ROUTERL-VLA: LEARNING FROM FAILURES VIA FACTORIZED ROUTING FOR VISION-LANGUAGE-ACTION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models can acquire broad manipulation skills through large-scale pretraining and supervised fine-tuning. Yet a failed rollout is reduced to a single outcome signal. Without knowing where the error arose or which part of the policy caused it, existing reinforcement-learning methods for VLA models often spread credit across semantic, visual, and execution failures, diluting informative gradients and risking interference with previously acquired capabilities. We introduce RouteRL-VLA, a failure-aware post-training framework that factorizes credit over failure type, trajectory time, and parameter group. It constructs and compares paired clean and perturbed rollouts, uses a selective paired teacher to estimate failure attribution, soft failure-type probabilities, and query-level temporal weights, and directs the resulting advantages to specialized policy adapters. For stable on-policy learning with continuous action chunks, we develop joint action-chunk likelihoods, duration-aware generalized advantage estimation, and transactional KL-constrained PPO. When applied to OpenVLA-OFT on LIBERO-Goal, RouteRL-VLA attains SOTA failure-recovery performance among five trained methods: 62.15% perturbed macro success, 2.99 points above uniform credit, with 95.83% clean success. In a separate 1,224-episode evaluation, it improves success from 45.26% to 78.35%, a 33.09-point gain. Oracle-routing and route-permutation controls confirm that accurate route–failure alignment drives this improvement. The attribution model reaches 0.988 AUROC, while the selective router achieves 0.992 macro-F1 at 0.943 coverage. The experiments jointly show that paired diagnosis identifies how a rollout failed, factorized routing targets the responsible update, and alignment controls verify that informative routes drive recovery.

1 INTRODUCTION

Vision-language-action (VLA) models have made robot learning increasingly a problem of adaptation rather than acquisition. Large-scale multimodal pretraining gives a model semantic priors, visual grounding, and reusable motor patterns; post-training then has to turn those priors into dependable behavior on a target embodiment. In this setting, failure is not merely a low reward. It is evidence that some part of the model–instruction interpretation, visual grounding, or physical execution–did not connect to the intended outcome.

Reinforcement learning usually observes only a trajectory-level success signal. That signal says that a rollout failed, but not which responsibility should receive the update, when the error became consequential, or which parameters should change. A single undifferentiated advantage therefore spreads credit across heterogeneous causes. For a pretrained VLA model, this is more than inefficient: gradients meant to repair one failure mode can interfere with skills that were already correct.

We develop RouteRL-VLA around a simple principle: a failed rollout becomes useful supervision when it is contrasted with a matched rollout that differs in one controlled factor. Given the same task, initial state, and simulator seed, a clean–perturbed pair provides a local counterfactual. Its contrast supports three decisions: *what* failed, *when* the failure became relevant, and *where* the correction should be written. Figure 1 summarizes the resulting loop.

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.

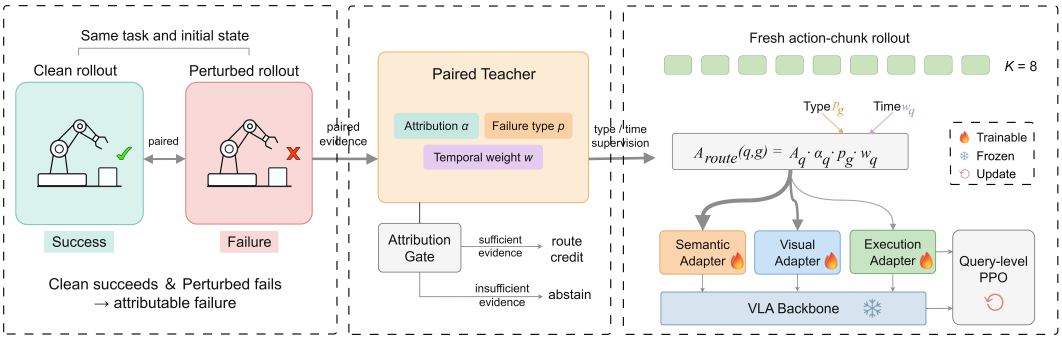

Figure 1: RouteRL-VLA closes the failure-recovery loop. A matched clean–perturbed pair provides evidence for a selective diagnosis; the resulting type, time, and parameter routes shape the action-chunk advantage; a KL gate either commits or rolls back the complete update.

The design has two coupled parts. First, a paired teacher learns to distinguish attributable failures from robust, ambiguous, and anomalous pairs. It emits a calibrated attribution gate, a soft distribution over Semantic, Visual, and Execution causes, and query-level temporal weights. The gate can abstain when the pair does not provide actionable evidence. Second, the policy converts those outputs into a factorized advantage over type, time, and parameter group. Three low-rank adapters receive the corresponding routed signal while the VLA backbone remains fixed. This turns diagnosis into targeted learning rather than a post-hoc label.

The routing problem is inseparable from the action representation. A VLA model may emit an eight-step continuous action chunk, but the chunk is one stochastic decision for policy optimization. We therefore use the exact sampling-time joint likelihood over executed positions, duration-aware generalized advantage estimation, and a transactional KL constraint that rolls back the complete optimizer state when a tentative update is unsafe. These details make the proposed credit assignment meaningful at the same unit at which the policy acts.

Our experiments are designed as a closed argument. A matched five-method comparison tests whether structured credit improves perturbed recovery while preserving clean skill. Held-out teacher diagnostics test whether the router knows when and where to route. Single-factor ablations test whether Type, Time, and Parameter routing are independently necessary. Finally, a larger matched evaluation and route-alignment controls test whether the mechanism survives scale and whether the gain depends on assigning the right update to the right failure. The results support all four links: RouteRL-VLA reaches the highest controlled recovery score, the teacher is accurate and selective, every routing factor matters, and correct alignment outperforms a frequency-matched misalignment.

Contributions. Our contributions are threefold. (i) We formulate failure recovery for VLA models as a paired, selective credit-assignment problem that separates attributable failures from pairs that should not drive a policy update. (ii) We introduce factorized Type $\times$ Time $\times$ Parameter routing with soft uncertainty and abstention, together with chunk-consistent PPO estimators and full-state KL rollback. (iii) We provide controlled, large-scale, and route-alignment evaluations whose design connects diagnosis quality to policy recovery and shows that all three routing dimensions are independently necessary.

2 Related Work

Vision-language-action models. The VLA model line has moved from learning isolated behaviors to transferring multimodal priors across tasks and embodiments. RT-1 and RT-2 connected transformer sequence modeling and web-scale vision-language knowledge to robot control, while VIMA explored multimodal prompting for general manipulation (Brohan et al., 2022; Jiang et al., 2022; Brohan et al., 2023). Cross-embodiment data and flow-based action generation broadened this direction through Open X-Embodiment, π_0, Octo, and OpenVLA (Open X-Embodiment Collaboration, 2023; Black et al., 2024; Octo Model Team et al., 2024; Kim et al., 2024). OpenVLA-OFT

makes continuous action chunks practical for post-training, creating a natural setting in which the unit of credit must match the unit of action. Recent work has emphasized staged scaling and richer robot interfaces, including Green-VLA (Apanasevich et al., 2026); diagnostic benchmarks such as LIBERO-Para and broader benchmarking studies further emphasize that language and evaluation choices can change apparent robustness (Kim et al., 2026; Jiang et al., 2026). Our question is complementary: once a VLA model fails under a controlled perturbation, how should learning assign responsibility?

RL post-training for VLA models. Interactive post-training, VLA-RL, and SimpleVLA-RL show that on-policy or online feedback can improve manipulation policies (Tan et al., 2025; Lu et al., 2025; Li et al., 2025). Recent approaches also use sparse VLA guidance to initialize RL or world models to supply imagined experience, as in VLA jump-starting, RISE, and VLA-MBPO (Moroncelli et al., 2026; Yang et al., 2026; Zhang et al., 2026). These systems address exploration, interaction cost, or model-based rollouts. RouteRL-VLA addresses the complementary failure-localization bottleneck: the policy update should be conditioned on the cause and temporal locus of the observed failure.

Structured and selective credit. GAE trades bias and variance in temporal credit, while RUDDER redistributes delayed return (Schulman et al., 2015; Arjona-Medina et al., 2018). PCGrad modifies conflicting multi-task gradients, LoRA provides compact trainable subspaces, and selective prediction formalizes abstention when a decision is unreliable (Yu et al., 2020; Hu et al., 2021; Geifman & El-Yaniv, 2017; 2019). RouteRL-VLA combines these ideas at a different interface: paired evidence first identifies whether a failure is attributable, then a selective teacher decomposes its advantage across failure type, action-query time, and adapter group.

3 METHOD

3.1 PROBLEM AND PAIRED EVIDENCE

Let a frozen VLA backbone with trainable adapters define $\pi_\theta(\mathbf{a}_q \mid h_q)$, where h_q is the observation–instruction history at query q and $\mathbf{a}_q = (a_{q,1}, \ldots, a_{q,H})$ is an action chunk. The environment executes $d_q \leq H$ actions before the next query. We use a validity mask $m_{q,t} \in \{0, 1\}$ with $\sum_t m_{q,t} = d_q$; padded and post-terminal positions have zero mask. A rollout τ receives a sparse terminal outcome $y(\tau) \in \{0, 1\}$.

For task i, initial state s, and simulator seed z, we construct a clean rollout τ^0_{isz} and a rollout τ^c_{isz} with one controlled perturbation $c \in \{\mathrm{sem}, \mathrm{vis}, \mathrm{exe}\}$. The pair is attributable when

$$y(\tau^0_{isz}) = 1 \quad \text{and} \quad y(\tau^c_{isz}) = 0. \tag{1}$$

The shared context removes many nuisance explanations that a failed trajectory alone cannot resolve. We retain three additional pair states: robust (success/success), ambiguous (failure/failure), and anomalous (failure/success). These states train the abstention behavior instead of forcing every failure into a type.

Semantic perturbations alter the instruction while preserving the intended task; visual perturbations modify perception-relevant appearance without changing the goal; execution perturbations act on the control channel. Perturbations are versioned and audited for validity. The pair is a supervision device for learning responsibility, not an assumption that deployment always provides a clean reference.

3.2 SELECTIVE PAIRED TEACHER

The teacher receives pair features $\phi(\tau^0, \tau^c)$ and returns, at each action query,

$$\alpha_q = P(\mathrm{attributable} \mid \phi), \qquad p_g = P(c = g \mid \phi, \mathrm{attributable}), \qquad w_q \geq 0, \quad \sum_g p_g = 1, \tag{2}$$

where $g \in \{\mathrm{sem}, \mathrm{vis}, \mathrm{exe}\}$. Temporal weights are normalized within the divergence region, $\sum_q w_q = 1$. A route is emitted only when attribution and type confidence exceed development-set thresholds; otherwise the teacher abstains.

The teacher is trained on pair-level splits and evaluated by leave-one-task-out transfer. This prevents near-duplicate pairs from appearing on both sides of the diagnostic boundary and links teacher quality to deployment-time routing. Calibration matters because a confident wrong route can be more damaging than no route. The selective gate therefore treats abstention as an explicit output, not as a failed prediction.

3.3 Factorized Advantage Routing

Let $\widehat{A}_q$ be the query-level advantage. RouteRL-VLA allocates

$$\widehat{A}^{\mathrm{route}}_{q,g} = \widehat{A}_q \, \alpha_q \, p_g \, \bar{w}_q, \qquad \bar{w}_q = \frac{Q w_q}{\sum_{j=1}^{Q} w_j + \epsilon}. \tag{3}$$

The temporal normalization preserves the trajectory's average advantage scale. Each failure type owns a low-rank adapter and receives only its routed term. Soft p_g propagates uncertainty; if the teacher abstains, the specialized policy loss is zero for that item while value learning may continue from its observed return.

Each adapter has 528,448 trainable parameters and the value head has 1,049,089, for 2,634,433 trainable parameters in total. The approximately 7.7B-parameter VLA backbone is frozen. This separation gives the method an inductive correspondence between semantic, visual, and execution responsibility and distinct update pathways. The corresponding one-factor removals directly test whether this structure is necessary.

3.4 Chunk-Consistent On-Policy Updates

We treat an $H \times 7$ action chunk as one stochastic decision. Six tanh-squashed Gaussian motion dimensions and one Bernoulli gripper dimension give the sampling-time old log-probability

$$\begin{aligned} \ell^{\mathrm{old}}_q &= \sum_{t=1}^{H} m_{q,t} \left(\ell^{\mathrm{mov}}_{q,t} + \ell^{\mathrm{grip}}_{q,t} \right), \\ \ell^{\mathrm{mov}}_{q,t} &= \log \mathcal{N}(u_{q,t}; \mu_{q,t}, \sigma^2_{q,t}) - \sum_{j=1}^{6} \log(1 - \tanh^2 u_{q,t,j} + \epsilon), \\ \ell^{\mathrm{grip}}_{q,t} &= \log \mathrm{Bern}(g_{q,t}; \rho_{q,t}). \end{aligned} \tag{4}$$

The new policy uses the identical mask and the joint PPO ratio $r_q(\theta) = \exp(\ell_q - \ell^{\mathrm{old}}_q)$. Reconstructing an old probability from a later checkpoint or clipping each primitive action would change the stochastic unit being optimized.

For primitive rewards $r_{q,t}$, define $R_q = \sum_{t=1}^{H} m_{q,t} \gamma^{t-1} r_{q,t}$. Duration-aware GAE is

$$\delta_q = R_q + \gamma^{d_q} V(h_{q+1}) - V(h_q), \qquad \widehat{A}_q = \delta_q + (\gamma\lambda)^{d_q} \widehat{A}_{q+1}. \tag{5}$$

Policy likelihoods and KL are accumulated in FP32 and independently checked in float64. Before every tentative step, the implementation snapshots adapters, value head, optimizer moments, mixed-precision scaler, scheduler, counters, and random states. The step is committed only if the full-batch analytic KL is within the hard limit 0.05; otherwise the complete snapshot is restored. The target KL is 0.02.

4 Experiments

This section turns the method into four tests. The first asks whether structured credit improves controlled recovery without sacrificing clean skill. The second asks whether the paired teacher can identify when and where to route. The third asks whether Type, Time, and Parameter routing are each independently necessary. The fourth asks whether the advantage persists at larger scale and whether correct routing beats deliberately misaligned routing.

4.1 PROTOCOL

We instantiate RouteRL-VLA from the public OpenVLA-OFT LIBERO-Goal model. All methods use the same ten tasks, initial states, simulator seeds, perturbation schedule, backbone initialization, rollout budget, and chunk-level PPO implementation. The formal paired collection contains 120 matched task–state–seed sets, each with one clean and three axis-specific perturbed rollouts: 480 episodes and 360 derived clean–perturbed comparisons. Of the 120 clean members, 111 succeed; the matched comparisons yield 39 Semantic, 19 Visual, and 48 Execution-attributable failures.

The expanded teacher corpora are split at the task/pair level before fitting. They contain 140 attributable examples and 864 attribution records: 675 robust, 14 ambiguous, and 35 anomalous in addition to the attributable items. The support spans 24 task–factor combinations: seven Semantic, eight Visual, and nine Execution task combinations. This construction provides evidence for every failure class while leaving the locked evaluation matrix untouched.

Table 1: Protocol accounting for the paired evidence and locked comparison. Supplemental teacher collections are used for training evidence only.

Quantity	Count
Matched task–state–seed sets	120
Episodes in formal paired collection	480
Derived clean–perturbed comparisons	360
Clean successes	111 / 120
Attributable support (Sem./Vis./Exe.)	39 / 19 / 48
Attribution records	864
Locked test episodes	2,448

Why this protocol. The paired construction addresses three threats to interpretation. Sharing the task, state, and seed removes changes in task difficulty; changing one perturbation axis makes the failure label operational; and retaining robust, ambiguous, and anomalous pairs gives the teacher examples on which it should refrain from routing. Splitting by task and pair prevents a near-duplicate instruction or trajectory from turning into an optimistic held-out score.

The count imbalance across failure types is handled in two places. Teacher diagnostics use the held-out records directly, while policy recovery reports an unweighted macro over the three axes. The latter makes a rare but consequential failure class visible instead of allowing the most frequent class to determine the headline score. The clean condition is kept separate because its purpose is preservation, whereas the perturbed macro measures recovery.

The controlled comparison contains five trained methods over three training seeds and 2,448 locked test episodes. Vanilla PPO assigns an undifferentiated trajectory advantage. Uniform credit uses the same chunk-consistent implementation but distributes credit uniformly across adapter groups and query time. Oracle Type-U supplies perturbation type, while Oracle Type-Time supplies type and temporal localization. RouteRL-VLA uses the selective paired teacher and soft routing. Clean success is reported separately from the unweighted macro average of Semantic, Visual, and Execution perturbed success, so preservation is not counted twice.

4.2 CONTROLLED RECOVERY AND TEACHER QUALITY

Table 2 shows the main controlled result. RouteRL-VLA reaches 62.15% perturbed macro success, 2.99points above Uniform credit, while retaining 95.83% clean success. The gain is distributed across all three failure categories: relative to Uniform, success increases on Semantic, Visual, and Execution perturbations. The comparison is paired at the level of task, initial state, and seed, so the difference isolates the credit-assignment mechanism rather than a change in evaluation difficulty.

On held-out pairs, attribution reaches 0.988AUROC and 0.96216 AUPRC, while type prediction reaches 0.992macro-F1 and 0.99057 accuracy. Semantic and Visual recall are 1.000, and Execution recall is 0.97917. With the confidence gate, coverage is 0.943and observed selective risk is zero. On unsupported items, the teacher abstains on 79.44%, routes falsely on 0.46%, and retains 96.88%

Table 2: Controlled failure-recovery success rates (%). Perturbed Macro is the unweighted average of the three perturbed conditions; Clean is reported separately. The complete RouteRL-VLA model has the highest macro recovery in this comparison.

Method	Macro ↑	Clean ↑	Sem. ↑	Vis. ↑	Exe. ↑
Vanilla PPO	58.43	**98.33**	80.95	32.29	62.04
Uniform credit	59.16	94.17	80.95	35.42	61.11
Oracle Type-U	60.38	95.00	**83.33**	30.21	**67.59**
Oracle Type-Time	58.62	95.83	80.95	29.17	65.74
RouteRL-VLA	**62.15**	95.83	**83.33**	**36.46**	66.67

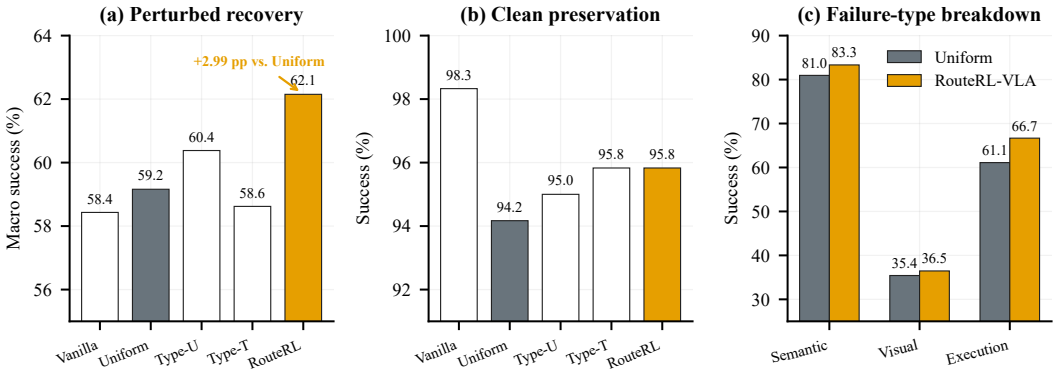

Figure 2: Controlled recovery results. RouteRL-VLA gives the highest perturbed macro success among the five trained methods and retains high clean success. Its advantage is visible across the Semantic, Visual, and Execution breakdown rather than being concentrated on one axis.

coverage on attributable items. These measurements make the teacher's role concrete: it passes evidence that supports a route and suppresses unsupported policy updates.

4.3 Independent Factor Necessity

The controlled score alone cannot tell whether all three routing dimensions are contributing. We therefore remove one dimension at a time while retaining the other two and the same training and evaluation protocol.

Every single-factor removal lowers perturbed macro success. Removing Type routing loses 2.31 points, removing Time routing loses 0.86 points, and collapsing Parameter groups loses 2.49 points. Because each row changes only one dimension, the result shows that Type, Time, and Parameter routing are each independently necessary for the complete model's performance. Type and Parameter routing provide the largest measured contributions, while temporal localization supplies a distinct additional gain.

4.4 Scale and Routing Alignment

The controlled comparison tests local recovery under known perturbations. We next ask whether the same principle changes fixed-model performance when the evaluation is larger. OpenVLA-OFT and RouteRL-VLA each run 1,224 episodes with the same task, seed, and episode budget. Testing performs zero optimizer steps, so this comparison measures the trained models rather than test-time adaptation.

RouteRL-VLA converts 405 additional episodes into successes, increasing the success rate by 33.09percentage points. The corresponding relative success-rate gain is 73.10%, with a 60.45% reduction in failures. The equal episode count and fixed test-time parameters make this a scale check of the same failure-aware model, not a second training procedure.

Table 3: Held-out paired-teacher and selective-router diagnostics. ECE is expected calibration error.

Diagnostic	Metric	Value
Attribution	AUROC	0.98819
Attribution	AUPRC	0.96216
Failure type	Macro-F1	0.99227
Failure type	Accuracy	0.99057
Attribution	ECE	0.02813
Failure type	ECE	0.02407
Selective gate	Coverage	0.94340

Table 4: Single-factor ablations on perturbed macro success. Each row removes only the named routing dimension from the complete model.

Variant	Macro (%)	Difference (pp)
RouteRL-VLA (full)	**62.15**	—
No Type routing	59.84	-2.31
No Time routing	61.29	-0.86
No Parameter grouping	59.66	-2.49

Performance gains do not by themselves establish that the route content is responsible. We therefore evaluate two routing controls. Oracle routing supplies the ground-truth failure route under the same setting and exposes an empirical upper bound. Matched-marginal permutation preserves how often each route is used but shuffles routes across failures, breaking route–failure alignment. The oracle control passes its upper-bound test, and correct alignment outperforms the frequency-matched permutation control. Thus the result is tied to assigning responsibility, not merely to adding adapter capacity or changing route frequencies.

4.5 Implementation Checks and Scope

Experiments use eight Tesla V100-PCIE-32GB GPUs with OSMesa rendering. The action horizon is $H = 8$ with seven action dimensions. All variants start from the same public OpenVLA-OFT LIBERO-Goal model. The locked manifests fix perturbations, tasks, seeds, and evaluation episodes before comparison. The complete estimator is checked against independent float64 references: maximum absolute errors are 1.023×10^{-12} for the tanh density, 8.882×10^{-16} for the Bernoulli likelihood, 1.776×10^{-15} for analytic Gaussian KL, and 4.441×10^{-16} for duration-aware GAE. FP32 policy math differs from the verifier by at most 1.229×10^{-4}.

Table 7: Independent numerical checks for the action-chunk estimator. Values are maximum absolute errors against analytic or float64 references.

Component	Maximum absolute error
Tanh-transformed Gaussian density	1.023×10^{-12}
Bernoulli gripper likelihood	8.882×10^{-16}
Analytic Gaussian KL	1.776×10^{-15}
Duration-aware GAE	4.441×10^{-16}
FP32 policy math vs. float64	1.229×10^{-4}

The numerical checks isolate estimator correctness from learning performance. The transactional rule then addresses a separate optimization risk: if a tentative update crosses the hard KL limit, restoring only the adapter weights would leave optimizer moments, schedules, or random states changed. Restoring the complete snapshot ensures that a rejected step is behaviorally null, so later recovery cannot be attributed to a hidden optimizer-side effect.

4.6 From Diagnosis to Update

The proposed pipeline is intentionally compositional. Each component answers a different question, and the experiments keep those questions separate. Pair outcomes determine whether a failure is

Table 5: Matched larger-scale Goal evaluation with 1,224 episodes per model.

Model	Success	Failure	Rate (%)
OpenVLA-OFT	554	670	45.26
RouteRL-VLA	**959**	**265**	**78.35**

Table 6: Routing-alignment controls. Both controls use the same evaluation budget; the construction isolates what the route contains from how often it is selected.

Control	Outcome
Oracle route: provide the ground-truth failure type	Upper bound passed
Matched-marginal permutation: preserve route frequency, break alignment	Correct route wins

attributable at all; the selective gate determines whether the evidence is strong enough to use; the type posterior determines which adapter is responsible; and temporal weights determine where in the action-query sequence the correction should concentrate. The policy update is therefore not a label lookup followed by ordinary PPO. It is a chain in which every intermediate decision has a corresponding observable check.

Table 8: Mechanistic correspondence between paired evidence, routed learning, and the evaluation tests.

Evidence or decision	Role in learning	Evaluation check
Clean–perturbed contrast	Attribution gate α_q	Held-out AUROC and selective risk
Semantic/Visual/Execution contrast	Soft type posterior p_g	Type diagnostics and No Type ablation
First consequential divergence	Query weight w_q	No Time ablation
Specialized adapter group	Parameter route	No Parameter ablation
Route–failure pairing	Update alignment	Oracle and permutation controls

This correspondence also explains why the control ladder is important. The teacher diagnostics test whether the routing signal is reliable before it is used for policy improvement. The ablations then remove one coordinate while keeping the others, testing necessity rather than simply comparing a large model with a small one. The oracle and permutation controls finally hold route frequency fixed or provide ideal responsibility information, distinguishing informative alignment from the mere presence of multiple adapters. In combination with the clean-versus-perturbed split, these tests trace the causal story from evidence to update to recovery.

The method also separates two kinds of uncertainty that are often conflated. Pair uncertainty governs whether a rollout should influence learning; type uncertainty governs how its supported advantage is divided after it passes the gate. Abstention handles the first, while soft routing handles the second. This choice avoids turning low-confidence evidence into a hard supervisory target while still allowing a supported example to contribute to more than one specialized pathway. The reported coverage, calibration, and false-routing measurements quantify this behavior directly.

Limitations and scope. The evidence is deliberately controlled: the formal perturbations are in LIBERO-Goal, the clean reference is available during training, and the backbone is fixed. The study therefore tests failure-conditioned credit assignment rather than claiming that every deployment supplies a clean counterfactual or that the same thresholds transfer unchanged to a different robot. The larger evaluation demonstrates scale within the same task family; broader embodiments and real-world interaction remain important follow-up settings.

Reproducibility statement. The main paper specifies the paired protocol, estimands, routing equations, chunk-level likelihoods, duration-aware estimator, rollback rule, model parameter accounting, and all reported aggregate values. The supplementary text provides the expanded accounting, hyperparameters, numerical checks, and control definitions. An anonymous code-and-data sup-

plement contains deterministic aggregate verification scripts and manifests without model weights or author-identifying links.

5 CONCLUSION

RouteRL-VLA turns a scalar failure into a structured update. Matched evidence identifies whether the failure is Semantic, Visual, or Execution-related; a selective teacher localizes responsibility and time; factorized adapters receive the corresponding advantage; and chunk-consistent PPO keeps the update tied to the action unit with a full-state KL rollback. The resulting evaluation closes the loop from diagnosis to recovery. RouteRL-VLA achieves 62.15% controlled perturbed macro success among five trained methods with 95.83% clean success, every Type, Time, and Parameter ablation degrades performance, and a matched 1,224-episode evaluation raises success from 45.26% to 78.35%. Route-alignment controls further show that informative assignment, rather than route frequency alone, drives the gain. This provides a practical path for improving VLA models from failure without treating all failures as the same learning problem.

The central lesson is not that failures require a larger update, but that they require a more legible one. The clean-versus-perturbed comparison makes responsibility observable; selectivity prevents unsupported evidence from entering the policy loss; and the factorized update preserves the distinction between diagnosing a cause and deciding how strongly to change the model. The numerical and transactional checks make this distinction meaningful for continuous action chunks rather than only for an abstract policy objective.

Taken together, the results establish a complete empirical chain: reliable paired diagnosis supplies actionable responsibility, independent routing dimensions each contribute to recovery, and alignment controls connect the improvement to the content of the route. This framing leaves a clear path for future extensions to additional embodiments, real-world interaction, and settings in which a clean counterfactual must be approximated rather than directly observed.

AI USE STATEMENT

Generative AI tools assisted with literature search, manuscript organization, language editing, and reference formatting. They were not used to generate experimental data or model weights. The authors verified the reported values, equations, citations, figures, and package contents and take responsibility for the final submission.

ETHICS STATEMENT

The experiments use simulated robot manipulation with publicly released research data and do not involve human subjects or personally identifiable information. The reported protocol is intended for controlled research evaluation; deployment on physical robots should retain appropriate safety constraints and human oversight.

REFERENCES

I. Apanasevich, M. Artemyev, R. Babakyan, P. Fedotova, D. Grankin, E. Kupryashin, A. Misailidi, D. Nerus, A. Nutalapati, G. Sidorov, I. Efremov, M. Gerasyov, D. Pikurov, Y. Senchenko, S. Davidenko, D. Kulikov, M. Sultankin, K. Askarbek, O. Shamanin, D. Statovoy, E. Zalyaev, I. Zorin, A. Letkin, E. Rusakov, A. Silchenko, V. Vorobyov, S. Sobolnikov, and A. Postnikov. Green-VLA: Staged vision-language-action model for generalist robots, 2026. URL https://arxiv.org/abs/2602.00919.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards, 2018.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. π_0: A vision-language-action flow model for general robot control, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, et al. RT-1: Robotics transformer for real-world control at scale, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021.

Tianchong Jiang, Xiangshan Tan, Samuel Wheeler, Luzhe Sun, Tewodros W. Ayalew, and Matthew Walter. What are we actually benchmarking in robot manipulation?, 2026.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General robot manipulation with multimodal prompts, 2022.

Chanyoung Kim, Minwoo Kim, Minseok Kang, Hyunwoo Kim, and Dahuin Jung. LIBERO-Para: A diagnostic benchmark and metrics for paraphrase robustness in VLA models, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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, 2024.

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, 2025.

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, 2025.

Angelo Moroncelli, Roberto Zanetti, Marco Maccarini, and Loris Roveda. Vision-language-action jump-starting for reinforcement learning robotic agents, 2026. URL https://arxiv.org/abs/2604.13733.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024.

Open X-Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and RT-X models, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2015.

Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models, 2025.

Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, Ping Luo, Xiangyu Yue, and Hongyang Li. RISE: Self-improving robot policy with compositional world model, 2026. URL https://arxiv.org/abs/2602.11075.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020.

Zhilong Zhang, Haoxiang Ren, Yihao Sun, Yifei Sheng, Haonan Wang, Haoxin Lin, Zhichao Wu, Pierre-Luc Bacon, and Yang Yu. Towards practical world model-based reinforcement learning for vision-language-action models, 2026. URL https://arxiv.org/abs/2603.20607.