SCORED BY ITS OWN REWARD: EVALUATOR-DEPENDENT DENSE-REWARD SPEED-UPS IN DIFFUSION RL

Anonymous authors
Paper under double-blind review

ABSTRACT

The measured advantage of a graded reward can depend on the evaluator used to define success. We compare a binary CLIP-threshold reward with a dense clipped-linear CLIP reward during DDPO fine-tuning of Stable Diffusion 1.5 on 14 compositional prompts. Two campaigns record CLIP scores; campaign B additionally scores each rollout with OWLv2 and BLIP-VQA, neither used for optimization. Hard-setting thresholds target a nominal 10% base-policy success rate. The dense reward reaches the CLIP success-rate target of 0.5 earlier by 8.45 and 7.19 epochs in the two campaigns (26.8% and 22.6% fewer epochs). In campaign B, on the 13 prompts with identifiable thresholds for all three evaluators, success changes from the initial block to the final ten-block mean from .104 to .914/.901 under CLIP, from .125 to .134/.146 under OWLv2, and from .121 to .182/.193 under BLIP-VQA (dense/binary). On these prompts, the dense-minus-binary success-rate AUC effects are +.0511, −.0101, and −.0006, respectively. The two CLIP-minus-external contrasts are +.0612 and +.0517, with simultaneous 95% confidence intervals [.0359, .0865] and [.0293, .0741]; complementary sign-flip tests give $p < .001$. Under these evaluators, the dense-reward speed-up is largely specific to the optimized CLIP score.

1 INTRODUCTION

Graded rewards assign different values to outcomes that receive the same binary success label. In text-to-image diffusion fine-tuning, objectives include CLIP similarity, aesthetic predictors, and learned preference models (Black et al., 2024; Fan et al., 2023; Xu et al., 2023; Clark et al., 2024). A comparison evaluated only through the optimized score measures progress on that score; it leaves open whether the reward-condition difference persists under another evaluation criterion. Analyses of proxy optimization distinguish proxy values from the intended objective (Gao et al., 2023; Skalse et al., 2022), motivating explicit comparisons across evaluators.

We test whether the dense-minus-binary effect persists under automated criteria not used for optimization, while holding the generator, optimizer, and prompt set fixed. Stable Diffusion 1.5 is fine-tuned with DDPO under a binary CLIP-B/32 threshold reward or a dense reward that replaces the step with a clipped linear ramp. In campaign B, the same rollouts are additionally scored by an OWLv2 detection criterion and a BLIP-VQA criterion based on a prompt-specific yes/no question. Neither auxiliary evaluator affects training. Each evaluator uses a prompt-specific threshold calibrated to a nominal 10% base-policy success target in the hard setting. These model-based criteria are not treated as human-validated measures of compositional correctness.

Let D_j denote the dense-minus-binary success-rate AUC effect under evaluator j. Separate significance tests for D_{CLIP} and an external-evaluator effect do not test whether the two effects differ (Gelman & Stern, 2006). We therefore estimate $D_{\mathrm{CLIP}} - D_{\mathrm{OWLv2}}$ and $D_{\mathrm{CLIP}} - D_{\mathrm{BLIP\text{-}VQA}}$ on the same 13 prompts and test these two contrasts jointly with paired prompt-level inference.

The experiments yield three observations. First, the dense reward reaches the CLIP success-rate target 8.45 and 7.19 epochs earlier in the two campaigns. Second, from the initial rollout block to the final ten-block mean in campaign B, success increases by 81.0/79.7 percentage points under CLIP, 1.0/2.1 under OWLv2, and 6.2/7.2 under BLIP-VQA (dense/binary). Third, both CLIP-minus-external

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.

reward-condition contrasts have simultaneous confidence intervals above zero. Their point estimates remain positive across the evaluated horizons and aggregation rules, and complementary sign-flip tests give p < .001.

2 Related work

Diffusion reinforcement learning. DDPO formulates denoising as a multi-step decision process and optimizes downstream rewards with policy gradients (Black et al., 2024); DPOK adds KL regularization (Fan et al., 2023), and ReFL and DRaFT backpropagate through differentiable rewards (Xu et al., 2023; Clark et al., 2024). Prior studies report automatic and, in some cases, human evaluations. We use Stable Diffusion 1.5 (Rombach et al., 2022) with the TRL DDPO implementation (von Werra et al., 2020), holding the optimization configuration fixed while changing the reward. Its prompt-specific rolling-buffer normalization differs from the current-group statistics of GRPO (Shao et al., 2024; Section 3.2).

Reward shaping and proxy optimization. Potential-based shaping preserves optimal policies (Ng et al., 1999); our ramp is not potential-based. Optimizing a learned proxy can raise the proxy while the intended objective stalls (Gao et al., 2023; Skalse et al., 2022). We ask a narrower question: whether the relative advantage of one reward over another, measured under the optimized proxy, is preserved under criteria that were not optimized.

Image-text and compositional evaluation. CLIP learns joint image–text representations from natural-language supervision (Radford et al., 2021); OWLv2 provides open-vocabulary detection (Minderer et al., 2023), and BLIP supports visual question answering (Li et al., 2022). Question-based faithfulness metrics such as TIFA and DSG (Hu et al., 2023; Cho et al., 2024) and compositional benchmarks such as T2I-CompBench and GenEval (Huang et al., 2023; Ghosh et al., 2023) assess objects, counts, attributes, and relations against human judgments. Our OWLv2 and BLIP-VQA scores are simplified, prompt-specific versions of such criteria.

Study design relative to prior work. Relative to prior work, we hold the generator, optimizer, and prompts fixed, compare two rewards derived from the same score, report the absolute trajectories of the same runs under each evaluator, and test the difference between evaluator-specific effects directly rather than comparing separate significance statements.

3 Experimental design and measurement

3.1 Training conditions and experimental units

The image experiments fine-tune Stable Diffusion 1.5 with DDPO under two reward conditions derived from the same CLIP-B/32 image–text similarity $s \in [0, 1]$. We compute s as the cosine similarity between normalized image and text embeddings, mapped from [−1,1] to [0,1]. We refer to this score as CLIP unless a different backbone is specified. The binary condition is labeled sparse in the implementation,

$$R_{\text{sparse}}(s) = \mathbf{1}\{s \geq \tau\}.$$

The dense reward replaces this step with a linear ramp of width $W = 0.03$,

$$R_{\text{dense}}(s) = \text{clip}\left(\frac{s - (\tau - W)}{W}, 0, 1\right).$$

The rewards differ on $\tau - W < s < \tau$ and coincide outside this interval. The ramp width is approximately three times the mean prompt-specific base-policy calibration standard deviation. In the separate 64-image-per-prompt base-policy evaluation set, 83.7% of hard-setting samples fall inside the ramp and 4.9% fall below it (Figure 1, right); the in-ramp fractions are 64.6% and 36.2% in the medium and easy settings. Thus, partial credit covers most below-threshold samples in this

hard-setting evaluation set. The threshold τ is calibrated per prompt and difficulty from 96 pre-RL base-policy images: the 40th, 70th, and 90th percentiles define the easy, medium, and hard settings, with nominal base success targets of 60%, 30%, and 10%.

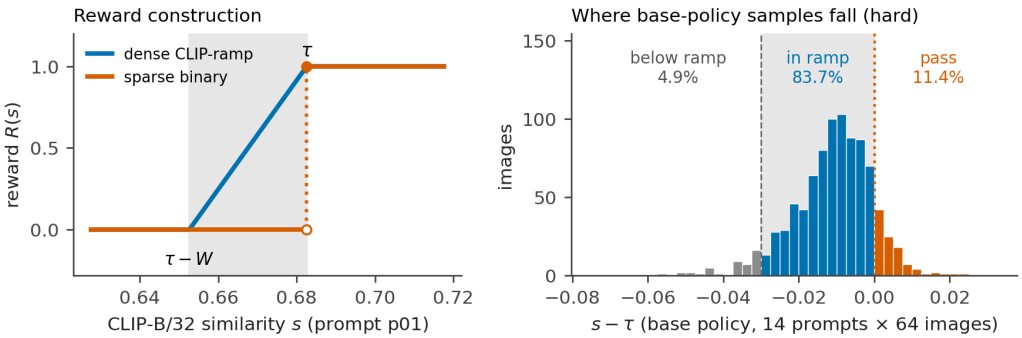


Figure 1: Reward construction. Left: the binary reward is an indicator at τ, and the dense reward uses a ramp of width $W = 0.03$; the rewards differ only on $\tau - W < s < \tau$. The plotted $\tau = 0.6826$ is the hard-setting CLIP threshold for p01. Right: $s - \tau$ for 64 base-policy evaluation images per prompt, pooled over 14 prompts. Of these images, 83.7% lie inside the ramp and 4.9% below it. The 96-image calibration set is separate. Prompt p01 is excluded only from the cross-evaluator comparison because its OWLv2 hard threshold is unidentifiable.

Each training process uses one fixed prompt. Campaign B's hard setting has 14 prompts, two reward conditions, and three nominal seeds (42, 1337, 2026), giving 84 processes. For all 42 campaign-B prompt–seed pairs, the logged CLIP mean scores agree across the eight reward calls in the initial rollout block and diverge in later blocks. This comparison concerns logged mean scores, rather than a saved pixel-level or initial-weight identity check.

Table 1: Image-experiment configuration. Campaign A includes easy, medium, and hard conditions; campaign B includes hard only. Each campaign–difficulty block has 84 separately trained policies. Listed settings describe the experimental implementation. Runtime figures are reported execution summaries, not measurements of reward-condition speed-up.

Component	**Setting**
Generator / optimizer	Stable Diffusion 1.5 / DDPO (TRL DDPOTrainer)
Reward conditions	Dense CLIP-ramp (W = 0.03) / sparse binary; same prompt-specific τ
Experimental unit	One prompt per separately trained policy
B-hard processes	14 prompts $\times$ 2 conditions $\times$ 3 seeds = 84
Training horizon	100 epochs; 64 images per epoch; 6,400 per run
Reward calls	8 calls per epoch; 8 images per call
Adapter	LoRA r = 4, α = 4 on UNet attention projections
Optimizer	AdamW; learning rate 10^{-4}; weight decay 10^{-4}; fp16
Policy update	1 inner epoch; 8 optimizer steps per epoch; PPO clip 10^{-4}; advantage clip ± 5; gradient-norm clip 1.0; no KL term
Normalization	Prompt-specific buffer of the last 16 rewards (population SD, $\epsilon = 10^{-6}$), applied to all 64
Calibration	96 base-policy images per prompt and evaluator
Generation settings	30 denoising steps; guidance 5.0; η = 1.0
Compute	One NVIDIA A40 per run; median 7.8 h; about 2,600 GPU-hours for 336 runs

Campaigns A and B are separate executions of the same protocol and nominal seeds, not copies of one trajectory. Campaign A records CLIP measurements; campaign B additionally records OWLv2 and BLIP-VQA. We do not pool the campaigns as six different seeds or describe evaluator interaction as independently replicated across both campaigns. Each epoch generates 64 training images. We report epoch-based progress and do not convert it into GPU-time or total-compute savings without accounting for optimization and reward-evaluation costs.

3.2 Reward normalization in the executed implementation

DDPO uses prompt-specific reward normalization with a rolling buffer of 16 rewards. At each 64-sample epoch, the buffer is updated before advantage computation, so its mean and standard deviation are determined by the final 16 rewards and then applied to all 64 samples:

$$A_i = \frac{r_i - \mu_{16}}{\sigma_{16} + 10^{-6}}, \qquad i = 1, \ldots, 64. \tag{1}$$

The normalized advantages are clipped to $[-5, 5]$. Because normalization uses 16 of the 64 samples, their sum over an epoch is not generally zero. Under the binary reward, an all-zero buffer assigns a clipped advantage of +5 to successes, whereas an all-one buffer assigns -5 to failures. In campaign B, all-zero buffers occur in 8.1% of the first 20 binary-reward epochs and 0% of dense-reward epochs. All-one buffers occur in 15.8% of dense-reward epochs and 12.7% of binary-reward epochs over the full horizon, predominantly in the second half. Appendix A.2 describes the reconstruction.

3.3 Calibration and the common prompt set

Because CLIP, OWLv2, and BLIP-VQA use different score scales, we calibrate a separate success threshold for each evaluator and prompt. In the hard setting used for our main cross-evaluator analysis, the threshold is the 90th percentile of scores from 96 pre-RL base-policy images, targeting an approximately 10% baseline success rate. Calibration uses the same nominal target while retaining each evaluator's definition of success. Easy and medium thresholds similarly target approximately 60% and 30% base-policy success, respectively. For CLIP, the achieved calibration rate was not stored; the first logged training block, generated before any update, has a CLIP success rate of .104 on the 13 common prompts. Per-image calibration scores were not retained for any evaluator. Appendix A.2 defines the OWLv2 and BLIP-VQA scores.

For p01, approximately 91% of OWLv2 calibration scores are zero, and the adopted calibration procedure flags the hard threshold as unidentifiable. We exclude p01 from all three evaluators in the cross-evaluator comparison, yielding 13 common prompts. CLIP-only analyses retain all 14 prompts.

3.4 Training trajectories and separate evaluation

Training trajectories are computed from the reward calls used during DDPO optimization: 8 calls of 8 images per epoch and 800 calls per run, for 336 runs (campaign A: easy, medium, hard; campaign B: hard; 84 runs per block). Campaign B additionally logs OWLv2 and BLIP-VQA success counts and mean scores for the same images. Final-checkpoint assets differ by campaign: campaign A has 64 newly generated images per cell scored by CLIP-B/32 and CLIP-L/14, and campaign B has an eight-image-per-cell archive scored by all three evaluators. Intermediate-epoch images are not stored. Appendix A.4 lists all assets.

3.5 Effect definitions and statistical analysis

For each epoch, we define an evaluator-specific success rate as the fraction of the 64 generated images that pass that evaluator's calibrated threshold. For condition c, prompt p, run r, evaluator j, and epoch t, let K_{cprjtb} be the success count in reward call b. The epoch success rate and trajectory summary at horizon H are

$$S_{cprjt} = \frac{1}{64} \sum_{b=1}^{8} K_{cprjtb}, \qquad A_{cprj}(H) = \frac{1}{H} \sum_{t=0}^{H-1} S_{cprjt}. \tag{2}$$

We call $A_{cprj}(H)$ the success-rate AUC: the arithmetic mean over logged blocks $t = 0, \ldots, H-1$, rather than the area under a receiver-operating-characteristic curve. Logged block t is generated after t completed DDPO epochs, before the next epoch's policy optimization. Each epoch contains eight optimizer steps under Table 1's configuration. Continuous-score analyses replace success rates with recorded within-call means. Standardized effects divide each prompt-level difference by that evaluator's prompt-specific score standard deviation in the 96-image base-policy calibration summary.

Our main cross-evaluator analysis uses campaign B in the hard setting. For each prompt and evaluator, we first compute the 100-block success-rate AUC for each of the three nominal seeds (42, 1337, and 2026) and then average the three seed values. Prompts receive equal weight, and p01 is excluded because its OWLv2 hard threshold is not identifiable. Let $\overline{A}_{c,p,j}$ denote this seed-averaged AUC and define the prompt-level dense-minus-binary effect as $d_{p,j} = \overline{A}_{dense,p,j} - \overline{A}_{sparse,p,j}$. We average $d_{p,j}$ equally over the 13 common prompts to obtain the evaluator-specific effect D_j. Our primary cross-evaluator contrasts are

$$I_{\text{OWLv2}} = D_{\text{CLIP}} - D_{\text{OWLv2}}, \qquad I_{\text{BLIP-VQA}} = D_{\text{CLIP}} - D_{\text{BLIP-VQA}}. \tag{3}$$

A positive contrast means that the dense-minus-binary success-rate AUC difference is numerically larger under CLIP. The contrast is defined on success-rate units after evaluator-specific threshold calibration; it does not assume that the three raw evaluator scores share a common scale.

For uncertainty, we resample the 13 prompts as paired units, keeping both reward conditions and all evaluator summaries for a prompt together. We use 10,000 bootstrap resamples and a max-t correction across the two contrasts to construct simultaneous 95% confidence intervals. The resampling unit is the prompt; images, epochs, and prompt–seed cells are not resampled as independent units. The intervals describe variation under prompt resampling within the tested 13-prompt set and do not include uncertainty from untested prompts or additional training seeds. We also report individual percentile bootstrap intervals for each D_j and, for each contrast, a complementary two-sided sign-flip test under a null of sign exchangeability, enumerating all 2^{13} sign assignments of the prompt-level contrasts.

First-passage summaries use an epoch success-rate target of 0.5, distinct from the image-level score threshold. The first logged rollout block is evaluated before any policy update and is indexed at $t = 0$; passage time therefore counts completed DDPO epochs before the criterion is met. For the three-epoch sustained and five-epoch smoothed criteria, we report the zero-based start index of the first qualifying window. Under the CLIP criterion, all 168 hard-setting runs across campaigns A and B satisfy each of these three passage criteria within the 100-block horizon, so no hard-setting run is right-censored.

4 Results

4.1 CLIP first-passage time in two campaigns

Figure 2 shows the hard-setting CLIP success trajectories. Under the zero-based completed-epoch convention in Section 3.5, campaign A has mean first-passage times of 23.14 epochs for dense and 31.60 for binary, a difference of -8.45 epochs and a 26.8% relative reduction. Campaign B has corresponding means of 24.60 and 31.79, a difference of -7.19 epochs and a 22.6% relative reduction. The prompt-level mean first-passage time is lower for dense on 13 of 14 prompts in campaign A and 14 of 14 prompts in campaign B.

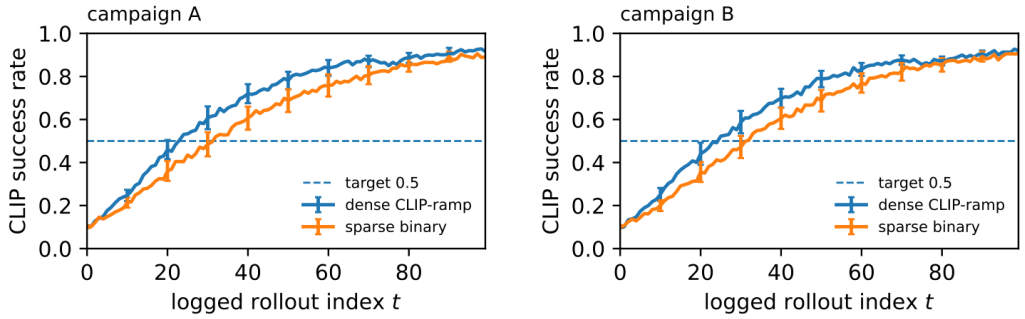


Figure 2: Hard-setting CLIP success trajectories. Left: campaign A; right: campaign B. The horizontal axis $t = 0, \ldots, 99$ counts completed DDPO epochs before each rollout block. Curves

show prompt means after averaging the three seeds within each prompt; error bars every ten blocks are one standard error across 14 prompts. Table 2 summarizes run-level passage times, not crossings of these mean curves.

4.2 Alternative trajectory summaries and difficulty

The dense-minus-binary differences in all three passage-time criteria are negative in both campaigns (Table 2). The 100-block AUC differences are +.068 [+.039, +.097] in campaign A and +.057 [+.034, +.084] in campaign B. In the final logged block ($t = 99$), the differences are +.025 [−.008, +.060] and +.014 [−.020, +.049], respectively. Both final-block intervals include zero.

Table 2: Hard-setting CLIP trajectory summaries on all 14 prompts. Δ= dense minus binary. Passage times are zero-based window-start indices (Section 3.5); AUC and final-block success are proportions. All values are recomputed from the stored training traces; intervals are paired prompt-level percentile bootstrap intervals (10,000 resamples, RNG seed 0). Differences are calculated before display rounding.

Campaign	Metric	Dense	Binary	Δ	95% CI
A	First passage	23.14	31.60	−8.45	[−12.81, −4.76]
A	Sustained, 3 epochs	27.90	36.43	−8.52	[−12.02, −5.43]
A	Smoothed, 5 epochs	24.95	33.26	−8.31	[−12.38, −5.00]
A	Success-rate AUC	.679	.611	+.068	[+.039, +.097]
A	Final block (t = 99)	.916	.891	+.025	[−.008, +.060]
B	First passage	24.60	31.79	−7.19	[−10.36, −4.62]
B	Sustained, 3 epochs	28.45	37.02	−8.57	[−12.86, −5.10]
B	Smoothed, 5 epochs	25.45	33.24	−7.79	[−11.62, −4.50]
B	Success-rate AUC	.672	.614	+.057	[+.034, +.084]
B	Final block (t = 99)	.920	.906	+.014	[−.020, +.049]

In campaign A, the sustained-passage differences are −8.52, −2.57, and −0.33 epochs in the hard, medium, and easy settings, respectively (Appendix Table 11). The easy setting has an initial success rate above the 0.5 passage target.

4.3 Absolute success trajectories under three evaluators

Figure 3 scores the same campaign-B rollouts under all three evaluators on the common 13-prompt set. Initial-rollout success rates are .104 (CLIP), .125 (OWLv2), and .121 (BLIP-VQA). The final ten-block means are .914/.901, .134/.146, and .182/.193, respectively (dense/binary). The corresponding increases are 81.0/79.7, 1.0/2.1, and 6.2/7.2 percentage points. These changes describe progress within each evaluator; the next subsection compares the dense-minus-binary effects directly. Appendix Figure 5 illustrates disagreement on a final-checkpoint image.

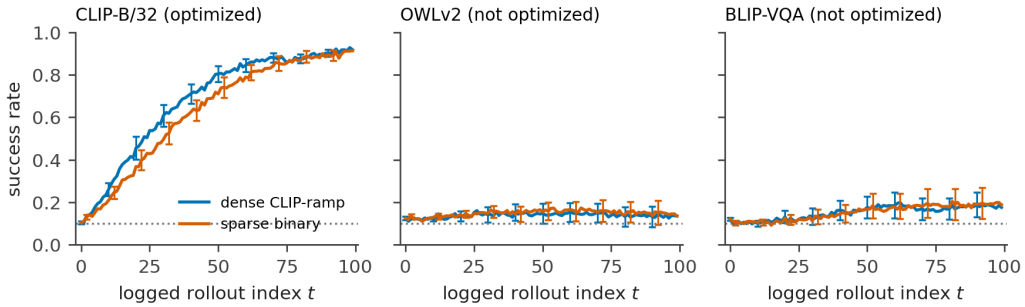


Figure 3: Success trajectories of the same campaign-B hard-setting rollouts (13 common prompts). The dotted line marks the nominal calibration target of 0.10; observed initial-rollout rates are .104,

.125, and .121 under CLIP, OWLv2, and BLIP-VQA. Curves show seed-averaged prompt means; error bars every ten blocks are one standard error across 13 prompts. Only CLIP enters the reward.

4.4 Evaluator-specific effects and their contrasts

On the common 13-prompt set, the dense-minus-binary success-rate AUC effects are $D_{\text{CLIP}} = +.0511$ [+.0301, +.0770], $D_{\text{OWLv2}} = -.0101$ [−.0204, −.0005], and $D_{\text{BLIP-VQA}} = -.0006$ [−.0206, +.0171] (individual percentile intervals). The OWLv2 effect is negative on 11 of 13 prompts. Its percentile interval excludes zero, whereas its sign-flip p-value is .078; inference for this individual effect depends on the procedure. The two CLIP-minus-external contrasts are +.0612 and +.0517, with simultaneous 95% intervals [.0359, .0865] and [.0293, .0741] (Figure 4) and sign-flip p-values .0002 and .0007.

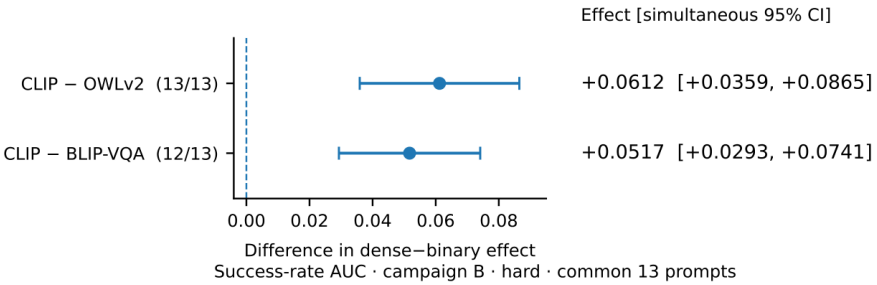


Figure 4: Evaluator-dependent effect differences in campaign B, hard setting. Each row compares the dense-minus-binary success-rate AUC effect under CLIP with the corresponding effect under one non-optimized evaluator on the common 13-prompt set. Positive values mean the relative dense-reward effect is larger under CLIP. Intervals are simultaneous 95% confidence intervals from a prompt-level paired bootstrap ($B = 10,000$, RNG seed 0) with a max-t correction across the two contrasts. Counts at left are prompts with a positive difference.

The CLIP effect exceeds the OWLv2 effect on 13 of 13 prompts and the BLIP-VQA effect on 12 of 13 (Appendix Table 8). Both contrasts are positive at all 20 evaluated horizons, $H = 5, 10, \ldots, 100$. Pointwise percentile intervals exclude zero for the evaluated $H \geq 10$ (Appendix B.5). Both contrasts also remain positive under the median-based summaries in Appendix B.1.

4.5 Continuous scores and a same-family evaluator

Continuous scores do not require a success threshold. On the common 13-prompt set, the 100-block mean CLIP score is higher under dense reward by +.0012 [+.0005, +.0022], with a sign-flip p-value of .004 for the raw-score difference (Table 3). The corresponding standardized effects are +0.13 [+0.06, +0.21] for CLIP, −0.05 [−0.12, +0.01] for OWLv2, and +0.03 [−0.04, +0.09] for BLIP-VQA. The latter two intervals include zero. These results concern trajectory-averaged scores, not changes in every part of the score distribution.

Table 3: Continuous-score dense-minus-binary effects over $t = 0, \ldots, 99$ (campaign B, hard, common 13 prompts). z divides each prompt-level difference by that prompt's 96-image base-policy calibration score standard deviation. Intervals are individual paired prompt-level percentile intervals; p is the sign-flip value for the raw difference, not z. Differences precede display rounding. Appendix Table 9 reports all-14 results.

Evaluator	Dense	Binary	Δ	95% CI	z	95% CI (z)	p
CLIP	.6762	.6749	+.0012	[+.0005, +.0022]	+0.13	[+0.06, +0.21]	.004
OWLv2	.3763	.3848	−.0085	[−.0210, +.0022]	−0.05	[−0.12, +0.01]	.203
BLIP-VQA	.7570	.7494	+.0076	[−.0124, +.0269]	+0.03	[−0.04, +0.09]	.485

At the campaign-A final checkpoint, 64 new images per cell were also scored by CLIP-L/14, which was not used for optimization. Pooling dense and binary final-checkpoint evaluations within each prompt, mean increases relative to base-policy images are 2.41 and 1.46 base-policy standard deviations under CLIP-B/32 and CLIP-L/14. This endpoint analysis uses campaign A and all 14 prompts; the OWLv2/BLIP-VQA trajectory analysis uses campaign B and the common 13. The final-checkpoint dense-minus-binary intervals include zero at the evaluated thresholds (Appendix B.7).

5 Discussion and limitations

What the dense speed-up measures. The dense reward reaches the CLIP-defined target 8.45 and 7.19 epochs earlier and yields a higher trajectory-averaged CLIP score in campaign B. For the same campaign-B rollouts, initial-to-final-ten-block success increases are 81.0/79.7 percentage points under CLIP, 1.0/2.1 under OWLv2, and 6.2/7.2 under BLIP-VQA (dense/binary). The reward-condition AUC effects also differ across evaluators. These are comparisons of model-based success definitions, not direct measurements of human-judged quality. Epoch-to-threshold estimates depend on both the evaluator and its threshold.

Validity of the external evaluators. OWLv2 and BLIP-VQA provide automated criteria rather than human judgments. The OWLv2 score is zero-inflated on several prompts, and a single BLIP-VQA question covers only part of each prompt. Evaluator disagreement can reflect differences in sensitivity or in the attributes being tested. Human evaluation would be needed to determine which measurements track compositional correctness.

Optimization details. In the recorded campaign-B runs, all-zero normalization buffers occur in 8.1% of the first 20 binary-reward epochs and in no dense-reward epoch (Section 3.2). This is a difference in the training signal under the fixed normalization rule, not a separately isolated cause of the speed-up. The configured PPO clip range is 10^{-4}; clip fractions were not logged.

Experimental scope. The estimates cover Stable Diffusion 1.5, DDPO, 14 fixed prompts, and the reward and normalization rules in Section 3; the cross-evaluator comparison uses the 13 prompts with identifiable thresholds. Unseen prompts, other generators, and other optimizers require separate experiments. Per-image scores for intermediate epochs, per-image calibration scores, and training-time advantage statistics were not stored. Supplementary autoregressive RLVR and discrete-diffusion experiments, which have no evaluator external to the reward and cannot test evaluator dependence, are reported in Appendices D and E.

6 Conclusion

Under the fixed DDPO configuration, a dense CLIP reward reaches a CLIP-defined target 8.45 and 7.19 epochs earlier than a binary reward in two campaigns. In campaign B, the same rollout trajectories yield different absolute success-rate changes under CLIP, OWLv2, and BLIP-VQA. On the common 13-prompt set, the dense-minus-binary AUC effects are +.0511, $-.0101$, and $-.0006$, respectively. The two CLIP-minus-external contrasts have simultaneous 95% intervals [.0359, .0865] and [.0293, .0741], both above zero.

A dense-reward speed-up measured with the optimized evaluator describes progress on that evaluator. Reporting absolute trajectories and reward-condition contrasts under both optimized and non-optimized evaluators makes the criterion dependence of the measured speed-up explicit.

Reproducibility statement

The manuscript specifies the experimental units, reward and evaluator definitions, calibration procedure, trajectory estimands, aggregation rules, and statistical procedures needed to reimplement the analyses. Table 1 lists the training configuration, including adapter, optimizer, policy-update, and normalization settings; Appendix A.2 and Table 5 give the 14 prompts, OWLv2 queries, count and relation specifications, BLIP-VQA questions, and the evaluator scoring rules; Appendix A.3

specifies the bootstrap and sign-flip procedures, including resampling units and random-number settings. All reported image-experiment results were recomputed by the authors from stored training traces, calibration summaries, and per-image scores, as documented in Appendices A.1, A.4, and A.5. The permutation experiment retains archived performance summaries; its data availability and interpretation are described in Appendices A.5 and C. The code, training traces, and per-image scores are not publicly released; the specifications above are provided for reimplementation.

AI USE STATEMENT

Generative AI tools assisted with manuscript restructuring, translation, prose editing, methodological and statistical review, interpretation of results, mathematical exposition, consistency checks, and preparation of figures and analysis code. AI-assisted revisions were checked against the underlying experimental records and source literature. The authors are responsible for the final manuscript, analyses, numerical results, and citations.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In International Conference on Learning Representations (ICLR), 2024. https://arxiv.org/abs/2305.13301.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian Scene Graph: Improving Reliability in Fine-Grained Evaluation for Text-to-Image Generation. In International Conference on Learning Representations (ICLR), 2024. https://arxiv.org/abs/2310.18235.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In International Conference on Learning Representations (ICLR), 2024. https://arxiv.org/abs/2309.17400.

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:2110.14168, 2021. https://arxiv.org/abs/2110.14168.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement Learning for Fine-Tuning Text-to-Image Diffusion Models. In Advances in Neural Information Processing Systems (NeurIPS), 2023. https://arxiv.org/abs/2305.16381.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202, 2023. https://arxiv.org/abs/2210.10760.

Andrew Gelman and Hal Stern. The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant. The American Statistician, 60(4):328–331, 2006.

Dhruba Ghosh, Hanna 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 Track, 2023. https://arxiv.org/abs/2310.11513.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2021. https://arxiv.org/abs/2103.03874.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. https://arxiv.org/abs/2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A Comprehensive Benchmark for Open-World Compositional Text-to-Image Generation. In Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2023. https://arxiv.org/abs/2307.06350.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162, 2022. https://arxiv.org/abs/2201.12086.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In Advances in Neural Information Processing Systems (NeurIPS), 2023. https://arxiv.org/abs/2306.09683.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In Proceedings of the 16th International Conference on Machine Learning (ICML), 1999.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models. arXiv:2502.09992, 2025. https://arxiv.org/abs/2502.09992.

Qwen. Qwen2.5 Technical Report. arXiv:2412.15115, 2024. https://arxiv.org/abs/2412.15115.

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. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8748–8763, 2021. https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR; arXiv:2112.10752, 2022. https://arxiv.org/abs/2112.10752.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300, 2024. https://arxiv.org/abs/2402.03300.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Hacking. In Advances in Neural Information Processing Systems (NeurIPS), 2022. https://arxiv.org/abs/2209.13085.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformer Reinforcement Learning. GitHub repository, 2020. https://github.com/huggingface/trl.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In Advances in Neural Information Processing Systems (NeurIPS), 2023. https://arxiv.org/abs/2304.05977.

A EXPERIMENTAL DETAILS AND REPRODUCIBILITY

This appendix documents the experimental configuration and numerical provenance of the main analyses. Stored code, manifests, training traces, calibration summaries, and per-image scores support the trajectory, evaluator-effect, continuous-score, and final-checkpoint analyses. The permutation experiment uses archived performance summaries. Appendix A.5 distinguishes retained records from unavailable artifacts.

A.1 REPRODUCIBILITY BASIS FOR REPORTED RESULTS

Table 4: Reproducibility basis for the reported result families.

Result family	Reproducibility basis
Reward construction and normalization	The stored implementation (TRL DDPOTrainer and the reward function) specifies the CLIP-B/32 score, binary reward, dense ramp with W = 0.03, prompt-specific calibration, 16-reward rolling normalization, and advantage clipping. Buffer degeneracy rates are reconstructed exactly from call-level records.
Training runs and traces	Manifests identify 336 unique runs with 100 logged blocks each, in four complete blocks: campaign A at easy/medium/hard and campaign B at hard.
CLIP trajectory summaries	Passage times, success-rate AUC, and final-block success, with paired prompt-level bootstrap intervals, are recomputed from the traces for both campaigns and all difficulties.
Evaluator trajectories and interaction	Campaign-B success trajectories, evaluator-specific effects, both contrasts, prompt-wise signs, max-t intervals, sign-flip tests, and horizon sensitivity are recomputed; the max-t intervals reproduce the reported values to four decimal places.
Continuous evaluator scores	Raw and standardized dense-minus-binary effects are recomputed from call-level score means and calibration standard deviations.
Final-checkpoint scores	Per-image CLIP-B/32 and CLIP-L/14 scores for 252 $\times$ 64 campaign-A images and 14 $\times$ 64 base-policy images; per-image CLIP, OWLv2, and BLIP-VQA scores for the 784-image campaign-B archive.
Ramp coverage	Computed from 64 base-policy images per prompt at evaluation seed 99999, not from the 96-image calibration set, whose per-image scores were not retained.
Reward permutation	Summary-level performance trend only; interpretation in Appendix C.
Text experiments	Stored evaluation summaries (Appendices D and E).

A.2 EXPERIMENTAL CONFIGURATION

The stored implementation specifies the reward score and shaping rule used in the image experiments. The input score is CLIP-B/32 image–text cosine similarity mapped to [0,1]. The sparse reward is $\mathbf{1}[s \geq \tau]$, and the dense reward uses a linear ramp of width $W = 0.03$ below the same prompt-specific threshold. Prompt-specific calibration thresholds are estimated from 96 base-policy images per prompt. Reward normalization uses a prompt-specific rolling buffer of 16 rewards whose statistics are applied to the 64 rewards generated in an epoch. The 60%, 30%, and 10% calibration rates are nominal targets; achieved rates and threshold-identifiability flags are separate fields in the calibration records.

Table 5 lists the 14 prompts with the OWLv2 queries, count and relation specifications, and BLIP-VQA questions used by the external evaluators (Section 3.3). Count targets refer to the named object; relations are evaluated between the centers of the two named objects.

Table 5: Prompts and external-evaluator specifications. Each prompt is trained in a separate process; p01 is excluded from the cross-evaluator comparison because its OWLv2 hard threshold is not identifiable.

ID	Prompt	OWLv2 queries	Count / relation	BLIP-VQA question
p01	a blue cube stacked on top of a red sphere	a blue cube; a red sphere	blue cube above red sphere	Is a blue cube stacked on top of a red sphere?
p02	four red apples arranged in a square on a plate	a red apple; a plate	count: red apple = 4	Are there four red apples on a plate?

p03	a green backpack to the left of a yellow umbrella	a green backpack; a yellow umbrella	green backpack left of yellow umbrella	Is a green backpack to the left of a yellow umbrella?
p04	three white candles next to a black book	a white candle; a black book	count: white candle = 3	Are there three white candles next to a black book?
p05	a small dog sitting under a large wooden chair	a dog; a wooden chair	dog below wooden chair	Is a dog sitting under a wooden chair?
p06	a red car and a blue bicycle on a street	a red car; a blue bicycle	—	Is there a red car and a blue bicycle on a street?
p07	two cats and one dog on a sofa	a cat; a dog; a sofa	count: cat = 2, dog = 1	Are there two cats and one dog on a sofa?
p08	a glass of orange juice on top of a stack of books	a glass of orange juice; a stack of books	glass of orange juice above stack of books	Is a glass of orange juice on top of a stack of books?
p09	a purple flower in a white vase on a windowsill	a purple flower; a white vase	—	Is a purple flower in a white vase on a windowsill?
p10	a silver fork to the right of a ceramic plate	a silver fork; a ceramic plate	silver fork right of ceramic plate	Is a silver fork to the right of a ceramic plate?
p11	five yellow rubber ducks in a row	a yellow rubber duck	count: yellow rubber duck = 5	Are there five yellow rubber ducks in a row?
p12	a brown teddy bear holding a red balloon	a teddy bear; a red balloon	—	Is a brown teddy bear holding a red balloon?
p13	a wooden table with a laptop and a coffee mug	a laptop; a coffee mug; a wooden table	—	Is there a laptop and a coffee mug on a wooden table?
p14	a black cat on top of a white refrigerator	a black cat; a white refrigerator	black cat above white refrigerator	Is a black cat on top of a white refrigerator?

OWLv2 (owlv2-base-patch16-ensemble) is queried with the objects named in each prompt (Table 5). Its score is the mean, over required objects, of the highest detection confidence above 0.10. For prompts that specify counts, this is multiplied by the mean over counted objects of max(0, 1 −ln − n*|/n*), where n is the number of detections and n* the target count; for prompts that specify a spatial relation, it is multiplied by clip(0.5 + 2m, 0, 1), where m is the signed normalized displacement between the two object centers, and it is zero if either object is undetected. The score is therefore zero-inflated for several prompts. The BLIP-VQA score (blip-vqa-base) is P(yes)/(P(yes) + P(no)) for the first generated token in response to a single prompt-specific yes/no question. During training, OWLv2 and BLIP-VQA score the same rollout images as CLIP but do not enter the reward.

Normalization buffer. The 16-reward buffer consists of the last two reward calls of each epoch, so its degenerate states can be reconstructed exactly from call-level records. All-zero buffers occur only under the binary reward: in 8.1% of the first 20 epochs and 1.8% of all epochs in campaign B (7.9% and 1.8% in the campaign-A hard setting). All-one buffers occur in 15.8% of dense and 12.7% of binary campaign-B epochs (18.0% and 13.2% in campaign A, hard), mostly in the second half of training. Averaged over epochs, the 64 clipped binary-condition advantages sum to −0.67.

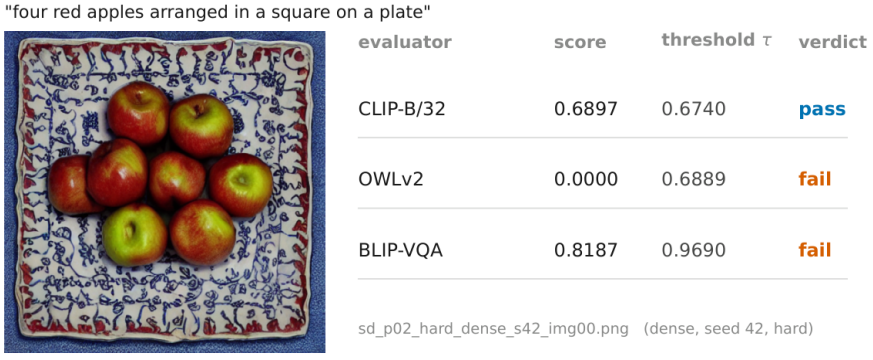


Figure 5: Illustrative evaluator disagreement on one campaign-B hard-setting final-checkpoint image. Selection excludes p01, requires scores and identifiable thresholds for all three evaluators, sorts by

prompt, condition, seed, and image index, and selects the first disagreement. For "four red apples arranged in a square on a plate," CLIP passes the image while OWLv2 and BLIP-VQA do not. The example is conditional on disagreement, not a representative sample. For p02, the count multiplier $\max(0, 1 - |n - 4|/4)$ equals zero when $n = 0$ or $n \geq 8$; the final score alone does not identify the detected count. The BLIP-VQA question is "Are there four red apples on a plate?"

A.3 Evaluator interaction aggregation and inference

The evaluator-interaction analysis uses campaign B in the hard setting. For each prompt and evaluator, success-rate AUC is computed as the unweighted mean over the logged blocks t = 0, ..., 99 for each seed, and the three seeds (42, 1337, and 2026) are then averaged arithmetically. Prompt-level dense-minus-binary effects are equally weighted over the common 13-prompt set, excluding p01. The evaluator-specific dense-minus-binary effects are 0.0511 for CLIP, -0.0101 for OWLv2, and -0.0006 for BLIP-VQA. The resulting CLIP-minus-external differences are 0.0612 for CLIP–OWLv2 and 0.0517 for CLIP–BLIP-VQA.

For simultaneous inference, prompts are resampled as paired units, keeping all conditions and evaluator summaries together. We draw 10,000 bootstrap samples with NumPy's `default_rng(0)` and use the same prompt indices for both contrasts. Let $\widehat{I}_k$ and $I_k^{(b)}$ denote the observed and resampled contrasts. We compute $\widehat{SE}_k$ as the sample standard deviation (`ddof=1`) across bootstrap contrasts, then set

$$M_b = \max_k \frac{\left| I_k^{(b)} - \widehat{I}_k \right|}{\widehat{SE}_k}, \qquad q = Q_{0.95}\left(M_b\right), \qquad CI_k = \widehat{I}_k \pm q\widehat{SE}_k.$$

The common critical value is approximately $q = 2.17$; confidence limits use its unrounded value. The simultaneous 95% intervals are [0.0359, 0.0865] for CLIP–OWLv2 and [0.0293, 0.0741] for CLIP–BLIP-VQA. Other reported image-analysis intervals, including those for individual D_j, use the 2.5th and 97.5th percentiles of paired prompt-level bootstrap effects, each with a fresh default_rng(0). These intervals are individual, not simultaneous across evaluators or horizons. As a sensitivity analysis, sign-flip tests enumerate all 2^{13} or 2^{14} sign assignments under a null of sign exchangeability, using the absolute mean difference as the statistic. Table 3 applies this test to raw-score differences, not standardized differences. Seed-level intervals in Appendices D and E are Student-t intervals for paired seed differences, with $n_{\text{seed}} - 1$ degrees of freedom.

A.4 Stored traces and final-checkpoint assets

The 336 training traces form four complete blocks: campaign A at easy, medium, and hard, and campaign B at hard only. Each block contains 14 prompts $\times$ 2 reward conditions $\times$ 3 seeds = 84 unique runs. Every trace contains 800 reward calls, corresponding to 8 calls per epoch for 100 epochs, with 8 images per call.

Table 6: Stored image-experiment assets. Campaigns A and B are separate training executions.

Asset	Coverage	Stored information
Training traces	A: easy/med/hard; B: hard	336 unique runs; call-level sample counts, reward means, CLIP means and success counts; campaign B additionally stores OWLv2 and BLIP-VQA means and success counts.
64-image final evaluation	Campaign A only	252 cells, each with 64 newly generated images at evaluation seed 99999; stored CLIP-B/32 and CLIP-L/14 scores.
Image-level archive	Campaign B hard + base references	672 final-checkpoint images (8 per prompt–condition–seed cell) plus 112 base-policy images; per-image CLIP, OWLv2, and BLIP-VQA scores.

The campaign-B eight-image cells are not subsets of campaign-A 64-image evaluations even when prompt, condition, and nominal seed match, because the two campaigns use separately trained adapters. The common evaluation seed 99999 specifies the random generator initialization for evaluation but does not imply cross-campaign image identity. Training traces retain call-level means and success counts rather than individual intermediate-rollout scores, so analyses that require arbitrary threshold sweeps over the training trajectory are not recoverable from the saved traces.

A.5 Reproducibility artifacts

The reconstruction uses explicit manifests for the 336 training traces, 252 campaign-A final-evaluation cells, and 784 image-level archive entries, together with prompt-level evaluator effects and the interaction-analysis output. These records separate campaign identity, difficulty, prompt, reward condition, seed, checkpoint, and evaluation seed, and prevent same-named cells from different campaigns from being treated as the same experimental unit.

The original interaction-analysis script was not preserved. We reconstructed the aggregation rule and simultaneous-inference procedure from the stored traces (Section 3.5). With NumPy default_rng(0) and 10,000 resamples, the reconstruction recovers the five evaluator-effect point estimates to within 3.3×10^{-5} and reproduces the simultaneous intervals to four decimal places.

The campaign-A CLIP AUC is recomputed on all 14 prompts; the +0.0591 value uses the common 13-prompt subset (Appendix B.4). All intervals in Table 2 and Appendix Table 11 use the same bootstrap convention.

For the ranking diagnostic in Appendix D.5, tie-corrected AUC point estimates are recovered from stored reward and Monte-Carlo arrays. Problem identifiers are absent from those arrays, and no stored execution record reproduces the legacy problem-cluster confidence intervals. Saved intervals use rollout-level independent resampling. We therefore report the corrected point estimates descriptively, without substituting rollout-level intervals for problem-cluster intervals.

The permutation performance trend is retained from the experimental record. Group-level reward vectors and the detailed intervention order are not available for rechecking the executed reward-multiset invariant; the conditional identity in Appendix C describes its mathematical assumptions.

Per-image calibration scores were not retained for any evaluator, and the CLIP calibration run did not record its achieved rate or identifiability flag. CLIP calibration and OWLv2/BLIP-VQA calibration were run separately with the same generation recipe and 96 images; whether the two runs used pixel-identical images cannot be verified from the stored records. Training-time advantage statistics, clip fractions, initial LoRA weights, and individual rewards within a reward call were not stored.

B Additional evaluator analyses

B.1 Sensitivity to seed and prompt aggregation

The main evaluator-interaction analysis averages seeds within each prompt and then averages prompts. Table 7 shows how the point estimates change under median-based alternatives while holding campaign B, hard difficulty, the 100-block horizon, and the common 13-prompt set fixed.

Table 7: Sensitivity of evaluator effects to seed and prompt aggregation. The first row is the main analysis.

Seed agg.	**Prompt agg.**	D_{CLIP}	D_{OWLv2}	$D_{\mathrm{BLIP\text{-}VQA}}$	I_{OWLv2}	$I_{\mathrm{BLIP\text{-}VQA}}$
Mean	Mean	.0511	−.0101	−.0006	+.0612	+.0517
Mean	Median	.0384	−.0047	+.0087	+.0431	+.0297
Median	Mean	.0493	−.0143	−.0021	+.0636	+.0514
Median	Median	.0328	−.0094	+.0028	+.0422	+.0300

Across the four aggregation rules, I_{OWLv2} ranges from .0422 to .0636 and $I_{\mathrm{BLIP\text{-}VQA}}$ ranges from .0297 to .0517. All eight reported contrast values are positive.

B.2 Prompt-level decomposition of cross-evaluator effects

Table 8 decomposes the cross-evaluator interaction by prompt. Each entry is the difference between the dense-minus-binary success-rate AUC effect measured by CLIP and the corresponding effect measured by an external evaluator. The CLIP–OWLv2 contrast is positive for all 13 prompts, while the CLIP–BLIP-VQA contrast is positive for 12 of 13. The sign counts are 13/13 for CLIP–OWLv2 and 12/13 for CLIP–BLIP-VQA; these counts describe direction only and do not measure the leverage of individual prompts on the aggregate effect.

Table 8: Prompt-level cross-evaluator effect differences on the common 13-prompt set. Each entry is the CLIP dense-minus-binary success-rate AUC effect minus the corresponding effect under the external evaluator.

Prompt	CLIP−OWLv2	CLIP−BLIP-VQA
p02	.0514	.0483
p03	.0890	.0539
p04	.1720	.1246
p05	.0276	−.0074
p06	.0418	.0983
p07	.0891	.0894
p08	.0163	.0049
p09	.1034	.0507
p10	.0499	.0197
p11	.0230	.0102
p12	.0738	.0784
p13	.0348	.0454
p14	.0236	.0560

B.3 Continuous-score sensitivity to the prompt set

Continuous evaluator scores do not require a calibrated success threshold, so we test whether inclusion of p01 changes the continuous-score results. Using campaign B in the hard setting and the same epoch, seed, and prompt aggregation as the main analysis, we repeat the comparison on both the common 13-prompt set and all 14 prompts.

Table 9: Continuous-score sensitivity to the prompt set. $\Delta$is dense minus binary. Confidence intervals are paired prompt-level percentile bootstrap intervals with 10,000 resamples; p is the exact sign-flip p-value.

Evaluator	Prompt set	Dense	Binary	Δ	95% CI	p
OWLv2	Common 13	.3763	.3848	−.0085	[−.0210, +.0022]	.203
OWLv2	All 14	.3512	.3593	−.0081	[−.0195, +.0019]	.187
BLIP-VQA	Common 13	.7570	.7494	+.0076	[−.0124, +.0269]	.485
BLIP-VQA	All 14	.7295	.7241	+.0054	[−.0132, +.0241]	.600

Adding p01 changes the OWLv2 effect from −.0085 (common 13) to −.0081 (all 14) and the BLIP-VQA effect from +.0076 to +.0054. The 95% confidence intervals include zero for all four prompt-set/evaluator combinations. For p01 alone, the seed-averaged dense-minus-binary differences are −.0033 for OWLv2 and −.0230 for BLIP-VQA.

B.4 Campaign-A CLIP AUC and prompt-set membership

The campaign-A hard-setting CLIP analysis uses all 14 prompts. We average the 100 epoch-level success rates within each run, then the three seeds within each prompt, and finally prompts with equal weight. Table 10 reconciles the all-14 analysis with the common-13 subset used for cross-evaluator comparisons. Both use the same traces, evaluator, horizon, and averaging rules. The exclusion of p01 concerns its OWLv2 hard threshold and is not required for the CLIP-only analysis.

Table 10: Campaign-A hard-setting CLIP success-rate AUC. Means and differences are shown to six decimal places. Intervals are paired prompt-level percentile bootstrap intervals with 10,000 resamples and RNG seed 0.

Prompt set	n	Dense	Binary	Δ	95% CI
All 14	14	.678891	.611257	+.067634	[+.038701, +.096735]
Common 13	13	.685950	.626835	+.059115	[+.032929, +.085012]

The all-14 values round to .679, .611, and +.068 as reported in the main text. The common-13 difference rounds to +.0591. Prompt p01 contributes a positive difference of +.178385; excluding it lowers the mean effect while increasing both condition means. Positive differences occur on 13/14 prompts in the full set and 12/13 in the reduced set.

B.5 Sensitivity to the trajectory horizon

The success-rate AUC averages over the first H logged blocks. Figure 6 reports effects and contrasts at $H = 5, 10, \ldots, 100$ using the main aggregation rule. D_{CLIP} reaches +.0716 at $H = 55$ and then

decreases as both CLIP trajectories approach their final levels; D_{OWLv2} and $D_{\text{BLIP-VQA}}$ remain between $-.0102$ and $+.0061$. Both contrasts are positive at all evaluated horizons. Their pointwise 95% percentile intervals exclude zero at every evaluated $H \geq 10$; these are not simultaneous bands over horizons.

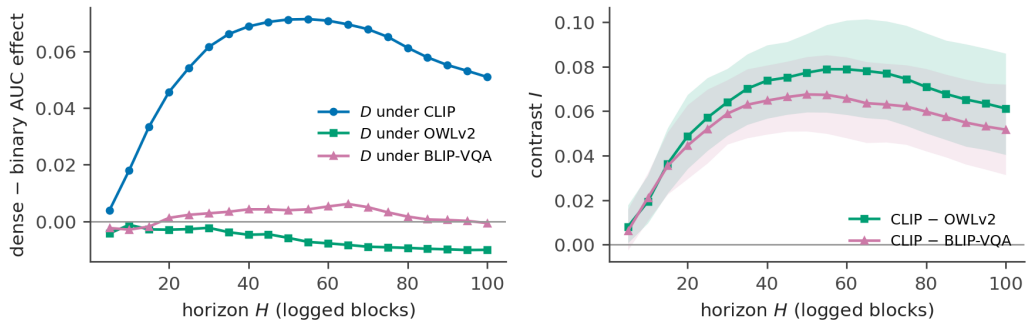


Figure 6: Horizon sensitivity (campaign B, hard, common 13 prompts). Left: dense-minus-binary success-rate AUC effect over the first H logged blocks under each evaluator. Right: CLIP-minus-external contrasts with pointwise 95% paired prompt-level percentile bootstrap intervals (10,000 resamples).

B.6 Campaign-A CLIP summaries by difficulty

Table 11 reports the campaign-A CLIP summaries for all three difficulties. The dense advantage in passage time and AUC is largest in the hard setting and smaller in the medium setting. In the easy setting, where the base-policy success rate already exceeds the 0.5 target, first and smoothed passage occur at or near $t = 0$ for nearly all runs, and the final-block success rate is lower under dense reward ($-.021$ [$-.044$, $-.003$]).

Table 11: Campaign-A CLIP trajectory summaries by difficulty, all 14 prompts. Δ= dense minus binary; intervals are paired prompt-level percentile bootstrap intervals (10,000 resamples, RNG seed 0).

Difficulty	Metric	Dense	Binary	Δ	95% CI
Easy	First passage	0.10	0.10	+0.00	[+0.00, +0.00]
Easy	Sustained, 3 epochs	0.24	0.57	−0.33	[−0.88, +0.05]
Easy	Smoothed, 5 epochs	0.00	0.00	+0.00	[+0.00, +0.00]
Easy	Success-rate AUC	.912	.918	−.006	[−.013, +.000]
Easy	Final block (t = 99)	.966	.987	−.021	[−.044, −.003]
Medium	First passage	7.90	9.26	−1.36	[−3.60, +0.57]
Medium	Sustained, 3 epochs	10.95	13.52	−2.57	[−4.57, −0.95]
Medium	Smoothed, 5 epochs	8.36	10.88	−2.52	[−4.62, −1.00]
Medium	Success-rate AUC	.809	.799	+.010	[−.003, +.025]
Medium	Final block (t = 99)	.937	.961	−.024	[−.070, +.006]
Hard	First passage	23.14	31.60	−8.45	[−12.81, −4.76]
Hard	Sustained, 3 epochs	27.90	36.43	−8.52	[−12.02, −5.43]
Hard	Smoothed, 5 epochs	24.95	33.26	−8.31	[−12.38, −5.00]
Hard	Success-rate AUC	.679	.611	+.068	[+.039, +.097]
Hard	Final block (t = 99)	.916	.891	+.025	[−.008, +.060]

B.7 Final-checkpoint thresholds and a same-family evaluator

Campaign A stores 64 newly generated final-checkpoint images per cell and 64 base-policy images per prompt, each scored by CLIP-B/32 and CLIP-L/14. For each prompt, thresholds are set at base-policy quantiles and dense-minus-binary success differences are computed (Table 12). Every displayed interval includes zero. Standardized mean-score differences are +0.07 [−0.03, +0.18] under CLIP-B/32 and +0.05 [−0.11, +0.21] under CLIP-L/14, using the sample standard deviation of the 64 base-policy scores within each prompt. Separately, pooling both reward conditions' final-checkpoint scores within each prompt gives mean increases over the base-policy mean of 2.41 and 1.46 base-policy standard deviations, respectively. These are all-14-prompt endpoint comparisons in campaign A, not trajectory-level speed-up estimates.

Table 12: Campaign-A hard-setting final-checkpoint success rates at thresholds set by each prompt's base-policy quantiles (64 base-policy images per prompt), all 14 prompts. The first row uses the training threshold τ. Intervals are paired prompt-level percentile bootstrap intervals.

Evaluator	Threshold	Dense	Binary	Δ	95% CI
CLIP-B/32	Training τ	.921	.904	+.0167	[−.0063, +.0406]
CLIP-B/32	Base 50th pct.	.990	.985	+.0048	[−.0004, +.0100]
CLIP-B/32	Base 70th pct.	.978	.966	+.0119	[.0000, +.0246]
CLIP-B/32	Base 90th pct.	.903	.884	+.0186	[−.0041, +.0406]
CLIP-B/32	Base 95th pct.	.856	.837	+.0190	[−.0164, +.0513]
CLIP-B/32	Base 99th pct.	.741	.708	+.0331	[−.0089, +.0740]
CLIP-L/14	Base 50th pct.	.898	.876	+.0223	[−.0138, +.0599]
CLIP-L/14	Base 70th pct.	.794	.769	+.0246	[−.0201, +.0696]
CLIP-L/14	Base 90th pct.	.579	.575	+.0041	[−.0487, +.0640]
CLIP-L/14	Base 95th pct.	.480	.469	+.0108	[−.0484, +.0770]
CLIP-L/14	Base 99th pct.	.298	.298	.0000	[−.0625, +.0636]

C Reward permutation (sensitivity analysis)

The permutation experiment changes the assignment of reward values to samples while leaving the within-group reward-value multiset unchanged by design. The recorded performance summaries decrease from 0.921 in the aligned condition to 0.223 under complete permutation, with intermediate reported conditions ordered between these endpoints.

We report the permutation experiment as a sensitivity analysis rather than an exact negative control. The executed DDPO implementation normalizes 64 rewards with statistics from a rolling 16-reward buffer, so the 64 normalized advantages are not guaranteed to sum to zero. Complete permutation also changes which samples receive success-associated and partial-credit values. The archived summaries therefore do not identify a unique credit-assignment mechanism. The derivation below states the fixed-vector identity and the conditions under which its expected linear update is zero.

The permutation experiment asks whether learning depends on the correspondence between samples and their reward values. The intervention is designed to shuffle reward values across samples while preserving their multiset, including repeated values. If the normalized advantages had zero mean over the full group and the sample-dependent update vectors were fixed, complete random permutation would give a zero expected linear update. The executed implementation does not generally satisfy the first condition, because 64 rewards are normalized using statistics from a rolling buffer of 16 rewards. Formally, let $A = (A_1, \ldots, A_n)$ be a fixed advantage vector, let h_i denote fixed sample-dependent vectors, and let π be a uniformly random permutation. Then each position receives every advantage value with equal probability, giving

$$\mathbb{E}_\pi \left[A_{\pi(i)} \right] = \frac{1}{n} \sum_{k=1}^{n} A_k = \overline{A}. \tag{4}$$

Therefore,

$$\mathbb{E}_\pi \left[\sum_{i=1}^{n} h_i A_{\pi(i)} \right] = \overline{A} \sum_{i=1}^{n} h_i. \tag{5}$$

The expected linear update is zero when $\overline{A} = 0$. In the executed DDPO implementation, however, normalization uses statistics from a rolling 16-reward buffer rather than the full 64-sample group, so this zero-mean condition is not guaranteed for the complete advantage vector. If normalization is recomputed after permutation, the normalized advantages can themselves depend on π; repeated policy updates can likewise change the h_i. Thus, the identity applies to the fixed-vector conditional case and motivates the permutation analysis, while the executed experiment is interpreted as a sensitivity test rather than an exact zero-update control. The fixed-vector identity addresses the linear sum above; clipping, additional losses, and nonlinear optimizer transformations require their own assumptions.

D AUTOREGRESSIVE RLVR EXPERIMENTS

These supplementary experiments have no evaluator that is external to the reward and therefore do not test evaluator dependence; they are reported for completeness. The autoregressive experiments use a 1.5B Qwen2.5 model on MATH and GSM8K (Qwen, 2024; Hendrycks et al., 2021; Cobbe et al., 2021) and compare several graded reward formulations with binary-reward baselines. Because task correctness is shared across these reward formulations, we use accuracy as the common outcome for comparing dense-minus-sparse effects across reward families and sparsity levels.

D.1 DENSE REWARD FAMILIES ON TASK ACCURACY

Table 13: Dense-minus-sparse task-accuracy effects for the four graded-reward families. Each comparison uses three seeds and 500 evaluation items. Effects are proportions; intervals are 95% Student-t intervals for paired seed differences. Arm labels follow the supplied evaluation records.

Level	Dense reward	Effect	95% CI
L3	Heuristic partial	$-.0107$	$[-.0314, +.0100]$
L3	Numeric proximity	$+.0113$	$[-.0012, +.0238]$
L3	Verified steps	$-.0020$	$[-.0283, +.0243]$
L3	Oracle-MC	$+.0060$	$[-.0347, +.0467]$
L5	Heuristic partial	$+.0040$	$[-.0091, +.0171]$
L5	Numeric proximity	$-.0067$	$[-.0192, +.0058]$
L5	Verified steps	$+.0080$	$[-.0137, +.0297]$
L5	Oracle-MC	$+.0007$	$[-.0267, +.0280]$

Across the eight reward-family comparisons, dense-minus-sparse accuracy effects range from $-.0107$ to +.0113. All reported 95% confidence intervals include zero; five point estimates are positive and three are negative.

D.2 OVERLAP-AWARE TASK-ACCURACY ANALYSIS

Because the MATH training and evaluation questions are drawn from the same `competition_math` training split, we partition evaluation items according to whether they appear in each run's RL fine-tuning set. The filtered question pools contain 2,723 items at L3 and 3,628 at L5, with 2,000 items used for RL fine-tuning in each run. The 500-item protocol has overlap rates of 69.4–76.2% at L3 and 50.2–56.6% at L5; the nested 150-item protocol has rates of 68.7–76.0% and 51.3–63.3%, respectively. These reported ranges span the eight nominal runs, whereas the reward-family comparisons use three seeds.

Table 14: Treatment-minus-baseline accuracy effects by RL training-set membership. N is the full evaluation-set size, and S is the seed count; effects are proportions. Non-overlap means absent from that run's RL fine-tuning set. Different protocols are not pooled.

Level	Treatment	Baseline	S	N	All	Overlap	Non-overlap
L3	DAPO-DS	Sparse-big	8	150	$+.0008$	$-.0069$	$+.0180$
L3	Heuristic partial	Sparse	3	500	$-.0107$	$-.0091$	$-.0151$
L3	Numeric proximity	Sparse	3	500	$+.0113$	$+.0118$	$+.0098$
L3	Verified steps	Sparse	3	500	$-.0020$	$-.0027$	$-.0003$
L3	Oracle-MC	Sparse	3	500	$+.0060$	$+.0046$	$+.0098$
L5	DAPO-DS	Sparse-big	8	150	$-.0050$	$-.0053$	$-.0041$
L5	Heuristic partial	Sparse	3	500	$+.0040$	$-.0025$	$+.0114$
L5	Numeric proximity	Sparse	3	500	$-.0067$	$-.0087$	$-.0043$
L5	Verified steps	Sparse	3	500	$+.0080$	$+.0048$	$+.0116$
L5	Oracle-MC	Sparse	3	500	$+.0007$	$-.0025$	$+.0041$

Table 14 reports treatment-minus-baseline accuracy effects on all, overlap, and non-overlap items. The four graded-reward families use three seeds and 500 evaluation items; the DAPO-DS rows use eight seeds and 150 items. The overlap and non-overlap effects have opposite signs for L3 DAPO-DS ($-.0069$ vs. +.0180), L5 Heuristic partial ($-.0025$ vs. +.0114), and L5 Oracle-MC ($-.0025$ vs. +.0041). The remaining seven rows retain the same sign between overlap and non-overlap subsets. These subgroup values are descriptive because the protocols and subgroup sample sizes differ.

D.3 SPARSITY LADDER AND GENERATION-BUDGET CONTROL

We use an eight-seed sparsity ladder to examine whether the relative effect of capped dynamic filtering (CDF) increases as all-failure groups become more common. The all-failure rate rises from 0.572 at L3 to 0.814 at L5, while the CDF-minus-sparse-big accuracy differences are +0.08, +0.08, and −0.50 percentage points at L3, L4, and L5, respectively. The confidence intervals include zero at all three levels.

Table 15: CDF-minus-sparse-big accuracy differences over eight seeds and 150 evaluation items. Intervals are 95% Student-t intervals for paired seed differences. p_{AF} is the all-failure group rate at group size 8. L3 is not generation-matched.

Level	p_{AF}	**Difference (pp)**	**95% CI (pp)**
L3	.572	+.08	[−1.39, +1.56]
L4	.675	+.08	[−1.56, +1.73]
L5	.814	−.50	[−1.76, +.76]

Generation counts are matched at L4 and L5, where both conditions generate eight candidate groups per step. The CDF-minus-sparse-big accuracy difference is +0.08 percentage points at L4 and −0.50 percentage points at L5, while p_{AF} increases from .675 to .814. At L3, sparse-big generates 9.09 candidate groups per step versus 8.00 for CDF, so the L3 comparison also differs in generation budget.

D.4 GROUP UTILIZATION UNDER CAPPED DYNAMIC FILTERING

Capped dynamic filtering (CDF) selects informative groups from a fixed set of generated candidates, without generating replacement groups after filtering.

Table 16: Per-step group counts and zero reward-gradient rates for the no-refill CDF implementation. Zero reward-gradient refers to the recorded reward-gradient diagnostic, not the entire optimizer update.

Level	**Condition**	**Candidate**	**Informative**	**In loss**	**Zero gradient**
L3	CDF	8.00	4.26	4.26	0.4%
L3	Sparse-big	9.09	5.06	9.09	0.5%
L4	CDF	8.00	3.46	3.46	0.9%
L4	Sparse-big	8.00	3.45	8.00	1.3%
L5	CDF	8.00	1.71	1.71	15.2%
L5	Sparse-big	8.00	1.71	8.00	14.1%

Table 16 reports candidate groups, groups with reward variation, groups entering the loss, and the recorded zero-gradient rate. At L4, CDF and sparse-big generate 8.00 candidate groups and contain 3.46 and 3.45 informative groups per step, respectively; CDF sends 3.46 groups to the loss and sparse-big sends 8.00. At L5, both generate 8.00 candidates and contain 1.71 informative groups; CDF sends 1.71 groups to the loss and sparse-big sends 8.00. The recorded zero-gradient rates are 0.9% versus 1.3% at L4 and 15.2% versus 14.1% at L5. At L3, sparse-big generates 9.09 candidate groups per step versus 8.00 for CDF, so the L3 comparison also differs in generation budget.

D.5 RANKING SIGNAL AMONG UNSUCCESSFUL ROLLOUTS

We evaluate whether graded reward components distinguish unsuccessful rollouts by observed Monte-Carlo completability. A single base-policy diagnostic samples eight rollouts for each of 50 problems per difficulty, yielding 314 failed rollouts at L3 and 370 at L5. For each failure, the prefix contains the first half of its non-empty reasoning lines (rounded down, with at least one line). We sample six continuations from the base policy. A failure is labeled *observed-positive* if at least one continuation is correct and *observed-zero* otherwise. This produces 98 observed-positive cases at L3 and 61 at L5. The observed-zero label records six unsuccessful samples, not a known zero underlying success probability.

For each difficulty and reward component, we pool failures across the 50 problems and calculate tie-corrected AUC,

$$AUC = \frac{1}{n_{\text{pos}} n_{\text{neg}}} \sum_{i:y_i=1} \sum_{j:y_j=0} \left(\mathbf{1}\{r_i > r_j\} + \frac{1}{2}\mathbf{1}\{r_i = r_j\} \right).$$

Here n_{pos} and n_{neg} are the observed-positive and observed-zero class sizes. This is the mid-rank convention: cross-class ties receive half credit, and 0.5 is the chance-ordering reference. The statistic is a level-pooled comparison, not an average of within-problem AUCs. Table 17 reports the corrected point estimates.

Table 17: Tie-corrected, level-pooled ranking AUC on failed rollouts. s_f is the fraction of checkable arithmetic steps that pass verification, and p_x is numeric-answer proximity. These are three scored components, not the four training arms in Appendix D.1. Values are descriptive point estimates; cluster-interval provenance is recorded in Appendix A.5.

Scored component	**Score**	**L3**	**L5**
Verified-step + proximity	$0.4s_f + 0.3p_x$	.4953	.5224
Numeric proximity	$0.7p_x$	.4774	.4672
Verified-step component	$0.7s_f$	.5100	.5466

Across the six comparisons, AUC ranges from .4672 to .5466, corresponding to deviations from the 0.5 chance-ordering reference of $-.0328$ to +.0466. These values describe association between each scored component and the observed continuation-success label under the pooled diagnostic.

E DISCRETE-DIFFUSION EXPERIMENTS

The supplementary LLaDA-8B experiments (Nie et al., 2025) compare symbolic dense and sparse rewards at medium and hard difficulty. Task accuracy is evaluated at checkpoints 100 and 150 over three seeds on 260 items that coincide with the experimental RL training set. The dense-minus-sparse differences are $-.0065$ and $-.0194$ in the medium setting and +.0090 and $-.0026$ in the hard setting, respectively.

Table 18: LLaDA dense-minus-sparse task-accuracy differences. Intervals are 95% Student-t intervals from three paired seed differences. Evaluation uses 260 items from the experimental training set. Differences are calculated before display rounding.

Difficulty	**Checkpoint**	**Dense**	**Sparse**	**Difference**	**95% CI**
Medium	100	.3689	.3754	$-.0065$	$[-.0802, +.0672]$
Medium	150	.3657	.3851	$-.0194$	$[-.1776, +.1387]$
Hard	100	.1564	.1474	$+.0090$	$[-.0246, +.0425]$
Hard	150	.1526	.1551	$-.0026$	$[-.0225, +.0173]$

The four dense-minus-sparse point estimates range from $-.0194$ to +.0090, and all four 95% confidence intervals include zero. These values are in-distribution task-accuracy comparisons because the 260 evaluation items are the experimental RL training items.