DISENTANGLING INTENT DISCLOSURE AND ATTRIBUTION IN MULTIMODAL VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting a generated image that fulfills a request requires checking what the image actually contains. Unified multimodal models combine image generation with visual understanding, offering a way to perform this verification within the same system. A key question is whether their judgments reflect the visible evidence or are influenced by the requested content and the image's stated source. We separate disclosure of the requested count (intent disclosure) from statements about the image's source (explicit source attribution) in a controlled study of Bagel on generated hand images. On 120 held-out images, intent disclosure produces the largest counting-accuracy change: accuracy falls from 77.50% to 37.50% with provenance unspecified. Self and other attribution each yield 35.83% accuracy, with six changed answers and metric-dependent score shifts. Selecting attention heads, individual components of the attention computation, for correction and setting their answer-processing outputs to zero improves intent-conditioned accuracy by 27.50 percentage points; comparison with attribution-based selection reveals distinct correction–capability trade-offs. All three intent-conditioned scoring variants reach the best achievable selection rate of 8/24 fixed candidate pools, each a set of images sharing a requested count. These findings distinguish the effects of intent and attribution, identify an effective correction criterion under the tested intervention, and show how answer accuracy and candidate selection capture complementary aspects of verification.

1 INTRODUCTION

A generated image can appear plausible while failing to satisfy a requested object count or spatial relation (Ghosh et al., 2023; Hu et al., 2023). Selecting useful outputs therefore requires assessing whether the requested content is actually present. Object-level checks and visual question answering address this requirement (Ghosh et al., 2023; Hu et al., 2023); complementary human-preference models, such as PickScore and ImageReward, provide scores for candidate ranking and model optimization (Kirstain et al., 2023; Xu et al., 2023). Unified multimodal models such as Bagel and Janus-Pro bring generation and visual understanding into a single system (Deng et al., 2025; Chen et al., 2025), creating an opportunity for a model to act as both generator and verifier. This setting raises a question about the verification process: how does the model weigh what is visible against what was requested, and does the image's stated source affect its judgment?

This concern is motivated by evidence that language priors and visual instructions can contribute to descriptions unsupported by the image (Rohrbach et al., 2018; Li et al., 2023). In generation verification, the requested content may therefore compete with the image as evidence. In the example in Figure 1, the request specifies seven fingers, but the image shows five. We call providing this target to the verifier *intent disclosure*. A separate input is *explicit source attribution*: a statement that the image was generated by the verifier itself or by another model. Distinguishing these inputs is essential for identifying whether an incorrect answer follows the requested content, responds to the stated source, or reflects both influences.

These two inputs connect our question to research on prompt following and self-attribution. Prompt-induced hallucination (PIH) describes answers that follow a prompt despite conflicting visual evidence (Rudman et al., 2026), while work on self-attribution investigates how the perceived source of an action affects evaluation by a monitor, a model that assesses actions or outputs (Khullar et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

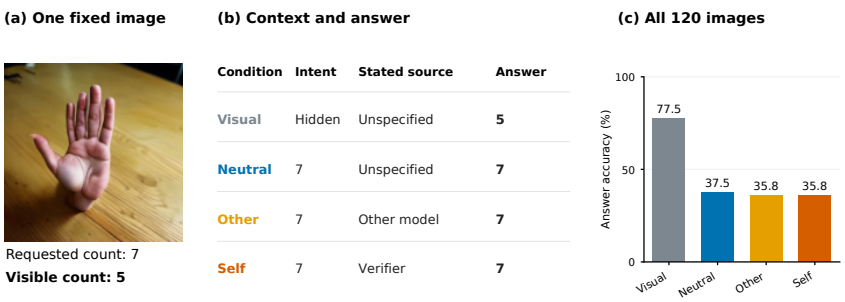


Figure 1: Separating intent disclosure from explicit attribution. (a–b) In this post-hoc illustration, a Janus-Pro image shows five digits (including the thumb) despite a request for seven. Bagel answers five without the target and seven in all three intent-disclosure conditions. Prompt and example-selection details appear in Appendices A and D. (c) Complete-answer accuracy across all 120 held-out images drops with intent disclosure; self and other attribution yield the same observed accuracy. Paired effects and uncertainty appear in Figure 2.

2026). These findings motivate a matched visual comparison in which target information and source wording can be varied separately. Our central question is how these two inputs affect three stages of verification: the model's numerical scores, its complete answers, and the candidate images selected using those scores.

These stages answer different practical questions. A score records how strongly the model favors a candidate count; a complete answer determines whether it reports the visible count correctly; candidate selection tests whether its scores identify a suitable image among available alternatives. A score shift need not change an answer or reverse a ranking. Measuring all three on the same images connects sensitivity to the supplied information with its consequences for verification.

We study this question with Bagel as the verifier and a fixed collection of Janus-Pro-generated hand images. Every image is evaluated under four conditions: visual-only, disclosed intent with unspecified provenance, disclosed intent attributed to another model, and disclosed intent attributed to the verifier. Image encoding and the counting question are held fixed. For intervention analysis, a discovery split selects attention heads (parallel components of the transformer's attention computation (Vaswani et al., 2017)), and a separate evaluation split measures the effects of setting selected head outputs to zero during answering.

With this design, our study makes three contributions. **(1) Intent–attribution separation:** matched conditions distinguish a large intent-disclosure effect from smaller, metric-dependent attribution effects in Bagel's counting. **(2) Correction–capability assessment:** correction-based head selection improves intent-conditioned accuracy with less visual capability cost (loss in visual-only accuracy) than attribution-based selection. **(3) Decision-level evaluation:** linking scores, answers, and candidate selection shows that all three intent-conditioned scores attain the best achievable selection in fixed candidate pools (sets of images sharing a requested count).

2 Related Work

Generated-image verification. Object-level evaluation and visual question answering test whether generated images satisfy a prompt (Ghosh et al., 2023; Hu et al., 2023; Lin et al., 2024), while preference models support ranking by human judgments (Xu et al., 2023; Kirstain et al., 2023). We study count verification, distinguishing correctness of the verifier's answer from success in selecting a candidate image.

Intent and source attribution. Rudman et al. (2026) study prompt-induced hallucination and identify attention heads whose ablation, an intervention that disables their contribution, reduces prompt-following errors. Khullar et al. (2026) examine self-attribution in language-model monitors, distinguishing effects of interaction history from explicit ownership statements. We compare intent

disclosure and explicit source wording on the same images, with the target held fixed for the self–other contrast. This connects prompt-following errors to a controlled test of provenance wording while keeping the comparison distinct from genuine generation history.

Interventions and evaluation criteria. Head pruning disables selected attention heads, while activation patching replaces internal values with those from a reference run; both motivate specifying the intervention and evaluation metric (Michel et al., 2019; Voita et al., 2019; Heimersheim & Nanda, 2024). Evaluating output distributions extends assessment beyond single responses. Calibration tests whether predicted confidence matches observed correctness (Scholten et al., 2025; Guo et al., 2017). We assess selected heads by both correction and retained visual capability, and compare scores, complete answers, and candidate choices on held-out images. Appendix F provides the full discussion of these research strands and additional grounding and interpretability literature.

3 Controlled Verification Protocol

The protocol separates the effects of disclosing a requested count and stating an image source while keeping visual input fixed. We first describe the images and data split, then define the matched conditions and the score, answer, and uncertainty measures used throughout the study.

3.1 Images, labels, and data separation

The evaluated collection was constructed for this study: 200 hand images generated with Janus-Pro-7B (Chen et al., 2025), with requested counts in $\{4, 5, 6, 7\}$. It is a generated evaluation collection rather than an existing named benchmark. We use its corresponding manually supplied visible-count labels. A manifest, or per-image metadata list, records each image path, requested count, label, and split. The labeling code applies a strict majority rule to the specified available annotator columns and excludes missing or tied labels. This run uses one resolved count column; the rule therefore does not imply multiple independent annotations for every image. We denote the requested count by t_i and the labeled visible count by v_i. Group A contains images for which $t_i = v_i$; group B contains conflicts, $t_i \neq v_i$. Agreement with the request is therefore distinguished from correctness about the image.

The split groups images sharing a seed/template identifier, so that such a group cannot cross discovery and evaluation. The resulting discovery set has 80 images, including 67 group-B images; evaluation has 120 images in 30 groups, with 21 A images and 99 B images. Each requested count contributes 30 evaluation images. Candidate pools and the split are constructed by deterministic hashing rather than by verifier scores or correctness. Label agreement in this collection is not used to claim that image truth is infallible; the reported conclusions remain conditional on the supplied labels.

3.2 Matched conditions

We use the Bagel checkpoint `ByteDance-Seed/BAGEL-7B-MoT` (Deng et al., 2025); MoT denotes its mixture-of-transformer-experts architecture. Images enter through the vision transformer (ViT) understanding branch (Dosovitskiy, 2020). The variational autoencoder (VAE) image branch is disabled (Kingma & Welling, 2013). Thus, ViT-only refers to a fixed image-input route, not a different verifier. Every image is evaluated in all four conditions in Table 1. The first prompt line either discloses the numeric generation target or gives a generic hand request. The next line specifies provenance, followed by an identical counting question. The contrast changes both information and wording; it does not isolate an abstract cognitive concept independently of its textual realization. Exact strings are in Appendix A.

We define the intent-disclosure contrast as neutral minus visual, and the explicit-attribution contrast as self minus other. The former tests adding a target while leaving provenance unspecified. The latter changes the stated source with target information held constant. Self minus neutral is a supplementary contrast, not a replacement selected after observing results. Because all images in this evaluation have an external source, the self condition is a framing manipulation rather than a truthful record of Bagel authorship. No real generation-history effect is identified by this comparison.

Table 1: Controlled inputs. Image, encoder, insertion position, and counting question are fixed. All numeric-intent conditions use the same target t_i.

Condition	Requested count disclosed	Provenance statement
Visual	No (generic hand request)	Unspecified
Neutral	Yes	Unspecified
Other	Yes	Generated by another AI model
Self	Yes	Generated by you

3.3 Scores, answers, and uncertainty

Let $z_i^c(d)$ be the first-answer-position logit (the unnormalized score before probability normalization) for count token d in condition c. The implementation retains distinct single-token representations of integers from zero through ten and verifies that the required target and visual counts are represented. For this candidate vocabulary D, define

$$q_i^c(t_i) = \frac{\exp z_i^c(t_i)}{\sum_{d \in D} \exp z_i^c(d)}, \qquad \ell_i^c = z_i^c(t_i) - z_i^c(v_i). \tag{1}$$

The first quantity is conditional on the candidate-count vocabulary, not the full language-model vocabulary. The second is the log ratio between target and visual-count probabilities. We average their paired differences on group B; group A would have identically zero target/visual log-odds. We separately record full-vocabulary probability mass on D.

Complete answers are greedily decoded, selecting the highest-scoring token at each step, with a maximum length of 64 tokens, using the same first-answer state as score extraction. A runtime check requires the highest-scoring token and the generated first token to agree. A conservative parser accepts bare counts and a limited set of unambiguous count statements. Unparsed answers remain in accuracy denominators and do not count as target acceptance. Let $\hat{v}_i^c$ denote the count parsed from the complete answer, with an unparsed answer treated as incorrect. We report answer accuracy $\mathrm{Acc}_c = \mathbb{E}[\hat{v}_i^c = v_i]$ over all evaluation images and the false-acceptance rate (FAR), $\mathrm{FAR}_c = \mathbb{E}[\hat{v}_i^c = t_i \mid i \in B]$ over conflicts. The area under the receiver operating characteristic curve (AUROC) measures how well $q_i^c(t_i)$ ranks intent-aligned images (A, the positive class) above conflicts (B, the negative class); candidate selection uses the same score within each fixed pool.

Paired cluster-bootstrap uncertainty estimates resample entire seed/template groups with replacement, retaining all conditions for each resampled group. Original contrast intervals use 1,000 bootstrap draws; supplementary accuracy, random-control, and AUROC checks use 5,000. All confidence intervals (CI) reported here are nominal 95% intervals without a multiple-comparison adjustment. The three sampled random sets for each intervention remain fixed during bootstrap resampling. These intervals capture image-group uncertainty conditional on those sets, rather than uncertainty over all possible random head selections.

4 Behavioral Effects of Intent and Attribution

We examine how intent disclosure and explicit attribution affect verification on the same held-out images. The analysis starts with complete-answer accuracy and wrong-target acceptance, then compares probability, log-odds, and ranking changes to establish which effects translate into observable errors.

4.1 Intent disclosure produces substantial visual counting errors

Providing a numeric target substantially impairs counting in this protocol. Table 2 reports all conditions on the same evaluation images. Accuracy falls from 93/120 in the visual condition to 45/120 in the neutral condition, a paired decrease of 40.00 percentage points (95% CI $[-50.00, -30.83]$). Across individual images, 58 initially correct answers become incorrect, while ten initially incorrect answers become correct. The net decrease therefore reflects actual answer errors, rather than a change limited to a confidence score.

The errors increasingly match the disclosed target. On group B, erroneous target acceptance rises from 5/99 without the numeric target to 57/99 with neutral disclosure, a difference of 52.53 points (CI

$[43.16, 61.06]$). The target is defined offline even when it is not shown to the verifier, so the visual condition supplies a comparable baseline for target-matching mistakes. The larger target-acceptance increase than overall accuracy decrease is not contradictory: target acceptance is measured on B, and a wrong target can replace an already wrong answer.

The neutral/visual contrast establishes a large intent-disclosure effect with provenance unspecified. Low accuracy in both attributed conditions therefore cannot by itself identify authorship as the cause. The self/other contrast tests the additional effect of source wording.

4.2 Explicit attribution changes some outputs without increasing aggregate errors

Self and other attribution each produce 43/120 correct answers, or 35.83% accuracy. The paired difference is zero points with CI $[-2.50, 2.50]$. This is not an equivalence test and does not establish that attribution has no effect. It shows that the observed sample contains no net increase in answer errors for self relative to other.

This aggregate equality masks changes in individual answers: the complete answers differ on six images. Four changes replace a wrong non-target count under other attribution with the wrong target under self attribution. One change is correct-to-incorrect and another is incorrect-to-correct. Consequently, B-group target acceptance increases by four cases, from 58/99 to 62/99, while the two correctness transitions offset each other. The target-acceptance difference is 4.04 points with CI $[0.00, 8.91]$. Calling this "four extra errors" would conflate the content of an error with whether an error occurred.

The neutral comparison helps expose dependence on the reference condition. Self minus neutral increases B-group target acceptance by 5.05 points (CI $[0.99, 10.53]$), while its accuracy difference is -1.67 points (CI $[-4.17, 0.00]$). These are useful supplementary measurements, but they answer a different question from self minus other. An unspecified source may differ from either explicit source statement. Reporting all conditions avoids choosing the reference that produces the strongest apparent attribution effect and then treating that effect as reference-independent.

4.3 Probability and log-odds changes have different meanings

Beyond these answer changes, attribution also affects scores, with directions that depend on the metric. On B, self minus other increases the conditional target probability by 0.00812 (CI $[0.00035, 0.01536]$), but decreases target/visual log-odds by 0.07765 (CI $[-0.11800, -0.03762]$). Figure 2 places the two contrasts on separate metric scales. In comparison, neutral minus visual increases target probability by 0.42170 and log-odds by 5.82576. These values indicate a large intent-disclosure shift and a much smaller, metric-dependent explicit-attribution shift.

The definitions explain why the score directions need not agree. Conditional target probability depends on the target relative to all retained count tokens. Target/visual log-odds compares only two candidates. For instance, probability mass can move away from other wrong counts toward both the target and the visual answer; target probability can rise even if the visual answer rises more. Across images, averaging further mixes different initial distributions and saturation levels. This is a mathematical possibility, not a fitted explanation of every observed transition; the data establish the metric disagreement, rather than one uniquely identified redistribution mechanism.

The probability is also not a direct measure of full-answer confidence. Average full-vocabulary mass on the retained count tokens is 0.8336 in the visual condition, compared with 0.9512–0.9623 in the three numeric-intent conditions. Conditioning on this vocabulary removes changes in the total mass allocated outside it. First-token statistics do not encode the complete continuation or its parseability. Retaining decoded answers is therefore necessary even when a conditional score is convenient for paired comparisons or ranking. Treating all these measurements as one scalar "anchoring strength" would erase distinctions that are visible in the current results.

4.4 Ranking sensitivity does not by itself establish behavioral harm

Self attribution has higher pooled AUROC than other attribution: 0.9149 versus 0.8961, a difference of 0.01876 with supplementary CI $[0.00433, 0.03916]$. That result concerns discrimination between

Table 2: Baseline outcomes. Accuracy uses all 120 images; wrong-target acceptance uses 99 conflict images. AUROC ranks intent-aligned versus conflicting images using the candidate-count conditional target score.

Condition	Accuracy (%) ↑	Wrong target (%) ↓	AUROC ↑
Visual	77.50	5.05	0.8913
Neutral	37.50	57.58	0.9259
Other	35.83	58.59	0.8961
Self	35.83	62.63	0.9149

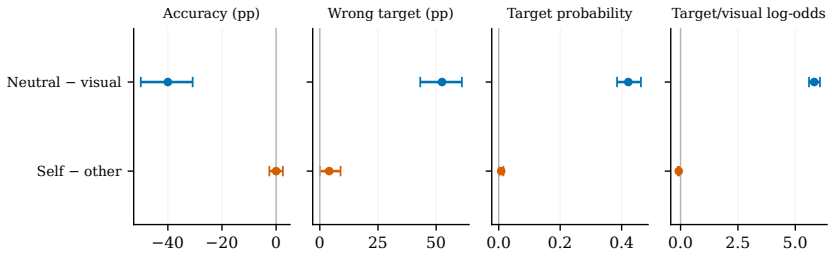


Figure 2: Paired effects with nominal 95% cluster-bootstrap intervals (1,000 draws). Accuracy uses 120 images; other panels use the 99 B images. Each panel has its own scale. A positive probability difference and a negative target/visual log-odds difference need not describe the same change in the output distribution.

the A and B images in this collection. It is compatible with equal answer accuracy because ranking images by target score and outputting the correct visible count are different tasks. Neutral attribution has an even higher AUROC of 0.9259 despite low counting accuracy, further illustrating that the ranking and answer criteria can diverge.

The AUROC difference depends on score changes across positive and negative examples, rather than modality alone. This distribution-specific ranking result motivates the candidate-selection test in Section 6; it does not establish closed-loop reliability.

5 Interventions: Correction and Capability Cost

Having separated intent and attribution effects behaviorally, we next test whether attention-head interventions can correct intent-driven errors while preserving visual counting ability. Heads are selected on discovery data using separate correction and attribution objectives, then evaluated on held-out images against matched random controls. Shared and exclusive subsets help distinguish the resulting effects.

The neutral and visual conditions provide complementary tests of an intervention. Neutral accuracy measures counting when the requested target is available; visual-only accuracy measures counting without that numeric cue. Because both conditions use the same images and labels, their paired changes show whether improved performance under intent disclosure is accompanied by a loss in ordinary visual counting.

5.1 Functional selection on discovery data

Functional selection ranks heads by their measured effect under intervention. Candidate screening first narrows the search using the 67 discovery conflicts. At the first answer position, the code captures head-sized slices of the input to each attention output projection. For each head, it computes the norm of the mean activation difference for self versus other and for neutral versus visual. Taking the top 40 heads for each contrast gives a union of 61 candidate heads in this run. This is a restricted search space, not an exhaustive evaluation of every head or a claim that activation magnitude establishes causal importance.

Each candidate is then zero-ablated individually on all four discovery conditions by setting its output slice to zero during answer processing. Two functional scores produce separate top-20 sets. The attribution score is the reduction in the signed self-minus-other log-odds contrast:

$$s_{\mathrm{attr}}(h) = \mathbb{E}_{i \in B_{\mathrm{disc}}} \left[(\ell_i^{\mathrm{self}} - \ell_i^{\mathrm{other}}) - (\ell_{i,-h}^{\mathrm{self}} - \ell_{i,-h}^{\mathrm{other}}) \right]. \tag{2}$$

Here B_{disc} is the discovery conflict subset, and the subscript $-h$ denotes zero-ablation of head h. The PIH score is the fraction corrected by deleting h among discovery images on which the intact neutral condition outputs the wrong target. Heads are ranked by descending score, with layer and head index breaking ties. Selecting a large value does not impose positive held-out performance or identify a semantic role independently of the scoring rule.

We call the two sets attribution-selected and PIH-selected; the archived code calls the former `anchor`. Their intersection contains nine heads, leaving eleven exclusive heads in each set. We evaluate these five sets and three size- and layer-matched random sets for each one, yielding twenty intervention sets. Random controls exclude the corresponding target set. All functional selection occurs on discovery data; the effects reported below use the held-out evaluation images. The random comparison is conditional on this candidate screen, selection procedure, and collection of controls.

Deletion is restricted to answer processing. The context is built before intervention hooks (functions that modify internal values during model execution) are installed, and selected head slices are set to zero before the attention output projection at every decoded answer position. Image and prompt prefill states, which encode the context before answer generation, are not recomputed under ablation. The same intervention is used for score extraction and decoding. This operator tests dependence on selected answer-processing contributions; it does not erase the heads throughout all earlier computation, replace them with mean activations, or remove information from the already constructed context.

5.2 Correction-selected heads improve held-out answers

PIH-selected ablation increases neutral-condition accuracy from 37.50% to 65.00%, a gain of 27.50 points (CI $[18.33, 36.67]$). B-group wrong-target acceptance decreases from 57/99 to 10/99. The accuracy gain is the direct evidence of correction: the acceptance decrease alone would also allow a model to replace the target with another wrong count. Reporting both measurements establishes that this selected set changes the error content and improves the evaluated answers.

To assess the role of targeted selection, the three matched PIH random sets provide a comparison, with neutral accuracies of 44.17%, 37.50%, and 44.17%. PIH-selected ablation exceeds their fixed mean by 23.06 points (CI $[15.28, 30.83]$). This comparison supports a benefit beyond these particular matched deletions. It does not estimate performance relative to every possible random set or establish superiority over a separately implemented mitigation method. The result is evidence for the usefulness of the discovery correction criterion within the evaluated intervention class.

Correction remains incomplete. Even after PIH-selected deletion, neutral-condition accuracy is 65.00%, below the intact visual-only accuracy of 77.50%. The intervention therefore does not recover a fully reliable verifier, and its benefit should not be described as eliminating prompt-induced error. The comparison also does not imply that dropping intent is always the preferred deployment policy: some verification tasks require the request to judge compliance. Our counting setting separates the effect of supplying that information from the ability to read a visible property.

5.3 Attribution-selected deletion incurs a visual capability cost

Attribution-selected deletion raises neutral accuracy to 51.67%, but lowers visual-only accuracy from 77.50% to 64.17%. The visual change is -13.33 points with CI $[-19.17, -7.50]$. Its neutral wrong-target acceptance falls sharply to 16.16%, yet the correction benefit is smaller than the drop in that error category. This is precisely why a reduction in target following cannot be treated as sufficient evidence of a successful repair.

PIH-selected deletion has a smaller observed visual cost: accuracy becomes 75.83%, a change of -1.67 points with CI $[-4.17, 0.00]$. This is not evidence of exact preservation or a test showing that the capability loss stays within a prespecified acceptable margin. Relative to the mean of its three fixed random controls, PIH-selected visual accuracy is lower by 4.72 points (CI $[-8.89, -1.11]$). The

Table 3: Held-out effects of selected head sets. Neutral and visual accuracy use 120 images; neutral wrong-target acceptance uses 99 B images. Set names describe selection rules, not independently established semantic circuits. Full matched random-control results appear in Appendix B.

Set	Heads	Neutral acc. (%)	Visual acc. (%)	Neutral wrong (%)
Intact model	0	37.50	77.50	57.58
Attribution-selected	20	51.67	64.17	16.16
PIH-selected	20	65.00	75.83	10.10
Shared	9	55.00	73.33	15.15
Attribution-only	11	37.50	75.00	58.59
PIH-only	11	46.67	79.17	43.43

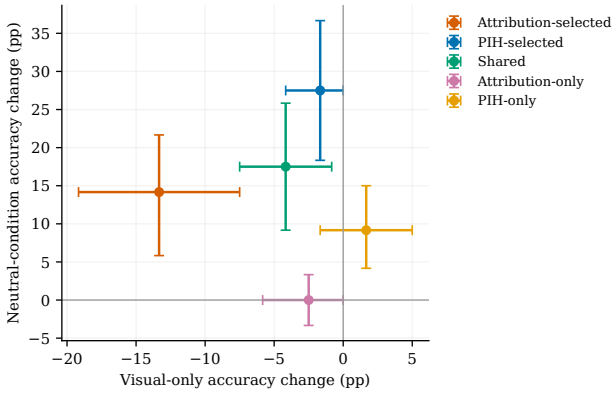


Figure 3: Correction and capability changes relative to the intact model. Vertical position measures neutral-condition correction; horizontal position measures visual-only capability change. Error bars are separate marginal 95% cluster-bootstrap intervals (5,000 draws), not joint confidence regions. Attribution-selected deletion has a substantial visual cost.

selected intervention's neutral-condition advantage and its visual-only cost must both remain visible when judging practical value.

These contrasting effects motivate the joint display in Figure 3, which keeps correction and capability separate. An intervention may be useful for a disclosed-intent workload while being harmful for ordinary counting, and the present data do not specify the frequency or value of these workloads in deployment. We consequently report the two coordinates rather than claiming a universally favorable trade-off. No general-purpose capability benchmark was evaluated, so visual-only counting is a task-matched capability check, not a comprehensive utility guarantee.

5.4 Shared and exclusive subsets do not establish an independent circuit

The two top-20 sets have Jaccard overlap $9/31 \simeq 0.290$, defined as intersection size divided by union size. Deleting their shared nine heads raises neutral accuracy to 55.00%, while deleting the attribution-only eleven leaves it at 37.50%. Deleting the PIH-only eleven raises it to 46.67%, a gain of 9.17 points with CI $[4.17, 15.00]$. These results show partially distinct intervention effects and a substantial shared component under this operator. Because the network is nonlinear, these subset effects do not form an additive decomposition of the full-set effect.

The signed attribution objective also requires care. The intact held-out self-minus-other log-odds contrast is already negative, at -0.07765. Attribution-selected deletion changes it to approximately -0.47222. A positive reduction under the selection convention therefore means a more negative contrast, not movement toward zero. Its absolute magnitude grows. Describing that signed change as removal of a positive self bias would reverse the meaning of the measured quantity.

Nor does the same deletion collapse the AUROC gap. The intact self-minus-other gap is 0.01876; it becomes 0.02597 after attribution-selected deletion and 0.02068 after shared-set deletion. PIH-

selected deletion yields 0.00529, but this is a different set and a smaller between-condition gap is not, by itself, improved reliability. The attribution-only subset leaves self and other aggregate accuracies at 35.83% and does not remove their excess target-acceptance difference. The present intervention results thus support correction-related head dependence and partial functional separation, but not an independently established harmful self-attribution mechanism.

6 From Scores to Candidate Selection

Returning to intact-model scores, we test whether they can select an image that satisfies the request, a distinct decision from counting each image correctly. We construct 24 disjoint pools of five evaluation images sharing a requested count, with the partition fixed independently of correctness. Each condition selects the image with the highest conditional target score; ties receive the average correctness of tied candidates. Success means selecting an image whose labeled count matches the request.

Self, other, and neutral scoring each succeed in 8/24 pools (33.33%), compared with 7/24 (29.17%) for visual scoring and an expected 17.50% for uniform selection. All three intent-conditioned scores attain the available oracle: only eight pools contain any correct image. Thus, despite the counting-accuracy loss under intent disclosure, these scores select a correct candidate whenever one is available in the current pools. Explicit self attribution produces no measured selection penalty here.

This outcome depends on the candidate distribution. Of the 21 correct images, nineteen have requested count four and two have requested count five; counts six and seven have no correct candidates. The result therefore establishes useful selection within these fixed pools, while leaving discrimination in the latter categories untested. Pooled AUROC cannot resolve that gap. Extending the decision-level claim requires pools with correct and incorrect candidates in each target category; a claim about iterative refinement additionally requires evaluating the update rule and resulting images.

7 Discussion and Limitations

The results distinguish intent-driven counting errors from the smaller, metric-dependent effects of explicit attribution. Correction-selected heads improve disclosed-intent answers, whereas attribution-selected heads incur a visual capability cost. These findings motivate measuring scores, complete answers, and selection decisions, as well as correction and retained capability.

The estimates apply to one Bagel checkpoint, Janus-Pro hand images, ViT-only encoding, one prompt construction, and answer-position zero-ablation. Attribution concerns explicit source wording rather than genuine generation history. Selection success is conditional on the fixed candidate pools, whose correct images occur in two target categories. Intervention comparisons are conditional on the candidate-head screen and three fixed random sets per target; confidence intervals are nominal and unadjusted for multiple comparisons.

Broader evaluations can test whether these distinctions persist across models, tasks, and candidate distributions. Parsing cases, earlier protocols, and the scope of independent verification and reproducibility are documented in Appendices C, E, and D, respectively.

8 Conclusion

Our controlled study of Bagel on Janus-Pro hand images distinguishes a large effect of intent disclosure from smaller, metric-dependent effects of explicit source attribution. Correction-selected attention heads improve disclosed-intent accuracy by 27.50 percentage points, with a smaller observed visual capability cost than attribution-selected heads. At the decision level, all three intent-conditioned scoring variants achieve the available oracle of 8/24 fixed candidate pools. These findings identify a useful correction objective and show the complementary value of answer-level and selection-level evaluation. Assessing scores, complete answers, and candidate choices, while measuring both correction and retained capability, provides a concrete basis for evaluating multimodal verification.

AI USE STATEMENT

In this work, we used generative AI tools for grammar correction and rewriting. We have not used generative AI tools for generating synthetic numerical results to substitute for measured experimental outcomes, and AI-assisted recruitment are not applicable to this work. Additionally, we used generative AI tools for literature discovery, manuscript organization and language editing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International conference on learning representations*, volume 2024, pp. 15625–15645, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, volume 2025, pp. 77965–78000, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

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

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020.

Dipika Khullar, Jack Hopkins, Rowan Wang, and Fabien Roger. Self-attribution bias: When ai monitors go easy on themselves. *arXiv preprint arXiv:2603.04582*, 2026.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Ritambhara Singh, Carsten Eickhoff, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision-language models. *arXiv preprint arXiv:2601.05201*, 2026.

Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 49420–49436, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 98891–98913, 2025.

A EXACT PROMPTS AND INTERVENTION IMPLEMENTATION

All conditions end with: "How many fingers are visible in this image? Answer with a single number." Neutral, other, and self begin with "The image was requested to show exactly t fingers." Visual begins with "The image was requested to show a hand." The provenance line is respectively "Image provenance: unspecified.", "Image provenance: generated by another AI model.", or "Image provenance: generated by you." Visual and neutral share unspecified provenance. Prompt text is inserted through the same Bagel context-building path after the same image representation.

For score extraction, the adapter checks that every target and labeled count has a distinct single-token candidate. It separately stores raw count logits, normalized count scores, the total full-vocabulary mass assigned to these tokens, first-token identity, the entire generated answer, and parser status. Greedy decoding uses `do_sample=False`; the configured automatic mixed-precision computation uses bfloat16, a 16-bit floating-point format. The first-answer score pass and decoding pass both start from the same constructed context. A mismatch between their first selected token triggers a runtime error.

At each selected layer, the hook acts on the input to `o_proj`, the attention output-projection layer. If the head dimension is d_h, head h corresponds to the slice $[hd_h, (h+1)d_h)$. This slice is replaced with zero for the current answer queries. Context prefill occurs before installing the hook. Deletion is applied throughout answer decoding, while activation screening records only the first answer position. This implementation is different from deleting a head at every prompt and image position, and from replacing its activation with a dataset mean.

B ALL MATCHED RANDOM CONTROLS

Each random control matches the number of selected heads in each layer and excludes its corresponding target set. Seeds are 20260921, 20260922, and 20260923. The table reports each realized set separately; comparisons to their average hold these choices fixed when resampling image groups.

Table 4: Complete random-control outcomes, using the same metrics and denominators as Table 3. Code names `anchor` and `pih` denote the attribution-selected and PIH-selected sets.

Control	Heads	Neutral acc. (%)	Visual acc. (%)	Neutral wrong (%)
anchor_random_0	20	48.33	75.83	31.31
anchor_random_1	20	50.83	79.17	35.35
anchor_random_2	20	35.00	78.33	67.68
pih_random_0	20	44.17	80.83	45.45
pih_random_1	20	37.50	85.00	62.63
pih_random_2	20	44.17	75.83	35.35
shared_random_0	9	48.33	75.83	37.37
shared_random_1	9	43.33	77.50	46.46
shared_random_2	9	36.67	78.33	60.61
anchor_only_random_0	11	30.00	78.33	79.80
anchor_only_random_1	11	25.00	67.50	86.87
anchor_only_random_2	11	41.67	75.00	48.48
pih_only_random_0	11	37.50	79.17	60.61
pih_only_random_1	11	40.83	75.00	52.53
pih_only_random_2	11	42.50	78.33	47.47

C ANSWER PARSING AND UNRESOLVED CASES

The original audit verified 26,748 records: 800 intact-model records over discovery and evaluation, 16,348 single-head records ($61 \times 67 \times 4$), and 9,600 set-intervention records ($20 \times 120 \times 4$). Of these, 26,746 were automatically parsed and two required review. Both unresolved answers occur in the PIH-selected visual condition and state: "There are four fingers and a thumb visible in the image." Both images have supplied label four. Treating the statement as a total count of five would leave both answers incorrect, so visual accuracy remains 91/120. One answer would newly match an incorrect

target, raising visual B-group target acceptance from 5/99 to 6/99. The published table here retains the original conservative parsing rather than silently relabeling the records.

D AUDIT SCOPE AND EXPLORATORY MATERIAL

Teaser example selection. The example in Figure 1 (`int7_tpl0_seed005`) was selected post hoc by a fixed trial-ID hash order among evaluation conflicts with a correct visual-only answer and target-following answers in all three numeric-intent conditions. It illustrates this response pattern rather than constituting a representative sample. The figure abbreviates prompt fields; exact strings appear in Appendix A.

The independent audit reproduced the complete analysis summary from the supplied records, reconstructed functional rankings, and checked recorded manifest fields and hashes. It did not rerun model inference, inspect the full image collection, or reconstruct missing activation dumps. The supplied result package also omits the upstream model revision, weight hash, and full environment lock. These missing materials limit fresh end-to-end reproduction from that package; the reported independent checks concern the saved records. Three original images were subsequently supplied and visually checked for illustrative figure selection; this does not constitute a new labeling of the evaluation set. A separate local evidence map records the source of each main claim and the status of limitations. Appendix E additionally reconstructs selected earlier behavioral and ablation summaries. The earlier model taxonomy, cross-task comparisons, and localization of effects to individual activation coordinates remain outside the evidence used for the main conclusions. They are not pooled with the matched-protocol estimates or used to claim nine-model confirmation of the present results. No pending attribution-v2 result is included in this paper.

E COMPLEMENTARY RESULTS UNDER THE EARLIER VERIFICATION PROTOCOL

Scope and measurement. Earlier experiments provide a broader descriptive comparison of verification scores and illustrate how different intervention objectives can diverge. The behavioral records cover nine multimodal verifiers on 200 Janus-generated hand images, including 166 conflicts between the labeled count and the requested count. The earlier self and other conditions, denoted C3 and C4, use the same Janus image but model-specific context constructions. In Bagel, the self condition uses a generation-style context with variational autoencoder (VAE) and ViT image representations, whereas the other condition uses stateless verification, evaluating the image without a retained generation context. These observations therefore characterize the earlier protocol; they do not isolate the explicit source-wording contrast of the main study. We report them separately from its held-out estimates.

Let $\widetilde{q}_i^c$ denote the saved candidate-count distribution in this earlier protocol. We distinguish the probability contrast

$$D_P = \frac{1}{|B|} \sum_{i \in B} \left[\widetilde{q}_i^{C3}(t_i) - \widetilde{q}_i^{C4}(t_i) \right] \tag{3}$$

from the mean log-odds contrast $D_L = \frac{1}{|B|} \sum_{i \in B} (\widetilde{m}_i^{C3} - \widetilde{m}_i^{C4})$, where $\widetilde{m}_i^c = \log \widetilde{q}_i^c(t_i) - \log \widetilde{q}_i^c(v_i)$, and from the ranking difference $\Delta_A = \mathrm{AUROC}_{C3} - \mathrm{AUROC}_{C4}$. These quantities measure target-score shifts, intent-versus-visual evidence, and between-image discrimination, respectively.

Score sensitivity varies across verifiers. Table 5 shows a range of responses to the earlier self/other constructions. Bagel, InternVL2.5, DeepSeek-VL2, and Janus-Pro have positive probability contrasts between 0.06511 and 0.10561; mPLUG-Owl3 and Qwen2.5-VL have smaller positive contrasts. Emu3 has a negative contrast, while LLaVA-OneVision (abbreviated LLaVA-OV in the table) is close to zero and Anole has a wider interval spanning zero. This variation motivates evaluating the score response for each verifier and protocol, without treating its sign as an invariant model category or as a measure of complete-answer errors.

Mean contrast and mean relative reduction answer different questions. The earlier set-ablation records show a clear reduction of Bagel's mean log-odds contrast: D_L changes from 1.09548 to 0.60008 on 163 recorded conflicts (Table 6), a 45.22% reduction of the mean. Averaging the individual relative reductions instead gives 42.90%. For Emu3, the mean signed contrast changes

Table 5: Earlier-protocol probability contrasts D_P on 166 Janus-image conflicts per verifier. Intervals use 2,000 paired-trial percentile bootstrap draws with seed zero; they are descriptive, unadjusted intervals, not the cluster-bootstrap intervals used in the main study.

Verifier	D_P	95% CI
Bagel (Deng et al., 2025)	$+0.10561$	$[+0.07741, +0.13418]$
InternVL2.5 (Chen et al., 2024)	$+0.08434$	$[+0.07376, +0.09560]$
DeepSeek-VL2 (Wu et al., 2024)	$+0.07879$	$[+0.06454, +0.09263]$
Janus-Pro (Chen et al., 2025)	$+0.06511$	$[+0.03710, +0.09541]$
mPLUG-Owl3 (Ye et al., 2025)	$+0.01319$	$[+0.00952, +0.01652]$
Emu3 (Wang et al., 2024)	-0.02014	$[-0.02632, -0.01423]$
Qwen2.5-VL (Bai et al., 2025)	$+0.00869$	$[+0.00333, +0.01448]$
LLaVA-OV (Li et al., 2024)	-0.00048	$[-0.00100, -0.00006]$
Anole (Chern et al., 2024)	-0.01241	$[-0.02364, +0.00082]$

from -0.01211 to -0.12694, moving farther from zero, even though the average individual relative reduction is positive (31.56%). For per-image contrasts d_i, the distinction follows from

$$1 - \frac{\mathbb{E}[d_i^{\text{ablated}}]}{\mathbb{E}[d_i^{\text{clean}}]} \neq \mathbb{E}\left[1 - \frac{d_i^{\text{ablated}}}{d_i^{\text{clean}}}\right] \tag{4}$$

in general. We therefore report the actual condition means rather than reconstructing them from a relative-reduction summary. Emu3's near-zero clean mean also makes a percentage reduction of that mean an unsuitable effect summary.

A smaller ranking gap can accompany different absolute outcomes. Earlier AUROC ablations distinguish convergence between conditions from an improvement in discrimination. For Bagel, the self-minus-other gap changes from 0.10276 to 0.00354, a 96.55% reduction using the baseline summary from the same result package. This change combines a small decrease in self AUROC with an increase in other AUROC. For Emu3, the gap falls by 71.13%, but both absolute AUROCs decrease (Table 7). Thus, reporting both absolute scores identifies whether a smaller gap reflects improvement, degradation, or a mixture of the two.

The probability contrast in these same AUROC-ablation records further distinguishes the objectives. On the 166 conflicts, Bagel's top-20 deletion leaves $D_P = 0.08960$, compared with 0.07278 averaged over its ten random sets. These descriptive means do not show a smaller residual probability contrast for the selected set, despite its smaller AUROC gap. A set's effect on ranking discrimination consequently cannot substitute for measuring its effect on the within-image score contrast.

Interpretive limits of the earlier records. The log-odds ablation files contain 163 Bagel and 129 Emu3 conflicts rather than all 166; the final tables do not establish the exclusion mechanism. The available common head lists contain 20 heads per model, but run-specific snapshots are absent, so they do not establish that the log-odds and AUROC runs used identical sets. The clean per-image scores underlying the AUROC baseline summaries are also unavailable in this result package; a separate earlier Bagel probability table yields a slightly different clean AUROC. We therefore retain the paired descriptive outcomes without pooling versions or inferring a shared causal mechanism across experiments. The main study addresses correction and capability directly under its matched protocol.

F Extended Related Work

Evaluating generated images. GenEval tests compositional properties such as object count and position (Ghosh et al., 2023). TIFA (Text-to-Image Faithfulness evaluation with question Answering) uses visual question answering to assess prompt faithfulness (Hu et al., 2023); Davidsonian Scene Graph (DSG) organizes dependent questions to reduce inconsistencies (Cho et al., 2024), while VQAScore uses answer probabilities for image–text alignment (Lin et al., 2024). Human-preference approaches target a complementary criterion: ImageReward learns preference rewards, Pick-a-Pic supplies preference data for PickScore, and Human Preference Score v2 (HPS v2) predicts

Table 6: Recorded mean log-odds contrasts D_L under the earlier set-ablation protocol. Random-20 averages ten sets for Bagel and one recorded set for Emu3. Each model's random records match its top-20 sample IDs.

Verifier	Recorded n	Clean	Top-20	Random-20
Bagel	163	+1.09548	+0.60008	+0.93493
Emu3	129	-0.01211	-0.12694	+0.07316

Table 7: Earlier-protocol AUROC for distinguishing intent-aligned from conflicting images, using the target-count score on all 200 images. Clean values are from the corresponding stored baseline summaries; ablated values are recomputed from paired per-image scores. Bagel random values average ten sets; Emu3 has one recorded set.

Verifier	Condition	Self	Other	Δ_A
Bagel	Clean	0.92186	0.81910	+0.10276
Bagel	Top-20	0.90592	0.90237	+0.00354
Bagel	Random-20	0.92812	0.82346	+0.10466
Emu3	Clean	0.88962	0.87243	+0.01719
Emu3	Top-20	0.85223	0.84727	+0.00496
Emu3	Random-20	0.90344	0.87651	+0.02693

human image preferences (Xu et al., 2023; Kirstain et al., 2023; Wu et al., 2023). Our task measures satisfaction of a labeled count and separates the verifier's answer from its candidate-selection decision.

Prompt-induced hallucination. Object-hallucination evaluations examine visual grounding in captions and model answers (Rohrbach et al., 2018; Li et al., 2023), while HallusionBench distinguishes language-hallucination and visual-illusion failures (Guan et al., 2024). These are related evaluation settings, not datasets used in our experiments. Rudman et al. (2026) study conflicts between prompt content and visual evidence, and identify attention heads whose ablation reduces prompt-following errors. Ablation here means intervening on an internal component to measure its contribution to behavior. This work provides the closest comparison for our disclosed-intent condition and correction-based head selection. We add matched provenance conditions and evaluate the selected components against both target-following errors and visual counting accuracy. The precise selection rule and zero-ablation operator are specified in Section 5.

Attribution and monitor reliability. Khullar et al. (2026) examine self-attribution in language-model monitors on coding and tool-use tasks. Their experiments distinguish actions presented through assistant or user history from explicit statements of ownership; the latter alone do not reproduce the same bias. Our study places explicit source statements alongside a fixed image and target description. The resulting comparison addresses provenance wording, complementing the history-based setting through a separately controlled input.

Mechanistic interpretation. Head-pruning studies show that attention heads contribute unequally to task performance (Michel et al., 2019; Voita et al., 2019), motivating direct tests of the consequences of deleting a selected set. Faithfulness work further emphasizes specifying what an explanation is intended to capture (Jacovi & Goldberg, 2020). Ferrando et al. (2025), for example, identify entity-recognition directions and examine their effects on knowledge refusal and attention. Heimersheim & Nanda (2024) discuss how activation-patching design and metric choice shape the conclusions of such experiments. Activation patching replaces internal values with values from a reference run; our experiment instead selects candidates by activation differences and then tests zero-ablation. We evaluate the resulting sets using distinct criteria for correction, attribution sensitivity, and retained counting capability.

Distributional and behavioral evaluation. Scholten et al. (2025) motivate evaluating output distributions beyond a single deterministic response. Calibration work distinguishes a model score from an empirically reliable probability of correctness (Guo et al., 2017). We accordingly treat the candidate-normalized target score as a ranking signal, not as an established calibrated correctness probability. For verification, distributional scores, complete answers, and candidate choices provide

complementary measurements. Our evaluation links these measurements on the same held-out images and adds a direct candidate-selection task. This organization allows the empirical relationship between scoring and decisions to be examined rather than built into the definition of a single success metric.