UNSUPERVISED MULTIMODAL REWARD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language reward models are trained on annotated preference pairs, which are expensive to collect and can disagree with each other, and are compared by their accuracy on preference benchmarks. We build one with no human preference label: a linear head on an intermediate layer of a frozen vision-language backbone, trained to score a web caption above a perturbed version of the same caption; labeled data enters only to select the layer. Because the backbone is frozen, the objective is convex and its optimum is a regularized approximation to the log density ratio between captions and the perturbation sampler, so the sampler is what specifies the reward. At 32B the head beats GPT-4o on VL-RewardBench and is level with it on Multimodal RewardBench and MMRB2. The benchmark score, however, does not identify the reward: two heads trained on halves of the same perturbations tie on the benchmark and rank hallucinations apart, and neither improves Best-of-N captioning over greedy decoding. Retraining the head on the policy's own verifier-labelled candidates does, and every reward benchmark scores that head below the probe it replaced. A reward model is therefore worth reporting with the comparison it was trained on and the deployment it serves, and selecting one by benchmark rank alone selects for neither.

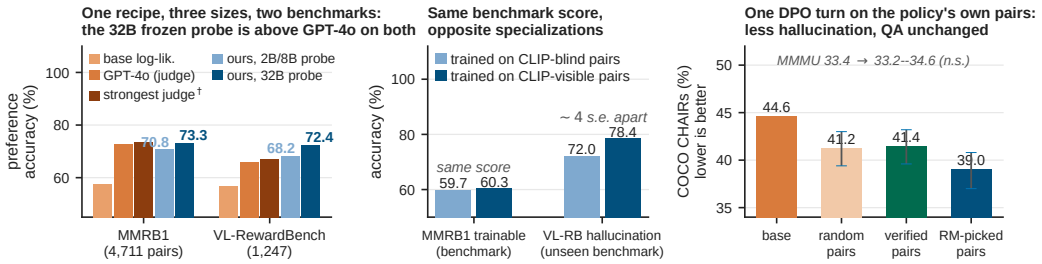

Figure 1: **(a)** One frozen-backbone probe per size, trained once on caption perturbations and evaluated with no re-tuning. **(b)** Two probes trained on halves of the same data, split by whether CLIP can see the perturbation, tie on the training benchmark and land 6.4 points apart on unseen hallucination judging. **(c)** One turn of the bootstrap loop on the policy: DPO on pairs the policy generated itself lowers its greedy hallucination rate (COCO CHAIRs) while MathVista and MMMU stay at the base level (Section 3.4, Table 8).

1 INTRODUCTION

Reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) has made the reward model the channel through which preferences enter post-training. For vision-language models the same design is now standard: a scalar scorer over an image, a prompt and a response, trained on annotated preference pairs and ranked on preference benchmarks (Yasunaga et al., 2025; Li et al., 2024; Hu et al., 2025). Two costs come with it. Multimodal preference pairs are expensive to annotate and the annotations disagree with each other (Yu et al., 2024b); and once the model is built, benchmark accuracy is the only currency in which it is compared, selected and trusted, although nobody has asked what sets that number for a vision-language reward model, or what it predicts.

Two observations from text guide the design. A reward is a comparison before it is a number: a score of .8 has no stable meaning across prompts, candidate pools or checkpoints, and the scalar hides the relative preference that was always the operative signal (Zhang, 2026), so what a reward model learns is set by which comparisons it is trained on. And those comparisons need no annotator:

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

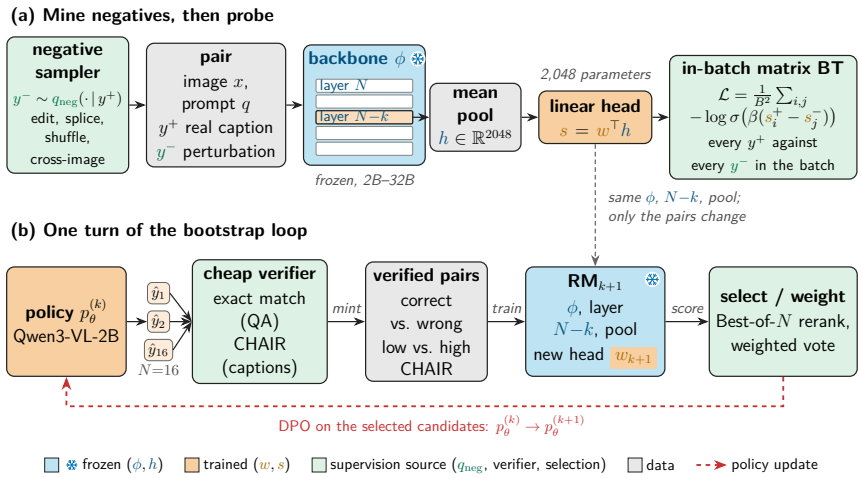

Figure 2: **(a) The reward from a linear probe.** A pair (image, prompt, real caption, same-image perturbation) passes once through the frozen backbone ϕ; the layer-$N-k$ states are mean-pooled over response tokens into h and scored by a bias-free head w, trained with in-batch matrix Bradley–Terry; the sampler q_{neg} is the only place supervision enters. **(b) One turn of the bootstrap loop** (Section 3.4): cheap verifiers label the policy's own Best-of-N candidates and the same head is retrained on the minted pairs. Blue frozen, orange trained, green supervision.

web continuations as chosen against in-batch continuations as rejected train a text reward model with no label at all (Fan et al., 2026). We take both to the vision-language setting, make the comparison the object of study, and remove the human labels entirely: the reward model is a linear head on the hidden states of an intermediate layer of a frozen backbone, trained by in-batch Bradley–Terry (Bradley & Terry, 1952) against same-image perturbations of web captions or same-seed re-renders of a perturbed prompt, with no human preference label (Fig. 2, Section 2). The claim is about the training data: the probe layer is selected on one labeled benchmark, which we therefore treat as development and never as transfer evidence, and the loop of Section 3.4 uses verifier labels, not annotators. Two things then decide what the head learns: the representation it reads, which sets the signal a linear head can extract, and the comparison distribution it is trained on, which sets the direction of the reward. Every choice in the paper is one of the two, and the verified policy errors of Section 3.4 are the second variable's final instance. A head of a few thousand parameters built this way beats GPT-4o by 6.6 points on VL-RewardBench at 32B, for about a thousandth of the compute of full backbone fine-tuning, and is level with GPT-4o on Multimodal RewardBench and MMRB2; at an equal pair count the mined pairs teach a frozen head more than curated RLAIF-V pairs do.

Because the backbone is frozen, the objective is transparent enough to say what that number is. It is ranking noise-contrastive estimation over fixed features (Gutmann & Hyvärinen, 2010; Ma & Collins, 2018), so it is convex with a unique minimizer, and that minimizer is an average of the negative sampler's difference vectors, weighted by how badly the head still misorders each pair (Section 2.2). No optimizer setting moves that minimizer, and in our runs nothing else on the training side, pool size, steps, curriculum or loss form, moves the benchmark either, so the representation and the sampler are the two choices that matter. The representation is the backbone's, and what we find about it repeats across backbones: on four Qwen3-VL sizes and on a second family, GLM-4.1V-9B, the best probe layer sits well below the top, twelve layers down for every backbone of 36 layers or more, and the same recipe with no re-tuning lifts each of them far above its own untrained judge (Section 3.2). The sampler is the other choice, and it is where the benchmark stops being informative.

Which pairs train the head decides which reward it learns. Two probes trained on halves of the same perturbations, split by whether CLIP can see the edit, tie on the benchmark and rank hallucinations 6.4 points apart (Fig. 1b). Deployed for Best-of-N reranking, benchmark rank orders utility in neither direction, and where the probe does beat majority vote, by 7 to 11 points, a training-free log-likelihood judge matches it. The analysis leaves one sampler to try: the policy's own errors. We mint those labels with cheap verifiers on Best-of-N candidates, for captions and for generated images; the resulting head is the only scorer in the study that beats greedy decoding on hallucination,

and every reward benchmark scores it below the probe it replaced. One DPO turn on the policy's own pairs also cuts its hallucination by 3 to 6 CHAIR points at no QA cost (Fig. 1c).

Our contributions combine probing, negative sampling and self-improvement, each studied on the policy side before (Section 4), into one reward-side recipe that uses no human preference label. (i) *Probed, not trained in:* a frozen vision-language backbone with a 2,048-parameter head at a depth set by the backbone, at or above GPT-4o on three benchmarks at 32B (Section 3.2). (ii) *Supervision as a sampler:* same-image perturbations written as a mixture of operators that apply to captions and to rendered images alike, with an argument for which reward direction they select and an ablation of the operators in both modalities (Section 2.2, Table 3). (iii) *What the score tracks:* the plateau is the convex optimum, the learned direction is set by the sampler, and benchmark rank predicts Best-of-N rank in neither direction (Section 3.3). (iv) *A bootstrap loop* that replaces annotators with cheap verifiers on the policy's own candidates, the one supervision that moved deployment, shown for captions and for image generation (Section 3.4).

2 Methodology

2.1 Algorithm

Recall that we hypothesize a frozen backbone already carries the reward. We now write that claim down as a model, and it is small enough to state in one line: a frozen Qwen3-VL backbone (Qwen Team, 2025a), a linear probe on layer $N-k$, mean pooling over the response tokens, and a bias-free linear head. The triple (layer, pooling, head) is the design space the rest of the paper sweeps (Fig. 2a; colours as in its caption). For an image x, a prompt q and a response y, let $h_t^{(N-k)}$ be the hidden state of layer $N-k$ at response token t; the reward is

$$s_w(x, q, y) \;=\; w^\top \frac{1}{|y|} \sum_{t \in y} h_t^{(N-k)}, \tag{1}$$

so the 2B model trains 2,048 parameters and the backbone none. A micro-batch of B preference pairs $\{(x_i, q_i, y_i^+, y_i^-)\}_{i=1}^B$ gives the in-batch matrix Bradley–Terry loss, every chosen response against every rejected one,

$$\mathcal{L} = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} -\log \sigma\big(\beta\,(s_i^+ - s_j^-)\big), \tag{2}$$

with β=0.1 the inverse temperature and B=8, so each positive meets eight negatives per step rather than one. Note that supervision enters only through the negative sampler, which is what makes the sampler the object Section 2.2 reasons about. For a positive (x, q, y^+) the rejected response is drawn from a mixture of perturbation operators applied to the positive itself,

$$y^- \;\sim\; q_{\text{neg}}(\cdot \mid x, q, y^+) \;=\; \sum_o \pi_o\, \mathcal{T}_o(y^+; x, q), \qquad \textstyle\sum_o \pi_o = 1, \tag{3}$$

and the modality fixes what the operators act on. Unlike the pixel-level perturbations of ImageNet-C (Hendrycks & Dietterich, 2019), which perturb the input, these operators edit the *response* and change a fact while leaving the input untouched. For text responses $\mathcal{T}_{\text{near}}$ substitutes k editable spans of y^+ (colour, number, size or spatial words), $\mathcal{T}_{\text{splice}}$ replaces its tail with another caption's, $\mathcal{T}_{\text{shuf}}$ permutes its clauses, and $\mathcal{T}_\times$ returns the caption of another image; the recipe uses $\pi_\times$=0.15 and k=1 (Fig. 6 and Appendix D). For example, a caption reading *two brown horses beside a red fence* becomes *two brown horses beside a blue fence*: one colour word moves, the token count is unchanged, and the pair differs only in a fact about the image. For image responses the operator acts on the prompt and a fixed-seed generator G renders both sides, $y^+ = G(q; s)$ and $y^- = G(\mathcal{T}_{\text{near}}(q); s)$. The usual label-free alternative is to have a model write the rejected response, as in constitutional and AI-feedback pipelines (Bai et al., 2022; Lee et al., 2024; Yu et al., 2024b). We do not, because such a negative differs from the human positive in author, length and register as well as in fact, and by Proposition 1 the reward learns every one of those differences; length alone is a known shortcut for reward models (Singhal et al., 2024). In our own test a head trained on human captions against the policy's captions for the same images scores .36 on Multimodal RewardBench, below chance, and picks the more hallucinated caption in every Best-of-N cell (Appendix F). Editing one

fact inside the same response holds author, length and register fixed by construction. These negatives cost nothing to label: the caption operators are string edits with no model in the loop and the 22k-pair bank builds in 10 seconds on one CPU core, while rendering the image arm's 1,212 pairs requires 5 GPU-hours (Appendix K).

2.2 What the Objective Identifies

We read Eq. (2) as ranking noise-contrastive estimation (Gutmann & Hyvärinen, 2010; Ma & Collins, 2018), with the reward a linear energy over frozen features (LeCun et al., 2006) and β the InfoNCE temperature (van den Oord et al., 2018). Two statements follow, on what training ends at and which direction the head picks; Appendix I reads them and Appendix J proves them. Write $d_{ij} = h_i^+ - h_j^-$, so that Eq. (2) is $\frac{1}{B^2}\sum_{ij} \ell(\beta\, w^\top d_{ij})$ with $\ell(z) = \log(1+e^{-z})$. Since ℓ is strictly convex and weight decay adds a quadratic penalty, the regularized objective has one minimizer: optimizer settings change the path to it, the sampler and β change the objective itself, and whether 200 AdamW steps reach it is an empirical question that Section 3.1 answers.

Proposition 1 (the optimum is a density ratio against the sampler)**.** *Let P and Q be the joint laws of a positive and of a negative triple (x, q, y), the latter with the sampler of Eq. (3) averaged over the positive it edits, $\bar{q}(y \mid x, q)$, and assume P and Q are mutually absolutely continuous. Over unrestricted scores, the cross-entry terms of Eq. (2) are minimized exactly by*

$$s^\star(x, q, y) \;=\; \frac{1}{\beta}\log\frac{\mathrm{d}P}{\mathrm{d}Q}(x, q, y) \;+\; C \;=\; \frac{1}{\beta}\log\frac{p(y \mid x, q)}{\bar{q}(y \mid x, q)} \;+\; C, \tag{4}$$

with C a constant. The B same-entry terms, where y^- is an edit of the very y^+ it is scored against, are coupled and are not minimized by Eq. (4).

Those coupled terms are 12.5% of the loss at $B{=}8$. The linear head fits the regularized minimizer of the same risk inside $\mathrm{span}\{h\}$, a weighted approximation to that log ratio (Eq. (5)). A nuisance the negatives match in distribution drops out of the ratio (Fan et al., 2026); two thirds of our pairs match length exactly and the rest (splice, shuffle, cross-image) leave the negative about 5% shorter on average, so the target is nearly but not exactly length-free (Appendix I, Appendix D).

Why perturbations provide useful supervision. Independent negatives from the caption marginal, our $\mathcal{T}_\times$, make $s^\star$ the image–caption pointwise mutual information up to a constant (Appendix I.2): a score of topic association, which can rank an accurate and an inaccurate response on the same topic equally. A same-context edit changes one fact and nothing else. In the stylized model of Appendix I.3 the quality factor cancels from the PMI score, while the edited pairs recover the correct ordering with margin $\log\frac{\eta}{1-\eta}$ even when a fraction $1-\eta$ of the originals is wrong. Perturbations thus supply one quality comparison, an original above its degraded copy, without requesting a judgment.

Corollary 1 (the sampler's pair law sets the direction)**.** *With the penalty $\frac{\lambda}{2}w^\top D w$, the population minimizer of Eq. (2) satisfies*

$$D\, w^\star \;=\; \frac{\beta}{\lambda}\,\mathbb{E}\big[\,\sigma(-\beta\, w^{\star\top} d)\; d\,\big], \qquad d = h^+ - h^-, \tag{5}$$

the expectation over the pair law of the matrix loss: the head is the mean difference vector of the sampler's pairs, each weighted by the probability the head still misorders it, so two samplers with the same perturbations but different pair laws give different heads in general.

The law of the differences, not their span, sets the direction; this is why the two CLIP-split heads of Section 3.3 can tie on a benchmark's sign pattern while ranking hallucinations 6.4 points apart. Nothing in the objective sets the head's length coefficient to zero, and Section 3.2 finds it with opposite signs at 2B and 8B under one sampler. Finally, Eq. (4) orders a policy's own samples by a ratio against the perturbation process, so it ranks them by how unlike a perturbation each one looks and has no term for how wrong it is. The design hypothesis this leaves is that a sampler drawn from the policy's own errors ranks them better: Section 3.3 measures how far the perturbation sampler falls short and Section 3.4 tests the hypothesis.

2.3 Experimental Setup

We build pairs from 70k PixMo web captions (Deitke et al., 2024) and train heads on Qwen3-VL Instruct backbones at 2B, 4B, 8B and 32B, with the same negative bank and a frozen vision tower at every size. The headline recipe is the 2B backbone at probe layer $N-4$, head only, 200 steps of 256 pairs with AdamW at learning rate 10^{-3} and weight decay 0.01; every later run changes one element of it, and the curated comparison set is RLAIF-V (Yu et al., 2024b). Every head trains on the same schedule of 256 pairs per step, so two runs are compared at equal tokens seen along the training timeline (Fig. 3, right).

We evaluate on Multimodal RewardBench (Yasunaga et al., 2025), VL-RewardBench (Li et al., 2024) and MMRB2 (Hu et al., 2025), with GPT-4o (OpenAI, 2024) as the anchor on all three and no re-tuning between them. Per-benchmark pair counts, the MMRB2 scoring construction and the confidence bands are in Appendix C.

For Best-of-N we sample $N{=}16$ candidates from a Qwen3-VL-2B or 8B policy over 2,500 prompts of COCO captioning (Lin et al., 2014), MathVista (Lu et al., 2024), MMMU (Yue et al., 2024) and Claude-written hard questions on COCO images, and measure CHAIRs (Rohrbach et al., 2018) or exact match against six training-free controls, random through the oracle, with 95% confidence intervals from a paired bootstrap over prompts.

3 Experiments

3.1 Two Choices Set the Accuracy: the Probe Layer and the Sampler

Two choices set the accuracy and neither is on the training side: the probe layer alone adds 5.8 trainable points over the last layer, and mined negatives beat curated RLAIF-V pairs by 3.5 at equal pair count. Table 2 puts those two and every other ablation of the paper on one backbone under one recipe. Every number is a real end-to-end run except the layer sweep that nominated the probe depth: linear heads trained briefly on precomputed hidden states of every layer, a relative instrument that appears only in Fig. 3 and Appendix D.

Probe depth. The best probe layer is never the top of the network: it is $N-4$ for 2B, $N-6$ for 4B and $N-12$ for both 8B and 32B. That depth is chosen on Multimodal RewardBench trainable accuracy, so we treat that benchmark as development throughout and VL-RewardBench and MMRB2, which no choice in the recipe has seen, as the transfer evidence. The gap over the last layer is 3 to 7 points in the layer sweep (Fig. 3) and 5.8 trainable points on the real 4B run. The optimum deepens from 2B to 8B and then holds at twelve layers below the top for every backbone of 36 layers or more, GLM-4.1V-9B included, a depth set by the backbone and not by its layer count; each peak is a plateau several layers wide (Appendix D).

Negative construction. We find that same-image perturbations transfer in both modalities and that cross-image negatives do not (Table 3). For text, $\mathcal{T}_\times$ alone reaches 99% training accuracy and falls below the untrained base on the benchmark, because it amplifies a length and style shortcut the same-image operators mostly remove: that head scores .224 on the length-sensitive safety/bias column against .451. For images the replica is exact, with the deranged negative transferring nothing to MMRB2 text-to-image (52.7 against a label prior of 53.2) while the same-seed planted edit accounts for all of it (56.0). Perturbation density and the operator mix are in Appendix D.

Every training-side change leaves the benchmark within .014 of the recipe. Every training-side change we tried leaves the benchmark where it was in real runs, beyond what convexity predicts: loss form, negative composition, four times the pairs, twice the steps and a curriculum that orders the negatives from easy to hard, or from hard to easy, each move a real 2B run by at most .014, and in the sweep setting the negative pool and a temperature schedule move it by less than .005. The sharpest row is a 540-configuration grid over head learning rate, β and weight decay: all 540 reach .995 or better in distribution, and selecting among them out of distribution changes the benchmark by .0003.

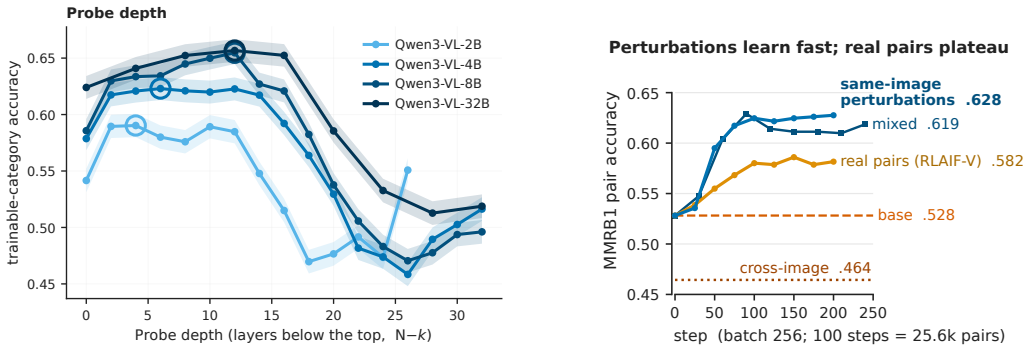

Figure 3: **Left:** layer-sweep trainable accuracy against layers below the top for 2B / 4B / 8B / 32B (bands: 75% binomial intervals); the optimum is never the last layer, and sweep numbers rank depths within a model only. **Right:** Multimodal RewardBench pair accuracy over the first 250 steps by negative type (2B full fine-tune, 674-pair eval subset): same-image perturbations transfer, cross-image negatives fall below the untrained base, curated pairs plateau lower.

Mined negatives against curated pairs at a matched pair budget. At 15k pairs the mined-negative head is 3 to 4 trainable points ahead of curated RLAIF-V pairs on Multimodal RewardBench and 2 to 5 on VL-RewardBench, including the hallucination split RLAIF-V was built for (Table 5). Unfreezing the backbone on the same 15k curated pairs recovers 11 points overall to .613: a frozen head matches full fine-tuning on mined-negative data where real pairs need the unfrozen model (Appendix D).

3.2 One Untuned Recipe Reaches GPT-4o Across Backbones and Benchmarks

The 32B probe is at or above GPT-4o on all three benchmarks and separably above it on one (Table 1). On VL-RewardBench it is the strongest entry on the table, 6.6 points above GPT-4o at 3.6 standard errors of the difference; on Multimodal RewardBench and on MMRB2 reasoning the benchmark cannot resolve the margin, and separating either at two standard errors would take 13 and 3.5 times the released pairs. The stronger reference than GPT-4o is the backbone itself as a generative judge, which on MMRB2, the one benchmark we ran it on, leads the probe by 4.7 points (64.6 against 59.9 on the four-task average, Appendix C); the probe gives up those points for a scorer that decodes nothing, one forward pass per response against a generated verdict in both presentation orders. Larger backbones improve the intermediate-layer probe on all three benchmarks, while the base log-likelihood judge stays near chance and the last-layer head trails at every size (Fig. 7). The recipe is not specific to Qwen3-VL: on GLM-4.1V-9B, a second family with 40 layers, the same recipe at $N-12$ beats its last layer and its untrained judge on both benchmarks (Table 1). A head trained only on caption perturbations also transfers a usable signal to response-image tokens once the backbone is large enough (MMRB2 editing, level with GPT-4o), which starts Section 3.4. Our training data overlaps none of the three benchmarks (Appendix B).

What the columns mean. The safety/bias column of Multimodal RewardBench is a length artifact, the one reward models are known to carry (Singhal et al., 2024). A head trained on same-image perturbations, two thirds of which match length exactly, scores .998 on it at 2B and .19 at 8B, where its correlation with response length flips sign. The gap is not seed noise: eight layer-sweep seeds give $.967 \pm .007$ at 2B and $.569 \pm .013$ at 8B, the same ordering (Appendix C). We therefore compare heads on the five *trainable* categories, where the 8B probe's .659 overall against .708 at 2B becomes .658 against .602.

3.3 Two Probes, One Score, Two Rewards

Same benchmark score, different reward. We train two probes on halves of the same mined-negative data, ranked by the drop in CLIP (Radford et al., 2021) image–text cosine from the real to the perturbed caption and split at the extreme quartiles, 15k pairs each at 2B. They score identically on Multimodal RewardBench, .597 against .603 trainable, and then separate three ways (Fig. 1b): VL-RewardBench hallucination 72.0 against 78.4, reasoning 63.4 against 56.8, and as Best-of-N scorers the CLIP-visible head beats probe-8B on CHAIRs by 3.8 points while losing 2.9 on QA.

Table 1: **One recipe on three reward benchmarks.** Each probe is one frozen linear head at its probe depth, trained once on caption perturbations and evaluated with no re-tuning. Published judges are quoted from each benchmark's paper; our Multimodal RewardBench rows (†) are over the 4,711 released pairs, which omit the withheld toxicity split (GPT-4o recomputed on them: .728 overall, Gemini 1.5 Pro .734), and safety/bias is a length artifact for every mined-negative head (Section 3.2). VL-RewardBench macro is the mean of its three splits. The 2B and 8B rows use the 60k-pair bank, the 4B and 32B rows the 15k-pair rebuild (Appendix D). Bold: column maximum among our Qwen3-VL rows. The GLM-4.1V-9B rows are a second backbone family under the identical recipe (15k pairs, 200 steps, $N-12$ of 40 layers, no re-tuning).

Model	**Overall**	**General**		**Knowl.**	**Reasoning**		**Safety**	**VQA**
		Corr.	Pref.		Math	Coding	Bias	
Claude 3.5 Sonnet	.720	.626	.678	.739	.686	.651	.768	.856
Gemini 1.5 Pro	.720	.635	.677	.663	.689	.555	.945	.872
GPT-4o	.715	.626	.690	.720	.676	.621	.748	.872
Ours: 2B probe, $N-4$	.708†	.669	.610	.587	.654	.490	**.998**	.850
Ours: 4B probe, $N-6$	.724†	.713	.621	.624	.716	.500	.951	.855
Ours: 8B probe, $N-12$	.659†	**.735**	**.638**	.635	.708	**.579**	.187	.860
Ours: 32B probe, $N-12$	**.733**†	.726	.636	**.643**	**.739**	.567	.742	**.911**
Ours: GLM-4.1V-9B probe, $N-12$	.692†	.636	.602	.514	.562	.612	.935	.857
base 4B (log-likelihood judge)	.571†	.449	.558	.581	.582	.364	.191	.892
base GLM-4.1V-9B (log-likelihood judge)	.565†	.440	.495	.573	.611	.404	.167	.892

VL-RewardBench	**Overall**	**Macro**	**General**	**Halluc.**	**Reason.**
Gemini-1.5-Pro (strongest published)	.672	–	–	–	–
GPT-4o (published)	.658	–	–	–	–
Ours: 32B probe, $N-12$	**.724**	**.649**	**.475**	**.810**	**.663**
Ours: 8B probe, $N-12$	.682	.607	.425	.765	.631
Ours: 4B probe, $N-6$	.680	.600	.420	.774	.606
Ours: 2B probe, $N-4$	.669	.592	.425	.762	.587
Ours: GLM-4.1V-9B probe, $N-12$	.602	.531	.348	.678	.568
base 8B (log-likelihood judge)	.542	.525	.459	.554	.561
base 4B (log-likelihood judge)	.525	.518	.464	.522	.568
base 2B (log-likelihood judge)	.569	.550	.525	.598	.527
base GLM-4.1V-9B (log-likelihood judge)	.570	.529	.425	.616	.546

Across sizes we observe "a small benchmark gap against a large hallucination gap". This is Proposition 1 made visible: the two samplers differ from the positives along different coordinates and the heads estimate different density ratios, and the benchmark scores neither coordinate.

Whether a scorer helps Best-of-N depends on the pool, not on its benchmark score. Across eight captioning and QA cells and a hard-QA ladder, a probe's benchmark accuracy does not predict how much it improves Best-of-N selection, in either direction (Table 6 and Appendix F). On captioning every mined-negative probe stays at greedy level where the oracle reaches .06 to .09 CHAIRs, and on MathVista and MMMU, where candidate agreement tracks correctness, majority vote is 1.4 to 3.8 points ahead of the probe, whose one positive role there is as a vote weighter. Where the 8B generator is confidently wrong instead (majority wrong on 32% of prompts), every external scorer beats majority vote: probe-2B by 10.8 points, probe-8B by 7.2 and the training-free log-likelihood judge by 9.2, all with 95% CIs excluding zero. The trained head is not what earns it: probe-2B against log-likelihood is $+1.6$ to $+2.1$ points with CIs including zero, and probe-2B beats probe-8B by 3.6 points, the reverse of their benchmark order. What decides the gain is the pool: whether the generator's majority can be trusted, and so whether any external scorer should replace the vote. The benchmark measures neither.

3.4 Learning from the Policy's Own Errors: Better Deployment, Lower Benchmarks

Figure 2b changes one element of the recipe of Section 2.1, the negative sampler: the policy's own Best-of-N candidates, labelled by a cheap verifier, are paired correct against wrong for the same

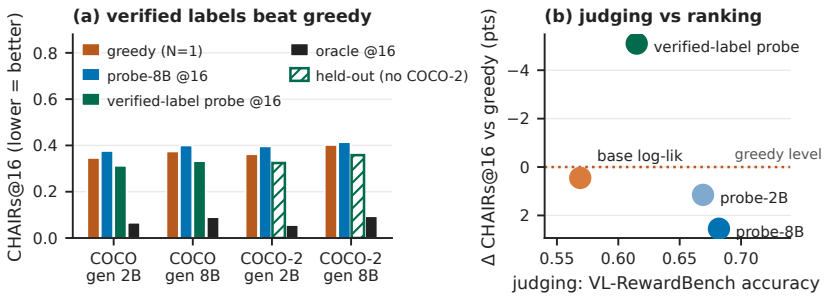

Figure 4: One turn of the loop. **(a)** CHAIRs@16 per captioning cell (hatched: verified probe trained without COCO-2). **(b)** Judging (VL-RewardBench) against ranking (ΔCHAIRs vs greedy).

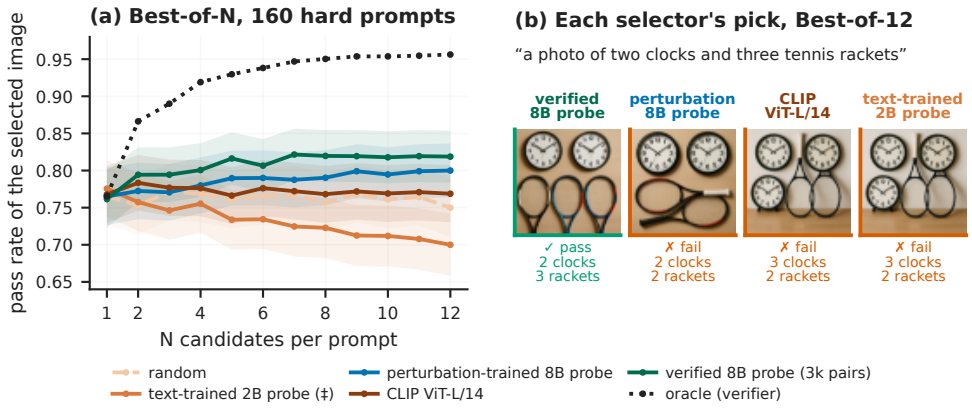

Figure 5: **The loop transposes to image generation: verifier-minted pairs gain 5.4 points over random selection at 8B, where caption perturbations alone lose 6.5.** **(a)** Best-of-N on 160 held-out hard compositional prompts: verifier pass rate of the image each selector picks from N Qwen-Image samples (bands: 75% binomial intervals; paired differences against random at N=12 in the text). **(b)** One prompt and each selector's pick out of twelve, judged by the GenEval-style OWLv2 verifier (Ghosh et al., 2023; Minderer et al., 2023) on object presence, count, colour and relative position; the selectors see only the image.

prompt. By Proposition 1 the head then learns the density ratio against the policy's own errors, which is the ratio Best-of-N selection needs.

It is the only change in the study that moved captioning: a head trained on 3,462 such pairs cuts CHAIRs by 5.1 points against greedy and by 4.1 points on a task it never trained on (Fig. 4a, Table 7). On QA it matches probe-8B as a selector, so the utility claim rests on captioning, the open-ended task where no vote is available. CHAIR is both harvest label and metric, so the gain is in absent COCO-object mentions rather than faithfulness at large (Appendix A).

Figure 4b puts the two results side by side. The verified head is the weakest judge of the three heads on VL-RewardBench, .615 against .669 and .682 for the perturbation probes, yet the only one that ranks the policy's captions below greedy's hallucination rate (the other benchmarks in Table 7). Both follow from the sampler: 70% of the verified pairs are QA correctness pairs, so the categories they do not cover fall while math rises. The two rewards differ because the two comparisons do. Benchmark pairs contrast other models' responses on annotator-chosen errors, close to what perturbations produce; verified pairs contrast this policy's own successes and failures; and by Proposition 1 a linear head learns only the one ratio it is shown, so it trades the benchmark's for the deployment's. Reconciling them is open: train on the union of the two samplers, give the head more than one direction, or draw benchmark pairs from the policy under deployment.

The actor side of the turn. One DPO turn on pairs the policy generated itself lowers its COCO hallucination rate from .446 to .390 CHAIRs with probe-picked pairs, .414 with verified and .412 with random pairs, at no cost on MathVista or MMMU (Fig. 1c, Table 8). The label adds little over random pairs, so the selector's effect on the actor is still to be shown (Appendix G).

The same turn on image generation. The loop transposes to image generation point for point (Fig. 5): Qwen-Image (Qwen Team, 2025b) produces pools of 12 on GenEval-style prompts (Ghosh et al., 2023) and a label-free OWLv2 verifier (Minderer et al., 2023) plays CHAIR's role. On 160 held-out hard prompts (random .765, oracle .956) the text-trained 2B head is 6.5 points below random (95% CI -11.6 to -1.6), the mined-negative image heads are $+2.5$, on par with CLIP ($+3.0$), and a head trained on 3,000 verifier-minted pairs is $+5.4$ at 8B (95% CI $+0.5$ to $+10.3$), 2.4 points above CLIP and 2.9 above the mined-negative image heads on the same prompts. The two verified heads cross over as MMRB2 judges, so here too the better ranker is the worse judge.

4 Related Work

Multimodal reward models and the benchmarks that rank them. A multimodal reward model is built by annotating preference pairs and fitting a scorer on them, from LLaVA-RLHF and RLHF-V (Sun et al., 2023; Yu et al., 2024a) to MM-RLHF's 120k human comparisons (Zhang et al., 2025) and the scoring heads of IXC-2.5-Reward and UnifiedReward (Zang et al., 2025; Wang et al., 2025). Removing the annotator usually means a model in their place (Bai et al., 2022; Lee et al., 2024; Yu et al., 2024b); we use a sampler. Multimodal RewardBench (Yasunaga et al., 2025), VL-RewardBench (Li et al., 2024) and MMRB2 (Hu et al., 2025) rank those models and generative judges (Zheng et al., 2023) by pairwise accuracy against human labels (cf. Lambert et al., 2024). We take them as given and ask what the number measures.

Probing and probe depth. The scorer we set against those judges is a probe, and linear probes are classically an analysis tool for encoders (Alain & Bengio, 2017; Tenney et al., 2019; Burns et al., 2023). LLaVA-Reward (Zhou et al., 2025) is the closest prior use of a probe as a reward, mixing depths through a cross-attention module trained on annotated preferences; ours is a 2,048-parameter head with no label.

Density-ratio estimation and negative sampling. A probe on a frozen backbone approximates a density ratio, which noise-contrastive estimation fits against a noise distribution (Gutmann & Hyvärinen, 2010; van den Oord et al., 2018) that the negative-sampler literature designs to shape the representation (He et al., 2020; Robinson et al., 2021; Kalantidis et al., 2020; Wang & Isola, 2020). With frozen features those choices act only through the objective. Closest to our objective is Fan et al. (2026), a label-free in-batch Bradley–Terry reward on text with web continuations as the negative process; their invariance theorem gives a nuisance's coefficient in the unrestricted optimum, which Proposition 1 restates for our matrix loss.

Best-of-N and learning from self-generated data. A reward built from that ratio is then put under optimization pressure, and proxy rewards degrade under it (Cobbe et al., 2021; Gao et al., 2023; Stiennon et al., 2020). We locate an earlier failure, before any pressure is applied, where the proxy's benchmark score does not predict its Best-of-N value against self-consistency (Wang et al., 2023). Expert iteration and STaR (Anthony et al., 2017; Zelikman et al., 2022) improve a policy on its own verified outputs, and GKD (Agarwal et al., 2024) diagnoses the train–inference mismatch on the policy side; ours is the reward-side analogue.

5 Concluding Remarks

Frozen vision-language representations support strong reward modeling with only a linear head and caption perturbations, but probe depth and training comparisons determine which signals the head extracts. Similar benchmark scores can conceal different specializations, and higher preference accuracy need not yield better Best-of-N selection. Retraining on verified policy outputs improves caption selection despite lowering benchmark accuracy, highlighting the importance of matching reward supervision to deployment errors. These findings motivate evaluating reward models through both preference agreement and downstream selection. Future work should explore how to combine broad preference learning with adaptation to policy-specific errors, and whether these rewards can improve both output selection and subsequent policy training.

AI USE STATEMENT

In this work we used generative AI tools to draft and revise prose, to write the experiment, analysis and plotting code, and as a component of two experimental artifacts that the paper describes as such: the 500 hard COCO questions and the 1,992 truth-preserving control edits of Appendix F were written by an LLM and verified in isolation, with 84 to 97% of decoys confirmed by edit type, and the same family served as generator and verifier there, which is stated wherever those numbers are used. We have not used generative AI tools to generate the research ideas or to choose which claims the paper makes. We have reviewed all AI-assisted work: every number reported here comes from a run whose artifact an author read, the contamination audit of Section 3.2 was written and checked against the released benchmark files, and each experiment's reading and its control were fixed before the run and are stated in the paragraph that reports it. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public image-caption corpora, public preference benchmarks and public model checkpoints, and involves no human subjects and no new data collection. One finding carries a foreseeable risk and we state it plainly: a reward model that is cheap to build and strong on benchmarks while blind to its own learned direction invites deployment without the controls that would catch the blindness. Section 3.3 shows two heads with the same benchmark score ranking hallucinations 6.4 points apart, and the safety/bias column of Section 3.2 is decided by response length rather than by anything the head was trained on, so a benchmark score is not evidence that such a model is safe to deploy. The controls the paper uses, a deployment-side Best-of-N cell and a sampler-split comparison, are the ones we would ask of any label-free reward model, our own included.

REPRODUCIBILITY STATEMENT

The reward model is specified completely by Eqs. (1) to (3), the recipe and hyper-parameters of Section 2.3, and the operator definitions and optimizer settings of Appendix D. Every input is public: positives are PixMo-Cap captions, negatives are string edits of them that rebuild in 10 seconds on one CPU core, the curated comparison set is RLAIF-V, and all three evaluation benchmarks are used as released with per-benchmark pair counts and scoring constructions given in Section 2.3. Proofs of Proposition 1 and Corollary 1 are in Appendix J. Appendix B lists what will be released and Appendix K the compute each stage costs; the contamination audit that licenses the benchmark comparisons is described in Section 3.2.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2017.

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *NeurIPS*, 2017.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Jingxuan Fan, Yueying Li, Zhenting Qi, Dinghuai Zhang, Kianté Brantley, Sham M. Kakade, and Hanlin Zhang. Scaling reward modeling without human supervision. *arXiv preprint arXiv:2603.02225*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *ICML*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. *arXiv preprint arXiv:2512.16899*, 2025.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, 2020.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning (ICML)*, 2024.

Lei Li, Yuancheng Wei, Zhihui Xie, et al. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

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 (ECCV)*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *EMNLP*, 2018.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. Qwen3-VL-2B/8B-Instruct vision-language models.

Qwen Team. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *NeurIPS*, 2020.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal reward-bench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*, 2025.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Tianyu Yu, Haoye Zhang, Yuan Yao, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, Kai Chen, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model. In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*, 2022.

Di Zhang. What Is RLCD? The Secret Behind Jev. Blog post, https://di-zhang-llm.github.io/blog/what-is-rlcd-the-secret-behind-jev, 2026. Posted September 21, 2026; accessed September 25, 2026.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.

Yanli Zhao, Andrew Gu, Rohan Varma, et al. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Shijie Zhou, Ruiyi Zhang, Huaisheng Zhu, Branislav Kveton, Yufan Zhou, Jiuxiang Gu, Jian Chen, and Changyou Chen. Multimodal llms as customized reward models for text-to-image generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Zhenxun Zhuang, Mingrui Liu, Ashok Cutkosky, and Francesco Orabona. Understanding AdamW through proximal methods and scale-freeness. *Transactions on Machine Learning Research*, 2022.

Appendix Contents

A Limitations and Broader Impacts **15**

B Reproducibility **15**

C Evaluation Protocol **15**

D Negative Mining and Probe Depth **16**

E Extended Results for RM Training **18**

F Extended Results for Best-of-N **19**

G Actor Training **21**

H CLIP Studies **22**

I Reading the Two Statements **22**

I.1 What Constructed Comparisons Supervise 23

I.2 Independent Negatives: Density Ratios and PMI 23

I.3 When Perturbations Expose an Additional Quality Signal 23

J Proofs **24**

K Compute Resources **26**

A Limitations and Broader Impacts

Our approach lowers the additional training and preference-annotation requirements for multimodal reward modeling by learning a small scoring head over a frozen backbone. This may make task-specific reward models more accessible to researchers and practitioners with limited annotation budgets, while retaining the inference cost of the underlying backbone. Constructing training comparisons explicitly also makes the intended supervision easier to inspect and revise.

The broader implication concerns how reward models are evaluated and selected. Our experiments show that higher preference-benchmark accuracy need not yield better selection of a policy's outputs. Reporting the training comparison distribution alongside deployment-specific evaluations can therefore support more informed model selection. Conversely, using benchmark performance alone to justify deployment risks overlooking systematic errors and shortcuts, including response-length biases.

These benefits have a limited demonstrated scope. Our experiments cover two backbone families, and verifier-based adaptation is evaluated for one round. Captioning improvements are measured with CHAIR, which also supplies the training labels; they establish fewer absent COCO-object mentions rather than general factuality or preserved informativeness, and a CHAIR-independent metric such as POPE (Li et al., 2023) would be needed to settle the latter. Best-of-N cells are single-seed, the hard-QA set was written and verified by one LLM family, and our Multimodal RewardBench overall omits the withheld toxicity split (Section 3.2). The method does not establish safety or fairness capabilities, and broader deployment requires independent evaluation of output quality, coverage, and task-specific failure modes.

B Reproducibility

The recipe is fully specified by Eqs. (1) to (3) and the hyper-parameters of Section 2.3 and Appendix D: a frozen Qwen3-VL checkpoint, a probe layer, mean pooling over response tokens, and one bias-free linear head trained with AdamW for 200 steps. Every input is public. Positives are PixMo-Cap captions (Deitke et al., 2024); negatives are string edits of those captions with no model in the loop, so the 22k-pair bank rebuilds in 10 seconds on one CPU core from the operator definitions in Appendix D. The curated comparison set is public RLAIF-V (Yu et al., 2024b), and all three evaluation benchmarks (Yasunaga et al., 2025; Li et al., 2024; Hu et al., 2025) are used as released, with the pair counts and the scoring construction stated per benchmark in Section 2.3. Contamination is checked by a 256-bit difference hash of every training image against every evaluation image, counting a pair as a near-duplicate at ≤ 10 differing bits, plus exact prompt matching. Against Multimodal RewardBench the perturbation set shares 5 near-duplicate images of 4,711 (0.1%) and no prompt; against VL-RewardBench, the 1,000 MMRB2 prompt images and the 1,000 COCO Best-of-N images it shares none. The image arm renders 1,212 PixMo-derived prompts with Qwen-Image, so its training images are generated and its prompts are not those of MMRB2. The same audit exposes the curated baseline: RLAIF-V pairs overlap VL-RewardBench on 326 of 1,247 rows and Multimodal RewardBench on 18, which is why every RLAIF-V-trained row also reports the decontaminated subset. Source for pair construction, training, every selector in the Best-of-N tables and the plotting scripts behind every figure will be released with the paper.

C Evaluation Protocol

Multimodal RewardBench releases 4,711 of its 5,211 pairs, the toxicity split being withheld, in eight categories; our overall number is therefore comparable to published ones only through the recomputation of Table 1. VL-RewardBench has 1,247 pairs (general 181, hallucination 749, reasoning 317), of which 326 overlap the RLAIF-V training set, so anything trained on RLAIF-V is also read on the 921 clean rows; on the 326 overlapping rows the RLAIF-V fine-tunes score .80 against .60 to .64 for models that never saw the data, and on the clean rows their lead disappears. MMRB2 has 1,000 judge-labelled pairs per task (text-to-image, editing, interleaved generation, reasoning; 996 reasoning pairs are scoreable). We score it with the prompt in the user turn and the response, images included, in the assistant turn, the construction the heads were trained on; a scalar reward is order-invariant, so its accuracy equals MMRB2's dual-order accuracy, exact ties count 0.5, and

reasoning is also reported on the 350-pair text-only-output subset, the convention that benchmark's paper uses for scalar reward models. On the four-task average under this construction the 32B probe reads 59.9 and the same backbone used as a generative judge 64.6; GPT-4o as a judge reads 59.7. Scoring prompt and response in one user turn instead costs the text heads 2 to 3 points on reasoning and 5 to 9 on its text-only subset. The GPT-4o margins in Table 1 are quoted as unpaired standard errors because the benchmark papers do not release per-item predictions, so a paired interval is not available. Shaded bands in every figure are 75% binomial intervals of the plotted accuracy over its evaluation set; seed-to-seed spread of the heads is narrower than the line width. Eight seeds of the layer-sweep heads put the safety/bias column of Section 3.2 at $.967 \pm .007$ at 2B and $.569 \pm .013$ at 8B, where the correlation between head score and response length is -0.57 at 2B and $+0.44$ at 8B. The Best-of-N intervals of Section 3.3 are over prompts at one selector seed.

The Best-of-N prompts are 500 each of COCO captioning (two prompt sets, COCO and COCO-2), MathVista and MMMU, plus 500 Claude-written hard questions on COCO images, of which 241 are counting or spatial; candidates are sampled at temperature 1.0 and top-p 0.95, plus the greedy decode. The mid-training runs of Section 3.4 continue from the mined-negative probe on 3,462 verified pairs harvested from those pools (MathVista 1,352, MMMU 1,072, COCO 511, COCO-2 527), with leave-one-task-out runs that drop every COCO-2 or every MMMU pair, and a non-convex check that unfreezes the backbone on 15k RLAIF-V pairs from either checkpoint. Heads train with AdamW, learning rate 10^{-3}, weight decay 0.01, β=0.1, micro-batch 8 and effective batch 256 or 1,024 as stated; the backbones have 28, 36, 36 and 64 layers.

D Negative Mining and Probe Depth

Two pair banks appear in the benchmark tables: the 2B and 8B rows of Table 1 use the original 60k-pair bank, the 4B and 32B rows and every ablation the rebuilt 15k-pair bank, which costs the 2B probe 2 to 3 points (.708 to .682 on Multimodal RewardBench), so the 32B row leads from the smaller bank. Perturbation density does not move either benchmark: one, three or all editable spans stay within .008 on both benchmarks (Table 4) while the training task gets much easier. The three token-edit heads beat the mixed-operator recipe on VL-RewardBench by 2.8 to 3.4 points at equal Multimodal RewardBench accuracy (Table 3); that comparison also drops the cross-image and the length-changing operators, so it does not isolate the mix.

Table 4: **Perturbation density,** k=1 / 3 / all editable spans per rejected caption (all = 13.6 edits on average), 21,521 identical images and captions per setting, token substitutions only. Frozen 2B $N-4$ head, 200 steps of 1,024, single seed. The training task gets much easier (val .82 $\rightarrow$ .996) and neither benchmark moves.

setting	val	MMRB1	VL-RewardBench			
		overall	overall	macro	general	halluc.
k=1 (1.0 edit)	.820	.673	.709	.665	.630	.781
k=3 (3.0 edits)	.973	.673	**.713**	.660	.613	.797
k=all (13.6 edits)	.996	**.681**	.707	.644	.547	.797
ref: mixed-operator probe	–	.682	.679	–	–	.769

Figure 3 gives the depth sweep per model, Table 3 the operator ablation in both modalities, Table 4 the density variants and Table 5 the matched-budget comparison against RLAIF-V. On that matched budget the mined-negative head reads .574 to .584 trainable on Multimodal RewardBench against .544 for RLAIF-V and .679 to .709 on VL-RewardBench against .660, with hallucination .769 to .781 against .737; its in-distribution validation accuracy is .82 to .92 against .55 for the curated pairs. The comparison is matched in pairs and not in tokens, since perturbation captions run 2.4 times longer, which leaves the RLAIF-V head 2.6M response tokens against 6.1M. Unfreezing the backbone on the same 15k real pairs gains 11 points overall and 7 trainable, to .613, which puts it above the mined-negative probe on the trainable categories and level with it on decontaminated VL-RewardBench (.727 against .747). Perturbation operators act on one caption at a time: a near-miss substitution of a colour, number or spatial word; a splice of another caption's tail; a clause shuffle; and, for 15% of pairs, the caption of another image. In the 15k-pair bank the operators and their share are near-miss substitution 50%, attribute substitution 15%, splice 17%, shuffle 10% and cross-image 8%. Measured as the ratio of rejected to chosen length in words, the two substitutions leave length

Table 2: **Every ablation and analysis of the paper on one backbone (Qwen3-VL-4B).** Frozen linear head, mean pooling, probe on layer $N-6$ unless stated, 200 steps of 256 pairs, single seed, no re-tuning between benchmarks. MMRB1 = Multimodal RewardBench (4,711 pairs; *trainable* = the five categories without safety/bias and VQA); VL-RB = VL-RewardBench (1,247 pairs, hallucination split 749). The Δ-CLIP heads are 5k-pair quartiles of the same perturbations split by whether CLIP can see the edit; the verified heads train on 2,400 verifier-minted pairs from a fresh head or from the mined-negative head; the base judge scores by response log-likelihood with no training.

setting	MMRB1 overall	MMRB1 trainable	VL-RB	VL-RB halluc.
mined-negative probe, $N-6$ (recipe)	**.724**	**.633**	.680	**.774**
last-layer control	.696	.575	.659	.700
frozen head on RLAIF-V 15k	.676	.598	.618	.677
Δ-CLIP blind / visible	.687 / .707	.615 / .631	.656 / .673	.721 / .765
verified, fresh / mid-training	.705 / .723	.615 / .632	.616 / **.690**	.650 / .762
base log-likelihood judge	.571	.507	.525	.522

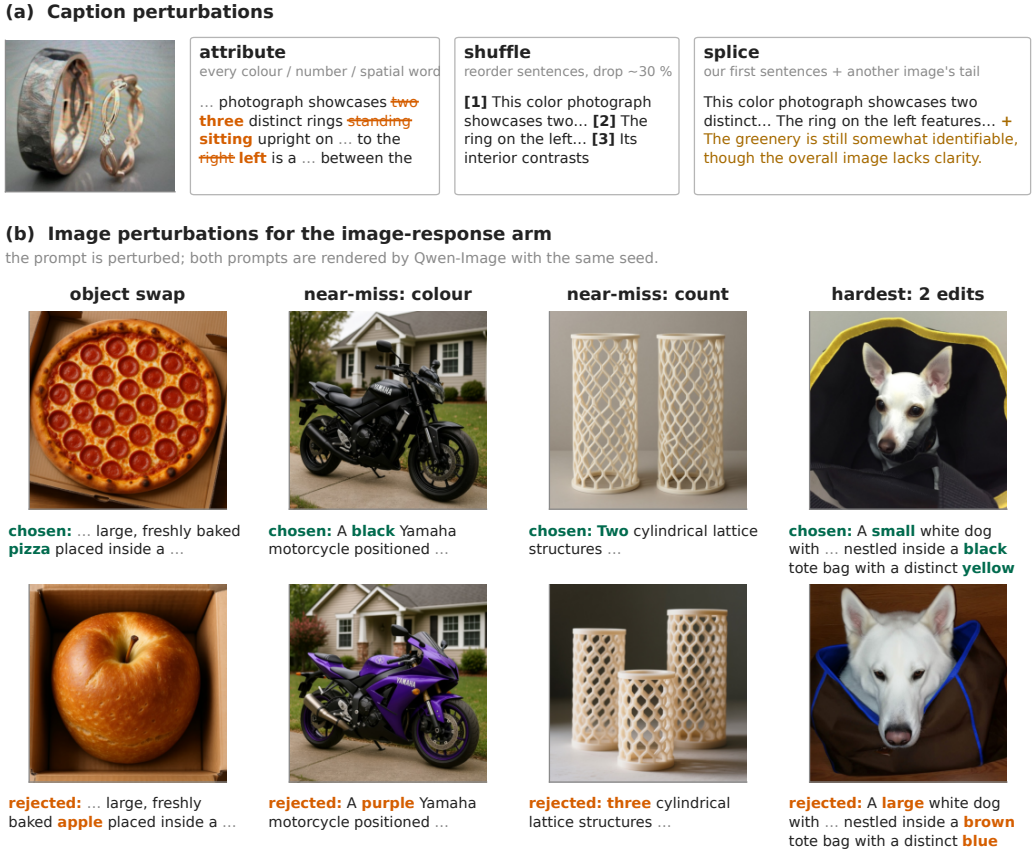

Figure 6: Negative mining by perturbation, laid out after ImageNet-C (Hendrycks & Dietterich, 2019). **(a)** A PixMo image and three operators applied to its human caption: attribute swaps every colour, number and spatial word; shuffle reorders and drops sentences; splice continues the true caption with another image's. Red marks the edit, orange the foreign text. **(b)** The image-response arm perturbs the prompt instead and renders both prompts with Qwen-Image under the same seed, so the two prompts differ only in the edited span.

unchanged (1.00), splice shortens to 0.82, shuffle to 0.69 (it drops about 30% of sentences) and cross-image captions are 1.10 times as long, so across the bank a negative is about 5% shorter than its positive. On GLM-4.1V-9B the same recipe at the last layer reads .628 on Multimodal RewardBench and .598 on VL-RewardBench against .692 and .602 at $N-12$ (Table 1); the untrained judge reads .565 and .570. Density variants perturb one, three or all editable spans (13.6 on average) of the same 21,521 captions. Heads train with AdamW, learning rate 10^{-3}, weight decay 0.01, β=0.1, micro-batch 8, effective batch 256 or 1,024 as stated, 200 steps, on precomputed vision-tower outputs; the 32B confirm run used FSDP (Zhao et al., 2023) over eight GPUs with 1,024 effective. The layer sweep trains a linear head on the precomputed hidden states of each layer for 8k pairs and five seeds;

Table 3: **Negative-mining ablation in both modalities.** Top: text responses, 2B backbone. The frozen-head rows use 15k pairs at $N-4$ on the official benchmarks; the two full-fine-tune rows are the June series that first separated the negative types (overall Multimodal RewardBench; their safety/bias column reads .224 and .451, the length shortcut). Bottom: image responses, 2B image-trained head at $N-12$ on 1,091 same-seed prompt pairs; *held-out* is accuracy on the head's own negative type, MMRB2 text-to-image is judge agreement (label prior 53.2). The *by kind* row splits held-out accuracy of the combined head by the planted edit.

text negatives y^-	**head**	**val**	**MMRB1**	**trainable**	**VL-RB**	**halluc.**
cross-image only, $\mathcal{T}_\times$	full FT	.99	.559	–	–	–
same-image perturbations, mixed	full FT	–	.582	–	–	–
token edit, k=1 span, $\mathcal{T}_{\text{near}}$	frozen head	.820	.673	.574	.709	.781
token edit, k=3 spans	frozen head	.973	.673	.563	**.713**	**.797**
token edit, all spans (13.6)	frozen head	.996	.681	.576	.707	**.797**
mixed operators, 15% $\mathcal{T}_\times$ (recipe)	frozen head	.92	**.682**	**.584**	.679	.769

image negatives y^-		**held-out, own type**	**MMRB2 T2I**
cross-image only, $G(q'; s')$		.93–1.00	52.7
planted edit only, $G(\mathcal{T}(q); s)$		.79–.85	56.0
both (recipe)		.79–.84	**56.2**
by kind: object swap / near-miss / multi-edit	2B	.944 / .714 / .837	
	8B	.917 / .786 / .884	

Table 5: **Mined negatives against curated RLAIF-V pairs at a matched pair budget,** 2B backbone, 15k pairs per setting. Frozen heads: $N-4$, 200 steps of 1,024 pairs. Full fine-tune: lr 10^{-5}, 400 steps of 256, vision tower frozen. *tokens* is the response-token count of the 15k pairs (both sides; Qwen3-VL tokenizer): perturbation captions are longer than RLAIF-V responses (204 against 86 tokens per side), so the RLAIF-V heads see fewer tokens at the same pair count. *val* is in-distribution pair accuracy on each setting's own held-out pairs. 326 of the 1,247 VL-RewardBench rows overlap the RLAIF-V training set, so RLAIF-V models are also read on the 921 clean rows; mined-negative pairs overlap none. Single seed; the comparison is matched in pairs and not in tokens.

training pairs (15k)	**head**	**tokens**	**val**	**MMRB1**	**trainable**	**VL-RB**	**VL-RB clean**
perturbations, token-edit	frozen head	6.4M	.820	.673	.574	.709	**.747**
perturbations, mixed operators	frozen head	6.1M	.92	**.682**	**.584**	.679	–
RLAIF-V	frozen head	2.6M	.551	.577	.544	.660	.672
RLAIF-V	full FT	2.6M	.637	**.686**	**.613**	**.735**	.727
base 2B (log-likelihood judge)	–	–	–	.563	–	.569	–

it ranks depths within a model and nothing else, and it misleads in two documented ways. Last-token pooling ranks first in the sweep at 4B, 8B and 32B and loses 7.5 points to mean pooling in the one real run. Sweep heads trained under a different regime than the real pipeline disagree with it by several points, in either direction.

E Extended Results for RM Training

Figure 7 separates the components behind Table 1 at each backbone size.

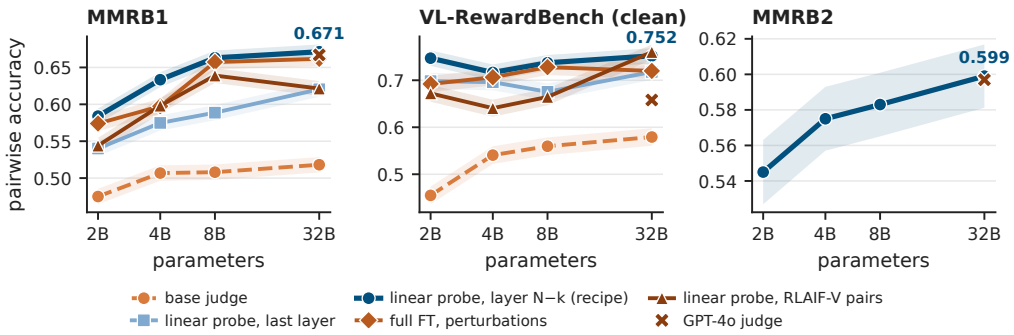

Figure 7: Three reward benchmarks against backbone size, by component (shaded: confidence bands). The linear probe at $N-k$ improves with size on all three and beats the last-layer probe at every size; the full fine-tune matches it on MMRB1 only from 8B up and trails it on VL-RewardBench at 32B (early-stopped at step 150 of 400); the RLAIF-V probe trails on MMRB1 everywhere and leads on VL-RewardBench only at 32B, where the two share source images.

F Extended Results for Best-of-N

Figure 8 and Table 6 give the hard-QA ladder with every control on every panel and Table 7 the loop turn per captioning cell. Initialization of the verified head enters only through how far a finite budget travels: at 2B a fresh head and one continued from the mined-negative probe land within 0.3 points everywhere (Table 7), and at 4B, where 2,400 pairs do not reach the optimum, the continued head is 7 points ahead (Table 2). A planted-decoy test (3.4 LLM-written single-error decoys and 4 template perturbations per captioning pool) shows that the mined-negative feature fires inside a real pool: every scorer ranks a source above its own one-edit decoy at least 93% of the time, yet the margin is 0.3 to 1.1 within-pool standard deviations and CHAIRs of the pick does not move, so the probe knows an edited sentence is worse than its source without knowing which of 16 real captions is the most truthful. Truth-preserving edits separate the signals: the probes do not prefer the source over a true edit (probe-2B .512, probe-8B .291) while log-likelihood prefers it 91% of the time.

Model-written negatives. Two 2B heads were trained with the policy's own captions in the pair instead of a string edit, same recipe otherwise. Self-corruption (chosen: the policy's caption; rejected: its perturbation) ties greedy and random in all four captioning cells and reads .689 on Multimodal RewardBench, the same as the off-policy head. Human-versus-policy (chosen: the PixMo caption; rejected: the policy's caption for the same image) reads .360 on Multimodal RewardBench, below chance, and picks a more hallucinated caption than random in all four cells (CHAIRs@16 .422 / .432 / .414 / .422 against random .370 / .384 / .406 / .402): it learned to tell a human author from a model, and on a benchmark where both responses are model-written that direction anti-correlates with the label.

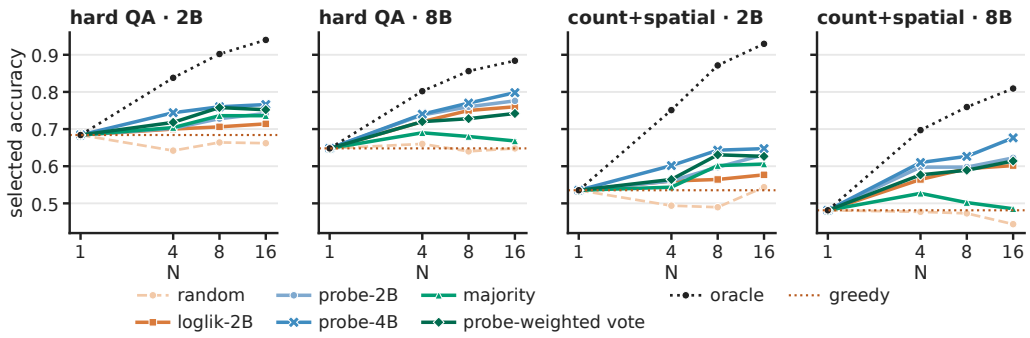

Figure 8: Accuracy of the selected candidate against N on 500 hard questions and the 241-question counting/spatial subset, for the 2B and 8B generators, with every training-free control on every panel. Where the generator is confidently wrong (8B) every external scorer beats majority vote; on the 2B pool probe, log-likelihood and vote are within noise. probe-8B omitted (it tracks probe-2B); its numbers are in Appendix F.

Table 6: **Best-of-16 on hard verifiable QA** (500 questions on COCO images, and the 241-question counting/spatial subset), accuracy of the selected candidate for each selector. *greedy* is $N{=}1$; *oracle* is pass@16. Paired bootstrap over prompts; single seed per selector. Where the 8B generator is confidently wrong (mean top-share .85) every external scorer beats majority vote; on the 2B pool the probes and log-likelihood are within noise of each other and of the vote.

cell	greedy	random	shortest	loglik-2B	probe-2B	probe-8B	majority	wvote	oracle
hard QA, gen 2B	.684	.672	.580	.714	.742	**.762**	.736	.752	.940
hard QA, gen 8B	.648	.648	.448	.760	**.776**	.740	.668	.742	.884
counting/spatial, gen 2B	.535	.502	.440	.577	**.631**	.622	.606	.627	.929
counting/spatial, gen 8B	.481	.436	.253	.602	**.622**	.610	.485	.614	.809

Table 7: **One turn of the loop, 2B.** Top: CHAIRs at $N{=}16$ per captioning cell (lower is better); *verified* is the head trained on 3,462 verifier-minted pairs, *held-out* the same recipe with every COCO-2 pair excluded, scored on COCO-2. Pooled paired bootstrap against greedy: -5.1 CHAIR points, 95% CI -7.1 to -3.1; held-out -4.1, 95% CI -6.9 to -1.1. Bottom: the same heads as judges on the two unseen reward benchmarks, against the mined-negative probe they replace (the 15k-pair rebuild that scores every Best-of-N cell). The best ranker is the worst recent judge.

captioning cell	greedy	probe-8B	verified	held-out	oracle
COCO, gen 2B	.346	.376	**.312**	–	.066
COCO, gen 8B	.374	.400	**.332**	–	.090
COCO-2, gen 2B	.362	.396	**.294**	.324	.056
COCO-2, gen 8B	.402	.414	**.344**	.358	.094

judge	VL-RB	halluc.	MMRB2 reas.
verified, fresh init	.615	68.8	.539
verified, perturbation init	.615	68.9	.540
mined-negative probe	**.679**	**76.9**	.533

G ACTOR TRAINING

We train the 2B policy on its own selected candidates against a size-matched random-selection control. DPO (Rafailov et al., 2023) uses β=0.1, learning rate 10^{-6}, two epochs, and three corpora from the same 500 prompts per task: verifier-minted pairs, probe-8B argmax-versus-argmin pairs, and random pairs as the control. Policies are re-evaluated greedily on all four tasks.

The actor half of the turn shows no selector effect, though it has a safe recipe. DPO costs nothing on QA: no change is outside ± 1.8 points, and the per-object hallucination rate on COCO falls for every corpus (CHAIRi .083 to .073, .079 and .082 for the verified, RM-picked and random pairs), with the label adding little over random pairs (Table 8). The size-matched random control plays the role of the untrained-initialization control in the text paper. The RM-picked corpus is the only one that shortens captions (7%), yet its per-object rate also falls, so the change is not purely length. One DPO turn on self-generated pairs is a mild, label-agnostic hallucination reducer, and the actor side of the loop still lacks a demonstrated selector effect.

Table 8: **Actor, DPO (2B policy).** Full-parameter DPO (β=0.1, lr 10^{-6}, 2 epochs, vision tower frozen) on three corpora from the same 500 prompts $\times$ 4 tasks: *verified* (3,462 pairs), *rm* (probe-8B argmax vs argmin, 2,000), *random* (two random candidates, 1,999; the control). Greedy regeneration at 512 tokens; Δ against the regenerated base, paired bootstrap over prompts (2,000 resamples), every QA CI including zero. CHAIRi is hallucinated object mentions over all object mentions, pooled over the 500 COCO captions, which shorter captions do not lower by themselves. Best value per column in bold.

	COCO		MathVista	MMMU
setting	words	CHAIRi ↓	acc (Δ)	acc (Δ)
base (regenerated)	245	.083	**.546**	.334
verified (3,462 pairs)	242	**.073**	.542 ($-$.004)	**.346** ($+$.012)
rm (2,000)	228	.079	.528 ($-$.018)	.338 ($+$.004)
random (1,999, control)	245	.082	.534 ($-$.012)	.332 ($-$.002)

The DPO base policy was regenerated at 512 new tokens under the identical protocol. Between 19 and 27% of DPO pairs were skipped by the 1,024-token cap, mostly long QA chains of thought, so the QA share of each corpus is below nominal; the update was small (loss .69 to .64 to .68). The image-generation counterpart (LoRA (Hu et al., 2022) on Qwen-Image (Qwen Team, 2025b) for 160 steps on 640 on-policy images, across seven Diffusion-DPO (Wallace et al., 2024) and related fine-tuning variants) has the same shape. No variant differs from the base generator on the hard prompts (all Δ within $[-.04, +.07]$), and every selection variant minus its random control is inside ± 2 points.

H CLIP STUDIES

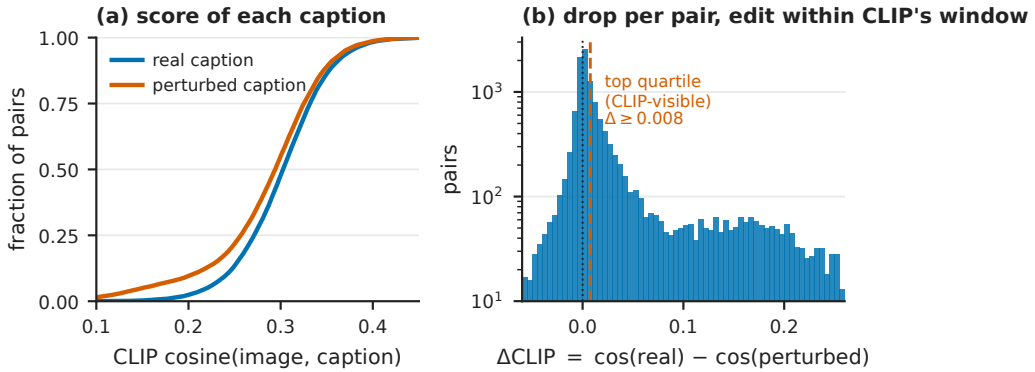

Figure 9: **CLIP view of the perturbation pairs** (22,000 same-image pairs, CLIP ViT-B/32). (a) Image–caption cosine of the real and the perturbed caption; the perturbation lowers the median from .303 to .294 and adds a low-score tail. (b) The drop ΔCLIP per pair, which the dissociation split of Section 3.3 ranks on, on a log count axis; the dashed line is the top-quartile cut ($\Delta \geq .008$) that defines the CLIP-visible arm. CLIP reads 77 tokens and the captions average 150 words, so for 43% of pairs the edit falls beyond its window and $\Delta = 0$ exactly; the CLIP-blind quartile is drawn from those pairs and is omitted from the histogram.

Figure 9 shows the score distribution the quartile cuts are taken from. Filtering the positives by image–caption CLIP score gives .685 / .698 / .678 on Multimodal RewardBench for the top, random and bottom quartile and .670 / .686 / .682 on VL-RewardBench. By contrast, the ΔCLIP split ranks each pair by the drop in CLIP cosine from the real to the perturbed caption. The CLIP-blind quartile has near-zero drop, part of it because CLIP truncates at 77 tokens and the edit falls outside the window.

I READING THE TWO STATEMENTS

The convexity fact of Section 2.2 predicts that optimizer settings cannot move the plateau, and predicts nothing about the negative pool, the hardness schedule or the temperature, each of which changes the population objective rather than the path to its minimizer. The temperature is the least obvious of these: writing $u = \beta w$ turns the weight decay into $\frac{\lambda}{2\beta^2}\|u\|^2$, so β acts as an effective regularization strength that rescales $w^\star$ and tilts its direction. That those interventions nonetheless leave the benchmark where it was (Section 3.1) is an empirical finding, and Eq. (5) says what it means: each re-weights the same difference vectors, and on the benchmark's pairs the re-weighted heads order alike; the CLIP split of Section 3.3 is the same equation read at a re-weighting large enough to separate on hallucination. The value at the minimizer is the best linear separation of positives from negatives in the layer $N-k$ feature space, which is why depth moves the plateau where no optimizer setting does, and convexity fixes the endpoint and not the distance to it, which is where initialization re-enters (Section 3.4).

Proposition 1 is stated for the decoupled objective because that is the one with a closed-form optimum, and for mutually absolutely continuous P and Q because that is what keeps the log ratio finite. Our sampler does not satisfy the hypothesis: a caption of another image has probability zero as a caption of this one, so on those pairs the unrestricted optimum runs to $\pm\infty$ and separates the two laws with unbounded margin. The in-distribution accuracy of the recipe is the finite-sample face of that, with all 540 configurations of Section 3.1 at .995 or better, and it is why the head that gets trained is pinned down by the weight decay and the linear class of Corollary 1 rather than by Eq. (4). The diagonal terms alone would leave a context-dependent offset free, making scores comparable within a prompt and not across prompts; the matrix form removes that freedom, and our same-image-only ablation (Section 3.1) confirms that the offset does not change within-prompt accuracy. The matrix form also decides what the target does with length. Fan et al. (2026) prove that a nuisance the negatives match in distribution enters the unrestricted optimum with coefficient zero, and remark that a nuisance the comparison holds fixed pairwise, as our substitution edits (two thirds of the bank) hold length on the diagonal, leaves the optimum free to use it in any way; splice and shuffle shorten

the negative, so on their pairs length is not held fixed either. The cross-entry pairs are the only ones that compare the negatives' length law with the positives', so it is the 87.5% of the loss and not the 12.5% that makes the target of Eq. (4) nearly length-free. A linear head cannot represent Eq. (4) exactly and fits the error-weighted mean difference of Eq. (5) instead, whose coefficient on length is not the target's; the sign flip of Section 3.2 is that head-side coefficient changing with the backbone under one sampler.

I.1 What Constructed Comparisons Supervise

Fix a context x (an image and its prompt) and a finite response set $\mathcal{Y}$, and let $J_x(y, y')$ be the probability that a training comparison designates y as chosen and y' as rejected, so that the conditional population loss is

$$\mathcal{L}_x(s) \;=\; \sum_{y,y' \in \mathcal{Y}} J_x(y, y')\, \ell\big(s(x, y) - s(x, y')\big), \qquad \ell(v) = \log(1 + e^{-v}). \tag{6}$$

If $\mu(y \mid x)$ is the distribution of real captions and $K_x(y' \mid y)$ constructs the negative, then $J_x(y, y') = \mu(y \mid x)K_x(y' \mid y)$: independent negatives use $K_x(y' \mid y) = \nu(y')$, our cross-image operator $\mathcal{T}_\times$, and a deterministic edit uses $K_x(y' \mid y) = \mathbf{1}\{y' = \mathcal{C}(y)\}$. Judged candidates give a J_x in which winner and loser are dependent after labeling, so their marginals alone do not determine the objective; the joint comparison law does. These are population statements about unrestricted scalar scores and carry no finite-sample or downstream guarantee.

I.2 Independent Negatives: Density Ratios and PMI

With $J_x(y, y') = \mu(y \mid x)\, \nu(y')$ and a common support S_x, the argument of Proposition 1 with the context held fixed gives the unrestricted minimizers of Eq. (6) on S_x,

$$s^*(x, y) \;=\; \log \frac{\mu(y \mid x)}{\nu(y)} + \phi(x), \qquad y \in S_x, \tag{7}$$

with $\phi(x)$ arbitrary. If $\nu = \mu_Y$, the caption marginal of the corpus law $\mu(x, y)$, then

$$s^*(x, y) \;=\; \log \frac{\mu(x, y)}{\mu_X(x)\, \mu_Y(y)} + \phi(x) \;=\; \mathrm{PMI}_\mu(x, y) + \phi(x), \tag{8}$$

the idealized case of using other images' captions as negatives. PMI measures image–caption association, not correctness or preference agreement; shared support is an assumption a finite corpus does not guarantee; and the result does not apply to edited copies or to judge-ordered candidates.

I.3 When Perturbations Expose an Additional Quality Signal

The following stylized example separates topic association from response quality. Represent a response as $y = (t, c)$, where t denotes its topic and $c \in \{0, 1\}$ indicates an incorrect or correct version. Suppose the corpus has distribution

$$\mu\big((t, c) \mid x\big) = p_x(t)\, b_\eta(c), \qquad b_\eta(1) = \eta, \quad b_\eta(0) = 1 - \eta, \quad \tfrac{1}{2} < \eta < 1. \tag{9}$$

Thus topics depend on the context, but the fraction of correct versions is the same across contexts and topics. Let $p(t)$ be the corpus topic marginal and assume $p_x(t) > 0$ on a common finite topic set. Independent corpus negatives have distribution $\mu_Y(t, c) = p(t)b_\eta(c)$, so Eq. (7) gives

$$s^*_{\mathrm{ind}}(x, t, c) \;=\; \log \frac{p_x(t)\, b_\eta(c)}{p(t)\, b_\eta(c)} + \phi(x) \;=\; \log \frac{p_x(t)}{p(t)} + \phi(x). \tag{10}$$

The quality factor cancels: the optimum distinguishes topics by their association with the context but assigns equal scores to correct and incorrect versions within a topic.

Now consider an idealized edit $\mathcal{C}(t, c) = (t, 1 - c)$ that switches the version while preserving the topic, with the original always labeled chosen. This label agrees with correctness with probability η and is reversed with probability $1 - \eta$; the corpus need not be perfectly correct. For a fixed topic,

write $\Delta_t = s(x,t,1) - s(x,t,0)$. Its conditional loss and minimizer are

$$\mathcal{L}_{\text{edit},t}(\Delta_t) = \eta\,\ell(\Delta_t) + (1-\eta)\,\ell(-\Delta_t), \qquad \Delta_t^* = \log\frac{\eta}{1-\eta} > 0. \tag{11}$$

Indeed, the derivative is $\sigma(\Delta_t) - \eta$, which vanishes uniquely at the stated margin. Perturbation training therefore ranks the correct version above the incorrect one, a distinction absent from the independent-negative optimum in this example.

A Bradley–Terry judge with reward $r(x,t,c) = \beta c$, $\beta > 0$, gives the same within-topic ordering with margin β; the edited pairs recover the ordering with a margin set by η instead. The example shows an additional quality distinction under the stated factorization and edit assumptions, not a general superiority of perturbations: changing quality-bearing content inside an otherwise matched response can supply a preference that corpus association alone does not teach, and whether actual edit rules do so, and whether it transfers, is what the experiments test.

J PROOFS

Convexity of the regularized objective. Write the population objective as

$$L(w) \;=\; \mathbb{E}\,\ell\big(\beta\, w^\top d\big), \qquad d = h^+ - h^-, \qquad \ell(z) = \log\big(1 + e^{-z}\big).$$

The function ℓ is twice differentiable with

$$\ell''(z) \;=\; \sigma(z)\,\sigma(-z) \;>\; 0,$$

so it is strictly convex, and a strictly convex function of an affine map is convex.

The Hessian of L is

$$\nabla^2 L(w) \;=\; \beta^2\,\mathbb{E}\Big[\ell''\big(\beta w^\top d\big)\, d\, d^\top\Big].$$

It is positive definite on $\operatorname{span}\{d\}$, because the weights ℓ'' are strictly positive, and zero on the orthogonal complement, because $w^\top d$ does not depend on that component. Hence L is strictly convex on the span and constant on its complement.

Adding the penalty turns the objective into

$$L_\lambda(w) \;=\; L(w) \;+\; \frac{\lambda}{2}\, w^\top D w, \qquad \lambda > 0,\; D \succ 0,$$

which is $\lambda\,\lambda_{\min}(D)$-strongly convex on all of $\mathbb{R}^d$. It therefore has exactly one minimizer, and any first-order method with a suitable step size converges to it from every initialization. Isotropic weight decay is $D = I$. AdamW's decoupled decay (Loshchilov & Hutter, 2019) updates $w \leftarrow w - \eta(P\nabla L(w) + \lambda w)$ with P the diagonal Adam preconditioner, so a stationary point satisfies $\nabla L(w) + \lambda P^{-1} w = 0$, the first-order condition of L_λ with $D = P^{-1}$ (Zhuang et al., 2022); P changes along the run, so the trained heads solve this problem only approximately, and their convergence is checked empirically in Section 3.1.

Initialization, step size, sample order and the number of steps change only the iterate sequence, so at fixed p, q_{neg}, β and λ none of them changes the minimizer.

The negative pool, its hardness and its composition change q_{neg} and therefore the law of d, so they may move the minimizer; Eq. (5) says how, through the error-weighted mean of d. The temperature is not neutral either. Writing $u = \beta w$,

$$L_\lambda(w) \;=\; \mathbb{E}\,\ell\big(u^\top d\big) \;+\; \frac{\lambda}{2\beta^2}\,\|u\|^2,$$

so β enters as an effective regularization strength λ/β^2 and a change in it rescales $w^\star$ and, since ridge shrinkage is isotropic while $\mathbb{E}[dd^\top]$ is not, tilts its direction as well. Only the accuracy of a *ranking* is invariant to the scale of $w^\star$, not to its direction. □

Proof of Proposition 1. Let $P = \mu \otimes p$ and $Q = \mu \otimes \bar{q}$ be the joint laws of a positive and of a negative triple $a = (x, q, y)$, with $\bar{q}$ the marginal law of a negative as in the statement. The

off-diagonal population loss over unrestricted scores s is

$$J(s) \;=\; \mathbb{E}_{a\sim P,\, b\sim Q}\, \ell\big(\beta\,(s(a)-s(b))\big),$$

where a and b are independent because they come from different micro-batch entries. We assume throughout that P and Q are mutually absolutely continuous, so that $\mathrm{d}P/\mathrm{d}Q$ is finite and positive almost everywhere. The hypothesis is what makes the minimizer a real-valued function: on a pair where one of the two densities vanishes, the stationarity condition below puts $\delta^\star$ at $\pm\infty$, the infimum of J over real-valued scores is not attained, and every sequence approaching it separates the two laws with unbounded margin. Our own sampler is of that kind, which is why the recipe's weight decay and its linear class, and not Eq. (4), are what pin down the head that gets trained (Corollary 1).

Group the contributions of each unordered pair $\{a, b\}$ of triples. The pair appears once with a as the positive, at density $P(a)Q(b)$, and once with b as the positive, at density $P(b)Q(a)$, so for a fixed difference $\delta = s(a) - s(b)$ its contribution is

$$P(a)Q(b)\,\ell(\beta\delta) \;+\; P(b)Q(a)\,\ell(-\beta\delta).$$

This is strictly convex in δ, and its derivative vanishes where

$$\sigma(\beta\delta) \;=\; \frac{P(a)Q(b)}{P(a)Q(b)+P(b)Q(a)}, \qquad \text{that is, at} \qquad \beta\,\delta^\star \;=\; \log\frac{P(a)}{Q(a)} \;-\; \log\frac{P(b)}{Q(b)}.$$

The function $s^\star = \beta^{-1}\log(\mathrm{d}P/\mathrm{d}Q) + C$ attains $\delta^\star$ for every pair simultaneously, so it minimizes every summand and therefore J. Any other minimizer attains $\delta^\star$ on almost every pair as well, because each summand is strictly convex in δ, so it differs from $s^\star$ by a constant on the support of P and Q. Since the two laws share the context law μ,

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x,q,y) \;=\; \frac{\mu(x,q)\,p(y\mid x,q)}{\mu(x,q)\,\bar{q}(y\mid x,q)} \;=\; \frac{p(y\mid x,q)}{\bar{q}(y\mid x,q)},$$

which is Eq. (4). The pairs run over triples and not over responses within one context, so no context-dependent offset is left free; only the constant C is. The factor β^{-1} is where the temperature enters the target, and a linear head absorbs it into the scale of w.

The B diagonal terms are outside this argument. There b is an edit of the same a it is scored against, so the pair law is $P(a)\,q_{\mathrm{neg}}(b\mid a)$ rather than the product $P(a)Q(b)$, the two orderings of $\{a, b\}$ no longer enter at weights that depend only on the marginals, and $\delta^\star$ above is not the pointwise minimizer. We therefore state the proposition for the decoupled objective, which the recipe optimizes on $1 - 1/B = 87.5\%$ of its terms at B=8. □

Proof of Corollary 1. The population objective with the penalty is

$$L_\lambda(w) \;=\; \mathbb{E}\,\ell\big(\beta\,w^\top d\big) \;+\; \frac{\lambda}{2}\,w^\top D w,$$

with the expectation over the pair law of the matrix loss: each micro-batch contributes B same-entry pairs and $B^2 - B$ cross-entry pairs, at total weights $1/B$ and $1 - 1/B$. Since $\ell'(z) = -\sigma(-z)$,

$$\nabla L_\lambda(w) \;=\; -\beta\,\mathbb{E}\big[\sigma(-\beta\,w^\top d)\,d\big] \;+\; \lambda\,Dw,$$

and L_λ is strongly convex (previous paragraph), so its unique minimizer is the unique zero of this gradient, which is Eq. (5). The inner product with a unit vector v gives the coefficient along v. In particular $Dw^\star$ is a positive combination of difference vectors and lies in their span, which with more pairs than dimensions is all of $\mathbb{R}^d$ and constrains nothing; what Eq. (5) fixes is the combination.

Now suppose a unit direction v satisfies $v^\top d = 0$ almost surely on the same-entry pairs, the pairwise form of length preservation (exact for the substitution operators, approximate for the bank). Those pairs then contribute nothing to the coefficient along v, and the cross-entry pairs, which compare captions of different images, contribute

$$\frac{\beta}{\lambda}\Big(1 - \frac{1}{B}\Big)\,\mathbb{E}_{\times}\big[\sigma(-\beta\,w^{\star\top} d)\,v^\top d\big],$$

an error-weighted mean of gaps between unrelated captions, which the objective does not set to zero. Pairwise preservation on the diagonal therefore leaves the coefficient along v to the cross-entry pairs

rather than fixing it at zero, the same conclusion as for the unrestricted optimum above, where a comparison that holds a nuisance fixed leaves its coefficient undetermined.

For two samplers q_1, q_2 with difference laws $D_1 \neq D_2$, Eq. (5) gives minimizers w_1, w_2 that coincide only if the two error-weighted means do, a condition on the laws and not on their spans. A benchmark pair set $\{e_k\}$ measures the sign pattern

$$\operatorname{sign}\left(w^\top e_k\right), \qquad k = 1, \ldots, K,$$

which coincides for w_1 and w_2 whenever the two heads order the benchmark's pairs alike, however far apart they are on other pairs. □

K Compute Resources

Training a head is the cheapest step in the study. Once the frozen backbone's hidden states are stored, a 540-configuration grid at 2B fits in about nine GPU-minutes, and a single head is 200 steps of a $d \to 1$ linear map. The 32B confirm run used FSDP over eight GPUs at effective batch 1,024. What costs is everything that is not the reward model: one forward pass of the backbone per pair to store its hidden states, Best-of-N sampling at $N{=}16$ over 2,500 prompts for two policy sizes, and the image arm's two same-seed Qwen-Image renders per prompt at about 7 GPU-seconds each, which is about 5 GPU-hours for its 1,212 training pairs. The caption negative bank itself is free: 10 seconds on one CPU core for 22k pairs, with no model in the loop. The figure the paper quotes against fine-tuning, about a thousandth of the compute, counts head training against the full fine-tune of the same backbone on the same pairs and excludes the backbone forward passes both settings share.