WHEN TO STOP PENALIZING ACTIONS? MARTINGALE GRPO FOR VLA REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A failed vision-language-action (VLA) rollout does not imply that every action contributed to failure. Yet in group-relative policy optimization (GRPO), a single trajectory-level advantage is assigned to all actions in a rollout. Potentially useful recovery attempts can thus be penalized alongside the errors they aim to correct. Ideal stopping conditions for failure feedback are characterized using success-probability and policy-score martingales. At these stops, the removed gradient contribution has zero conditional expectation given the retained information. Truncation therefore preserves the expectation of a reference GRPO group-gradient estimator and can reduce its covariance. One such stop occurs when the conditional probability of success is zero, motivating error and recovery cues for selecting candidate feedback boundaries. Martingale GRPO (MART-GRPO) is introduced with a vision-language model (VLM) for event-guided truncation. The VLM is fine-tuned on our annotated training set using low-rank adaptation (LoRA). The last visible error without observed recovery is localized by the frozen VLM in complete rollout videos. Failure feedback is retained through the selected error interval and removed afterward, while successful-rollout loss terms and original loss normalization are preserved. For retrospective boundaries, changes in expectation and variance are analyzed to assess departures from ideal stopping. Experiments with OpenVLA-OFT on RoboTwin 2.0 show that MART-GRPO improves success rates over full-trajectory GRPO across multiple tasks.

1 INTRODUCTION

Vision-language-action (VLA) models map visual observations and language instructions to actions (Zitkovich et al., 2023; Kim et al., 2025b). Pretrained policies are adapted through supervised fine-tuning on expert demonstrations (Kim et al., 2025a) and further optimized through reinforcement learning from environment interaction (Lu et al., 2025; Liu et al., 2025a). Group-relative policy optimization (GRPO) (Shao et al., 2024) turns terminal task rewards into trajectory-level advantages by comparing rollouts collected for the same task. Each advantage is then applied across the rollout's action chunks.

A terminal failure does not imply that every action contributed to failure. For example, a policy may drop an object, attempt to regrasp it, and exhaust its interaction budget (Figure 1). In a mixed-outcome group, the failed rollout's negative advantage penalizes both the error and the subsequent recovery attempts. This motivates the question: *When to stop penalizing actions?*

Several methods already change which actions receive feedback, each at a different point in the pipeline. Temporal GRPO reconstructs stage-relative advantages but requires task-stage detection and rollout alignment (Zhou et al., 2026). PCM allocates gradient computation through chunk sampling guided by phase-level success–failure divergence (Bagaria et al., 2026). RedFlow's corrective targets require suitable successful matches in offline data (Yan et al., 2026). In LLM reinforcement learning, negative-advantage updates are masked in μ-GRPO using importance-ratio triggers (Tian et al., 2026). ESPO terminates rollouts early (Li et al., 2026b), whereas PS-PPO samples update prefixes with importance weighting to preserve the reference gradient expectation (Hwang & Kim, 2026). Here, failed-rollout suffixes are masked using visible errors and recovery cues after com-

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

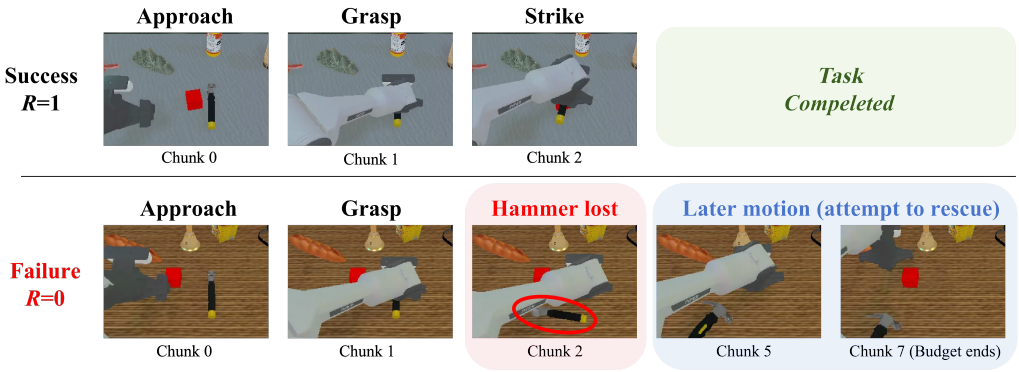


Figure 1: **Errors and recovery attempts within a failed rollout.** Selected RoboTwin hammer frames from two scenes, a success (top) and a failure (bottom). The red outline marks the hammer after it is lost at chunk 2, and the later frames show recovery attempts. In mixed-outcome groups, full GRPO assigns the error and subsequent recovery attempts the same negative advantage.

plete execution. Group advantages and original loss normalization are retained without additional correction weights for truncation.

To identify a principled truncation boundary, we ask when future actions make no expected gradient contribution. At a fixed policy, the conditional success probability and the cumulative policy score are both martingales, and their increments carry that contribution. Together they define the *remaining conditional credit*: the expected policy-gradient contribution of future actions given the observed history. For example, when failure is already determined by the observed history, the remaining credit is zero. If the decision to stop reads only the observed prefix and the remaining credit is zero, the removed failure contribution has zero conditional expectation. For conditionally independent rollouts whose normalization factors are measurable with respect to the retained information, the truncated GRPO group-gradient estimator is a conditional projection of the full reference estimator. It therefore has the same expectation and no greater covariance.

Motivated by this analysis, we propose *Martingale GRPO* (MART-GRPO) for event-guided failure-feedback selection. A vision-language model (VLM) selects feedback boundaries based on visible errors and whether recovery follows. The VLM is fine-tuned on error and recovery annotations and frozen during policy optimization. It reads each complete failed rollout and localizes the onset of the last visible error without observed recovery to a chunk interval. Later frames help interpret earlier events, so the boundary is selected retrospectively. The interval's right endpoint defines B: failure feedback is retained from chunk 0 through B and removed afterward. Successful-rollout loss terms are retained. A conditional residual relates these retrospective boundaries to the conditional projection above and bounds their effects on gradient expectation and total variance.

Our contributions are threefold.

- **A martingale analysis of failure feedback.** Ideal stopping rules are characterized through remaining conditional credit and a conditional projection of the complete GRPO group-gradient estimator. The projection preserves the reference expectation without increasing covariance, while residual and bias–variance analyses characterize the effects of retrospective boundaries.
- **Event-guided feedback truncation with MART-GRPO.** A VLM is fine-tuned with LoRA on our annotated training set and frozen to localize the last visible error without observed recovery in complete rollouts. These event intervals are converted into failure-prefix masks that preserve successful-rollout loss terms, group advantages, and original loss normalization.
- **Empirical evaluation of policy performance and feedback selection.** Seven-task OpenVLA-OFT comparisons on RoboTwin 2.0 show higher mean success than Full GRPO, complemented by localization and feedback-use measurements. Two-task ablations favor event-guided truncation over uniform penalty scaling and random prefixes.

2 A MARTINGALE ANALYSIS OF FAILURE FEEDBACK

We analyze feedback truncation through the gradient contribution of the discarded suffix. Under the conditions below, removing a contribution with zero conditional expectation given the retained information preserves the full-feedback estimator's expectation without increasing covariance. Two martingales make this statement precise for GRPO, and a projection residual then covers boundaries selected from complete rollouts.

2.1 GRPO WITH A FEEDBACK MASK

We express feedback truncation as a mask on GRPO's trajectory-level advantage. Trajectory i contains K_i action chunks and terminal reward $R_i \in \{0, 1\}$. History $h_{i,k}$ precedes chunk $a_{i,k}$, indexed from zero. A policy conditions on observations, task instructions, and available proprioception. For rollout group $g(i)$, the reference advantage is

$$A_i = \frac{R_i - \overline{R}_{g(i)}}{s_{g(i)} + \epsilon_{\text{num}}}, \qquad \overline{R}_g = \frac{1}{|g|} \sum_{j \in g} R_j. \tag{1}$$

Here s_g is the group's reward standard deviation and $\epsilon_{\text{num}} > 0$ ensures numerical stability. The mask modifies the temporal support of A_i as follows.

$$\begin{aligned} &\widetilde{A}_{i,k} = w_{i,k} A_i, \qquad B_i \in \{-1, 0, \ldots, K_i - 1\}, \\ &w_{i,k} = \begin{cases} 1, & R_i = 1, \\ \mathbf{1}\{k \le B_i\}, & R_i = 0. \end{cases} \end{aligned} \tag{2}$$

In the analysis, $B_i = -1$ represents an empty failure prefix. Let $\mathcal{V}_i$ be the valid action tokens and $N_i = |\mathcal{V}_i|$. Masking preserves the original N_i, since renormalizing by the retained tokens would scale up every surviving term. Section 3 gives the clipped policy objective.

2.2 REMAINING CONDITIONAL CREDIT

The ideal analysis concerns on-policy score gradients at a fixed differentiable collection policy π_θ, before PPO clipping or optimizer transformations. Use a finite horizon H, padding after termination with zero-score dummy actions. The objective is undiscounted terminal success. The environment and initial distribution do not depend on θ. Let $\mathcal{F}_k$ contain the complete pre-action history, including remaining budget, and let $\mathcal{F}_H$ contain the outcome. Assume common policy support, valid differentiation under expectation, and square-integrable scores

$$P_k = \mathbb{E}[R \mid \mathcal{F}_k], \qquad Z_k = \nabla_\theta \log \pi_\theta(a_k \mid h_k), \qquad S_t = \sum_{k=0}^{t-1} Z_k. \tag{3}$$

The score identity $\mathbb{E}[Z_k \mid \mathcal{F}_k] = 0$ makes S_t a martingale, while P_k is the Doob martingale of R (Durrett, 2019). Write $\Delta P_k = P_{k+1} - P_k$. Its expectation given the history and the selected action is the local action advantage, distinct from the group coefficient A_i (Appendix B.2). This advantage records how the action moved the success probability, so an action inside a failed rollout can still carry positive credit.

Pairing their increments gives $\nabla J = \sum_k \mathbb{E}[Z_k \Delta P_k]$ for $J(\theta) = \mathbb{E}[R]$ by score centering (Williams, 1992; Sutton et al., 1999) (Appendix B.2). Summing the same products from chunk t onward and conditioning on the observed history gives the remaining conditional credit

$$\chi_t = \mathbb{E}[R(S_H - S_t) \mid \mathcal{F}_t] = \mathbb{E}\left[\sum_{k=t}^{H-1} Z_k \Delta P_k \,\middle|\, \mathcal{F}_t \right]. \tag{4}$$

Holding the observed prefix fixed, χ_t is the first-order sensitivity of continuation success to the policy parameters (Appendix B.3). With random initial conditions, $\mathbb{E}[\chi_0] = \nabla J$, while χ_0 itself is conditional on $\mathcal{F}_0$. Contributions can cancel, so zero credit does not require every later action advantage to vanish.

2.3 CONDITIONAL PROJECTION OF A COMPLETE GRPO GROUP

We first consider rules that decide from the observed prefix. Let $T \in \{0, \dots, H\}$ be a stopping time, with $\{T \le k\} \in \mathcal{F}_k$. It selects the last penalized failure chunk $B = T - 1$. Successes retain their full scores, giving $\widetilde{S}_T = (1 - R)S_T + RS_H$. Let $\mathcal{K}_T$ record the stopped history and final outcome, together with the complete trajectory on success. This retained information generally exceeds $\mathcal{F}_T$.

Proposition 1 (A conditional projection characterization)**.** *Under the score assumptions above,*

$$\chi_T = 0 \quad \Longleftrightarrow \quad \widetilde{S}_T = \mathbb{E}[S_H \mid \mathcal{K}_T]. \tag{5}$$

Optional sampling gives $\mathbb{E}[(1 - R)(S_H - S_T) \mid \mathcal{F}_T] = -\chi_T$, from which both directions follow (Appendix B.4). The condition is imposed history by history. Preserving the unconditional expectation alone permits cancellation across histories and does not require $\chi_T = 0$ (Appendix C).

Theorem 2 (Stopping failure feedback in a complete GRPO group)**.** *Conditionally on the initial scene X, assume $m \ge 2$ independent rollouts, each with X in its local initial information and a local stopping time T_i satisfying $\chi^{(i)}_{T_i} = 0$. Let $\mathcal{K} = \bigvee_i \mathcal{K}^{(i)}_{T_i}$, and assume positive token counts N_i are $\mathcal{K}$-measurable. The square-integrable group estimators satisfy*

$$G_g = \frac{1}{m} \sum_i \frac{A_i}{N_i} S^{(i)}_H, \qquad G_{g,T} = \frac{1}{m} \sum_i \frac{A_i}{N_i} \widetilde{S}^{(i)}_{T_i} = \mathbb{E}[G_g \mid \mathcal{K}]. \tag{6}$$

Their expectations agree. With $D_g = G_g - G_{g,T}$,

$$\mathrm{Cov}(G_g) - \mathrm{Cov}(G_{g,T}) = \mathbb{E}[D_g D_g^\top] \succeq 0. \tag{7}$$

The same projection and covariance identity hold conditional on any positive-probability acceptance event determined by X and the group outcomes.

Since $\mathcal{K}$ contains all outcomes and the token counts N_i are $\mathcal{K}$-measurable, the coefficients A_i/N_i factor out of the conditional expectation. This accommodates the dependence among group advantages (Appendix B.5).

The theorem preserves the *population expectation* $\mu = \mathbb{E}[G_g]$. Group-relative weighting and token normalization can make μ differ from ∇J. Figure 2a illustrates the projection by grouping full-gradient realizations by retained information, with hollow markers showing conditional expectations.

The finite-batch estimate is the sample average $\overline{G} = n^{-1} \sum_j G_{g_j}$. On the same n independent groups, ideal masking gives $\overline{G}_T = \overline{G} - \overline{D}$, where $\overline{D} = n^{-1} \sum_j D_{g_j}$. Since $\mathbb{E}[\overline{D}] = 0$, the two estimates share expectation μ even when their realized values differ. Figure 2b shows the contributions forming a batch estimate. Masking zeros the orange suffix terms while preserving the original GRPO coefficients and normalization. Large markers show the full and masked batch estimates, and the star marks μ. Ideal masking reduces mean-squared error (MSE) relative to μ by $\mathbb{E}\|D_g\|^2/n$ on average over batches, while individual batch errors can increase or decrease.

Corollary 3 (Failure certification as a special case)**.** *If $P_T = 0$ on $\{T < H\}$, then $\chi_T = 0$, $\widetilde{S}_T = S_T$, and $\mathcal{K}_T = \mathcal{F}_T$ up to null events. The group theorem therefore includes local failure-certifying stopping times.*

For stopping times satisfying $\chi_T = 0$, earlier stopping has no larger covariance for the same coefficients and denominators (Appendix B.7). Certification is sufficient but unnecessary, since an action-independent outcome can have $P_T > 0$ and zero remaining credit (Appendix C).

Selecting candidate boundaries. Failure certification motivates looking for errors after which recovery is unlikely within the remaining budget. Visible errors and recovery observations serve as cues for candidate boundaries, whose statistical effects are characterized below.

2.4 FROM IDEAL STOPS TO RETROSPECTIVE BOUNDARIES

Later observations can reveal whether an apparent error is followed by recovery. A boundary selected from the complete rollout can therefore depend on events beyond the retained prefix. Such

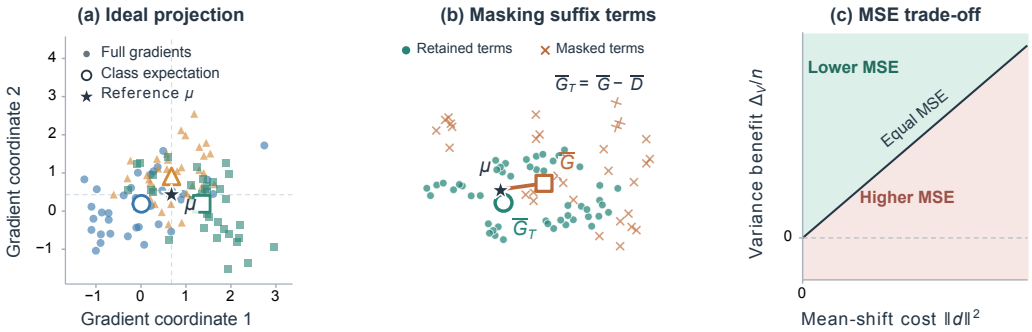


Figure 2: **Feedback truncation and gradient estimation.** Schematic views of (a) ideal conditional projection, (b) a finite-batch estimate before and after suffix masking, and (c) the trade-off between variance reduction and expectation shift.

selection can change the estimator's expectation even when each fixed candidate time has zero remaining credit (Appendix C). We quantify this effect through a conditional projection residual.

Fix the policy, boundary rule, and reference sampling distribution. Let G be the full group estimator, G_B its masked version, and $D_B = G - G_B$. Let $\mathcal{K}_B$ contain the scene, all outcomes, the selected boundaries, actual denominators, retained failure prefixes, and complete success records. Thus G_B is measurable in $\mathcal{K}_B$. The boundaries themselves must be included because they can convey future information. Define

$$\eta_B = \mathbb{E}[D_B \mid \mathcal{K}_B], \qquad \epsilon_B^2 = \mathbb{E}\|\eta_B\|^2. \tag{8}$$

Proposition 4 (Projection residual of a retrospective rule)**.** *For square-integrable G and G_B,*

$$G_B = \mathbb{E}[G \mid \mathcal{K}_B] - \eta_B. \tag{9}$$

Consequently $\eta_B = 0$ is equivalent to this projection property and implies equal expectations and covariance nonincrease. For nonzero residual, write $V_B = \operatorname{tr}\operatorname{Cov}(G_B)$, $V_D = \operatorname{tr}\operatorname{Cov}(D_B)$, and $\Delta_V = \operatorname{tr}\operatorname{Cov}(G) - V_B$. Then

$$\|\mathbb{E}[G_B] - \mathbb{E}[G]\|^2 \le \epsilon_B^2, \qquad \Delta_V \ge V_D - 2\sqrt{V_B \epsilon_B^2}. \tag{10}$$

The proposition follows from conditional expectation and Cauchy–Schwarz (Appendix B.8) under a fixed unfiltered or accepted-group law. The ideal stops of Theorem 2 satisfy $\eta_B = 0$, and the residual measures a rule's departure from their conditional projection property.

Statistical benefit and truncation cost. For a finite batch, estimation accuracy depends on the balance between variance reduction and a shift in expectation. Write $G_w = G - D$ for a masked estimator and its removed part, so that G_B and D_B above are one instance. Set $\mu = \mathbb{E}[G]$ and $d = \mathbb{E}[D]$. Here bias is measured relative to the reference expectation μ.

With $\Delta_V = \operatorname{tr}\operatorname{Cov}(G) - \operatorname{tr}\operatorname{Cov}(G_w)$, the standard bias–variance decomposition gives, for averages of n independent copies,

$$\mathbb{E}\|\overline{G}_w - \mu\|^2 - \mathbb{E}\|\overline{G} - \mu\|^2 = \|d\|^2 - \frac{\Delta_V}{n}. \tag{11}$$

For an ideal stop in Theorem 2, $d = 0$ and the MSE reduction is $\mathbb{E}\|D_g\|^2/n$, which is strict when the deleted component is nonzero. For a general boundary, strict MSE improvement is equivalent to $\Delta_V > n\|d\|^2$, the region above the diagonal in Figure 2(c) (Proposition 7 in the appendix).

Equation 11 balances variance reduction against the squared expectation shift at a given batch size. Appendix B.11 compares complete-rollout selection with rules based on the observed prefix. The next section constructs boundaries from visible errors and recovery.

3 MARTINGALE GRPO

Section 2 identifies conditions for preserving the reference gradient expectation when removing a failure suffix. The failure-certification case (Corollary 3) motivates using visible errors and recovery as cues for candidate boundaries. MART-GRPO retains feedback on localized errors and uses subsequent recovery observations to select the suffix to mask (Figure 3). A frozen VLM reviews each complete failed rollout and localizes the last visible error without observed recovery. A fixed mapping converts this interval into a failure-prefix mask for GRPO.

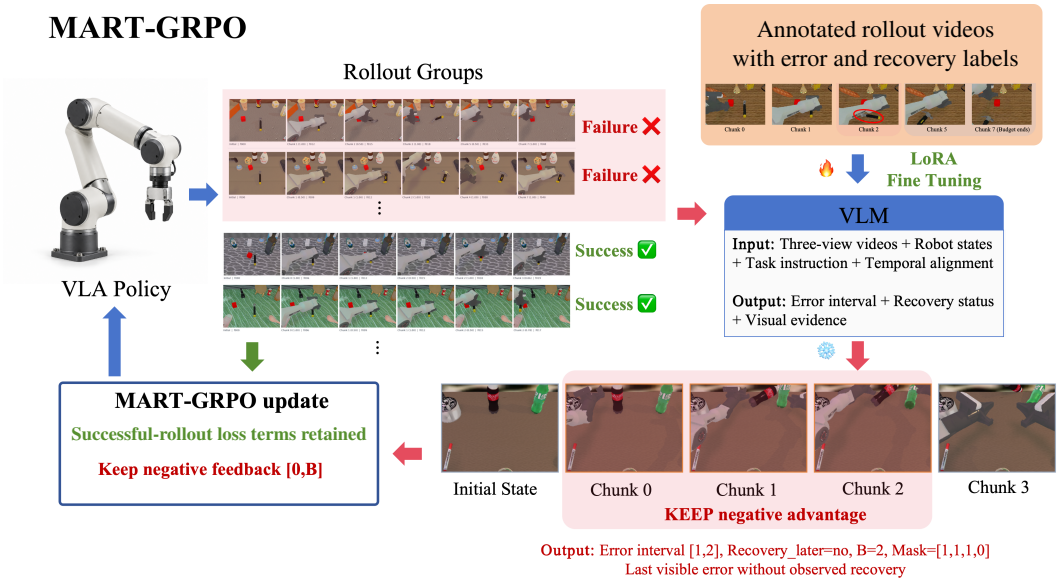


Figure 3: **MART-GRPO: from error localization to feedback truncation.** A VLM is fine-tuned offline on event annotations, then frozen during RL. It localizes the last visible error without observed recovery in complete failed rollouts. The interval's right endpoint sets B: retain negative feedback through B and mask the suffix. Success terms and original token denominators are preserved.

3.1 LEARNING TO LOCALIZE FAILURE EVENTS

Event supervision. The VLM input x combines task instructions, three synchronized camera views, proprioceptive observations, and frame-to-chunk alignment. Structured targets record visible outcomes, error types, localization intervals, supporting frames, and observed recovery. The interval brackets the onset of the target error, since its exact chunk is uncertain.

The corpus contains 1,024 failed rollout videos with event annotations, supplemented by successful reference videos. Training, validation, and test partitions have disjoint task–scene identities. Annotation details and split counts are documented in Appendix E.

VLM fine-tuning. A single LoRA adapter (Hu et al., 2022) is trained with the Qwen3.8-27B vision-language backbone (Qwen Team, 2026) frozen. For each structured response y, we average the negative log likelihood over response tokens, then over training examples:

$$\mathcal{L}_{\text{event}}(\phi) = -\frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \frac{1}{|y|} \sum_{v=1}^{|y|} \log p_\phi(y_v \mid x, y_{<v}). \tag{12}$$

We select the checkpoint by validation task-macro interval intersection-over-union (IoU), breaking ties by lower response loss and then earlier update, and freeze it during RL. Appendix E.3 gives the training and evaluation protocol.

3.2 Selecting the feedback boundary

For a failed rollout, the boundary uses the following fields from the VLM's structured response:

$$\widehat{e}_i = f_\phi(x_i) = (o_i, \mathrm{type}_i, [\ell_i, u_i], r_i, \mathcal{I}_i). \tag{13}$$

Here o_i is the predicted visible outcome, r_i records subsequent recovery from the localized error, and $\mathcal{I}_i$ identifies supporting frames for inspection. Response validation checks the schema and requires any non-null interval to contain two integer indices satisfying $0 \le \ell_i \le u_i < K_i$. Invalid responses and failed inference requests raise implementation errors (Appendix E.1). For a valid response, truncation requires a predicted failure, a specific error type, a non-null interval, and an explicit report of no recovery. Other valid responses retain full feedback, which we call a semantic fallback:

$$B_i = \begin{cases} u_i, & \text{a localized event passes the checks,} \\ K_i - 1, & \text{a semantic fallback applies.} \end{cases} \tag{14}$$

For a qualifying event, $[\ell_i, u_i]$ localizes the error onset, while feedback is retained on the full prefix $0, \ldots, B_i$ with $B_i = u_i$. This choice includes every candidate onset chunk. For example, $[1, 2]$ in a four-chunk failed rollout retains chunks $0, 1, 2$, giving mask $(1, 1, 1, 0)$ under Equation 2.

3.3 Policy optimization

At each policy update, rollouts execute to their terminal condition or budget. Following Full GRPO, we retain groups containing both successes and failures and collect further groups to fill the batch. Terminal rewards determine group advantages. The frozen VLM and fixed boundary mapping determine masks for the new failed rollouts, while successes bypass event inference and retain all their loss terms.

Let $k(u)$ denote the chunk containing token u, and $\rho_{i,u}(\theta)$ the token probability ratio relative to the collection policy. For M rollouts and clipping limits $1 - \varepsilon_-, 1 + \varepsilon_+$, we optimize

$$\mathcal{J}_w(\theta) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N_i} \sum_{u \in \mathcal{V}_i} \min\Big\{ \rho_{i,u} \widetilde{A}_{i,k(u)},\ \mathrm{clip}(\rho_{i,u}, 1 - \varepsilon_-, 1 + \varepsilon_+) \widetilde{A}_{i,k(u)} \Big\}. \tag{15}$$

Masks remain fixed and detached throughout each PPO update, and the original valid-token counts N_i are retained. Deployment uses the actor alone. Appendix D.2 defines the retained coefficient mass used to quantify this feedback.

Connection to the analysis. The VLM uses later observations to interpret errors and recovery, so B is retrospective. Proposition 4 characterizes these masks through a conditional projection residual. For these boundaries, the ideal projection guarantee requires $\eta_B = 0$, an additional condition beyond visible-event localization. Policy comparisons in Section 4 assess learning under the clipped PPO objective.

4 Evaluation

We evaluate VLM event localization, VLA policy performance, and two ablations on the amount and temporal placement of failure feedback.

4.1 VLM event localization

We evaluate 220 unrecovered test failures with annotated intervals from the seven policy tasks, using the task–scene-disjoint split in Section 3. Interval intersection-over-union (IoU) compares discrete chunk sets. Endpoint accuracy measures agreement within a given chunk distance of the annotated right endpoint.

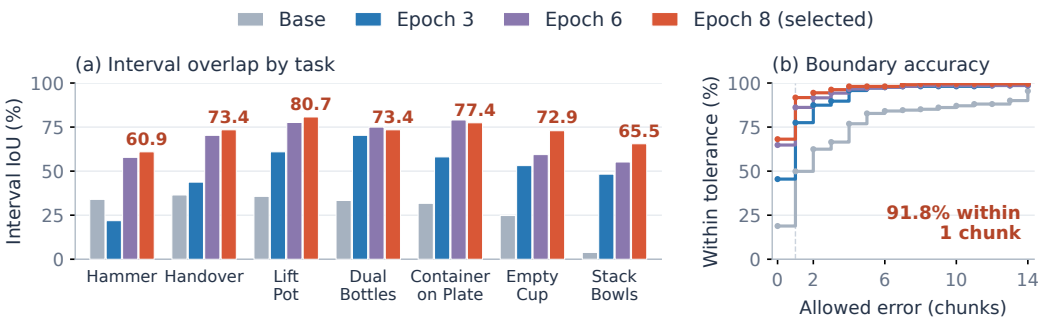


Figure 4: **VLM event localization.** Per-task interval IoU (a) and cumulative endpoint accuracy (b) on test failures. Epoch 8 is validation-selected.

Table 1: **Seven-task policy performance.** Success (%) on 128 training-seed (Seen) and 128 held-out-seed (Held-out) episodes. All pools both splits. RL scores are at update 40 except Best, the maximum All over updates $5, 10, \ldots, 40$. SFT is the initial score. Δ_{40} and Δ_{Best} give MART-GRPO gains over Full GRPO (percentage points). Means and differences are computed before rounding.

	SFT	Full GRPO (Li et al., 2026a)				MART-GRPO					
Task	All	Seen	Held-out	All	Best	Seen	Held-out	All	Best	Δ_{40}	Δ_{Best}
Hammer	30.47	78.12	70.31	74.22	74.61	82.81	79.69	**81.25**	**81.25**	+7.03	+6.64
Handover	21.09	38.28	31.25	34.77	35.55	40.62	46.88	**43.75**	**44.53**	+8.98	+8.98
Lift pot	5.47	60.16	56.25	58.20	58.20	57.03	60.16	**58.59**	**58.59**	+0.39	+0.39
Dual bottles	20.70	50.00	39.84	44.92	44.92	53.91	43.75	**48.83**	**51.17**	+3.91	+6.25
Container on Plate	87.11	92.97	94.53	**93.75**	**94.92**	92.97	90.62	91.80	**94.92**	-1.95	+0.00
Empty cup	80.86	91.41	86.72	89.06	90.62	92.19	90.62	**91.41**	**91.41**	+2.34	+0.78
Stack bowls	39.06	63.28	57.03	60.16	63.28	63.28	66.41	**64.84**	**67.19**	+4.69	+3.91
Mean (7 tasks)	40.68	67.75	62.28	65.01	66.02	68.97	68.30	68.64	69.87	+3.63	+3.85

Both metrics use predictions before semantic fallback, score missing intervals as zero, and average within tasks and then across tasks. Validation IoU selects update 304 (epoch 8). These scores measure exploratory agreement with the reference annotations (Appendix E.3).

Figure 4 compares the base model with epochs 3, 6, and 8. Task-macro IoU rises from 28.5% to 50.9%, 67.8%, and 72.1%, respectively. The selected checkpoint achieves 68.1% exact endpoint agreement and 91.8% agreement within one chunk. Per-task values appear in Table 8.

4.2 VLA POLICY PERFORMANCE

Setup. We compare SFT, Full GRPO, and MART-GRPO on seven RoboTwin 2.0 tasks (Chen et al., 2025), using the action-token implementation of OpenVLA-OFT from SimpleVLA-RL (Kim et al., 2025a; Li et al., 2026a). Both RL methods start from the same task-specific SFT checkpoint. Each update uses 64 accepted groups of eight rollouts, mixed-outcome filtering with adaptive refill, global PPO minibatch 128, and the original token normalization. The actor receives a head-camera view and proprioception, while the frozen VLM reviews three views. Experiments run on servers equipped with eight H100 GPUs. Greedy evaluation uses 128 training-seed (Seen) and 128 held-out-seed episodes under the same task configuration every five updates. The primary endpoint is update 40. Full settings and scheduled evaluations appear in Appendices D.1 and F.1.

Results. At update 40, Table 1 shows mean success rising from 65.01% to 68.64% (+3.63 points), with gains of 1.23 points on seen seeds and 6.03 on held-out seeds. Six tasks improve, led by Handover (+8.98) and Hammer (+7.03). Mean per-task best scores rise from 66.02% to 69.87%. Container on Plate starts at 87.11% SFT success, leaving limited headroom, and both methods peak at 94.92%. On Container on Plate, MART's pooled endpoint success is 1.95 points below Full GRPO, with the difference confined to held-out seeds.

4.3 FEEDBACK ABLATIONS

We conduct ablations on two tasks, Hammer and Handover. Two ablations use statistics frozen from all 40 updates of the corresponding MART run. Historical MART retains 49.2% and 32.7% of negative coefficient mass (Equation 67), respectively, measuring feedback weight rather than gradient norm. *Uniform scaling* uses these ratios (0.4921 and 0.3269) to scale every failed-trajectory advantage. *Random prefix* samples boundaries independently from the historical MART boundary distribution. Both ablations preserve success weights and original token denominators, use no online VLM, and follow the same 40-update budget and evaluation schedule. Protocol details appear in Appendix F.2.

Table 2 reports success at updates 10, 20, 30, and 40, and Figure 5 compares final results alongside the SFT references. Uniform scaling reaches 77.73% on Hammer and 40.62% on Handover, improving over Full GRPO. MART remains higher by 3.52 and 3.12 points, respectively. Random prefix reaches 71.88% and 41.41%, trailing MART by 9.38 and 2.34 points. VLM-guided truncation thus achieves higher success than simply reducing failure penalties on both tasks. Its advantage over random prefixes further supports selecting boundaries from visible events.

Table 2: **Feedback ablations.** Success (%) pooled across both seed splits at updates 10, 20, 30, and 40. Column bests are bold. Reported differences are computed before rounding.

Method	Failure feedback	Hammer				Handover			
		10	20	30	40	10	20	30	40
Full GRPO	Full trajectory	48.05	53.91	65.23	74.22	**32.03**	31.25	35.55	34.77
Uniform scaling	Uniformly scaled	**51.56**	63.28	65.23	77.73	29.69	33.98	37.89	40.62
Random prefix	Random prefix	51.17	67.19	75.00	71.88	25.78	32.03	38.67	41.41
MART-GRPO	Event-guided prefix	49.61	**69.14**	**78.12**	**81.25**	25.39	**35.16**	**42.58**	**43.75**

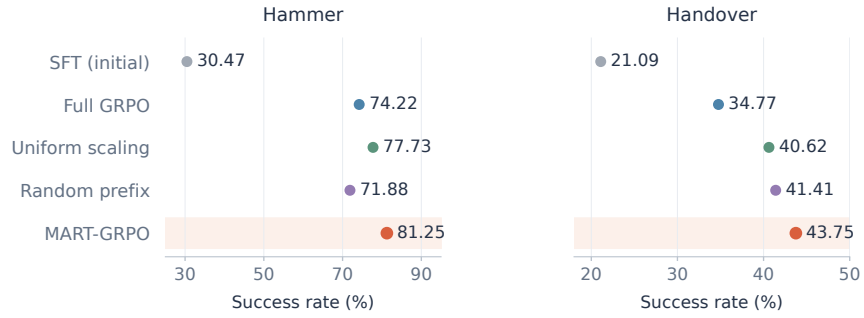


Figure 5: **Final performance on Hammer and Handover.** Success after 40 updates, with SFT references and task-specific horizontal scales.

5 CONCLUSION AND LIMITATIONS

We study where failure feedback should end in VLA reinforcement learning. Two martingales characterize remaining conditional credit and identify ideal stops that preserve the reference group-gradient expectation without increasing covariance. Guided by this analysis, MART-GRPO selects feedback boundaries from visible errors and recovery in complete rollout videos. Retrospective residuals characterize departures of these boundaries from the ideal projection. Policy evaluations show success-rate improvements over Full GRPO across multiple tasks. On both ablation tasks, VLM-guided truncation achieves higher endpoint success than uniform penalty scaling and random prefix selection. A key open question is how reliably visible events track remaining credit as the policy evolves. Future work will examine this relationship across policy stages and develop prefix-based event predictors.

AI USE STATEMENT

Generative AI tools assisted method implementation, experiment planning, and result analysis. They also supported language polishing, figure preparation, and checks of mathematical proofs. The authors reviewed the AI-assisted work and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Appendix B supplies the stopping and retrospective-projection proofs, together with the hindsight decomposition and the approximate-stopping bound. Finite-model checks supplement the proofs. Appendix D.1 gives the policy training and evaluation settings, and Appendix E documents the boundary mapping and the annotation corpus. Appendix F.1 provides scheduled policy evaluations and feedback statistics, and Appendix F.2 details the ablation protocol.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/16105fb9cc614fc29e1bda00dab60d41-Abstract.html.

Vaidehi Bagaria, Nikshep Grampurohit, and Pulkit Verma. Learn where outcomes diverge: Efficient VLA RL via probabilistic chunk masking. *arXiv preprint arXiv:2605.16154*, 2026. URL https://arxiv.org/abs/2605.16154.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. URL https://arxiv.org/abs/2506.18088.

Yanzhe Chen, Zhijun Cao, and Mike Zheng Shou. Where success breaks: Failure-boundary learning for robust vision-language-action models. *arXiv preprint arXiv:2609.06114*, 2026. URL https://arxiv.org/abs/2609.06114.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034. URL https://sites.math.duke.edu/~rtd/PTE/PTE5_011119.pdf.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-Group Policy Optimization for LLM Agent Training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408, 2025. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/195f15384c2a79cedf293e4a847ce85c-Abstract.html.

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.

Doo Hwan Hwang and Kee-Eung Kim. PS-PPO: Prefix-sampling PPO for critic-free RLHF. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.29758.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, 2025a. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. SimpleVLA-RL: Scaling VLA training via reinforcement learning. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cbfbcb4da14235bd69b134070898ae9d-Abstract-Conference.html.

Zihang Li, Rui Zhou, Yingcheng Shi, Wenhan Yu, Zhewen Tan, Zixiang Liu, Zeming Li, Binhua Li, Yongbin Li, Tong Yang, and Jieping Ye. ESPO: Early-stopping proximal policy optimization. *arXiv preprint arXiv:2605.29860*, 2026b. URL https://arxiv.org/abs/2605.29860.

Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can RL bring to VLA generalization? an empirical study. In *Advances in Neural Information Processing Systems*, volume 38, pp. 97121–97151, 2025a. doi: 10.52202/085713-3248. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8c0fabe372177d2aded596be2d3b4544-Abstract-Conference.html.

Zeyi Liu, Arpit Bahety, and Shuran Song. REFLECT: Summarizing robot experiences for failure explanation and correction. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3468–3484. PMLR, 2023. URL https://proceedings.mlr.press/v229/liu23g.html.

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. *arXiv preprint arXiv:2503.20783*, 2025b. URL https://arxiv.org/abs/2503.20783.

Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint arXiv:2505.18719*, 2025. URL https://arxiv.org/abs/2505.18719.

Thomas Mesnard, Théophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Éric Moulines, Marcus Hutter, Lars Buesing, and Rémi Munos. Counterfactual credit assignment in model-free reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664. PMLR, 2021. URL https://proceedings.mlr.press/v139/mesnard21a.html.

Qwen Team. Qwen3.8-27B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7a7f6cc5dc2a84fb4edf0feb8e5cfd50-Abstract-Conference.html.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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.

Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. RoboCLIP: One demonstration is enough to learn robot policies. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55681–55693, 2023. doi: 10.52202/075280-2430. URL https://papers.nips.cc/paper_files/paper/2023/hash/ae54ce310476218f26dd48c1626d5187-Abstract-Conference.html.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://papers.nips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Minghao Tian, Yunfei Xie, and Chen Wei. How off-policy can GRPO be? Mu-GRPO for efficient LLM reinforcement learning. *arXiv preprint arXiv:2605.17570*, 2026. URL https://arxiv.org/abs/2605.17570.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://link.springer.com/article/10.1007/BF00992696.

Zhengyang Yan, Junhao Li, Fangqi Zhu, Zijun Wang, Quanxin Shou, Yikun Miao, Xiaoyi Pang, Zicong Hong, and Song Guo. RedFlow: Redirect failure into action-level corrections for flow-matching VLA policy. *arXiv preprint arXiv:2607.27782*, 2026. URL https://arxiv.org/abs/2607.27782.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244, 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Yao Zhou, Hang Gao, Fengge Wu, Changwen Zheng, and Wenwen Qiang. Temporal GRPO: Beyond trajectory-level credit in vision-language-action reinforcement learning. *arXiv preprint arXiv:2608.13026*, 2026. URL https://arxiv.org/abs/2608.13026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

Appendix

Contents

A. Related Work . . . 14
B. Proofs and the Scope of the Martingale Argument . . . 15
B.1 Notation and regularity . . . 15
B.2 Standard martingale identities . . . 15
B.3 Remaining conditional credit . . . 16
B.4 Proof of Proposition 1 . . . 17
B.5 Proof of Theorem 2 . . . 17
B.6 Proof of Corollary 3 . . . 18
B.7 Ordering ideal stopping rules . . . 19
B.8 Proof of Proposition 4 . . . 19
B.9 Exact truncation tradeoff . . . 20
B.10 Approximate stopping and conditional bias . . . 20
B.11 Comparing retrospective and prefix boundaries . . . 21
B.12 Partial suffix removal . . . 22
C. Examples Clarifying the Projection Conditions . . . 23
D. Policy Training and Evaluation . . . 24
D.1 Training and evaluation settings . . . 24
D.2 Feedback accounting and normalization . . . 24
E. VLM Event Model: Mapping, Corpus, and Localization . . . 25
E.1 VLM inputs, prompting, and boundary mapping . . . 25
E.2 Annotated video corpus . . . 27
E.3 Event model training . . . 28
F. Policy Results and Feedback Ablations . . . 29
F.1 Seven-task policy results . . . 29
F.2 Feedback ablations . . . 30

A Related Work

VLA models and reinforcement learning. Vision-language pretraining supports robot control in RT-2 and OpenVLA (Zitkovich et al., 2023; Kim et al., 2025b), while π_0 uses flow-based continuous action generation (Black et al., 2025). OpenVLA-OFT studies efficient supervised adaptation (Kim et al., 2025a). VLA-RL and SimpleVLA-RL develop RL pipelines for pretrained policies (Lu et al., 2025; Li et al., 2026a), with further work examining RL generalization (Liu et al., 2025a). We select which action chunks receive terminal failure feedback while retaining the policy architecture and environment reward.

Temporal credit assignment. RUDDER redistributes delayed rewards through return prediction (Arjona-Medina et al., 2019). Hindsight and counterfactual credit assignment use future information to construct action-credit estimates or baselines (Harutyunyan et al., 2019; Mesnard et al., 2021). Temporal GRPO reconstructs stage-relative advantages using task-stage detection and rollout alignment, while GiGPO estimates local credit at repeated anchor states (Zhou et al., 2026; Feng et al., 2025). These methods build a new credit signal. We keep the existing trajectory-level advantage and change only the actions it reaches.

Truncating negative feedback. A second group changes which update terms survive. PCM samples chunks using phase-level success–failure divergence to allocate gradient computation (Bagaria et al., 2026). μ-GRPO uses importance-ratio triggers to veto suffix updates, motivated by off-policy prefix mismatch (Tian et al., 2026). ESPO uses a regret surrogate to terminate failing rollouts early and assign terminal failure (Li et al., 2026b), whereas PS-PPO samples update prefixes with importance weighting to preserve the reference gradient expectation and study the variance–computation tradeoff (Hwang & Kim, 2026). MART-GRPO masks failed-rollout suffixes after complete execution, selecting boundaries from visible errors and recovery. Existing group advantages and loss normalization are retained without additional truncation correction weights.

Vision-language feedback and corrective learning. Vision-language feedback has been used to construct rewards from image–text similarity (Rocamonde et al., 2024), video or textual demonstrations (Sontakke et al., 2023), and process-level predictions (Lu et al., 2025). REFLECT summarizes robot experiences for failure explanation and correction (Liu et al., 2023). DLS uses privileged simulator states for progress-localized signals and directional shaping in flow-based VLA adaptation (Chen et al., 2026). RedFlow estimates chunk quality and derives corrective action targets from successful matches in offline data (Yan et al., 2026). In MART-GRPO, the fine-tuned VLM identifies the last visible error without observed recovery in complete videos. Its interval selects where an existing negative advantage ends, while the scalar task reward remains unchanged.

Gradient variance and normalization. REINFORCE and the policy-gradient theorem give the score-function foundation (Williams, 1992; Sutton et al., 1999). Baselines and value-based advantage estimates address gradient variance (Greensmith et al., 2004; Schulman et al., 2016), while GRPO and RLOO use groups of sampled outcomes (Shao et al., 2024; Ahmadian et al., 2024). Dr. GRPO analyzes normalization bias, and DAPO studies token-level aggregation and dynamic sampling (Liu et al., 2025b; Yu et al., 2025). These loss and normalization choices affect the reference update itself. We characterize failure-only truncation of complete GRPO groups through optional sampling and conditional-expectation projection (Durrett, 2019). The analysis covers dependent advantages, token denominators, and outcome filtering. Under the stopping and measurability conditions, the projection preserves the reference expectation without increasing covariance. This guarantee concerns the reference estimator before PPO clipping (Schulman et al., 2017). Conditional projection residuals characterize expectation and variance changes under retrospective boundary selection.

B Proofs and the Scope of the Martingale Argument

B.1 Notation and regularity

We use the fixed-policy setting of Section 2. A trajectory has finite horizon H and binary terminal outcome R. Its distribution factors into an initial distribution, policy probabilities, and environment transitions, with only the policy factors depending on θ. Differentiation is assumed to commute with expectation. For continuous actions, the policy has a differentiable density with common support. Scores are square-integrable, and terminal padding uses parameter-independent dummy actions with zero score.

Let h_k denote the complete pre-action history used in the analysis, so that $\mathcal{F}_k = \sigma(h_k)$, and let $\mathcal{F}_H$ include R. The notation $\pi_\theta(a_k \mid h_k)$ denotes the action law conditional on this history, even when the policy uses only its observable component. Additional environment information may be included in h_k provided that the conditional action law remains π_θ. Thus $P_k = \mathbb{E}[R \mid \mathcal{F}_k]$ conditions on the analytical history, while the policy input specifies which part of that history is available to the agent. Under these assumptions, Z_k is $\mathcal{F}_{k+1}$-measurable and satisfies $\mathbb{E}[Z_k \mid \mathcal{F}_k] = 0$.

B.2 Standard martingale identities

Proposition 5 (Outcome innovations and local credit)**.** *Under the fixed-policy score assumptions in Section 2, $(P_k)_{k=0}^{H}$ is a bounded martingale. For $J(\theta) = \mathbb{E}[R]$ and every stopping time $T \le H$,*

$$\begin{aligned} \nabla J(\theta) &= \sum_{k=0}^{H-1} \mathbb{E}[Z_k \Delta P_k], \\ \mathbb{E}\left[\sum_{k=T}^{H-1} (\Delta P_k)^2 \,\middle|\, \mathcal{F}_T\right] &= P_T(1 - P_T). \end{aligned} \tag{16}$$

Proof. Since $0 \le R \le 1$, each P_k is integrable and lies in $[0, 1]$. The tower property gives

$$\mathbb{E}[P_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[\mathbb{E}[R \mid \mathcal{F}_{k+1}] \mid \mathcal{F}_k] = P_k. \tag{17}$$

Moreover $P_H = R$, because the terminal outcome is observed at H. The likelihood-ratio identity yields

$$\nabla J(\theta) = \mathbb{E}\left[R \sum_{k=0}^{H-1} Z_k\right]. \tag{18}$$

For each k, measurability of Z_k with respect to $\mathcal{F}_{k+1}$ and the score identity imply

$$\mathbb{E}[RZ_k] = \mathbb{E}[P_{k+1}Z_k], \tag{19}$$
$$\mathbb{E}[P_k Z_k] = \mathbb{E}[P_k \mathbb{E}[Z_k \mid \mathcal{F}_k]] = 0. \tag{20}$$

Subtracting gives the first identity by a conditional-expectation substitution at the fixed policy parameter.

For a deterministic t, increments ΔP_k after t are mutually orthogonal conditional on $\mathcal{F}_t$. For $t \leq j < k$, the tower property gives

$$\mathbb{E}[\Delta P_j \Delta P_k \mid \mathcal{F}_t] = \mathbb{E}[\Delta P_j \mathbb{E}[\Delta P_k \mid \mathcal{F}_k] \mid \mathcal{F}_t] = 0. \tag{21}$$

The same equality at a bounded stopping time follows by partitioning over $\{T = t\}$. Expanding $R - P_T = \sum_{k=T}^{H-1} \Delta P_k$ therefore gives

$$\mathbb{E}\left[\sum_{k=T}^{H-1} (\Delta P_k)^2 \,\middle|\, \mathcal{F}_T\right] = \mathbb{E}[(R - P_T)^2 \mid \mathcal{F}_T] = P_T - P_T^2, \tag{22}$$

where $R^2 = R$ is used in the last step. On any event where $P_T = 0$, nonnegativity of R implies $R = 0$ almost surely conditional on $\mathcal{F}_T$. □

Action-conditioned credit. To interpret ΔP_k, define $Q^\pi(h, a) = \mathbb{E}[R \mid h, a]$ for the undiscounted terminal-success objective. Since $\mathcal{F}_k = \sigma(h_k)$ and the action a_k is observed by $k + 1$, the tower property gives

$$\mathbb{E}[P_{k+1} \mid \mathcal{F}_k, a_k] = \mathbb{E}[R \mid \mathcal{F}_k, a_k] = Q^\pi(h_k, a_k). \tag{23}$$

Subtracting the current success probability P_k yields the local action advantage

$$\mathcal{A}^\pi(h_k, a_k) = Q^\pi(h_k, a_k) - P_k = \mathbb{E}[\Delta P_k \mid \mathcal{F}_k, a_k]. \tag{24}$$

Conditioning on $\mathcal{F}_k$ gives $\mathbb{E}[\mathcal{A}^\pi(h_k, a_k) \mid \mathcal{F}_k] = 0$. Because Z_k is a function of (h_k, a_k), its expected credit is $\mathbb{E}[Z_k \mathcal{A}^\pi(h_k, a_k)] = \mathbb{E}[Z_k \Delta P_k]$. This connects each term in the gradient identity to the action's expected effect on terminal success.

Policy-relative success probability. The process P_k is defined by the collection policy, the analytical history, and the remaining budget. Accordingly, $P_T = 0$ certifies failure under that continuation law. Its first-order interpretation uses the common policy support assumed above.

B.3 Remaining conditional credit

The same score substitution can be conditioned on an earlier history. For deterministic $t \leq k$, score centering and the measurability of Z_k at $k + 1$ give

$$\begin{aligned} \mathbb{E}[RZ_k \mid \mathcal{F}_t] &= \mathbb{E}[P_{k+1}Z_k \mid \mathcal{F}_t], \\ \mathbb{E}[P_k Z_k \mid \mathcal{F}_t] &= 0. \end{aligned} \tag{25}$$

Subtracting the two equalities and summing over $k \geq t$ proves Equation 4. Partitioning over $\{T = t\}$ extends this identity to a bounded stopping time. Optional sampling centers the remaining score sum, so

$$\begin{aligned} \mathbb{E}[S_H - S_T \mid \mathcal{F}_T] &= 0, \\ \chi_T &= \mathbb{E}[(R - P_T)(S_H - S_T) \mid \mathcal{F}_T]. \end{aligned} \tag{26}$$

Thus χ_T is the conditional covariance between terminal success and the remaining score sum. Since it sums vector contributions, later action credits may cancel.

Continuation sensitivity and initial conditioning. For a fixed complete prefix h_t, let $V_\vartheta(h_t)$ be the success probability when future actions are sampled from π_ϑ. The recorded prefix and conditional environment law are held fixed. Under the differentiation assumptions above, the suffix likelihood-ratio identity gives

$$\begin{aligned}\nabla_\vartheta V_\vartheta(h_t)|_{\vartheta=\theta} &= \mathbb{E}_\theta\left[R\sum_{k=t}^{H-1} Z_k \,\middle|\, h_t\right] \\ &= \chi_t(h_t).\end{aligned} \tag{27}$$

This derivative measures local sensitivity at the current policy parameter, so $\chi_t = 0$ can arise through cancellation or stationarity. At the initial time, $\mathcal{F}_0 = \sigma(X)$ makes χ_0 the gradient conditional on scene X. Averaging over the parameter-independent initial distribution then gives the population gradient $\mathbb{E}[\chi_0] = \nabla J$.

B.4 Proof of Proposition 1

Fix a stopping time T for the local filtration. To retain the stopped history on failure and the complete history on success, define

$$\mathcal{K}_T = \sigma(\mathcal{F}_T,\, R,\, \{R\mathbf{1}_A : A \in \mathcal{F}_H\})\,. \tag{28}$$

On failure, the last collection is identically zero, leaving the stopped history and failure label. On success, it reveals the complete trajectory. Therefore $\widetilde{S}_T = (1-R)S_T + RS_H$ is $\mathcal{K}_T$-measurable. This information set records what is retained after execution, while the stopping decision itself is determined by the local prefix.

Proof. Bounded optional sampling (Durrett, 2019, Section 4.8) yields

$$\mathbb{E}[(1-R)(S_H - S_T) \mid \mathcal{F}_T] = -\chi_T. \tag{29}$$

Suppose $\chi_T = 0$. At histories with $P_T < 1$,

$$\mathbb{E}[S_H - S_T \mid \mathcal{F}_T, R=0] = \frac{-\chi_T}{1-P_T} = 0. \tag{30}$$

Histories with $P_T = 1$ have no failure branch. Hence the removed score has zero conditional mean given $\mathcal{K}_T$ on failure, and is identically zero on success. Since $S_H - \widetilde{S}_T = (1-R)(S_H - S_T)$, this proves

$$\mathbb{E}[S_H \mid \mathcal{K}_T] = \widetilde{S}_T. \tag{31}$$

Conversely, the projection implies $\mathbb{E}[S_H - \widetilde{S}_T \mid \mathcal{K}_T] = 0$. Conditioning on $\mathcal{F}_T \subseteq \mathcal{K}_T$ and using Equation 29 gives $\chi_T = 0$. □

Weighted estimators and scope of necessity. For a bounded $\mathcal{K}_T$-measurable coefficient C, the condition gives

$$\begin{aligned} G = CS_H, \qquad G_T = C\widetilde{S}_T, \\ G_T = \mathbb{E}[G \mid \mathcal{K}_T]. \end{aligned} \tag{32}$$

Square integrability then gives expectation preservation and covariance reduction. Necessity in Proposition 1 concerns the unweighted score, since a zero coefficient can hide nonzero conditional credit in a weighted estimator. Unconditional expectation preservation is weaker because conditional shifts can cancel across histories.

B.5 Proof of Theorem 2

Let X denote the shared initial scene. The complete trajectory sigma-fields are conditionally independent given X. Each local filtration contains X and that trajectory's history, without revealing sibling futures. Each $T_i \le H$ is a stopping time for its local filtration, with $\chi^{(i)}_{T_i} = 0$. The N_i are positive integer token counts known in $\mathcal{K} = \bigvee_i \mathcal{K}^{(i)}_{T_i}$.

Proof. Every R_i is retained in $\mathcal{K}$, so each bounded GRPO coefficient A_i in Equation 1 is measurable there. This shared information accommodates the dependence among group advantages.

Conditional independence of the local trajectories, followed by Proposition 1, gives

$$\begin{aligned}\mathbb{E}[S_H^{(i)} - \widetilde{S}_{T_i}^{(i)} \mid \mathcal{K}] &= \mathbb{E}[S_H^{(i)} - \widetilde{S}_{T_i}^{(i)} \mid \mathcal{K}_{T_i}^{(i)}] \\ &= 0.\end{aligned} \tag{33}$$

The first equality uses the fact that each sibling's retained record is a function of its own trajectory and X. Taking the known coefficients A_i/N_i outside the conditional expectation and summing proves $G_{g,T} = \mathbb{E}[G_g \mid \mathcal{K}]$. With $D_g = G_g - G_{g,T}$ and $\mu_g = \mathbb{E}[G_g]$, conditional centering implies

$$\mathbb{E}[(G_{g,T} - \mu_g)D_g^\top] = \mathbb{E}[D_g(G_{g,T} - \mu_g)^\top] = 0. \tag{34}$$

Expanding the covariance gives

$$\begin{aligned}\mathbb{E}[G_{g,T}] &= \mathbb{E}[G_g], \\ \mathrm{Cov}(G_g) - \mathrm{Cov}(G_{g,T}) &= \mathbb{E}[D_g D_g^\top] \succeq 0.\end{aligned} \tag{35}$$

Let $\mathcal{E}$ be an acceptance event determined by X and $(R_1, \ldots, R_m)$, with $\Pr(\mathcal{E}) > 0$. Since $\mathcal{E} \in \mathcal{K}$, the same projection holds under $\Pr(\cdot \mid \mathcal{E})$. Therefore

$$\begin{aligned}\mathbb{E}[G_{g,T} \mid \mathcal{E}] &= \mathbb{E}[G_g \mid \mathcal{E}], \\ \mathrm{Cov}(G_g \mid \mathcal{E}) - \mathrm{Cov}(G_{g,T} \mid \mathcal{E}) &= \mathbb{E}[D_g D_g^\top \mid \mathcal{E}] \succeq 0.\end{aligned} \tag{36}$$

The reference distribution on both sides is the same accepted-group distribution. □

Retained normalization. Measurability of N_i allows its reciprocal to be absorbed into the coefficient. Successful trajectories retain their complete records, so their token counts may depend on the completed trajectory. For failures, a fixed-length protocol or a count determined by the stopped record and outcome satisfies the same requirement. A future-dependent failure count introduces additional information, under which the removed score would have to be centered again. The theorem therefore uses the actual token count known in the retained record.

Binary coefficients and action credit. The relation to local action credit can also be made explicit. Let U collect sibling outcomes, which are independent of the local trajectory conditional on X. Including U in the initial analytical history preserves score centering. For a bounded coefficient $C = c(R, U)$, set $c_0(U) = c(0, U)$ and $\delta c(U) = c(1, U) - c(0, U)$. The binary outcome then gives $C = c_0(U) + \delta c(U)R$. Substituting the action-conditioned advantage and using score centering gives

$$\mathbb{E}[CZ_k \mid \mathcal{F}_k] = \delta c(U)\mathbb{E}[Z_k \mathcal{A}^\pi(h_k, a_k) \mid \mathcal{F}_k]. \tag{37}$$

Thus the coefficient difference $\delta c(U)$ weights local action credit. This identity uses the unfiltered on-policy law, whereas the accepted-group result above follows from conditioning the established projection. The two steps account separately for coefficient dependence and group selection.

B.6 Proof of Corollary 3

Proof. On $\{T < H\}$, the condition $P_T = 0$ and nonnegativity of R imply $R = 0$ almost surely. Hence $\chi_T = 0$ there. At $T = H$ the suffix is empty. Moreover, successes have $T = H$, so $\widetilde{S}_T = S_T$. The outcome is already $\mathcal{F}_T$-measurable, and any additionally retained successful history is already complete. Thus $\mathcal{K}_T = \mathcal{F}_T$ up to null events. Applying Theorem 2 proves the corollary. □

For comparison, suppose that terminal success is determined by an independent fair coin revealed at H. At every earlier history, $P_t = 1/2$, while independence and score centering give $\chi_t = 0$. Thus zero remaining credit can hold before the outcome is determined.

B.7 Ordering ideal stopping rules

Corollary 6 (Ordering ideal stopping rules)**.** *Consider two rules satisfying Theorem 2, with the same coefficients and denominators and $T_{i,1} \le T_{i,2}$ for every trajectory. Their group estimators obey*

$$\mathrm{Cov}(G_{g,T_1}) \preceq \mathrm{Cov}(G_{g,T_2}). \tag{38}$$

Proof. For $T_{i,1} \le T_{i,2}$, a later stopped prefix determines the earlier stopping record. Both rules retain the final outcome and complete successful trajectories. Therefore $\mathcal{K}^{(i)}_{T_{i,1}} \subseteq \mathcal{K}^{(i)}_{T_{i,2}}$, and their joint information sets satisfy $\mathcal{K}_1 \subseteq \mathcal{K}_2$. With the same full estimator and coefficients measurable in both information sets, the tower property gives

$$\begin{aligned} \mathbb{E}[G_{g,T_2} \mid \mathcal{K}_1] &= G_{g,T_1}, \\ \mathrm{Cov}(G_{g,T_2}) - \mathrm{Cov}(G_{g,T_1}) &= \mathbb{E}[(G_{g,T_2} - G_{g,T_1})(G_{g,T_2} - G_{g,T_1})^\top] \succeq 0. \end{aligned} \tag{39}$$

The same argument holds under a shared acceptance event in $\mathcal{K}_1$. □

Earliest zero-credit stop. For each trajectory, define

$$\begin{aligned} T_* &= \min\big(\{t \in \{1, \dots, H-1\} : \chi_t = 0\} \cup \{H\}\big), \\ B_* &= T_* - 1. \end{aligned} \tag{40}$$

Adaptedness of χ_t makes this first zero, with fallback H, a stopping time. It precedes every stop with $T \ge 1$ and $\chi_T = 0$. Within the class satisfying Theorem 2 with the same coefficients and denominators, the ordering result therefore favors earlier stops. Whenever the original denominators are measurable in the joint information retained at T_*, this rule belongs to that class and attains its minimum covariance. The construction uses exact conditional credit at the fixed collection policy.

B.8 Proof of Proposition 4

All expectations below use the same fixed reference law, which may be the law conditional on group acceptance.

Proof. Since G_B is $\mathcal{K}_B$-measurable,

$$\mathbb{E}[G \mid \mathcal{K}_B] = G_B + \mathbb{E}[D_B \mid \mathcal{K}_B] = G_B + \eta_B. \tag{41}$$

Taking expectations gives $\mathbb{E}[G_B] - \mathbb{E}[G] = -\mathbb{E}[\eta_B]$, and Jensen's inequality bounds the squared expectation shift. Moreover,

$$\mathrm{Cov}(G_B, D_B) = \mathrm{Cov}(G_B, \eta_B), \tag{42}$$

because $D_B - \eta_B$ has conditional mean zero given $\mathcal{K}_B$. Expanding $\mathrm{Cov}(G_B + D_B)$ therefore yields

$$\begin{aligned} \mathrm{Cov}(G) - \mathrm{Cov}(G_B) = {} & \mathrm{Cov}(D_B) + \mathrm{Cov}(G_B, \eta_B) \\ & + \mathrm{Cov}(\eta_B, G_B). \end{aligned} \tag{43}$$

The absolute trace of either cross term is at most $\sqrt{V_B \,\mathrm{tr}\,\mathrm{Cov}(\eta_B)} \le \sqrt{V_B \epsilon_B^2}$, proving Equation 10. If $\eta_B = 0$, then $\mathbb{E}[D_B] = 0$ and the cross terms vanish, giving

$$\mathrm{Cov}(G) - \mathrm{Cov}(G_B) = \mathbb{E}[D_B D_B^\top] \succeq 0. \tag{44}$$

Conversely, the projection identity forces $\eta_B = 0$, completing the equivalence in Proposition 4. The general bound controls total variance through its trace. When the residual vanishes, the matrix covariance ordering above also holds. □

The definition of $\mathcal{K}_B$ includes the selected boundaries and actual denominators, so η_B accounts for the information they convey. Conditioning on a coarser record $\mathcal{C} \subseteq \mathcal{K}_B$ gives $\mathbb{E}[D_B \mid \mathcal{C}] = \mathbb{E}[\eta_B \mid \mathcal{C}]$. Jensen's inequality therefore yields

$$\mathbb{E}\|\mathbb{E}[D_B \mid \mathcal{C}]\|^2 \le \epsilon_B^2. \tag{45}$$

This lower bound measures the residual component visible through $\mathcal{C}$, while ϵ_B^2 uses the full retained information. In particular, a positive coarse moment establishes a nonzero projection residual. Finite-model checks verify these identities and bounds across 18 examples, including retrospective selection, outcome filtering, and future-dependent failure denominators.

B.9 EXACT TRUNCATION TRADEOFF

Proposition 7 (Exact truncation tradeoff)**.** *For square-integrable vectors G, D, let $G_w = G - D$, $\mu = \mathbb{E}[G]$, $d = \mathbb{E}[D]$, and $\Delta_V = \operatorname{tr}\operatorname{Cov}(G) - \operatorname{tr}\operatorname{Cov}(G_w)$. Averages of n independent identically distributed copies satisfy Equation 11. Strict MSE improvement relative to μ occurs exactly when*

$$\Delta_V > n\|d\|^2. \tag{46}$$

Proof. The expectation of G_w is $\mu - d$. For an average of n independent identically distributed gradient observations,

$$\mathbb{E}\|\overline{G}_w - \mu\|^2 = \|d\|^2 + \frac{1}{n}\operatorname{tr}\operatorname{Cov}(G_w), \tag{47}$$

$$\mathbb{E}\|\overline{G} - \mu\|^2 = \frac{1}{n}\operatorname{tr}\operatorname{Cov}(G). \tag{48}$$

Subtracting gives Equation 11, with $\Delta_V = \operatorname{tr}\operatorname{Cov}(G) - \operatorname{tr}\operatorname{Cov}(G_w)$. To relate this to the removed update, define $V_D = \operatorname{tr}\operatorname{Cov}(D)$ and $C_{wD} = \mathbb{E}\langle G_w - \mathbb{E}G_w, D - d\rangle$. The identity $G = G_w + D$ gives

$$\operatorname{tr}\operatorname{Cov}(G) = \operatorname{tr}\operatorname{Cov}(G_w) + \operatorname{tr}\operatorname{Cov}(D) + 2\mathbb{E}\langle G_w - \mathbb{E}G_w, D - d\rangle. \tag{49}$$

Thus $\Delta_V = V_D + 2C_{wD}$, and strict improvement is equivalent to Equation 46. The cross term records any noise cancellation between retained and removed contributions. Independence is used across the n copies, while the retained and removed parts of each copy may be dependent. □

Applying the projection-residual bounds. For the retrospective estimator, set $G_w = G_B$ and $D = D_B$. Substituting $\|d\|^2 \le \epsilon_B^2$ and the total-variance bound from Proposition 4 gives

$$\begin{aligned}&\mathbb{E}\|\overline{G}_B - \mu\|^2 - \mathbb{E}\|\overline{G} - \mu\|^2 \\ &\qquad \le \epsilon_B^2 - \frac{V_D}{n} + \frac{2\sqrt{V_B\epsilon_B^2}}{n}.\end{aligned} \tag{50}$$

A negative right-hand side is sufficient for MSE improvement. This compares the residual magnitude with the removed variance and the batch size.

Connecting the decomposition to clipped PPO. For a fixed nonnegative weight w and fixed advantage A, positive homogeneity gives

$$\min\{\rho w A, \operatorname{clip}(\rho) w A\} = w \min\{\rho A, \operatorname{clip}(\rho) A\}. \tag{51}$$

Consequently, at the same parameters and on the same data, detached masking with the original denominator decomposes the raw surrogate gradient into retained and removed contributions. Consistent subgradients are used at clipping boundaries, and common regularizers belong to the retained reference part. The decomposition therefore applies to paired surrogate gradients evaluated before an optimizer update.

A statistical design consequence. Under the ideal projection, $d = 0$ and $C_{wD} = 0$, so $\Delta_V = V_D \ge 0$. For a fixed general rule with $\Delta_V > 0$ and $\|d\| > 0$, strict MSE improvement occurs when $n < \Delta_V/\|d\|^2$. Thus increasing the batch size reduces the variance benefit relative to the fixed expectation shift. These comparisons use population moments of the fixed estimators, with empirical estimates subject to sampling uncertainty.

B.10 APPROXIMATE STOPPING AND CONDITIONAL BIAS

The zero-credit condition gives an exact projection. To quantify a nearby case, we bound the conditional expectation of the removed suffix at an adapted stop using the remaining outcome uncertainty.

Proposition 8 (Conditional bias of failure-only suffix removal)**.** *Under the finite-horizon score assumptions above, let $T \le H$ be a stopping time, c_0 a bounded $\mathcal{F}_T$-measurable failure coefficient, and $S_T^+ = \sum_{k=T}^{H-1} Z_k$. For $D = (1-R)c_0 S_T^+$,*

$$\begin{aligned}\mathbb{E}[D \mid \mathcal{F}_T] &= -c_0\mathbb{E}[(R - P_T)S_T^+ \mid \mathcal{F}_T] \\ &= -c_0\mathbb{E}\left[\sum_{k=T}^{H-1} Z_k \Delta P_k \,\middle|\, \mathcal{F}_T\right].\end{aligned} \tag{52}$$

The conditional norm obeys

$$\|\mathbb{E}[D \mid \mathcal{F}_T]\| \le |c_0|\sqrt{P_T(1-P_T)\mathbb{E}[\|S_T^+\|^2 \mid \mathcal{F}_T]}. \tag{53}$$

If $P_T \le \eta$, $|c_0| \le L$, and $\mathbb{E}[\|S_T^+\|^2 \mid \mathcal{F}_T] \le v^2$ on the early-stop event, the conditional expectation shift there is at most $Lv\sqrt{\eta}$, and at $T = H$ it is zero.

Proof. For each possible stopping value t, conditional score centering and the tower property imply $\mathbb{E}[\sum_{k=t}^{H-1} Z_k \mid \mathcal{F}_t] = 0$. Splitting over $\{T = t\}$ yields $\mathbb{E}[S_T^+ \mid \mathcal{F}_T] = 0$. Hence

$$\mathbb{E}[D \mid \mathcal{F}_T] = c_0\big(\mathbb{E}[S_T^+ \mid \mathcal{F}_T] - \mathbb{E}[RS_T^+ \mid \mathcal{F}_T]\big) \tag{54}$$

$$= -c_0\mathbb{E}[(R - P_T)S_T^+ \mid \mathcal{F}_T]. \tag{55}$$

The remaining-credit identity established above gives the second equality in Equation 52. Applying conditional Cauchy–Schwarz to the sum of paired increments bounds the squared norm by

$$c_0^2\mathbb{E}\left[\sum_{k=T}^{H-1}\|Z_k\|^2 \,\middle|\, \mathcal{F}_T\right]\mathbb{E}\left[\sum_{k=T}^{H-1}(\Delta P_k)^2 \,\middle|\, \mathcal{F}_T\right]. \tag{56}$$

Conditional orthogonality of the score increments identifies the first expectation as $\mathbb{E}[\|S_T^+\|^2 \mid \mathcal{F}_T]$. Proposition 5 identifies the second as $P_T(1 - P_T)$, yielding Equation 53. The stated bounds and $P_T(1 - P_T) \le P_T$ give the final inequality. The two factors quantify the remaining score variation and outcome uncertainty. □

Since $G_w = G - D$, the unconditional shift is

$$\mathbb{E}[G_w] - \mathbb{E}[G] = \mathbb{E}\left[c_0 \sum_{k=T}^{H-1} Z_k \Delta P_k\right]. \tag{57}$$

The coefficient c_0 supplies the failure weight and its sign. For a constant baseline b, setting $c_0 = -b$ multiplies the expected suffix credit by $-b$. The factor $P_T(1 - P_T)$ measures residual outcome uncertainty, while the paired increments determine its contribution to the gradient shift.

The stopping time is chosen from the observed prefix, and $(1-R)$ selects removal after the outcome is known. Equation 52 explicitly accounts for this failure selection under the on-policy score law. Its use requires an adapted stop and a failure coefficient measurable at that stop.

B.11 Comparing retrospective and prefix boundaries

A rule based on the observed prefix provides a reference for examining how complete-video selection changes the removed contribution. We compare the two rules under the same fixed policy and group law, using the same GRPO coefficients and original denominators. This comparison characterizes the combined change in available information and boundary-selection procedure.

Prefix comparison rule. Evaluate the frozen event model on successive observed prefixes. At each nonterminal time t, an error qualifies according to the event-mapping conditions in Section 3, with the recovery field omitted. Let T' be the first qualifying time, with fallback H, and set $B' = T' - 1$. Here the boundary is fixed at the detection time, so it is determined by the available prefix. This choice makes T' a stopping time when decoding randomness is fixed in advance or included in the local filtration. By comparison, the complete-video rule selects the right endpoint of the last unrecovered error.

For each failed trajectory i, write B_i and B_i' for the two boundaries. Define the group-level disagreement event $\mathcal{E}_{\neq} = \{\exists i : R_i = 0,\ B_i \neq B_i'\}$. Both rules retain all successful-rollout contributions. Hence, with $D_{B'} = G - G_{B'}$, their removed contributions agree outside $\mathcal{E}_{\neq}$, and

$$\delta^2 = \mathbb{E}\|D_B - D_{B'}\|^2 = \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_{\neq}}\|D_B - D_{B'}\|^2\big]. \tag{58}$$

If $\Pr(\mathcal{E}_{\neq}) > 0$, this equals $\Pr(\mathcal{E}_{\neq})\mathbb{E}[\|D_B - D_{B'}\|^2 \mid \mathcal{E}_{\neq}]$. If the event has probability zero, then $\delta = 0$.

Proposition 9 (Hindsight decomposition)**.** *For square-integrable estimators, write* $V_D = \operatorname{tr}\operatorname{Cov}(D_B)$*,* $V_{B'} = \operatorname{tr}\operatorname{Cov}(G_{B'})$*, and* $V_{D'} = \operatorname{tr}\operatorname{Cov}(D_{B'})$*. With* $\Delta_V = \operatorname{tr}\operatorname{Cov}(G) - \operatorname{tr}\operatorname{Cov}(G_B)$*, the two rules satisfy*

$$\begin{aligned} \|\mathbb{E}[D_B]\| &\le \|\mathbb{E}[D_{B'}]\| + \delta, \\ \Delta_V &\ge V_D - 2\big|\operatorname{tr}\operatorname{Cov}(G_{B'}, D_{B'})\big| - 2\delta\big(\sqrt{V_{B'}} + \sqrt{V_{D'}}\big) - 2\delta^2. \end{aligned} \tag{59}$$

Proof. Write $D' = D_{B'}$ and $\Delta_D = D_B - D'$, so that $D_B = D' + \Delta_D$ and $G_B = G_{B'} - \Delta_D$. The expectation bound follows from $\mathbb{E}[D_B] = \mathbb{E}[D'] + \mathbb{E}[\Delta_D]$ and $\|\mathbb{E}[\Delta_D]\| \le \sqrt{\mathbb{E}\|\Delta_D\|^2} = \delta$. For the variance, $\Delta_V = \operatorname{tr}\operatorname{Cov}(D_B) + 2\operatorname{tr}\operatorname{Cov}(G_B, D_B)$ and

$$\operatorname{Cov}(G_B, D_B) = \operatorname{Cov}(G_{B'}, D') + \operatorname{Cov}(G_{B'}, \Delta_D) - \operatorname{Cov}(\Delta_D, D') - \operatorname{Cov}(\Delta_D). \tag{60}$$

Since $\operatorname{tr}\operatorname{Cov}(\Delta_D) \le \mathbb{E}\|\Delta_D\|^2 = \delta^2$, the Cauchy–Schwarz inequality for traces bounds the last three terms by $\delta\sqrt{V_{B'}}$, $\delta\sqrt{V_{D'}}$, and δ^2, which proves Equation 59. All moments use the fixed reference group law, so the same algebra applies when this is an accepted-group law. □

The first bound separates the prefix rule's expectation shift from the additional discrepancy δ. For a local trajectory under the on-policy law, Proposition 8 bounds the prefix shift when its failure coefficient is measurable at T'. At group level, Equation 59 retains $\mathbb{E}[D_{B'}]$ under the chosen reference law, accounting for the actual coefficients and any group selection.

Diagnostic quantities. These terms specify a possible paired evaluation on saved rollout groups. Querying the frozen model on successive prefixes gives each B'_i, which can be compared with the complete-video boundary B_i on failed trajectories. The fraction of groups with a differing failure boundary estimates $\Pr(\mathcal{E}_{\neq})$, and the distribution of $B_i - B'_i$ describes timing differences. Paired gradients on disagreeing groups estimate the conditional moment in δ^2, while gradients under the prefix rule estimate $\operatorname{tr}\operatorname{Cov}(G_{B'}, D_{B'})$. Finite models with known disagreement events verify the decomposition, including the zero-disagreement case.

B.12 Partial suffix removal

The same MSE decomposition also permits a theoretical comparison of removal strengths at a fixed boundary. Let D be the removed suffix contribution and w_{prefix} the corresponding binary prefix mask on failures. A scalar $\lambda \in [0, 1]$ defines $G_\lambda = G - \lambda D$, with temporal weights on failed trajectories given by

$$w_\lambda = 1 - \lambda(1 - w_{\mathrm{prefix}}). \tag{61}$$

The prefix keeps weight one and the suffix gets $1 - \lambda$. Successful-rollout loss terms and the original denominator are preserved. This family varies the strength of suffix removal, whereas the uniform-scaling ablation in Section 4 scales the entire failed trajectory.

Set $\mu = \mathbb{E}[G]$, $d = \mathbb{E}[D]$, $V_D = \operatorname{tr}\operatorname{Cov}(D)$, and $C_{GD} = \mathbb{E}\langle G - \mu, D - d\rangle$. The same variance expansion gives

$$\begin{aligned} \mathbb{E}\|\overline{G}_\lambda - \mu\|^2 - \mathbb{E}\|\overline{G} - \mu\|^2 &= \lambda^2\|d\|^2 + \frac{\lambda^2 V_D - 2\lambda C_{GD}}{n}, \\ \lambda^* &= \operatorname{clip}_{[0,1]} \frac{C_{GD}}{V_D + n\|d\|^2}. \end{aligned} \tag{62}$$

The second line minimizes the quadratic in the first when its denominator is positive. If the denominator is zero, then $D = 0$ almost surely and all strengths are equivalent. For a fixed joint distribution of G and D, increasing n makes the optimal removal strength λ^* nonincreasing.

Under the exact projection, D is centered and uncorrelated with $G - D$, giving $d = 0$ and $C_{GD} = V_D$. If $V_D > 0$, Equation 62 gives $\lambda^* = 1$. Thus complete suffix removal minimizes MSE within this continuous family under the exact projection.

C EXAMPLES CLARIFYING THE PROJECTION CONDITIONS

The following examples separate remaining conditional credit from success probability, boundary selection, and cancellation across scenes. Each episode has two chunks. Chunk 0 is parameter independent, and chunk 1 samples $a \sim \mathrm{Bernoulli}(p)$ with logit parameter θ, giving scores $Z_0 = 0$ and $Z_1 = a - p$. For a fixed baseline $b \in [0, 1]$, the reference estimator is $G = (R - b)Z_1$. Successes retain both chunks. On failure, $T = 1$ removes Z_1, while $T = 2$ retains it, with boundary $B = T - 1$. Bars denote averages of n independent episodes, and MSE is measured relative to $\mu = \mathbb{E}[G]$.

Rare success and expectation shift. Set $R = a$ and choose the fixed candidate $T = 1$. The discarded action determines success, so $\chi_1 = p(1 - p) > 0$ for $0 < p < 1$. The full and masked estimators satisfy

$$G = (R - b)(a - p), \qquad \mathbb{E}[G] = p(1 - p), \tag{63}$$

$$G_w = \mathbf{1}\{a = 1\}(1 - b)(1 - p), \qquad \mathbb{E}[G_w] = (1 - b)p(1 - p). \tag{64}$$

The expectation shifts by $-bp(1-p)$, a fraction b of the reference expectation. Thus, as p approaches zero, the absolute shift vanishes while its relative size remains fixed. The two possible values of G are bp and $(1 - b)(1 - p)$, which give

$$\begin{aligned} \mathrm{Var}(G) &= p(1 - p)(1 - b - p)^2, \\ \mathrm{Var}(G_w) &= p(1 - p)^3(1 - b)^2. \end{aligned} \tag{65}$$

At $b = 1/2$ and $p = 0.01$, the expectations are 0.0099 and 0.00495, and the variances are 0.00237699 and 0.0024257475. Both squared bias and variance increase, so masking raises MSE for every n. For $n = 64$, MSE rises from 0.00003714046875 to 0.0000624048046875.

The same construction makes the source of the variance increase explicit at $p = b = 1/2$. Here $G = 1/4$ on both outcomes, while $G_w = 1/4$ on success and zero on failure. Therefore

$$\begin{aligned} \mathbb{E}[G_w] &= \frac{1}{8}, \\ \mathrm{Var}(G_w) &= \frac{1}{64}, \\ \mathbb{E}\|\overline{G}_w - \mu\|^2 &= \frac{1}{64} + \frac{1}{64n}. \end{aligned} \tag{66}$$

Writing $D = G - G_w$ gives $\mathrm{Var}(D) = 1/64$ and $\mathrm{Cov}(G_w, D) = -1/64$. The full estimator's zero variance comes from cancellation between retained and removed contributions.

The approximate-stopping bias bound is also attained in this construction. The failure coefficient is $-b$, the remaining score is $a - p$, and $P_1 = p$. Its squared bound is therefore $b^2p^2(1 - p)^2$, exactly the squared expectation shift. This equality illustrates why a small absolute bias bound can coexist with a fixed relative change.

Zero credit with uncertain outcomes. Set $p = b = 1/2$, let an independent fair coin determine R, and choose $T = 1$. Success is independent of the action, so $P_1 = 1/2$ and $\chi_1 = \mathbb{E}[RZ_1 \mid \mathcal{F}_1] = 0$. The full estimator has expectation zero and variance $1/16$. Masking failures gives $G_w = R(a - 1/2)/2$, with expectation zero and variance $1/32$. On failure, the retained information contains the outcome but leaves the action uniform, making the conditional expectation of G zero. On success, the complete record reveals G. Thus G_w is the conditional projection even though success remains uncertain at $T = 1$.

Future-dependent boundary selection. Keep $p = b = 1/2$ and the independent outcome coin, but choose $T = 1$ when $a = 1$ and $T = 2$ otherwise. Both deterministic candidates have zero remaining credit. However, $\{T = 1\} = \{a = 1\}$ depends on an action unobserved at time 1, so T is not a stopping time. Its failure-only mask removes exactly the event $(a, R) = (1, 0)$, which has probability $1/4$ and full contribution $-1/4$. The expectation therefore changes from zero to $1/16$. Here the shift arises from selecting on the future action, despite zero credit at each fixed candidate time.

Cancellation across initial scenes. Let an observed initial scene $X \in \{0, 1\}$ be uniform, with $p = b = 1/2$ and $T = 1$. Set $R = a$ for $X = 0$ and $R = 1 - a$ for $X = 1$. The two conditional credits are $1/4$ and $-1/4$. Within each scene, G is constant at the corresponding value, while G_w retains it only on success. Averaging across scenes gives expectation zero for both estimators and variances $1/16$ and $1/32$, respectively. Yet on failure, the retained scene identifies $G = \pm 1/4$, whereas $G_w = 0$. The equal global expectations therefore result from cancellation across scenes rather than conditional projection.

Finite enumeration check. Exact rational enumeration verifies the expectations, variances, and MSEs in these examples. Further checks cover local and group projections, nested stops, variable denominators, outcome filtering, approximate bias, and partial suffix removal. These finite calculations complement the analytic proofs.

D Policy Training and Evaluation

Experiments are conducted in RoboTwin 2.0 (Chen et al., 2025), a SAPIEN-based simulation benchmark for bimanual manipulation. The seven tasks cover tool use, handover, lifting, placement, and stacking in randomized scenes. A rollout ends when the task-specific success condition is satisfied or its execution budget is exhausted. The simulator assigns a binary terminal reward $R \in \{0, 1\}$ from this outcome.

D.1 Training and evaluation settings

The policy uses the action-token OpenVLA-OFT implementation from SimpleVLA-RL (Kim et al., 2025a; Li et al., 2026a). For each task, the RL methods start from the same SFT checkpoint. Table 3 summarizes the rollout collection, optimization, and evaluation settings. One policy update collects 64 accepted groups (512 trajectories) and makes one PPO pass in minibatches of 128, giving four optimizer steps.

Table 4 lists task-specific episode budgets and Full GRPO success at every scheduled evaluation from update 0 to 40. Its initial and update-40 evaluations supply the SFT and Full GRPO columns in Table 1.

D.2 Feedback accounting and normalization

The collection policy applies temperature 1.6 on the action-token support, and token likelihoods use the same distribution. In Equation 1, s_g is the sample standard deviation of group rewards and $\epsilon_{\text{num}} = 10^{-6}$. Each chunk score $Z_{i,k}$ sums its constituent token scores. At the collection parameters, the unclipped full-feedback objective therefore contributes $A_i Z_{i,k}/N_i$ per chunk, matching the reference estimator's weighting.

The loss averages action-token terms within each trajectory before gradient accumulation, using the original valid-token count N_i as its denominator. This implements the normalization in Equation 15. To quantify the feedback retained by a mask w, define its pre-clipping negative coefficient mass as

$$D_-(w) = \frac{1}{M} \sum_{i:R_i=0} \frac{|A_i|}{N_i} \sum_{u \in \mathcal{V}_i} w_{i,k(u)}. \tag{67}$$

This quantity combines the retained token weights with each failed rollout's advantage and original denominator. The ratio $D_-(w)/D_-(1)$ measures the retained fraction of feedback weight used in the ablation analysis. For example, retaining three of eight equally populated chunks leaves $3/8$ of a failed trajectory's original coefficient mass. Across trajectories, the aggregate fraction also depends on their advantage weights.

Theorem 2 additionally requires N_i to be measurable from the retained information. In the implementation, the valid-token mask is determined by the executed control steps, and N_i is its sum. Feedback truncation sets suffix advantages to zero while retaining this mask and its denominator. A fully valid eight-chunk Hammer failure has $N_i = 8 \times 25 \times 14 = 2800$. Earlier termination reduces the valid-token count before feedback truncation is applied.

Table 3: **Policy training and evaluation settings.** Configuration for the reported policy comparisons and feedback ablations.

Parameter	Setting
Policy and rollout collection	
Server capacity	8 NVIDIA H100 GPUs
Initialization	Task-specific SFT checkpoint
Actor observations	Head-camera view and proprioception
Environment	Randomized RoboTwin 2.0 scenes, seen instructions
Training initial conditions	1,000 per task
Action chunk	25 control steps, 14 action dimensions per step
Accepted rollout batch	64 groups of 8 trajectories per update
Group filtering	Success fraction in $[0.1, 0.9]$, with adaptive refill
Sampling temperature	1.6
Policy optimization	
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Learning rate	5×10^{-6}
Weight decay	0.01
Actor LoRA	Disabled
Gradient clipping	Global norm capped at 1
PPO passes per rollout batch	1
Global PPO minibatch	128 trajectories
Microbatch per GPU	1 trajectory
PPO clipping	$\varepsilon_- = 0.2$, $\varepsilon_+ = 0.28$
KL / entropy coefficients	0 / 0
Loss normalization	Original valid-token count of each trajectory
Training budget	40 policy updates
Policy evaluation	
Decoding	Greedy
Schedule	Updates $0, 5, \ldots, 40$
Episodes per evaluation	128 seen-seed and 128 held-out-seed
Scene splits	Training-seed and held-out-seed scenes, with the same task configuration

Table 4: **Task budgets and Full GRPO learning progress.** Success (%) pooled over 128 seen-seed and 128 held-out-seed episodes at each update. Update 0 is the SFT evaluation. Episode budgets are maxima, with 25 action steps per chunk.

Task	Episode budget		Full GRPO success (%) at update								
	Max. action steps	Max. chunks	0	5	10	15	20	25	30	35	40
Hammer	200	8	30.5	35.5	48.0	46.9	53.9	63.7	65.2	74.6	74.2
Handover	400	16	21.1	24.6	32.0	30.5	31.2	34.4	35.5	33.6	34.8
Lift pot	200	8	5.5	20.7	29.3	35.9	44.5	46.1	52.0	54.3	58.2
Dual bottles	100	4	20.7	21.5	32.0	30.1	35.9	44.1	43.0	41.4	44.9
Container on Plate	150	6	87.1	91.4	91.4	93.8	92.6	93.0	93.8	94.9	93.8
Empty cup	200	8	80.9	82.8	86.7	83.6	85.9	90.6	89.8	89.5	89.1
Stack bowls	400	16	39.1	46.1	47.7	56.2	57.0	56.2	60.2	63.3	60.2

E VLM Event Model: Mapping, Corpus, and Localization

E.1 VLM inputs, prompting, and boundary mapping

Input construction. Each request contains observations from one completed rollout. Table 5 lists their sources and representation. The three camera streams are supplied as separate video

inputs, while task context, robot observations, and temporal alignment enter the user message as text. Terminal rewards are used by the training loop to select failed rollouts for inference.

Table 5: **VLM inputs.** Sources and representation for a single rollout request.

Input	Source	Representation
Task context	Task configuration and rollout metadata	Natural-language instruction, task clarification, executed chunk count, and initial action budget.
Camera streams	Head and two wrist cameras	Three synchronized video sequences. Native 320×240 RGB frames are extended to 320×288 with camera, frame, and chunk labels. Sampling uses the initial frame and up to six frames per chunk.
Robot state	Robot observation records	Gripper commands, measured openings, and end-effector poses at each sampled frame. Joint positions and velocities are included at the initial frame and chunk ends. Values are serialized as JSON text.
Temporal alignment	Rollout recorder	Frame and chunk indices, within-chunk progress, and simulation timestamps.

Each full action chunk contains 25 control steps. Frame 0 precedes chunk 0, and a frame with within-chunk fraction 1 records the end of that chunk. These indices locate events on the policy timeline.

Prompt structure. A shared template combines the task instruction and a fixed task-specific clarification with the executed chunk count, initial action budget, and synchronized observations. The clarification specifies the intended manipulation, such as stacking the bowls and releasing the moved bowl. The system instruction and an abridged version of the user instruction are shown below.

VLM event-localization prompt

System instruction. Review robot video evidence. Return the requested concise JSON.

User instruction (abridged). Review the synchronized camera views and robot observations for the given task. Track objects across views and check whether apparent errors are followed by regrasping or task completion. Identify the last visible error sequence without subsequent recovery and localize its onset using the recorded frame-to-chunk mapping. Return a chunk interval when timing is ambiguous and a null interval when no specific error can be localized.

Required response. Return a compact JSON response containing the observed outcome, grasp and handover status, error type, error interval, recovery status, final holder, up to six supporting frame indices, and a brief factual summary.

Frozen-model inference uses greedy decoding with at most 512 generated tokens. The resulting event record supplies the outcome, interval, and recovery information used by the boundary rule.

Boundary mapping and fallbacks. The mapping in Equations 13 and 14 is applied in the following order.

1. **Select rollouts.** Terminal environment rewards determine which rollouts require inference. Failed rollouts are passed to the event model, while successful rollouts bypass it and retain all their loss terms.
2. **Validate responses.** The mapper checks the response schema and any non-null interval. A valid interval contains two integer chunk indices satisfying $0 \leq \ell_i \leq u_i < K_i$. Malformed responses, identity mismatches, and failed inference requests raise implementation errors rather than becoming no-event predictions.
3. **Apply semantic fallbacks.** Full feedback is retained for a predicted success or uncertain outcome, no clear error, an uncertain error type, or a missing localization interval. It is also retained when subsequent recovery from the localized error is reported or uncertain.

4. **Set the boundary.** For a valid failure prediction with a localized error and no observed recovery, the mapper returns the upper interval endpoint as $B_i = u_i$.

Supporting frame identifiers provide evidence for inspection. The validated event fields determine the feedback boundary.

E.2 Annotated video corpus

The annotated corpus contains 1,151 rollout videos, comprising 1,024 environment-failed trajectories and 127 successful references. Table 6 summarizes the training, validation, and test partitions.

Table 6: **Annotated video corpus.** Video counts by environment outcome and data split.

Video type	Training	Validation	Test	Total
Failed rollouts	640	128	256	1,024
Successful references	118	5	4	127
Total	758	133	260	1,151

Task–scene identity determines the split even when the same scene appears at several policy checkpoints. The full collection has 1,045 unique task–scene identities with zero overlap across partitions. The corpus also excludes the held-out scene seeds used for policy evaluation. Gradient fitting uses the training partition, and checkpoint ranking uses validation metrics.

Each video has structured event annotations recording observed outcome, grasp establishment, handover completion, error type, error localization interval, observed recovery, final holder, evidence frames, and a summary. The error interval expresses uncertainty about the onset of the last visible error without observed recovery. A null interval is allowed when no error can be localized, and unknown fields are preserved. Figure 6 pairs a video excerpt with its annotation. In this example, the initial pickup difficulty is resolved, and the interval identifies the later placement error in chunk 12.

(a) Video excerpt and instruction Stack bowls

"Stack the ceramic bowl carefully over the blue bowl with slightly rounded base."



(b) Event annotation

Observed outcome	Failure	**Grasp established**	Yes
Error type	Placement failure	**Later recovery**	Not observed
Error interval	Chunks [12, 12]	**Final holder**	Neither arm

Summary. After the initial grasping difficulty is resolved, releasing the bowl during stacking displaces the bowls. No subsequent recovery is observed.

Figure 6: **An annotated rollout.** Selected head-camera frames and event labels from the video corpus. The interval identifies the placement error after the initial grasping difficulty is resolved. Chunk indices are zero-based.

Table 7: **LoRA training configuration.** Settings for the shared VLM event adapter.

Parameter	Setting
Adapter	
Adaptation scope	One shared adapter across tasks
Target layers	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `in_proj_qkv`, `in_proj_z`, `in_proj_b`, `in_proj_a`, `out_proj`
Rank	8
LoRA scaling parameter	16
Dropout	0
Frozen components	Base model, vision encoder, and projector
Optimization	
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0
Gradient clipping	Global norm capped at 1
Effective batch size	20 trajectories
Compute precision	Bfloat16
Loss normalization	Mean over response tokens within each trajectory, then over trajectories in the batch

Environment outcomes are stored separately from versioned visible-event labels. The collection's sampling design and recording length can reveal outcome information. The complete corpus manifest contains 1,024 predicted visible failures, 71 visible successes, and 56 uncertain outcomes, with 1,011 non-null intervals and 140 null intervals. These label counts cover all 1,151 videos.

The event model trained on this corpus is documented in the next subsection, and the policy runs that use it in Appendix F.1.

E.3 Event model training

Adapter configuration. One adapter is shared across all annotated tasks. Table 7 summarizes its LoRA and optimization settings, which remain fixed throughout training. The response loss follows Equation 12, with the data partitions in Table 6. Training and localization evaluation use the chat template with thinking disabled.

Checkpoint selection. Checkpoints are ranked by validation task-macro IoU, with lower response NLL and then earlier update as tie-breakers. Both the highest validation IoU (78.2%) and lowest NLL occur at update 304 (epoch 8), which is retained. A fresh patience counter starts at update 361, with validation every 19 updates (half an epoch). Four checks without an IoU improvement of at least 0.005 over the last reset anchor trigger stopping at update 437 (11.5 epochs). This patience rule controls training duration separately from checkpoint ranking.

Localization evaluation. Test diagnostics also informed training duration. Table 8 reports agreement with versioned reference intervals on 220 localizable, unrecovered test failures from the seven tasks in the policy comparison. Eligibility is determined from reference annotations independently of model predictions. The reported checkpoint is selected by validation IoU.

Qualitative localization. Figure 7 shows a failed Handover in which the selected model localizes the transfer error to chunks $[4, 5]$, matching the reference annotation. Head-camera frames place this interval between the initial grasp and later motion without recovery.

Localization across checkpoints. Figure 8 follows the prediction for the same rollout across checkpoints. The base model predicts $[6, 8]$, and epoch 3 predicts $[5, 5]$. At epochs 6 and 8, both endpoints match the reference interval $[4, 5]$.

Table 8: **Seven-task event localization.** Interval IoU (%) against versioned reference intervals on 220 localizable, unrecovered test failures from the tasks in the policy comparison. Update 304 is the validation-selected checkpoint. Scores measure label agreement under the evaluation protocol in Appendix E.3.

Task	n	Base	Epoch 3	Epoch 6	Epoch 8
beat_block_hammer	32	34.0	21.9	57.8	60.9
handover_block	32	36.5	43.8	70.3	73.4
lift_pot	32	35.7	60.9	77.6	80.7
pick_dual_bottles	32	33.3	70.3	75.0	73.4
place_container_plate	31	31.7	58.1	79.0	77.4
place_empty_cup	32	24.8	53.1	59.4	72.9
stack_bowls_two	29	3.8	48.3	55.2	65.5
Task-macro mean	220	28.5	50.9	67.8	72.1

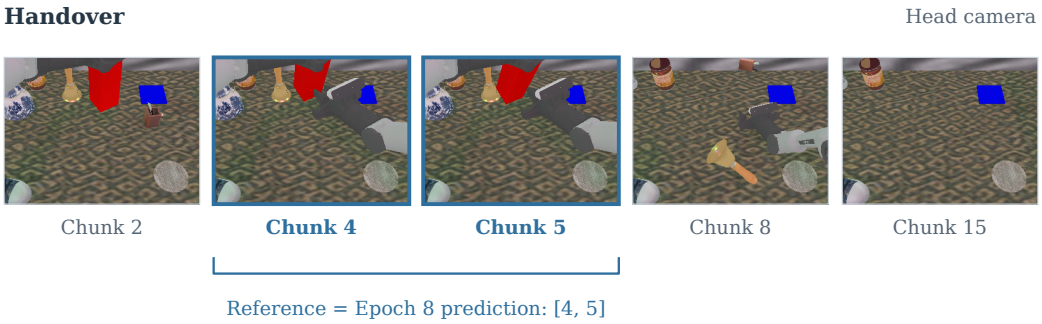


Figure 7: **Event localization on a test rollout.** Head-camera frames of a failed Handover. Blue outlines mark the interval identified by both the reference annotation and the selected model. Chunk indices are zero-based.

F Policy Results and Feedback Ablations

This appendix provides scheduled policy evaluations, feedback statistics, and ablation details.

F.1 Seven-task policy results

The frozen event model is the validation-selected epoch-8 adapter (update 304) of Appendix E.3. Table 1 reports the seven-task policy comparisons at update 40 and their best scheduled scores.

Training configuration. All methods follow the training and evaluation settings of Appendix D.1. Microbatches for actor updates and log-probability computation contain one trajectory per rank, and gradient accumulation preserves the global PPO minibatch of 128 trajectories.

The fixed-endpoint results use the scheduled evaluation after 40 updates, with 128 seen-seed and 128 held-out-seed episodes and greedy decoding. The Best columns additionally report each run's maximum pooled success over the same eight evaluations at updates $5, 10, \ldots, 40$, and their means average the per-task maxima. The SFT column uses the Full-GRPO run's initial evaluation. Both methods share the task checkpoint and evaluation seed splits. Separate simulator executions retain some variability under greedy action decoding, so their initial success rates can differ. Table 9 provides all scheduled scores across the seven tasks. The update budget fixes the number of accepted training groups. Adaptive refill can require different numbers of generated trajectories to reach this budget.

Feedback usage. Table 10 summarizes the feedback retained by the event masks.

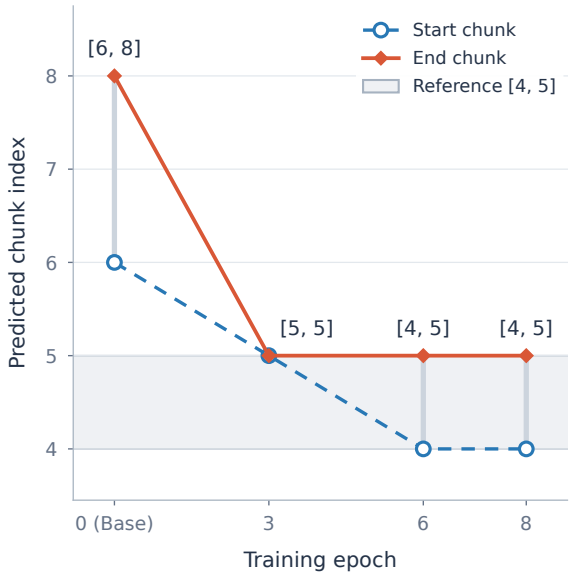


Figure 8: **Predicted intervals across checkpoints.** Interval endpoints for the Handover rollout in Figure 7. The shaded band marks the reference interval $[4, 5]$.

Table 9: **Scheduled policy evaluations.** Combined success (%) at each scheduled update. Table 1 reports update 40 and the per-run maximum over updates $5, 10, \ldots, 40$. SFT evaluations at update 0 use the same task checkpoint in separate simulator executions. The dash denotes the unavailable initial Handover MART evaluation.

	Hammer		Handover		Lift pot		Dual bottles		Container		Empty cup		Stack bowls	
Update	Full	MART	Full	MART	Full	MART	Full	MART	Full	MART	Full	MART	Full	MART
0	30.5	30.9	21.1	—	5.5	5.9	20.7	20.3	87.1	86.3	80.9	80.9	39.1	37.1
5	35.5	42.6	24.6	25.8	20.7	24.6	21.5	23.0	91.4	88.7	82.8	84.4	46.1	40.2
10	48.0	49.6	32.0	25.4	29.3	31.2	32.0	25.0	91.4	91.4	86.7	87.5	47.7	52.3
15	46.9	63.3	30.5	39.1	35.9	38.3	30.1	34.4	93.8	94.9	83.6	89.5	56.2	56.2
20	53.9	69.1	31.2	35.2	44.5	42.2	35.9	37.1	92.6	94.1	85.9	89.5	57.0	59.0
25	63.7	77.0	34.4	37.1	46.1	48.8	44.1	37.9	93.0	94.1	90.6	87.5	56.2	60.9
30	65.2	78.1	35.5	42.6	52.0	58.2	43.0	41.4	93.8	92.6	89.8	88.3	60.2	59.8
35	74.6	80.5	33.6	44.5	54.3	58.6	41.4	51.2	94.9	94.1	89.5	87.1	63.3	67.2
40	74.2	**81.2**	34.8	**43.8**	58.2	**58.6**	44.9	**48.8**	**93.8**	91.8	89.1	**91.4**	60.2	**64.8**

The statistics pool event records and loss-mask coefficients over all 40 updates of each MART run, using the same frozen adapter and preserving success weights and valid-token denominators. Retained coefficient mass in Table 10 is the ratio of summed post-mask to summed pre-mask mass. A localized event at the final chunk retains full feedback without a semantic fallback.

Success rates are computed from integer episode counts, with pooled success calculated over all 256 episodes across both seed splits.

F.2 Feedback Ablations

The ablations cover Hammer and Handover, selected from the seven-task policy comparisons. For each task, statistics are pooled over all 40 MART updates and frozen before ablation training. Uniform scaling divides the retained negative coefficient mass summed over these updates by the corresponding sum of original mass, using Equation 67. This gives $\gamma = 0.4920678764$ for Hammer and 0.3268904598 for Handover. Random prefix draws independently from the empirical boundary his-

Table 10: **What the event mask removes.** Pooled over all 40 updates of each MART run. N_{-} counts accepted failed trajectories. Shortened denotes a boundary before the final chunk. Retained mass is the ratio of summed negative coefficient mass after and before masking. Fallbacks retain full feedback. These statistics quantify loss weights.

Task	N_{-}	Shortened (%)	Retained mass (%)	Fallback (%)
Hammer	8,304	90.6	49.2	8.3
Handover	12,955	99.4	32.7	0.5
Lift pot	11,577	99.7	47.4	0.2
Dual bottles	13,296	97.9	59.3	1.9
Container on Plate	4,364	97.8	51.8	2.0
Empty cup	5,612	94.7	53.6	4.8
Stack bowls	10,176	96.4	37.9	2.7

togram of accepted failures, including full-feedback fallbacks. The ablation RNG is separate from policy sampling. Each sampled boundary is a zero-based chunk index, and its prefix is intersected with the new rollout's valid-token range. Both ablations preserve success weights, original token denominators, and terminal rewards, and run without VLM inference.

All ablations use the same task SFT checkpoints, 64 accepted groups of eight rollouts per update, global PPO minibatch 128, and the update-40 evaluation endpoint with 128 seen-seed and 128 held-out-seed episodes. Both ablations reuse frozen statistics, with uniform scaling matching the aggregate coefficient ratio and random prefix matching the boundary distribution. Their realized coefficient-mass ratios are measured on each run's own rollout groups.

At update 40, uniform scaling scores $100/128$ on seen-seed and $99/128$ on held-out-seed scenes for Hammer ($199/256$ combined), and $50/128$ and $54/128$ for Handover ($104/256$ combined). Random prefix scores $96/128$ and $88/128$ for Hammer ($184/256$ combined), and $46/128$ and $60/128$ for Handover ($106/256$ combined), in the same split order.

Figure 9 compares the endpoints alongside each run's observed retained negative coefficient mass, pooled over all 40 updates using Equation 67. Under these similar feedback-weight budgets, event-guided truncation achieves higher success than both ablations on both tasks.

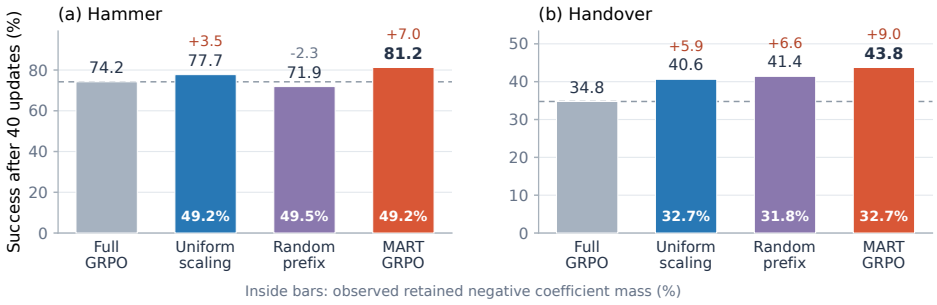


Figure 9: **Feedback ablations.** Success at update 40, with dashed Full GRPO references and percentage-point changes relative to Full GRPO above the bars. Percentages inside the bars show each run's observed retained negative coefficient mass, pooled over all 40 updates.