations p. 35
E.1 Matched relations; E.2 Typed visual tasks; E.3 Visual-state error and scope.

A Benchmark Construction and Evaluation

CROSS-Bench pairs ordinary candidate decisions with interventions on their auxiliary evidence relations. Figure 4 summarizes the construction and evaluation sets; the following sections define matched inputs, readouts, and statistical estimates.

A.1 Data, access, and matched interventions

The 28,000 roots comprise 24,000 semantic and 4,000 stress roots, with 2–5 ordered candidates.

Sources and construction. Object, attribute, and relation roots use PhraseCut, Visual Genome, GQA, and VAW (Wu et al., 2020; Krishna et al., 2017; Hudson & Manning, 2019; Pham et al., 2021). We convert public V*/SEAL training examples (Wu & Xie, 2024) into candidate-based localized visual-QA roots for the training partition; V*Bench remains an external evaluation set. Count/set roots use GQA and COCO (Lin et al., 2014). TextOCR supplies scene text (Singh et al., 2021); Places365, Stanford40 Actions, and Transient Attributes supply global states (Zhou et al., 2018; Yao et al., 2011; Laffont et al., 2014); SPair-71k supplies correspondences (Min et al., 2019). Deterministic renderings provide documents, symbols, and charts/tables; image transforms provide style, visual-change, and jigsaw tasks.

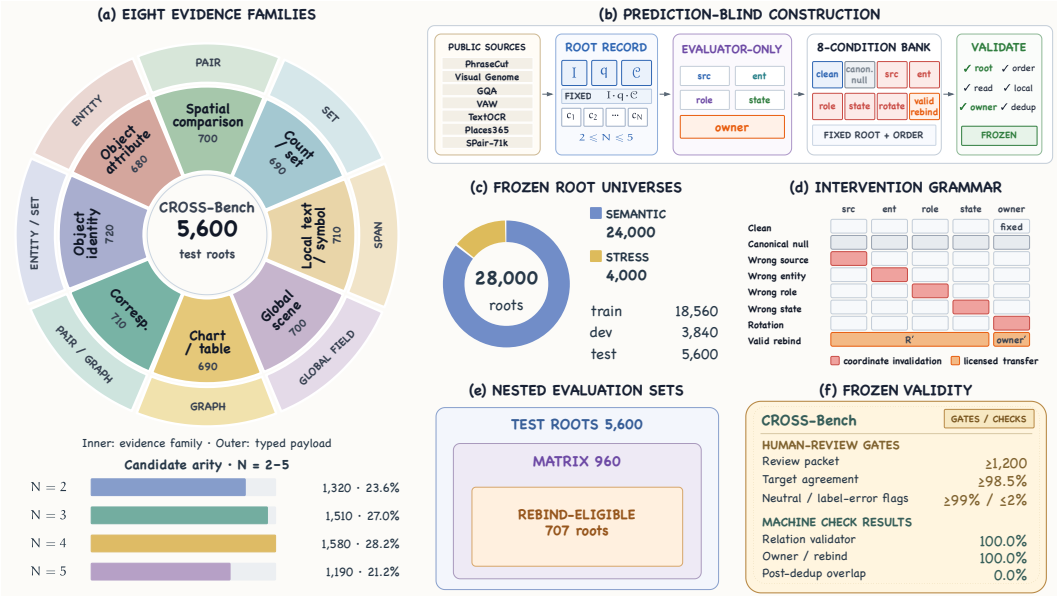


Figure 4: **CROSS-Bench construction and evaluation.** The intervention matrix lies within the full test set; rebinding applies to its eligible subset.

Table 4: **Benchmark splits and validation.** (a) Root counts; indented subsets are already included in Semantic. (b) Machine checks. (c) Property confirmation and candidate-label error flags before adjudication.

(a) Split hierarchy

Root set	Train	Dev	Test
Semantic	16,000	3,200	4,800
Complete-intervention subset of semantic	3,200	640	960
Rebinding-eligible test subset	–	–	707
Stress	2,560	640	800
Full suite (Semantic + Stress)	18,560	3,840	5,600

(b) Release-level checks

Check	Evaluated scope	Result
Paired-input and identity validation	28,000 roots; 56,000 route rows	0 errors
Independent stratified validation	192 roots in 48 cells	0 errors
Source-cluster split overlap	Train, dev, and test	0 shared clusters
Exact / perceptual image overlap	25,350 local images vs. 16,301 public reference images	0 / 0 matches

(c) Human-review summary

Review item	Reviewed roots	Rate (%)
Both reviewers confirm a unique target	1,200	98.8
Both reviewers confirm answer neutrality	1,200	99.3
Either reviewer flags a candidate-label error	1,200	1.4

Semantic roots query visual properties or relations specified by source annotations or rendering states. We pair the supported candidate with alternatives from the task vocabulary, numerical values, text strings, or correspondence choices. Each root retains its source, entity, ordered roles, and typed state for constructing matched conditions. Stress roots cover absent, ambiguous, mismatched, or insufficient evidence, with explicit ambiguity, rejection, or insufficient-evidence answers. App. B.3 describes component-specific training supervision.

Splits and intervention subsets. A source cluster groups a parent image or constructed instance with its related views and derivatives, all within one train/dev/test partition. Table 4(a) gives the split hierarchy. The 960-root test matrix is selected from construction records before model evaluation. Every matrix root has clean, null, coordinate-invalid, and candidate-rotation conditions. Valid rebinding requires an alternative relation supporting another candidate and applies to 707 roots.

Validation. Machine checks verify paired inputs, identities, and split separation. Independent stratified validation samples four roots per combination of domain (semantic/stress), split, and evidence family (Table 4(b)). Two reviewers independently assess the 1,200-root review packet: 100 roots per family and 400 high-risk cases. Review covers target uniqueness, answer neutrality, candidate labels, readability, relation preservation, invalidation locality, and alternative-owner validity. All three reported review rates meet the release criteria: at least 98.5% joint confirmation of target uniqueness, at least 99.0% joint confirmation of neutrality, and at most 2.0% candidate-label errors flagged by either reviewer before adjudication.

Image overlap checks compare exact payload hashes and flag a perceptual match when both pHash and dHash Hamming distances are at most 4. The reference collection covers V*Bench, HRBench4K/8K, MMStar, RealWorldQA, MME-RealWorld-Lite, MMVP, CV-Bench-2D/3D, CountBenchQA, OCRBench, ChartQA, BLINK, and gRefCOCO. Table 4(b) reports zero matches within this collection.

Access routes. Both routes receive the ordinary image, question, and ordered candidates. Controlled also provides the answer-neutral required-fact Program, carrier/source identity, and ordered roles; Ground predicts this evidence-access structure. Both estimate visual state and candidate effects. The two route rows for each root share its split assignment. Gold answers, true states, owner annotations,

Table 5: **Observable evidence interventions and evaluator-only relations.**

Condition	Observable auxiliary input u_i^ι	Evaluator relation
Clean	The request and observations specify a valid source–entity–role–state relation.	Valid; owner $o_i = y_i$
Canonical null	Remove the auxiliary payload; keep the ordinary input.	No auxiliary support
Wrong source	Substitute the auxiliary source carrier or its declared source association.	Source invalid
Wrong entity	Replace the target observation or its declared entity association.	Entity invalid
Role	Change an ordered role relative to the same carrier.	Ordered relation invalid
State	Change the state-bearing observation relative to its relation.	State relation invalid
Candidate rotation	Rotate the observable evidence–candidate association without a matching relation change.	No new valid owner
Valid rebinding	Change the auxiliary request and its carrier/association consistently.	Valid; new owner $n_i \neq o_i$

and intervention labels remain evaluator-side in these ordinary routes. State-only Oracle is a separate diagnostic that replaces estimated visual states while holding access and bindings fixed.

Observable interventions. Primary matched replays fix the ordinary root r_i, candidate identities and order, and Base. Each prediction call receives the auxiliary request, carrier, and association in u_i^ι; the evaluator separately records the supported candidate through $(\mathrm{source}, \mathrm{entity}, \mathrm{role}, \mathrm{state}) \mapsto \mathrm{owner}$. This owner annotation is not a prediction or gating input. Table 5 defines the observable changes. A clean relation licenses support for $o_i = y_i$, while its visual state is still estimated by the model. Invalid carriers may remain readable; specificity measures their incremental effect on the decision.

Valid rebinding changes the auxiliary request and its supporting relation together. The label y_i remains attached to the original question q_i; it is not a correctness target for the rebound request q_i'. Candidate order remains fixed. Consistent reindexing of candidates and state mappings alone is not valid rebinding.

Worked observable input. In the spatial case of Fig. 9, the request is "Where is the dark jug P1 relative to the striped pot P2?" Options are A: right, B: left, C: above, D: below. The image and options remain unchanged. The interface answers the auxiliary request; the original question defines the reference decision.

- **Clean:** use that request with subject P1, reference P2, and their ordered observations. The visual relation supports A.
- **Invalid role:** retain the request but exchange the subject/reference association of the observations. The carrier remains readable, but its roles no longer match the request.
- **Valid rebinding:** ask "Where is the striped pot P2 relative to the dark jug P1?" and supply the consistently reversed roles. The visual relation now supports B. Neither request reveals the directional state or the supported option.

Replay boundaries and Base references. We record each replay's changed input, intervention boundary, and reused upstream stages, then rerun affected downstream stages. For RIVET, role changes reuse physical observations and recompute reading, projection, composition, and scalar control. Observation replacements restart acquisition; typed-state replay fixes bindings and replaces the state before the shared decision suffix. Ground predicts access before replay. Full-chain accuracy evaluates the complete prediction, while replay measures the response to the specified evidence change.

The primary replay retains the original $\mathbf{p}_i^0$ to track how an established clean effect follows a changed relation. The fixed-new-request comparison instead evaluates $\mathbf{p}_i^{0,q'}$ and recomputes null, unmatched, and matched calls with that posterior inside composition and strength control. It measures evidence relevance at a fixed request; the new Base participates in the response computation itself.

A.2 Evaluation sets and statistical estimation

Ordinary accuracy. Each backbone and route is evaluated on the 5,600-root full test suite in Table 4(a). Every root contributes one equally weighted ordinary prediction. With $c_i^M = 1$ for a correct answer and zero otherwise, including absent or unparseable answers,

$$\mathrm{Acc}(M;\mathcal{E}) = \frac{100}{|\mathcal{E}|}\sum_{i\in\mathcal{E}} c_i^M, \qquad \Delta\,\mathrm{Acc} = \frac{100}{|\mathcal{E}|}\sum_{i\in\mathcal{E}}(c_i^M - c_i^{\mathrm{Base}}).$$

Intervention calls are repeated conditions of these roots, not additional accuracy examples. Family and arity breakdowns retain root-level weighting; averages over backbones or suites are labeled.

Answer and posterior readout. Ordinary accuracy uses the final option from native inference for released reasoners and the argmax for vector-valued interfaces. For autoregressive systems, we sum option-identifier token log likelihoods at the terminal answer slot given the pre-answer context, then normalize over candidates at unit temperature. Identifiers, templates, and stopping rules are fixed across conditions; no length normalization or fitted calibration is applied. Learned interfaces return candidate vectors; generated answers are never converted to one-hot posteriors. Each released reasoner's Base is its own checkpoint's no-auxiliary posterior on the ordinary input. Canonical null reuses it under identical rendered inputs and scoring rules. Matched interfaces share the same per-root Base.

Scoreability and analysis sets. A scoreable vector is complete, finite, and in the fixed candidate order. $\mathcal{O}_m^M$ contains roots with all posteriors required by metric m: Base/clean for G, Base/all invalid conditions (including canonical null and candidate rotation) for A, and Base/clean/rebind for ownership. Missing vectors are not imputed; exact-Base returns remain scoreable. Inclusion in $\mathcal{O}_m^M$ depends on posterior availability, without filtering on answer correctness or transfer success. Empty sets yield no estimate.

Ownership uses $\mathcal{O}_{T,L}^M \cap \mathcal{D}_M^+$ within the 707 rebinding-eligible roots. Activity requires the clean owner's probability increment to exceed $\delta_\Delta = 0.010$ (1 pp), before ownership normalization; exact-Base clean responses are not active. App. C.1 examines thresholds 0.005, 0.010, 0.020, 0.030, and 0.050.

Ownership comparison references. Cross-system comparisons use each method's clean-active membership and clean-effect normalizers b_i. Architectural ablations in Table 3(b) fix Full RIVET's membership and b_i, retaining roots whose effects disappear after intervention. State-only Oracle (Table 3a) fixes each interface's predicted-state membership and b_i on roots scoreable in both replays. Other backbones retain their own memberships and normalizers. Primary comparisons use these method-specific sets; common-set sensitivity uses roots scoreable and clean-active for both methods.

Failures and coverage. Table 6 reports ordinary-answer failures on 5,600 roots, complete-call scoreability on the 960-root matrix, and clean-active coverage among 707 rebinding-eligible roots. Absent or unparseable ordinary responses are answer failures; parsed incorrect answers also score zero. Auxiliary-replay failures affect posterior availability. Complete-call coverage requires all applicable conditions; only eligible roots require rebinding calls. Each metric requires only its own posteriors: a root missing an invalid-condition call can still contribute to clean utility.

Qwen2.5-VL-7B Ground RIVET uses $n_G = 948$ and $n_A = 932$ matrix roots. Among rebinding-eligible roots, 682 have scoreable Base/clean/rebind posteriors, and 521 exceed the clean-effect threshold. Under the same rule, Direct Evidence Prompt, Pooled Support Fusion, and Confidence-Gated Fusion have 248, 574, and 558 clean-active roots; uncalibrated and calibrated aligned sums have 474 and 487.

Units and interpretation. Accuracy is reported in percent (%) and changes in percentage points (pp). G_{clean} averages natural-log probability ratios, detecting support changes without an argmax change. The floor $\epsilon = 10^{-8}$ stabilizes probability logs and ownership denominators; δ_Δ separately defines clean activity. TV measures total posterior displacement. Ownership normalizes movement at named candidates by the clean increment, measuring relative magnitude rather than successful-rebinding frequency. TV and T_{auth} lie in $[0,1]$; L_{orig} is nonnegative and uncapped. Transfer and leakage need not sum to one because probability can move to other candidates and total positive movement can change.

Table 6: **Ordinary-answer failures and posterior coverage.** Root counts use the full test suite, intervention matrix, and rebinding-eligible subset, respectively. All-call scoreability requires every applicable matrix call; each metric uses its required posteriors. RIVET uses Qwen2.5-VL-7B Ground.

System	Answer failures / 5,600	All calls scoreable / 960	Clean-active / 707
Qwen2.5-VL-7B (Base)	0	960	0
VLM-R^3-7B	112	938	355
Pixel-Reasoner-7B	168	931	332
DeepEyes-7B	84	944	386
TreeVGR-7B	140	936	367
RIVET	185	928	521

Uncertainty. Percentile 95% confidence intervals (CIs) use 10,000 source-cluster bootstrap resamples shared across systems. Each resample recomputes estimates and paired contrasts under fixed scoreability and activity rules. Common sets retain roots meeting all compared methods' stated requirements. Active counts describe coverage.

Five Composer members yield one posterior per root. CIs quantify test-sampling uncertainty conditional on the fitted systems; App. D.4 evaluates training variation. Marginal CIs describe individual estimands; paired CIs use their within-resample contrasts. Estimates and CI endpoints are rounded only for display.

App. D.5 reports full-suite outcome categories, and App. E.1 illustrates the original and new owner coordinates.

B RIVET Implementation and Training

The implementation and component fitting are shared across four frozen Base backbones.

B.1 Acquisition, binding, and typed reading

Planning and acquisition. The all-MiniLM-L6-v2 planner (Wang et al., 2020) reads the complete question–candidate input. Its 40 executable signatures specify facts and roles; 22 acquisition contracts specify observations. Each contract sums its signature probabilities. At most three contracts retain their original mass; unselected registered mass remains Program-tail. The interface returns Base when unknown probability is at least that of every contract.

The nine Reader specialists cover 41 visual operators. Acquisition combines GroundingDINO Swin-T (Liu et al., 2025a) and OWLv2-base (Minderer et al., 2023) proposals with SAM2.1 Hiera-small (Ravi et al., 2025) masks and native full-image, span, and graph views. All structurally valid atoms from retained acquisition and binding hypotheses enter Compose. Shared physical observations retain separate semantic memberships.

Shared Binder. Binder scores observation-to-role assignments under registered singleton, set, ordered-pair, span, global, and graph cardinalities. Width-128 role and observation encoders feed a shared rank-64 compatibility term and a cardinality-specific structural term. Ordered pairs add a directed rank-32 subject–object term. Structural features have width 32 and dropout 0.1. Scores normalize over legal assignments; retained hypotheses preserve their absolute masses. Blocks of 128 ordered pairs are computation units, not a top-128 selection. Readers estimate each bound structure's state.

Typed Readers. Table 7 maps specialists to state vocabularies and frozen dependencies. Their visual operations differ; binding, candidate alignment, uncertainty, and mass interfaces are shared.

The frozen variants are DINOv2-base (Oquab et al., 2024), SigLIP2-base-patch16-224 (Tschannen et al., 2025), AlphaCLIP GRIT-MIM B/16 (Sun et al., 2024), SAM2.1 Hiera-small, and Qwen2.5-VL-7B-Instruct.

Spatial state readout. An ordered-pair binding supplies subject/object boxes and views. DINOv2 provides the frozen visual features. Horizontal/vertical logits use normalized center offsets;

Table 7: **Typed Reader specialists and their frozen dependencies.**

Specialist	Typed estimate	Frozen model dependencies
Object identity	Entity / instance state	DINOv2, SigLIP2, SAM2.1
Color	Local color state	AlphaCLIP, SigLIP2, SAM2.1
Other object states	Material, pose, appearance	AlphaCLIP, SigLIP2, Qwen2.5-VL
Spatial relation	Direction / depth / proximity / support	DINOv2
Count and set	Cardinality / membership	DINOv2, SigLIP2, SAM2.1
Local text and symbol	Text / symbol state	Qwen2.5-VL
Global visual state	Scene / global-state field	SigLIP2, Qwen2.5-VL
Chart and table	Axis / legend / cell state	Qwen2.5-VL
Correspondence	Cross-view / structural state	DINOv2, SigLIP2, Qwen2.5-VL

proximity and support use distance, overlap, and contact. Relative-depth states use a directed score based on apparent size and vertical position, with opposite signs for front and behind. Exchanging entities changes these directed features before candidate projection. State margins and view quality contribute to reading sufficiency.

Text and structured readout. Frozen Qwen2.5-VL readouts supply recognition hypotheses on bound spans. The text Reader normalizes strings, combines hypothesis probabilities through normalized edit-distance matching, and pools expert distributions in log space within each view. Log probabilities form state logits; localization, span completeness, and lattice coverage supply quality summaries.

For charts and tables, frozen readouts supply mark-to-series/category or cell-to-row/column links. The Reader accumulates value distributions along compatible paths, normalizing numbers before lookup or comparison. View marginalization combines estimates using their masses; unreadable or uncovered mass stays outside candidate logits.

Neither Base probabilities nor evaluator owners enter these state readouts.

B.2 Support representation and posterior computation

Retained assignment j receives mass $m_j^{\mathrm{bind}} = m(1-u)b_j$ from acquisition mass m and binding uncertainty u; b_j is its pre-retention probability. Unresolved or unretained mass enters Binder-null. With readable probability r_j^{read}, its Reader passes $w_j = m_j^{\mathrm{bind}} r_j^{\mathrm{read}}$ to the atom and the remainder to Reader-null. Figures 8 and 9 trace these assignments.

An atom $e_j = (\rho_j, \beta_j, \boldsymbol{\ell}_j, \mathbf{s}_j, u_j, w_j)$ retains provenance, ordered binding, candidate-aligned logits, centered relative evidence, uncertainty summaries, and represented mass. For native Reader logits L_j and registered candidate-state map φ_j,

$$\ell_{j,k} = L_j[\varphi_j(k)], \qquad s_{j,k} = \ell_{j,k} - \frac{1}{N}\sum_{h=1}^{N} \ell_{j,h}.$$

Registered maps preserve candidate identity using unique, in-range indices.

The deterministic Frame builds queries from candidate values, operators, and ordered roles. Candidate–atom compatibility uses the aligned logits $\boldsymbol{\ell}_j$ and typed metadata, preserving binding and quality features for learned composition. The centered $\mathbf{s}_j$ expresses relative candidate support.

With $m_{\mathrm{represented}} = \sum_j w_j$, the terminal mass partition is

$$\begin{aligned} 1 = {} & m_{\mathrm{represented}} + m_{\mathrm{reader,null}} + m_{\mathrm{reader,tail}} + m_{\mathrm{binder,null}} \\ & + m_{\mathrm{acquisition,null}} + m_{\mathrm{acquisition,tail}} + m_{\mathrm{program,unknown}} + m_{\mathrm{program,tail}}. \end{aligned}$$

These channels group into represented/null/unknown/tail. Unresolved mass is never redistributed to surviving atoms; bindings that reuse pixels retain separate identities and masses.

A valid empty bank returns Base exactly; a valid nonempty bank enters Composer and Bound. Malformed identity, projection, provenance, or schema produces failure. Visual-state errors remain possible within valid atoms (App. E.3).

Candidate-conditioned aggregation. Each candidate forms its slot representation from the same valid evidence bank. Let $\mathbf{h}_j$ encode atom j, $\mathbf{x}_k$ encode candidate k, and $\mathbf{b}_{jk}$ contain their compatibility features. Shared key maps H, Q, normalized slot keys $\mathbf{e}_v$, and compatibility map C define eight slots of width $d = 128$:

$$a_{jkv} = \frac{\langle H(\mathbf{h}_j) + Q(\mathbf{x}_k), \mathbf{e}_v \rangle + C(\mathbf{b}_{jk})_v}{\sqrt{d}},$$
$$\omega_{jkv} = w_j \frac{\exp a_{jkv}}{\sum_{u=1}^{8} \exp a_{jku}}, \qquad M_{kv} = \sum_j \omega_{jkv},$$
$$\boldsymbol{\mu}_{kv} = \frac{\sum_j \omega_{jkv}[V(\mathbf{h}_j) + U(\mathbf{b}_{jk})]}{\max(M_{kv}, \epsilon)}.$$

Shared value maps V, U form slot means; this feature normalization leaves atom weights unchanged. Absolute slot mass M_{kv} remains a separate input to attention and mass summaries. Softmax is over slots: $\sum_v \omega_{jkv} = w_j$ and $\sum_v M_{kv} = m_{\text{represented}}$ for every candidate.

Slot embeddings and query keys augment the means; weighted second moments measure disagreement. Cross-attention uses these features, assignment entropy, and represented/unresolved mass summaries. Its logits include $\log M_{kv}$, with zero-mass slots masked out. A shared head maps the attended query and candidate embedding to a residual.

Composer. Each of five members has 2,030,507 trainable parameters. With $\mathbf{q}_t = \max(\mathbf{p}_t, \epsilon)$ for $t \in \{0, 1\}$, shape scale $B = 4.0$, and learned positive α_m,

$$\mathbf{r}_m = F_{\theta_m}(\text{Frame}(\mathcal{S}_{\text{valid}}, \mathcal{C})), \qquad \mathbf{c}_m = \text{ctr}(\mathbf{r}_m),$$
$$\mathbf{d}_m = \alpha_m B \tanh(\mathbf{c}_m), \qquad \mathbf{p}_1^{(m)} = \text{softmax}(\log \mathbf{q}_0 + \mathbf{d}_m),$$
$$\mathbf{p}_1 = \frac{1}{5} \sum_{m=1}^{5} \mathbf{p}_1^{(m)}, \qquad \mathbf{z} = \text{ctr}(\log \mathbf{q}_1 - \log \mathbf{q}_0).$$

Base anchors each member after the residual head; the ensemble averages probabilities. Pre-$\tanh$ centering affects the nonlinear transformation, whose output $\mathbf{d}_m$ is generally uncentered. Final centering fixes the additive constant of the ensemble-relative log response. The learned α_m scales the shape-bounded residual, so B alone is not a bound on its magnitude.

Bound. The 24 features form four groups: four arity indicators; eight posterior summaries (entropy, normalized entropy, top probability, and top-two margin for each of $\mathbf{p}_0, \mathbf{p}_1$); seven distances (two directed KL divergences, JS divergence, TV, and $L_1/L_2/L_\infty$); and five response summaries (mean absolute value, RMS, maximum absolute value, positive fraction, and negative fraction of $\mathbf{z}$). Candidate masks exclude padding. A width-64 encoder with a width-128 residual feed-forward layer and dropout 0.1 predicts one sigmoid strength. Structural checks select exact Base, execution with valid support, or failure. The learned controller does not receive their outcomes.

For valid support and $0 < a < 1$,

$$\mathbf{p}_2 = \text{Normalize}(\mathbf{q}_0^{1-a} \odot \mathbf{q}_1^{a}), \qquad a = \sigma[A_\eta(\text{Geo}_N(\mathbf{p}_0, \mathbf{p}_1, \mathbf{z}))].$$

Endpoints $a = 0$ and $a = 1$ directly return stored $\mathbf{p}_0$ and $\mathbf{p}_1$. Bound scales all evidence-induced pairwise log-odds displacements by one scalar, without another candidate-wise correction. Candidate ranks can change; T_{auth} and L_{orig} need not improve monotonically because they compare normalized probability increments across clean and rebind calls. Refitting candidate-wise strengths tests removal of the common-scalar constraint.

B.3 Training objectives and implementation details

Figure 5 summarizes component fitting. Acquisition and Readers remain frozen; projection and authorization are parameter-free. The fitted components and checkpoint choices remain fixed across all four Base backbones and external tasks.

Evaluation joins owner maps, rebinding targets, and condition labels to fixed model outputs. Table 8 gives optimization settings.

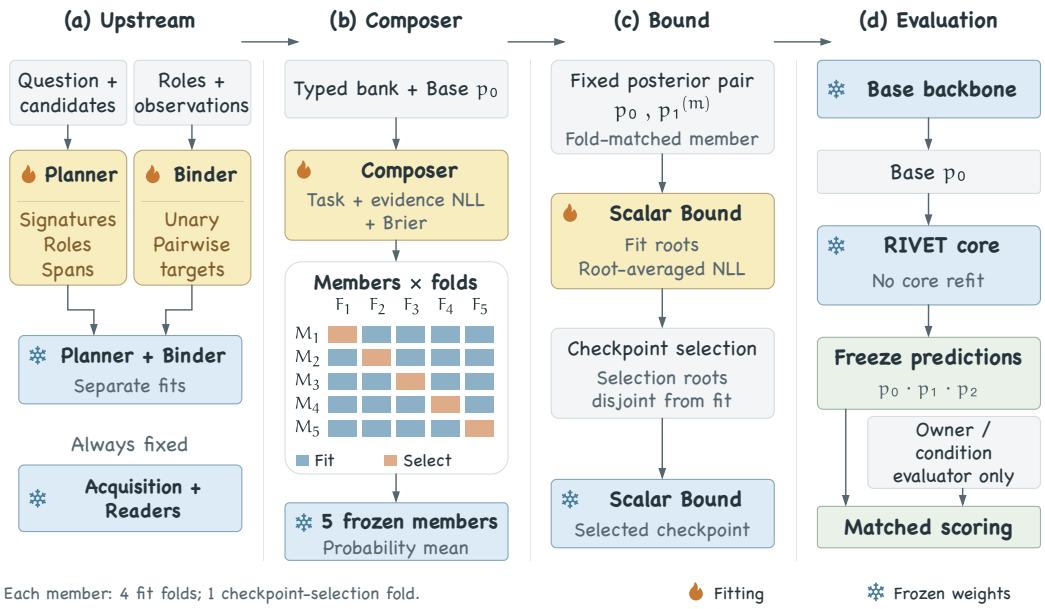


Figure 5: **Component fitting and frozen evaluation.** Flames mark fitting; snowflakes mark fixed weights or inputs. Each Composer member fits four folds and selects on the fifth. Bound uses the corresponding member's pairs from that fifth fold, with disjoint fitting/selection roots. Evaluation averages five members' probabilities.

Planner: program supervision. Benchmark training annotations and adapted GQA, TextVQA, TallyQA, human-authored ChartQA, TAT-QA, and Visual Genome records (Hudson & Manning, 2019; Singh et al., 2019; Acharya et al., 2019; Masry et al., 2022; Zhu et al., 2021; Krishna et al., 2017) supervise signatures, candidate compatibility, roles, and spans. Table renderings and Visual Genome pair/jigsaw transforms extend coverage. Semantic bootstrap precedes complete-set training. Fitting uses records assigned to training in all three partitions: source, program-template, and domain. Held-out records enter neither fitting nor checkpoint selection.

Binder: role-assignment supervision. Required-fact requests and training-image entity proposals form 13,400 unary and 3,400 ordered-pair role-assignment problems. Role-to-entity targets train entity selection and subject–object assignment through equally weighted cross-entropy objectives.

Composer: CROSS-Bench task supervision. Composer fits the 16,000 semantic train roots in Table 4(a), using predicted supports, task labels, and Base posteriors for four backbones. The 2,560 stress roots in the 18,560-root training split are excluded. Each member fits four root/image-disjoint folds and selects its checkpoint on the fifth. For target y_i, let $\boldsymbol{\pi}_{i,m} = \mathrm{softmax}(\mathbf{c}_{i,m})$ be the evidence-only distribution before $\tanh$ and Base anchoring. Member m minimizes the mean over its fitting roots of the following loss, with $K = 4$ Base records per root:

$$\mathcal{L}_{i,m} = \frac{1}{K}\sum_{b=1}^{K}\left[-\log p_{1,i,y_i}^{(m,b)} + \frac{0.05}{N_i}\|\mathbf{p}_{1,i}^{(m,b)} - \mathbf{e}_{y_i}\|_2^2\right] - 0.25\log \pi_{i,m,y_i}.$$

Table 8: **Optimization schedules and checkpoint selection.** All rows use AdamW; Planner rates are encoder/head rates. Composer and Bound ranges give minimum–maximum epochs, with stopping patience of 10 and 3 epochs, respectively.

Component	Learning rate	Schedule	Selected checkpoint
Planner	$2\times10^{-5}/10^{-4}$	1,000 steps per stage	Final step; no heldout selection
Binder	10^{-3}	25 epochs	Epoch 25
Composer	3×10^{-4}	12–60 epochs	Minimum fold-validation loss
Bound	3×10^{-4}	3–20 epochs	Minimum root NLL

Here $\mathbf{e}_{y_i}$ is the one-hot target over N_i candidates. Task NLL and candidate-averaged Brier loss act on the anchored member posterior $\mathbf{p}_1^{(m,b)}$ before Bound. Backbone averaging precedes equal weighting of roots; the evidence NLL supervises the shared residual independently of Base. Checkpoint selection evaluates this same combined objective on the member's heldout fold.

Bound: posterior-pair supervision. For each semantic train root in fold f, the Composer member fitted on the other four folds and selected on f supplies fixed $(\mathbf{p}_0, \mathbf{p}_1)$ pairs. Bound fitting and checkpoint selection use disjoint root subsets and NLL on $\mathbf{p}_2$, averaged over pairs within each root and then across roots. Ties favor earlier epochs.

Batching and regularization. Planner uses batches of 128, 100 warmup steps, and weight decay 0.01; Binder batches contain 64 role groups. Composer uses a packing budget of 2,048 supports; an oversized root forms its own batch, preserving the complete bank. Bound batches contain at most 256 roots. Binder, Composer, and Bound use weight decay 10^{-4}.

Matched decision interfaces. The interfaces in Tables 1(b) and 12 share Qwen2.5-VL-7B, evidence, candidate order, auxiliary-evidence budget, roots, and scoring. Direct Evidence Prompt serializes supports without fitting. Pooled Support Fusion learns pooled residuals (0.604M); Confidence-Gated Fusion adds a scalar gate (0.642M); Unconstrained Residual Fusion (Unconstrained Fusion) learns free candidate-wise residuals (1.514M). RIVET has five 2.03M Composer members (10.15M total) and a Bound below 0.10M. The capacity-matched control below additionally matches model and training budgets.

Calibrated aligned summation. Registered projection and absolute mass feed an additive response:

$$d_c = \sum_{j \in \mathcal{S}_{\text{valid}}} w_j \tau_{t(j)} s_{j,c}, \qquad \mathbf{p}_{\text{sum}} = \text{softmax}(\log \mathbf{q}_0 + \mathbf{d}).$$

Specialist $t(j)$ has one positive scale τ_t shared across its atoms and candidates, accommodating native logit magnitudes. Fitting and selection use Bound's disjoint root partitions and root-averaged NLL; scales remain fixed across evaluation. The uncalibrated variant sets $\tau_t = 1$.

Both variants use all valid atoms with unchanged absolute masses and return Base for an empty bank. Only Reader-scale calibration differs; neither variant learns cross-evidence interactions or posterior-dependent scalar control.

Capacity-matched Fusion. Each member encodes all valid atoms and pools independently of the queried candidate. All candidates receive the same pooled representation; candidate conditioning enters the downstream query–representation head. Missing mass and empty-bank Base return are preserved. Member width matches RIVET's per-member parameter budget; five-member averaging, training roots, folds, loss, and checkpoint selection also match. Scalar Bound is fitted separately under the same protocol. This isolates candidate-conditioned aggregation under matched capacity and structural fallback.

Execution scope. Cold execution covers image decoding, acquisition, binding, reading, and decision computation; warm execution reuses image features. Cached replay starts from preserved supports and covers the decision suffix, including Base returns and failures. Bound acts after reading, leaving upstream computation intact.

C Additional Experimental Results

C.1 Paired comparisons and statistical uncertainty

Unconstrained Fusion exposes the separation between task utility and ownership; the capacity-matched comparison tests candidate-conditioned aggregation under matched model and training budgets. Common-root, effect-scale, and threshold analyses then examine this comparison. Unless a backbone average is specified, results use Qwen2.5-VL-7B Ground and the units in App. A.2. Each table identifies its sample set and Base reference.

Comparison with unconstrained fusion. Table 9(a) reports RIVET minus Unconstrained Fusion on Qwen2.5-VL-7B Ground: 960 roots for accuracy and method-specific scoreable/active sets for posterior metrics. Panel (b) reports method-specific and common clean-active results within 707 eligible roots, with each method's b_i. Both follow the paired estimation protocol in App. A.2.

Table 9: **RIVET versus Unconstrained Fusion.** Qwen2.5-VL-7B Ground, original-root Base. (a) Matrix accuracy on 960 roots and method-specific posterior metrics. (b) Ownership with each method's own clean-effect normalizers.

(a) Paired differences

Metric	Point difference	Paired 95% CI
Matrix accuracy (pp)	-1.25	$[-3.54, +0.93]$
G_{clean}	-0.015	$[-0.073, +0.042]$
A_{inv}	-0.346	$[-0.378, -0.315]$
T_{auth}	+0.579	$[+0.511, +0.640]$
L_{orig}	-0.452	$[-0.519, -0.387]$

(b) Clean-active coverage and common-set sensitivity

Evaluation set	Interface	Active $n/707$	T_{auth}	L_{orig}
Method-specific	Unconstrained Fusion	612/707	0.072	0.657
	RIVET	521/707	0.651	0.205
Common clean-active	Unconstrained Fusion	478/707	0.082	0.623
	RIVET	478/707	0.644	0.213

Relative to Unconstrained Fusion, RIVET has lower invalid activity, higher transfer, and lower leakage, with paired 95% CIs excluding zero for all three contrasts. The paired accuracy interval includes zero: the evidence-use differences are resolved, whereas the accuracy difference is not.

Transfer under matched capacity and training. RIVET exceeds Capacity-matched Fusion in transfer by $+0.139$ on the primary metric-specific sets (paired 95% CI $[+0.057, +0.213]$). The control replaces candidate-conditioned aggregation with candidate-independent pooling, matching evidence, capacity, ensemble, and training (App. B.3).

The difference remains positive on 489 common clean-active roots: $+0.117$ (paired 95% CI $[+0.035, +0.201]$; Table 10b). With each method's own normalizers, transfer/leakage are $0.526/0.274$ for Capacity-matched Fusion and $0.643/0.218$ for RIVET.

The other primary-set RIVET–control differences are accuracy -0.94 pp (paired 95% CI $[-3.23, +1.35]$), clean utility -0.008 ($[-0.063, +0.047]$), and leakage -0.081 ($[-0.164, +0.007]$). All three CIs include zero.

Effect scale and normalization sensitivity. On these 489 roots, RIVET's unnormalized positive new-owner change is larger (7.82 versus 6.68 pp), although its clean-owner change is smaller (11.92 versus 12.64 pp; Table 10a). The positive original-owner residual is also smaller (2.63 versus 3.55 pp), complementing the normalized ownership measures.

Table 10: **Effect scale and threshold sensitivity.** (a) Positive candidate-probability changes on 489 common clean-active roots, using the original Base. (b) RIVET minus Capacity-matched Fusion on the common active set at each threshold, retaining each method's normalizers.

(a) Candidate-probability change by destination (pp)

Interface	Clean owner	Rebound new owner	Rebound original owner
Capacity-matched Fusion	12.64	6.68	3.55
RIVET	11.92	7.82	2.63

(b) Paired ownership contrasts with 95% CIs

δ_Δ	Common n	ΔT_{auth} [95% CI]	ΔL_{orig} [95% CI]
0.005	531	$+0.128$ $[+0.049, +0.211]$	-0.064 $[-0.132, +0.003]$
0.010	489	$+0.117$ $[+0.035, +0.201]$	-0.056 $[-0.125, +0.010]$
0.020	457	$+0.108$ $[+0.019, +0.198]$	-0.047 $[-0.118, +0.022]$
0.030	424	$+0.124$ $[+0.022, +0.228]$	-0.061 $[-0.139, +0.019]$
0.050	374	$+0.103$ $[-0.014, +0.230]$	-0.049 $[-0.139, +0.036]$

Table 11: **Evidence-use estimates and clean-active coverage** for the 960-root matrix in Table 1. Brackets give marginal 95% CIs; ownership uses method-specific active sets and original-Base references. Shared-evidence interfaces appear below the divider.

(a) Clean utility and invalid activity

System	G_{clean} [95% CI]	A_{inv} [95% CI]
Qwen2.5-VL-7B	$0.000\ [0.000, 0.000]$	$0.000\ [0.000, 0.000]$
VLM-R^3-7B	$0.153\ [0.092, 0.217]$	$0.162\ [0.124, 0.207]$
Pixel-Reasoner-7B	$0.128\ [0.059, 0.201]$	$0.143\ [0.103, 0.192]$
DeepEyes-7B	$0.184\ [0.126, 0.247]$	$0.171\ [0.134, 0.216]$
TreeVGR-7B	$0.161\ [0.101, 0.224]$	$0.187\ [0.148, 0.237]$
Capacity-matched Fusion	$0.289\ [0.234, 0.347]$	$0.000\ [0.000, 0.000]$
RIVET	$0.281\ [0.231, 0.333]$	$0.000\ [0.000, 0.000]$

(b) Ownership and clean-active sample size

System	Active $n/707$	$T_{\text{auth}} \uparrow$ [95% CI]	$L_{\text{orig}} \downarrow$ [95% CI]
VLM-R^3-7B	355/707	$0.413\ [0.332, 0.489]$	$0.274\ [0.182, 0.383]$
Pixel-Reasoner-7B	332/707	$0.427\ [0.338, 0.519]$	$0.253\ [0.164, 0.371]$
DeepEyes-7B	386/707	$0.462\ [0.384, 0.541]$	$0.283\ [0.191, 0.394]$
TreeVGR-7B	367/707	$0.497\ [0.416, 0.577]$	$0.308\ [0.211, 0.427]$
Capacity-matched Fusion	547/707	$0.512\ [0.444, 0.578]$	$0.286\ [0.207, 0.387]$
RIVET	521/707	$0.651\ [0.612, 0.689]$	$0.205\ [0.177, 0.236]$

Table 10(a) reports three positive candidate-probability changes: $100\mathbb{E}_i([\Delta^{\text{clean}}_{i,o_i}]_+)$, $100\mathbb{E}_i([\Delta^{\text{rebind}}_{i,n_i}]_+)$, and $100\mathbb{E}_i([\Delta^{\text{rebind}}_{i,o_i}]_+)$, with $[x]_+ = \max(x, 0)$ taken before averaging. All readouts use the original-root Base and are in pp. Transfer divides each positive new-owner increment by that root's method-specific clean-effect normalizer b_i, then clips before averaging; dividing column means does not recover T_{auth}. Leakage remains uncapped. The later signed new-owner readouts (App. C.2) also retain probability decreases.

Transfer point estimates remain higher across the five evaluated thresholds as common-set coverage decreases from 531 to 374 roots (Table 10b). The transfer CI spans zero only at the highest threshold, 0.050; leakage CIs span zero throughout. At each threshold, we intersect scoreable clean-active sets, retaining each method's b_i under shared cluster resampling.

RIVET's method-specific sets contain 401–557 active roots across these thresholds, with $T_{\text{auth}} \in [0.632, 0.667]$ and $L_{\text{orig}} \in [0.189, 0.224]$. Paired contrasts use the common sets in the table.

Table 11 reports marginal CIs and clean-active coverage for the released systems and shared-evidence interfaces. Capacity-matched Fusion combines clean utility with zero matrix invalid activity, yet transfers less than RIVET. This separates suppression of invalid contributions from successful redirection under a valid relation change.

Additive evidence and confidence gating. Aligned summation already yields substantial transfer (Table 12). Its two variants isolate Reader-scale calibration while retaining structural checks and absolute represented mass. Calibration raises the transfer point estimate from 0.581 to 0.603. For pooled evidence, confidence gating leaves invalid activity nearly unchanged: 0.476 versus 0.477 without the gate (Table 1b), despite positive clean utility. Calibrating aligned support and scaling a candidate-independent pool thus yield different evidence-use behavior.

Table 12: **Aligned summation and confidence gating.** Accuracy on 960 roots and metric-specific evidence-use estimates; ownership uses each interface's clean-active original-Base reference.

Interface	Acc. (%)	G_{clean}	A_{inv}	Active $n/707$	$T_{\text{auth}}/L_{\text{orig}}$
Confidence-Gated Fusion	81.56	0.213	0.476	558/707	0.176 / 0.357
Aligned Sum ($\tau_t = 1$)	78.85	0.203	0.000	474/707	0.581 / 0.249
Calibrated Aligned Sum	79.79	0.221	0.000	487/707	0.603 / 0.237

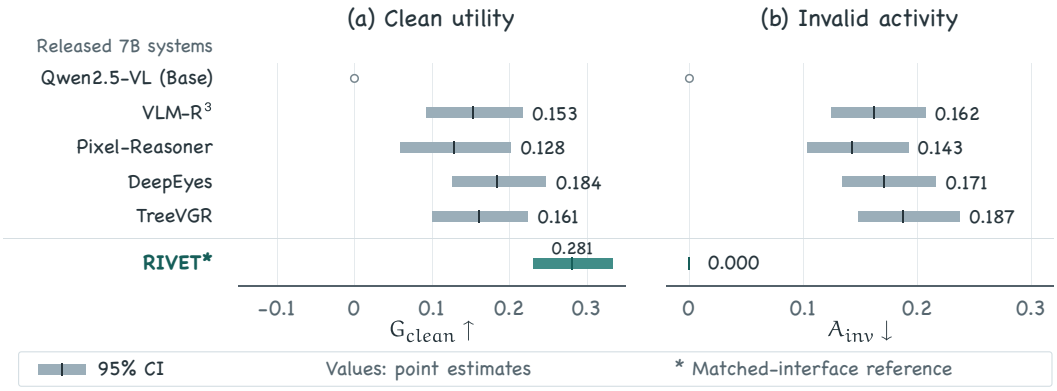


Figure 6: **Clean utility and invalid activity.** Released 7B systems are separated from the matched-interface RIVET reference (*). Markers show point estimates; shaded bands show marginal 95% CIs. Base defines zero auxiliary effect.

C.2 Semantic specificity and fixed-request controls

Released systems combine positive clean utility (Figure 6) with nonzero mean responses at every displayed invalid relation coordinate (Table 13). RIVET retains Base on these matrix controls.

Condition means use the same scoreable roots $\mathcal{O}_A^M$; A_{inv} takes the within-root maximum, including canonical null, and can exceed every condition mean. The following controls examine schema-valid errors and request-specific contributions.

Schema-valid semantic errors. RIVET retains positive utility from clean carriers while responding less to their erroneous counterparts (Table 14a). Each of the 400 clean/error pairs shares a request, valid schema, source/entity identifiers, candidate order, and Base, without an invalidity token. All pairs contribute to the clean-utility and invalid-response estimates. For example, a substituted chart-cell value retains its identifiers. RIVET and Capacity-matched Fusion share structural checks and matched model budgets.

Response frequency and magnitude. RIVET's lower mean invalid TV (0.022 versus 0.058) primarily reflects fewer responding roots than Capacity-matched Fusion (12.8% versus 34.8%). Conditional mean TV among responding roots is similar: approximately 0.172 versus 0.167, compared with 0.207 for Unconstrained Fusion.

For invalid evidence, let $v_i = \text{TV}(\mathbf{p}_i^{\text{inv}}, \mathbf{p}_i^0)$. Non-Base response rate measures response occurrence; mean TV also reflects its magnitude:

$$R_{\neq} = \mathbb{E}_i \mathbf{1}[v_i > 0], \qquad \overline{V} = \mathbb{E}_i v_i = R_{\neq}\, \mathbb{E}[v_i \mid v_i > 0] \quad (R_{\neq} > 0),$$

Here $\overline{V} = 0$ when $R_{\neq} = 0$; $100R_{\neq}$ is a percentage, with no threshold discarding small responses. Selective VQA evaluates abstention alongside answer quality (Eisenschlos et al., 2024). Here, clean utility and invalid activity distinguish selective evidence use from auxiliary-channel suppression.

Zero TV includes both direct Base return and an unchanged nonempty-bank output.

Table 13: **Invalid-evidence response by relation coordinate.** Each system uses its matrix set $\mathcal{O}_A^M$ throughout. Columns give condition-wise mean TV; A_{inv} averages the within-root maximum over these conditions and canonical null.

	Condition-wise mean TV					Aggregate
System	Source	Entity	Role	State	Rotation	A_{inv}
VLM-R^3-7B	0.096	0.117	0.138	0.088	0.124	0.162
Pixel-Reasoner-7B	0.079	0.102	0.119	0.090	0.108	0.143
DeepEyes-7B	0.101	0.123	0.146	0.095	0.132	0.171
TreeVGR-7B	0.115	0.141	0.164	0.120	0.153	0.187
RIVET	0.000	0.000	0.000	0.000	0.000	0.000

Table 14: **Evidence use under structurally valid controls.** (a) 400 clean/error pairs. (b) Natural-log gains on the 638 common roots of Table 20; n is the number of roots. (c) Candidate-probability changes on a separate 654-root common cohort. Panels (b–c) use the new request's own Base without clean-active filtering.

(a) Semantic specificity

Decision interface	Clean $G_{\text{clean}} \uparrow$	Invalid mean TV$\downarrow$	Non-Base rate (%)$\downarrow$
Unconstrained Fusion	0.245	0.168	81.3
Capacity-matched Fusion	0.231	0.058	34.8
RIVET	0.224	0.022	12.8

(b) Request-matched evidence contribution

Decision interface	n	Matched $G_{q'}$	Crossed $G_{q'}$	Matched$-$crossed
Unconstrained Fusion	638	0.244	0.116	0.128
Capacity-matched Fusion	638	0.225	0.063	0.162
RIVET	638	0.229	0.022	0.207

(c) Matched-evidence candidate-probability changes (pp)

Decision interface	Signed new-owner change	Positive original-owner residual
Unconstrained Fusion	+8.43	4.72
Capacity-matched Fusion	+8.08	2.11
RIVET	+8.19	1.03

Metadata-only control. Removing state-bearing pixels/text leaves a balanced identifier-and-format probe ($n = 800$; accuracy 50.8%). This near-chance shortcut check (Hsieh et al., 2023) indicates little validity signal for the tested probe.

Selectivity at a fixed request. Table 14(b) fixes the image, new request q'_i, candidate order, and freshly evaluated Base, comparing null, crossed, and matched evidence. Crossed carriers retain the registered type and valid schema but refer to the wrong entity or ordered role. The request identifies the relation without supplying its target state.

For the new owner n_i, define

$$G_{q'}(e) = \mathbb{E}_i \log \frac{\max(p(n_i \mid q'_i, e), \epsilon)}{\max(p(n_i \mid q'_i, \varnothing), \epsilon)}.$$

Panel (b) reuses Table 20's 638 common roots from 707 rebinding-eligible roots, requiring scoreability across six calls, four interfaces, and the additional anchor calls. No clean-active filtering is applied; null gain is zero. RIVET and Capacity-matched Fusion have similar matched log-gain point estimates (0.229 versus 0.225), while RIVET has lower crossed gain (0.022 versus 0.063). Their matched–crossed contrasts are 0.207 and 0.162.

Candidate-level effects. Panel (c) uses 654 roots scoreable across all three calls and interfaces, without clean-active filtering. For matched-evidence change $d^M_{i,c} = p_M(c \mid q'_i, e_i^{\text{match}}) - p_M(c \mid q'_i, \varnothing)$, it reports signed new-owner change $100\mathbb{E}_i d^M_{i,n_i}$ and positive original-owner residual $100\mathbb{E}_i([d^M_{i,o_i}]_+)$. The signed mean includes decreases in new-owner probability; the residual averages only positive increments at the original owner.

RIVET and Capacity-matched Fusion have similar new-owner point estimates (difference $+0.11$ pp; paired 95% CI $[-0.91, +1.13]$). RIVET's original-owner residual point estimate is lower (difference -1.08 pp; CI $[-2.19, +0.04]$); both paired CIs span zero.

The semantic-error pairs measure whether a carrier changes the posterior; the fixed-request comparisons locate its effect at the requested candidate. App. C.4 extends the latter comparison to both requests on common roots, and App. D.3 applies it to independently constructed relations. The following controls examine how evidence representation and response strength affect these behaviors.

Table 15: **State and interface interventions in Table 3.**

Control	Change to the shared interface	Execution mode
Hard program selection	Keep the highest-mass Program and record removed mass as tail; downstream models stay fixed.	Frozen replay
Shuffled evidence bindings	Apply one nonidentity permutation of legal role assignments where possible; preserve pixels, mass, and joins.	Frozen replay
Renormalize represented mass	Set represented mass to one and unresolved channels to zero; recompute mass-derived Composer features.	Frozen replay
Pre-Compose evidence selection	Keep the highest-mass atom; put removed mass in tail and preserve empty-bank Base.	Frozen replay
No scalar Bound (D_1)	Return stored D_1, preserving supports and Composer.	Endpoint replay
Candidate-wise Bound	Fit candidate-wise strengths on the same posterior inputs; keep the Base bypass.	Controller refit
State-only Oracle	Replace only state logits by the annotated state; fix bindings, mass, and the predicted-state reference.	State replay

C.3 Mechanism, evidence access, and task structure

Representation and state controls. Table 15 specifies each intervention, the retained inputs, and whether parameters are refitted. All controls preserve the ordinary root, Base, and candidate order; ownership references follow App. A.2.

Shuffling retains most accuracy gain (4.63 versus 4.96 pp) but reduces transfer from 0.651 to 0.296 and increases leakage from 0.205 to 0.402 (Figure 7; Table 3b). In the fitted interface, represented-mass renormalization ("Mass norm.") and preselection also reduce transfer and increase leakage. Candidate-wise control raises clean utility while degrading ownership. Hard program selection yields both low transfer (0.083) and low leakage (0.056); low leakage alone therefore does not establish successful redirection.

Preselection ranks absolute represented mass, breaking ties by support identity. Shuffling uses a root-seeded nonidentity permutation when available; single-assignment roots remain in evaluation unchanged. At Composer's input, renormalization transforms a validated terminal bank with $r = m_{\text{represented}} > 0$ to $w'_j = w_j/r$, $m'_{\text{represented}} = 1$, and zero null/unknown/tail mass; other atom fields stay fixed. Mass-derived features and summaries are recomputed. Structural validity, nonnegative mass, and unit total mass are checked before frozen replay. This jointly changes absolute weights and explicit mass summaries while preserving relative atom weights.

These evidence-bank transformations preserve empty banks: $\mathcal{T}(\varnothing) = \varnothing$, so induced invalid activity requires nonempty support. Frozen replay tests reliance on retained evidence structure; controller refitting tests additional candidate-wise freedom after response formation.

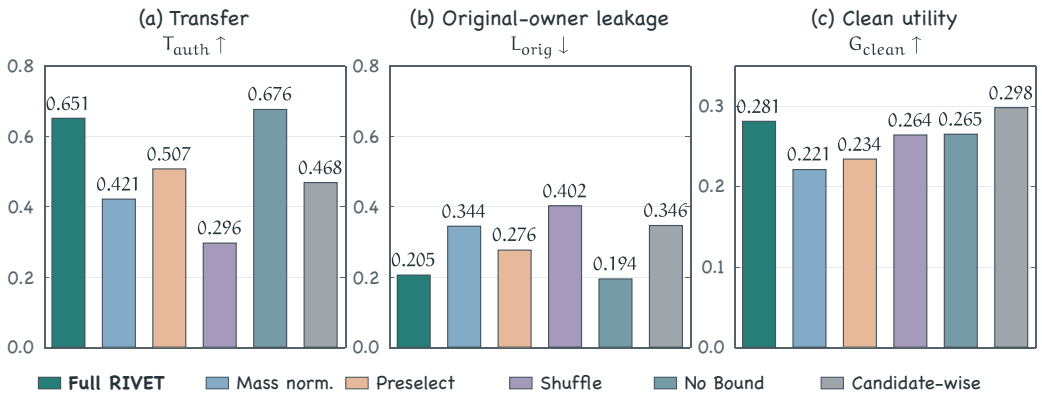


Figure 7: **Selected representation and control ablations.** Point estimates of transfer, original-owner leakage, and clean utility on Qwen2.5-VL-7B Ground; ownership uses the fixed Full RIVET reference.

Table 16: **Response-strength controls with fixed evidence and Composer.** Qwen2.5-VL-7B Ground: accuracy gain on 5,600 roots; matrix utility; ownership on Full RIVET's fixed 521 active roots and clean-effect normalizers.

Controller	ΔAcc. (pp)	G_{clean}	$T_{\text{auth}}/L_{\text{orig}}$
No scalar Bound (D_1)	+4.64	0.265	0.676 / 0.194
Global scalar $a^\star$	+4.82	0.274	0.661 / 0.201
Learned scalar Bound	+4.96	0.281	0.651 / 0.205

State estimation at fixed bindings. State-only Oracle increases transfer from 0.512 to 0.564 for Capacity-matched Fusion and from 0.651 to 0.685 for RIVET, retaining their point-estimate ordering (Table 3a). It sets the annotated native state's logit to zero and others to $\log \epsilon$ (App. B.2), then reapplies registered projection. Bindings, mass, quality features, Base, and the predicted-state ownership reference remain fixed; owner annotations stay hidden. Entity or role-assignment errors remain unchanged.

For each interface M, both state conditions use $\mathcal{D}^{+}_{M,\text{pred}} \cap \mathcal{O}^{M}_{\text{pred,oracle}}$, the predicted-state active roots scoreable in both replays, with $b^{\text{pred}}_{i,M}$ fixed. This reference contains 612, 547, and 521 roots for Unconstrained Fusion, Capacity-matched Fusion, and RIVET, respectively. The paired evaluation retains disappearing effects.

Global versus posterior-dependent response strength. Observed accuracy and clean utility increase from D_1 through the global scalar to learned Bound, while ownership point estimates favor the unattenuated response (Table 16). The ownership pattern is already present before Bound, which calibrates response strength.

These controls hold the complete evidence bank, Base, and composed endpoint fixed. The global scalar $a^\star \in [0, 1]$ minimizes candidate NLL on Bound's selection roots and applies unchanged to every evaluation call through Eq. 7. Learned Bound predicts a scalar from each posterior pair. Both retain the endpoint and empty-bank bypasses. Accuracy uses all 5,600 roots; ownership fixes Full RIVET's 521 active matrix roots and normalizers, as in Table 3(b). On identical clean roots with the same weights and clipping, $G_{\text{clean}} = \text{NLL}(\mathbf{p}_0) - \text{NLL}(\mathbf{p}_M)$, so clean utility and NLL reduction express the same log-score change.

Response formation and evidence access. About 95% of the route-balanced four-backbone accuracy gain is present at D_1, which includes the full upstream chain and Compose; Bound adds 0.32 pp (Table 17a; 5,600 roots). Qwen2.5-VL-7B Ground has 515 active roots at D_1 and 521 at D_2. These endpoint-specific sets differ from the fixed 521-root reference used for ablations.

Restoring Program, acquisition, and binding cumulatively raises the four-backbone mean gain from 5.70 to 6.29, 7.03, and 7.54 pp (panel b), with Readers and decision modules fixed. The total increase is 1.84 pp; each step is conditional on preceding restorations and does not rank independent module contributions. The State-only Oracle separately tests reading with access fixed.

Table 17: **Endpoint gains and ordered access restoration.** Four-backbone mean accuracy changes (pp) on 5,600 roots. Panel (a) weights the two routes equally; panel (b) restores access cumulatively.

(a) Endpoint contrasts

Contrast	Controlled	Ground	Mean over routes
Full response $D_1 - D_0$	+7.21	+5.39	+6.30
Scalar-controlled response $D_2 - D_0$	+7.54	+5.70	+6.62

(b) Cumulative access restoration

Access setting	ΔAcc. (pp)
Ground: predicted upstream	+5.70
+ verified Program	+6.29
+ verified acquisition	+7.03
Controlled: verified Program, acquisition, and binding	+7.54

Table 18: **Ground gains by evidence family, candidate count, and split.** Each partition covers the same 5,600 roots. Gains average over four backbones; root-weighted subgroup means recover the full-suite gain.

(a) Evidence family

Family	Roots	ΔAcc. (pp)
Object identity	720	+7.2
Object attribute	680	+6.0
Spatial comparison	700	+7.5
Count / set	690	+5.4
Local text / symbol	710	+6.5
Global scene	700	+6.2
Chart / table	690	+4.2
Correspondence structure	710	+2.6

(b) Arity and split

Subset	Roots	ΔAcc. (pp)
$N = 2$	1,320	+6.4
$N = 3$	1,510	+6.0
$N = 4$	1,580	+5.8
$N = 5$	1,190	+4.4
Semantic	4,800	+5.8
Stress	800	+5.1

Accuracy by evidence family, candidate count, and split. Table 18 partitions the same 5,600 roots by family, candidate count, and split. The four-backbone mean Ground gains are positive in every displayed subgroup; root-weighted aggregation recovers +5.70 pp up to rounding.

Mean gains are smaller for correspondence and five-candidate problems. Family, candidate count, and split provide complementary views of the same test roots. These subgroup gains describe task utility; the access, state-replacement, and relation controls examine the mechanisms and ownership of that utility.

Diagnostic-subset concentration. Qwen2.5-VL-7B's Ground accuracy gain is larger on the diagnostic matrix than outside it: 18.44 versus 2.18 pp. These correspond to 177 net additional correct decisions on 960 matrix roots and 101 on the remaining 4,640 roots, yielding 278 on the full 5,600-root suite ($+4.96$ pp). Table 1 measures matrix utility and ownership; Table 2(a) reports full-suite benefit.

The matrix is selected to admit complete paired interventions, emphasizing roots whose evidence relations can be systematically changed. Reporting the full suite alongside it keeps the diagnostic subset's sensitivity separate from overall task benefit.

C.4 Crossed request–evidence analysis

We extend the fixed-new-request comparison in App. C.2 to both requests on one common cohort. Each request uses its own no-evidence Base and both carriers, separating matched from crossed contributions. The same roots also support a comparison of candidate-level effects under the two Base anchors.

Requests, carriers, and common roots. The requests $q_{i0} = q_i$ and $q_{i1} = q_i'$ have different supported candidates, $t_{i0} = o_i$ and $t_{i1} = n_i$. Carriers E_{i0} and E_{i1} match their respective requests and remain readable but mismatched to the other request. They retain valid schemas and registered types (App. C.2); supported candidates remain evaluator annotations. All six calls in Table 19 preserve the image, candidate identities and order, and inference budget.

Both relations and their crossed mismatches must be unambiguous. A common cohort of 638 of the 707 rebinding-eligible roots is scoreable across all six calls for the four interfaces and the additional anchor calls reported below. Every panel uses these roots without clean-active filtering, including valid Base returns. Each root contributes both requests to every interface.

Table 19: **Crossed request–evidence calls.** Each row fixes the request and its Base; columns vary only the auxiliary carrier.

Request	Null ∅	Original E_{i0}	Rebound E_{i1}
q_{i0}	Base for q_{i0}	Matched; target o_i	Unmatched to q_{i0}
q_{i1}	Base for q_{i1}	Unmatched to q_{i1}	Matched; target n_i

Table 20: **Per-request contribution and reference sensitivity.** (a) Natural-log gains under each request's own Base. (b) Signed new-owner changes and positive original-owner residuals (pp); composition and scalar control are recomputed for each Base anchor. All panels use the same 638 scoreable roots without clean-active filtering.

(a) Matched and crossed log gains by request

	Original request q_{i0}		Rebound request q_{i1}	
Interface	Matched $G_{0,0}$	Crossed $G_{0,1}$	Matched $G_{1,1}$	Crossed $G_{1,0}$
Unconstrained Fusion	0.304	0.144	0.244	0.116
Calibrated Aligned Sum	0.219	0.035	0.183	0.031
Capacity-matched Fusion	0.289	0.065	0.225	0.063
RIVET	0.283	0.020	0.229	0.022

(b) Candidate changes by Base anchor (pp)

	Original Base anchor		New-request Base anchor	
Interface	New owner	Old residual	New owner	Old residual
Capacity-matched Fusion	6.71	3.42	7.97	2.08
RIVET	7.58	2.49	8.13	1.14

Within-request contribution. Let $p_M(c \mid q, E)$ use the no-evidence Base for request q. Extending $G_{q'}$ to $r, h \in \{0, 1\}$ gives matched $(r = h)$ and crossed $(r \neq h)$ gains:

$$G_{r,h}(M) = \mathbb{E}_i \log \frac{\max(p_M(t_{ir} \mid q_{ir}, E_{ih}), \epsilon)}{\max(p_M(t_{ir} \mid q_{ir}, \varnothing), \epsilon)},$$

Null log gain is zero. All four entries average over the same equally weighted roots at the candidate supported by the fixed request. Within each request, the matched–crossed contrast cancels the shared log-Base term and compares log target probabilities under the two carriers. It measures selectivity at a fixed target, complementing T_{auth}, which measures redirection relative to the original clean effect.

Panel (a) compares matched and crossed natural-log gains. Panel (b) reports matched-evidence probability changes in pp, using the signed changes and positive residuals defined in App. C.2.

Similar matched gain, lower crossed gain. On the common roots in Table 20(a), RIVET and Capacity-matched Fusion have close matched log-gain point estimates: 0.283 versus 0.289 under q_i, and 0.229 versus 0.225 under q_i'. RIVET's crossed gains at the same target candidates are lower: 0.020 versus 0.065, and 0.022 versus 0.063. The larger matched–crossed contrasts therefore arise mainly from smaller target-candidate gains under mismatched evidence. Unconstrained Fusion has the largest matched gains but also the largest crossed gains; Calibrated Aligned Sum reduces both.

Reference sensitivity. Panel (b) compares q_i' with its matched carrier under the original and new-request Base anchors. Both branches recompute composition and scalar control with their specified anchor. On the common cohort, each reports the signed new-owner mean $100\mathbb{E}_i(p^M_{i,n_i} - p^{\text{anchor}}_{i,n_i})$ and the positive original-owner residual $100\mathbb{E}_i([p^M_{i,o_i} - p^{\text{anchor}}_{i,o_i}]_+)$. Holding roots and readouts fixed distinguishes anchor sensitivity from changes in cohort or normalization.

RIVET's new-owner point estimate exceeds Capacity-matched Fusion's by 0.87 pp under the original Base and 0.16 pp under the new-request Base. Its positive original-owner residual is lower by 0.93 and 0.94 pp, respectively. Thus, the new-owner difference depends on the reference, while the residual separation is similar across anchors.

D Transfer, Robustness, and Evaluation Coverage

We evaluate task utility on external benchmarks, ownership and fixed-request selectivity on independently authored relations, and the stability of the CROSS-Bench comparison across decision-layer refits. Full-set accuracy retains Base fallbacks and counts failed responses as incorrect; posterior analyses use their stated scoreable cohorts.

Table 21: **Per-backbone accuracy (%) under frozen Ground transfer.** Enhanced interfaces share predicted evidence. CV-Bench uses source-balanced 2D and official Overall scores; Δ is RIVET minus Base (pp), computed before rounding.

		V*Bench	MMStar	CV-Bench			MME-RW-L
Backbone	Endpoint			2D	3D	Overall	
Qwen2.5-VL-7B	Base	76.44	60.47	75.40	73.25	74.33	44.76
	Capacity-matched	81.68	63.20	81.01	77.58	79.30	52.21
	RIVET	82.72	63.47	82.01	78.17	80.09	53.26
	Δ	+6.28	+3.00	+6.61	+4.92	+5.76	+8.49
Qwen3-VL-8B	Base	82.20	54.47	75.43	91.75	83.59	50.13
	Capacity-matched	87.43	59.40	80.41	92.75	86.58	57.27
	RIVET	86.91	59.00	80.50	92.92	86.71	57.01
	Δ	+4.71	+4.53	+5.07	+1.17	+3.12	+6.88
InternVL3.5-8B	Base	69.11	68.53	78.80	85.92	82.36	49.77
	Capacity-matched	76.44	70.67	81.83	88.25	85.04	56.70
	RIVET	77.49	70.20	82.35	88.58	85.46	57.53
	Δ	+8.38	+1.67	+3.54	+2.67	+3.11	+7.76
LLaVA-OneVision-2-8B	Base	82.20	63.60	80.07	92.33	86.20	52.11
	Capacity-matched	85.86	66.40	83.79	93.17	88.48	57.43
	RIVET	86.39	66.20	83.50	93.08	88.29	57.11
	Δ	+4.19	+2.60	+3.44	+0.75	+2.09	+5.00
Mean over backbones	Base	77.49	61.77	77.43	85.81	81.62	49.19
	Capacity-matched	82.85	64.92	81.76	87.94	84.85	55.90
	RIVET	83.38	64.72	82.09	88.19	85.14	56.23
	Δ	+5.89	+2.95	+4.66	+2.38	+3.52	+7.03

D.1 Frozen Ground transfer

Frozen Ground improves the accuracy point estimate over Base in all 16 backbone–benchmark pairs. Table 21 expands Table 2(b) into per-backbone accuracies on full V*Bench (191 questions), MMStar (1,500), CV-Bench (1,438 2D and 1,200 3D), and English MME-RealWorld-Lite (MME-RW-L; 1,919) evaluation sets (Wu & Xie, 2024; Chen et al., 2024a; Tong et al., 2024; Zhang et al., 2025).

For each backbone, the Base branch retains its preprocessing, prompts, and decoding; native candidate order and scoring stay fixed across interfaces. Capacity-matched Fusion and RIVET consume the same predicted evidence bank with frozen decision cores. Each interface evaluates every question once, retaining all candidates. Unsupported arities or states return Base; missing/unparseable answers score zero. CV-Bench uses the source-balanced aggregation below.

Aggregation. Each interface's per-backbone CV-Bench 2D score equally weights ADE20K and COCO; Overall equally weights 2D and Omni3D (source weights $1/4$, $1/4$, $1/2$). Table 2(b) and Table 21's final block report descriptive four-backbone means on shared evaluation questions. Gains and means are computed before rounding within each benchmark.

Task-level transfer. The per-backbone rows show how gain magnitudes vary alongside each model's starting accuracy. All four backbones improve on both CV-Bench partitions, with 3D gains ranging from $+0.75$ to $+4.92$ pp and Overall gains from $+2.09$ to $+5.76$ pp.

Capacity-matched Fusion also improves from the same evidence, with a higher mean MMStar score and higher LLaVA CV-Bench scores than RIVET. Both interfaces retain task utility, with relative ordering varying by benchmark and backbone. Apps. C.1 and C.4 separately examine effect destination and selectivity.

D.2 Ownership on independently authored relations

Task transfer measures answer quality; the following probes test ownership independently of CROSS-Bench templates. Using Flickr30k Entities images and region–phrase annotations (Plum-

Table 22: **Ownership on 320 independently authored visual relations.** (a) Full-set accuracy (%), scoreable Base/clean/rebind pair counts with active counts in parentheses, and original-Base ownership. (b) Natural-log gains at the new request's supported candidate, using its own Base on scoreable pairs without active filtering.

(a) Clean utility and ownership

Interface	Acc. (%)	G_{clean}	Pairs (active)	T_{auth}	L_{orig}
Unconstrained Fusion	75.31	0.208	308 (238)	0.174	0.492
Capacity-matched Fusion	74.38	0.201	307 (220)	0.447	0.309
RIVET	72.81	0.187	304 (209)	0.563	0.254

(b) Request-matched evidence contribution

Interface	Pairs	Matched $G_{q'}$	Unmatched $G_{q'}$
Unconstrained Fusion	308	0.182	0.083
Capacity-matched Fusion	307	0.177	0.053
RIVET	304	0.169	0.029

mer et al., 2015), annotators author 320 roots: 160 attribute and 160 ordered-spatial relations, each with 2–5 candidates and a distinct alternative owner. They inspect images without model predictions, resolve clean and rebound referents, and exclude ambiguous relations and source-image overlap with CROSS-Bench or training collections. Eligibility follows the visual relation without consulting RIVET's registry or model predictions.

Shared inputs and scoring. Qwen2.5-VL-7B Ground supplies one frozen predicted evidence bank to the three decision interfaces. Annotation boxes, owner labels, and true states stay evaluator-side; prediction receives the image, questions, ordered candidates, and observable evidence. No component is adapted to the probe set. Unsupported operations return Base, and failed ordinary answers count as incorrect. Clean accuracy includes all 320 roots; G_{clean} uses the method-specific roots with scoreable Base/clean/rebind posteriors, reported as pairs in Table 22. Ownership additionally requires each method's clean effect on these roots to exceed δ_{Δ}. The threshold and candidate normalizers follow App. A.2.

Capacity-matched Fusion has higher clean accuracy and utility, while RIVET has higher transfer and lower leakage point estimates. On method-specific active sets, RIVET's differences from the control are $+0.116$ (paired 95% CI $[-0.009, +0.234]$) and -0.055 ($[-0.141, +0.029]$), respectively. Both intervals include zero. On 197 common clean-active roots, RIVET and the control have transfer 0.551 and 0.458, and leakage 0.263 and 0.296, retaining each method's clean normalizer. Both sets retain the main comparison's ownership point-estimate ordering (Table 1).

Panel (b) fixes the new request q' and its freshly evaluated Base. Its null, matched, and unmatched calls are jointly scoreable on each method's panel (a) pair set. As in App. C.2, $G_{q'}$ measures contribution to the candidate supported by q', with zero null gain. At this fixed request, RIVET has a slightly lower matched-gain point estimate than Capacity-matched Fusion (0.169 versus 0.177) and a lower unmatched gain (0.029 versus 0.053).

D.3 Selectivity on an independent relation collection

Complementing the ownership probes in App. D.2, this collection tests fixed-request evidence selectivity on 900 independently constructed roots spanning attribute, ordered-spatial, count/set, and local-text relations. Native-system diagnosis and a shared-evidence comparison examine released reasoners and decision interfaces under their respective input conditions. The roots are disjoint from the preceding 320-root study.

Construction and evaluation. Flickr30k Entities provides localized referents, COCO provides instance/set annotations, and TextOCR provides text spans (Plummer et al., 2015; Lin et al., 2014; Singh et al., 2021). Paired requests are authored from images and annotations without model outputs or RIVET's registry. Two reviewers resolve the requested referents and states; each root admits two unambiguous, answer-neutral requests with distinct supported candidates and readable, schema-valid cross-mismatched carriers. Source-image clusters are disjoint from CROSS-Bench,

Table 23: **Independent-relation diagnosis and fixed-request selectivity.** (a) Accuracy on all 900 roots, clean utility, and original-Base ownership; counts give scoreable Base/clean/rebind roots with clean-active counts in parentheses. (b) Natural-log gains averaged equally over two requests on 830 common roots, with request-specific Bases and no active filtering.

(a) Native-system diagnosis

System	Acc. (%)	G_{clean}	Scoreable (active)	T_{auth}	L_{orig}
DeepEyes-7B	75.11	0.184	858 (439)	0.398	0.327
TreeVGR-7B	74.33	0.176	847 (421)	0.426	0.341

(b) Shared evidence: means over both requests

Interface	Mean matched	Mean crossed	Matched−crossed
Calibrated Aligned Sum	0.176	0.035	0.141
Capacity-matched Fusion	0.204	0.059	0.145
RIVET	0.199	0.026	0.173

component-training collections, and the preceding probes. Unsupported operations remain included and follow each system's declared fallback.

The native arm evaluates DeepEyes and TreeVGR with images, requests, candidates, and observable carriers. Ordinary accuracy uses all 900 roots; posterior and clean-active sets are method-specific. The shared-evidence arm compares Calibrated Aligned Sum, Capacity-matched Fusion, and RIVET on one frozen Qwen2.5-VL-7B Ground evidence bank, holding fitted components fixed. Inference and scoring follow App. A.2.

Common cohort and aggregation. The shared-evidence arm uses a fixed common cohort of 830 roots scoreable for all three interfaces across the six calls in App. C.4. Every method and condition uses these same roots, without clean-active filtering, including valid Base returns. Each request has its own Base. Panel (b) reports equal-request means: matched $(G_{0,0} + G_{1,1})/2$ and crossed $(G_{0,1} + G_{1,0})/2$, with equal root weights within each term. The final column subtracts these two means. All three readouts concern the candidate supported by the fixed request.

Native-system diagnosis. The systems have different point-estimate orderings across metrics. DeepEyes has higher accuracy than TreeVGR (75.11% versus 74.33%). On their respective clean-active sets, TreeVGR has both higher transfer (0.426 versus 0.398) and higher original-owner leakage (0.341 versus 0.327).

Shared-evidence selectivity. The mean matched log-gain point estimates are close for RIVET and Capacity-matched Fusion (0.199 versus 0.204), while the mean crossed gain is lower for RIVET (0.026 versus 0.059). Their matched–crossed contrasts are 0.173 and 0.145. RIVET's larger contrast reflects a greater reduction in crossed gain than in matched gain relative to the control. Calibrated Aligned Sum has lower matched gain and a contrast of 0.141. On this independently constructed collection spanning four relation types, the equal-request means retain the aggregate point-estimate pattern in App. C.4.

D.4 Robustness to decision-layer training randomness

Across three seeds, we refit the complete five-member Composer and scalar Bound while fixing the Planner, acquisition, Binder, Readers, Base posteriors, evidence banks, and data partitions. Capacity-matched Fusion shares seeds, training budget, and checkpoint selection; each complete fitted system is one repetition.

Table 24 reports Qwen2.5-VL-7B Ground: accuracy gain on 5,600 roots, and clean utility and ownership on the 960-root matrix. Run 1 is the primary fit reported in Tables 1–3. Runs share Base but use their own scoreable and clean-active sets; mean and sample standard deviation (SD) summarize the three fits.

Transfer and leakage retain the same method ordering across all three paired decision-layer fits with fixed upstream evidence: RIVET has higher mean transfer (0.636 versus 0.513) and lower mean leakage (0.220 versus 0.290). Paired transfer differences range from $+0.104$ to $+0.139$, and leakage

Table 24: **Three-seed variation with fixed upstream evidence.** Accuracy gain on 5,600 roots; utility and method/run-specific ownership on the 960-root matrix. Summaries are mean ± sample SD across three fits.

Interface	Run	ΔAcc. (pp)	G_{clean}	T_{auth}	L_{orig}
Capacity-matched Fusion	1	+4.98	0.289	0.512	0.286
	2	+5.30	0.275	0.491	0.310
	3	+4.79	0.296	0.535	0.273
	Mean ± SD	5.02 ± 0.26	0.287 ± 0.011	0.513 ± 0.022	0.290 ± 0.019
RIVET	1	+4.96	0.281	0.651	0.205
	2	+4.61	0.264	0.618	0.237
	3	+5.11	0.276	0.639	0.219
	Mean ± SD	4.89 ± 0.26	0.274 ± 0.009	0.636 ± 0.017	0.220 ± 0.016

differences from -0.081 to -0.054. Accuracy-gain ordering changes across fits, with means of 4.89 and 5.02 pp.

D.5 Evaluation coverage and failure handling

On 5,600 Qwen2.5-VL-7B Ground roots, the mutually exclusive execution outcomes are structurally valid nonempty support (78.4%), empty-bank exact-Base fallback (16.9%), structural rejection (2.8%), unsupported acquisition/Reader states (1.4%), and infrastructure failure (0.5%). The empty-bank category covers otherwise supported execution; unsupported states are classified separately. Empty banks and unsupported states both return Base. Valid nonempty support proceeds to the decision interface.

Structural and infrastructure failures account for the 185 absent ordinary responses in Table 6 (3.3% after rounding). These roots remain in the 5,600-root accuracy denominator and score zero. Missing posteriors follow the metric-specific scoreability rules in App. A.2 and are not imputed. Non-Base response rate measures posterior change: valid nonempty support can also leave Base unchanged.

E Mechanism Illustrations and Limitations

These mechanism illustrations pair real images with schematic numerical traces to explain typed visual reading and candidate-specific evidence effects. Clean and wrong-match conditions share a request; valid rebinding changes the request and its evidence relation together. Candidate order and the original Base remain fixed references.

Reading the eight relation illustrations. The upper trace follows acquisition, binding, typed reading, and composition. Multiple binding hypotheses there belong to one clean execution and enter the same support bank. A dashed Reader branch instead depicts a separate rebind condition. Native state profiles describe the visual estimate; candidate-aligned centered logits express its relative support for each answer option. Represented, null, unknown, and tail mass retain uncertainty. Compose combines the bank into a candidate-conditioned response, and Bound scales its strength.

The lower bars show $100(p^{\iota}_{2,c} - p_{0,c})$ in pp for the original and new owners. They visualize the response's destination, rather than the normalized T_{auth} or L_{orig} metrics. Wrong-match and rebind reuse the same readable observation with different requests. Their contrast shows how that observation's effect changes with its relevance to the requested relation.

The upper $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2$ vectors describe the clean trajectory. Read each highlighted probability with its candidate label: the leading candidate can change between endpoints. The lower rows show separate clean, wrong-match, and rebind calls. For each condition, probability conservation gives

$$\sum_{c \in \mathcal{C}} (p^{\iota}_{2,c} - p_{0,c}) = 0.$$

Only the two owner coordinates are drawn; their bars need not cancel because other candidates carry the balancing change. A negative bar indicates a probability decrease relative to Base, not an unreadable observation. These bars reflect the complete bank, Base, and scalar control, unlike atom-level centered support. See Apps. C.4 and D.3 for quantitative comparisons.

E.1 MATCHED RELATIONS AND EVIDENCE EFFECTS

Object-bound color. In Figure 8, a red umbrella is readable in both the wrong-match and rebind conditions. It supports A only when the request also changes to its holder P4. The clean request instead concerns the orange-shirted boy P2 and his blue umbrella P1. The contrast separates recognizing a color from using it for the requested object.

The two clean binding hypotheses also expose a different source of uncertainty. Their probabilities, 0.82 and 0.18, describe alternative holder–umbrella assignments; their represented masses, 0.623 and 0.137, additionally account for availability and readability. Both enter the same bank with their own object identities. The remaining mass stays outside candidate support, so competing hypotheses do not exhaust the uncertainty. The bank therefore carries both possible assignments into composition, together with the mass left unresolved. In the lower comparison, the red observation stays fixed while the requested holder changes. Its candidate effect depends on that holder identity as well as the color state, keeping assignment uncertainty distinct from color-reading confidence. The blue P1 and red P3 readings retain separate holder assignments within the bank.

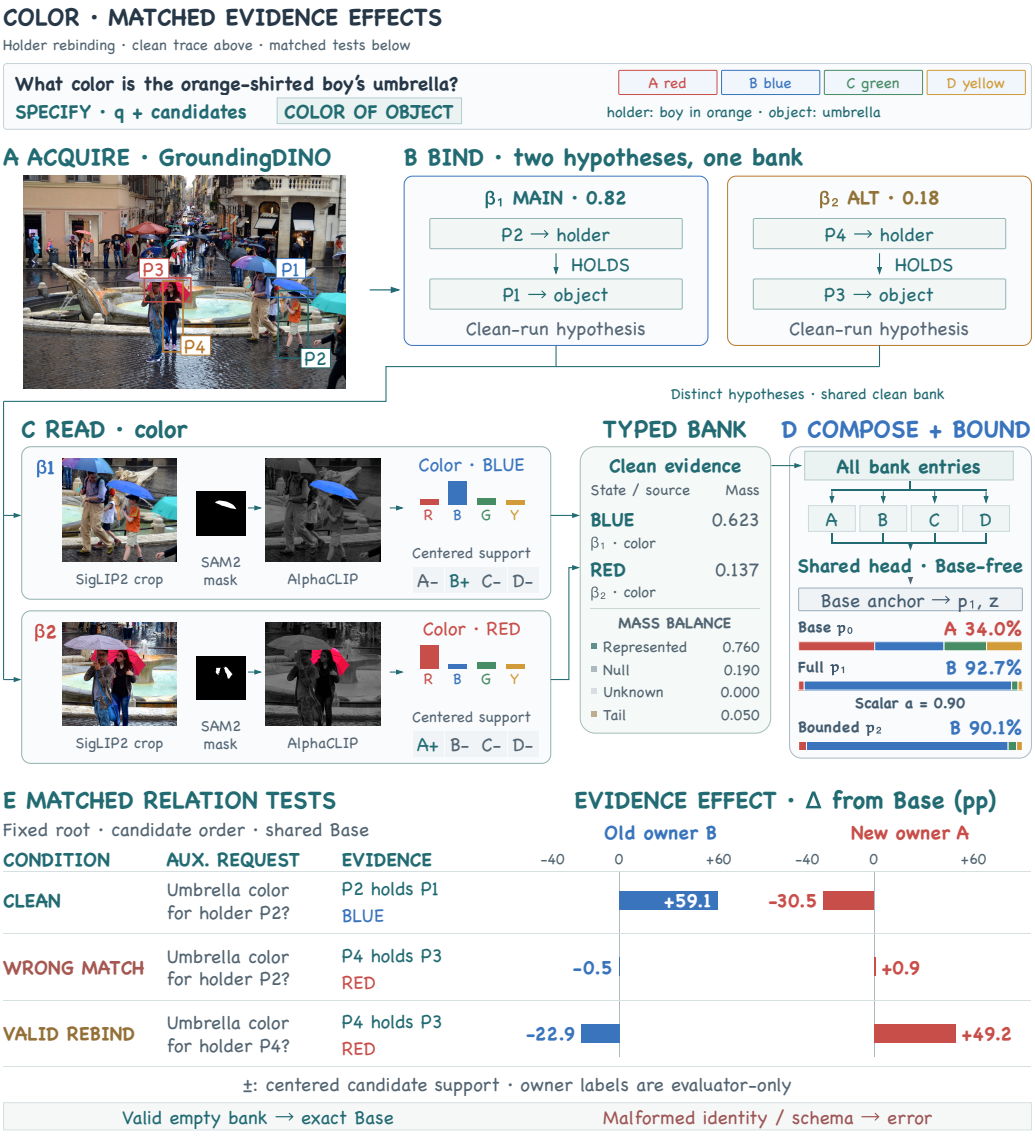


Figure 8: **Color evidence under matched relations.** Schematic numerical traces: changing the requested holder from P2 to P4 makes the red umbrella support A.

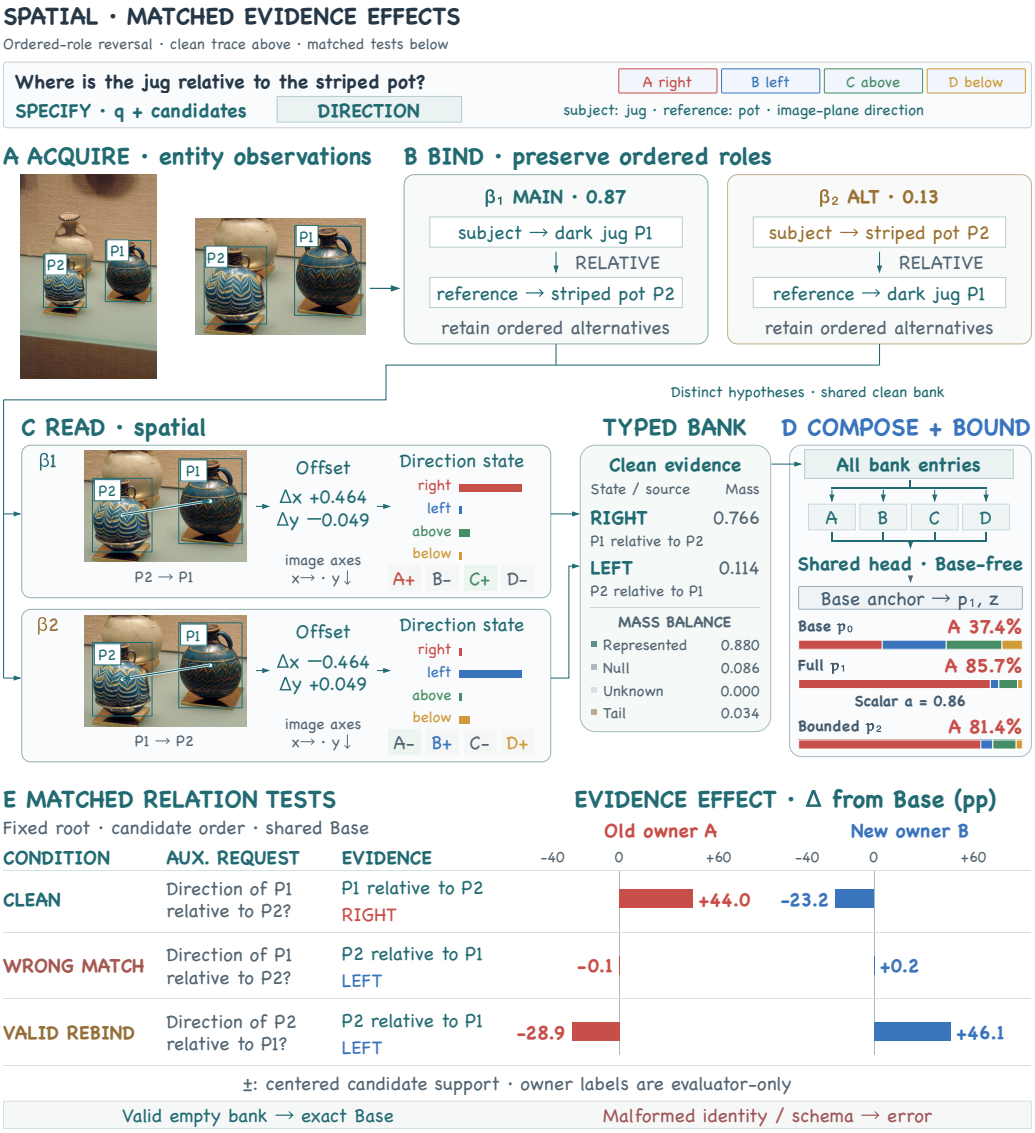


Figure 9: **Spatial evidence under role reversal.** Request and ordered roles reverse together; the physical observations and candidate order remain unchanged. Schematic numerical traces.

Ordered spatial roles. Figure 9 keeps the same two objects and image. Clean asks for the dark jug P1 relative to the striped pot P2; rebind asks for P2 relative to P1. Reversing both the request and the ordered binding changes the read state from right to left, aligning it with B rather than A. Reversing the evidence roles alone leaves them mismatched to the original request.

The horizontal offset changes sign under role reversal even though the boxes and their separation are unchanged. An unordered pair would retain both objects but lose this distinction. The ordered binding carries it into the Reader, whose direction estimate can then be aligned with the fixed right/left candidate vocabulary. This isolates a relational change from a change in image content or object recognition. The wrong-match row supplies the complementary contrast: the reversed observation is visually meaningful, yet answers the opposite directed question.

The upper alternative binding and lower rebind trial use the same reversed order for different purposes: the former represents uncertainty within the clean request, whereas the latter evaluates a changed request with matching roles.

E.2 TYPED EVIDENCE ACROSS VISUAL TASKS

Entity identity and region scope. Figure 10 distinguishes the laptop at R1 from the desktop at R2 using their device context. The hinge and keyboard support the laptop state; the rear screen and stand support the desktop state. Both observations are informative, but they answer different region-specific requests.

The wrong-match condition supplies R2 while retaining the R1 request. Valid rebinding changes the requested region as well, making A the supported candidate. The scene contains two distinct computer instances; the requested region selects the relevant category estimate.

A screen alone is compatible with either category. The hinge, keyboard, and stand provide device context for the category estimate; R1 and R2 preserve which device supplied it. The two distinctions remain separate in the support bank. A confident desktop estimate at R2 answers the rebound request about that device, while the clean request about R1 requires the laptop observation. Rebinding changes which category reading answers the auxiliary request while preserving the image and the device context used for each estimate.

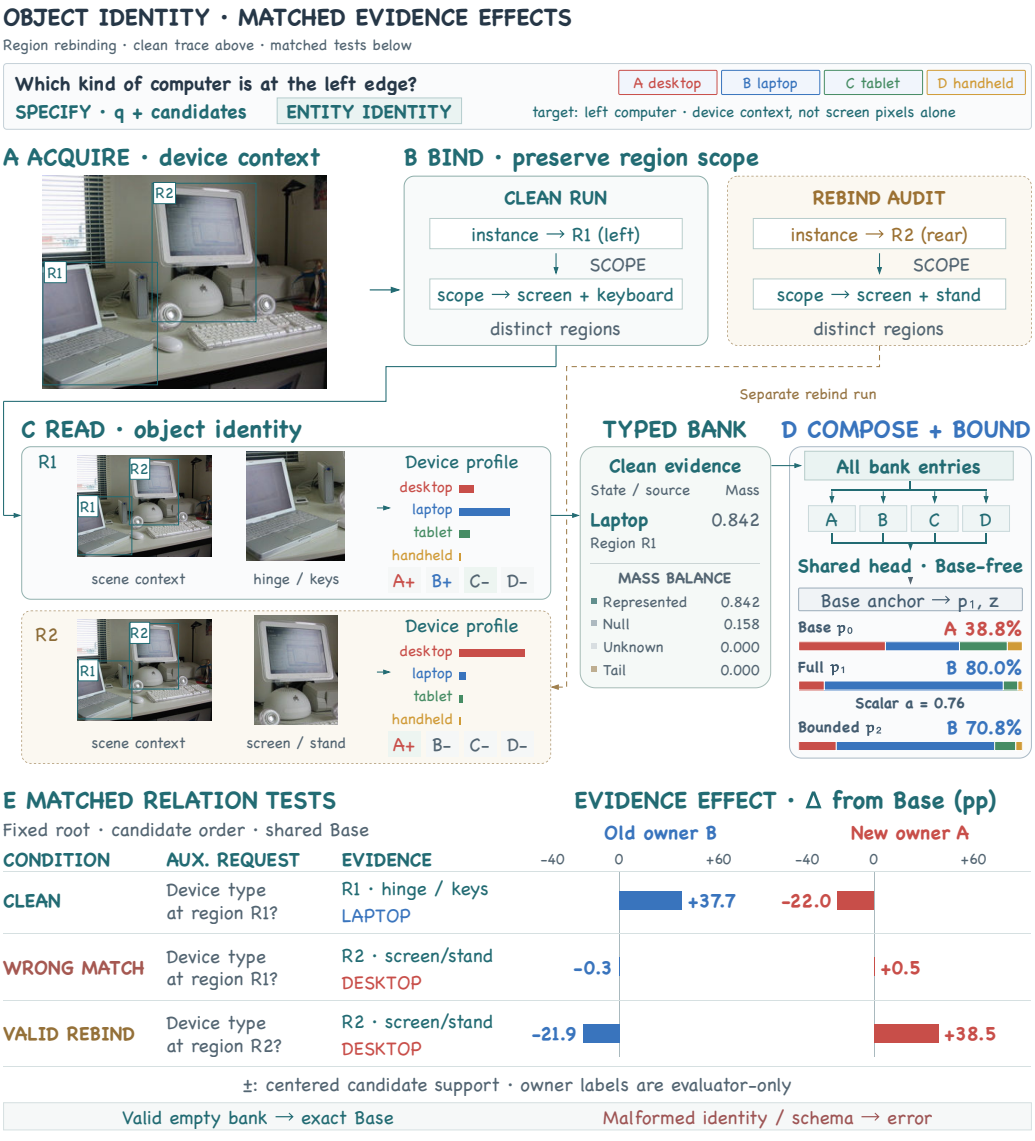


Figure 10: **Entity identity under region rebinding.** Device context distinguishes R1 from R2; the auxiliary request determines which region is relevant. Schematic numerical traces.

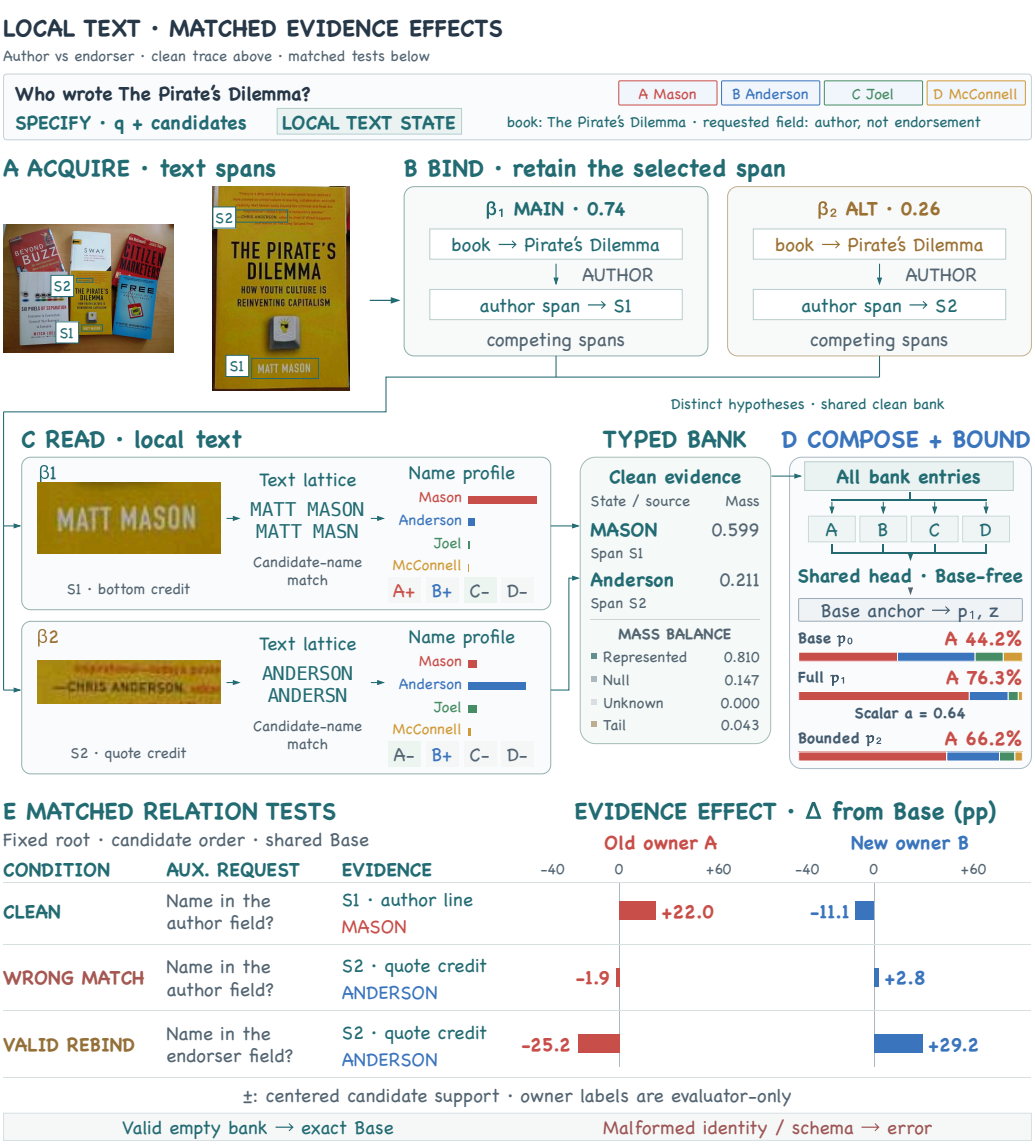


Figure 11: **Local text under field rebinding.** The author and endorsement credits contain readable names but license different answers to their respective requests. Schematic numerical traces.

Local text and field identity. In Figure 11, the same book contains an author credit, Matt Mason, and an endorsement credit, Chris Anderson. The Reader retains a recognition lattice for each localized span and aligns the resulting name estimates with the candidate vocabulary. Both strings can be read correctly while only one answers the author request.

Changing the evidence span from S1 to S2 without changing that request creates a wrong match. Asking for the endorser makes S2 relevant and redirects the illustrated support toward B. This separates uncertainty about the recognized characters from uncertainty about which semantic field a span fills.

The recognition lattice contains alternative transcriptions of a localized name. Candidate alignment combines these readings according to the name they express, while the span identity records where the text was found. Neither operation replaces the other: spelling variants can share a candidate meaning, whereas two clearly recognized names still occupy different fields on the same cover. The bank preserves both distinctions for composition.

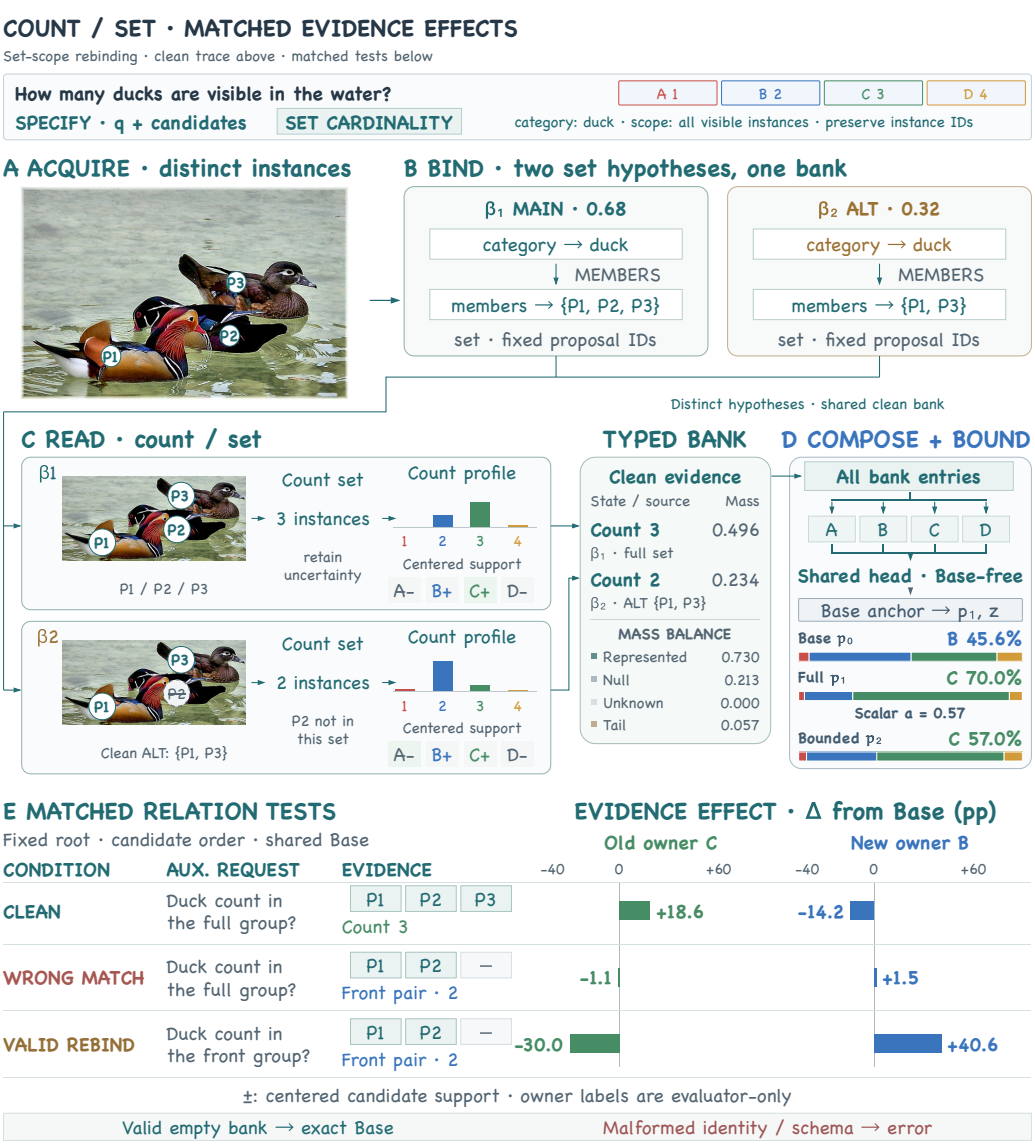


Figure 12: **Counting under a change of set scope.** The request changes from the full group to the front pair. The alternative clean set $\{\mathrm{P1}, \mathrm{P3}\}$ and rebound set $\{\mathrm{P1}, \mathrm{P2}\}$ have equal counts but different membership. Schematic numerical traces.

Set membership and counting scope. Figure 12 keeps proposal identities P1–P3 fixed. The full-group request concerns all three ducks, whereas the rebound request concerns the front group, P1 and P2. The latter two-member observation is incomplete for the original request but relevant to the new scope.

The upper clean trace also retains an alternative set, $\{\mathrm{P1}, \mathrm{P3}\}$, alongside the full set. That alternative and the lower rebind set $\{\mathrm{P1}, \mathrm{P2}\}$ both have cardinality two, but different members. Preserving set identity keeps these equal-cardinality observations distinguishable for composition; preserving both clean hypotheses carries membership uncertainty into the shared Composer.

This case therefore separates detecting instances from deciding which instances belong to the query. A two-member set can be counted exactly and still omit a relevant duck from the full-group request. Under the front-group request, exclusion of the third duck is instead appropriate. The change occurs in query-defined membership, with the image, detected instances, and numeric candidate vocabulary held fixed.

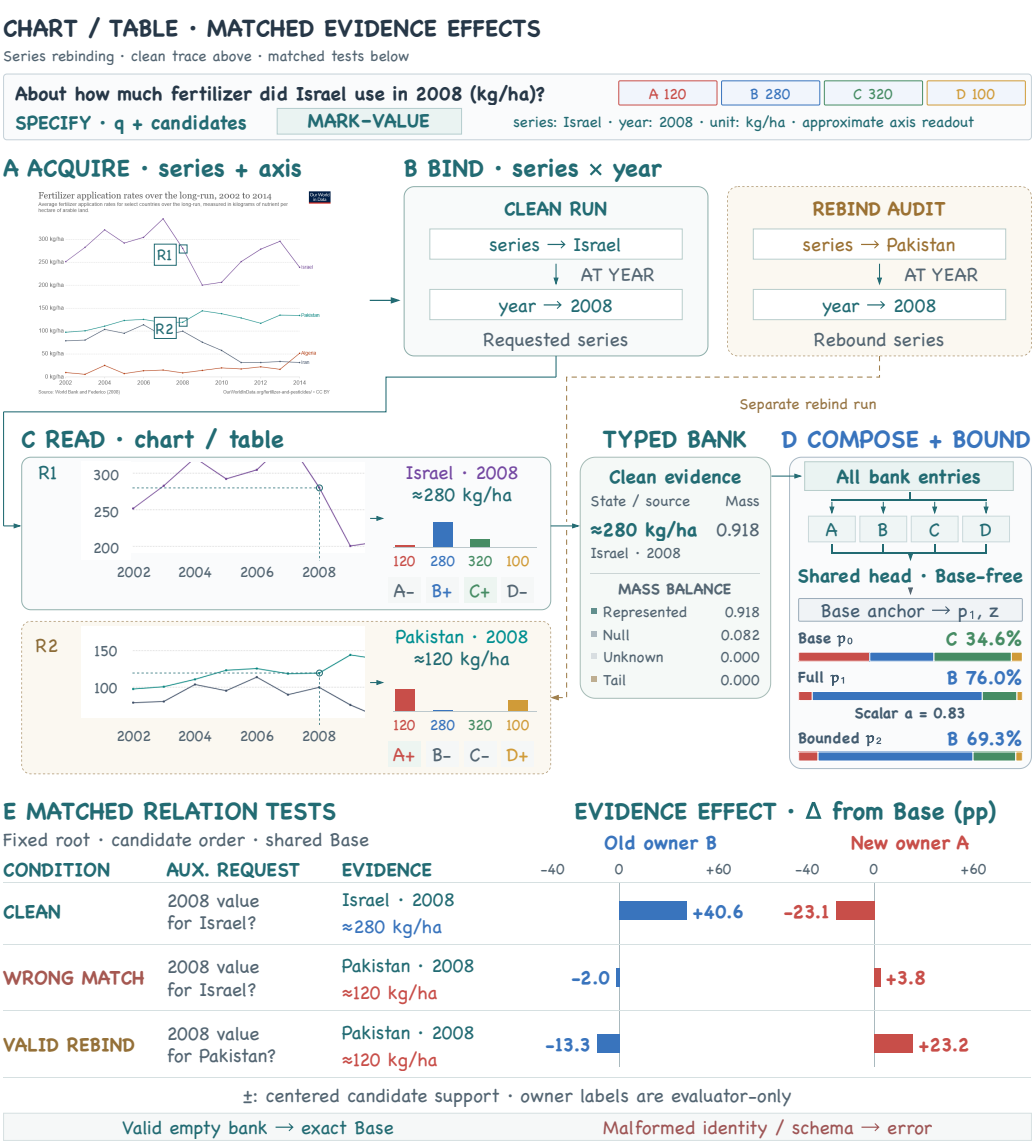


Figure 13: **Chart evidence under series rebinding.** Changing Israel to Pakistan changes the relevant value while preserving the year, axis, and unit. Schematic numerical traces.

Chart values and series identity. Figure 13 holds the year, axis, and unit fixed while changing the requested series. The illustrated readouts are approximately 280 kg/ha for Israel and 120 kg/ha for Pakistan in 2008. Reading the latter value accurately does not make it an answer to the former request.

The wrong-match and rebind conditions use the same Pakistan observation. Their difference is whether the auxiliary request still names Israel or has changed to Pakistan. The Reader's mark-to-series and mark-to-axis associations retain which observation produced each value. Candidate alignment then maps that value to the fixed numeric options. The series identity distinguishes the two readings even when they share the same year, axis, and unit.

The shared axis makes this a series-selection comparison rather than a unit conversion or scale change. Both values are plausible readings in the same plot; the legend and requested year identify which mark is relevant. Retaining those associations lets the bank carry a numeric state together with its chart context, instead of treating every readable plotted value as interchangeable evidence.

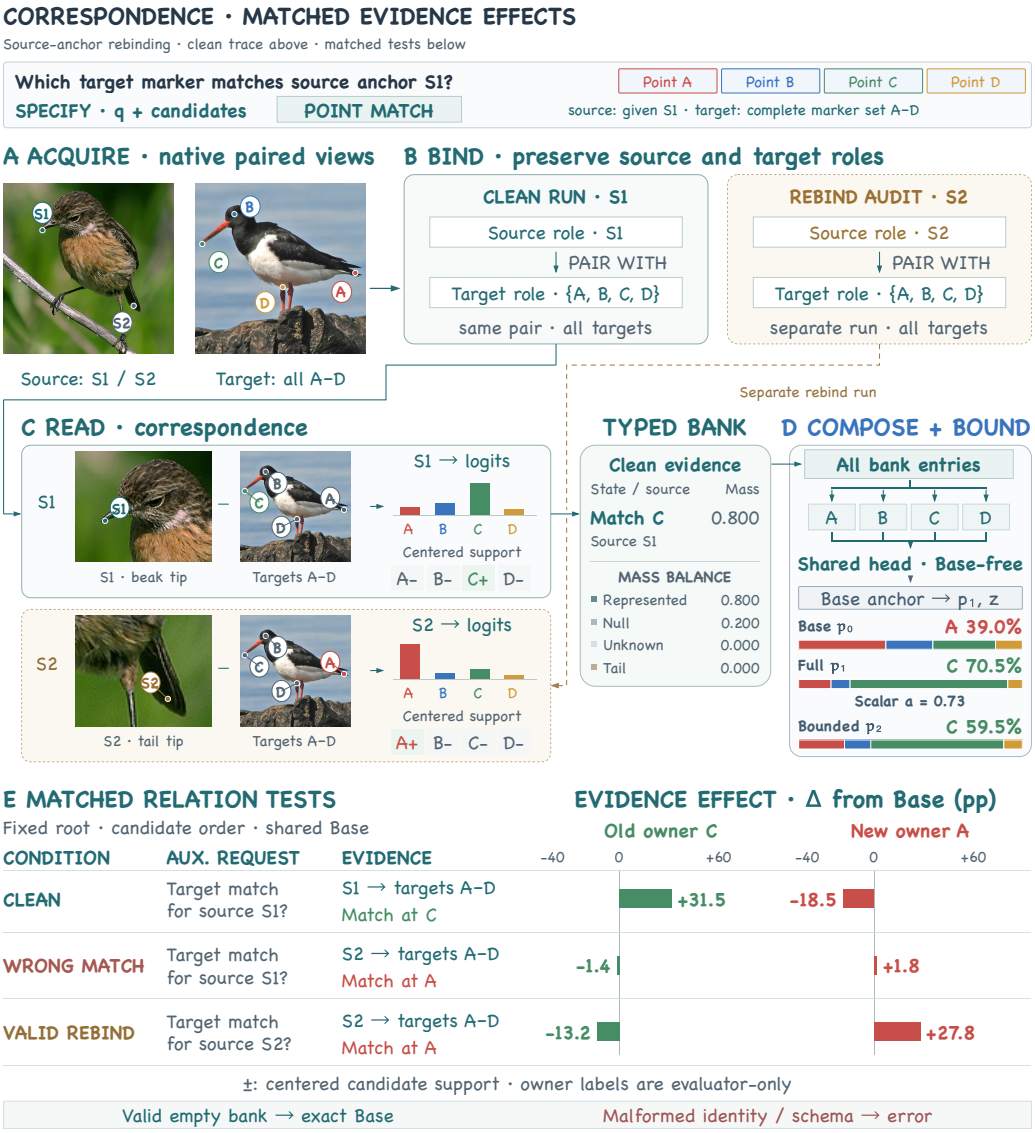


Figure 14: **Correspondence under source-anchor rebinding.** S1 and S2 query different semantic parts against the same target-marker set. Schematic numerical traces.

Cross-view correspondence and source anchors. Figure 14 fixes the image pair and the complete target-marker set. Source anchor S1 marks the beak, corresponding to target C; source anchor S2 marks the tail, corresponding to target A. Here correspondence follows semantic parts across different birds. It identifies a corresponding part, not the same physical bird appearing in both images.

The wrong-match condition uses S2 evidence for an S1 request. Valid rebinding changes the requested source anchor to S2, so that the same correspondence becomes relevant. No target marker is inserted or reordered. The separate dashed branch shows the changed source relation; the upper trace continues to show the clean execution.

Recognizing a beak in the source image is only part of the task: the answer identifies its corresponding location in the target image. The source anchor and target-marker identity therefore play different roles in the binding. Keeping every target marker fixed makes the effect shift interpretable as a response to the queried source part, rather than to a changed answer vocabulary or target layout.

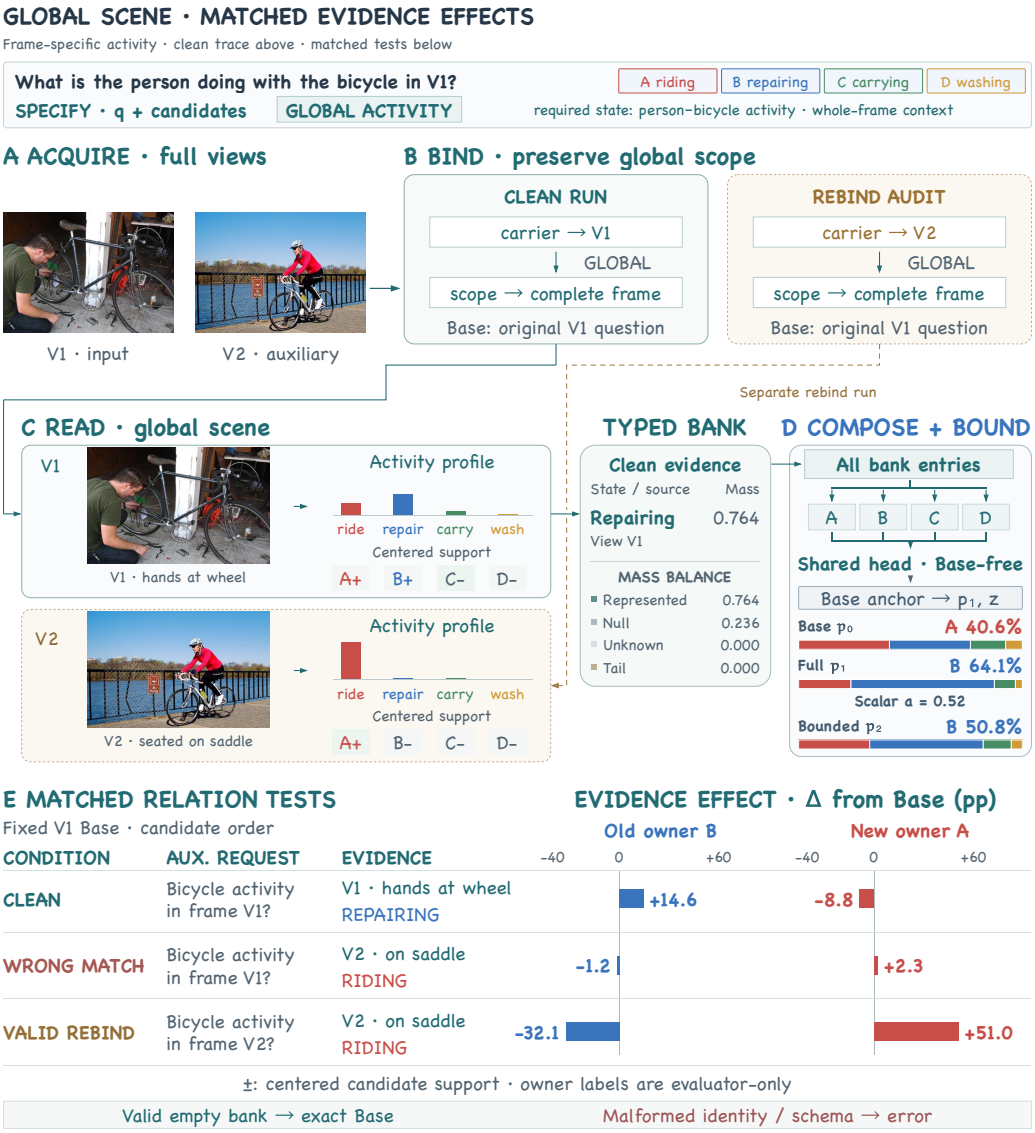


Figure 15: **Global activity under source-frame rebinding.** The auxiliary request switches from V1 to V2; the original input and Base remain references. Schematic numerical traces.

Global activity and source frames. Figure 15 contrasts a person repairing a bicycle in V1 with a person riding in V2. A whole-frame Reader estimates the activity using person–bicycle context. V1 remains the ordinary input and Base reference, while the auxiliary request explicitly identifies which frame supplies the relevant activity.

V2 is a wrong match when that request still concerns V1. It becomes relevant when the request changes to V2, redirecting support from repairing to riding. Source identity determines which activity state answers the auxiliary request. The carrier covers a whole frame, so the binding identifies the source scene while the Reader estimates the person–bicycle activity within it. The bank retains both the source-frame identity and its activity estimate for composition.

Both frames contain a person and a bicycle, so object presence alone does not resolve the activity. Their configuration supplies the relevant visual state, while the frame identity selects which configuration the request concerns. This is the gallery's global counterpart to the localized examples: the evidence carrier is a whole frame, but its contribution still depends on an explicit relation between that frame and the request.

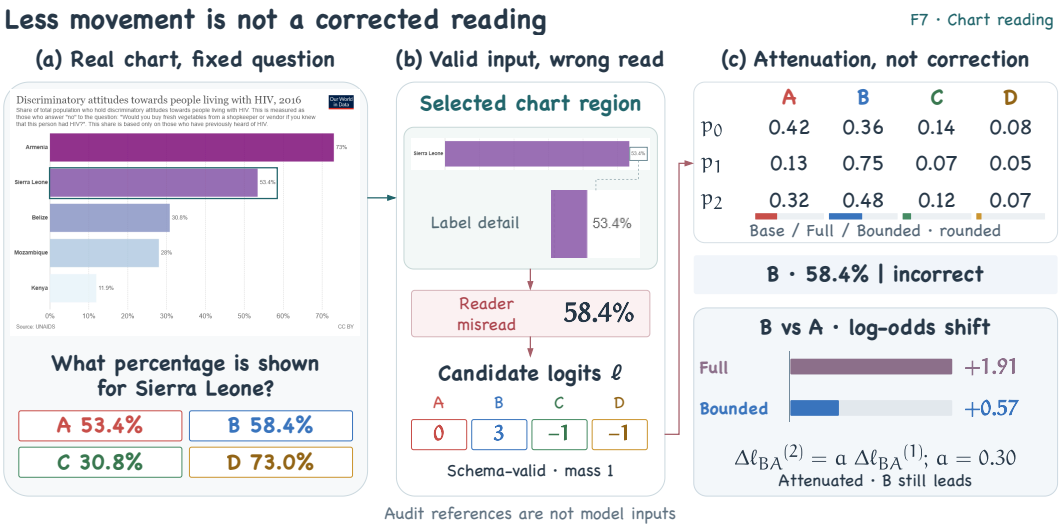


Figure 16: **Illustrative chart-reading error.** Attenuation leaves incorrect candidate B ahead.

E.3 A VISUAL-STATE ERROR SURVIVES SCALAR ATTENUATION

In Figure 16, the Sierra Leone row shows 53.4%, but the Reader is illustrated as reading 58.4%. Binding, schema, and entity–label association remain legal; projection maps the mistaken state to B. The error reaches Compose through a valid interface.

The Base-relative B-versus-A log-odds displacement obeys

$$\Delta\ell_{BA}^{(t)} = \log\frac{p_{t,B}}{p_{t,A}} - \log\frac{p_{0,B}}{p_{0,A}}, \qquad \Delta\ell_{BA}^{(2)} = a\Delta\ell_{BA}^{(1)}.$$

Strength $a = 0.30$ reduces this displacement from $+1.91$ to $+0.57$, yet B still leads in $\mathbf{p}_2$. Correcting the state requires revisiting the visual estimate (App. C.3).

Because Base favors A, sufficiently small strength can restore A's ranking. The pairwise tie occurs at

$$a^\star = \frac{\log(p_{0,A}/p_{0,B})}{\Delta\ell_{BA}^{(1)}} \approx 0.081,$$

using the displayed rounded posteriors. Thus $a = 0.30$ attenuates the error but remains above the A–B crossover. Below $a^\star$, A leads B by retreat toward Base, not by recovery of the chart value. This distinction matters when reading a corrected final answer: scalar control can reverse this pair's ranking while the underlying visual-state estimate remains wrong.

The successful chart illustration changes which series is requested; this case instead keeps the requested relation legal and changes the read value. Together they separate candidate-association errors from visual-state errors. Preserving identity makes that distinction explicit, while scalar attenuation alone does not recover the missing visual fact.

The state-only Oracle intervention targets the visual estimate while retaining its binding and downstream interface. It therefore addresses a different error source from varying scalar strength. Read together, the gallery and this boundary case distinguish three outcomes: using the correct observation for the request, estimating its state correctly, and converting that state into an appropriate candidate effect.

Scope. CROSS-Bench evaluates 2–5-candidate decisions on static images with annotatable evidence relations and one-step endpoints. RIVET implements these relations through registered finite state algebras. Ownership tests positive contributions, not signed contradictions or hidden-state causal mediation. Open-ended generation, temporal evidence, and new operations require extending the corresponding relation and evaluation.