FACTORGRAPH-RLVR: DIAGNOSING REWARD INFORMATIVENESS IN FACTORIZED VISUAL REASONING CURRICULA

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has improved multimodal reasoning, but gains attributed to curriculum design may instead reflect how often training produces informative group-relative updates. We introduce FACTORGRAPH-RLVR, a controlled visual-graph testbed that crosses visual masking and reasoning load in a matched 3×3 factorial design, together with a parser-gated dense verifier. To separate curriculum composition from optimization opportunity, we compare reasoning-heavy and uniform curricula under a fixed optimizer-step budget, match checkpoints by cumulative informative updates, and ablate dense supervision with exact-only rewards. A preregistered evaluation covers 28 frozen model checkpoints, three final evaluation splits, and four decoding conditions. Under greedy decoding, reasoning-heavy dense RLVR improves exact success over the unadapted model on all three splits. The gains are 4.1, 4.2, and 4.7 percentage points on in-distribution, style-shift, and length-shift data, respectively. All 36 seed–split–decoding contrasts are positive, and every 95% bootstrap confidence interval computed from paired source groups excludes zero. Uniform training yields smaller gains, informative-update matching retains most of the reasoning-heavy effect, and exact-only training remains near the base model. These results provide controlled evidence that curriculum composition influences visual RLVR beyond optimizer-step budget and measured informative-update opportunity.

1 INTRODUCTION

In reinforcement learning with verifiable rewards (RLVR), curriculum design changes which tasks a vision-language model encounters and can also change how often an optimizer step carries a reward-ranking signal. GRPO compares rewards among responses sampled for the same prompt; when all responses receive the same reward, the group provides no such signal (Shao et al., 2024; Yu et al., 2025). Thus, even at equal training budgets, gains attributed to curriculum design may reflect both task content and the number of informative training steps.

This raises the question of whether curriculum composition still matters when optimizer steps and the number of informative training steps are controlled separately. Prior RLVR methods adaptively filter, select, or reweight examples using difficulty estimates or rollout statistics to find useful training signal (Yu et al., 2025; Bae et al., 2026; Jiang et al., 2025; Leng et al., 2026; Jeddi et al., 2025; Ning et al., 2026). Because these choices can change both task content and reward informativeness, their performance gains alone do not resolve this question.

To study this question, we compare fixed curriculum distributions rather than introducing a new adaptive scheduling algorithm. We introduce FACTORGRAPH-RLVR, a visual-graph testbed in which a model follows typed transitions using two rendered relation tables. Visual masking and path length vary independently in a matched 3×3 design. Each source program yields nine task instances, allowing comparisons across visual and reasoning loads over the same underlying relation tables. A parser-gated verifier supplies partial credit for intermediate transitions, while exact-trace success remains the evaluation endpoint.

Using Qwen3-VL-8B-Instruct with LoRA-based GRPO, we compare four curriculum distributions (visual-heavy, reasoning-heavy, mixed-control, and uniform) at 512 optimizer steps and at checkpoints

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.

approximately matched by cumulative informative training steps. We also compare dense and exact-only rewards under the same mixed-control curriculum, where exact-only training yields little improvement over the base model. Reasoning-heavy dense training achieves the largest mean gains over the base model across three final splits and retains most of those gains after informative-step matching. Its mean advantage over uniform training also persists.

Our contributions are:

- We introduce a source-matched visual-graph testbed that independently varies destination-digit masking and required path length in a 3×3 design, together with a parser-gated dense verifier.
- We provide a diagnostic comparison of four fixed curriculum distributions at a common optimizer-step budget and at checkpoints with approximately matched cumulative informative-step counts.
- We show that reasoning-heavy dense training retains a mean advantage over uniform training after informative-step matching, with variation across training seeds.

2 Related Work

Learnability and reward variance in RLVR. GRPO estimates advantages from within-prompt reward contrasts (Shao et al., 2024). Groups whose rollouts receive homogeneous rewards therefore provide no ranking signal, motivating dynamic filtering (Yu et al., 2025; Bae et al., 2026), variance-based curricula (Jiang et al., 2025), and variance-aware multimodal sampling (Leng et al., 2026). Related analyses identify question-difficulty and response-length biases in group-relative optimization (Liu et al., 2025a). We use logged reward dispersion to identify informative training steps, rather than assuming that nominal difficulty or high dispersion implies better training.

Verifiable multimodal reasoning. Recent work applies programmatic rewards to perception and multimodal reasoning (Liu et al., 2025b; Zhang et al., 2025). Synthetic puzzle environments provide scalable supervision and exact verification (Feng et al., 2025; Alam et al., 2026). Most closely, PuzzleCraft combines graded visual-puzzle rewards with a dynamic curriculum emphasizing medium-difficulty, non-collapsed rollout groups (Jeddi et al., 2025), while Wiki-R1 uses a data and sampling curriculum for knowledge-based visual question answering (Ning et al., 2026). Our goal is complementary: rather than proposing another general adaptive sampler, we hold source programs fixed and independently intervene on visual masking and reasoning load.

Controlled visual diagnostics. Curriculum learning studies how training order and composition affect optimization (Bengio et al., 2009). Programmatically generated benchmarks such as CLEVR isolate compositional visual skills with controlled scene programs (Johnson et al., 2017), while PuzzleVQA diagnoses perception, induction, and deduction on abstract visual patterns (Chia et al., 2024). Our matched factorial construction extends this diagnostic perspective to RLVR: every source program appears at all nine visual–reasoning cells, permitting a paired response-surface analysis of task sensitivity and reward informativeness.

3 FactorGraph-RLVR

3.1 Overview

Figure 1 summarizes the training and evaluation protocol. A deterministic graph program is rendered into matched visual–reasoning cells, and each prompt is expanded into a group of four rollouts. A strict parser first rejects malformed traces; only parsed traces are decomposed into field, step, final-answer, and exact-match signals. Group-relative normalization then converts these rewards into the GRPO update, while training logs record reward dispersion. We use these logs to match checkpoints by measured informative training steps, as defined below.

3.2 Task construction

Each instance contains a BLUE and an ORANGE numeric relation table. A graph program defines two-digit node labels and a deterministic outgoing transition for each color. Given an initial state and

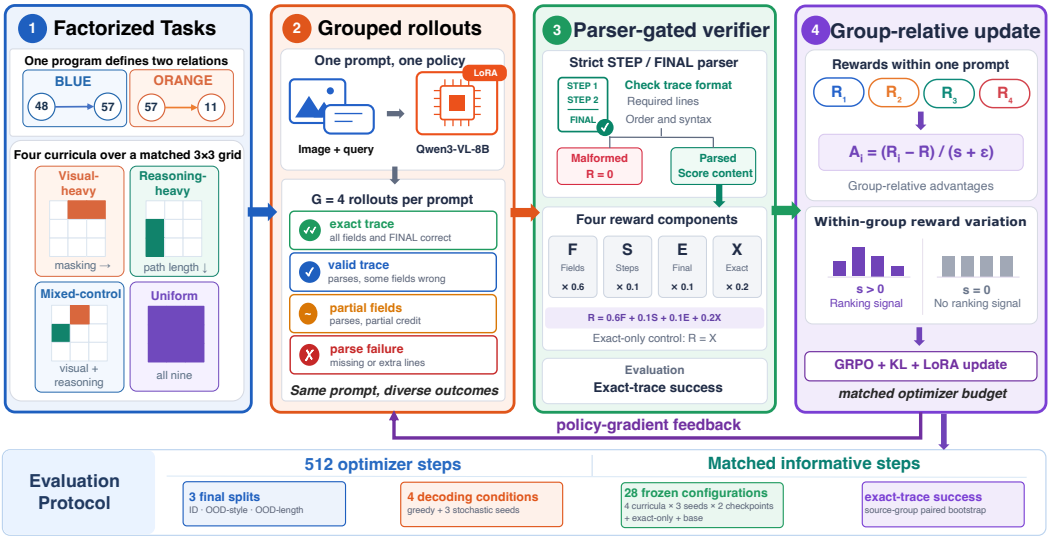

Figure 1: Overview of FACTORGRAPH-RLVR. Matched source programs generate a 3×3 visual–reasoning factorial grid and curriculum pools. Qwen3-VL-8B-Instruct produces four completions per prompt; a strict parser gates decomposed dense rewards, and GRPO normalizes rewards within prompt groups. Here s is the standard deviation within one rollout group; the logged training-step diagnostic uses $\sigma_t > 10^{-8}$ (Section 3.5). Fixed-step and informative-step-matched checkpoints are evaluated with source-group paired bootstrap contrasts.

an ordered relation sequence, the target trace is

$$s_i = T_{r_i}(s_{i-1}), \qquad a_i = (a_{i-1} + s_i) \bmod 100, \quad a_0 = 0, \tag{1}$$

where $r_i \in \{\text{BLUE}, \text{ORANGE}\}$. The answer must contain one line per transition, including the destination and cumulative accumulator. The last line has the form `FINAL: VALUE`, where `VALUE` is the last destination.

The generator varies two independently assigned factors. Visual stress $d_v \in \{0, 1, 2\}$ controls the width of a horizontal mask through every destination digit (0, 4, or 7 pixels), while reasoning stress $d_r \in \{0, 1, 2\}$ requires 2, 3, or 4 transitions. Every source group contains all nine cells, enabling matched comparisons across curricula. Because every additional transition adds a required `STEP` line, d_r also changes output length, the number of scored fields, and the opportunity for a format error. We therefore interpret it as a compound reasoning-load intervention rather than a pure measure of latent reasoning ability, and include exact-only and length-shift controls in the frozen protocol.

3.3 REWARD

Let y be a completion and $y^\star = (t_1^\star, \ldots, t_N^\star, e^\star)$ the verifier-generated target, where each step tuple contains source, relation, destination, and accumulator fields. A completion parses only if it contains exactly N consecutively numbered `STEP` lines and one `FINAL` line, with no additional non-empty text. Missing, extra, duplicate, non-consecutive, or syntactically invalid lines give zero reward. For a parsed completion with predicted tuples t_i and final token e, we define

$$F = \frac{1}{4N} \sum_{i=1}^{N} \sum_{k \in \{\text{src,rel,dst,acc}\}} \mathbf{1}[t_{i,k} = t_{i,k}^\star], \tag{2}$$

$$S = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[t_i = t_i^\star], \qquad E = \mathbf{1}[e = e^\star], \tag{3}$$

$$X = \mathbf{1}[(t_1, \ldots, t_N, e) = (t_1^\star, \ldots, t_N^\star, e^\star)]. \tag{4}$$

The dense reward is

$$R_{\text{dense}}(y) = 0.6F + 0.1S + 0.1E + 0.2X. \tag{5}$$

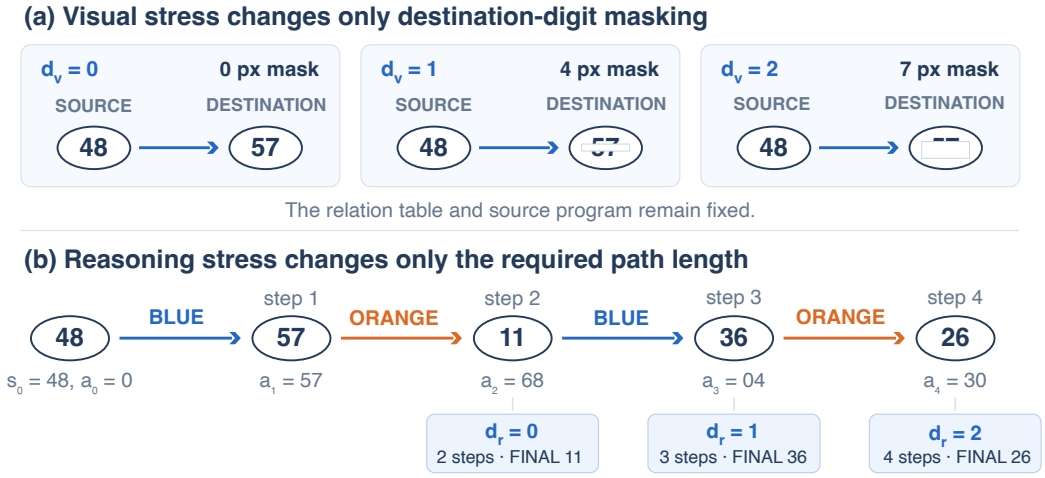

Figure 2: Factorized interventions for one matched source program. Visual stress adds a horizontal mask through every destination digit while keeping the table fixed; reasoning stress extends the required typed-edge path while keeping the image fixed. The complete 3×3 grid is shown in Appendix F.

Thus $R_{\text{dense}} \in [0, 1]$. The sparse baseline uses $R_{\text{exact}} = X$. Both rewards share the parser and exact-trace target; the dense reward additionally credits correct fields and intermediate steps. Evaluation uses X, not the training reward or final-token accuracy alone.

Development runs also included a binary format bonus, $0.1\,\mathbf{1}[y \text{ parses}]$. A three-seed validation audit failed its prespecified retention criterion, so all main runs use dense-only reward without subsequent weight tuning. We report that audit separately from final-test results.

3.4 Curriculum assignments

We compare three targeted assignments alongside a uniform pool, using the same model, optimizer, reward code, and training budget:

- **Visual-heavy**: samples from $(d_v, d_r) \in \{(1, 0), (2, 0)\}$.
- **Reasoning-heavy**: samples from $(d_v, d_r) \in \{(0, 1), (0, 2)\}$.
- **Mixed-control**: pairs one visual and one reasoning cell per source group, alternating stress levels 1 and 2 across source indices (Appendix B.3).

An additional **uniform** pool contains all nine cells. Here curriculum denotes a fixed training distribution, not a progression from easy to hard examples. Uniform training uses the same 512-step budget as the targeted curricula, but pool sizes differ. We match optimizer exposure rather than numbers of distinct examples or epochs.

3.5 Reward informativeness and checkpoint matching

Let σ_t denote the trainer's logged within-group reward standard deviation at optimizer step t. The prespecified diagnostic is

$$I_t = \mathbf{1}[\sigma_t > 10^{-8}], \qquad C(t) = \sum_{u=1}^{t} I_u. \tag{6}$$

Thus $C(t)$ counts informative *logged training steps*, not individual completions or separately enumerated prompt groups. With group reward scaling, the trainer averages the sample standard deviations of the four-rollout prompt groups, then averages collected values at logging. Logging occurs every optimizer step; Appendix A specifies the reduction. Each step contributes at most one to $C(t)$. This diagnostic measures reward dispersion, not gradient magnitude or the causal value of an update.

The target of 96 was prespecified below the minimum informative-step count observed in the earlier visual-heavy development runs. Alongside step 512, we evaluate the saved checkpoint whose $C(t)$ is closest to this target. Checkpoints are saved every 32 steps; ties select the earlier checkpoint. If a run never reaches the target, the rule selects step 512 and labels the target unmet. Selected counts range from 91 to 100, so matching is approximate. Appendix B.1 lists every selected step and count. Selection uses training diagnostics and is fixed before final evaluation.

4 Experimental Setup

Model and optimization. We use Qwen3-VL-8B-Instruct (Bai et al., 2025) with LoRA rank 16 (Hu et al., 2022) and target modules set to all linear layers. GRPO uses four generations per prompt, learning rate 5×10^{-6}, temperature 1.2, top-p 0.95, maximum completion length 240, and 512 optimization steps. The visual encoder is trainable. Each targeted curriculum contains 512 records, while the uniform pool contains 2,304 records balanced across all nine cells. All four arms use the same optimizer-step budget, factorized prompt, and frozen dense-only reward implementation in the main experiment.

Data and distribution shifts. Development evaluation uses 1,152 examples from 128 source groups and informed protocol design. The final sets were generated afterward, checked against training, smoke-test, and development source signatures and generation keys, and checked against each other. Final-ID contains 2,304 examples from 256 source groups; OOD-style and OOD-length each contain 1,152 examples from 128 groups. Every group contributes nine cells, giving 4,608 unique final samples.

Final-ID uses paths of 2, 3, and 4 transitions. OOD-style retains these lengths and masking levels but changes rendering parameters, including color palette, output resolution, blur, and pixel noise. OOD-length retains the ID rendering configuration and changes the three path lengths to 2, 6, and 10. Its aggregate therefore includes a short-path reference level as well as longer paths. Appendix B.2 specifies the configurations. These splits test generalization within the visual-graph generator.

Models and decoding. The final matrix contains the base model, four dense curriculum arms at step 512 and at informative-step-matched checkpoints for seeds 48–50, and three exact-only mixed-control models. These 28 model configurations are evaluated on all three splits. Primary evaluation is greedy decoding. Sensitivity evaluation repeats stochastic decoding with three seeds at temperature 0.7 and top-p 0.8; these are repeated draws, not different hyperparameter settings. Each condition generates one completion per sample with a 240-token limit, yielding 336 jobs.

Endpoints and uncertainty. The primary metric is exact-trace success under greedy decoding. For model–base and direct trained-condition comparisons, 20,000 bootstrap resamples draw source groups with replacement, preserving paired predictions and the nine cells within each group. We report paired 95% intervals within each training seed and mean $\pm$ sample standard deviation across three training seeds. These intervals characterize source-sampling uncertainty conditional on a trained checkpoint, not uncertainty over training randomness.

The primary direct contrast is reasoning-heavy minus uniform at step 512 on final-ID with greedy decoding. The prespecified directional criterion requires a positive mean and positive effects in at least two of three training seeds; it is not a significance test over training seeds. Mixed-control and visual-heavy versus uniform, informative-matched comparisons, OOD splits, and stochastic decodings are secondary analyses. The reward contrast compares dense with exact-only under mixed-control at step 512. All intervals are pointwise, without multiplicity adjustment; we make no family-wise significance claim. Appendix G reports all direct contrasts.

5 Results

5.1 Reasoning-heavy dense RLVR gives the largest confirmatory gains

Table 1 reports the complete primary greedy-decoding comparison. The complete matrix passed integrity validation before aggregation. The unadapted base model obtains 35.8%, 31.8%, and 28.5%

Table 1: Confirmatory exact success (%) under primary greedy decoding. Entries for trained conditions are mean ± sample standard deviation across frozen seeds 48–50; the base model is evaluated once on the same samples. Reported gains use unrounded paired differences, not subtraction of rounded table entries. "Dense informative" uses the checkpoint matched closest to 96 cumulative informative training steps (actual counts 91–100; Appendix B.1).

Checkpoint	Arm	Final ID	OOD style	OOD length
Base	–	35.8	31.8	28.5
Dense step 512	Visual-heavy	36.7 ± 0.6	32.2 ± 0.3	29.4 ± 0.4
Dense step 512	Reasoning-heavy	39.9 ± 1.8	36.0 ± 2.0	33.2 ± 2.5
Dense step 512	Mixed-control	36.9 ± 0.5	32.6 ± 0.9	30.0 ± 0.8
Dense step 512	Uniform	37.6 ± 0.9	33.4 ± 0.4	30.5 ± 0.5
Dense informative	Visual-heavy	36.7 ± 0.2	32.2 ± 0.2	29.3 ± 0.4
Dense informative	Reasoning-heavy	39.2 ± 2.2	35.2 ± 2.5	32.5 ± 2.3
Dense informative	Mixed-control	37.0 ± 0.3	33.1 ± 0.6	29.7 ± 0.5
Dense informative	Uniform	37.2 ± 0.8	33.2 ± 0.5	29.7 ± 0.8
Exact-only step 512	Mixed-control	36.4 ± 0.2	31.8 ± 0.6	28.9 ± 0.1

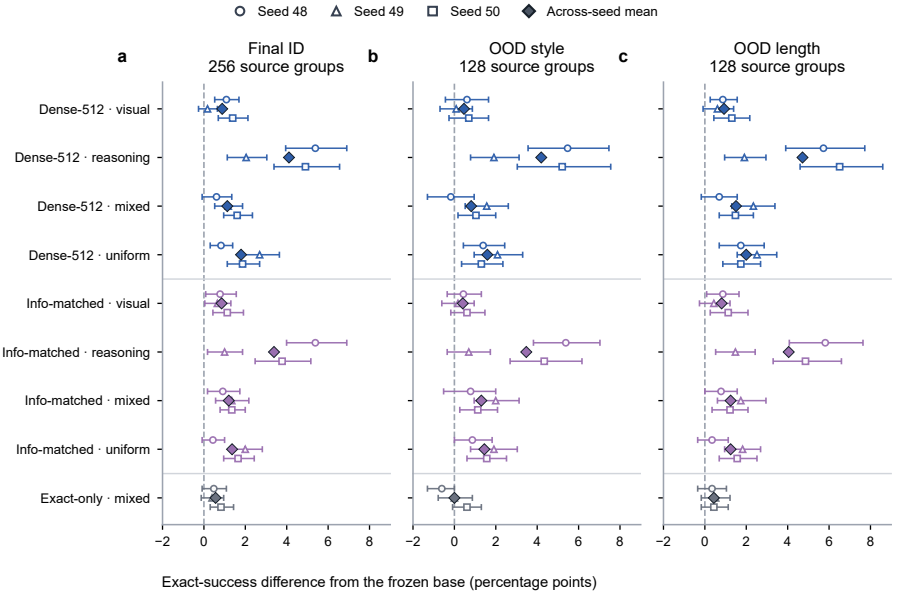

Figure 3: Complete confirmatory effects under greedy decoding. Points show seed-specific exact-success differences from the same unadapted base model; horizontal bars are source-group paired 95% bootstrap intervals from 20,000 resamples. Diamonds show the across-seed mean. Panels correspond to all three frozen final splits, and no checkpoint group or seed is omitted.

exact success on final-ID, OOD-style, and OOD-length. Dense reasoning-heavy training raises these values to $39.9 \pm 1.8\%$, $36.0 \pm 2.0\%$, and $33.2 \pm 2.5\%$ across three seeds, corresponding to mean gains of 4.1, 4.2, and 4.7 percentage points. Every one of the 36 reasoning-heavy seed–split–decoding contrasts is positive, and every source-group paired 95% bootstrap interval excludes zero.

At the same optimizer-step budget, reasoning-heavy training exceeds uniform by 2.31 ± 2.68 points on final-ID. The seed-48/49/50 effects are $+4.56$, -0.65, and $+3.04$ points, with paired 95% intervals $[3.30, 5.90]$, $[-1.30, -0.04]$, and $[1.78, 4.38]$. This meets the prespecified directional criterion, but the advantage is not consistent across training seeds. The mean differences on OOD-style and OOD-length are 2.60 and 2.72 points. Mixed-control and visual-heavy improve over base, yet their greedy means are below uniform on all three splits (Appendix G).

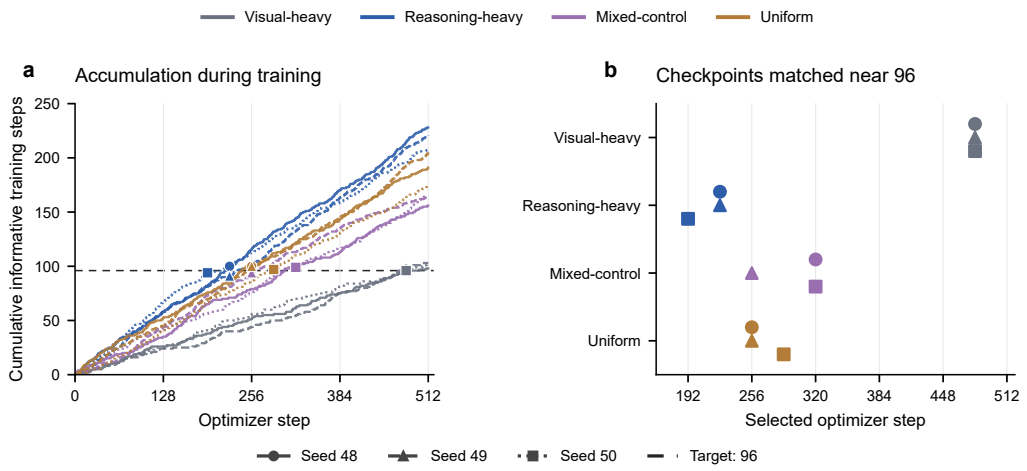

Figure 4: Reward informativeness and checkpoint matching. (a) Cumulative informative training steps for all four curricula and seeds 48–50; curves show every logged step without smoothing. The dashed reference marks the target of 96, and symbols mark selected checkpoints. (b) Selected optimizer steps for the same runs. Counts at selection range from 91 to 100 because checkpoints are saved every 32 steps. Colors identify curricula; line styles and symbols identify individual seeds, not aggregate estimates.

5.2 Informative-step matching preserves most of the effect

Figure 4 shows how the diagnostic accumulates over training. At step 512, visual-heavy runs contain 98–103 informative steps, compared with 208–228 for reasoning-heavy, 156–165 for mixed-control, and 175–205 for uniform. Matching near 96 therefore compares checkpoints at different optimizer steps rather than equating nominal training duration.

At checkpoints matched near 96 cumulative informative training steps, reasoning-heavy training remains 3.4, 3.5, and 4.1 points above the base model under greedy decoding. Across all 36 seed–split–decoding comparisons, the matched reasoning-heavy effect is positive in every case and its paired interval excludes zero in 32. The retained gain is consistent with a curriculum effect beyond this measured count of informative steps. Informative-step matching controls this logged count, while training duration and example order can still differ.

Dense uniform training is consistently beneficial relative to base (36/36 positive comparisons; 34/36 paired intervals exclude zero). At informative-matched checkpoints, the direct reasoning-heavy minus uniform differences are 2.03, 2.03, and 2.81 points on the three greedy splits. Each is positive in seeds 48 and 50 but negative in seed 49, so matching preserves the mean advantage rather than establishing a seed-independent effect.

5.3 Exact-only reward yields small gains under mixed-control

The reward ablation holds the mixed-control curriculum fixed. Dense mixed-control reaches 36.9%, 32.6%, and 30.0% greedy exact success, compared with 36.4%, 31.8%, and 28.9% for exact-only reward. The direct dense-minus-exact effects are 0.56 ± 0.38, 0.81 ± 0.65, and 1.07 ± 0.74 points. All nine greedy seed effects are positive, but only five paired intervals exclude zero (Appendix G). Exact-only mixed-control stays close to the base model: its greedy gains are 0.6, 0.0, and 0.4 points, and only 5 of its 36 seed–split–decoding intervals against base exclude zero in the positive direction. Six exact-only contrasts against base are negative and three are exactly zero, as determined from integer success-count differences.

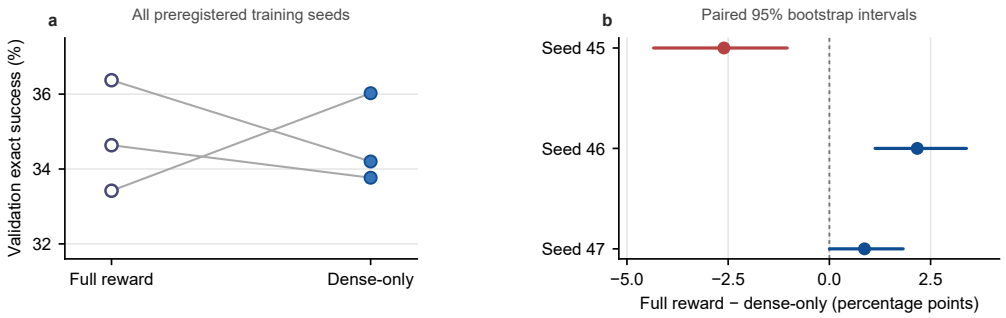

Figure 5: Frozen reward audit on validation exact success. (a) Matched seed-level values for the full reward, which includes the auxiliary format bonus, and the dense-only reward. (b) Full-minus-dense contrasts with source-group paired 95% bootstrap intervals. Each seed contributes 1,152 examples from 128 source groups; intervals use 20,000 resamples. All preregistered seeds 45–47 are shown.

5.4 The conclusion is stable across decoding conditions

Across the three stochastic decoding seeds, reasoning-heavy dense training retains the largest family mean gain. Dense reasoning-heavy family means range from +3.7 to +4.9 points over the base model across all 12 split–decoding cells; informative-matched reasoning-heavy means range from +3.0 to +4.1 points. Of all 108 preregistered checkpoint-group–split–decoding family means, 106 are positive, one is exactly zero, and one is negative. The zero and negative means occur for exact-only mixed-control on OOD-style under greedy and one stochastic decoding, respectively (the latter is -0.4 points). Signs are determined from integer success-count differences before rounding. Appendix D displays the complete sensitivity matrix, including near-zero and reversed effects.

A post-hoc factorial decomposition shows that gains are heterogeneous within each split, with larger reasoning-heavy gains in unmasked, higher-reasoning-load cells. Strongly masked cells often show small or near-zero changes. Appendix E reports every cell for all four decodings; this decomposition is descriptive, without cell-level significance tests.

5.5 The auxiliary format bonus fails its frozen retention rule

Figure 5 shows the prespecified validation audit that determined the reward used in the main study. Retaining the format bonus required a mean gain of at least 1.0 percentage point and a positive gain in at least two of three seeds. The mean paired full-minus-dense difference was $+0.14$ percentage points (sample SD 2.47 points) and was positive in two of three training seeds. The directional condition was met, but the frozen $+1.0$-point mean threshold was not; we therefore retained dense-only without retuning. The negative seed-45 contrast and positive seed-46 contrast provide no evidence that the format bonus improves exact success robustly across training seeds.

6 Discussion

The fixed-step and informative-step-matched comparisons separate nominal training duration from the frequency of logged reward dispersion. Neither comparison equalizes token exposure, gradient magnitudes, or optimizer history. The mean reasoning-heavy advantage over uniform survives matching, but reverses in seed 49. Thus curriculum rankings remain sensitive to training randomness; the reported contrasts do not identify the cause of that reversal.

The factorial construction clarifies what reasoning load means here. Longer paths require more transitions, more scored fields, and longer formatted responses. The advantage may reflect adaptation to this compound demand rather than an isolated change in latent reasoning ability. Gains concentrate in unmasked cells, whereas strong masking yields little improvement. The results therefore support adaptation to structured execution more clearly than improved robustness to visual occlusion. Because

OOD-length mixes 2-, 6-, and 10-transition instances, its aggregate gain does not imply improvement at every path length.

The mixed-control ablation holds training content fixed while changing reward granularity. Its mean advantage for dense reward does not distinguish richer feedback from a higher frequency of informative steps. Testing exact-only reward under reasoning-heavy training would further determine whether the reward comparison depends on curriculum composition.

7 LIMITATIONS

The study covers one synthetic task family, one 8B backbone, and one LoRA training budget; transfer to external visual-reasoning benchmarks is untested. The 512-record targeted pools and 2,304-record uniform pool differ in coverage and potential sample reuse at a fixed step budget. Informative-step matching is approximate and does not remove these differences. Path length also changes formatting demands under the shared 240-token cap, so long-path errors cannot be attributed exclusively to reasoning. The exact-only ablation covers mixed-control only. These constraints limit causal and cross-domain interpretations.

8 CONCLUSION

FACTORGRAPH-RLVR provides a controlled setting for studying curriculum composition alongside reward informativeness. Reasoning-heavy dense training yields the largest mean exact-success gains in the evaluated matrix and retains most of its improvement at informative-step-matched checkpoints. Uniform training produces smaller gains on average across seeds, but the direct curriculum ranking reverses in seed 49. Exact-only mixed-control remains near the base model. These findings support reporting curriculum content, optimizer exposure, and informative training steps separately when interpreting visual RLVR results.

ETHICS STATEMENT

This work uses synthetic images and does not collect personal or sensitive information. The main risks concern reproducibility and overclaiming; we therefore distinguish development from confirmatory evidence and retain frozen manifests for every reported run.

REPRODUCIBILITY STATEMENT

Experiment code, datasets, and per-example evaluation results are available in the anonymous repository.

The training records all random seeds and uses frozen JSONL manifests, deterministic graph programs, and paired source groups; GRPO execution is not assumed bitwise deterministic across launch paths or devices. The accompanying artifact bundle includes checkpoint freezes, final-dataset inventory and validation reports, a pre-outcome evaluation manifest, a runtime lock, the direct launcher, completion records from all 336 evaluation jobs, and the frozen statistics outputs. Appendix B summarizes the inputs and analysis rules. Every final value in the manuscript is generated from the complete validated matrix rather than manually transcribed from partial jobs.

AI USE STATEMENT

Generative AI tools assisted with software implementation and debugging, analysis of logged experimental results, literature search and summarization, and reference formatting. All AI-assisted work was reviewed and verified by the authors, who take responsibility for the final manuscript, code, claims, and released artifacts.

REFERENCES

Md Tanvirul Alam, Saksham Aggarwal, Justin Yang Chae, and Nidhi Rastogi. SPHINX: A synthetic environment for visual perception and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, 2026.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, Jeongyeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of EACL*, pp. 700–719, 2026. doi: 10.18653/v1/2026.eacl-long.30. URL https://aclanthology.org/2026.eacl-long.30/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. PuzzleVQA: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 16259–16273, 2024. doi: 10.18653/v1/2024.findings-acl.962. URL https://aclanthology.org/2024.findings-acl.962/.

Yichen Feng, Zhangchen Xu, Fengqing Jiang, Yuetai Li, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Visualsphinx: Large-scale synthetic vision logic puzzles for RL. *arXiv preprint arXiv:2505.23977*, 2025. URL https://arxiv.org/abs/2505.23977.

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

Ahmadreza Jeddi, Hakki C. Karaimer, Hue Nguyen, Zhongling Wang, Ke Zhao, Javad Rajabi, Ran Zhang, Raghav Goyal, Konstantinos G. Derpanis, Babak Taati, and Radek Grzeszczuk. Puzzlecraft: Exploration-aware curriculum learning for puzzle-based RLVR in VLMs, 2025. URL https://arxiv.org/abs/2512.14944. Version 2, revised March 2026.

Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. VCRL: Variance-based curriculum reinforcement learning for large language models. *arXiv preprint arXiv:2509.19803*, 2025. URL https://arxiv.org/abs/2509.19803.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. MMR1: Enhancing multimodal reasoning with variance-aware sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9075–9087, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025a. URL https://arxiv.org/abs/2503.20783.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025b.

Shan Ning, Longtian Qiu, and Xuming He. Wiki-R1: Incentivizing multimodal reasoning for knowledge-based VQA via data and sampling curriculum. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TH1Tgbjkm7.

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 preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.14476.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1859–1869, 2025.

A Implementation details

The reference configuration uses `rlvr-py311-cu124`, Qwen3-VL-8B-Instruct, and 512 optimization steps. LoRA uses rank 16, scaling factor 32, dropout 0.05, and all-linear target modules, with the visual encoder unfrozen. Training uses bfloat16 and SDPA attention, maximum sequence length 4,096, and maximum completion length 240. Rollout settings are four generations per prompt, temperature 1.2, top-p 0.95, and repetition penalty 1.1. All twelve dense and three exact-only final checkpoint configurations specify per-device training batch size one and four gradient-accumulation steps. Both the model seed and data seed equal the designated training seed.

KL regularization and learning rate. All fifteen runs use KL coefficient $\beta = 0.04$, applied to the per-token loss rather than subtracted from the verifier reward (`kl_in_reward=False`). Reference log probabilities are evaluated with the policy's LoRA adapter disabled, rather than a separately loaded reference model. The learning rate increases linearly to 5×10^{-6} over $\lceil 0.05 \times 512 \rceil = 26$ warmup steps, then follows cosine decay to zero over the remaining steps. Checkpoints matched by informative-step count therefore need not have the same instantaneous learning rate or optimizer history.

Logged reward dispersion. All runs use `scale_rewards=group` and `logging_steps=1`. For collection event k, let $B_{t,k}$ be the number of prompt groups and $r_{t,k,b,g}$ their verifier rewards, with $G = 4$ rollouts per group. The implemented reduction is

$$s_{t,k,b} = \sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(r_{t,k,b,g} - \bar{r}_{t,k,b})^2}, \qquad \sigma_t = \frac{1}{K_t}\sum_{k=1}^{K_t}\frac{1}{B_{t,k}}\sum_{b=1}^{B_{t,k}} s_{t,k,b}, \tag{7}$$

where $\bar{r}_{t,k,b}$ is the group mean and K_t counts metric collections since the preceding log. The metric buffer is cleared after each log. Thus `reward_std` averages within-group sample standard deviations; it is neither a pooled standard deviation nor a count of non-degenerate groups. The frozen diagnostic thresholds this scalar at 10^{-8}. Zero dispersion removes the reward-ranking signal, but need not imply zero parameter change because KL regularization and optimizer state can still contribute.

Prompt templates. The system instructions below are shared across curricula. The user message supplies the rendered image and the text template below. Placeholder `STATE` is the initial two-digit state, `N` is the requested transition count, and `RELATIONS` joins the requested relation names with the literal separator ` then `. The ID and OOD-length generators use the same wording. Line wrapping below is for presentation.

```
Use the BLUE numeric relation table on the left for BLUE
steps and the ORANGE numeric relation table on the right for
ORANGE steps.
```

```
Each row maps its two-digit SOURCE state to its two-digit
DESTINATION state.
STATE and ACC are separate variables even though both are
two digits.  Start with ACC=00.
At every step, the next STATE is the previous DESTINATION.
Separately add that numeric destination to the previous ACC
and keep the last two digits (mod 100).
For every requested step, output exactly one line in this
form:
STEP 1:  17-[BLUE]->42 | ACC: 42
After all requested STEP lines, output exactly one final
line containing only the last destination:
FINAL: 42
Always preserve this line structure even when a glyph is
uncertain.  Do not add prose, headings, arithmetic
expressions, extra labels, or unrequested steps.
```

User text template:

```
The image contains a BLUE numeric relation table on the left
and an ORANGE numeric relation table on the right.  Each row
maps one two-digit SOURCE to one two-digit DESTINATION.
STATE and ACC are separate variables even though both are
two digits.  Start with ACC=00.  The next STATE is the
previous DESTINATION. Separately update ACC by adding that
numeric destination and keeping the last two digits (mod
100).  Start at STATE={STATE} and follow these {N} relations
in order:  {RELATIONS}.
Output exactly {N} STEP lines with every destination and
cumulative ACC, then one FINAL line containing only the last
destination.
```

The evaluation implementation supplies system text, then a user message containing the image followed by the task text, through the model processor's chat template. It uses `add_generation_prompt=True`. The instructions prohibit additional prose, but this is not a constrained decoder or a measured guarantee against truncation. Removing text before a final `</think>` tag is a parser rule, not a generation-time thinking-token suppression mechanism.

Infrastructure amendment. All configured arms use effective world size one. A hardware failure interrupted mixed-control seed 49 after the run had progressed beyond its last complete optimizer-boundary checkpoint at step 416; steps 417–440 from the failed attempt are non-authoritative. We restored the full trainer state from checkpoint 416 on a healthy device and retained both trajectories for auditability. Because the failed device remains unavailable, all seed-50 arms use a frozen direct single-process entry path with unchanged datasets, seeds, configured sampling, effective batch size, optimizer, model, reward, and evaluation rules. The entry path and rank metadata differ from the earlier world-size-one launcher, so bitwise equivalence across launch paths is not claimed.

Parser. The parser scores the text following the last literal `</think>` tag, or the entire completion when the tag is absent, after removing blank lines. A format-valid completion has exactly N consecutively numbered `STEP` lines followed by one `FINAL` line, for $N + 1$ nonblank lines in total:

```
STEP i:  SOURCE-[RELATION]->DESTINATION | ACC: VALUE
FINAL: VALUE
```

where `RELATION` is `BLUE` or `ORANGE`. Matching is case-insensitive. Missing, additional, misnumbered, or duplicated lines and any extra answer-region commentary cause parse failure. The grammar does not independently enforce the two-digit semantic domain of source, destination, accumulator, or final tokens; syntactically valid tokens are judged by equality with the verifier-generated target.

Manifest serialization. Training manifests store the complete target trace plus `FINAL` for the reward scorer. Evaluation manifests store the final token and graph metadata, from which the

evaluator independently reconstructs the ordered target trace. The frozen evaluation gate treats explicit conversion and round-trip validation as required checks rather than treating the two serializations as interchangeable.

B FROZEN PROTOCOL AND REPRODUCIBILITY CHECKS

The evaluation inputs and analysis rules were fixed before confirmatory outcomes were opened. Full manifests, checksums, and completion records are retained in the artifact bundle; they establish provenance, not model quality.

B.1 CHECKPOINT SELECTION

All main training conditions retain seeds 48–50. Checkpoints were saved every 32 optimizer steps through step 512. An informative training step has logged reward standard deviation greater than 10^{-8}, as defined in Section 3.5. The matching rule selects the saved checkpoint nearest to a cumulative count of 96, with ties resolved toward the earlier step; if the run does not reach 96, step 512 is retained and the target is marked unmet. All twelve dense runs reach the target. Table 2 gives the selected checkpoints, whose counts range from 91 to 100. The exact-only comparison uses all three step-512 checkpoints, without outcome-based selection.

Table 2: Informative-step matching. Each entry is optimizer step / cumulative informative training steps. Selection uses training diagnostics, not final-test performance.

Curriculum	Seed 48	Seed 49	Seed 50
Visual-heavy	480 / 96	480 / 99	480 / 96
Reasoning-heavy	224 / 100	224 / 91	192 / 94
Mixed-control	320 / 99	256 / 94	320 / 99
Uniform	256 / 100	256 / 100	288 / 97

B.2 FINAL SPLIT CONSTRUCTION

Each source program produces nine instances, crossing three visual masking levels with three path lengths. Table 3 lists the split sizes and configured rendering changes. OOD-length retains the two-transition level while replacing the three- and four-transition levels with six and ten transitions. Its aggregate score therefore combines short and longer paths; it is not a long-path-only estimate.

Table 3: Frozen final split settings. Resolution is output width $\times$ height in pixels. All splits use masking band widths of 0, 4, and 7 pixels at the native rendering resolution of 1152×1050.

Split	Instances	Sources	Path lengths	Resolution
final-ID	2,304	256	2, 3, 4	1152×1050
OOD-style	1,152	128	2, 3, 4	1024×934
OOD-length	1,152	128	2, 6, 10	1152×1050

OOD-style changes the palette, font sizes, and line geometry, resizes the image, and adds blur and point noise. Its configured blur radius is 0.35 with 300 noise points, versus zero for final-ID and OOD-length. Node and legend font sizes change from 26 and 29 to 24 and 27; edge width changes from 6 to 5 pixels. Relation semantics, masking levels, and path lengths remain unchanged. OOD-length uses the final-ID rendering settings.

All 4,608 sample identifiers are globally unique. Saved validation reports pass the static gate, verify recomputed solutions and absence of answer leakage, and record no frozen source-signature or generation-key overlap. The generation checks include the smoke, training, and development reference sets and previously generated final splits. These checks validate data integrity; they do not filter model results.

B.3 CURRICULUM CONSTRUCTION

Each training curriculum uses the same 256 source programs, indexed from 0 to 255. Mixed-control takes cells $(d_v, d_r) = (1, 0)$ and $(0, 1)$ for even indices, and $(2, 0)$ and $(0, 2)$ for odd indices. Each source therefore contributes one visual-stress instance and one reasoning-stress instance at the same nonzero stress level. The resulting 512 records contain 128 instances of each selected cell. Visual-heavy takes $(1, 0)$ and $(2, 0)$ for every source; reasoning-heavy takes $(0, 1)$ and $(0, 2)$. Uniform takes all nine cells for every source. These are fixed training pools, not an alternating temporal schedule of optimizer updates.

B.4 EVALUATION AND STATISTICAL REPORTING

The matrix contains one unadapted base model, twelve dense step-512 checkpoints, twelve informative-step-matched dense checkpoints, and three exact-only step-512 checkpoints. Crossing these 28 configurations with three splits and four decoding conditions gives 336 jobs. Each job produces one completion per instance with a 240-token generation cap.

Greedy decoding is the primary condition (seed 2046090301). The three stochastic conditions share temperature 0.7 and top-p 0.8 and use seeds 2046090701–2046090703. They assess sampling-seed sensitivity at one sampling configuration, not robustness over a temperature sweep.

The frozen analysis uses 20,000 paired bootstrap resamples of source groups with random seed 2046090301 and 95% intervals. A resampled group retains all nine factorial instances, paired between a trained model and the base model. These intervals condition on a particular training seed; means and sample standard deviations across the three training seeds are reported separately. Strict-gate outcomes are descriptive diagnostics, not criteria for dropping jobs or rerunning conditions.

Descriptive diagnostic coverage. The preregistration also specified a 12-point plot of exact-success gain against informative-step fraction, one point per dense curriculum and training seed. That scatter plot is not included in this manuscript. Figure 4 instead reports cumulative informative-step trajectories and checkpoint locations; it is a different descriptive diagnostic, not the specified gain–fraction analysis. No correlation-based inference is claimed.

B.5 EXECUTION RECORDS

The artifact bundle retains runtime locks, model and tokenizer resources, project-code and evaluator records, checkpoint manifests, and per-job completion checks. Hardware interruptions and recovery are documented separately from statistical selection. Completed jobs and all positive, null, and negative outcomes are retained.

C DEVELOPMENT-STAGE EVIDENCE

The earlier development runs used the combined dense-plus-format reward and seeds 42–44; they are kept separate from the frozen dense-only confirmatory study. The base model reached 32.29% validation exact success. Reasoning-heavy gains ranged from 2.34 to 2.86 points, mixed-control gains from 1.74 to 2.78 points, and the visual-heavy effect was not robust across seeds. These results informed the final protocol but are not pooled with confirmatory outcomes.

Table 4: Development validation exact success (%). Parentheses give paired gains over the same 32.29% base model. Mean ± sample standard deviation summarizes seeds 42–44.

Condition	Seed 42	Seed 43	Seed 44	Mean ± SD
Base	32.29	32.29	32.29	32.29 ± 0.00
Visual-heavy	33.51 (+1.22)	32.47 (+0.17)	32.90 (+0.61)	32.96 ± 0.52
Reasoning-heavy	34.64 (+2.34)	35.16 (+2.86)	34.64 (+2.34)	34.81 ± 0.30
Mixed-control	34.03 (+1.74)	35.07 (+2.78)	34.55 (+2.26)	34.55 ± 0.52

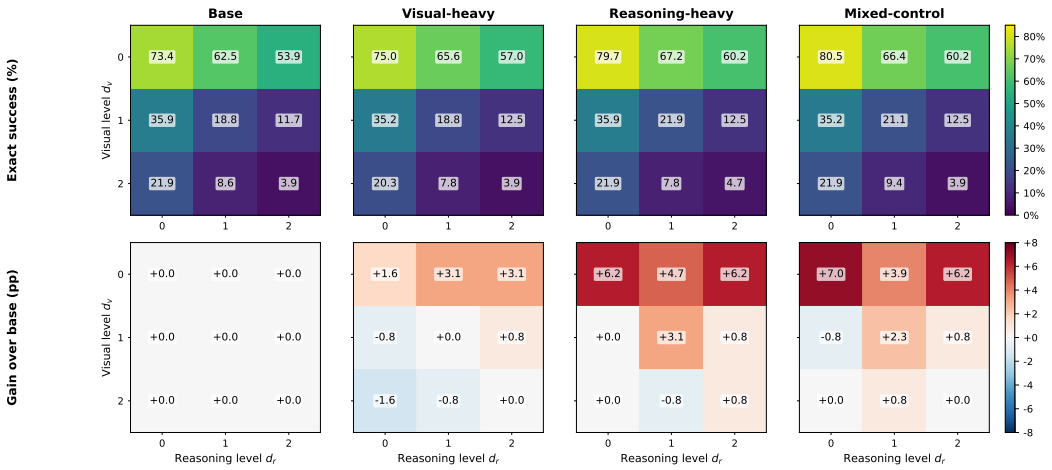

Figure 6: Complete development-stage validation response surfaces for combined-reward seed 44. The top row shows exact success and the bottom row shows paired gain over the base model for each visual/reasoning cell. Values are percentages.

D COMPLETE DECODING-SENSITIVITY MATRIX

Figure 7 reports every checkpoint-group–split–decoding family mean from the frozen analysis. It complements the seed-level intervals in Figure 3; the heat map is descriptive and does not replace the source-group paired bootstrap analysis.

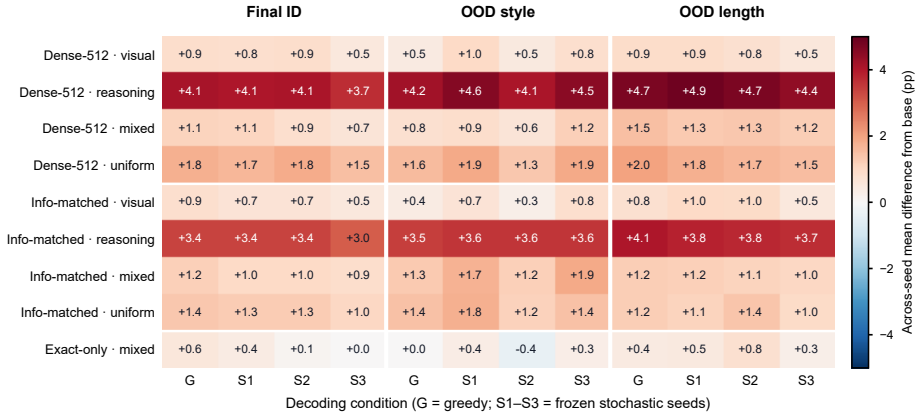

Figure 7: Complete decoding sensitivity. Each cell is the mean exact-success difference from the unadapted base model across the three frozen training seeds for one checkpoint group, split, and decoding condition. All 108 preregistered family means are shown, including near-zero and negative values.

E FINAL FACTORIAL DECOMPOSITION

Figures 8–11 decompose the final outcomes by visual and reasoning level. This is a post-hoc descriptive analysis, separate from the preregistered split-level bootstrap comparisons. Each entry averages the paired model–base difference across seeds 48–50. No cell-level hypothesis tests are performed. Source tables retain all seed-level scores, base scores, and across-seed standard deviations.

Columns enumerate the nine (d_v, d_r) cells. For final-ID and OOD-style, reasoning levels correspond to 2, 3, and 4 transitions; for OOD-length they correspond to 2, 6, and 10. Every trained condition is included, with one observation per source group in each cell. Averaging the nine unrounded cell differences within a displayed row gives the corresponding across-seed mean model–base difference for that split and decoding. The shared color scale covers all four decodings.



Figure 8: Greedy final factorial decomposition. Cells show mean model–base exact-success differences (percentage points) across all three training seeds. All nine trained groups and all visual/reasoning cells are shown for each split. Colors encode descriptive means, not significance.



Figure 9: Complete factorial decomposition for stochastic repeat 1 (seed 2046090701). Temperature is 0.7 and top-p is 0.8. Entries, panels, and color scale follow Figure 8.

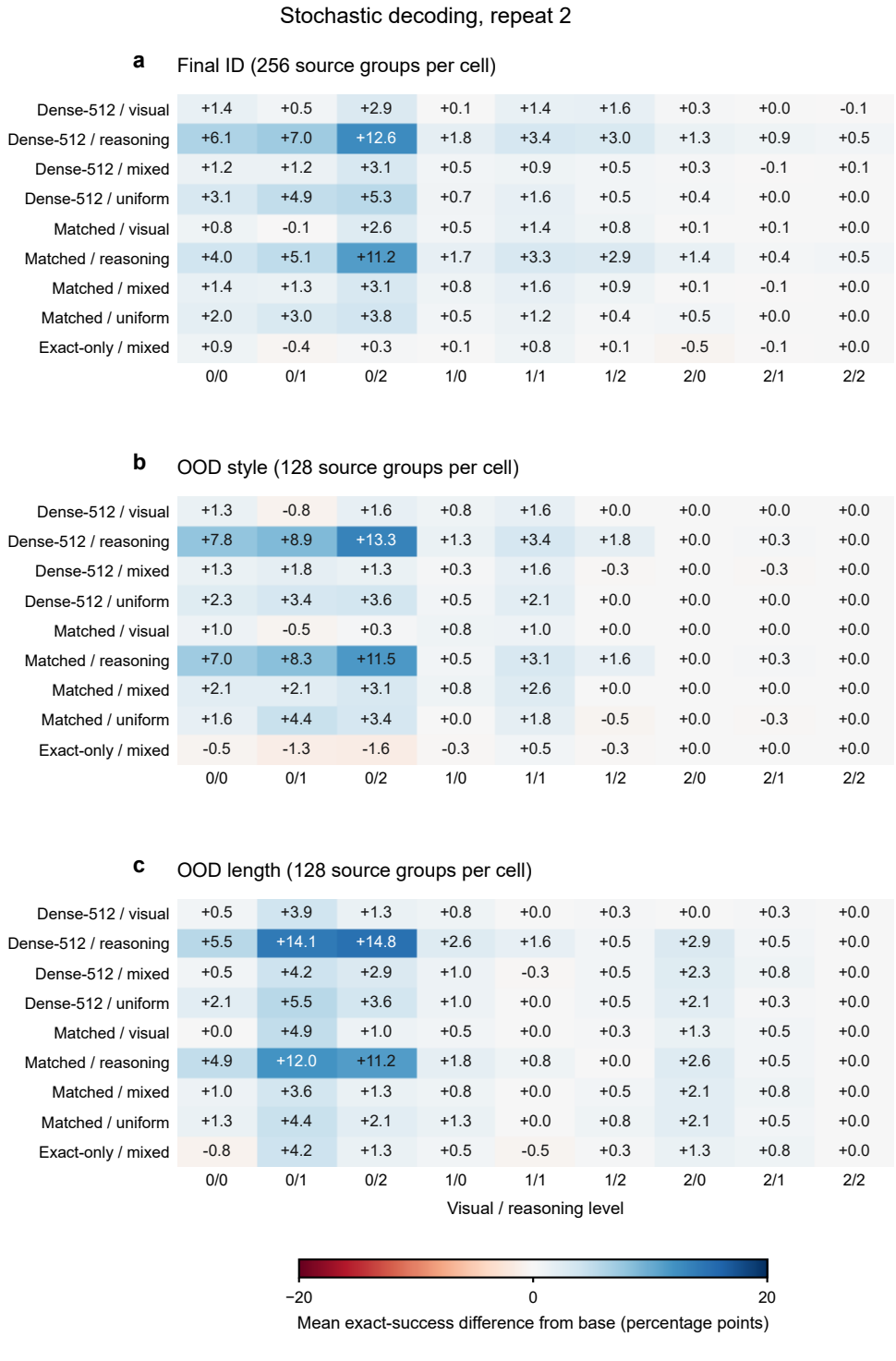

Figure 10: Complete factorial decomposition for stochastic repeat 2 (seed 2046090702), with the same sampling parameters and shared scale as Figure 9. All cells and checkpoint groups are retained.

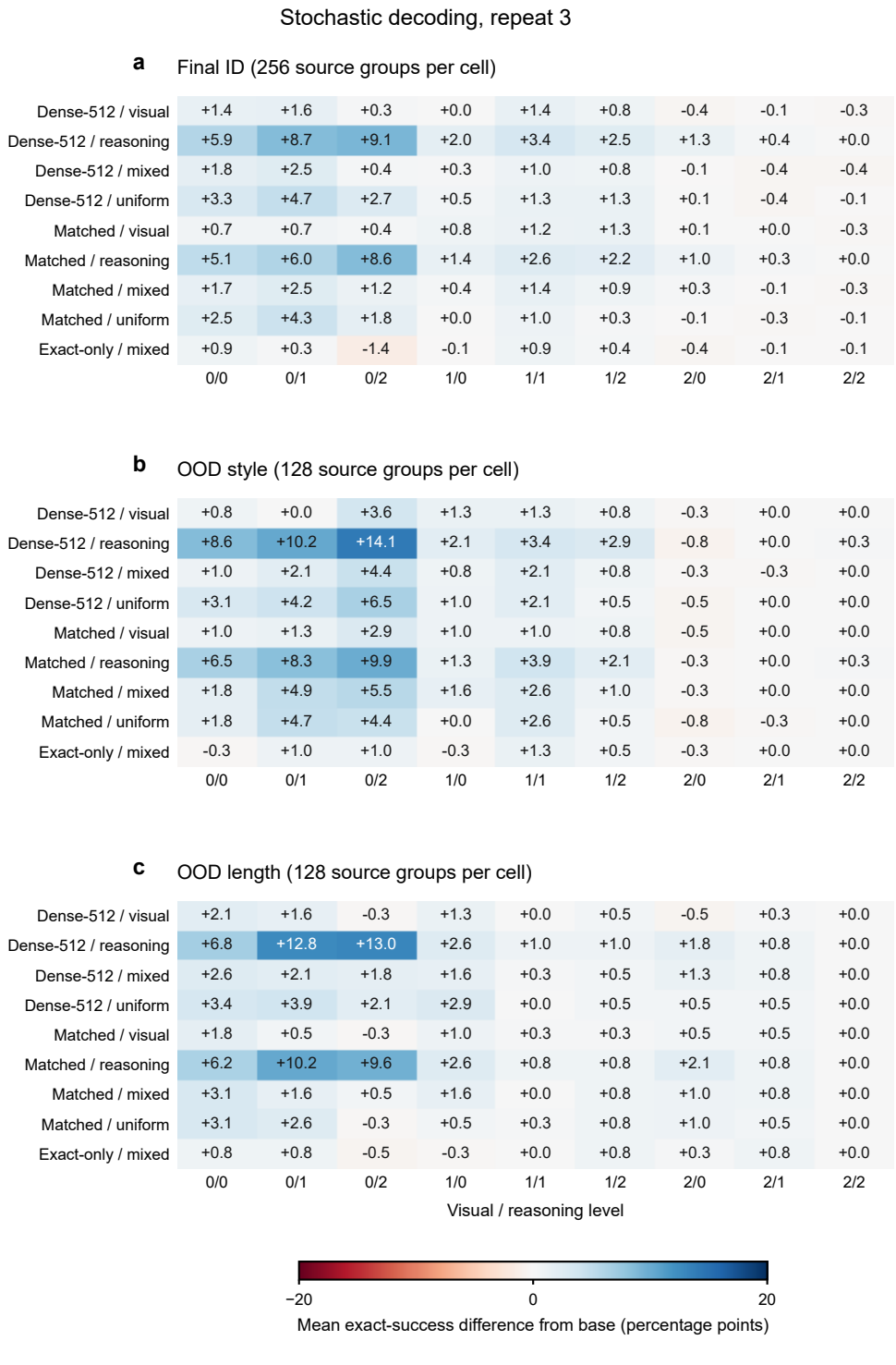

Figure 11: Complete factorial decomposition for stochastic repeat 3 (seed 2046090703), with the same sampling parameters and shared scale as Figure 9. All cells and checkpoint groups are retained.

F COMPLETE FACTORIAL PREVIEW

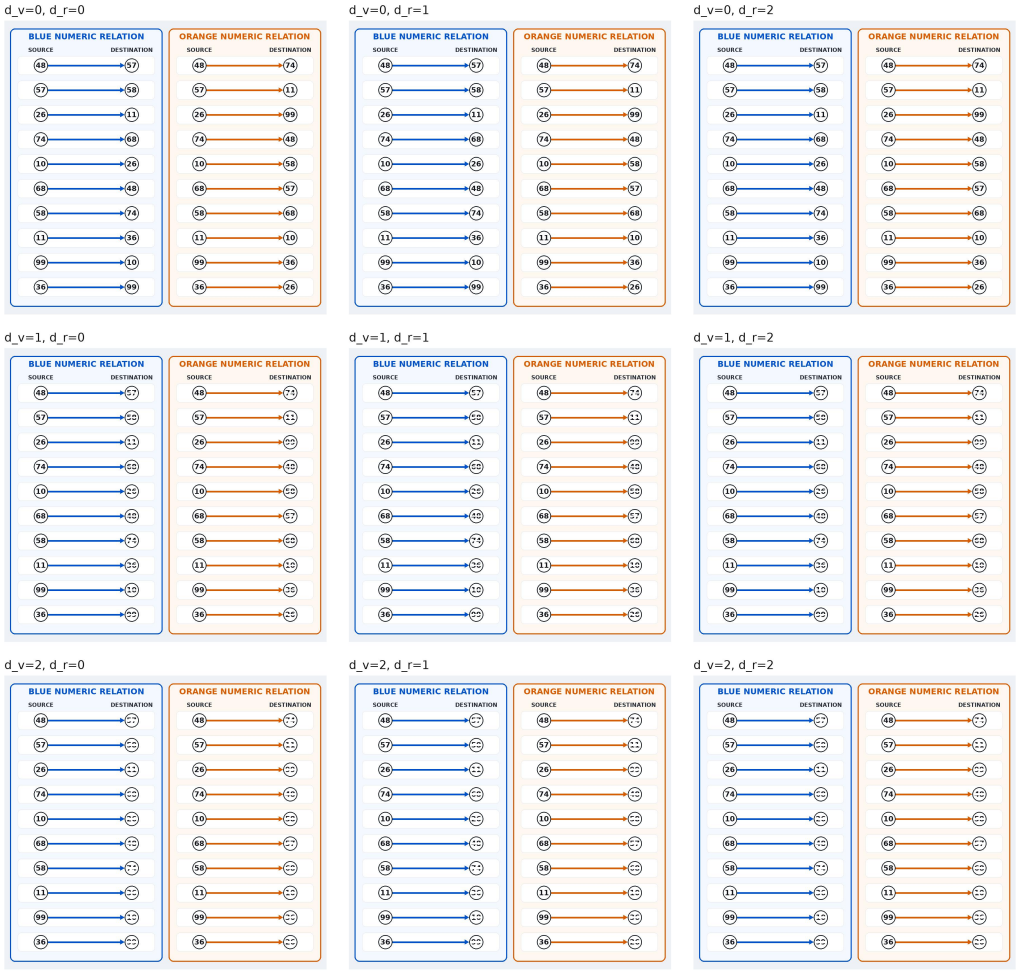

Figure 12: All nine matched cells for one source program. Rows increase destination-digit masking (d_v); columns increase the number of required transitions (d_r). The apparent truncation of destination digits in the lower rows is the controlled visual intervention, not a rendering error.

G COMPLETE DIRECT CURRICULUM AND REWARD CONTRASTS

We report seven direct comparisons for each of three splits and four decodings: three targeted dense curricula minus uniform at step 512, the same comparisons at informative-matched checkpoints, and dense minus exact-only at step 512 under mixed-control. This gives 84 comparison families and 252 training-seed effects, with no exclusions. The primary comparison is step-512 reasoning-heavy minus uniform on final-ID under greedy decoding; other comparisons are secondary or sensitivity analyses.

These comparisons implement the original analysis plan using existing frozen predictions. Their computation was completed after the model–base summaries; the supplemental enumeration is not a new pre-outcome registration. Pairing uses sample and rollout identifiers within the same source groups and training seed. Each source contributes nine cells; final-ID contains 256 sources and each OOD split 128. The frozen implementation uses 20,000 paired source-cluster resamples, bootstrap seed 2046090301, and linear-interpolated percentile endpoints. An independent integer-count implementation reproduced all 252 intervals to numerical precision. Intervals are pointwise and conditional on each trained checkpoint, with no multiple-comparison correction. Across-seed SD uses three training seeds, not three stochastic decoding repetitions.

Table 5: Direct paired comparisons, greedy. Values are percentage points. Seed columns give the effect [pointwise 95% source-cluster bootstrap interval]; mean ± SD summarizes the three effects. R/V/M are reasoning-heavy/visual-heavy/mixed-control dense training, U is uniform dense training, and D–E is dense minus exact-only under mixed-control.

Contrast	Mean ± SD	Seed 48	Seed 49	Seed 50
final-ID				
512: R–U	2.31 ± 2.68	4.56 [3.30, 5.90]	-0.65 [-1.30, -0.04]	3.04 [1.78, 4.38]
512: M–U	-0.67 ± 0.74	-0.22 [-0.82, 0.39]	-1.52 [-2.21, -0.87]	-0.26 [-0.78, 0.22]
512: V–U	-0.91 ± 1.44	0.26 [-0.26, 0.78]	-2.52 [-3.43, -1.69]	-0.48 [-1.00, 0.00]
Matched: R–U	2.03 ± 2.97	4.95 [3.65, 6.38]	-1.00 [-1.65, -0.39]	2.13 [1.09, 3.21]
Matched: M–U	-0.16 ± 0.58	0.48 [-0.30, 1.30]	-0.65 [-1.30, -0.04]	-0.30 [-0.82, 0.22]
Matched: V–U	-0.51 ± 0.85	0.35 [-0.35, 1.09]	-1.35 [-2.13, -0.65]	-0.52 [-1.04, -0.04]
512: D–E	0.56 ± 0.38	0.13 [-0.43, 0.69]	0.78 [0.09, 1.52]	0.78 [0.04, 1.52]
OOD-style				
512: R–U	2.60 ± 2.41	4.08 [2.26, 5.99]	-0.17 [-0.95, 0.61]	3.91 [2.00, 5.99]
512: M–U	-0.78 ± 0.69	-1.56 [-2.60, -0.61]	-0.52 [-1.48, 0.35]	-0.26 [-1.04, 0.52]
512: V–U	-1.13 ± 0.76	-0.78 [-1.82, 0.17]	-2.00 [-3.04, -1.04]	-0.61 [-1.48, 0.26]
Matched: R–U	2.03 ± 2.94	4.51 [2.86, 6.25]	-1.22 [-2.08, -0.43]	2.78 [1.48, 4.17]
Matched: M–U	-0.14 ± 0.27	-0.09 [-1.39, 1.22]	0.09 [-0.52, 0.78]	-0.43 [-1.48, 0.52]
Matched: V–U	-1.04 ± 0.66	-0.43 [-1.30, 0.43]	-1.74 [-2.69, -0.87]	-0.95 [-1.74, -0.26]
512: D–E	0.81 ± 0.65	0.43 [-0.43, 1.39]	1.56 [0.52, 2.69]	0.43 [-0.43, 1.39]
OOD-length				
512: R–U	2.72 ± 2.91	3.99 [2.34, 5.73]	-0.61 [-1.30, 0.09]	4.77 [3.04, 6.60]
512: M–U	-0.49 ± 0.48	-1.04 [-1.91, -0.17]	-0.17 [-0.87, 0.52]	-0.26 [-1.13, 0.61]
512: V–U	-1.07 ± 0.76	-0.87 [-1.82, 0.00]	-1.91 [-2.78, -1.04]	-0.43 [-1.13, 0.26]
Matched: R–U	2.81 ± 2.94	5.47 [3.73, 7.29]	-0.35 [-0.87, 0.17]	3.30 [2.00, 4.69]
Matched: M–U	0.00 ± 0.40	0.43 [-0.61, 1.39]	-0.09 [-0.87, 0.78]	-0.35 [-1.30, 0.61]
Matched: V–U	-0.43 ± 0.95	0.52 [-0.43, 1.39]	-1.39 [-2.17, -0.69]	-0.43 [-1.22, 0.35]
512: D–E	1.07 ± 0.74	0.35 [-0.61, 1.30]	1.82 [0.87, 2.86]	1.04 [0.09, 2.08]

Table 6: Direct paired comparisons, stochastic repeat 1 (generation seed 2046090701). Units, abbreviations and interval definitions follow Table 5.

Contrast	Mean ± SD	Seed 48	Seed 49	Seed 50
final-ID				
512: R–U	2.37 ± 2.79	4.64 [3.30, 6.12]	-0.74 [-1.43, -0.09]	3.21 [1.95, 4.56]
512: M–U	-0.61 ± 0.72	-0.48 [-1.17, 0.17]	-1.39 [-2.17, -0.65]	0.04 [-0.48, 0.56]
512: V–U	-0.84 ± 1.61	0.26 [-0.26, 0.78]	-2.69 [-3.56, -1.87]	-0.09 [-0.65, 0.48]
Matched: R–U	2.08 ± 3.13	4.95 [3.65, 6.34]	-1.26 [-2.04, -0.52]	2.56 [1.52, 3.69]
Matched: M–U	-0.30 ± 0.48	0.17 [-0.61, 1.00]	-0.78 [-1.52, -0.09]	-0.30 [-0.95, 0.30]
Matched: V–U	-0.61 ± 0.77	0.22 [-0.48, 0.95]	-1.30 [-2.13, -0.52]	-0.74 [-1.35, -0.13]
512: D–E	0.65 ± 0.75	-0.22 [-0.82, 0.35]	1.09 [0.30, 1.91]	1.09 [0.39, 1.82]
OOD-style				
512: R–U	2.69 ± 2.95	4.69 [2.78, 6.68]	-0.69 [-1.74, 0.26]	4.08 [2.26, 6.08]
512: M–U	-1.04 ± 0.76	-1.39 [-2.26, -0.52]	-1.56 [-2.78, -0.43]	-0.17 [-1.04, 0.69]
512: V–U	-0.90 ± 1.33	-0.17 [-0.95, 0.61]	-2.43 [-3.56, -1.39]	-0.09 [-0.87, 0.69]
Matched: R–U	1.79 ± 2.83	4.25 [2.69, 5.90]	-1.30 [-2.34, -0.26]	2.43 [0.95, 3.99]
Matched: M–U	-0.06 ± 0.13	-0.09 [-1.22, 1.04]	0.09 [-0.69, 0.95]	-0.17 [-1.04, 0.69]
Matched: V–U	-1.13 ± 1.05	-0.17 [-1.04, 0.69]	-2.26 [-3.30, -1.30]	-0.95 [-1.65, -0.26]
512: D–E	0.49 ± 0.49	0.35 [-0.35, 1.04]	0.09 [-0.87, 0.95]	1.04 [0.26, 1.91]
OOD-length				
512: R–U	3.10 ± 2.77	4.95 [3.21, 6.77]	-0.09 [-0.95, 0.78]	4.43 [2.78, 6.16]
512: M–U	-0.46 ± 0.36	-0.17 [-1.04, 0.69]	-0.87 [-1.65, -0.17]	-0.35 [-1.13, 0.35]
512: V–U	-0.84 ± 0.98	0.00 [-0.69, 0.69]	-1.91 [-2.78, -1.13]	-0.61 [-1.30, 0.09]
Matched: R–U	2.75 ± 2.95	5.21 [3.56, 6.94]	-0.52 [-1.13, 0.00]	3.56 [2.34, 4.86]
Matched: M–U	0.09 ± 0.61	0.78 [-0.17, 1.74]	-0.35 [-0.95, 0.26]	-0.17 [-0.78, 0.43]
Matched: V–U	-0.12 ± 1.01	0.78 [0.09, 1.56]	-1.22 [-2.00, -0.52]	0.09 [-0.69, 0.87]
512: D–E	0.81 ± 0.39	0.43 [-0.43, 1.30]	0.78 [-0.09, 1.74]	1.22 [0.35, 2.17]

Table 7: Direct paired comparisons, stochastic repeat 2 (generation seed 2046090702). Units, abbreviations and interval definitions follow Table 5.

Contrast	Mean ± SD	Seed 48	Seed 49	Seed 50
final-ID				
512: R–U	2.24 ± 2.90	4.69 [3.34, 6.21]	-0.95 [-1.69, -0.26]	2.99 [1.69, 4.38]
512: M–U	-0.98 ± 0.96	-0.52 [-1.22, 0.17]	-2.08 [-2.91, -1.35]	-0.35 [-1.00, 0.26]
512: V–U	-0.94 ± 1.68	0.26 [-0.39, 0.87]	-2.86 [-3.82, -2.00]	-0.22 [-0.82, 0.39]
Matched: R–U	2.13 ± 3.32	5.21 [3.86, 6.73]	-1.39 [-2.08, -0.69]	2.56 [1.52, 3.69]
Matched: M–U	-0.25 ± 0.59	0.30 [-0.39, 1.04]	-0.87 [-1.52, -0.26]	-0.17 [-0.74, 0.39]
Matched: V–U	-0.56 ± 1.02	0.48 [-0.22, 1.22]	-1.56 [-2.34, -0.82]	-0.61 [-1.26, 0.00]
512: D–E	0.72 ± 0.52	0.22 [-0.39, 0.82]	0.69 [-0.04, 1.48]	1.26 [0.52, 2.04]
OOD-style				
512: R–U	2.75 ± 2.46	4.34 [2.60, 6.25]	-0.09 [-0.95, 0.69]	3.99 [2.43, 5.73]
512: M–U	-0.69 ± 0.45	-1.22 [-2.17, -0.43]	-0.43 [-1.39, 0.43]	-0.43 [-1.39, 0.43]
512: V–U	-0.84 ± 0.95	-0.52 [-1.39, 0.26]	-1.91 [-3.04, -0.87]	-0.09 [-0.95, 0.78]
Matched: R–U	2.43 ± 2.64	4.77 [3.12, 6.60]	-0.43 [-1.30, 0.43]	2.95 [1.74, 4.25]
Matched: M–U	0.03 ± 0.66	0.17 [-1.13, 1.48]	0.61 [-0.17, 1.39]	-0.69 [-1.65, 0.17]
Matched: V–U	-0.87 ± 0.85	-0.17 [-1.13, 0.78]	-1.82 [-2.86, -0.78]	-0.61 [-1.30, 0.00]
512: D–E	1.01 ± 0.41	0.87 [0.17, 1.56]	1.48 [0.43, 2.60]	0.69 [-0.17, 1.65]
OOD-length				
512: R–U	3.04 ± 2.56	4.60 [2.78, 6.60]	0.09 [-0.78, 1.04]	4.43 [2.69, 6.25]
512: M–U	-0.35 ± 0.23	-0.52 [-1.39, 0.35]	-0.09 [-0.87, 0.69]	-0.43 [-1.22, 0.35]
512: V–U	-0.90 ± 0.18	-0.69 [-1.56, 0.09]	-0.95 [-1.74, -0.17]	-1.04 [-2.00, -0.17]
Matched: R–U	2.37 ± 2.57	4.60 [2.95, 6.34]	-0.43 [-1.22, 0.35]	2.95 [1.82, 4.17]
Matched: M–U	-0.26 ± 0.15	-0.35 [-1.22, 0.52]	-0.09 [-0.69, 0.61]	-0.35 [-1.04, 0.26]
Matched: V–U	-0.43 ± 0.78	0.35 [-0.43, 1.13]	-1.22 [-1.91, -0.52]	-0.43 [-1.22, 0.35]
512: D–E	0.55 ± 0.28	0.43 [-0.43, 1.30]	0.87 [0.00, 1.82]	0.35 [-0.52, 1.22]

Table 8: Direct paired comparisons, stochastic repeat 3 (generation seed 2046090703). Units, abbreviations and interval definitions follow Table 5.

Contrast	Mean ± SD	Seed 48	Seed 49	Seed 50
final-ID				
512: R–U	2.20 ± 2.81	4.64 [3.30, 6.12]	-0.87 [-1.61, -0.17]	2.82 [1.56, 4.17]
512: M–U	-0.84 ± 0.49	-0.69 [-1.39, 0.00]	-1.39 [-2.17, -0.65]	-0.43 [-0.95, 0.04]
512: V–U	-0.97 ± 1.65	0.35 [-0.26, 0.95]	-2.82 [-3.78, -1.95]	-0.43 [-1.04, 0.17]
Matched: R–U	1.98 ± 3.41	4.86 [3.52, 6.29]	-1.78 [-2.65, -1.00]	2.86 [1.87, 3.95]
Matched: M–U	-0.16 ± 0.76	0.56 [-0.22, 1.43]	-0.95 [-1.69, -0.26]	-0.09 [-0.74, 0.56]
Matched: V–U	-0.51 ± 1.31	0.48 [-0.22, 1.22]	-2.00 [-2.82, -1.26]	0.00 [-0.65, 0.69]
512: D–E	0.62 ± 0.76	-0.26 [-0.91, 0.39]	1.04 [0.30, 1.82]	1.09 [0.43, 1.78]
OOD-style				
512: R–U	2.63 ± 2.36	4.17 [2.43, 6.08]	-0.09 [-1.22, 1.04]	3.82 [2.08, 5.64]
512: M–U	-0.69 ± 0.60	-1.39 [-2.26, -0.52]	-0.35 [-1.48, 0.87]	-0.35 [-1.22, 0.52]
512: V–U	-1.04 ± 1.05	-0.43 [-1.22, 0.26]	-2.26 [-3.56, -1.04]	-0.43 [-1.22, 0.35]
Matched: R–U	2.11 ± 2.35	3.99 [2.52, 5.56]	-0.52 [-1.56, 0.52]	2.86 [1.56, 4.34]
Matched: M–U	0.46 ± 0.67	-0.09 [-1.48, 1.22]	1.22 [0.26, 2.26]	0.26 [-0.61, 1.22]
Matched: V–U	-0.61 ± 0.61	-0.35 [-1.30, 0.52]	-1.30 [-2.17, -0.43]	-0.17 [-0.95, 0.61]
512: D–E	0.84 ± 0.48	0.52 [-0.26, 1.30]	1.39 [0.26, 2.52]	0.61 [-0.17, 1.48]
OOD-length				
512: R–U	2.89 ± 2.58	4.43 [2.78, 6.16]	-0.09 [-0.78, 0.69]	4.34 [2.78, 5.99]
512: M–U	-0.32 ± 0.30	-0.35 [-1.22, 0.52]	0.00 [-0.69, 0.69]	-0.61 [-1.56, 0.26]
512: V–U	-0.98 ± 0.98	0.09 [-0.78, 0.95]	-1.82 [-2.86, -0.87]	-1.22 [-2.17, -0.35]
Matched: R–U	2.72 ± 2.71	5.03 [3.21, 6.94]	-0.26 [-1.04, 0.43]	3.39 [2.26, 4.69]
Matched: M–U	0.09 ± 0.76	0.95 [0.09, 1.91]	-0.26 [-0.87, 0.26]	-0.43 [-1.13, 0.17]
Matched: V–U	-0.43 ± 0.83	0.35 [-0.35, 1.04]	-1.30 [-2.08, -0.52]	-0.35 [-1.13, 0.35]
512: D–E	0.93 ± 0.89	0.17 [-0.78, 1.13]	1.91 [0.95, 2.86]	0.69 [-0.09, 1.48]