SOURCE REFERENCES FAIL TO CONSTRAIN INFERENCE: DIAGNOSING SOURCE MISBINDING IN MULTI-IMAGE CONTEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-image question answering, even when the question explicitly specifies the target image, the model may still answer based on other images in the context. We define this failure as **source misbinding**: the model answers correctly from the target image when it is presented alone but produces an answer consistent with a distractor source when other images are introduced. This discrepancy leads us to hypothesize that the model may fail to translate the specified source into a constraint on evidence use during inference. To test this hypothesis, we construct D-SrcBench and T-SrcBench and conduct source readout and intervention experiments. The results support the explanation that source information remains decodable within the model, yet it does not reliably constrain the evidence used for answering. Based on this diagnosis, we propose **Source-First Inference**: with the backbone frozen, the model first predicts the relevant image source and then downweights or masks attention to distractor-image tokens at sensitive layers. On the fixed confirmatory test set, internal hard masking (Hard) improves LLaVA-1.5-7B accuracy from 30.4% to 59.0%, outperforming soft biasing (Soft) with the same source reader (51.2%) and the best-performing evaluated baseline, our reproduction of MIA-DPO (36.0%).

1 INTRODUCTION

Multimodal large language models (MLLMs) are increasingly used to answer questions in multi-image contexts (Meng et al., 2025; Wang et al., 2025a; Xu et al., 2025). When a prompt explicitly refers to a particular image, the model must ground its inference in that source. However, we observe that even when the model answers correctly with the referenced image alone, adding distractor images can cause it to produce an answer supported by another image. We call this failure **source misbinding**.

We define source misbinding and the evaluation subset as follows: consider an input $x = (\mathbf{I}, q_t)$, where $\mathbf{I} = (I_1, \ldots, I_K)$ contains multiple images, and q_t explicitly refers to the target image I_t, with $t \in \{1, \ldots, K\}$. Let $\mathcal{A}_i$ denote the set of accepted answers to the question for image I_i under a fixed scoring rule. For the model's answer $\hat{a} = f_\theta(x)$, we define source misbinding $\mathrm{SM}_\theta(x)$ as:

$$\mathrm{SM}_\theta(x) = \mathbf{1}\left[\hat{a} \notin \mathcal{A}_t \ \wedge \ \sum_{j \neq t} \mathbf{1}[\hat{a} \in \mathcal{A}_j] = 1\right], \quad j \in \{1, \ldots, K\}. \tag{1}$$

The equation indicates that, during multi-image inference, the model's answer does not belong to the target image's accepted answer set $\mathcal{A}_t$ but matches an accepted answer for exactly one distractor image. It characterizes the observable correspondence between answers and image sources. Conceptually, source misbinding can occur both when the model answers correctly with only the target image and when the target-only answer is already incorrect and shifts to a distractor-consistent answer after distractor images are added. However, an incorrect target-only answer alone is insufficient to clearly determine its evidence source.

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.

To control for answering ability when the target image is provided alone, we define the capability-controlled subset $\mathcal{D}^{\theta}_{\text{cap}}$:

$$\mathcal{D}^{\theta}_{\text{cap}} = \left\{ x \in \mathcal{D} \;\middle|\; \hat{a}^{\text{solo}} \in \mathcal{A}_t \right\}. \tag{2}$$

Here, $\mathcal{D}$ denotes the full set of evaluated multi-image instances, and $\hat{a}^{\text{solo}}$ denotes the answer obtained with only the target image I_t. This subset contains only instances that the model can answer correctly using the target image alone. Target-only correctness is therefore a control for diagnosis, not a requirement of the behavioral definition.



Figure 1: **Analysis of source misbinding examples.** (a,b) Source misbinding question–answer (QA) examples. (c) Non-target image attention preference and image-origin decoding performance. (d) Target-image attention and mean distractor-image attention in a four-image probe.

The motivating case in Figure 1(a) uses an example from MIA-DPO (Liu et al., 2025) to illustrate $\text{SM}_\theta(x)$. When asked about the color of the bench in Image 2, the model answers "white" with only the target image but "green" after the distractor image is added. We reproduce this behavior using a minimal two-image color probe: despite the unchanged reference to Image 1 in the question, adding the distractor image changes the answer from "red" to "green" (Figure 1(b)). The heatmaps in Figure 1 show attention from the answer token "green" to visual patches. These attention weights are extracted from a forward pass with the multi-image prompt and generated answer provided together, and averaged over attention heads at decoder block 16 (zero-indexed). In the color probe, summing these attention weights over each image's visual tokens yields 4.45% for the target image and 8.88% for the distractor image. Thus, although the prompt explicitly refers to Image 1, the answer token assigns nearly twice as much attention to the competing image as to the target image. These cases lead us to further consider whether image-origin information remains recoverable when the answer matches a distractor image.

This question is closely related to multimodal reliability and the organization of information flow. SHIFT (Wang et al., 2025b) smooths layer-wise feature changes to mitigate hallucinations, DAB (Li et al., 2025) balances attention across images, and DTS (Lee et al., 2026) limits unwanted cross-image information leakage. However, joint inference requires integrating multiple sources, whereas an explicit image reference in the prompt already constrains which source should support the answer. MIA-DPO (Liu et al., 2025) already incorporates such references into attention-based preference selection. These methods organize visual information, but they have not established whether an explicit source reference can control evidence use in the presence of distractor images. We therefore examine whether source identity information remains recoverable from the model's internal representations and whether enforcing the use of the specified source restores correct answering.

These analyses reveal a discrepancy between attention preference and source recoverability. Among the 11 MIA-DPO cases, non-target images receive the most question-token attention in 10 cases at layer 16 (Figure 1(c), left). Meanwhile, image-origin decoding accuracy reaches 75% on the two-image subset, above the 50% chance level (Figure 1(c), right). In a separate four-image color probe, the mean answer-token attention received by non-target images exceeds that received by the target image at both layer 16 (3.23% vs. 1.79%) and layer 31 (4.42% vs. 4.33%) (Figure 1(d)). Thus,

even when query- or answer-token attention favors competing images, source information for the target image remains recoverable. Can the recovered source information still route evidence to the specified target image?

To distinguish source recoverability from effective source guidance, we keep the images and question unchanged and block text-to-visual attention to non-target images at selected decoder layers. We find that enforcing the known target source repairs source misbinding, whereas controls that preserve access to a wrong source or random visual tokens do not achieve comparable repair. On a repairable subset, exchanging intermediate attention outputs between baseline runs and runs using the correct source transfers correction and misbinding in opposite directions. Together, these interventions indicate that source information remains available but is not translated into an effective constraint during answering. This diagnosis motivates Hard, which predicts the source specified by the question and restricts direct text-to-visual access to that source at repair-sensitive layers during prompt processing.

Additionally, we construct a benchmark with explicit image references and per-image answer annotations. Paired target-only evaluations identify the source-valid subset, while the full test set supports overall evaluation. This design enables us to distinguish uniquely distractor-matched answers from other errors and measure their repair.

Our contributions are threefold:

(1) We formalize **source misbinding** as a source-binding failure in multi-image question answering and experimentally identify a failure of information-flow constraints underlying this behavior. (2) We propose **Source-First Inference**, which separates source prediction from evidence constraints. With the backbone frozen, a lightweight source reader first predicts the image specified by the question. This prediction is then explicitly translated into constraints on evidence use during inference at layers sensitive to repair interventions. (3) We introduce D-SrcBench and T-SrcBench, two strict source-resolved benchmarks with explicit image references, per-image answer annotations, and paired target-only evaluations. Under a unified evaluation protocol, we systematically evaluate source misbinding, error repair, and preservation of previously correct answers, and show that our method achieves higher accuracy than the evaluated existing baselines.

2 Problem Formulation and Related Work

Multi-Image Hallucination. General multi-image benchmarks broaden evaluation beyond single-image question answering. MMIU covers seven types of inter-image relationships and 52 tasks, while MuirBench evaluates 12 multi-image tasks across 10 relation categories using paired answerable and unanswerable instances (Meng et al., 2025; Wang et al., 2025a). MC-Bench focuses on open-ended multi-context visual grounding, where text prompts identify target instances across image pairs (Xu et al., 2025). MIA-DPO adds unrelated images to single-image questions while retaining explicit target references, and uses attention patterns to construct preference pairs for multi-image alignment (Liu et al., 2025). MIHBench evaluates object existence, count agreement, and identity consistency across images, revealing the effects of image count, distractor position, and single-image hallucination tendencies (Li et al., 2025). MIOH further evaluates object existence, counting, attributes, and position under comprehensive, comparative, and selective inference, showing that hallucinations can arise during cross-image information integration (Min et al., 2026). These works study broad multi-image competence, model hallucinations, or visual grounding; we instead isolate source misbinding as a source-binding failure in which the answer uniquely matches a distractor image under an explicit image reference, and evaluate it using source-resolved data with target-only evaluation controls.

Multimodal Information Routing. Several works modify how visual information is allocated during inference. VAR identifies visual attention sinks and redistributes their surplus attention toward informative visual tokens (Kang et al., 2025). At the image level, SoFA interpolates between causal and bidirectional inter-image attention masks to reduce position bias (Tian et al., 2025); DAB balances attention across images while preserving the total visual-attention budget (Li et al., 2025); and DTS scales delimiter-token hidden states to strengthen intra-image interactions and limit cross-image information leakage (Lee et al., 2026). FOCUS directly addresses cross-image information leakage through image-wise masking and contrastive logit aggregation (Park et al., 2025). Focus-Centric Visual Chains organize multi-image inference through structured image-wise context (Zhang et al., 2025). These methods improve inter-image balance, positional robustness, or context struc-

ture; our work tests whether explicit source references are actually honored and enforces constraints by applying Hard at repair-sensitive layers.

3 DIAGNOSING SOURCE MISBINDING

Section 1 reveals that although image-origin information remains recoverable, attention may still favor competing images. We therefore conduct a controlled mechanistic diagnosis to investigate how available source information becomes ineffective during answering. We use LLaVA-1.5-7B (Liu et al., 2024) as our subject of study. Synthetic color probes provide single-variable tests, while D/T analyses examine whether the same diagnosis applies to real images. Full benchmark construction and evaluation are presented in Section 5.

3.1 CAPABILITY-CONTROLLED FAILURES

We first use a single-color probe, where Image 1 contains a red circle and Image 2 contains a circle of another color (green, orange, or pink). Although LLaVA-1.5-7B answers all 150 target-only questions correctly in each condition, its multi-image answers match the distractor in 105, 147, and 145 cases, respectively. Thus, the errors vary with the competing visual content rather than arising from a fixed preference for a particular color word.

Next, we conduct experiments on 707 questions from 295 dual- and triple-image groups, ensuring that the model can answer each question correctly with only the target image. During multi-image testing, 266 answers remain correct and 441 become incorrect; among these incorrect answers, 365 match an answer uniquely associated with a distractor image, while the remaining 76 do not. In the experimental subset, we analyze the token-count-normalized pre-softmax attention margin between the target and competing sources. This margin is positive across layers 0–31 in non-misbound cases and negative in misbound cases. Within layers 16–23, at least 28/32 D attention heads and at least 29/32 T attention heads favor the competing source in misbound cases. These results indicate that most instances of source misbinding are associated with a specific competing source; however, whether the specified source is unavailable or simply not effectively used remains to be determined.

3.2 SOURCE INFORMATION REMAINS RECOVERABLE

To determine whether source misbinding reflects a loss of source information, we test two necessary conditions: visual tokens must retain information about the image they belong to (*source identity*), and the hidden state must retain the image index specified in the question (*reference index*). We evaluate both conditions on 129 target-only-correct D/T instances, with 55 instances in the discovery split and 74 in the held-out split, with no overlap in image families between the two splits.

For source identity, we construct a cosine nearest-centroid classifier using normalized visual-token representations from the discovery split. On held-out misbound cases, source-index accuracy at layer 16 reaches 76.4% for 26 dual-image instances and 50.8% for 27 triple-image instances, exceeding chance levels of 50.0% and 33.3%, respectively. Thus, visual tokens still encode image-origin information when the model answers based on a distractor image. Because the classifier may exploit positional cues, this result establishes source-identity decodability rather than the model's ability to select the correct source (Elazar et al., 2021).

We next test whether reference-index information remains available beyond the token that explicitly names the index. We use a linear probe to predict the specified image from the hidden state at the final prompt position. For each instance, we pair the original question with a variant that changes only the referenced image. At layer 16, the probe achieves 100% accuracy on the 68 dual-image variants and 100% accuracy on the triple-image variants. These results exceed the corresponding majority-class baseline of 50% in both settings; the uniform-chance levels are 50% and 33.3%, respectively. Thus, the disappearance of either source identity or the specified index cannot explain misbinding. The failure occurs when using this information to select evidence. We next test whether enforcing the specified source changes the model's answer and at which layers this constraint has an effect.

3.3 Attention Bias vs. Repair Sensitivity

To this end, we use one red target and three green distractors, placing the target first or second, and measure attention from teacher-forced answer tokens to each visual source on 16 instances that are answered correctly with the target image alone but exhibit distractor-consistent errors under multi-image input. Figure 2(a, top) summarizes the combined share of visual attention assigned to the three distractors. After averaging over attention heads and answer positions, we find that distractor attention reaches 92.2% at layers 7–14, 85.7% at layers 16–23, and 85.1% at layers 24–31. The earliest band shows the strongest distractor preference, but can intervening there repair the answer? Attention magnitude alone may not establish causal importance (Serrano & Smith, 2019). We apply the same source constraint at these depths on 150 distractor-consistent errors: during prompt processing, text-to-visual attention to non-target visual tokens is set to zero, and the remaining weights are renormalized, after which decoding continues without further edits. Error-recovery rates are 0%, 100%, and 0% for layers 7–14, 16–23, and 24–31, respectively. The control with eight scattered layers repairs 0/150 cases, whereas constraining all 32 layers repairs 150/150 cases (Figure 2(b)). The strongest attention bias therefore does not identify the effective intervention band. Instead, layers 16–23 form a repair-sensitive window (our eight-layer window lies within a complete-repair plateau spanning four window positions and is insensitive to window shifts; Appendix C.1) in which enforcing the known source restores answering; this result does not establish the unique or earliest origin of the error. Does recovery depend on restoring access to the requested source, or could generic suppression of visual input produce the same effect?

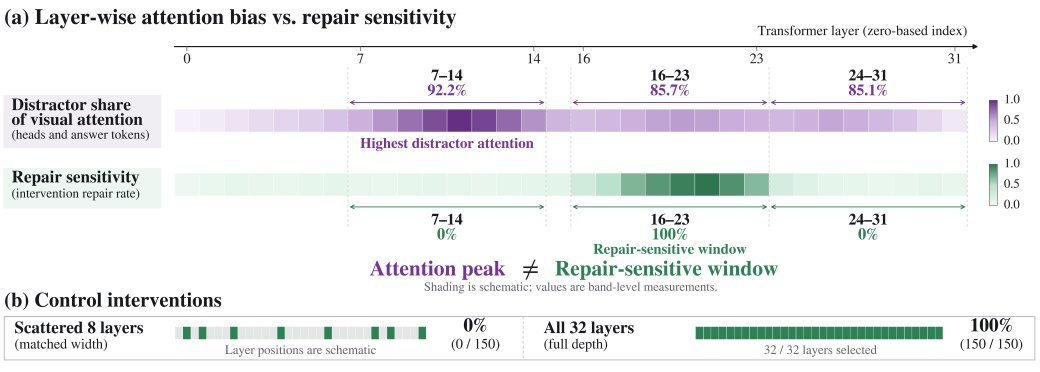


Figure 2: **Layer-wise dissociation between attention bias and repair sensitivity.** (a) Distractor share of visual attention on 16 instances and intervention error-recovery rates on 150 distractor-consistent errors. (b) Scattered eight-layer and full-depth controls on the same 150 instances. Layer indices are zero-based.

3.4 Source-specific Causal Controls

To distinguish source-specific recovery from generic suppression, we keep the number of accessible visual tokens fixed while varying their source. We select one question from each selected D/T image group, balancing across D/T, entity/color questions, and initially correct or misbound answers. At layers 16–23, text queries retain access only to the target image, a wrong image, or randomly selected visual tokens. Random retention includes target-image tokens and matches token count, but not total attention mass. Among initially misbound cases, retaining the target source yields a repair rate of 93.75%, retaining a wrong source yields 0%, and random retention yields only 6.25%. Among initially correct cases, target-source and random retention both preserve 100% of correct answers, whereas wrong-source retention preserves only 37.50%. Moreover, 96.15% of the errors produced under wrong-source retention match the retained image's annotated answer. Recovery therefore depends on which source remains accessible, rather than merely on reducing the amount of visual input.

Finally, we apply the same source constraint to all questions in the real-image diagnostic set. Using the source specified by each question repairs 67.12% of source-misbinding cases while causing only 1.88% of initially correct answers to become incorrect. Taken together, these results show that source information remains available, but competing visual evidence can still determine the answer; source-specific access constraints and aligned state transfers (Appendix D.5) can reverse this behavior. This

motivates an inference framework that makes the specified source effective during answering; full method comparisons and confirmatory evaluation are presented in Section 5.

4 SOURCE-FIRST INFERENCE

4.1 SOURCE PREDICTION

The central idea of our method named **Source-First Inference** is to predict the source first and then apply internal hard masking (Hard) or soft biasing (Soft). Using the frozen LLaVA-1.5-7B backbone, we extract source-local features from decoder layer 16 through a single forward pass. At question-token positions, we average the keys and values over each image's visual-token span and normalize the attention logits within that span. The resulting weights aggregate the source values, which are then pooled over question tokens and attention heads. This produces a token-preserving question representation $\mathcal{Q}$ and 128-dimensional key, value, and evidence vectors for each source.

Our lightweight reader g_ϕ combines the question representation with the source representations and produces one logit per source. For N input images, the source belief and selected source are:

$$\mathbf{b} = \mathrm{softmax}\Big(g_\phi(\mathcal{Q}, \mathcal{X})\Big), \quad \hat{s} = \underset{s \in \{1,\dots,N\}}{\arg\max}\ b_s. \tag{3}$$

Here, $\mathcal{Q}$ denotes the question representation and $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ denotes the source-specific representations of the N images. The reader g_ϕ produces an unnormalized logit vector

$$g_\phi(\mathcal{Q}, \mathcal{X}) = (z_1, \dots, z_N) \in \mathbb{R}^N \tag{4}$$

4.2 HARD SOURCE ENFORCEMENT

Hard turns the selected source into an access constraint during a fresh generation pass. We apply it to every attention head and every nonvisual prefill query in the repair-sensitive window $\mathcal{L} = \{16, \dots, 23\}$. For the selected source $\hat{s}$, Hard blocks all visual keys belonging to the other input images. Let $\mathcal{V}_s$ denote the visual-key positions of source s. The blocked visual-key set is $\mathcal{D}(\hat{s}) = \bigcup_{s \in \{1,\dots,N\},\, s \neq \hat{s}} \mathcal{V}_s$, where N is the number of input images and $\mathcal{D}(\hat{s})$ contains the visual keys from all nonselected sources.

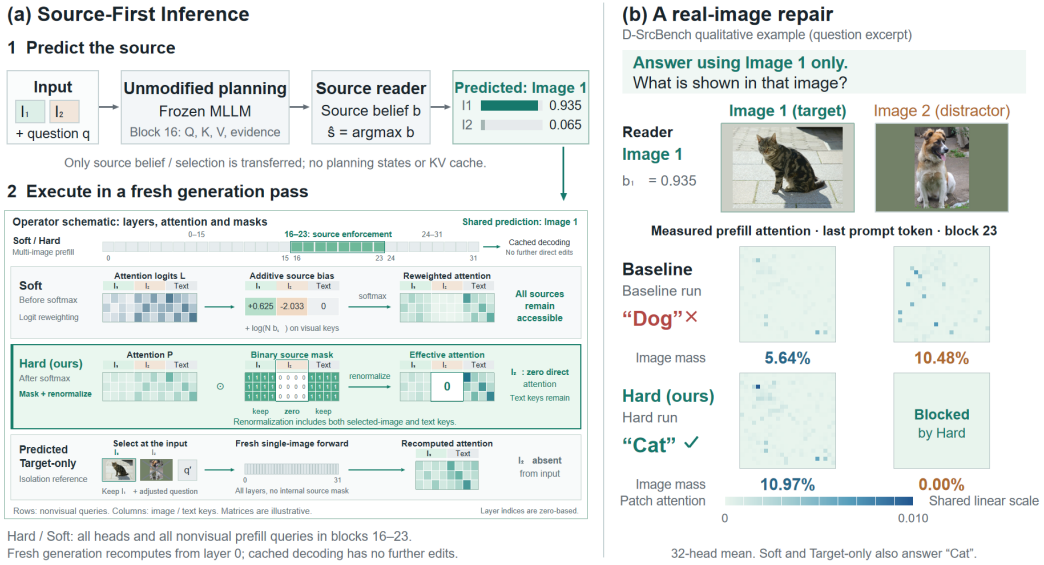


Figure 3: Source-First Inference architecture and example. (a) Source-First Inference method. (b) Source-First Inference example.

For an affected layer, head, and nonvisual query, let $\mathbf{p}$ be the ordinary attention row after causal or padding masking and softmax. Hard zeros the weights assigned to $\mathcal{D}(\hat{s})$ and renormalizes the

remaining row as $\widetilde{p}_j = p_j \mathbf{1}[j \notin \mathcal{D}(\hat{s})] / \left(\sum_{k \notin \mathcal{D}(\hat{s})} p_k\right)$. Here, p_j and $\widetilde{p}_j$ are the original and modified attention weights at key position j, respectively; j and k index all text and visual key positions; and $\mathbf{1}[\cdot]$ is the indicator function, equal to one when its condition holds and zero otherwise.

Separating source prediction from subsequent evidence constraints allows the same source reader to support different execution strategies. In addition to Hard, Soft adjusts visual attention using source confidence while retaining access to all images. As an input-isolation control, Predicted Target-only retains only the predicted source image. These settings use the same source reader, allowing us to compare the effects of different execution strategies. Figure 3 illustrates the Source-First Inference architecture and a real-image repair example; Appendix E.6 provides the detailed inference procedure, execution controls, and example analysis.

5 Benchmarks and Experiments

5.1 Existing Data and Evaluation Gaps

Existing multi-image data expose related errors but do not provide the joint controls needed for source attribution. Among 325 refined color-attribute candidates from MIA-DPO's preference data (Liu et al., 2025), 91 are answered correctly with the target image alone; 41 of these elicit the rejected attribute under multi-image input. These conditional counts do not identify a competing source, since the preference labels do not link rejected answers to non-target image indices. In our explicit left/right-reference subset of NLVR2 (Suhr et al., 2019), 722 of 1,501 examples are incorrect (48.1%), but its binary labels likewise do not identify the image supporting an error. Our benchmarks combine explicit source references, target-only capability controls, and per-image answer annotations.

5.2 D/T Benchmarks

Each question names one source in an ordered group of two or three images. Every image has a canonical answer, aliases, and a frozen single-image response. Within target-only-capable questions, a *gold-consistent* misbinding uniquely matches a non-target source's annotation; a *model-relative* misbinding uniquely matches its frozen single-image response.

D/T diagnostic set. Images are collected from Wikimedia Commons (Wikimedia Commons contributors, n.d.). D-SrcBench contains 178 two-image groups and 356 QA pairs; T-SrcBench contains 117 three-image groups and 351 pairs. Together, they contain 707 distinct images and 707 questions, with each image queried once. All questions pass the target-only capability control.

Confirmatory test set. A separate set was constructed from COCO 2017 validation images (Lin et al., 2014) with the reader and method frozen. Method-blind quality control retained 129 of 150 candidate groups: 65 two-image groups with 130 QA pairs and 64 three-image groups with 192 pairs. All 322 pairs form the *full test set*; the 261 answered correctly by the base model under Parser Target-only form the fixed *source-valid*, or capability-controlled, subset. Of the 322 questions, 39 were evaluated in an earlier pilot and 283 were first evaluated in the formal run. The capability subset is fixed across methods, not selected separately for each model.

Baselines and protocol. We compare HA-DPO (Zhao et al., 2023), POVID (Zhou et al., 2024), MIA-DPO (Liu et al., 2025), and SoFA (Tian et al., 2025) with the base model and Hard, using LLaVA-1.5-7B as the common backbone. We use author-released HA-DPO and POVID weights, our reproduction of MIA-DPO, and our implementation of SoFA's published attention rule. Hard follows Section 4; implementation and compatibility notes are given in Appendix B.6. All runs use the same images, order, questions, and scorer. Accuracy uses the full specified set; gold repair, recovery, and damage use fixed subsets defined by the base model. Confidence intervals use paired bootstrap resampling over image groups or shared-image clusters.

5.3 Comparison with Existing Methods

Table 1 compares existing mitigation methods under the same evaluation protocol. On the D/T diagnostic set, Hard reaches 65.9% accuracy, compared with 37.6% for the base model and 38.5% for HA-DPO, the strongest evaluated baseline on this set. The gains hold for both D-SrcBench (44.7% to 67.7%) and T-SrcBench (30.5% to 64.1%).

Table 1: **Comparison with existing mitigation methods.** All entries are percentages. On the D/T diagnostic set, Gold repair uses 365 gold-consistent baseline errors, Recovery uses all 441 baseline errors, and Damage uses 266 baseline-correct questions. Full and Capable denote the 322-question confirmatory test and its fixed 261-question capability-controlled subset. †: our training reproduction; ‡: our implementation of the published method.

	D/T diagnostic set						Confirmatory test	
Method	D ↑	T ↑	All ↑	Gold repair ↑	Recovery ↑	Damage ↓	Full ↑	Capable ↑
Base LLaVA-1.5-7B	44.66	30.48	37.62	0.00	0.00	0.00	30.43	34.87
HA-DPO	46.35	30.48	38.47	6.58	7.03	9.40	31.06	34.10
POVID	45.51	30.20	37.91	20.27	21.09	34.21	32.61	39.46
MIA-DPO†	27.53	33.05	30.27	41.92	41.27	87.97	36.02	41.38
SoFA‡	44.10	29.63	36.92	4.93	5.22	10.53	30.43	34.48
Hard (ours)	**67.70**	**64.10**	**65.91**	**52.33**	**51.02**	**9.40**	**59.01**	**67.05**

Figure 4(a) separates repair from damage to previously correct answers. Hard repairs 52.3% of 365 gold-consistent errors with 9.4% damage to 266 baseline-correct answers, matching HA-DPO's damage rate while repairing substantially more errors. POVID and our MIA-DPO reproduction repair 20.3% and 41.9%, but damage 34.2% and 88.0%, respectively. Their recovery gains are therefore largely offset by regressions. SoFA likewise yields little net improvement, recovering 23 baseline errors while introducing 28 new errors.

On the confirmatory test, Hard improves accuracy from 30.4% to 59.0%, a gain of 28.6 percentage points (pp) (95% confidence interval (CI): [23.3, 33.8]; Figure 4(b)). MIA-DPO is the strongest evaluated baseline on this set at 36.0%. Within the source-valid subset, Hard reaches 67.0%, compared with 41.4% for MIA-DPO. On the full test, Hard repairs 75 of 121 gold-consistent errors and damages 17 of 98 previously correct answers; across all baseline errors, it recovers 109 of 224, yielding a net gain of 92 correct answers.

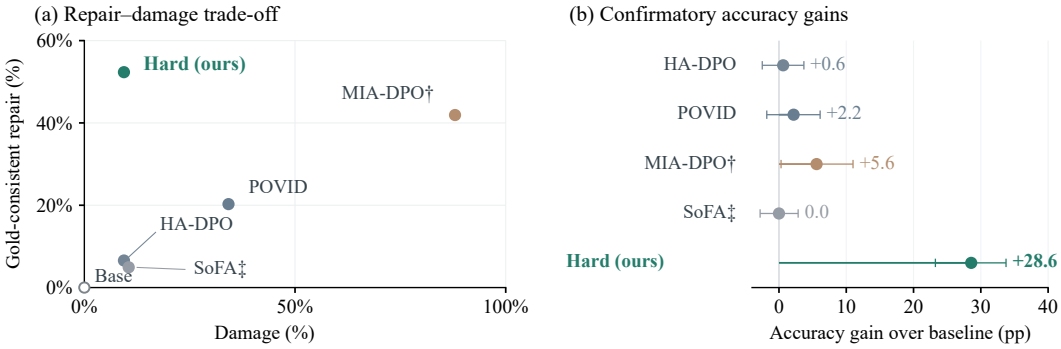


Figure 4: **Repair–damage trade-off and confirmatory gains.** (a) Gold repair versus damage on the D/T diagnostic set. (b) Accuracy gains over Base on the confirmatory set; error bars show 95% paired, group-clustered bootstrap confidence intervals. †: our MIA-DPO reproduction; ‡: our SoFA implementation.

5.4 Further Analysis

Execution under a shared reader. Soft, Hard, and Predicted Target-only reach 51.2%, 59.0%, and 64.0% on the full test with the reader fixed; Table 21 in Appendix C.2 gives the complete comparison.

Source selection and residual errors. Figure 5 shows the D/T outcomes: Hard recovers 225 of 441 baseline errors and damages 25 of 266 correct answers. Among 604 questions with a correct source prediction, 142 remain incorrect; case D illustrates one such failure. On the full test, replacing predicted sources with parsed or annotated sources raises Hard from 59.0% to 74.2% under the same executor. Selection does not explain the entire gap, however. In the source-valid subset, the reader selects correctly on 206 questions, but Hard still fails on 34 despite the target image being

individually answerable. This separates errors in choosing a source from errors that persist after the correct source has been chosen.

Inference cost. A separate timing study over 24 image groups, with a median output length of two tokens, gives cold end-to-end medians of 375.6 ms for Baseline, 702.0 ms for Soft, 694.3 ms for Hard, and 465.6 ms for Predicted Target-only on an A100 graphics processing unit (GPU). Predicted-method timings include the planning pass but exclude model loading, with no reuse of visual features or decoder key–value caches.

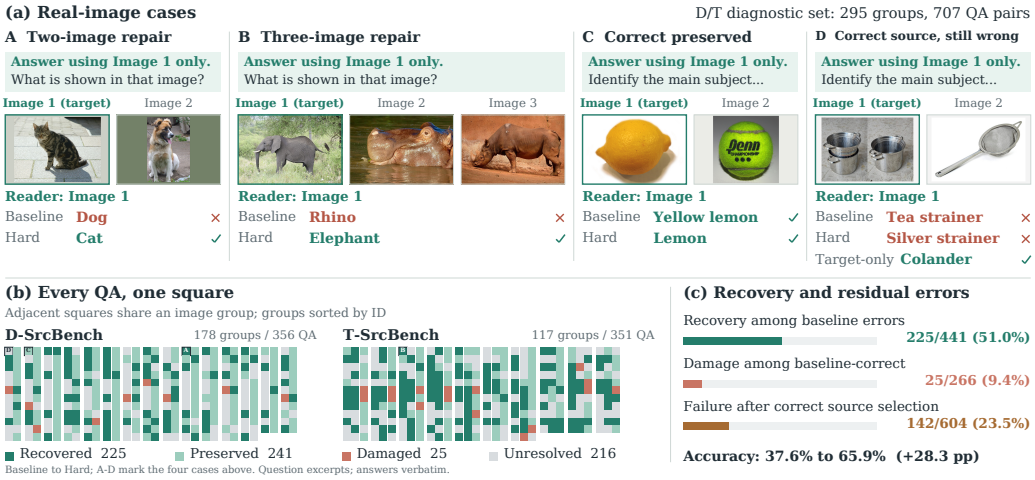


Figure 5: **Recovery and residual errors under Hard source enforcement.** Results on the 707-question D/T diagnostic set. (a) Real-image examples of recovery, preservation, and residual failure. (b) One square per question, colored by the baseline-to-Hard outcome. (c) Recovery, damage, and failure after correct source selection.

Conclusion

Reliable multi-image question answering requires not only identifying the requested image but also using that image as the source of evidence. In this paper, we formalize source misbinding as a distinct failure in which a model that can answer from the target image alone follows a competing image when distractors are present. Controlled experiments identify the failure as a layer-specific information-flow problem: the requested source remains encoded, yet competing evidence can still control the answer; enforcing access to the requested source in the repair-sensitive window selectively restores source-consistent answering. Based on these findings, we introduce Source-First Inference, which predicts the requested source in an unmodified planning pass and enforces it during a fresh generation pass. Our primary Hard executor improves accuracy from 30.4% to 59.0% on the independent test set and from 37.6% to 65.9% on the real-image diagnostic set, while retaining measurable limitations from imperfect source prediction and residual errors after correct source selection.

AI Use Statement

We used LLMs to assist with text translation, language polishing, and relevant literature searches. The authors reviewed the translated and polished text and verified that the cited references support the corresponding claims. The authors take full responsibility for the final content of the paper.

Portions of the content in the appendix were drafted by LLMs to organize experimental data. The authors manually inspected, validated, and revised all LLM-generated content to ensure correctness.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,

Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Imre Csiszár and František Matúš. Information Projections Revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003. doi: 10.1109/TIT.2003.810633. URL https://doi.org/10.1109/TIT.2003.810633.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing Multi-Image Understanding through Delimiter Token Scaling. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aebf6284fe85a8f44b4785d41bc8249a-Abstract-Conference.html.

Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. MIHBench: Benchmarking and Mitigating Multi-Image Hallucinations in Multimodal Large Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3754993. URL https://doi.org/10.1145/3746027.3754993.

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 *Computer Vision – ECCV 2014*, 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.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. MIA-DPO: Multi-Image Augmented Direct Preference Optimization For Large Vision-Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/557a20663907ed637c2807f608d5bec2-Paper-Conference.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5f92032278b1c70946e0b753068de51e-Paper-Conference.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Joonki Min, Chaeyun Kim, Hyungwook Choi, Yejin Kim, Kihyun Kim, Yohan Jo, and Joonseok Lee. Fine-Grained Multi Image Object Hallucination Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18295–18305, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Min_Fine-Grained_Multi_Image_Object_Hallucination_Benchmark_CVPR_2026_paper.html.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating Cross-Image Information Leakage in LVLMs for Multi-Image Tasks. *arXiv preprint arXiv:2508.13744v1*, 2025. URL https://arxiv.org/abs/2508.13744v1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sofia Serrano and Noah A. Smith. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, 2019. doi: 10.18653/v1/P19-1282. URL https://aclanthology.org/P19-1282/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, 2019. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and Mitigating Position Bias of Multi-image Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10599–10609, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tian_Identifying_and_Mitigating_Position_Bias_of_Multi-image_Vision-Language_Models_CVPR_2025_paper.html.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. In *International Conference on Learning Representations*,

2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9cf6139382f98623d08cc595622f3fb1-Paper-Conference.pdf.

Sudong Wang, Yunjian Zhang, Yao Zhu, Enci Liu, Jianing Li, Yanwei Liu, and Xiangyang Ji. SHIFT: Smoothing Hallucinations by Information Flow Tuning for Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3639–3649, 2025b. doi: 10.1109/ICCV51701.2025.00347. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_SHIFT_Smoothing_Hallucinations_by_Information_Flow_Tuning_for_Multimodal_Large_ICCV_2025_paper.html.

Wikimedia Commons contributors. Wikimedia Commons, n.d. URL https://commons.wikimedia.org/wiki/Commons:Welcome. Accessed September 21, 2026.

Yunqiu Xu, Linchao Zhu, and Yi Yang. MC-Bench: A Benchmark for Multi-Context Visual Grounding in the Era of MLLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17675–17687, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_MC-Bench_A_Benchmark_for_Multi-Context_Visual_Grounding_in_the_Era_ICCV_2025_paper.html.

Xiaochen Yang, Hao Fang, Jiawei Kong, Yaoxin Mao, Bin Chen, and Shu-Tao Xia. Looking Back and Forth: Cross-Image Attention Calibration and Attentive Preference Learning for Multi-Image Hallucination Mitigation. *arXiv preprint arXiv:2603.07048*, 2026. URL https://arxiv.org/abs/2603.07048.

Juntian Zhang, Chuanqi Cheng, Yuhan Liu, Wei Liu, Jian Luan, and Rui Yan. Weaving Context Across Images: Improving Vision-Language Models through Focus-Centric Visual Chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27782–27798, 2025. doi: 10.18653/v1/2025.acl-long.1347. URL https://aclanthology.org/2025.acl-long.1347/.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization. *arXiv preprint arXiv:2311.16839*, 2023. URL https://arxiv.org/abs/2311.16839.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. URL https://arxiv.org/abs/2402.11411.

A Datasets and Evaluation Protocols

A.1 Image groups, questions, and source annotations

Each ordered image group has an explicit target reference, an open-ended question, and canonical answers with accepted aliases for every source. Image order, target index, single-image questions, and base-model single-image responses are frozen. The diagnostic images come from Wikimedia Commons (Wikimedia Commons contributors, n.d.); the independent test uses COCO 2017 validation images (Lin et al., 2014).

Table 2: Dataset sizes in question–answer (QA) pairs and baseline accuracy. D and T are disjoint within a collection; the capability subset is nested within its parent.

Collection	Groups	QA	Capable QA	Baseline correct
Diagnostic D	178	356	356	159/356 (44.7%)
Diagnostic T	117	351	351	107/351 (30.5%)
Diagnostic ALL	295	707	707	266/707 (37.6%)
Test D	65	130	107	48/130 (36.9%)
Test T	64	192	154	50/192 (26.0%)
Test ALL	129	322	261	98/322 (30.4%)

A.2 Quality control, exclusions, and answer ambiguity

Screening checks subject visibility, annotation support, question specificity, and image suitability. Diagnostic refinement permits alias corrections, question revisions, and group removal; final-test review is blind to method outputs, source beliefs, and repair outcomes. Table 3 records the exclusions.

Table 3: Data construction and quality-control counts. Each count retains its original unit. Test membership was frozen without label or question overrides; an ordinary wrong answer was not relabeled as misbinding.

Stage	Record
Diagnostic precursor	315 groups, 749 questions. Exclude 40 questions outside the core track and one two-image group that fails the single-image recognition check.
Diagnostic retained	295 groups, 707 questions: 518 subject, 187 object-color, 2 bottle-body-color. Previously complete / repaired groups contribute 365 / 342 questions.
Test candidates	150 groups, 375 basic-entity questions; no implicit-reference or color track.
Test early review	16 of 19 groups retained.
Test later review	Review all 70 high-risk groups and 25 of 61 low-risk groups (seed 20260911). One issue in the low-risk sample meets the threshold of at most one; retain the other 36 low-risk groups automatically.
Test exclusions	Later review: 77 retained, 17 removed, 1 unresolved and excluded. Together with 3 earlier removals: 21 excluded groups; 129 groups / 322 questions retained.

A.3 Diagnostic, test, and capability-controlled populations

The diagnostic set overlaps reader development and is conditioned on correct target-only answering. The independent test fixes membership before capability filtering: 261 of 322 questions are answered correctly by the frozen base model under Parser Target-only. This same subset is used for every method; distractor-only correctness is not required. The full test includes 39 questions from 16 previously evaluated pilot groups and 283 from 113 newly opened groups.

A.4 Open-generation scoring and outcome definitions

Evaluation uses greedy open generation with at most 32 new tokens and no answer candidates. The scorer lowercases responses, replaces non-alphanumeric characters with spaces, and matches frozen aliases at word boundaries. Longer matches suppress contained shorter ones. A unique target match is correct; a unique non-target match is distractor-consistent. Multiple source matches, negation,

uncertainty, explicit alternatives, or multiple recognized colors in a color question are ambiguous. Unmatched answers are ordinary errors. Both remain in the denominator; aliases admit lexical variants, not generic superclasses.

Let B_i, C_i, and A_i^m indicate correct baseline multi-image, base-model target-only, and method-m answers. $G_i = 1$ requires $C_i = 1$, $B_i = 0$, and a unique match to a non-target source annotation. Model-relative misbinding instead uses that source's frozen single-image response.

$$\mathrm{Acc}(m) = \frac{\sum_i A_i^m}{N}, \qquad \mathrm{Repair}_G(m) = \frac{\sum_i G_i A_i^m}{\sum_i G_i}, \tag{5}$$

$$\mathrm{Recovery}(m) = \frac{\sum_i (1 - B_i) A_i^m}{\sum_i (1 - B_i)}, \qquad \mathrm{Damage}(m) = \frac{\sum_i B_i (1 - A_i^m)}{\sum_i B_i}. \tag{6}$$

A changed wrong answer is not a repair. Gold-consistent and model-relative flags can overlap: on the capable test their counts are 121 and 132, with intersection 111, leaving 28 of the 170 baseline errors outside their union.

Table 4: Scoring boundary examples. The first row is an observed diagnostic response; the other rows specify deterministic parser behavior for the displayed annotations, rather than additional experimental cases.

Target / distractor	Response	Scoring consequence
Colander / tea strainer	Silver strainer	Other: no accepted source alias matches.
Bell pepper / eggplant	Capsicum	Correct through the frozen synonym list.
Red / green	Red or green	Ambiguous: alternatives match two sources.
Red / green	Not green	Ambiguous; not a strict misbinding.
Bell pepper / eggplant	Vegetables	Other: a broad category does not identify a source.

A.5 Splits, overlap checks, and statistical units

Of 707 diagnostic questions, 352 match historical training groups, 161 validation, and 132 a historical locked-test partition; the remaining 62 are not certified image-disjoint. Independent-test checks cover image identity, perceptual duplicates, groups, and templates relative to development assets, not foundation-model pretraining.

Accuracy is a question micro-average. Paired bootstrap and sign-flip tests use shared-image connected components (129 full-test clusters; 128 capable-test clusters). We use 10,000 bootstrap draws with seed 20260911; sign-flip tests enumerate at most 16 clusters, otherwise use 10,000 draws. Holm correction (Holm, 1979) covers the five planned contrasts within each cohort. Repair and damage retain their separate denominators.

B Complete Experimental Results

B.1 Independent-test method comparison

Hard is the main executor. Soft and Predicted Target-only share its source reader; the latter is an input-isolation reference. Tables 5 and 6 show both populations. Moving from Hard to predicted isolation corrects 34 full-test answers and loses 18; isolation is not a per-question upper bound.

B.2 Diagnostic-set comparison

Hard exceeds matched-scope Soft on the diagnostic set (Table 7). Parser Target-only reaches 707/707 by the collection's eligibility criterion, not as an unconditional capability estimate.

Table 5: Complete independent-test accuracy. Entries are correct QA / all QA (percent). †: training reproduction; ‡: implementation of the published rule. Parser and Oracle Hard use identical sources on these explicit-reference questions.

Method	D	T	All
Baseline	48/130 (36.9%)	50/192 (26.0%)	98/322 (30.4%)
Soft	72/130 (55.4%)	93/192 (48.4%)	165/322 (51.2%)
Hard	93/130 (71.5%)	97/192 (50.5%)	190/322 (59.0%)
Predicted Target-only	102/130 (78.5%)	104/192 (54.2%)	206/322 (64.0%)
Parser Hard	96/130 (73.8%)	143/192 (74.5%)	239/322 (74.2%)
Parser Target-only	107/130 (82.3%)	154/192 (80.2%)	261/322 (81.1%)
Oracle Hard	96/130 (73.8%)	143/192 (74.5%)	239/322 (74.2%)
HA-DPO	48/130 (36.9%)	52/192 (27.1%)	100/322 (31.1%)
POVID	52/130 (40.0%)	53/192 (27.6%)	105/322 (32.6%)
MIA-DPO†	53/130 (40.8%)	63/192 (32.8%)	116/322 (36.0%)
SoFA‡	51/130 (39.2%)	47/192 (24.5%)	98/322 (30.4%)

Table 6: Accuracy on the fixed base-model capability subset of the independent test. Membership is shared by all methods.

Method	D	T	All
Baseline	45/107 (42.1%)	46/154 (29.9%)	91/261 (34.9%)
Soft	67/107 (62.6%)	82/154 (53.2%)	149/261 (57.1%)
Hard	88/107 (82.2%)	87/154 (56.5%)	175/261 (67.0%)
Predicted Target-only	102/107 (95.3%)	104/154 (67.5%)	206/261 (78.9%)
Parser Hard	91/107 (85.0%)	132/154 (85.7%)	223/261 (85.4%)
Parser Target-only	107/107 (100.0%)	154/154 (100.0%)	261/261 (100.0%)
Oracle Hard	91/107 (85.0%)	132/154 (85.7%)	223/261 (85.4%)
HA-DPO	43/107 (40.2%)	46/154 (29.9%)	89/261 (34.1%)
POVID	51/107 (47.7%)	52/154 (33.8%)	103/261 (39.5%)
MIA-DPO†	50/107 (46.7%)	58/154 (37.7%)	108/261 (41.4%)
SoFA‡	47/107 (43.9%)	43/154 (27.9%)	90/261 (34.5%)

Table 7: Complete diagnostic accuracy. Entries are correct QA / all QA (percent). †: training reproduction; ‡: implementation of the published rule. Parser and Oracle Hard use identical sources on these explicit-reference questions.

Method	D	T	All
Baseline	159/356 (44.7%)	107/351 (30.5%)	266/707 (37.6%)
Soft	229/356 (64.3%)	203/351 (57.8%)	432/707 (61.1%)
Hard	241/356 (67.7%)	225/351 (64.1%)	466/707 (65.9%)
Predicted Target-only	285/356 (80.1%)	320/351 (91.2%)	605/707 (85.6%)
Parser Hard	294/356 (82.6%)	247/351 (70.4%)	541/707 (76.5%)
Parser Target-only	356/356 (100.0%)	351/351 (100.0%)	707/707 (100.0%)
Oracle Hard	294/356 (82.6%)	247/351 (70.4%)	541/707 (76.5%)
HA-DPO	165/356 (46.3%)	107/351 (30.5%)	272/707 (38.5%)
POVID	162/356 (45.5%)	106/351 (30.2%)	268/707 (37.9%)
MIA-DPO†	98/356 (27.5%)	116/351 (33.0%)	214/707 (30.3%)
SoFA‡	157/356 (44.1%)	104/351 (29.6%)	261/707 (36.9%)

B.3 TWO-IMAGE, THREE-IMAGE, AND PRIOR-EXPOSURE STRATA

D and T contain different image groups, so their difference is not a paired image-count effect. On the 283 newly opened test questions, Baseline, Soft, Hard, and predicted isolation answer 84, 146, 168, and 180 correctly. Full-test confidence intervals do not apply to this stratum.

B.4 REPAIR, PRESERVATION, DAMAGE, AND ERROR TRANSITIONS

Hard gains more answers through recovery than it loses through damage (Tables 8–10). Table 11 distinguishes repair from transitions between wrong-answer categories; an unmatched answer alone does not identify its visual source.

Table 8: Repair and damage on the 707-QA diagnostic set (denominators: gold repair 365; recovery 441; damage 266). Gold repair is conditional on baseline gold-consistent misbinding; recovery covers every baseline error.

Method	Gold repair	Recovery	Damage
Baseline	0/365 (0.0%)	0/441 (0.0%)	0/266 (0.0%)
Soft	153/365 (41.9%)	182/441 (41.3%)	16/266 (6.0%)
Hard	191/365 (52.3%)	225/441 (51.0%)	25/266 (9.4%)
Predicted Target-only	292/365 (80.0%)	362/441 (82.1%)	23/266 (8.6%)
Parser Hard	245/365 (67.1%)	280/441 (63.5%)	5/266 (1.9%)
Parser Target-only	365/365 (100.0%)	441/441 (100.0%)	0/266 (0.0%)
Oracle Hard	245/365 (67.1%)	280/441 (63.5%)	5/266 (1.9%)
HA-DPO	24/365 (6.6%)	31/441 (7.0%)	25/266 (9.4%)
POVID	74/365 (20.3%)	93/441 (21.1%)	91/266 (34.2%)
MIA-DPO†	153/365 (41.9%)	182/441 (41.3%)	234/266 (88.0%)
SoFA‡	18/365 (4.9%)	23/441 (5.2%)	28/266 (10.5%)

Table 9: Repair and damage on the 322-QA independent test (denominators: gold repair 121; recovery 224; damage 98). Gold repair is conditional on baseline gold-consistent misbinding; recovery covers every baseline error.

Method	Gold repair	Recovery	Damage
Baseline	0/121 (0.0%)	0/224 (0.0%)	0/98 (0.0%)
Soft	54/121 (44.6%)	81/224 (36.2%)	14/98 (14.3%)
Hard	75/121 (62.0%)	109/224 (48.7%)	17/98 (17.3%)
Predicted Target-only	97/121 (80.2%)	132/224 (58.9%)	24/98 (24.5%)
Parser Hard	97/121 (80.2%)	144/224 (64.3%)	3/98 (3.1%)
Parser Target-only	121/121 (100.0%)	170/224 (75.9%)	7/98 (7.1%)
Oracle Hard	97/121 (80.2%)	144/224 (64.3%)	3/98 (3.1%)
HA-DPO	9/121 (7.4%)	15/224 (6.7%)	13/98 (13.3%)
POVID	36/121 (29.8%)	47/224 (21.0%)	40/98 (40.8%)
MIA-DPO†	62/121 (51.2%)	91/224 (40.6%)	73/98 (74.5%)
SoFA‡	4/121 (3.3%)	12/224 (5.4%)	12/98 (12.2%)

Table 10: Paired answer outcomes (QA counts): retained correct, recovered error, damaged correct, and residual error. Columns form a partition within each row.

Collection	Method	Retained	Recovered	Damaged	Residual
Diagnostic	Soft	250	182	16	259
Diagnostic	Hard	241	225	25	216
Diagnostic	Predicted Target-only	243	362	23	79
Full test	Soft	84	81	14	143
Full test	Hard	81	109	17	115
Full test	Predicted Target-only	74	132	24	92
Capable test	Soft	77	72	14	98
Capable test	Hard	76	99	15	71
Capable test	Predicted Target-only	74	132	17	38

Table 11: Baseline-to-Hard error-category transitions. The post-intervention distractor-match column reports the scorer category; it is not an additional count of baseline strict misbindings.

Collection	Baseline	Correct	Distractor	Ambig.	Other
Diagnostic	Correct	241	21	0	4
Diagnostic	Gold-consistent	191	151	0	23
Diagnostic	Other error	34	9	0	33
Test	Correct	81	12	0	5
Test	Gold-consistent	75	34	0	12
Test	Other error	34	14	0	55

B.5 Source selection and answer correctness

Correct source selection does not guarantee a correct Hard answer: 142/604 diagnostic and 34/206 capable-test cases still fail (Table 12). On the full test, 53 of 259 correctly selected targets also fail under isolation; all 53 fall outside the capability subset.

Table 12: Joint source and answer outcomes (QA counts). S denotes correct source selection and A correct answering; source correctness is measured independently of answer correctness.

Collection	Method	S, A	$S, \neg A$	$\neg S, A$	$\neg S, \neg A$
Diagnostic	Soft	419	185	13	90
Diagnostic	Hard	462	142	4	99
Diagnostic	Predicted Target-only	604	0	1	102
Full test	Soft	152	107	13	50
Full test	Hard	187	72	3	60
Full test	Predicted Target-only	206	53	0	63
Capable test	Soft	137	69	12	43
Capable test	Hard	172	34	3	52
Capable test	Predicted Target-only	206	0	0	55

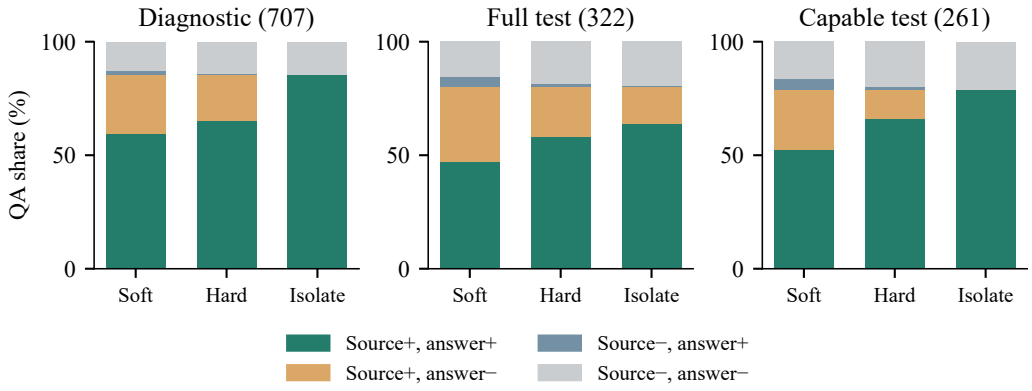


Figure 6: Source selection and answer correctness under the shared reader. "Isolate" denotes Predicted Target-only. Bars use each full displayed cohort. Correct answers with wrong source choices and wrong answers with correct choices are counted separately. The capable cohort excludes target-only failures by construction.

B.6 External baselines and cross-model measurements

All final LLaVA comparisons use identical questions, image order, frozen scoring, greedy decoding, and a 32-token limit. Table 15 records checkpoint handling and deviations from the original methods.

The MIA-DPO preference-data analysis is separate from its trained-baseline evaluation. Rejected-attribute matches lack official distractor-source labels and are not a strict misbinding rate (Table 16).

Table 13: Paired independent-test differences (first method minus second), in percentage points (pp). Confidence intervals (CIs) use 10,000 shared-image-cluster bootstrap draws. Sign-flip p-values are Holm-adjusted within the five planned contrasts for each cohort.

Subset	Contrast	Gain	95% CI	p
Full	Soft / Base	+20.81	[15.86, 25.62]	0.0005
Full	Hard / Soft	+7.76	[3.64, 11.99]	0.0020
Full	Pred. isolation / Hard	+4.97	[0.95, 9.03]	0.0237
Full	Parser isolation / Pred. isolation	+17.08	[13.79, 20.36]	0.0005
Full	Oracle Hard / Hard	+15.22	[12.09, 18.35]	0.0005
Capable	Soft / Base	+22.22	[16.54, 27.88]	0.0005
Capable	Hard / Soft	+9.96	[4.91, 15.07]	0.0005
Capable	Pred. isolation / Hard	+11.88	[8.14, 15.91]	0.0005
Capable	Parser isolation / Pred. isolation	+21.07	[17.19, 25.00]	0.0005
Capable	Oracle Hard / Hard	+18.39	[14.67, 22.26]	0.0005

Table 14: Independent-test accuracy differences from Base, with the archived paired group-bootstrap intervals. These intervals do not constitute a claim that every pair of methods differs significantly.

Method	Gain (pp)	95% CI
HA-DPO	+0.62	[-2.48, 3.70]
POVID	+2.17	[-1.81, 6.12]
MIA-DPO	+5.59	[0.31, 11.01]
SoFA	+0.00	[-2.80, 2.85]
Hard	+28.57	[23.25, 33.75]

Table 15: Baseline implementations and exclusions. Comparable checkpoints use LLaVA-1.5-7B and CLIP ViT-L/14 at 336 pixels (Liu et al., 2024; Radford et al., 2021). † denotes our training reproduction; ‡ denotes our implementation of a published rule.

Method	Implementation
HA-DPO (Zhao et al., 2023)	Author adapter `juliozhao/hadpo-llava-1.5`; recorded multimodal-projector name mapping applied.
POVID (Zhou et al., 2024)	The release labeled merged stage 1 resolves to a low-rank adaptation (LoRA) adapter. Load it over the base, merge in memory, then apply the author stage-2 adapter.
MIA-DPO† (Liu et al., 2025)	Public `liuhaotian/llava-v1.5-7b`, revision `4481d270cc22fd5c4d1bb5df129622006ccd9234`; private author initialization unavailable. Official preference data; modified LLaVA-Hound-DPO code.
Training	LoRA (Hu et al., 2022) rank/scaling/dropout: 128/256/0.05; learning rate 5×10^{-5}; $\beta = 0.1$, $\gamma = 0$; 1 epoch; bfloat16 (BF16); maximum length 4096; frozen projector. Batch 64: two A100 40-GB devices, 1 example/device, 32 accumulation steps; 421 optimizer steps. Follow the final author script, which differs from the paper (3 epochs, $\gamma = 0.1$).
Output check	Diagnostic / test: 284/707 and 30/322 answers reach the token limit; 665 and 311 distinct responses. Fixed scorer retained.
SoFA‡ (Tian et al., 2025)	Interpolate separately normalized causal/bidirectional inter-image attention at even layers. Coefficient 0.5 selected on 32 separate development examples. No executable author implementation was available; the published interpolation location is ambiguous.
Excluded	Inspected LLaVA-RLHF (Sun et al., 2024): 224-pixel visual configuration. DTS (Lee et al., 2026): required delimiter states unavailable in the same form. CAPL (Yang et al., 2026): no comparable checkpoint evaluation.

Table 16: MIA-DPO preference-data behavior and known-source intervention. The 91 target-only-correct cases come from 325 refined color-attribute candidates; 41/91 (45.1%) match the rejected attribute. Source-localized probe and patching cohorts are separate (Section D).

Condition	Count
Multi-image answer: chosen / rejected / other	36 / 41 / 14
Chosen-attribute recovery on 41 rejected matches, layers 7–14	8/41
Chosen-attribute recovery, layers 16–23	15/41
Chosen-attribute recovery, layers 24–31	3/41

Qwen2.5-VL-7B (Bai et al., 2025) uses its own reader, visual spans, and intervention window. Tables 17 and 18 use different cohorts. The oracle scan is diagnostic and does not replace the frozen LLaVA executor. Sequence-matched and target-only activation donors have different alignments and are reported separately.

Table 17: Earlier Qwen2.5-VL-7B evaluation on historical D/T and localized MIA cases, with its historical Query-conditioned Evidence Routing (QER) source reader and layer window (strict-repair denominators: D 2, T 4, MIA 1 baseline misbindings). These cohorts are separate from both final LLaVA collections and from the 11-error actuator study.

Data	Execution	Accuracy	Strict repair
D	baseline	31/34 (91.2%)	0/2 (0.0%)
D	oracle	31/34 (91.2%)	0/2 (0.0%)
D	always on QER	31/34 (91.2%)	0/2 (0.0%)
D	soft selective QER	31/34 (91.2%)	0/2 (0.0%)
MIA-DPO	baseline	6/11 (54.5%)	0/1 (0.0%)
MIA-DPO	oracle	6/11 (54.5%)	0/1 (0.0%)
MIA-DPO	always on QER	6/11 (54.5%)	0/1 (0.0%)
MIA-DPO	soft selective QER	6/11 (54.5%)	0/1 (0.0%)
T	baseline	35/40 (87.5%)	0/4 (0.0%)
T	oracle	39/40 (97.5%)	4/4 (100.0%)
T	always on QER	39/40 (97.5%)	4/4 (100.0%)
T	soft selective QER	38/40 (95.0%)	3/4 (75.0%)

In the sequence-matched conditions ("seq.-matched"), the donor retains the recipient's multi-image prompt, input IDs, and visual-token layout, but its non-target pixel blocks are zeroed before the vision encoder. At each selected layer, patching replaces the last prefill text position's residual state or attention output with the corresponding donor activation. The target-only conditions instead use a single-image donor with a different input sequence. These donor constructions should therefore be distinguished when interpreting repair rates.

Table 18: Archived Qwen oracle actuator and sequence-matched activation-patching results on 11 fixed errors. All rows use the same samples and are not independent replications. The 14–27 attention mask and sequence-matched residual patch each repair 4/11 errors, but repair different subsets. The 21–27 target-only residual patch repairs 10/11 under a different donor-input configuration. Visual attention changes alone are not counted as answer repair.

Layers/condition	Operator	Correct	Misbound	Other
all 0 27	attention mask	5/11 (45.5%)	4	2
all 0 27	mask plus value	6/11 (54.5%)	3	2
all 0 27	value zero	6/11 (54.5%)	2	3
early 0 6	attention mask	1/11 (9.1%)	9	1
early 0 6	mask plus value	2/11 (18.2%)	8	1
early 0 6	value zero	1/11 (9.1%)	8	2
early 7 13	attention mask	2/11 (18.2%)	9	0
early 7 13	mask plus value	5/11 (45.5%)	5	1
early 7 13	value zero	4/11 (36.4%)	5	2
input	pixel zero	6/11 (54.5%)	3	2
late 21 27	attention mask	1/11 (9.1%)	10	0
late 21 27	mask plus value	1/11 (9.1%)	10	0
late 21 27	target only attention output	2/11 (18.2%)	8	1
late 21 27	target only residual	10/11 (90.9%)	1	0
late 21 27	value zero	0/11 (0.0%)	11	0
mid late 14 27	attention mask	4/11 (36.4%)	5	2
mid late 14 27	mask plus value	5/11 (45.5%)	5	1
mid late 14 27	value zero	5/11 (45.5%)	4	2
middle 14 20	attention mask	3/11 (27.3%)	6	2
middle 14 20	mask plus value	5/11 (45.5%)	5	1
middle 14 20	value zero	4/11 (36.4%)	5	2
early 7 13	seq.-matched residual	0/11 (0.0%)	11	0
middle 14 20	seq.-matched residual	2/11 (18.2%)	9	0
late 21 27	seq.-matched residual	3/11 (27.3%)	8	0
mid late 14 27	seq.-matched residual	4/11 (36.4%)	7	0
late 21 27	seq.-matched attention output	2/11 (18.2%)	9	0

C Ablations and Sensitivity Analyses

C.1 Window width and position

The original layer-set controls compare contiguous, dispersed, and full-depth interventions on 150 synthetic distractor-consistent errors (Table 20; Figure 7). We subsequently scan widths 1–8 at every legal starting layer with stride one: 228 windows on the same fixed cases, using the known target source and unchanged model, inputs, decoding, and nonvisual-prefill masking.

Table 19 summarizes the scan. Width eight yields complete repair at four consecutive starts, placing the adopted layers 16–23 on a maximum-repair plateau with tolerance to placement. Width six also reaches 150/150, so eight is not the minimum effective width and 16–23 is not uniquely optimal. This is a sensitivity analysis on the existing synthetic cohort, not held-out window selection or a real-image benchmark comparison.

Table 19: Exhaustive contiguous-window scan on 150 fixed four-image synthetic distractor-consistent errors. Each window is evaluated on all cases; source selection uses the known target, without a reader. Layer ranges are zero-based (add one for one-based numbering). The best ranges list every maximizing position; widths 1–3 fail at all positions. The baseline repairs 0/150 and full-depth masking repairs 150/150.

Width	Windows	Best repair	Best layer ranges
1	32	0/150 (0.0%)	All positions (0 repairs)
2	31	0/150 (0.0%)	All positions (0 repairs)
3	30	0/150 (0.0%)	All positions (0 repairs)
4	29	39/150 (26.0%)	19–22
5	28	144/150 (96.0%)	18–22
6	27	150/150 (100.0%)	19–24
7	26	150/150 (100.0%)	16–22, 18–24, 19–25
8	25	150/150 (100.0%)	16–23, 17–24, 18–25, 19–26

Table 20: Tested layer sets on the same 150 four-image cases, using the known target and zero-based layer indices. Raw answers, membership, visual spans, layer lists, and prefill/decode counters were checked. Dispersed set 1 was previously mislabeled as a mechanism band; its actual layers are shown. These named bands are not a regular sliding scan or held-out window selection.

Condition	Layers	Width	Correct / tested
No intervention	–	0	0/150 (0.0%)
Dispersed set 1	0, 2, 7, 8, 9, 10, 11, 14	8	0/150 (0.0%)
Early contiguous	7–14	8	0/150 (0.0%)
Middle contiguous	16–23	8	150/150 (100.0%)
Late contiguous	24–31	8	0/150 (0.0%)
Dispersed set 2	1, 4, 6, 12, 18, 21, 27, 30	8	0/150 (0.0%)
All layers	0–31	32	150/150 (100.0%)

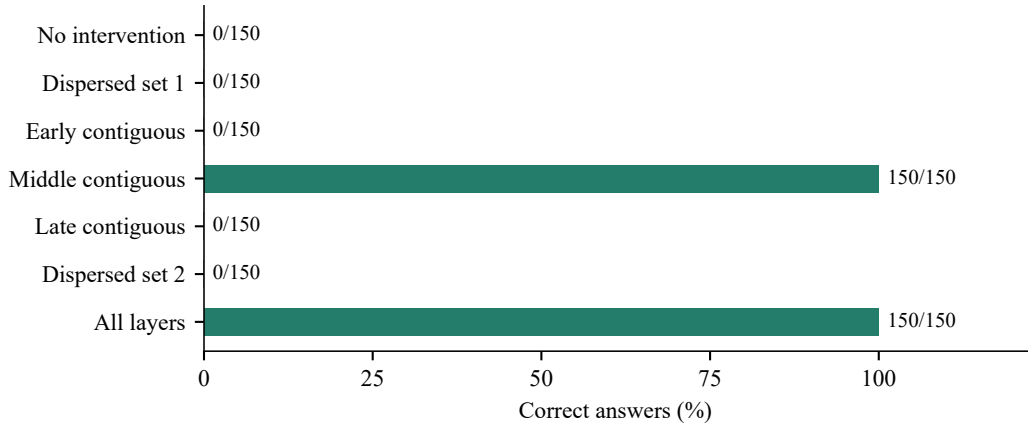


Figure 7: Known-source layer controls on 150 synthetic distractor-consistent errors: one red target in position 1 or 2 and three green distractors, 576 visual tokens per image. Every case is target-only correct and baseline green. The constraint affects nonvisual prefill queries, not cached decoding. Zero bars are measured failures; the categorical axis is not an interpolated sweep.

C.2 SHARED-READER EXECUTION CONTROLS

Table 21 fixes source prediction while comparing logit bias, support restriction, and input isolation. Soft and Hard share layers 16–23 and all nonvisual prefill queries. Historical multi-layer-planner or text-after-image-only Soft results are excluded.

Table 21: Shared-reader execution comparison, using the final matched-scope Soft. Entries are correct QA / cohort size (percent).

Execution	Diagnostic	Full test	Capable test
Soft	432/707 (61.1%)	165/322 (51.2%)	149/261 (57.1%)
Hard	466/707 (65.9%)	190/322 (59.0%)	175/261 (67.0%)
Predicted Target-only	605/707 (85.6%)	206/322 (64.0%)	206/261 (78.9%)

C.3 SOURCE-SELECTION CONTROLS

With the Hard executor fixed, parsed/annotated sources raise accuracy from 466/707 to 541/707 on the diagnostic set and from 190/322 to 239/322 on the full test. Parser and Oracle indices coincide here. Parser Target-only reaches 261/322, showing the remaining target-image answering limit.

C.4 REFERENCE, ORDER, CARDINALITY, AND COMPONENT CONTROLS

Reference swaps hold pixels and order fixed; image rotations also update the reference to preserve the target. The latter changes both position and index wording. Historical routing uses the maximum-attention source; a separate 32-group reader-belief check reuses recorded answers rather than regenerating them.

Table 22: Answer outcomes after changing only the explicit image reference, on historical held-out cases. Correctness is scored against the newly requested image. No ambiguous responses occur in these rows.

Data	Original answer	QA	Correct after swap	Misbound	Other
D	correct	8	3/8 (37.5%)	5	0
D	misbound	26	18/26 (69.2%)	8	0
MIA-DPO	misbound	11	9/11 (81.8%)	0	2
T	correct	13	0/13 (0.0%)	13	0
T	misbound	27	2/27 (7.4%)	24	1

Adding a third image changes correct source selection to wrong on 29/174 locked-partition conditions, versus 1/174 in reverse (Table 23). The 922 conditions come from 461 questions; they are not independent questions. An extra softmax term can reduce target probability even when its raw score rises.

Table 23: Source-selection transitions from two to three images for a historical reader. Each question contributes two paired conditions. Columns partition conditions, not independent questions. C/W denote correct/wrong source selection.

Partition	QA	Pairs	C→C	C→W	W→C	W→W
Training	253	506	384	121	0	1
Validation	121	242	182	60	0	0
Locked development	87	174	143	29	1	1

Question-only exceeds visual-only in the historical controls (Table 24). These separately trained networks differ from the final reader, so the comparison is not a component-removal ablation of that reader.

Table 24: Historical source-selection controls on the 170-question locked development partition. Question-only: mean block-20 query; visual-only: zero query in a shared source scorer; position-only: three learned constants; fixed three class: query plus source evidence padded to three images. Shuffled pairing permutes training queries; random labels use deterministic valid source labels. Each control is trained separately and evaluated against true labels, not obtained by ablating the final reader.

Control	D	T	All
fixed three class	75/83 (90.4%)	53/87 (60.9%)	128/170 (75.3%)
position only	38/83 (45.8%)	24/87 (27.6%)	62/170 (36.5%)
question only	77/83 (92.8%)	52/87 (59.8%)	129/170 (75.9%)
random labels	41/83 (49.4%)	31/87 (35.6%)	72/170 (42.4%)
shuffled pairing	49/83 (59.0%)	35/87 (40.2%)	84/170 (49.4%)
visual only	48/83 (57.8%)	39/87 (44.8%)	87/170 (51.2%)

D Additional Mechanistic Analyses

D.1 Source-identity and reference-index probes

Source labels remain decodable on misbound cases, while the final-prefill reference probe is correct on all held-out D/T variants at block 16 (Figure 8; Table 26). These readouts measure recoverability, not generator use or answer accuracy. Probe cohorts and fitting procedures are listed in Table 25.

Table 25: Historical probe cohorts and fitting procedures. These probes are distinct from the final Source-First reader; preprocessing and classifiers use discovery data only.

Probe / cohort	Definition
D/T partition	129 target-only-correct instances: 55 discovery, 74 held out; image families stay within partitions. Held out: D 34 (8 correct, 26 misbound), T 40 (13 correct, 27 misbound).
Source identity	At each block 0–31, fit cosine nearest centroids to normalized discovery visual-token states. Labels are image of origin. Token accuracy is separate from within-image majority voting.
Reference index	Original/swapped questions share partitions. Training-standardized features; balanced logistic regression; at most 2,000 iterations; binary `liblinear`, multiclass `lbfgs`. Five spans: reference phrase, "Image," numeral, question, final prefill. Full curves are in the comma-separated values (CSV) file.
Localized MIA	11 held-out cases with 2–5 images; block-16 attention argmax selects the target in 1/11 cases and a non-target in 10/11. Reference probe: 16/22 variants correct. Source-identity score on four two-image cases: 75% (chance 50%), not QA accuracy. Localization uses additional single-image evidence, not official preference labels.

Table 26: Reference-index probe at layer 16. Counts include original/reference-swapped variants; they are not independent image groups.

Data	Readout	Train	Test	Accuracy	Majority
D	reference	56	68	100.0%	50.0%
D	image word	56	68	50.0%	50.0%
D	number	56	68	100.0%	50.0%
D	question	56	68	100.0%	50.0%
D	final prefill	56	68	100.0%	50.0%
T	reference	54	80	100.0%	50.0%
T	image word	54	80	30.0%	50.0%
T	number	54	80	100.0%	50.0%
T	question	54	80	100.0%	50.0%
T	final prefill	54	80	100.0%	50.0%
MIA-DPO	reference	42	22	100.0%	40.9%
MIA-DPO	image word	42	22	36.4%	40.9%
MIA-DPO	number	42	22	100.0%	40.9%
MIA-DPO	question	42	22	100.0%	40.9%
MIA-DPO	final prefill	42	22	72.7%	40.9%

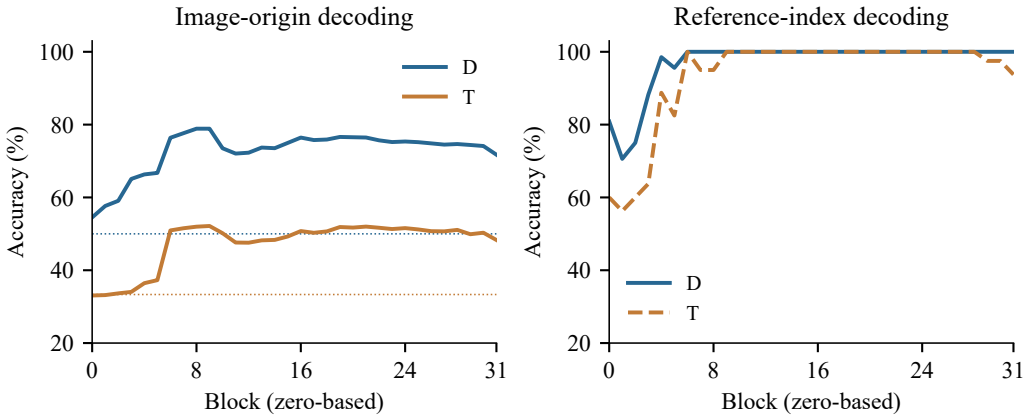


Figure 8: Historical held-out D/T probes. Left: visual-token image-origin accuracy on misbound cases (D: 26 instances / 29,952 tokens; T: 27 / 46,656). At block 16, accuracy is 76.44% / 50.76%; chance and patch-position-only baselines are 50% / 33.33%. Right: final-prefill reference-index accuracy on original/swapped questions, 68/68 D and 80/80 T at block 16. Paired variants are not independent image groups.

D.2 Attention measurements and query–key (QK) margins

For query q, head h, and layer ℓ, let $z_k = Q_q K_k^\top / \sqrt{d}$ be the post-RoPE (Su et al., 2024) logit before masking and softmax. The token-count-normalized source margin is

$$\Delta_{\ell hq} = \log\left(\frac{1}{|\mathcal{V}_t|}\sum_{k\in\mathcal{V}_t} e^{z_k}\right) - \log\left(\frac{1}{|\mathcal{V}_c|}\sum_{k\in\mathcal{V}_c} e^{z_k}\right), \tag{7}$$

Here t is the target; c is the localized competitor for misbound cases or strongest non-target for correct cases. Figure 9 averages final-prefill margins over heads and instances. Separation begins at the earliest measured block, not specifically at block 16.

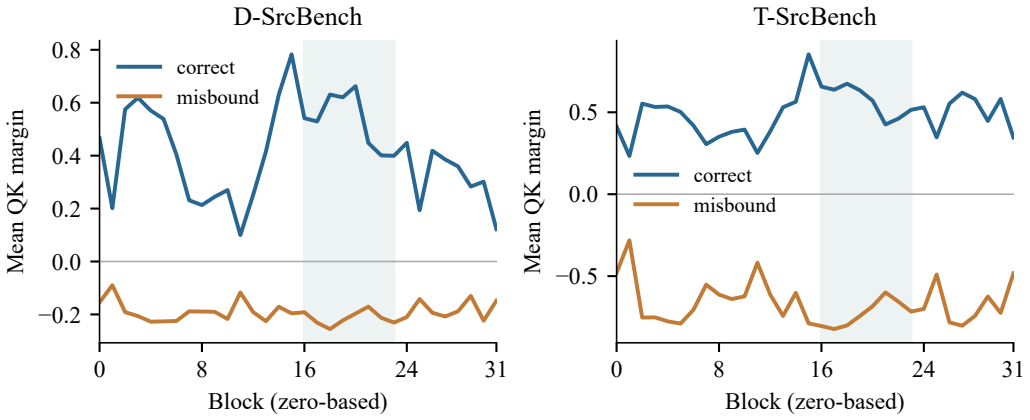


Figure 9: Pre-softmax target–competitor margins on historical held-out instances: correct/misbound counts are 8/26 for D and 13/27 for T. Shading marks the intervention band, not the error origin. Across blocks 16–23, 28–32 of 32 D heads and 29–32 T heads have negative mean margins; these need not be the same heads at every layer. Individual queries can differ from the stratum mean.

Image mass $m_s = \sum_{k\in\mathcal{V}_s} p_k$ differs from visual share $m_s / \sum_j m_j$. Table 27 separates the qualitative measurements. The released QK trajectories also include question-query, arithmetic-mean, and maximum-logit variants.

Table 27: Attention measurements used in different illustrations. Values are raw attention masses unless explicitly stated. Local zero distractor mass after Hard does not exclude distractor information in residual states or later decoding.

Illustration	Measurement
Two-image red/green	Teacher-forced stored answer token "green," block 16, head mean; expanded position 1182, token identifier (ID) 7933. Target / distractor mass: 0.0444771 / 0.0888309. Not final-prompt generation prefill.
Four-image probe	Blocks 16 / 31: target mass 0.01794 / 0.04327; total distractor mass 0.09680 / 0.13251. Per-distractor division by three gives plotted target/competitor pairs 1.79%/3.23% and 4.33%/4.42%.
Cat/Dog method figure	Actual generation prefill, last valid prompt token, separate Baseline/Hard runs. Blocks 16 and 23; per-head effective weights after masking and full-row renormalization, before attention times value. Raw image sums; no per-image rescaling or smoothing.

D.3 Source-specific retention controls

Target retention exceeds wrong-source and random-token retention on the mixed 32-question development subset (Table 28). Random retention matches token count, not attention mass; text and visual keys are renormalized jointly.

Table 28: Source retention on 32 development questions: 16 baseline-correct and 16 baseline-error cases, balanced across D/T and entity/color. All conditions keep the full input and prefill band. Random retention uses a fixed per-example seed to retain 576 visual tokens across sources. Correct counts are accuracy on this mixed cohort, not repair over 32 errors. No separately mass-matched control is included.

Retention	Accuracy	Error recovery	Correct retained
Target retained	31/32 (96.9%)	15/16 (93.8%)	16/16 (100.0%)
Wrong source retained	6/32 (18.8%)	0/16 (0.0%)	6/16 (37.5%)
Random 576 tokens	17/32 (53.1%)	1/16 (6.2%)	16/16 (100.0%)

D.4 Distractor replacement and reference changes

Changing distractor content can change the wrong answer while the target and reference stay fixed (Table 29). Reference swaps and image rotations change different input components and are evaluated separately (Section C.4). The same dependence is evident at the input level. In paired distractor-replacement cases, we keep the target image, question, source count, and target position fixed while changing only the distractor image. Among initially incorrect answers, 63.64% switch to the new distractor image's annotated answer; among initially correct answers, 22.22% become incorrect. Thus, changing only the competing source can redirect the model's answer without changing the question or target image.

Table 29: Distractor replacement with target image, target position, question, and source count fixed. Of 32 planned development cases, 20 had valid replacement images; the other 12 are excluded, not counted as failures. Rows partition the 20 completed cases.

Outcome	Cases
Correct baseline remains correct	7
Baseline error follows new distractor	7
Baseline error becomes target-correct	1
Correct baseline becomes wrong	2
Remains another error	3

D.5 Bidirectional activation patching

We replace internal activations across runs (Meng et al., 2022). B is the baseline and C the known-target 16–23 constrained run on the same prompt. Patches replace post-output-projection self-attention vectors at matching nonvisual prompt positions after the last image. Correction patches C into B; reversal patches B into an otherwise constrained recipient. Cached decoding is not

patched. We next test whether intermediate attention-module outputs are sufficient to change the model's answers. On this D/T cohort, the model answers with content from a distractor image during normal operation but answers correctly under target-source restriction. We keep the images and question identical across the two runs and exchange only the attention-module outputs at text positions following the final image in layers 16–23. Replacing the baseline outputs with those from the source-restricted run repairs all instances; the reverse replacement restores misbinding in all instances. As a control, we preserve the norm of each token's intervention delta while permuting its feature coordinates, which repairs no instances. These results show that, on this preselected cohort of repairable misbound instances, replacing these intermediate outputs is sufficient to switch answers bidirectionally between the target and distractor images, whereas equivalent random perturbations do not have the same effect.

Table 30: Aligned attention-output patching on held-out repairable cohorts (D 23, T 26, localized MIA 5). Eligibility requires target-only correctness, baseline distractor matching, and a correct constrained answer. Correction is $C \to B$; reversal is $B \to C$. Randomized control permutes hidden coordinates of the donor-minus-recipient vector with a deterministic sample/layer seed, preserves its per-token norm, and adds it to the recipient; token positions stay fixed.

Data	Layers	Correction	Reversal	Randomized
D	7–14	0/23 (0.0%)	0/23 (0.0%)	0/23 (0.0%)
D	16–23	23/23 (100.0%)	23/23 (100.0%)	0/23 (0.0%)
D	24–31	5/23 (21.7%)	0/23 (0.0%)	0/23 (0.0%)
MIA localized	7–14	0/5 (0.0%)	0/5 (0.0%)	0/5 (0.0%)
MIA localized	16–23	5/5 (100.0%)	5/5 (100.0%)	0/5 (0.0%)
MIA localized	24–31	0/5 (0.0%)	0/5 (0.0%)	0/5 (0.0%)
T	7–14	0/26 (0.0%)	0/26 (0.0%)	0/26 (0.0%)
T	16–23	26/26 (100.0%)	26/26 (100.0%)	0/26 (0.0%)
T	24–31	1/26 (3.8%)	1/26 (3.8%)	0/26 (0.0%)

Patching 16–23 transfers correction and reversal throughout the eligible cohorts. This does not show that every layer or token is necessary. Single-image donors change sequence context and are excluded. A separate eight-case query/key/value (Q/K/V) input check found identical block-16 donors (no-op patches) and nonzero block-23 differences; it is a different intervention and does not locate a unique error-origin layer.

E Implementation, Efficiency, and Failure Cases

E.1 Reader architecture, training, and freezing

The reader uses cached block-16 features from a frozen backbone (Table 31). Source labels, stored as `teacher_source`, follow the explicit reference under order/cardinality variants. Answers and attribute labels are not reader inputs; dataset identity is used only for training-domain balancing. Here, z_s is the score assigned to source I_s. Softmax converts these logits into the normalized source belief $\mathbf{b}$, with $b_s \geq 0$ and $\sum_{s=1}^{N} b_s = 1$. Thus, b_s represents the reader's relative belief that I_s is the source specified by the question; it is used as a normalized source score. The predicted source $\hat{s}$ is the index of the source with the highest score.

For source logits z, label y, and old-reader distribution p^{old}, the per-variant loss with Kullback–Leibler (KL) divergence is

$$L_v = -\log \operatorname{softmax}(z)_y + 0.25\,[0.35 - z_y + \max_{s \neq y} z_s]_+ \\ + 0.10\,\mathbf{1}[\arg\max p^{\text{old}} = y,\ \max p^{\text{old}} \geq 0.70]\, D_{\text{KL}}(p^{\text{old}} \| \operatorname{softmax}(z)). \tag{8}$$

The family loss averages L_v and adds 0.10 times the mean identity-wise variance of aligned raw source logits across variants. The hard negative is the highest-logit wrong source. Table 32 gives the frozen training and selection settings.

E.2 Two-stage inference, attention operators, and caches

Only the source belief and selected index pass from planning to generation (Table 33); generation starts a fresh forward and its own cache. Hard removes nonselected visual weights after

Table 31: Final reader architecture. All feature dimensions follow head averaging. The token gate, projections, and scorer are shared across sources; softmax and argmax are applied outside the reader. No confidence threshold, abstention, reselection, or output-dependent fallback is used.

Stage	Operation / size
Block-16 inputs	Question queries $H \times T_q \times 128$. Per-source keys/values: mean over visual tokens. Evidence: source-local softmax of query logits, value-weighted sum, then question-token mean.
Head average	Queries $T_q \times 128$; key, value, evidence each $N \times 128$.
Pooling / normalization	Learned scalar token gate with softmax pooling; parameter-free layer normalization of query and source features.
Source representation	Concatenate normalized key/value/evidence with their source-centered versions: 768 dimensions.
Projections	Bias-free query $128 \to 32$, source $768 \to 32$; both L2-normalized.
Scorer	Concatenate query, source, and elementwise product (96); Linear$(96, 64)$, Gaussian error linear unit (GELU) (Hendrycks & Gimpel, 2016), Linear$(64, 1)$.
Parameters	35,074 stored scalars, including one frozen layer-weight parameter.

Table 32: Reader training and model selection. The selected checkpoint is fixed before the independent test; backbone weights stay frozen. Diagnostic overlap is reported in Section A.5. Distinct training-question/family counts are not inferred from epoch-average accuracies.

Setting	Value
Domains	Historical D, T, localized MIA; refined D and T. Reference variants update labels under original/reversed order and changed cardinality.
Sampling	Semantic-target families with related order/cardinality variants. Balance domain and target position by repeating smaller buckets. One family per optimizer step; mean variant loss.
Enabled options	Hard-negative, anti-forgetting, permutation, cardinality, position balancing, source normalization.
AdamW (Loshchilov & Hutter, 2019)	Learning rate 8×10^{-4}; weight decay 0.01; gradient-norm clipping 1.0; 8 epochs; seed 20260912.
Architecture selection	Rank by worst-domain leave-one-domain-out accuracy, then mean. Layer 16 mean/worst: 75.76%/57.14%; layers 16–17: 82.55%/36.23%.
Final selection	Seed chosen on development validation; independent-test results do not select the checkpoint.

causal/padding masking and softmax, then renormalizes all remaining keys, including text. Its denominator is clamped to the dtype's smallest positive normal value. Soft adds $\log(Nb_s)$ before softmax with coefficient one.

Table 33: Source-First Inference. Hard and Soft affect all heads and all nonvisual prefill queries in the specified blocks. The subsequent decoding stage uses the cache created by that generation pass.

Input:	ordered images $I_{1:N}$, multi-image question q, frozen single-image question q^{solo}, reader g_ϕ.
1.	Prepare the full prompt and image tensors; verify the visual spans $\mathcal{V}_{1:N}$.
2.	Run an unmodified planning forward without a decoder cache; collect block-16 features F.
3.	Compute $b = \text{softmax}(g_\phi(F))$ and $\hat{s} = \arg\max_s b_s$.
4.	Start a fresh generation pass: *Hard:* keep all input images; edit effective attention rows in blocks 16–23. *Soft:* keep all input images; add $\log(Nb_s)$ to source-s visual logits. *Predicted Target-only:* generate from $(I_{\hat{s}}, q^{\text{solo}})$ without an attention edit.
5.	Emit the first token from the fresh prefill; continue ordinary cached decoding.
6.	Restore the attention hook and return the generated answer.

For original attention p and retained support K, the idealized row is an information projection (Csiszár & Matúš, 2003):

$$\min_{r \in \Delta} D_{\text{KL}}(r \| p) \quad \text{subject to} \quad r_j = 0 \text{ for } j \notin K, \qquad r_j = \frac{p_j \mathbf{1}[j \in K]}{\sum_{k \in K} p_k}. \tag{9}$$

Here Δ is the row simplex and K contains text and selected-image keys. A normalization multiplier gives $r_j \propto p_j$ on K. This is a row-level result, not a projection of the model's entire information state.

E.3 Prompts, decoding, and runtime configuration

Evaluation uses frozen LLaVA-1.5-7B (Liu et al., 2024), CLIP ViT-L/14 (Radford et al., 2021), 336-pixel images, 576 visual tokens/image, 16-bit floating point (FP16), and eager attention. The test seed is 20260910; decoding is greedy with a 32-token limit. Hooks require verified contiguous visual spans; FlashAttention/scaled dot-product attention (SDPA) equivalence is not assumed.

The LLaVA conversation wrapper preserves image order. Single-image conversion uses the stored question, removes the explicit-reference prefix, and changes "that image" to "this image," without adding labels. For example, "Answer using Image 1 only. Identify the main subject of that image" becomes "Identify the main subject of this image." Both retain the specific-common-name instruction. Full prompts accompany the data.

Question-token extraction requires a unique text match, allowing recorded leading-space/newline variants; a failed match raises an error. Gold labels are used for supervision/scoring, never as a prediction fallback. Parser and Oracle are separate evaluation branches.

E.4 Latency and measurement scope

Table 34 measures the recorded short-output implementation; the median answer is two tokens. Cold execution recomputes visual features for both passes. Warm execution precomputes visual features, without sharing decoder caches.

Table 34: Median latency (ms) from archived trials on an isolated A100-SXM4-40GB. Each method/cache condition uses one question from each of 24 independent groups, one measured repetition/group, and two warm-up questions. Compute Unified Device Architecture (CUDA) synchronization brackets image reading, processing/tokenization, graphics processing unit (GPU) transfer, planning, generation, and response decoding; model loading is excluded. These short-output timings do not estimate long-output throughput.

Method	Cold	Cached visual features
Baseline	375.6	354.5
Soft	702.0	661.0
Hard	694.3	651.1
Predicted Target-only	465.6	439.8
Parser Hard	392.1	370.6
Parser Target-only	143.6	129.4
Oracle Hard	392.1	369.4

E.5 Representative failures and task scope

Table 35 shows an answer error after correct selection, a wrong source choice, and damage to a correct answer. Their frequencies are given by the full paired counts, not these selected examples.

Table 35: Diagnostic failure examples. IDs abbreviate `DSB_Gxxxx_RC2_QS1`. "Solo" uses the target alone. Source probabilities are shown where recorded; all answers retain the frozen scorer. "Silver strainer" is an ordinary error, not a repair.

Case	Target / distractor	Reader choice	Baseline / Hard / Solo
G0042	Colander / tea strainer	Target	"Tea strainer" / "Silver strainer" / "Colander"
G0057	Image 1: comb; Image 2: fork	Image 2; $(0.1925, 0.8075)$	"Fork" / "Fork" / "Blue comb"
G0256	Image 1: red train; Image 2: green	Image 2; $(0.2104, 0.7896)$	"Red" / "Green" / "Red"

Hard constrains direct attention only in the specified prefill layers; earlier hidden states may already contain distractor information. The evaluated scope is explicit single-source answering, not implicit references or questions requiring joint evidence.

E.6 Inference Procedure and Controls

Figure 3(a) summarizes the two-pass execution. Given the complete multi-image input and question, Source-First Inference proceeds as follows: 1) preprocess the images and question, and locate each image's visual-token span; 2) run an unmodified planning forward with `use_cache=False`, compute $\mathbf{b}$, and select $\hat{s}$ using Equation 3; 3) start a fresh generation pass on the same multi-image input and apply hard source enforcement during prefill at layers 16–23; 4) generate the answer and continue cached decoding without further direct Hard edits.

The Soft and Predicted Target-only branches in Figure 3(a) provide execution controls under the same source prediction. Soft adds $\log(Nb_s)$ to the visual logits of source s before softmax, with all sources remaining accessible. Predicted Target-only uses the same selected source, removes the other images at the input, applies the fixed single-image question transformation, and reruns ordinary single-image generation. It serves as an input-isolation control under the same source prediction. Parser- and oracle-selected variants retain their respective source choices as additional references. Full comparisons appear in Section 5.

Panel (b) shows a qualitative real-image repair example. The reader assigns $b_1 = 0.935$ to the target Image 1, but the unmodified baseline answers "Dog" and assigns greater block-23 prefill attention to the distractor than to the target (10.48% versus 5.64%). Under Hard, the answer becomes "Cat"; target-image attention mass is 10.97%, while direct attention to the nonselected image is zero within the enforced window. The heatmaps compare separate Baseline and Hard runs, and overall repair and damage rates are reported in Section 5.