MOON: REDUCING OBJECT HALLUCINATION WITH REPAIR-AWARE VISUAL GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models often describe objects that are absent from an image. External visual guidance can correct these errors, but it can also introduce false mentions or remove correct content. We introduce MOON, a controller that learns when and how to guide a frozen vision-language model. Its central idea is to predict what guidance will change. Adding a mention repairs an omission when the object is present and causes damage when it is absent; deleting a mention reverses these roles. MOON combines these effects with estimates of object presence and the unguided model's behavior to choose an action before generation. Across three backbone updates, it reduces average weighted object loss over 8–128-image paired-audit budgets by 1.9–2.3% relative to a direct risk predictor trained on the same observations. Both methods also use a pool of unguided target responses. The advantage extends to prompt and dataset changes and narrows as paired supervision increases. Controlled comparisons support repair/damage structure as a useful inductive bias for adapting visual guidance with limited paired data.

1 INTRODUCTION

Vision-language models produce fluent descriptions that sometimes mention objects absent from the image (Rohrbach et al., 2018; Li et al., 2023). Visual experts offer a way to correct these hallucinations without updating the language model. A list of detected objects can help recover a missed traffic light, for example, but an incorrect label can introduce a nonexistent object, and a long hint can distract from content the model already describes correctly. Effective guidance therefore requires two decisions: which evidence to trust, and whether using it will improve the answer.

Existing frozen-model interventions change decoding, attention, or evidence prompts (Leng et al., 2024; Huang et al., 2024; Zhao et al., 2025). Their usefulness depends on the response of the guided model. A backbone upgrade or a new prompt can change which objects the unguided model mentions and how it reacts to the same evidence. Learning a new intervention policy from a small set of target responses is difficult because each action can both repair errors and damage correct content.

MOON addresses this problem by modeling these two effects separately. Its coverage-aware evidence module estimates object presence while distinguishing unavailable expert observations from negative ones. Its repair-aware controller represents changes from the unguided model, called Base, through bounded addition and deletion factors. An added traffic-light mention repairs an omission when the object is present and causes damage when it is absent. Weighting these changes by presence and Base behavior gives the expected loss of each guidance action. The controller selects an action before decoding a single answer.

The decomposition is particularly useful when adapting to a new model or prompt. The expected loss separates into object presence, Base behavior, and the effects of guidance. Each component describes a different way the setting can change. Their combination gives the controller a structured risk estimate to learn from a small paired audit, in which the same images are evaluated with and without guidance.

Our experiments find the largest benefit in this limited-data regime. Accounting for the conditions under which mentions change improves on direct risk prediction, and the bounded repair/damage

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.

factors provide a further gain. As more paired responses become available, flexible predictors catch up. This pattern supports the decomposition as an inductive bias for adaptation.

We contribute a repair-aware controller that separates evidence validity from intervention utility (Section 3), and an empirical account of how its structure helps under limited paired supervision (Section 5.1). Transfer and measurement studies establish where the benefit persists, what adaptation costs, and how the fitted factors relate to observed mention changes (Sections 5.2–5.4).

2 Related work

Hallucination and intervention. CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023), and AMBER (Wang et al., 2023) measure different hallucination behaviors. VCD modifies token distributions (Leng et al., 2024); OPERA modifies attention-related decoding decisions (Huang et al., 2024); MARINE supplies external image-grounded guidance (Zhao et al., 2025). Recent methods reinforce visual attention (ILVAD and AFIP) (Xie et al., 2026; Li et al., 2026), attenuate feed-forward language priors (FADE) (Guo et al., 2026), or encourage and suppress mentions using an intrinsic object inventory (PatchGate) (Ko et al., 2026). MOON complements these mechanisms by learning which available intervention is likely to improve a particular response.

Evidence and transfer. DETR, RAM++, and Grounding DINO provide the evidence interfaces (Carion et al., 2020; Huang et al., 2025; Liu et al., 2024c). Their predictions remain fallible observations. Label-shift correction concerns changed class proportions (Lipton et al., 2018); MOON additionally allows Base and intervention behavior to change. Effect-invariant policy generalization likewise makes transport depend on invariant components (Saengkyongam et al., 2024), whereas our conditions concern decoder-specific addition/deletion effects. This connects the decision value of retained information (Blackwell, 1953) to a concrete decomposition of guidance effects. Recent evaluation cautions that lower hallucination can accompany lost informativeness or capability (Fazli et al., 2026); we therefore report recall, native endpoints, and downstream checks alongside weighted loss.

3 Evidence validity and intervention utility

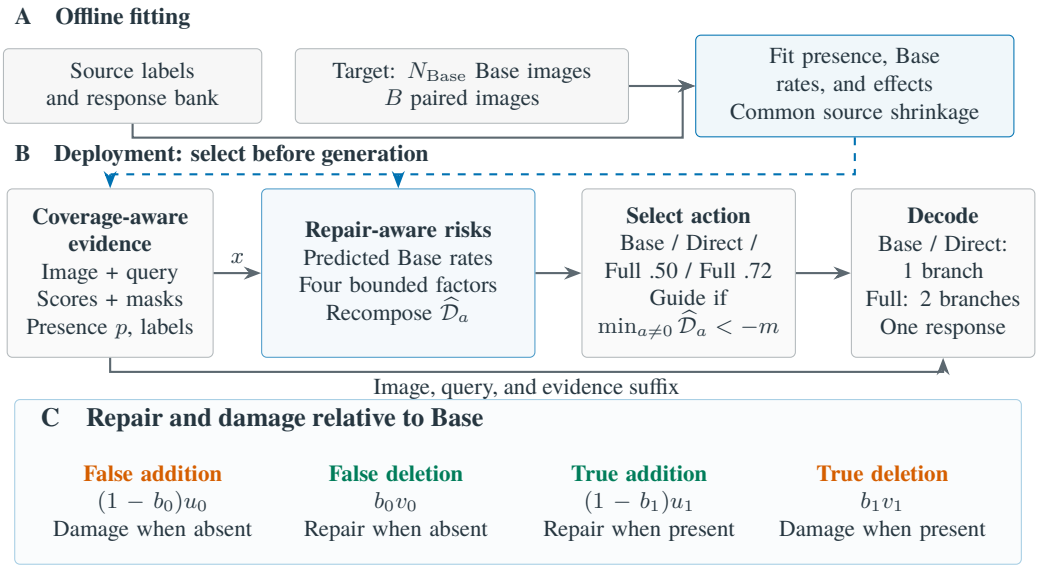
Figure 1 separates offline fitting from deployment. For image I, query q, and canonical object class o, let $H \in \{0, 1\}$ denote presence. An evidence expert j provides a score c_j and applicability mask a_j. A zero mask means the class is outside the expert's coverage or was not queried. Let x denote context available before generation: the query/task, expert scores and masks, image quality, and predictions from fitted presence and Base-state estimators. Class conditioning is implicit in the probabilities below. Appendix B.1 specifies the 16-dimensional implementation.

3.1 Coverage-aware evidence

We calibrate an applicable expert score to $g_j(c_j)$, center its log odds by the fitting prior π_j, and mask unavailable observations:

$$z_j = a_j \left[\operatorname{logit} g_j(c_j) - \operatorname{logit} \pi_j\right], \qquad \widehat{p}(x) = \operatorname{sigmoid}\left(\frac{\beta_0 + \sum_j \theta_j z_j + \theta_v v(I)}{T}\right). \tag{1}$$

An unavailable expert contributes zero; an applicable low score remains informative. Calibration and threshold selection use disjoint calibration data. Linear CA uses additive fusion. Joint CA additionally models interactions among the support bits, allowing joint false-support patterns to be represented. A support gate requires at least one applicable positive expert observation before a label is selected. The evidence set contains deduplicated canonical labels above the threshold. This is a coverage-aware interface, while the empirical object audit in this paper is restricted to 80 classes.



Weight events by presence p and loss weight ρ; aggregate over classes relative to Base.

Figure 1: MOON separates evidence validity from the utility of applying guidance. Offline fitting uses source observations and target Base/paired pools. At deployment, coverage-aware evidence (CA) supplies presence estimates and supported labels; repair-aware action selection (IG+) predicts excess risk and dispatches one frozen-decoder action. False additions and true deletions contribute damage; false deletions and true additions contribute repair. Base and Direct use one decoder branch, while Full combines two branches into one response.

3.2 Actions and paired loss

The default action set is $\mathcal{A} = \{0, \text{Full}_{.50}, \text{Full}_{.72}, \text{Direct}\}$. Action 0 is the original Base decoder. Full mixes next-token logits from two branches with the same image and generated prefix:

$$\ell_t^{\text{Full}_\gamma} = (1-\gamma)\ell_t^{\text{Base}} + \gamma \ell_t^{\text{guide}}. \tag{2}$$

Equation 2 defines the two-branch Full action. Direct uses the guided prompt as its sole branch. The prompt presents supported labels as an incomplete list of visible objects. Empty evidence falls back to Base. Actions are selected before generating the chosen answer; collecting all actions is needed only for supervised fitting and evaluation.

Let $Z^a \in \{0,1\}$ indicate whether the response from action a mentions class o. The per-class loss is

$$\ell(H, Z) = \rho(1-H)Z + H(1-Z), \tag{3}$$

where $\rho = 4$ for the caption audit. The larger false-positive weight prioritizes avoiding unsupported mentions; Section 5.2 tests sensitivity to that choice. For $h \in \{0,1\}$, define joint flip probabilities

$$A_{h,a} = P(Z^0 = 0, Z^a = 1 \mid H = h, x), \quad D_{h,a} = P(Z^0 = 1, Z^a = 0 \mid H = h, x). \tag{4}$$

For the following population identities, let $p = P(H = 1 \mid x)$; deployment substitutes the CA estimate $\widehat{p}$. The conditional excess loss of guidance is exactly

$$\Delta_a(x) = \rho(1-p)(A_{0,a} - D_{0,a}) + p(D_{1,a} - A_{1,a}). \tag{5}$$

The joint events in Equation 4 separate repair from damage. False additions and true deletions are damage; false deletions and true additions are repair. Their contribution to action risk depends jointly on presence and the model's response to guidance.

3.3 FACTORIZED ACTION RISKS

Write the Base mention rate and conditional flip rates as

$$b_h = P(Z^0 = 1 \mid H = h, x), \tag{6}$$

$$u_{h,a} = P(Z^a = 1 \mid Z^0 = 0, H = h, x), \qquad v_{h,a} = P(Z^a = 0 \mid Z^0 = 1, H = h, x). \tag{7}$$

Equations 6–7 give $e_{h,a} = A_{h,a} - D_{h,a} = (1 - b_h)u_{h,a} - b_h v_{h,a}$. For class weights $w_o \geq 0$ summing to one, normalized image-level action risk is

$$\mathcal{D}_a(I, q) = \frac{1}{\rho + 1} \sum_o w_o \left[\rho(1 - p_o)e_{0,a,o} - p_o e_{1,a,o}\right]. \tag{8}$$

The empirical audit uses $w_o = 1/80$ and reports $100\mathcal{D}_a$ in loss points. At deployment, Base rates are conditional predictions from a fitted 18-cell task/coverage/quality table, available before decoding (Appendix B.1). Equation 8 is the risk-recomposition interface: the same class probability can imply different actions under different Base or flip rates.

Adaptation and fitted readouts. We call the repair-aware action-selection rule IG+. It supports three levels of adaptation. IG source freezes the fitted effects; IG RT updates target presence/Base statistics while reusing source conditional flips; local adaptation also learns from paired target action outcomes. Reusing source flips gives exact target risks when $u_{h,a,t} = u_{h,a,s}$ and $v_{h,a,t} = v_{h,a,s}$ at the retained conditioning resolution.

In the attribution study, MOON-R fits four bounded sigmoid factors per action to paired excess risk and combines them through Equation 8. The implementation shares these four factors across classes within each image, and a common source mixture regularizes adaptation. The heads are latent factors under this risk-only objective. MOON-S adds supervision on the corresponding flip events, allowing us to examine their calibration separately from policy quality. Appendix B specifies the shared Base-state estimators, risk readouts, and source-prior ablations.

Single-anchor adaptation. If only an anchor action j is labeled on the target, source contrasts can supply the remaining actions:

$$e_{h,a,t} = e_{h,j,t} + (1 - b_{h,t})(u_{h,a,s} - u_{h,j,s}) - b_{h,t}(v_{h,a,s} - v_{h,j,s}). \tag{9}$$

Equation 9 transfers action-minus-anchor flip contrasts that remain stable across domains. We use Full .50 as the anchor and measure transport residuals on held-out images.

Selection and error. Let $a_* \in \arg\min_{a \neq 0} \widehat{\mathcal{D}}_a$. IG+ selects a_* if its predicted excess risk is below $-m$, otherwise Base. A nonnegative margin m trades intervention frequency against estimated gain. If $|\widehat{\mathcal{D}}_a - \mathcal{D}_a| \leq \epsilon$ uniformly over feasible actions, excess risk relative to the best conditional-risk action is at most $2\epsilon + m$ (Appendix E). The bound links action quality to risk estimation accuracy. Joint recipe selection extends the action space to threshold–decoder pairs and includes Base.

4 EXPERIMENTAL DESIGN

Models, data, and budget. We evaluate three fixed release updates: Qwen2-VL to Qwen2.5-VL (7B) (Wang et al., 2024; Bai et al., 2025), InternVL2 to InternVL2.5 (8B) (Chen et al., 2024b), and LLaVA-1.5 to LLaVA-1.6-Vicuna (7B) (Liu et al., 2024a;b). Their aliases are Q2→Q25, I2→I25, and L15→L16; Appendix A lists exact checkpoint identifiers. We also use stationary Q25, fixed prompt changes, and COCO-to-NoCaps/AMBER transfer (Lin et al., 2014; Agrawal et al., 2019; Wang et al., 2023). Source fitting/calibration/development uses 1,600/400/400 images. Disjoint target pools supply adaptation and 1,600 COCO, 3,000 NoCaps, or 604 AMBER test images. Twenty shared image permutations define nested paired budgets $B = 8, 16, 32, 64, 128, 256, 512$ (AMBER: at most 256). Backbone and expert weights remain frozen. Standard decoding is greedy, capped at 128 tokens; tuning excludes target test outcomes.

Supervision and endpoints. The standard target Base pool has 512 images (AMBER: 400); each paired image adds three guided responses. At $B = 32$, this costs 608 target responses and 672 decoder branches on COCO, in addition to source acquisition. With N test images, the shared audit reports

$$L_{80} = 100\frac{4\mathrm{FP} + \mathrm{FN}}{5 \cdot 80N}, \quad U_{80} = 100\frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \quad R_{80} = 100\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{10}$$

The loss-weight study replaces 4 and 5 by ρ and $\rho+1$ to obtain L_ρ. The common audit in Equation 10 is distinct from native CHAIR, CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), and downstream task scores.

Comparison and measurement. The attribution controls share source information and target responses while varying how they predict action risk. We evaluate the fitted heads against held-out flip events and audit 800 responses from 200 images against human mention labels. Each audited image retains its four paired action responses.

Statistical analysis. Learning-curve loss (AULC) averages trapezoidal loss over $\log_2 B$ in the 8–64 and 8–128 windows. Repeated-fit comparisons preserve image/fit pairing and use a two-axis uncertainty approximation with t_{19} intervals. Across 24 main-text readout contrasts, crossed bootstrap gives the same effect directions and pointwise 95% interval zero-exclusion decisions, with a median bootstrap/two-axis SE ratio of 1.02 (range 1.00–1.09). Significance uses the correction of Holm (1979) within the specified families. At $B = 512$, two one-sided tests (Schuirmann, 1987) assess equivalence at a ± 0.10 loss-point tolerance, corresponding to ten net FP-equivalent errors per 100 images. Appendix G gives the full bootstrap comparison and tighter-tolerance results.

5 Results

5.1 Learning guidance with limited paired data

MOON's largest gains occur when few target images have paired action responses. Figure 2 compares the risk-fitted controller, MOON-R, with a direct risk predictor called Prior-matched Flat. Both use the same features and response labels, have 204 readout parameters in addition to shared feature estimators, and mix their fitted target risk f_B with the same source prediction μ_s:

$$\widehat{f}_B = (1 - \alpha_B) f_B + \alpha_B \mu_s, \qquad \alpha_B = 7/(B + 7). \tag{11}$$

The difference is how they form the target risk. MOON-R combines four sigmoid factors through Equation 8; Flat averages four tanh heads to predict risk directly. A third control, Random partition, replaces the repair/damage organization with ten fixed random head-routing maps, averaged within each fit.

Across the three backbone updates, MOON-R reduces average loss over the 8–128 learning curve by 1.9–2.3% relative to Flat. Lower curve loss also holds on stationary Q25, against Random partition, and when the common source mixture is removed. At 512 paired images, Flat and Random partition are practically equivalent to MOON-R within the stated margin. The gain is clearest over the low-budget curve as a whole; individual budgets give less uniform evidence (Appendix B).

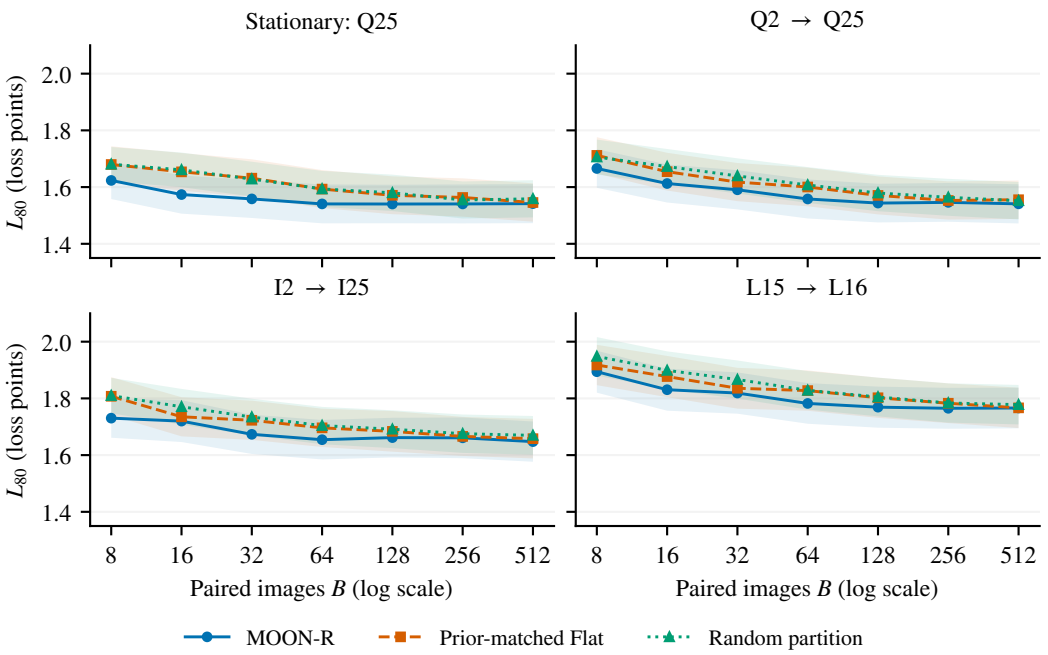


Figure 2: MOON's advantage is concentrated at low paired budgets. Means span 20 shared fits; bands are pointwise 95% intervals. Random partition averages ten mappings within fit. Each method also uses 512 target Base images.

How the structure helps. The link-matched control, Flat-σ, replaces Flat's tanh activation with a sigmoid-based readout and produces nearly the same low-budget curve loss. MOON-R improves on both controls across the two windows, supporting a contribution from the structured readout beyond the activation choice.

Conditional weighting makes a larger difference. Signed-Recompose weights signed heads by object presence and Base behavior, retaining this part of MOON's structure. It improves on Flat across all three updates, but the bounded repair/damage readout improves further. Figure 3 follows these two steps: conditional recomposition and bounded factors each contribute improvements of comparable size. The ordering holds in both low-budget windows.

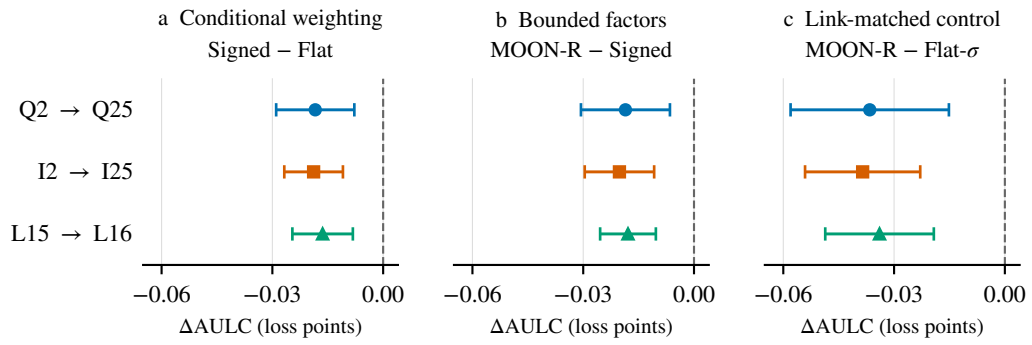


Figure 3: Conditional recomposition improves low-budget risk, with a further gain from MOON's repair/damage readout. Panels follow the two structural changes and the link-matched comparison. Signed denotes Signed-Recompose. Points show 8–128 AULC differences with pointwise 95% intervals; negative values favor the first method. All plotted contrasts are significant under the correction defined in Section 4 and Appendix F.

MOON-R also improves on FreeMix, which replaces fixed recomposition with a learned combiner. Both the source-frozen and target-adapted versions use 228 parameters and have higher curve loss in both windows across all three updates. At 512 paired images they meet the stated equivalence tolerance; the tighter ± 0.05-point criterion also holds for both versions (Appendix G). The fixed structure is most useful while target data are scarce.

What the heads learn. Risk-only fitting trains factors that support action selection. Direct event supervision, used by MOON-S, gives those factors a clearer probabilistic interpretation: every head has lower Brier score and ECE across the three updates, with the largest Brier improvement for false-mention deletion (v_0). Correctly aligned event targets also outperform shuffled targets in policy loss. However, the policy difference between MOON-S and MOON-R remains statistically unresolved. Better-calibrated heads need not produce a measurable improvement in the selected actions. Appendix F reports these diagnostics and the readout comparisons.

5.2 Transfer and supervision cost

The low-budget advantage extends beyond backbone updates. MOON-R has lower curve loss than Flat and Random partition under a prompt change and three transfers from COCO to NoCaps or AMBER (Table 1). The result holds over both low-budget windows. These settings change the text query or the image domain while retaining the same risk-prediction structure.

MOON retains its advantage with a smaller pool of unguided target responses. Reducing the Base pool from 512 to 64 images preserves the gain over Flat (Figure 4, left). At $B = 32$, this reduces target acquisition from 608 to 160 responses and from 672 to 224 decoder branches. Both settings share the same source acquisition. The 32-image Base pool gives less stable results, reported in Appendix I.

Finally, the gain persists when the cost of unsupported mentions changes. Refitting the policies for false-positive weights $\rho \in \{1, 2, 4, 8\}$ retains lower curve loss for MOON-R (Figure 4, right). Thus the low-budget result spans both equal error costs and stronger penalties for hallucination.

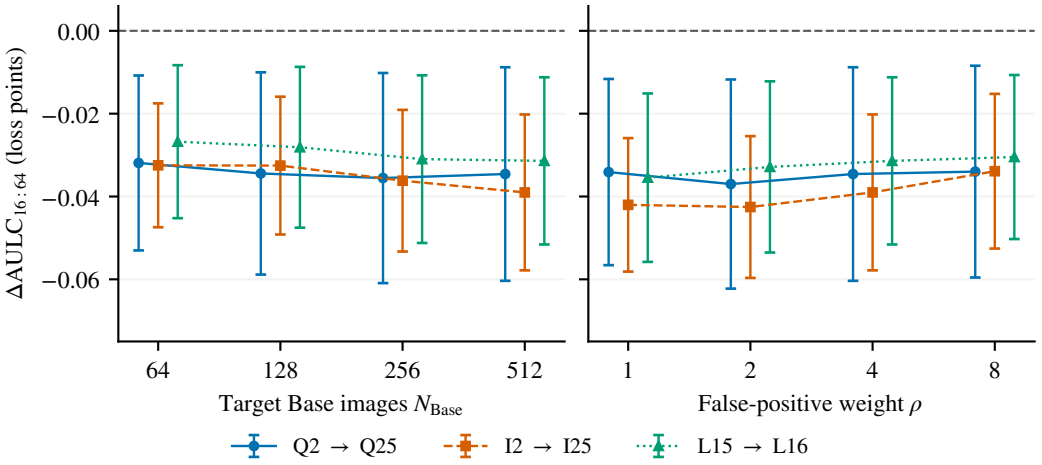


Figure 4: The low-budget advantage persists across Base-pool sizes and loss weights. MOON-R minus Flat AULC over $B = 16$–64, with pointwise 95% intervals. Left: $\rho = 4$; right: 512 Base images and refitting at each ρ. The panels share the $512, \rho = 4$ anchor; all contrasts pass the tests specified in Appendix C.

Table 1: Prompt and domain changes retain the low-budget benefit. MOON-R minus comparator AULC over $B = 8$–128; intervals are pointwise 95%. Adjusted p values cover the 16 contrasts across both low-budget windows. Appendix C defines the Prompt-P4* source query.

Condition	Comparator	ΔAULC [95% CI]	p_{H16}
Prompt-P4*	Prior-matched Flat	-0.052 [-0.069, -0.036]	<0.001
Prompt-P4*	Random partition	-0.054 [-0.068, -0.039]	<0.001
Q25 → NoCaps	Prior-matched Flat	-0.040 [-0.058, -0.023]	<0.001
Q25 → NoCaps	Random partition	-0.052 [-0.064, -0.040]	<0.001
I25 → AMBER	Prior-matched Flat	-0.030 [-0.054, -0.006]	0.0335
I25 → AMBER	Random partition	-0.040 [-0.059, -0.021]	0.0013
L16 → AMBER	Prior-matched Flat	-0.077 [-0.105, -0.049]	<0.001
L16 → AMBER	Random partition	-0.074 [-0.097, -0.052]	<0.001

5.3 Intervention prediction and transport

More accurate effect predictions accompany the improvement in action selection. Across the three backbone updates, IG+ reduces action-effect MAE by 0.067–0.093 loss points relative to Flat residual and also improves on DeepSets. Its selected policies lie 0.023–0.049 points above reference policies fitted on an independent 3,200-image pool (Appendix H).

Stable conditional effects allow adaptation with fewer action labels. With a single labeled anchor, factorization has lower mean loss in all 36 comparisons with Flat controls at the same acquisition budget, although only nine differences are statistically resolved. This makes anchor transfer useful where action effects can be reused.

What must transfer. Base-only adaptation requires stable conditional addition and deletion rates. Anchor adaptation instead requires stable action-minus-anchor contrasts, so an effect change shared by the anchor and another action can cancel in Equation 9. These are assumptions on the decoder's response to guidance. Accurate presence estimates alone cannot establish them: two target mechanisms can have identical presence and Base observations but different guided outcomes. Moreover, a conditional flip rate with zero source support is not identified by source data. Appendix E gives the corresponding transport and non-identifiability results.

Substantial changes in those effects still require paired target observations. For Q25-to-NoCaps, the Full .72 effect predicted from the Full .50 anchor has a residual of -0.495 $[-0.589, -0.401]$ loss points. Very small paired audits can also be unreliable: at $B = 8$, local adaptation is worse than the source policy in 3 of 20 Qwen fits and 7 of 20 InternVL fits. Updating presence and Base statistics alone cannot account for every change in guidance behavior.

5.4 Evidence, external methods, and practical endpoints

Evidence fusion and action selection make complementary contributions. With Joint CA held fixed, replacing Flat residual by IG+ lowers loss from 1.603 to 1.536. Changing Linear to Joint CA with IG+ lowers it from 1.796 to 1.536. Joint threshold–action selection further reduces weighted loss, with a small reduction in recall relative to staged selection.

On L15 with common canonical evidence, MOON has loss 1.928 versus 2.052 for ILVAD+C and 2.050 for AFIP-public+C, which evaluates AFIP's head-guidance component. Native NoCaps CIDEr/SPICE improve and AMBER CHAIR decreases across the three target backbones.

We also test whether the controller preserves broader capabilities. Frozen-controller evaluation on MMStar, RealWorldQA, GQA, POPE, and OCRBench (Chen et al., 2024a; xAI, 2024; Hudson & Manning, 2019; Li et al., 2023; Liu et al., 2024d) meets a -2 point non-inferiority margin in all 15 comparisons. Grouped HallusionBench (Guan et al., 2024) results include no measured Q25 gain. Warm IG+ latency on an RTX A6000 ranges from 1.368 to 1.734 seconds across the three backbones, between Base and always applying Full .50. Appendix H reports the full endpoints and paired differences.

Deployment inputs and computation. The controller's Base-state inputs are conditional predictions from fitted tables; they do not require a trial Base response for the current image. It constructs the 16 features from the request, expert observations, image statistics, and frozen estimators, evaluates the guided-action risks, and dispatches the selected decoder. Full retains two token branches within this single response, whereas Base and Direct each use one. In the reported warm measurements, controller computation is at most 0.75 ms per image and is already included in IG+ latency. Expert loading, cached evidence, and cold execution are accounted for separately in Appendix I.

Mention-extraction validation. Human correction of parsed mentions leaves the estimated policy difference nearly unchanged. The parser achieves 99.30% micro-F1 on 800 audited responses, with action-specific F1 differences within the ± 1 percentage-point equivalence margin. Replacing its outputs with adjudicated mentions changes MOON-R minus Flat loss from -0.03394 to -0.03506 points and preserves the direction in all five conditions. The audit supports stability to mention-extraction errors, while its human-corrected 95% interval, $[-0.08467, 0.01454]$, spans zero (Appendix F).

Practical scale of convergence. The large-budget equivalence result depends on the resolution at which differences matter. Under the 80-class audit with $\rho = 4$, a tolerance of 0.10 loss points corresponds to ten net false-positive-equivalent errors per 100 images; false negatives contribute one quarter as much. Across the 12 final-budget comparisons with Flat-σ, Signed-Recompose, and the two FreeMix controls, tightening the tolerance from ± 0.10 to ± 0.05 retains equivalence in all 12 tests; at ± 0.025, seven pass (Appendix G). Thus convergence persists at half the declared tolerance, while finer distinctions remain less consistently resolved. The low-budget curve comparisons address the separate question of which structure learns useful guidance sooner.

6 Limitations and conclusion

The controlled study covers 80 object classes and assigns a larger cost to unsupported mentions than to omissions. Native task metrics and loss-weight experiments extend the evaluation, but open-vocabulary, attribute, and relation errors remain outside the attribution study. Adaptation also requires source and target Base observations; transferring effects with Base-only or anchor labels works only to the extent that those effects remain stable.

The population decomposition and the fitted implementation have different scopes. Linearity of expectation permits summing class risks without assuming class independence, while the implemented heads share flip factors across classes within an image. This restriction reduces the fitted parameter count but cannot represent arbitrary class-specific effects. Likewise, risk-only supervision does not uniquely identify the latent heads as event probabilities. The selection-error bound concerns conditional expected loss under a uniform estimation-error assumption; it does not guarantee that guidance improves every generated response.

MOON turns visual guidance into a decision about expected repair and damage. Predicting how evidence changes the model's mentions makes that decision easier to learn from limited paired supervision. The benefit diminishes as target data increase, giving a clear role for the decomposition: adapting guidance when new action labels are scarce and the vision-language model remains frozen.

Reproducibility statement

Code is available at https://anonymous.4open.science/r/MooN_-1C3E. Appendix A specifies checkpoints, prompts, actions, and data partitions. Appendix B defines the 16 inputs, Base estimators, fitting grid, readouts, and shared supervision. Appendices C, F, and G provide acquisition counts, comparison families, and uncertainty checks; Appendix E gives the assumptions and proofs.

AI use statement

Generative AI assisted with experimental-control design, analysis of tabulated results, manuscript editing, LaTeX, scientific figures, and literature searches. The authors reviewed the analyses, citations, and presentation and take responsibility for the final manuscript.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Agrawal_nocaps_novel_object_captioning_at_scale_ICCV_2019_paper.html.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pp. 382–398, 2016. URL https://link.springer.com/chapter/10.1007/978-3-319-46454-1_24.

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.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2): 265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pp. 213–229, 2020. URL https://arxiv.org/abs/2005.12872.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f8ee6a3d766b426d2618e555b5aeb39-Paper-Conference.pdf.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2024b. URL https://arxiv.org/abs/2412.05271.

Mehrdad Fazli, Sina Mansouri, Mohit Marvania, and Ziwei Zhu. Does playing it safe count as faithfulness? reassessing LVLM hallucination mitigation methods. *arXiv preprint arXiv:2609.01888*, 2026. URL https://arxiv.org/abs/2609.01888.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://arxiv.org/abs/2310.14566.

Yichen Guo, Kai Tang, Jinhao You, Fenglai Lin, Yiding Sun, Dongxu Zhang, Wenya Wang, Lin William Cong, and Shanghang Zhang. FADE: Mitigating Hallucinations by Reducing Language-Prior Dominance in Large Vision-Language Models. *arXiv preprint arXiv:2606.29431*, 2026. URL https://arxiv.org/abs/2606.29431.

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.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4117–4126, 2025. URL https://doi.org/10.1145/3746027.3755316.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6693–6702, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Jihyung Ko, Eunji Jung, Hyeongsub Kim, Ziseok Lee, Jae Won Cho, Sanghyun Jo, and Kyungsu Kim. PatchGate: Narrowing the verbalization gap with intrinsic object inventories in frozen vision-language models. *arXiv preprint arXiv:2608.21819*, 2026. URL https://arxiv.org/abs/2608.21819.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Quanjiang Li, Zhiming Liu, Wei Luo, Tingjin Luo, and Chenping Hou. Correcting visual blur induced by attention distraction to reduce hallucinations: Algorithm and theory. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.24602.

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. URL https://aclanthology.org/2023.emnlp-main.20/.

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

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

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*, 2024a. URL https://arxiv.org/abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. Official project announcement, 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55, 2024c. URL https://arxiv.org/abs/2303.05499.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67:220102, 2024d. doi: 10.1007/s11432-024-4235-6. URL https://link.springer.com/article/10.1007/s11432-024-4235-6.

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. URL https://aclanthology.org/D18-1437/.

Sorawit Saengkyongam, Niklas Pfister, Predrag Klasnja, Susan Murphy, and Jonas Peters. Effect-invariant mechanisms for policy generalization. *Journal of Machine Learning Research*, 25(34):1–36, 2024. URL https://jmlr.org/papers/v25/23-0802.html.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. URL https://link.springer.com/article/10.1007/BF01068419.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.html.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv preprint arXiv:2311.07397*, 2023. URL https://arxiv.org/abs/2311.07397.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

xAI. RealWorldQA. Dataset card, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Yutong Xie, Zhenglin Hua, Ran Wang, Wing W. Y. Ng, Xizhao Wang, and Yuheng Jia. Finding the correct visual evidence without forgetting: Mitigating hallucination in LVLMs via inter-layer visual attention discrepancy. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.20965.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 77461–77486, 2025. URL https://proceedings.mlr.press/v267/zhao25j.html.

QUALITATIVE EXAMPLES OF ACTION SELECTION

Figure 5 illustrates how selecting among guidance actions can repair incorrect answers, avoid damage from noisy evidence, and recover missed objects.

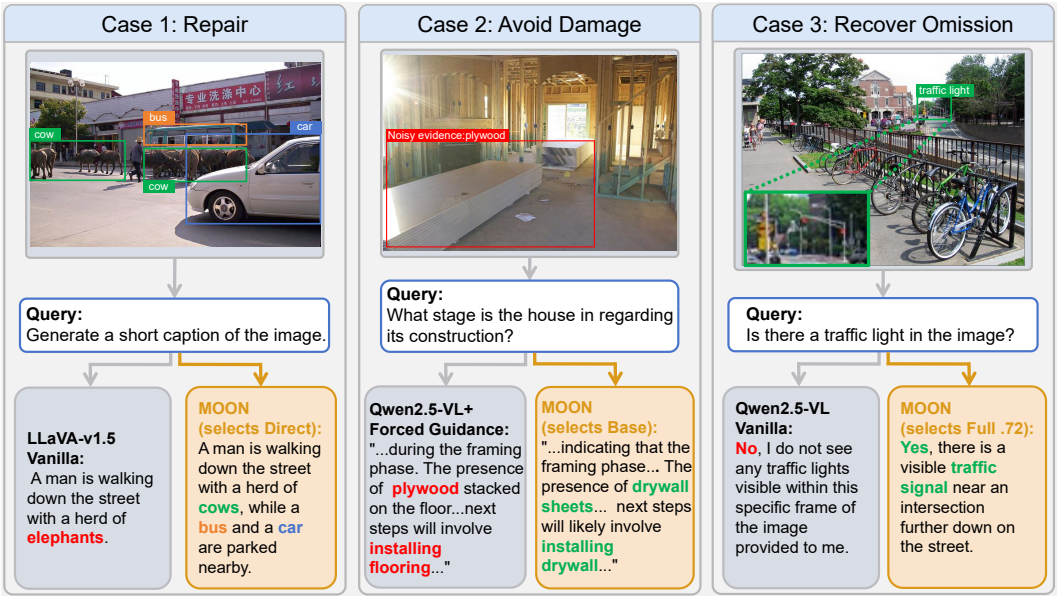


Figure 5: **Qualitative examples of repair-aware action selection.** Left (LLaVA-v1.5): MOON selects Direct, replacing the unsupported elephant mention with cows and adding mentions of the visible bus and car. Middle (Qwen2.5-VL): forced Direct guidance adopts the noisy plywood cue, whereas MOON selects Base and preserves the drywall description. Right (Qwen2.5-VL): MOON selects Full with $\gamma = 0.72$ and recognizes the distant traffic light missed by Base. Responses in the middle panel are excerpted as indicated.

A PROTOCOL AND IMPLEMENTATION

Checkpoint identities. Table 2 identifies the six backbone aliases used throughout the paper.

Table 2: Public checkpoints corresponding to the backbone aliases.

Alias	Public checkpoint identifier
Q2	`Qwen/Qwen2-VL-7B-Instruct`
Q25	`Qwen/Qwen2.5-VL-7B-Instruct`
I2	`OpenGVLab/InternVL2-8B`
I25	`OpenGVLab/InternVL2_5-8B`
L15	`llava-hf/llava-1.5-7b-hf`
L16	`llava-hf/llava-v1.6-vicuna-7b-hf`

Each backbone uses its native image processor and conversation template. Backbone and expert weights remain frozen. The three expert interfaces are DETR, RAM++, and Grounding DINO. Their outputs are mapped to the shared canonical vocabulary before fusion. Out-of-coverage and unqueried entries retain unavailable masks; they are not replaced with negative detections. Supervised image–class presence labels and parsed response mentions are distinct variables.

Prompts and decoding. Table 3 gives the fixed source caption prompt P0 and four target prompts.

Table 3: Exact unguided queries for the source and target prompt settings.

ID	Exact base query
P0	Describe the image.
P1	Write a concise caption for this image.
P2	Describe the visible objects in the image.
P3	Give a factual description of this image.
P4	Describe this image in two sentences.

The guided branch appends the following text after the original query:

```
Visual evidence: <alphabetically ordered canonical
labels>.
Use the image to answer the question. The evidence
list is incomplete.
Do not infer that an unlisted object is absent.
```

Selected labels are deduplicated, sorted, and capped at 32. Empty evidence invokes Base without the suffix. The default evidence threshold is 0.8 at $\rho = 4$, with the support gate. Full .50 and Full .72 specify logit-mixing coefficients and share this evidence set. They interpolate the two branches at each shared generated prefix; Direct uses only the guided branch. Each Full response uses two branch trajectories at a common generated prefix. Standard generation is greedy with EOS enabled and a 128-new-token cap. The length experiment changes only this cap to 256. The decoding-shift experiment uses temperature 0.7 and top-$p = 0.9$, with one fixed image-identifier seed per image.

Image covariate and fitting. The image-only quality proxy is fixed before policy fitting. RGB is converted to luminance $Y = 0.2126R + 0.7152G + 0.0722B$ in $[0, 1]$ and resized bilinearly to 256×256. Let s_Δ be the mean squared four-neighbor discrete Laplacian on interior pixels and e the fraction with $Y < 0.02$ or $Y > 0.98$. We use $v(I) = \mathrm{clip}_{[0,1]}[0.8/(1 + 400s_\Delta) + 0.2e]$, with a split at 0.5 for coarse quality strata. Computing this proxy does not modify the image. Calibration fits only on the calibration partition, and action-effect fitting uses the shared paired-response pools. Sparse conditional cells are regularized toward broader/source estimates. Hyperparameter selection is confined to source development data or cross-validation within the stated target audit; a target test label is never used to choose a fit or an adaptation rule.

Information-matched controls. Flat cell estimates direct action losses within contextual cells; Flat residual adapts a source loss prediction; Flat DeepSets pools label-level evidence with a permutation-invariant predictor (Zaheer et al., 2017). IG+ uses the same allowed source and target outcomes to estimate the factorized quantities. Presence-only control removes action-effect conditioning, pooled-context control reduces conditional information, and no-damage control drops false-addition and true-deletion contributions. The no-prior variant fits without a source effect prior. The no-margin variant sets $m = 0$. An exact same-head expansion of Equation 8 is an implementation identity and is not counted as an independently learned Flat baseline.

Acquisition accounting. For the standard four-action set, one fully observed image produces four responses and six decoder-branch trajectories: one for Base, two for each Full action, and one for Direct. Source effect fitting on 1,600 images therefore costs 6,400 responses and 9,600 branches, excluding separate calibration/development acquisition. With a 512-image target Base pool and B fully paired target images, the additional target collection contains $512+3B$ responses and $512+5B$ branches. Base plus one Full .50 anchor instead requires $512 + B$ responses and $512 + 2B$ branches. When the anchor pool exceeds the original Base pool, the additional Base responses are charged. AMBER substitutes its 400-image Base pool. Equal guided-output budgets are 96/384/1,536; equal branch caps are 1,024/2,048/4,096. Table 35 reports the matched-budget contrasts.

Joint recipes and external methods. Joint recipe selection uses four thresholds: 0.70, 0.80, 0.90, and 0.95. Each is paired with Direct or Full .50, alongside Base. This is nine feasible actions and 13 branch trajectories if every action is acquired. The staged recipe fixes the evidence threshold at 0.80

before evaluation. Global selects one recipe with finite supervised data. The L15 external-baseline panel uses native loading, a common 400-image source development pool, and at most eight tuning trials per method. The +C versions use identical canonical evidence, ordering, and suffix. AFIP-public denotes the released head-guidance component evaluated here.

Labels, metric denominators, and failures. The common audit contains 80 canonical classes per image. Presence and mention extraction use a fixed ontology and synonym mapping, with annotation disagreement resolved independently of controller predictions. Repeated mentions do not multiply class counts. True negatives enter the fixed class denominator of L_{80} but not U_{80} or R_{80}. An empty caption contributes false negatives for present classes. Failed and empty outputs remain in the evaluation. Native NoCaps and AMBER metrics are computed with their native evaluators and ontology. The NoCaps test split contains 432 in-domain, 1,959 near-domain, and 609 out-of-domain images.

Statistical units and diagnostics. The 20 fit-output slots index shared nested training subsets. A deterministic frozen output reused across slots contributes no extra independent image or fit. Paired inference treats images and fits as crossed sampling units. Appendix G reports the numerical agreement of two-axis and crossed-bootstrap intervals. Calibration tests use 4,800 image–action pairs per fit but keep the three actions on an image together. Downstream resampling respects image clusters, including repeated questions. The 15-test non-inferiority family is distinct from the grouped HallusionBench evaluation. The latter contains 1,129 questions, 346 figures, and 167 scene sets. Its grouped results are not pooled with question-independent accuracy tests.

The independent-reference experiment fits a policy with 3,200 separate reference images and uses cross-fitted action selection/evaluation. Its loss measures the performance of a finite-data reference policy. The hindsight minimum instead chooses the best realized action on each test image and lower-bounds loss for that finite action set. The theoretical optimum minimizes conditional expected risk.

Timing. Timing uses one NVIDIA RTX A6000, BF16, and batch size one. After 30 warm-up images, five rounds of 200 images use CUDA synchronization and reset peak-memory statistics. We measure mean/p95 wall time, peak allocated memory, controller time, and output length, separating cold loading, warm execution, and cached modes. Controller computation is included in the measured policy latency; the separately reported controller time must not be added to that total again.

B PRIOR-MATCHED AND RANDOM-PARTITION CONTROLS

Study design. The attribution study separates the benefit of a common function-space source prior from differences between target risk parameterizations. It covers stationary Q25 and the three natural release updates on COCO. The source train split contains 1,600 fitting, 400 calibration, and 400 development images; the target validation split contains the disjoint 512-image Base pool, 1,600-image test set, and 100-image QA set. Target paired images are nested subsets of the adaptation pool at $B = 8, 16, 32, 64, 128, 256, 512$. Twenty shared image permutations define the repeated fits. The action bank, canonical evidence, greedy decoding, and 128-token cap are unchanged.

B.1 FEATURES AVAILABLE BEFORE GENERATION

The controller uses the image, original query, task metadata, expert outputs, and fitted presence and Base-state estimators. Table 4 defines its 16 inputs. For each of the $K = 80$ classes c, let p_c be the fused presence probability, and let a_{ec}, t_{ec}, s_{ec} denote expert e's applicability, calibrated score, and support bit. An applicable negative has $a_{ec} = 1, s_{ec} = 0$; unavailable observations have $a_{ec} = 0$. Define

$$m_c = (1 - p_c)b_{0,c} + p_c b_{1,c}, \qquad \bar{p} = K^{-1} \sum_c p_c, \qquad S_c = \mathbf{1}\left\{ \sum_e s_{ec} > 0 \right\},$$

$$\bar{b}_0 = \frac{\sum_c (1 - p_c) b_{0,c}}{\sum_c (1 - p_c)}, \qquad \bar{b}_1 = \frac{\sum_c p_c b_{1,c}}{\sum_c p_c}.$$

Here m_c predicts the probability of a Base mention. Probabilities are clipped to $[10^{-6}, 1-10^{-6}]$ before weighted averaging. We use $\langle r_c \rangle = K^{-1} \sum_c r_c$, binary entropy $h(p) = -p \log p - (1-p) \log(1-p)$, and the support-gated, deduplicated, sorted, 32-label-capped evidence set E_τ.

Table 4: The complete input vector before source-fitted preprocessing. All features are available before decoding. CA and Base estimators are frozen for evaluation; $\mathrm{Var}_{\mathrm{pop}}$ denotes population variance.

Feature	Definition	Source	At inference?
x_1: presence	$\bar{p}$	CA/Joint CA	Yes
x_2: absent Base rate	$\bar{b}_0$	Base table, p_c	Yes
x_3: present Base rate	$\bar{b}_1$	Base table, p_c	Yes
x_4: mention density	$\langle m_c \rangle$	Predicted Base rates	Yes
x_5: query length	$\min(\lvert \mathrm{NFC}(q) \rvert, 256)/256$; Unicode code points in the original query	Query	Yes
x_6: presence entropy	$\langle h(p_c) \rangle / \log 2$	CA/Joint CA	Yes
x_7: disagreement	$\langle \mathrm{Var}_{\mathrm{pop}}\{t_{ec} : a_{ec} = 1\} \rangle$; variance is zero with fewer than two applicable experts	Expert scores, masks	Yes
x_8: mention-weighted evidence	$\sum_c m_c p_c / \sum_c m_c$; zero for a zero denominator	Predicted rates, p_c	Yes
x_9: omission tendency	$\max_c p_c(1-m_c)$	Predicted rates, p_c	Yes
x_{10}: evidence size	$\lvert E_\tau \rvert/32$; $\tau = 0.8$ for the default four-action setting at $\rho = 4$	Evidence list	Yes
x_{11}: support coverage	$\langle S_c \rangle$	Expert support bits	Yes
x_{12}: applicability	$\sum_{e,c} a_{ec}/(3K)$	Expert masks	Yes
x_{13}: degradation	$v(I) = \mathrm{clip}_{[0,1]}[0.8/(1 + 400 s_\Delta) + 0.2e]$	Image pixels	Yes
x_{14}: image area	$\log(1 + W H_{\mathrm{img}})$	Image dimensions	Yes
x_{15}: evidence failure	One for expert execution failure, invalid values, or fusion failure; zero otherwise	Runtime status	Yes
x_{16}: task family	Description: 0; presence/exhaustive enumeration: 0.5; reasoning: 1	Request metadata or fixed text router	Yes

The quality statistic $v(I)$ uses the luminance construction in Appendix A, with bilinear resizing. Valid unavailable observations and normal nondetections leave $x_{15} = 0$. Failed experts receive unavailable masks and set $x_{15} = 1$; unavailable fused probabilities use the source-fitting prior and set the same flag. An empty evidence set selects Base. An unreadable image is recorded as a request failure.

Base-state cells. Class-dependent Base predictions use the mapping

$$b_{h,c} = b_{h,j(c)}, \qquad j(c) = (\mathrm{task}(q), \mathrm{coverage}(c), \mathrm{quality}(I)).$$

The 18 cells combine three task families, three coverage states, and two quality states. Coverage is shared for at least two applicable experts, exclusive for one, and uncovered for none. Quality is clear when $v(I) < 0.5$ and degraded otherwise. With mention counts k_{hj} among n_{hj} eligible source records, the task-level rate and cell estimates are

$$\mu_{ht} = \frac{k_{ht} + 0.5}{n_{ht} + 1}, \quad b^s_{hj} = \frac{k^s_{hj} + 20\mu_{ht}}{n^s_{hj} + 20}, \quad b^t_{hj} = \frac{k^t_{hj} + 20 b^s_{hj}}{n^t_{hj} + 20}.$$

Empty source cells fall back to the corresponding task rate and then the global source rate. Presence and Base labels fit these tables offline. At deployment, the frozen table supplies a conditional prediction from the cell index before a response is generated. The count-smoothing strength 20 and the risk-mixture weight $\alpha_B = 7/(B+7)$ regularize different estimators.

Cross-fitting and preprocessing. Source risk-head training uses image-grouped five-fold out-of-fold fusion and state predictions. Target adaptation rows use Base tables estimated with the corresponding image excluded. Evaluation uses the frozen table fitted on the full allowed adaptation pool. The source calibration/development and target test partitions remain outside the risk-head fitting records. Features x_1–x_{14} are clipped at source-fitting 1%/99% quantiles and standardized with source-fitting means and population standard deviations; constant columns become zero. Features x_{15}, x_{16} retain their stated values.

Image-level recomposition and parameter count. The four heads for an action share their values across classes within an image. With $g_a(x) = \mathrm{sigmoid}(W_a x + \eta_a)$, ordered as (u_0, v_0, u_1, v_1), the loss-point implementation of Equation 8 is

$$100\widehat{\mathcal{D}}_a = C^{\mathsf{T}} g_a(x), \qquad C = \frac{100}{\rho+1}\begin{bmatrix} \rho\langle(1-p_c)(1-b_{0,c})\rangle \\ -\rho\langle(1-p_c)b_{0,c}\rangle \\ -\langle p_c(1-b_{1,c})\rangle \\ \langle p_c b_{1,c}\rangle \end{bmatrix}.$$

Raw probabilities supply these coefficients; standardized features supply the heads. The weighted definitions of $\bar{b}_0, \bar{b}_1$ give the same coefficients from $\bar{p}$. This implementation assumes shared within-image flip factors; image averages alone do not identify arbitrary class-specific flip heterogeneity. Three guided actions with four affine heads each give $3 \times 4 \times (16+1) = 204$ fitted risk-head parameters. Shared components comprise 36 Base conditional rates, task/global fallback statistics and counts, and the fitted CA/Joint-CA and source-prediction modules.

B.2 Candidate selection and fixed-epoch fitting

The eight candidates in Table 5 use the same source-development split and initialization seeds. The selection score is normalized action excess-risk MSE, averaged equally across the three guided actions, development images, seeds, and conditions, then across the matched readouts to select one common configuration. Scores within 10^{-8} are tied; ties select fewer epochs, then the smaller learning rate, then smaller L2 coefficient. The reported fits use C6, fixed before target fitting.

Table 5: The full source-development candidate grid.

Candidate	Adam learning rate	L2 coefficient	Epochs
C1	0.003	0	100
C2	0.003	0	200
C3	0.003	0.001	100
C4	0.003	0.001	200
C5	0.01	0	100
C6	0.01	0	200
C7	0.01	0.001	100
C8	0.01	0.001	200

Each candidate completes its specified number of full-batch Adam updates with $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. Weights are initialized from $\mathcal{N}(0, 0.01^2)$ and biases at zero. The penalty is $\lambda\|W\|_F^2/2$ on weights. Scoring uses the final epoch; target fits use epoch 200. The shared source mixture remains fixed at $\alpha_B = 7/(B+7)$. FreeMix-Source freezes its 24 combiner parameters, while FreeMix-Target updates all 228 parameters. Each setting has 20 fits.

B.3 Risk readouts and comparison design

Shared supervision and acquisition. All readouts use the same target excess-risk labels and the fixed C6 configuration from Appendix B.2. The margin is $m = 0.005$ in normalized risk units, or

0.5 loss points. It is shared across methods and is distinct from the ± 0.10 loss-point equivalence margin. All methods use the same target action labels. The 512 Base responses plus three guided responses on each of B images cost $512 + 3B$ responses and $512 + 5B$ decoder branches. At $B = 8, 32, 128, 512$, these are respectively 536/608/896/2,048 responses and 552/672/1,152/3,072 branches. Source fitting and separate calibration/development acquisition are additional common costs.

Structured and direct-risk readouts. For a given guided action, the structured readout has four sigmoid affine heads, $(\widehat{u}_0, \widehat{v}_0, \widehat{u}_1, \widehat{v}_1)$. They are combined with the shared p, b_0, b_1 through Equation 8. The direct-risk readout averages four tanh affine heads on the image-level input:

$$f_B^{\mathrm{Flat}}(x, a) = \frac{1}{4} \sum_{k=1}^{4} \tanh(w_{a,k}^{\mathsf{T}} x + c_{a,k}).$$

Flat learns risk directly from the same p, b inputs. MOON-R constrains the risk through the addition/deletion decomposition. Its heads are latent factors under risk-only fitting; the event-supervised MOON-S variant evaluates their probabilistic interpretation separately.

Common prior and on/off comparison. Every learner uses the identical source prediction function $\mu_s(x, a)$. Each model first fits its unshrunk target predictor f_B, then applies Equation 11 with the identical fixed $\alpha_B = 7/(B + 7)$. This matches the final risk-space prior and shrinkage, in addition to matching source data. The on/off comparison sets α_B to zero without retraining f_B or changing the draw. The "no prior" condition removes this output mixture while retaining the shared frozen features, preprocessing, and stopping protocol. The IG no-prior ablation in the cell-based study instead removes the source effect prior.

Random-partition placebo. For each map identifier $j \in \{0, \ldots, 9\}$, an image–class pair is routed to head

$$g_j(I, o) = \mathrm{SHA256}(j, \mathrm{imageID}(I), \mathrm{classID}(o)) \bmod 4.$$

Four tanh affine risk heads are trained on these random groups; selected class risks are averaged to form the action prediction. Results average over all ten fixed mappings. Each mapping uses the same 204-parameter count, response observations, source prediction, and shrinkage coefficient. Random routing changes the training partition and effective support of each head, providing a non-semantic risk-partition control. By comparison, permuting heads and inverting that permutation in recomposition would leave the predictor unchanged.

Aggregation and inference. The curve endpoint is normalized trapezoidal AULC over $\log_2 B$, evaluated on 8–64 and 8–128 windows. For K successive doubling budgets it is

$$\mathrm{AULC} = \frac{1}{K-1} \sum_{k=1}^{K-1} \frac{L_{B_k} + L_{B_{k+1}}}{2}.$$

AULC has units of loss points. Placebo results average the ten mappings within each of the 20 fits, preserving the original image and fit sampling units. The principal family has $4 \times 2 \times 2 = 16$ contrasts. The secondary families contain 56 pointwise budget contrasts, 24 prior on/off contrasts, and 40 individual-map AULC contrasts. All use two-sided Holm correction within the stated family. Twenty of the 56 individual-budget contrasts pass correction, compared with all 16 integrated-curve contrasts.

At $B = 512$, eight equivalence tests compare MOON separately with Flat and placebo. Each TOST p value is the maximum of its two one-sided tests; Holm correction is applied across all eight. The equivalence margin is ± 0.10 loss points, and the reported 90% intervals are pointwise, not simultaneous. Equivalence decisions use the Holm-adjusted TOST values. Appendix G expresses this tolerance in object-error counts and examines tighter margins for the additional readout controls.

Shared response banks. The 28 MOON-prior curve points share the IG+ local policy results on the same action bank, providing a common reference across studies. Flat source, cell, residual, and DeepSets are distinct from Prior-matched Flat; the 65-comparison transfer family evaluates the former controls. The four-condition COCO study fixes $\rho = 4$. The matched prompt/domain and loss-weight studies extend these comparisons to additional settings.

C SEMANTIC SUPERVISION AND TRANSFER DESIGN

Risk-only and semantic objectives. MOON-R denotes the risk-only structured readout used in the matched-prior study. MOON-S uses the same four sigmoid heads per action and adds supervision on their repair/damage event targets. Both MOON-S and Semantic-shuffled minimize the same risk MSE plus β_B times the mean four-head target MSE, with $\beta_B = 0.10 \cdot 16/(B + 16)$. Head targets are image-level event rates computed from the same (H, Z^0, Z^a) labels: an event count k among n eligible class states gives $(k + 0.5)/(n + 1)$. The risk-only model receives the same raw outcomes but does not use this auxiliary objective. The target risk is the paired difference in Equation 3, aggregated by Equation 8.

Semantic shuffling. Twenty fixed mappings are retained. For each mapping, fit, paired budget, and head, a permutation of training images changes the association between input features and head targets; the three actions share that head's permutation. Risk labels, examples, head architecture, auxiliary weight, source prediction, and optimization are unchanged. This preserves the marginal target collection while removing its original alignment with input features. The diagnostic measures the value of aligning auxiliary targets with inputs. Shuffling can also create conflict between the risk and auxiliary objectives.

Optimization and shared estimators. We use full-batch Adam (Kingma & Ba, 2015), learning rate 0.01, coefficients $(0.9, 0.999)$, numerical constant 10^{-8}, and 200 epochs under configuration C6 (Appendix B.2). The risk-only and semantic readouts have 204 trainable parameters and the 16 inputs in Table 4. The output mixture remains $(1-\alpha_B)f_B+\alpha_B\mu_s$, with $\alpha_B = 7/(B+7)$. Expanding the recomposition algebra with the same heads, initialization, prior, and optimizer gives the identical estimator, counted once. Flat's tanh readout and MOON's sigmoid recomposition impose different functional constraints.

Additional transfer settings. The matched-prior study evaluates Q25 prompt transfer, Q25 COCO-to-NoCaps, and I25/L16 COCO-to-AMBER. The prompt experiment, denoted Prompt-P4*, uses source query "Describe the visible objects in detail." and target query "Describe this image in two sentences." The four-prompt study uses P0 as its source query. NoCaps uses 1,000 adaptation, 3,000 test, and 500 reserve images from its 4,500-image public validation set; the first 512 adaptation images supply Base observations. AMBER uses 400 adaptation and 604 test images, with paired budgets up to 256. Splits and repeated fitting draws are shared by methods.

Acquisition and objective changes. For requested Base size N and paired budget B, the actual Base size is $N' = \max(N, B)$. Target acquisition costs $N' + 3B$ responses, $N' + 5B$ decoder branches, and $80N'$ presence labels per fit. The requested 32/64 allocation aliases 64/64. Source fitting costs 6,400 responses and 9,600 branches; calibration and development together add 3,200 responses and 4,800 branches, giving a common 9,600/14,400 overhead. These counts measure acquired responses and branches; runtime is measured separately. Changing $\rho \in \{1, 2, 4, 8\}$ recomputes targets and source risks, refits controllers, and reselects actions using the saved response bank; it does not require regenerating the four responses.

Inference and repeated anchors. The interval estimator uses $\mathrm{SE}^2 = \widehat{\mathrm{Var}}(\bar{d}_i)/N + \widehat{\mathrm{Var}}(\bar{d}_f)/20$ with t_{19} critical values. Here $\bar{d}_i$ averages a paired contrast over fits and $\bar{d}_f$ averages it over test images. AULC is computed within the paired structure before inference. Images and fits provide the two variance components. Appendix G reports crossed-bootstrap sensitivity on the same paired records. Separate Holm families contain 3 semantic, 16 transfer, 12 Base-pool, and 12 loss-weight comparisons. Secondary budget-point families contain 15/52/42/36 tests; all semantic mappings form a 60-test family. The three MOON-S minus MOON-R comparisons are auxiliary. The Base-pool $N = 512$ and loss-weight $\rho = 4$ settings share an anchor. The semantic MOON-R rows use the same common-reference curves as the prior-matched study.

D INFERENCE ALGORITHM AND INTERPRETATION

Inputs: image I, query q, frozen decoder and experts, fitted CA and action-risk readouts, evidence threshold, and action margin m.
Output: one generated response.

1. Map expert outputs to canonical labels and retain scores and applicability masks. Encode unavailable observations with a zero mask and applicable nondetections with an active mask and negative support.
2. Compute presence probabilities and supported labels with CA. Deduplicate and alphabetically order the selected labels; enforce the 32-label cap. If the evidence list is empty, return Base.
3. Construct the 16 features in Table 4 from the image, query, expert observations, and frozen conditional estimators. Predict every feasible action's excess risk with the fitted readout and source mixture.
4. Choose the minimum predicted non-Base risk. If it fails to fall below $-m$, select Base; otherwise select that action.
5. For Base, decode with the original query. For Direct, decode with the guided suffix. For Full, evaluate Base and guided logits on the same image and generated prefix and interpolate them at the selected strength. Append the chosen token and stop at EOS or the fixed cap.

The controller predicts risk before decoding. Base and Direct use one autoregressive branch; Full uses two logit computations and KV caches at a shared generated prefix to produce one final sequence. Deployment cost includes the expert, fusion, feature, and controller computations plus the selected decoder; selecting Base still incurs expert cost on a cache miss. Under Base-only or anchor adaptation, source effects are transported under the invariance conditions in Appendix E; held-out residuals quantify the resulting approximation error.

E DERIVATIONS AND DECISION LIMITS

Proposition 1 (Neutrality of unavailable evidence)**.** *In Equation 1, changing the numerical placeholder for an unavailable expert score does not change the fused posterior, provided the applicability mask remains zero and the calibrated log odds are finite.*

Proof. The entire centered score is multiplied by $a_j = 0$. This proves neutrality with respect to the score placeholder. Explicit mask and context features can describe informative observation patterns. □

Proposition 2 (Repair/damage decomposition)**.** *For a fixed class, input, and action, Equation 5 is the conditional expectation of $\ell(H, Z^a) - \ell(H, Z^0)$. Equation 8 follows by a weighted sum over classes, without a conditional independence assumption between classes.*

Proof. When $H = 0$, the loss change is $\rho(Z^a - Z^0)$, with expectation $\rho(A_{0,a} - D_{0,a})$. When $H = 1$, it is $Z^0 - Z^a$, with expectation $D_{1,a} - A_{1,a}$. Weighting these cases by $1 - p$ and p proves the first claim. Linearity of expectation proves the second. Finally, conditioning on Z^0 gives $A_{h,a} = (1 - b_h)u_{h,a}$ and $D_{h,a} = b_h v_{h,a}$. □

Proposition 3 (Finite-action selection error)**.** *Include Base with $\mathcal{D}_0 = \widehat{\mathcal{D}}_0 = 0$. Suppose $|\widehat{\mathcal{D}}_a - \mathcal{D}_a| \le \epsilon$ for every action. The margin rule in Section 3.3 selects an action $\widehat{a}$ satisfying*

$$\mathcal{D}_{\widehat{a}} - \min_{a \in \mathcal{A}} \mathcal{D}_a \le 2\epsilon + m.$$

Proof. If the selected non-Base action clears the margin, it minimizes estimated risk over all actions. For a true minimizer $a^\star$, insert and subtract the two estimates to obtain $\mathcal{D}_{\widehat{a}} \le \widehat{\mathcal{D}}_{\widehat{a}} + \epsilon \le \widehat{\mathcal{D}}_{a^\star} + \epsilon \le \mathcal{D}_{a^\star} + 2\epsilon$. If the rule selects Base, every non-Base estimate is at least $-m$ and every true risk is at least $-m - \epsilon$. Base regret is thus at most $m + \epsilon$, which implies the stated bound. □

The comparator minimizes conditional expected loss. Applying the bound to a fitted policy requires a valid uniform estimation-error envelope.

Proposition 4 (Information retained for action selection)**.** *Let X denote the available context and $T = T(X)$ a coarser summary. For finite integrable action risks $\mathcal{D}_a(X)$,*

$$\mathbb{E}\left[\min_a \mathbb{E}[\mathcal{D}_a(X) \mid T]\right] \;\geq\; \mathbb{E}\left[\min_a \mathcal{D}_a(X)\right].$$

Calibration of a presence probability based on T is not sufficient to force equality.

Proof. For every action, $\mathcal{D}_a(X) \geq \min_b \mathcal{D}_b(X)$. Taking conditional expectations, then the minimum on the left, and finally expectations yields the inequality. To see why calibration alone is insufficient, consider two equally likely contexts with the same presence probability but opposite signs of guidance excess risk, $+c$ and $-c$, while Base risk is zero. The coarsened minimum risk is zero, whereas contextual selection has expected excess risk $-c/2$. Such signs can arise from different true-addition/deletion rates at the same presence probability. This example establishes a strict decision-value gap at a fixed presence probability. □

Proposition 5 (Base-state transport and its error)**.** *If $u_{h,a,t} = u_{h,a,s}$ and $v_{h,a,t} = v_{h,a,s}$ conditional on the retained context, substituting target p_t, b_t into Equation 8 gives target action risk. Reweighting the same source flip effects by target Base-state probabilities gives the same expression. If the approximation errors obey $|p_t - \widetilde{p}_t| \leq \epsilon_p$, $|b_{h,t} - \widetilde{b}_{h,t}| \leq \epsilon_b$, and both conditional flip errors are at most ϵ_f, then normalized image-risk error is bounded by*

$$|\mathcal{D}_{a,t} - \widetilde{\mathcal{D}}_{a,t}| \leq \epsilon_p + 2\epsilon_b + \epsilon_f.$$

Proof. The first claims follow from $e = (1-b)u - bv$. For the bound, holding the flip rates fixed, changing b alters e by at most $(u+v)\epsilon_b \leq 2\epsilon_b$. Holding b fixed, changing u, v alters e by at most $[(1-b)+b]\epsilon_f = \epsilon_f$. Since $|e_h| \leq 1$, changing p in the normalized class risk has coefficient at most $(\rho+1)/(\rho+1) = 1$. Changes in the two e terms contribute at most $2\epsilon_b + \epsilon_f$ because their normalized nonnegative weights sum to at most one. The class-weighted sum preserves the bound. □

If a state has zero source support, its conditional flip rate is not identified from that source alone. Smoothing supplies a regularized estimate, whose transport still requires the stated invariance assumption. Held-out natural-shift residuals quantify the empirical transport error.

Proposition 6 (Anchor transport)**.** *Equation 9 holds if $u_{h,a,t} - u_{h,j,t} = u_{h,a,s} - u_{h,j,s}$ and $v_{h,a,t} - v_{h,j,t} = v_{h,a,s} - v_{h,j,s}$. Without a restriction on conditional flips or their contrasts, target Base observations alone do not identify guided action risks.*

Proof. Subtract $e_{h,j,t}$ from $e_{h,a,t}$, use the common target b_h, and substitute the two invariant contrasts. For non-identifiability, fix any target distribution of (H, Z^0, X) and let two possible target mechanisms have different $P(Z^a \mid H, Z^0, X)$. Both generate identical Base observations but can have different action risks. An anchor supplies information for its own action, but unrestricted non-anchor mechanisms remain unidentified. □

Joint recipes. If a fixed-threshold action set is contained in a joint threshold–action set, the minimum true conditional risk over the joint set cannot be larger. For fitted policies, the gain from a larger action set also depends on estimation error. The nine-action experiment compares learned policies with common information and separately reports changes in unsupported mentions and recall.

F Readout attribution and measurement validation

The readout comparisons use three natural backbone updates, seven paired budgets, 20 shared fits, and 1,600 test images per update. MOON-R and Prior-matched Flat provide the common-reference curves. All readouts use the same target observations and source mixture. The statistical families below distinguish low-budget loss, final-budget equivalence, head calibration, and measurement sensitivity.

F.1 LINK, CONDITIONAL, AND LEARNED READOUTS

Table 6: Readout parameter counts. Frozen shared feature estimators and the common source predictor are excluded throughout.

Readout	Total	Target-updated
MOON-R / Flat / Flat-σ / Signed	204	204
FreeMix-Source	228	204
FreeMix-Target	228	228

Flat-σ changes the direct readout's activation. Signed-Recompose retains conditional weighting by presence and Base behavior with signed heads. FreeMix uses a learned combination, either source-fitted and frozen or adapted on target. These controls examine complementary restrictions on the risk function. Each of the five curve-contrast types below forms a six-test Holm family (three updates and two windows). Budget-point comparisons form a separate family of 84. Final-budget equivalence combines all four new controls in one 12-test family with margin ± 0.10 loss points; FreeMix contributes six of these tests. The six Flat-σ–Flat curve-equivalence tests use margin ± 0.03 AULC points.

Table 7: All readout curve contrasts. Intervals are pointwise 95%; p_H uses the six-test family for each contrast type.

Update	Window	ΔAULC	95% CI	p_H
MOON-R minus Flat-σ				
Q2 $\rightarrow$ Q25	8–64	-0.0373	[-0.0613, -0.0134]	0.0041
Q2 $\rightarrow$ Q25	8–128	-0.0366	[-0.0580, -0.0152]	0.0041
I2 $\rightarrow$ I25	8–64	-0.0407	[-0.0588, -0.0227]	<0.001
I2 $\rightarrow$ I25	8–128	-0.0385	[-0.0541, -0.0229]	<0.001
L15 $\rightarrow$ L16	8–64	-0.0321	[-0.0484, -0.0158]	0.0018
L15 $\rightarrow$ L16	8–128	-0.0339	[-0.0486, -0.0192]	<0.001
MOON-R minus Signed-Recompose				
Q2 $\rightarrow$ Q25	8–64	-0.0187	[-0.0323, -0.0051]	0.0095
Q2 $\rightarrow$ Q25	8–128	-0.0186	[-0.0306, -0.0065]	0.0089
I2 $\rightarrow$ I25	8–64	-0.0211	[-0.0320, -0.0101]	0.0021
I2 $\rightarrow$ I25	8–128	-0.0202	[-0.0296, -0.0108]	0.0012
L15 $\rightarrow$ L16	8–64	-0.0179	[-0.0262, -0.0096]	0.0012
L15 $\rightarrow$ L16	8–128	-0.0179	[-0.0254, -0.0103]	<0.001
MOON-R minus FreeMix-Source				
Q2 $\rightarrow$ Q25	8–64	-0.0200	[-0.0381, -0.0020]	0.0368
Q2 $\rightarrow$ Q25	8–128	-0.0203	[-0.0356, -0.0049]	0.0368
I2 $\rightarrow$ I25	8–64	-0.0433	[-0.0652, -0.0214]	0.0028
I2 $\rightarrow$ I25	8–128	-0.0371	[-0.0546, -0.0196]	0.0017
L15 $\rightarrow$ L16	8–64	-0.0213	[-0.0374, -0.0052]	0.0368
L15 $\rightarrow$ L16	8–128	-0.0204	[-0.0335, -0.0073]	0.0167
MOON-R minus FreeMix-Target				
Q2 $\rightarrow$ Q25	8–64	-0.0212	[-0.0383, -0.0041]	0.0180
Q2 $\rightarrow$ Q25	8–128	-0.0236	[-0.0398, -0.0074]	0.0148
I2 $\rightarrow$ I25	8–64	-0.0452	[-0.0654, -0.0249]	<0.001
I2 $\rightarrow$ I25	8–128	-0.0409	[-0.0577, -0.0242]	<0.001
L15 $\rightarrow$ L16	8–64	-0.0301	[-0.0499, -0.0103]	0.0148
L15 $\rightarrow$ L16	8–128	-0.0313	[-0.0493, -0.0133]	0.0070

Continued on next page

Table 7 continued

Update	Window	ΔAULC	95% CI	p_H
Signed-Recompose − Flat				
Q2 → Q25	8–64	-0.0190	[-0.0307, -0.0073]	0.0061
Q2 → Q25	8–128	-0.0184	[-0.0290, -0.0078]	0.0053
I2 → I25	8–64	-0.0204	[-0.0294, -0.0113]	<0.001
I2 → I25	8–128	-0.0188	[-0.0268, -0.0109]	<0.001
L15 → L16	8–64	-0.0147	[-0.0240, -0.0055]	0.0061
L15 → L16	8–128	-0.0164	[-0.0246, -0.0082]	0.0020

Table 8: Complete mean-loss curves for all readout controls. Every cell averages 20 fits. S/T denote the source-frozen and target-adapted FreeMix combiners.

B	MOON-R	Flat	Flat-σ	Signed	Mix-S	Mix-T
Q2 → Q25						
8	1.6656	1.7118	1.7102	1.6899	1.6900	1.6973
16	1.6126	1.6544	1.6539	1.6355	1.6456	1.6329
32	1.5905	1.6176	1.6179	1.6007	1.5928	1.6007
64	1.5581	1.6004	1.6001	1.5799	1.5833	1.5926
128	1.5436	1.5707	1.5705	1.5581	1.5601	1.5706
256	1.5461	1.5528	1.5520	1.5485	1.5429	1.5484
512	1.5409	1.5554	1.5556	1.5464	1.5414	1.5434
I2 → I25						
8	1.7304	1.8075	1.8081	1.7668	1.8093	1.8005
16	1.7199	1.7351	1.7336	1.7274	1.7427	1.7524
32	1.6733	1.7228	1.7214	1.6993	1.7252	1.7221
64	1.6544	1.6962	1.6976	1.6773	1.6861	1.6927
128	1.6618	1.6837	1.6822	1.6740	1.6669	1.6799
256	1.6609	1.6660	1.6655	1.6629	1.6770	1.6691
512	1.6476	1.6580	1.6591	1.6511	1.6645	1.6602
L15 → L16						
8	1.8941	1.9178	1.9189	1.9123	1.9208	1.9164
16	1.8306	1.8771	1.8759	1.8558	1.8717	1.8622
32	1.8189	1.8358	1.8353	1.8287	1.8194	1.8456
64	1.7822	1.8275	1.8265	1.8012	1.8002	1.8238
128	1.7691	1.8022	1.8039	1.7859	1.7864	1.7976
256	1.7651	1.7829	1.7828	1.7751	1.7687	1.7751
512	1.7661	1.7666	1.7690	1.7661	1.7689	1.7710

Table 9: Equivalence tests. Intervals are pointwise 90%; TOST p values are Holm-adjusted within the six- or twelve-test family.

Update	Window/control	Δ	90% CI	p_H
Low-budget Flat-σ minus Flat; margin ± 0.03				
Q2 → Q25	8–64	-0.0004	[-0.0017, 0.0009]	<0.001
Q2 → Q25	8–128	-0.0004	[-0.0015, 0.0007]	<0.001
I2 → I25	8–64	-0.0007	[-0.0022, 0.0009]	<0.001
I2 → I25	8–128	-0.0005	[-0.0019, 0.0009]	<0.001
L15 → L16	8–64	-0.0005	[-0.0019, 0.0008]	<0.001
L15 → L16	8–128	-0.0003	[-0.0015, 0.0009]	<0.001

Continued on next page

Table 9 continued

Update	Window/control	Δ	90% CI	p_H
$B = 512$; **MOON-R minus control; margin** ± 0.10				
Q2 $\rightarrow$ Q25	Flat-σ	-0.0147	[-0.0287, -0.0007]	<0.001
Q2 $\rightarrow$ Q25	Signed-Recompose	-0.0055	[-0.0130, 0.0021]	<0.001
Q2 $\rightarrow$ Q25	FreeMix-Source	-0.0005	[-0.0099, 0.0088]	<0.001
Q2 $\rightarrow$ Q25	FreeMix-Target	-0.0025	[-0.0144, 0.0095]	<0.001
I2 $\rightarrow$ I25	Flat-σ	-0.0115	[-0.0245, 0.0015]	<0.001
I2 $\rightarrow$ I25	Signed-Recompose	-0.0035	[-0.0125, 0.0055]	<0.001
I2 $\rightarrow$ I25	FreeMix-Source	-0.0169	[-0.0299, -0.0039]	<0.001
I2 $\rightarrow$ I25	FreeMix-Target	-0.0126	[-0.0237, -0.0015]	<0.001
L15 $\rightarrow$ L16	Flat-σ	-0.0028	[-0.0122, 0.0065]	<0.001
L15 $\rightarrow$ L16	Signed-Recompose	-0.0000	[-0.0060, 0.0060]	<0.001
L15 $\rightarrow$ L16	FreeMix-Source	-0.0028	[-0.0151, 0.0095]	<0.001
L15 $\rightarrow$ L16	FreeMix-Target	-0.0048	[-0.0191, 0.0094]	<0.001

Three of the 84 individual-budget contrasts pass Holm correction. Tables 8 and 7 show that the advantage is clearer across the evaluated low-budget windows than at individual budgets.

F.2 HEAD-EVENT CALIBRATION

Calibration evaluates each factor on its eligible Base state: u_h uses $H = h, Z^0 = 0$ and v_h uses $H = h, Z^0 = 1$. The target event is addition for u and deletion for v. Table 10 pools the guided actions within each update and reports event support alongside Brier score, ten-bin ECE, and calibration slope/intercept. Twenty fits share the test observations. Brier comparisons form one 12-test Holm family; the three policy-loss comparisons form a separate family.

Table 10: Held-out head calibration averaged across fits. Lower Brier/ECE is better; ideal slope/intercept are 1/0.

Method	Head	Brier	ECE10	Slope	Intercept	AUROC
Q2 $\rightarrow$ Q25						
MOON-R	u_0	0.010088	0.00472	1.259	0.601	0.759
MOON-S	u_0	0.010051	0.00015	1.006	0.012	0.777
MOON-R	v_0	0.229821	0.12994	0.572	-0.238	0.663
MOON-S	v_0	0.211967	0.01053	0.935	0.015	0.666
MOON-R	u_1	0.234919	0.04376	0.817	-0.103	0.620
MOON-S	u_1	0.232342	0.01254	0.976	0.049	0.624
MOON-R	v_1	0.044753	0.01174	1.075	-0.044	0.686
MOON-S	v_1	0.044471	0.00112	0.960	-0.089	0.697
I2 $\rightarrow$ I25						
MOON-R	u_0	0.009961	0.00526	1.242	0.495	0.747
MOON-S	u_0	0.009918	0.00013	0.973	-0.107	0.769
MOON-R	v_0	0.220346	0.12624	0.596	-0.264	0.685
MOON-S	v_0	0.203317	0.00716	0.957	0.001	0.690
MOON-R	u_1	0.233323	0.05091	0.806	-0.123	0.628
MOON-S	u_1	0.229805	0.00843	0.932	0.025	0.632
MOON-R	v_1	0.040975	0.01712	1.156	0.040	0.691
MOON-S	v_1	0.040534	0.00249	1.016	-0.014	0.706
L15 $\rightarrow$ L16						
MOON-R	u_0	0.009927	0.00602	1.272	0.549	0.746
MOON-S	u_0	0.009871	0.00017	0.977	-0.105	0.770
MOON-R	v_0	0.223898	0.12506	0.573	-0.218	0.670

Continued on next page

Table 10 continued

Method	Head	Brier	ECE10	Slope	Intercept	AUROC
MOON-S	v_0	0.207367	0.00856	0.968	0.021	0.674
MOON-R	u_1	0.234890	0.05045	0.875	-0.159	0.624
MOON-S	u_1	0.231466	0.00676	1.017	-0.018	0.629
MOON-R	v_1	0.042774	0.01661	1.227	0.237	0.702
MOON-S	v_1	0.042345	0.00119	1.066	0.192	0.718

Table 11: MOON-S minus MOON-R Brier differences, event supports, and corrected tests. Counts refer to eligible label–action observations, shared across fits; they are not independent image counts.

Head	Eligible	Events	ΔBrier	95% CI	p_{H12}
Q2 → Q25					
u_0	328605	3391	-0.000037	[-0.000041, -0.000032]	<0.001
v_0	11901	7589	-0.017855	[-0.020186, -0.015524]	<0.001
u_1	8388	4877	-0.002576	[-0.003753, -0.001400]	<0.001
v_1	35106	1688	-0.000282	[-0.000359, -0.000205]	<0.001
I2 → I25					
u_0	326691	3319	-0.000043	[-0.000048, -0.000037]	<0.001
v_0	13815	9025	-0.017029	[-0.019113, -0.014944]	<0.001
u_1	9801	5755	-0.003518	[-0.004602, -0.002435]	<0.001
v_1	33693	1467	-0.000441	[-0.000524, -0.000358]	<0.001
L15 → L16					
u_0	325251	3292	-0.000055	[-0.000061, -0.000049]	<0.001
v_0	15255	9925	-0.016531	[-0.018649, -0.014414]	<0.001
u_1	10851	6293	-0.003424	[-0.004366, -0.002482]	<0.001
v_1	32643	1496	-0.000429	[-0.000525, -0.000332]	<0.001

Table 12: MOON-S minus MOON-R policy-loss contrasts in the head-calibration study. None passes the three-test correction.

Update	ΔL	95% CI	p_{H3}
Q2 → Q25	-0.00610	[-0.01346, 0.00126]	0.1981
I2 → I25	-0.00914	[-0.01776, -0.00052]	0.1164
L15 → L16	-0.00618	[-0.01370, 0.00134]	0.1981

F.3 Human mention audit

The audit contains 200 images with all four action responses per image (800 responses): 50 images from each natural backbone update and 25 each from Q25-to-NoCaps and L16-to-AMBER. Human reference labels describe mentions in the response text over the canonical 80 classes. They are separate from image-presence labels. Per-action precision/recall/F1 and exact-set agreement evaluate the parser against these reference mention sets. Exact-set accuracy requires all 80 class decisions in a response to agree, a stricter criterion than class-level F1. The action comparisons retain image pairing and use a ± 1 percentage-point F1 equivalence margin, with Holm correction over six pairs.

Table 13: Parser micro precision, recall, and F1 (%); exact-set accuracy (%). Overall macro-F1 is 99.29%.

Action	Resp.	P	R	F1	F1 95% CI	Exact
All actions	800	98.98	99.63	99.30	[99.07, 99.54]	87.875
Base	200	99.10	99.70	99.40	[99.15, 99.65]	89.000
Full .50	200	98.77	99.52	99.14	[98.82, 99.46]	86.500
Full .72	200	98.83	99.64	99.23	[98.96, 99.51]	86.000
Direct	200	99.20	99.68	99.44	[99.20, 99.68]	90.000

Table 14: Action-pair F1 differences in percentage points. Pointwise intervals are 95%; superiority and TOST p values use separate Holm-6 corrections. All pairs meet equivalence.

Action pair	ΔF1	95% CI	p_H	$p_{H,\mathrm{TOST}}$
Base − Full .50	0.255	[0.023, 0.486]	0.1566	<0.001
Base − Full .72	0.163	[-0.073, 0.399]	0.5229	<0.001
Base − Direct	-0.043	[-0.227, 0.140]	1.0000	<0.001
Full .50 − Full .72	-0.092	[-0.376, 0.192]	1.0000	<0.001
Full .50 − Direct	-0.298	[-0.527, -0.068]	0.0677	<0.001
Full .72 − Direct	-0.206	[-0.442, 0.030]	0.3489	<0.001

F.4 Human-corrected risk

Human-corrected evaluation changes the mention labels used to score the audited responses. The overall contrast weights conditions by their image counts. The human-corrected estimate retains the direction and magnitude of the automatic estimate, with a 95% interval that spans zero. The change in the pooled contrast is -0.00113 loss points with interval $[-0.00778, 0.00553]$.

Table 15: MOON-R minus Flat loss before and after human correction. The direction is retained in all five conditions.

Condition	Images	Auto ΔL	Human ΔL	Human 95% CI
Q2 → Q25	50	-0.05350	-0.05350	[-0.13309, 0.02609]
I2 → I25	50	-0.00150	-0.00625	[-0.11224, 0.09974]
L15 → L16	50	-0.02525	-0.01950	[-0.11331, 0.07431]
Q25 → NoCaps	25	-0.09800	-0.10400	[-0.22651, 0.01851]
L16 → AMBER	25	-0.01300	-0.01800	[-0.15018, 0.11418]
Overall	200	-0.03394	-0.03506	[-0.08467, 0.01454]

G Statistical sensitivity

G.1 Uncertainty across images and repeated fits

For each contrast, let D_{ri} be the paired AULC difference for fit r and test image i, after integrating over $\log_2 B$. With $R = 20$ fits and $N = 1{,}600$ images, the two-axis approximation is

$$\widehat{\mathrm{SE}}^2 = \frac{\mathrm{Var}_i(\overline{D}_{\cdot i})}{N} + \frac{\mathrm{Var}_r(\overline{D}_{r\cdot})}{R}, \qquad \mathrm{CI}_{95} = \overline{D} \pm t_{.975,R-1}\widehat{\mathrm{SE}}.$$

Images and fits are the two sampling units; the RN entries retain their crossed dependence. Coverage relies on this approximation to the sampling structure.

We assess sensitivity with 50,000 crossed-bootstrap replicates of the same paired records. Each replicate independently resamples the 20 fit slots and 1,600 images, then computes their Cartesian weighted mean. Methods and budgets share the resampling weights. Bootstrap intervals use the

2.5th and 97.5th percentiles, and bootstrap standard errors use the standard deviation of replicate means.

The 24 main-text readout contrasts have identical effect directions and pointwise 95% interval zero-exclusion decisions under both procedures. Their bootstrap/two-axis SE ratio has median 1.0226 and range 0.9977–1.0881. The six adjacent Signed-Recompose–Flat contrasts also agree throughout, with median ratio 1.0650 and range 1.0293–1.0952. All 30 intervals exclude zero. This is a same-record sensitivity analysis of pointwise intervals; Holm-adjusted tests retain the two-axis procedure and the comparison families in Appendix F.

Table 16 gives every bootstrap interval alongside its SE ratio. Table 7 gives the corresponding effects and two-axis intervals. Dividing the replicates into ten blocks of 5,000 gives a maximum estimated Monte Carlo standard error of 0.000191 loss points for interval endpoints and 0.000051 for bootstrap SEs across the 24 main-text contrasts; the corresponding maxima for the six adjacent contrasts are 0.000075 and 0.000023. These quantify numerical precision of the bootstrap calculation.

Table 16: Crossed-bootstrap sensitivity for all 30 readout curve contrasts. Intervals are pointwise 95% in loss points. The last column is bootstrap SE divided by two-axis SE.

Update	Window	Bootstrap 95% CI	SE ratio
MOON-R minus Flat-σ			
Q2 → Q25	8–64	[-0.059464, -0.014621]	0.9977
Q2 → Q25	8–128	[-0.056347, -0.016054]	0.9978
I2 → I25	8–64	[-0.058800, -0.023798]	1.0276
I2 → I25	8–128	[-0.053981, -0.023965]	1.0329
L15 → L16	8–64	[-0.048143, -0.016517]	1.0362
L15 → L16	8–128	[-0.048283, -0.019931]	1.0310
MOON-R minus Signed-Recompose			
Q2 → Q25	8–64	[-0.031676, -0.005771]	1.0183
Q2 → Q25	8–128	[-0.029928, -0.006907]	1.0170
I2 → I25	8–64	[-0.032105, -0.010648]	1.0520
I2 → I25	8–128	[-0.029675, -0.011098]	1.0555
L15 → L16	8–64	[-0.026475, -0.009569]	1.0881
L15 → L16	8–128	[-0.025649, -0.010261]	1.0793
MOON-R minus FreeMix-Source			
Q2 → Q25	8–64	[-0.037098, -0.002599]	1.0154
Q2 → Q25	8–128	[-0.034585, -0.005413]	1.0169
I2 → I25	8–64	[-0.064901, -0.023460]	1.0125
I2 → I25	8–128	[-0.054845, -0.021490]	1.0174
L15 → L16	8–64	[-0.037665, -0.006505]	1.0318
L15 → L16	8–128	[-0.033800, -0.008353]	1.0316
MOON-R minus FreeMix-Target			
Q2 → Q25	8–64	[-0.037600, -0.004588]	1.0222
Q2 → Q25	8–128	[-0.039042, -0.007949]	1.0211
I2 → I25	8–64	[-0.064513, -0.025697]	1.0229
I2 → I25	8–128	[-0.057261, -0.024937]	1.0295
L15 → L16	8–64	[-0.049641, -0.011992]	1.0120
L15 → L16	8–128	[-0.049286, -0.015057]	1.0142
Signed-Recompose minus Prior-matched Flat			
Q2 → Q25	8–64	[-0.030426, -0.007758]	1.0348
Q2 → Q25	8–128	[-0.028544, -0.008196]	1.0293
I2 → I25	8–64	[-0.029685, -0.011347]	1.0794
I2 → I25	8–128	[-0.027024, -0.010734]	1.0952
L15 → L16	8–64	[-0.024267, -0.005592]	1.0705
L15 → L16	8–128	[-0.024500, -0.008309]	1.0596

G.2 Equivalence on the scale of object errors

With 80 classes and false-positive weight $\rho = 4$, the reported loss has the direct error-count interpretation

$$L_{80} = \frac{100(4\mathrm{FP} + \mathrm{FN})}{5 \cdot 80N} = \frac{\mathrm{FP}}{N} + 0.25\frac{\mathrm{FN}}{N}.$$

A tolerance of 0.10 loss points therefore permits a net difference of ten FP-equivalent errors per 100 images. With only FP changes, this is ten FP per 100 images; with only FN changes, it is 40 FN per 100 images. The tolerance applies to weighted loss, so opposing FP and FN changes can offset one another. It is an application tolerance assumption, corresponding to about 5.6–6.7% of a loss of 1.5–1.8, whose suitability depends on the intended use.

Table 17 reports post-hoc sensitivity to tighter tolerances on the same 12 final-budget readout comparisons. For each margin, TOST takes the larger of the two one-sided p values, followed by Holm correction across the 12 comparisons. All comparisons meet the 0.05-point tolerance; seven meet 0.025. Equivalence conclusions consequently depend on the accepted weighted-error tolerance.

Table 17: Sensitivity of $B = 512$ equivalence to the loss tolerance. Each row uses the same 12 comparisons and its own Holm-12 adjustment.

Tolerance	Net FP-equivalents per 100 images	Passing	Largest adjusted p
± 0.100	10	12/12	1.114×10^{-9}
± 0.050	5	12/12	0.00030906
± 0.025	2.5	7/12	0.217552

H Complementary results and failure cases

H.1 Additional policy comparisons

Table 18: Natural backbone updates on COCO. L_{80} (lower is better), $B = 32$ for adapted policies; Base uses no audit. All methods share source information and target image subsets.

Method	Q2 → Q25	I2 → I25	L15 → L16
Base	2.916	3.389	3.743
Global	1.820	1.974	2.112
Flat source	1.698	1.753	1.959
Flat cell	1.702	1.797	1.928
Flat DeepSets	1.662	1.786	1.936
Flat residual	1.680	1.766	1.899
IG source	1.697	1.758	1.959
IG RT	1.661	1.712	1.905
IG+ local	**1.591**	**1.673**	**1.819**

Table 19: All 13 transfer conditions at $B = 32$. The displayed Flat comparator is residual regression. The last column is the largest adjusted p among five comparisons in that condition (source, cell, residual, DeepSets, and Global); correction covers all 65 comparisons.

Condition	IG+	Flat res.	ΔL_{80}	Max p_{Holm65}
Q2 → Q25	1.591	1.680	-0.090	0.013
I2 → I25	1.673	1.766	-0.093	0.003
L15 → L16	1.819	1.899	-0.080	0.003

Continued on next page

Table 19 continued

Condition	IG+	Flat res.	ΔL_{80}	Max p_{Holm65}
P1	1.564	1.616	-0.051	0.020
P2	1.522	1.598	-0.076	0.002
P3	1.535	1.612	-0.076	0.003
P4	1.533	1.629	-0.095	0.003
Q25-NoCaps	1.719	1.814	-0.095	<0.001
Q25-AMBER	1.746	1.855	-0.109	0.011
I25-NoCaps	1.827	1.909	-0.082	0.020
I25-AMBER	1.839	1.945	-0.106	0.020
L16-NoCaps	1.943	2.061	-0.118	0.003
L16-AMBER	2.011	2.148	-0.137	0.004

At $B = 32$, IG+ local improves on Flat source, cell, and residual predictors across all three natural updates; all nine paired contrasts pass Holm correction (Table 31). Against Flat residual, the losses are lower by 0.090/0.093/0.080 points. Table 19 extends these comparisons to all 13 prompt and domain conditions. All 65 contrasts against the five information-matched policies pass their family-wise correction. Prior-matched Flat is evaluated separately in Section 5.1 and the matched transfer study.

The stationary setting marks a useful boundary: at $B = 32$, IG+ local has loss 1.558 versus 1.550 for frozen Flat source. Adaptation can also fail under shift. At $B = 8$, local adaptation is worse than its source policy in 3/20 Qwen and 7/20 InternVL fits; several prompt shifts show the same behavior. Table 32 retains every tested condition and budget, including these failures and the cross-validation gate.

H.2 Intervention value and transport

Table 20: Paired action-effect prediction, averaged over 20 fits. All errors are in loss points. Bin error is a signed-utility reliability diagnostic, not probability ECE. All six IG+–Flat MAE contrasts pass Holm correction.

Update	Predictor	MAE	RMSE	Bin error
Q2 → Q25	Flat residual	1.294	1.666	0.347
Q2 → Q25	Flat DeepSets	1.297	1.666	0.330
Q2 → Q25	IG+ local	1.212	1.562	0.109
I2 → I25	Flat residual	1.361	1.745	0.303
I2 → I25	Flat DeepSets	1.355	1.736	0.312
I2 → I25	IG+ local	1.268	1.634	0.087
L15 → L16	Flat residual	1.419	1.804	0.298
L15 → L16	Flat DeepSets	1.563	1.981	0.576
L15 → L16	IG+ local	1.352	1.719	0.120

The structured readout lowers action-effect MAE and signed-risk reliability-bin error (Table 20); all six paired MAE contrasts pass correction (Table 33). Independently fitted reference policies have losses 1.546/1.650/1.769, placing IG+ local 0.045/0.023/0.049 points above them. These finite-data reference gaps measure headroom relative to the independently fitted policies.

Natural domain shifts affect several components at once. Class-presence rates are 11.33% on COCO, 9.02% on NoCaps, and 13.61% on AMBER. The four-component comparison favors Base updates more consistently than presence updates. On L16→AMBER, Base-only recomposition has loss 1.919 versus 1.925 with both components updated. Under greedy-to-sampling transfer, local adaptation improves on Base-only RT from 1.558 to 1.482, 1.825 to 1.722, and 1.987 to 1.878 for Q25/I25/L16. Table 34 displays all held-out transport residuals, including the large errors that limit exact reuse of source effects.

With only Base and Full .50 target responses, anchor factorization has lower mean loss in all 36 matched output/branch comparisons with Flat controls; nine pass Holm correction (Table 35). Macro differences against Flat anchor residual are -0.022 [95% CI: $-0.031, -0.014$] for guided-output budgets and -0.017 [$-0.025, -0.010$] for branch caps. Both pass correction within the exploratory 93-test family (Table 37). Guided-output matching and branch matching measure different acquisition resources because Full requires two decoder branches.

H.3 Evidence and external methods

Table 21: Evidence selection on the same 80-class candidate pool. Learned selectors target 90% precision on an independent calibration split; test precision is unconstrained.

Model	Selector	Precision (%)	Recall (%)
Q25	Intersection	95.09	47.56
Q25	Linear CA	90.05	79.20
Q25	Joint CA	90.38	81.25
Q25	Unknown-negative	88.11	63.02
I25	Intersection	94.46	49.82
I25	Linear CA	90.88	77.15
I25	Joint CA	90.36	82.67
I25	Unknown-negative	86.46	63.52
L16	Intersection	94.09	51.10
L16	Linear CA	90.52	77.28
L16	Joint CA	91.83	80.09
L16	Unknown-negative	86.28	62.60

Table 22: Selector–controller composition on Q25, $B = 128$. Both components contribute to the final loss. U_{80} and R_{80} are percentages.

Selector	Controller	$L_{80} \downarrow$	$U_{80} \downarrow$	$R_{80} \uparrow$
Linear CA	Full .50	1.912	17.03	87.29
Linear CA	Flat residual	1.856	16.54	87.14
Linear CA	IG+ local	1.796	16.02	87.40
Joint CA	Full .50	1.744	15.60	88.16
Joint CA	Flat residual	1.603	14.36	88.69
Joint CA	IG+ local	1.536	13.75	88.90

Joint CA has the highest evidence recall among the tested selectors on all three target backbones, with precision 90.38%/90.36%/91.83% (Table 21). Intersection provides higher precision with substantially lower recall. The composition panel separates evidence quality from action selection: replacing Linear with Joint CA lowers IG+ loss from 1.796 to 1.536; holding Joint CA fixed, replacing Flat residual with IG+ lowers it from 1.603 to 1.536.

Joint selection over Base and four thresholds paired with Direct or Full .50 gives loss 1.556, versus 1.638 for Joint Flat and 1.634 for staged selection at threshold 0.80. The paired differences pass the exploratory-family correction (Table 37). Relative to the staged controller, unsupported mentions decrease from 14.53% to 13.76%, while recall decreases from 87.06% to 86.61%.

Table 23: Shared-evidence comparison on L15/COCO at $B = 128$. "+C supplies the same canonical evidence. AFIP-public denotes the released head-guidance component.

Method	$L_{80} \downarrow$	$U_{80} \downarrow$	$R_{80} \uparrow$
Base	4.354	36.54	70.78
ILVAD	2.593	22.92	76.33
ILVAD+C	2.052	18.08	80.09
AFIP-public	2.526	22.34	76.33
AFIP-public+C	2.050	18.04	79.54
Flat residual	1.975	17.54	86.19
IG+ local	1.928	17.14	86.37

On LLaVA-1.5 with a native loader, MOON's loss of 1.928 is 6.05% below ILVAD+C and 5.93% below AFIP-public+C with the same canonical evidence (Xie et al., 2026; Li et al., 2026). Both paired intervals exclude zero. AFIP-public evaluates the released head-guidance component; the comparison uses the checkpoint and action budget specified in Appendix A.

H.4 Measurement, native endpoints, and runtime

Presence-annotation disagreement ranges from 3.20% to 5.10%, with Cohen's $\kappa = 0.797$–0.848 (Table 36). Training on single-annotator labels degrades IG+ less than Flat residual in all three updates, including after correction across the 93-test family (Table 37). This measures robustness to the observed annotation disagreements. The separate response-mention audit in Appendix F evaluates the parser used for policy scoring.

Native NoCaps CIDEr and SPICE improve across the three backbones; AMBER CHAIR decreases with higher coverage (Table 38). All 15 frozen-controller task comparisons meet the -2 point non-inferiority margin, with a worst simultaneous lower bound of -1.13 points (Table 39). The grouped HallusionBench results include no measured Q25 gain (Table 40).

Table 24: Measured warm latency (ms/image) on one NVIDIA RTX A6000, BF16, batch size 1. Controller time is included in IG+. Thirty warm-up images precede 200 images over five rounds.

Model	Base	Full .50	Flat res.	IG+	Controller
Q25	927.0	1682.7	1432.5	1367.7	0.74
I25	1034.8	1919.9	1593.9	1555.9	0.75
L16	1169.3	2156.5	1792.9	1733.7	0.73

Warm IG+ latency on an RTX A6000 is 1.368/1.556/1.734 seconds, above Base and below always applying Full .50 (Table 24). Controller computation is at most 0.75 ms in the displayed measurements. Table 41 separates cold expert loading from cached execution. Acquiring all four actions costs six decoder-branch trajectories per image; Table 30 includes the additional Base-pool supervision.

I Additional paired evidence

The following comparisons characterize low-budget adaptation, transport error, and practical performance. Loss differences are in points, and intervals are pointwise unless identified as simultaneous. IG+ denotes the cell-based controller in these panels; MOON-R denotes the matched neural readout. Each displayed main comparison family includes all its comparisons, and the transport panel covers every tested condition and action.

I.1 Common-prior attribution

Flat denotes Prior-matched Flat and Random denotes random partition with the common prior. Table 25 contains all 16 curve contrasts. Table 26 reports the separate eight-test TOST family. Re-

moving the common output mixture preserves all eight MOON–Flat curve advantages, whereas the mixture itself has no resolved effect for MOON on the six natural-update comparisons (Table 27).

Table 25: MOON-R minus each matched readout: all curve contrasts. Holm correction covers 16 tests.

Window	Comparator	ΔAULC	95% CI	p_{H16}
Stationary Q25				
8–64	Flat	-0.0691	[-0.0870, -0.0513]	1.80e-06
8–64	Random	-0.0706	[-0.0866, -0.0547]	2.59e-07
8–128	Flat	-0.0622	[-0.0782, -0.0463]	1.70e-06
8–128	Random	-0.0645	[-0.0790, -0.0501]	2.40e-07
Q2→Q25				
8–64	Flat	-0.0377	[-0.0614, -0.0140]	0.0035
8–64	Random	-0.0511	[-0.0665, -0.0358]	8.47e-06
8–128	Flat	-0.0370	[-0.0582, -0.0157]	0.0034
8–128	Random	-0.0490	[-0.0630, -0.0349]	5.20e-06
I2→I25				
8–64	Flat	-0.0414	[-0.0595, -0.0233]	5.22e-04
8–64	Random	-0.0587	[-0.0748, -0.0426]	3.27e-06
8–128	Flat	-0.0390	[-0.0547, -0.0233]	3.08e-04
8–128	Random	-0.0540	[-0.0680, -0.0400]	1.80e-06
L15→L16				
8–64	Flat	-0.0326	[-0.0488, -0.0165]	0.0013
8–64	Random	-0.0553	[-0.0705, -0.0401]	3.27e-06
8–128	Flat	-0.0343	[-0.0487, -0.0198]	4.25e-04
8–128	Random	-0.0518	[-0.0655, -0.0381]	2.22e-06

Table 26: All eight $B = 512$ comparisons meet equivalence within ± 0.10 loss points. TOST and Holm p values use the original eight-test family.

Comparator	ΔL	90% CI	TOST p	p_{H8}
Stationary Q25				
Flat	-0.0040	[-0.0159, 0.0078]	9.12e-12	4.56e-11
Random	-0.0171	[-0.0282, -0.0059]	4.06e-11	1.62e-10
Q2→Q25				
Flat	-0.0145	[-0.0284, -0.0007]	9.31e-10	9.31e-10
Random	-0.0110	[-0.0217, -0.0004]	5.17e-12	3.10e-11
I2→I25				
Flat	-0.0104	[-0.0235, 0.0027]	1.73e-10	5.19e-10
Random	-0.0221	[-0.0336, -0.0105]	2.16e-10	5.19e-10
L15→L16				
Flat	-0.0005	[-0.0103, 0.0092]	1.54e-13	1.23e-12
Random	-0.0118	[-0.0220, -0.0016]	2.80e-12	1.96e-11

Table 27: All 24 output-mixture contrasts. On–off switches only the common mixture; shared source-derived features remain. Fifteen tests pass Holm correction.

Window	Contrast	ΔAULC	95% CI	p_{H24}
Stationary Q25				
8–64	MOON on–off	-0.0245	[-0.0302, -0.0187]	6.19e-07
8–64	Flat on–off	-0.0325	[-0.0375, -0.0274]	8.60e-10
8–64	MOON–Flat, off	-0.0772	[-0.0946, -0.0597]	3.90e-07
8–128	MOON on–off	-0.0188	[-0.0233, -0.0142]	1.03e-06
8–128	Flat on–off	-0.0257	[-0.0298, -0.0216]	1.30e-09
8–128	MOON–Flat, off	-0.0692	[-0.0850, -0.0534]	4.34e-07
Q2→Q25				
8–64	MOON on–off	-0.0011	[-0.0173, 0.0152]	1.0000
8–64	Flat on–off	-0.0202	[-0.0333, -0.0070]	0.0455
8–64	MOON–Flat, off	-0.0568	[-0.0854, -0.0282]	0.0060
8–128	MOON on–off	-0.0002	[-0.0139, 0.0135]	1.0000
8–128	Flat on–off	-0.0157	[-0.0266, -0.0047]	0.0663
8–128	MOON–Flat, off	-0.0524	[-0.0773, -0.0275]	0.0036
I2→I25				
8–64	MOON on–off	-0.0161	[-0.0317, -0.0004]	0.3583
8–64	Flat on–off	-0.0283	[-0.0390, -0.0176]	3.55e-04
8–64	MOON–Flat, off	-0.0536	[-0.0733, -0.0339]	2.56e-04
8–128	MOON on–off	-0.0126	[-0.0262, 0.0009]	0.4617
8–128	Flat on–off	-0.0222	[-0.0314, -0.0130]	9.35e-04
8–128	MOON–Flat, off	-0.0486	[-0.0662, -0.0311]	2.26e-04
L15→L16				
8–64	MOON on–off	0.0098	[-0.0151, 0.0346]	1.0000
8–64	Flat on–off	-0.0175	[-0.0364, 0.0014]	0.4617
8–64	MOON–Flat, off	-0.0599	[-0.0814, -0.0383]	2.26e-04
8–128	MOON on–off	0.0088	[-0.0125, 0.0302]	1.0000
8–128	Flat on–off	-0.0144	[-0.0308, 0.0019]	0.4617
8–128	MOON–Flat, off	-0.0575	[-0.0761, -0.0390]	5.78e-05

I.2 Semantic supervision, transfer, and acquisition

The semantic, transfer, Base-pool, and loss-weight families contain 3, 16, 12, and 12 primary contrasts; MOON-S versus MOON-R is a separate three-test auxiliary family. MOON-R is the first method except in the semantic rows. The $N_{\text{Base}} = 512, \rho = 4$ anchor is shared by both sensitivity families. The smallest Base pool is a secondary boundary check rather than part of the 12-test primary Base-pool family.

Table 28: All 46 primary and auxiliary extension curve contrasts. The p_H column uses the family named in each panel.

Condition	Setting	Window	ΔAULC	95% CI	p_H
Semantic: MOON-S minus Semantic-shuffled (Holm-3)					
Q2→Q25	S–shuffled	8–128	-0.0386	[-0.0515, -0.0258]	4.76e-06
I2→I25	S–shuffled	8–128	-0.0473	[-0.0600, -0.0347]	6.67e-07
L15→L16	S–shuffled	8–128	-0.0424	[-0.0540, -0.0309]	6.67e-07
Prompt and domain transfer (Holm-16)					
Prompt-P4*	Flat	8–64	-0.0572	[-0.0763, -0.0381]	4.55e-05
Prompt-P4*	Random	8–64	-0.0598	[-0.0766, -0.0430]	6.18e-06

Continued on next page

Table 28 continued

Condition	Setting	Window	ΔAULC	95% CI	p_H
Prompt-P4*	Flat	8–128	-0.0524	[-0.0687, -0.0361]	1.99e-05
Prompt-P4*	Random	8–128	-0.0536	[-0.0684, -0.0388]	5.19e-06
Q25 → NoCaps	Flat	8–64	-0.0455	[-0.0668, -0.0242]	0.0013
Q25 → NoCaps	Random	8–64	-0.0554	[-0.0692, -0.0416]	1.23e-06
Q25 → NoCaps	Flat	8–128	-0.0404	[-0.0583, -0.0225]	8.69e-04
Q25 → NoCaps	Random	8–128	-0.0517	[-0.0638, -0.0396]	4.82e-07
I25 → AMBER	Flat	8–64	-0.0323	[-0.0604, -0.0041]	0.0335
I25 → AMBER	Random	8–64	-0.0421	[-0.0639, -0.0204]	0.0020
I25 → AMBER	Flat	8–128	-0.0298	[-0.0536, -0.0060]	0.0335
I25 → AMBER	Random	8–128	-0.0396	[-0.0586, -0.0206]	0.0013
L16 → AMBER	Flat	8–64	-0.0864	[-0.1216, -0.0512]	4.09e-04
L16 → AMBER	Random	8–64	-0.0832	[-0.1085, -0.0579]	1.61e-05
L16 → AMBER	Flat	8–128	-0.0769	[-0.1053, -0.0485]	1.45e-04
L16 → AMBER	Random	8–128	-0.0744	[-0.0966, -0.0523]	1.29e-05
Base-pool sensitivity, $\rho = 4$ (Holm-12)					
Q2→Q25	$N = 64$	16–64	-0.0319	[-0.0530, -0.0108]	0.0333
Q2→Q25	$N = 128$	16–64	-0.0344	[-0.0588, -0.0100]	0.0342
Q2→Q25	$N = 256$	16–64	-0.0355	[-0.0609, -0.0102]	0.0342
Q2→Q25	$N = 512$	16–64	-0.0346	[-0.0604, -0.0088]	0.0342
I2→I25	$N = 64$	16–64	-0.0325	[-0.0474, -0.0175]	0.0027
I2→I25	$N = 128$	16–64	-0.0325	[-0.0492, -0.0159]	0.0056
I2→I25	$N = 256$	16–64	-0.0362	[-0.0533, -0.0191]	0.0032
I2→I25	$N = 512$	16–64	-0.0390	[-0.0578, -0.0202]	0.0035
L15→L16	$N = 64$	16–64	-0.0268	[-0.0452, -0.0083]	0.0342
L15→L16	$N = 128$	16–64	-0.0281	[-0.0475, -0.0087]	0.0342
L15→L16	$N = 256$	16–64	-0.0310	[-0.0512, -0.0107]	0.0333
L15→L16	$N = 512$	16–64	-0.0314	[-0.0516, -0.0112]	0.0333
Loss-weight sensitivity, $N_{\text{Base}} = 512$ (Holm-12)					
Q2→Q25	$\rho = 1$	16–64	-0.0341	[-0.0566, -0.0116]	0.0249
Q2→Q25	$\rho = 2$	16–64	-0.0370	[-0.0622, -0.0117]	0.0249
Q2→Q25	$\rho = 4$	16–64	-0.0346	[-0.0604, -0.0088]	0.0249
Q2→Q25	$\rho = 8$	16–64	-0.0340	[-0.0596, -0.0084]	0.0249
I2→I25	$\rho = 1$	16–64	-0.0420	[-0.0581, -0.0259]	3.47e-04
I2→I25	$\rho = 2$	16–64	-0.0425	[-0.0597, -0.0254]	5.59e-04
I2→I25	$\rho = 4$	16–64	-0.0390	[-0.0578, -0.0202]	0.0035
I2→I25	$\rho = 8$	16–64	-0.0339	[-0.0526, -0.0152]	0.0109
L15→L16	$\rho = 1$	16–64	-0.0355	[-0.0558, -0.0151]	0.0136
L15→L16	$\rho = 2$	16–64	-0.0329	[-0.0535, -0.0122]	0.0248
L15→L16	$\rho = 4$	16–64	-0.0314	[-0.0516, -0.0112]	0.0249
L15→L16	$\rho = 8$	16–64	-0.0305	[-0.0503, -0.0107]	0.0249
Auxiliary: MOON-S minus MOON-R (Holm-3)					
Q2→Q25	S–R	8–128	-0.0089	[-0.0235, 0.0057]	0.2159
I2→I25	S–R	8–128	-0.0108	[-0.0232, 0.0016]	0.1707
L15→L16	S–R	8–128	-0.0079	[-0.0153, -0.0006]	0.1093

Table 29: All realizable points at 32 Base images: MOON-R minus Flat. Adjusted values retain the original 42-test secondary Base-pool family. Requesting $B = 64$ necessarily acquires 64 Base images and is the shared $64, 64$ allocation.

Update	B	ΔL	95% CI	p_{H42}
Q2→Q25	16	-0.0353	[-0.0593, -0.0113]	0.1862
Q2→Q25	32	-0.0214	[-0.0467, 0.0039]	1.0000

Continued on next page

Table 29 continued

Update	B	ΔL	95% CI	p_{H42}
I2→I25	16	-0.0051	[-0.0357, 0.0254]	1.0000
I2→I25	32	-0.0436	[-0.0676, -0.0196]	0.0467
L15→L16	16	-0.0326	[-0.0591, -0.0061]	0.2965
L15→L16	32	-0.0092	[-0.0374, 0.0190]	1.0000

Table 30: Target acquisition per fit. Actual Base size is at least B; every paired image adds three guided responses and five branches. Source acquisition adds 9,600 responses and 14,400 branches to every row. Labels count 80 presence judgments per actual Base image.

Base req.	B	Base actual	Responses	Branches	Labels
32	16	32	80	112	2560
32	32	32	128	192	2560
32	64	64	256	384	5120
64	16	64	112	144	5120
64	32	64	160	224	5120
64	64	64	256	384	5120
128	16	128	176	208	10240
128	32	128	224	288	10240
128	64	128	320	448	10240
256	16	256	304	336	20480
256	32	256	352	416	20480
256	64	256	448	576	20480
512	16	512	560	592	40960
512	32	512	608	672	40960
512	64	512	704	832	40960

I.3 Adaptation, effect prediction, and transport

The information-matched comparisons evaluate adaptation across different effect estimators. Negative adaptation is counted within the 20 paired fits; the cross-validation gate uses only the target audit. Its loss is not uniformly below always adapting. The action-effect MAE contrasts form a separate six-test family.

Table 31: All nine information-matched update contrasts at $B = 32$: IG+ local minus comparator, with Holm-9 adjustment.

Comparator	ΔL	95% CI	p_{H9}
Q2→Q25			
Flat source	-0.108	[-0.150, -0.066]	1.76e-04
Flat cell	-0.112	[-0.150, -0.074]	5.59e-05
Flat residual	-0.090	[-0.133, -0.047]	5.70e-04
I2→I25			
Flat source	-0.079	[-0.115, -0.044]	5.12e-04
Flat cell	-0.124	[-0.172, -0.076]	1.76e-04
Flat residual	-0.093	[-0.128, -0.058]	1.52e-04
L15→L16			
Flat source	-0.141	[-0.187, -0.094]	4.24e-05
Flat cell	-0.109	[-0.151, -0.068]	1.52e-04
Flat residual	-0.080	[-0.118, -0.042]	5.70e-04

Table 32: Local minus source loss at all tested adaptation conditions and budgets. Negative fits count local loss above source loss. Intervals are pointwise.

B	ΔL	95% CI	CV gate L	Neg./20
Q2→Q25				
8	-0.032	[-0.065, 0.000]	1.667	3
32	-0.108	[-0.150, -0.066]	1.618	1
128	-0.154	[-0.198, -0.111]	1.589	0
I2→I25				
8	-0.022	[-0.055, 0.010]	1.729	7
32	-0.079	[-0.115, -0.044]	1.695	0
128	-0.091	[-0.134, -0.048]	1.701	0
L15→L16				
8	-0.065	[-0.099, -0.032]	1.908	1
32	-0.141	[-0.187, -0.094]	1.875	0
128	-0.190	[-0.244, -0.136]	1.851	0
Prompt-P1				
8	-0.019	[-0.053, 0.014]	1.600	7
32	-0.063	[-0.102, -0.024]	1.593	3
128	-0.105	[-0.147, -0.062]	1.588	0
Prompt-P2				
8	-0.026	[-0.056, 0.005]	1.594	7
32	-0.091	[-0.123, -0.058]	1.584	0
128	-0.117	[-0.155, -0.079]	1.576	0
Prompt-P3				
8	-0.034	[-0.067, 0.000]	1.594	6
32	-0.079	[-0.115, -0.042]	1.588	0
128	-0.102	[-0.145, -0.059]	1.583	0
Prompt-P4				
8	-0.026	[-0.062, 0.010]	1.586	7
32	-0.084	[-0.121, -0.048]	1.544	1
128	-0.100	[-0.143, -0.057]	1.538	0

Table 33: All paired action-effect MAE contrasts: IG+ local minus comparator.

Comparator	ΔMAE	95% CI	p_{H6}	Wins/20
Q2→Q25				
Flat residual	-0.083	[-0.128, -0.037]	0.006	17
Flat DeepSets	-0.086	[-0.151, -0.021]	0.013	17
I2→I25				
Flat residual	-0.093	[-0.144, -0.041]	0.006	19
Flat DeepSets	-0.086	[-0.131, -0.042]	0.004	19
L15→L16				
Flat residual	-0.067	[-0.111, -0.023]	0.010	19
Flat DeepSets	-0.211	[-0.343, -0.078]	0.010	18

Held-out transport. Calibration/held-out image counts are 400/1,200 for COCO, 400/2,600 for NoCaps, and 200/404 for AMBER. Residual is observed minus predicted action effect, with the fitted calibration heads held fixed. Table 34 retains all 54 residuals. The largest absolute residual is the Q25–NoCaps anchor prediction for Full .72: -0.495 with 95% CI $[-0.589, -0.401]$. This is

a descriptive maximum across the panel; its interval is pointwise and excludes calibration-refitting uncertainty.

Table 34: Complete held-out residual panel, in loss points. Base RT reuses conditional flips; Anchor .50 uses the target anchor. Small residuals in some cells coexist with substantial transport error in others.

Condition	Estimator	Full .50	Full .72	Direct
Q2→Q25	Base RT	-0.002	0.137	0.011
Q2→Q25	Anchor .50	-0.161	0.089	-0.139
I2→I25	Base RT	-0.002	-0.150	0.012
I2→I25	Anchor .50	-0.123	-0.240	-0.096
L15→L16	Base RT	-0.036	0.205	0.119
L15→L16	Anchor .50	0.012	0.422	0.186
Q25-NoCaps	Base RT	-0.044	-0.183	-0.053
Q25-NoCaps	Anchor .50	-0.173	-0.495	-0.247
Q25-AMBER	Base RT	0.129	0.093	-0.035
Q25-AMBER	Anchor .50	-0.194	-0.433	-0.413
I25-NoCaps	Base RT	0.026	-0.208	-0.083
I25-NoCaps	Anchor .50	-0.101	-0.413	-0.219
I25-AMBER	Base RT	0.297	0.202	0.218
I25-AMBER	Anchor .50	0.010	-0.338	-0.097
L16-NoCaps	Base RT	-0.004	0.086	-0.008
L16-NoCaps	Anchor .50	-0.102	-0.090	-0.108
L16-AMBER	Base RT	0.050	-0.022	0.109
L16-AMBER	Anchor .50	-0.215	-0.463	-0.203

Single-anchor budgets. Full .50 is the only guided target label for the anchor controllers. The output budget counts guided responses, while the branch cap charges Base and decoding branches. All-action fitting uses three guided responses per paired image; Full uses two branches and Direct one. The shared source prior is supplied to all compared controllers. Every condition-specific mean difference in Table 35 favors anchor factorization, and 9/36 survive correction.

Table 35: All single-anchor contrasts. F-A is Flat anchor residual; F-all is the all-action Flat residual. The original 36-test correction is retained.

Budget	Comparator	ΔL	95% CI	p_{H36}
Guided outputs: Q2→Q25				
96	F-A	-0.023	[-0.049, 0.002]	0.871
96	F-all	-0.117	[-0.151, -0.083]	<0.001
384	F-A	-0.038	[-0.059, -0.018]	0.033
384	F-all	-0.057	[-0.086, -0.028]	0.021
1536	F-A	-0.018	[-0.041, 0.004]	0.986
1536	F-all	-0.034	[-0.060, -0.008]	0.308
Guided outputs: I2→I25				
96	F-A	-0.035	[-0.065, -0.006]	0.494
96	F-all	-0.093	[-0.125, -0.061]	<0.001
384	F-A	-0.009	[-0.022, 0.004]	0.986
384	F-all	-0.051	[-0.084, -0.018]	0.125
1536	F-A	-0.007	[-0.019, 0.005]	0.986
1536	F-all	-0.028	[-0.048, -0.007]	0.266
Guided outputs: L15→L16				
96	F-A	-0.046	[-0.081, -0.010]	0.331
96	F-all	-0.122	[-0.165, -0.078]	<0.001

Continued on next page

Table 35 continued

Budget	Comparator	ΔL	95% CI	p_{H36}
384	F-A	-0.019	[-0.043, 0.005]	0.986
384	F-all	-0.055	[-0.086, -0.024]	0.041
1536	F-A	-0.006	[-0.021, 0.009]	0.986
1536	F-all	-0.015	[-0.032, 0.003]	0.986
Branch cap: Q2→Q25				
1024	F-A	-0.022	[-0.044, -0.001]	0.687
1024	F-all	-0.063	[-0.096, -0.030]	0.026
2048	F-A	-0.016	[-0.032, 0.001]	0.850
2048	F-all	-0.032	[-0.055, -0.008]	0.266
4096	F-A	-0.012	[-0.028, 0.003]	0.986
4096	F-all	-0.021	[-0.043, 0.000]	0.749
Branch cap: I2→I25				
1024	F-A	-0.026	[-0.059, 0.008]	0.986
1024	F-all	-0.058	[-0.091, -0.026]	0.038
2048	F-A	-0.003	[-0.020, 0.014]	0.986
2048	F-all	-0.029	[-0.055, -0.003]	0.588
4096	F-A	-0.020	[-0.040, -0.001]	0.687
4096	F-all	-0.016	[-0.042, 0.011]	0.986
Branch cap: L15→L16				
1024	F-A	-0.022	[-0.043, -0.001]	0.640
1024	F-all	-0.054	[-0.080, -0.029]	0.009
2048	F-A	-0.016	[-0.035, 0.003]	0.977
2048	F-all	-0.037	[-0.069, -0.004]	0.588
4096	F-A	-0.019	[-0.036, -0.002]	0.588
4096	F-all	-0.022	[-0.043, -0.002]	0.588

I.4 Annotation noise and joint action selection

For annotation reliability, two annotators independently label 300 images and 80 classes per condition; a third adjudicates disagreements. This yields 24,000 binary comparisons per row of Table 36. Training on single-annotator labels instead of adjudicated labels raises IG+ loss by 0.053/0.059/0.062 points and Flat residual loss by 0.113/0.114/0.142 across the three updates. Table 37 preserves the corresponding paired degradation contrasts, joint-recipe comparisons, and anchor macro contrasts.

These analyses are exploratory. The 93-test family was assembled after the initial results and comprises 24 AULC, five composition, three joint-recipe, three annotation, six decoding, nine reference-gap, 36 source-budget, four anchor-macro, and three annotation-robustness contrasts. Table 37 retains adjustment over this complete family. Holm correction addresses multiplicity within that selected family; the exploratory interpretation also accounts for its post-hoc selection. The crossed-bootstrap sensitivity in Appendix G reports intervals and standard errors.

Table 36: Presence-annotation agreement. Count columns denote paired positive/negative judgments; disagreement is a percentage.

Condition	$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$	Disagree	κ
Q2→Q25	2418	344	424	20814	3.20	0.845
I2→I25	2814	384	471	20331	3.56	0.848
L15→L16	3296	450	541	19713	4.13	0.845
Q25-NoCaps	2495	473	454	20578	3.86	0.821
Q25-AMBER	2822	503	535	20140	4.32	0.820
I25-NoCaps	3258	518	524	19700	4.34	0.836

Continued on next page

Table 36 continued

Condition	$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$	Disagree	κ
I25-AMBER	2384	540	498	20578	4.32	0.797
L16-NoCaps	2871	568	571	19990	4.75	0.807
L16-AMBER	3257	609	615	19519	5.10	0.811

Table 37: Joint IG minus comparator at $B = 128$; IG+ anchor minus comparator averaged equally over families/budgets; and IG+ minus Flat degradation under annotation noise at $B = 32$. All rows of these three comparison types are shown. Exploratory comparisons retain Holm adjustment over the complete 93-test family.

Comparison	Comparator	ΔL	95% CI	p_{H93}
Joint recipes	Global	-0.180	[-0.221, -0.139]	<0.001
Joint recipes	Staged recipe	-0.078	[-0.120, -0.037]	0.010
Joint recipes	Joint Flat	-0.083	[-0.115, -0.050]	0.001
Anchor: outputs	Flat anchor	-0.022	[-0.031, -0.014]	0.001
Anchor: outputs	Flat residual	-0.063	[-0.075, -0.052]	<0.001
Anchor: branches	Flat anchor	-0.017	[-0.025, -0.010]	0.004
Anchor: branches	Flat residual	-0.037	[-0.046, -0.027]	<0.001
Noise: Q2→Q25	Flat degradation	-0.060	[-0.099, -0.022]	0.018
Noise: I2→I25	Flat degradation	-0.054	[-0.085, -0.023]	0.013
Noise: L15→L16	Flat degradation	-0.081	[-0.112, -0.049]	0.001

I.5 Native benchmarks and operating costs

These endpoints use each benchmark's native metric, separately from the shared 80-class loss. Controllers remain frozen after COCO fitting. For the five downstream tasks, the non-inferiority margin is -2 percentage points and the one-sided lower bounds use Bonferroni adjustment across all 15 model–task comparisons. POPE uses image-cluster resampling; GQA is the 1,000-question validation subset. OCRBench sums are divided by ten for percentage reporting. HallusionBench retains its question, figure, and scene grouping.

Table 38: All native caption endpoints: NoCaps (3,000 images) and AMBER (604 images). AMBER CHAIR and coverage are percentages.

Model	Method	CIDEr	SPICE $\times 100$	CHAIR	Coverage
Q25	Base	95.05	17.07	10.68	75.73
Q25	Flat residual	97.16	17.49	8.02	79.44
Q25	IG+ local	99.57	17.98	7.38	80.91
I25	Base	94.94	17.23	11.08	77.72
I25	Flat residual	97.03	17.62	8.39	80.55
I25	IG+ local	99.66	18.14	7.65	80.71
L16	Base	95.57	17.34	10.48	76.81
L16	Flat residual	97.69	17.72	7.79	79.12
L16	IG+ local	100.19	18.22	7.36	80.94

Table 39: All downstream task comparisons. Scores and differences are percentage points. The 95% intervals are pointwise; NI lower bounds are simultaneous across 15 comparisons.

Task	N	Base	MOON	Δ pp	95% CI	NI lower
Q25						
MMStar	1500	49.80	50.87	1.07	[0.33, 1.80]	0.07
RealWorldQA	765	64.71	65.10	0.39	[-0.65, 1.44]	-1.13
GQA subset	1000	65.90	66.70	0.80	[-0.10, 1.70]	-0.40
POPE adv.	3000	87.57	88.37	0.80	[0.30, 1.33]	0.10
OCRBench	1000	67.60	68.20	0.60	[-0.10, 1.30]	-0.30
I25						
MMStar	1500	49.93	51.33	1.40	[0.67, 2.20]	0.40
RealWorldQA	765	62.75	63.53	0.78	[-0.26, 1.83]	-0.65
GQA subset	1000	60.60	61.40	0.80	[0.10, 1.60]	-0.10
POPE adv.	3000	84.43	85.40	0.97	[0.40, 1.53]	0.18
OCRBench	1000	65.20	65.60	0.40	[-0.30, 1.10]	-0.60
L16						
MMStar	1500	46.13	47.33	1.20	[0.53, 1.93]	0.27
RealWorldQA	765	57.78	58.69	0.92	[0.00, 1.96]	-0.39
GQA subset	1000	55.00	55.90	0.90	[0.10, 1.70]	-0.20
POPE adv.	3000	81.23	82.57	1.33	[0.80, 1.90]	0.53
OCRBench	1000	60.30	60.80	0.50	[-0.30, 1.30]	-0.60

Table 40: HallusionBench: 1,129 questions, 346 figures, and 167 scene sets. Paired scores are Base/MOON; intervals concern the change in aAcc.

Model	aAcc (%)	fAcc (%)	qAcc (%)	ΔaAcc 95% CI
Q25	72.01 / 72.01	42.20 / 42.20	45.71 / 45.71	[-0.80, 0.85]
I25	71.39 / 71.74	40.75 / 41.33	42.86 / 43.52	[-0.36, 1.13]
L16	68.73 / 69.44	38.73 / 38.73	39.78 / 41.10	[-0.17, 1.59]

Warm Base, Full .50, Flat residual, and IG+ latencies appear in Table 24. Table 41 separates expert loading and cached evidence for the same three target models. Measurements use one RTX A6000, BF16, batch size one, 200 images, and five rounds; round variation is descriptive.

Table 41: Cold and cached IG+ operating modes for all three target backbones. Cold time includes expert loading; warm no-cache is the standard deployment measurement.

Model	Mode	Mean ms	p95 ms	Round SD
Q25	Cold incl expert load	6063.71	7803.47	45.32
Q25	Warm no cache	1367.65	1841.47	19.58
Q25	Warm evidence cached	1155.24	1507.11	13.88
I25	Cold incl expert load	6044.56	7791.57	63.71
I25	Warm no cache	1555.85	2161.88	11.89
I25	Warm evidence cached	1163.93	1519.21	12.22
L16	Cold incl expert load	6007.66	7691.28	69.15
L16	Warm no cache	1733.66	2393.15	25.35
L16	Warm evidence cached	1168.61	1488.65	27.36