NOT ALL EXPERIENCES ARE EQUAL: MEASURING BUDGET-CONSERVING LEARNING VALUE FOR MULTI-TURN AGENT RL

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy reinforcement learning for multi-turn language agents largely optimizes the update rule (advantages, group comparisons, KL) while treating *which* experiences deserve update mass as a default. Return R and advantage A record how an episode ended and how strongly the current surrogate wants to reinforce it. Neither answers the counterfactual that matters under a conserved budget: whether giving one experience more update mass, taken from its neighbors, improves held-out capability. We introduce **EVA** (**E**xperience **V**aluation and **A**ugmentation), a data-centric layer in front of an unchanged $\mathcal{U}_{\mathrm{RL}}$, and its measurement core **IA-ELV**. IA-ELV scores each experience by a budget-preserving probe that returns gold coordinates $(\Delta J, \Delta H)$, namely meta-performance change and decision-uncertainty reduction. On ALFWorld and WebShop, at 1.5B and 7B, learning value is sparse and peaks at mid-training, where return and advantage fail to recover ΔJ. A leak-free predictor ranks the gold object with the same sign, so EVA can sit in front of GiGPO without a probe at every step. Gold Mid allocation, and the same rule under Mid-continue GiGPO, beat uniform weighting across both environments and both scales; Mid is the strongest stage. Soft probes keep learning value distinct from augmentation gain. Legally recollected four-action routing then adds held-out value beyond reweighting alone. EVA therefore separates outcome, credit, learning value, and augmentation value under explicit conservation constraints. Our code is available at `https://anonymous.4open.science/r/eva-3B8C`.

1 INTRODUCTION

On-policy reinforcement learning is now a standard way to train language-model agents for multi-turn interaction Yao et al. (2023); Shridhar et al. (2021); Yao et al. (2022); Ouyang et al. (2022); Shao et al. (2024); Feng et al. (2025). The usual focus is the update rule: how to estimate advantages, compare completions inside a group, and set a KL penalty Schulman et al. (2017; 2016); Ahmadian et al. (2024); Shao et al. (2024); Feng et al. (2025). That work has made credit assignment more stable. It has not asked a prior question. Once a batch has been collected, which trajectories should receive more of a *fixed* update budget, and which should be left alone or transformed?

The default answers are uniform weights, or a rank by return R or absolute advantage $|A|$. Neither is a statement about future capability. A successful episode can be almost a copy of others already in the group. A large advantage can lock in a local fix that does not transfer. A failure with return zero can still mark a decision state the policy has not learned. R says how the episode ended. A says how strongly the current surrogate wants to reinforce an action Schulman et al. (2016). The quantity we need is different: if this experience is given more mass, taken from its neighbors, does held-out behavior improve? On Mid-stage pools that gap is not academic. Both R and $|A|$ are negatively correlated with the measured effect of a budget-conserving reweight, so following them moves mass onto the wrong trajectories (Appendix C.1).

Trainers also rewrite goals or branch from an intermediate state. Those operators cost extra interaction, can break validity, and can leave the batch off policy. An experience that is worth learning from in raw form need not repay that cost. We keep four objects separate,

$$R \neq A \neq \mathrm{ELV} \neq \mathrm{EAG}, \tag{1}$$

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.

where ELV is the learning value of the original experience and EAG is the extra net value of applying an operator.

This paper measures those objects and puts the scores in front of an unchanged $\mathcal{U}_{\mathrm{RL}}$. The core is **IA-ELV** (Information-Aware Experience Learning Value). For an experience e_i inside a local context C_i of on-policy rollouts that share a task prompt, a budget-preserving probe compares a reference update with default weights to a focus update that moves a fixed mass δ onto e_i. The gold coordinates are held-out performance change ΔJ and decision-uncertainty reduction ΔH, with a policy-change term kept as a risk signal rather than as utility. The two branches spend the same optimization budget, so their difference is reallocation, not extra training (Figure 2, Appendix B). A leak-free predictor then ranks $(\Delta J, \Delta H)$ from features that exist before the probe (Appendix D). **EVA** (**E**xperience **V**aluation and **A**ugmentation) is the layer around that estimator. It can downweight, keep, paraphrase, or branch; the main closed loops use conservative reweighting, and a separate ALFWorld 1.5B Mid study tests legally recollected four-action routing (Appendix F). We evaluate on ALFWorld and official WebShop with 1.5B and 7B GiGPO agents. Our contributions are summarized as follows.

- We show that return and advantage are not learning value under a matched budget. At mid-training both are negatively correlated with the held-out effect of a conserved reweight, so ranking by R or $|A|$ moves mass onto the wrong trajectories.
- We measure that effect with IA-ELV, a paired probe that returns $(\Delta J, \Delta H)$ without extra update mass, so the score is reallocation rather than more training. A leak-free predictor then ranks the same object from features available before the probe.
- We place EVA in front of unchanged GiGPO. Gold Mid allocation, and the same rule under Mid-continue training, beat uniform weighting on both environments and both scales; Mid is the strongest stage.
- We keep worth learning distinct from worth transforming ($R \neq A \neq \mathrm{ELV} \neq \mathrm{EAG}$). Soft probes separate ELV from EAG, and on ALFWorld 1.5B Mid, legally recollected four-action routing still adds held-out value after reweighting is already in place.

2 Related Work

On-policy RL for multi-turn language agents. Language agents that interleave reasoning and interaction follow the ReAct paradigm (Yao et al., 2023), with ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022) as standard long-horizon testbeds. Preference tuning and on-policy RL then improve behavior via PPO (Ouyang et al., 2022; Schulman et al., 2017), leave-one-out REINFORCE (Ahmadian et al., 2024), group-relative GRPO (Shao et al., 2024), hierarchical GiGPO credit (Feng et al., 2025), and turn-level PPO (Li et al., 2026). WebRL trains open LLM web agents with a self-evolving curriculum and an outcome reward (Qi et al., 2025). A curriculum chooses which goals to attempt next; an advantage estimator such as GAE (Schulman et al., 2016) says how strongly the current surrogate should reinforce an action already inside the batch. All of these methods redesign *how* credit is assigned, or *which tasks* to sample. We leave $\mathcal{U}_{\mathrm{RL}}$ unchanged (GiGPO in our experiments) and ask, after those tasks have already been rolled out, which completed trajectories should receive more mass under a matched interaction and update budget.

Experience valuation and data selection. Non-uniform data utility is a long-standing theme. Prioritized Experience Replay (Schaul et al., 2016) samples off-policy transitions by TD error, not by the effect of reallocating a fixed on-policy update. DSIR resamples pretraining documents to match a target n-gram distribution (Xie et al., 2023). Rho-1 drops tokens that a reference model already fits (Lin et al., 2024). Data Shapley (Ghorbani & Zou, 2019), influence functions (Koh & Liang, 2017), TracIn (Pruthi et al., 2020), and LESS (Xia et al., 2024) score examples by contribution, influence, or gradient similarity. Those scores answer influence, corpus matching, token selection, or sampling priority. IA-ELV asks a narrower counterfactual: whether giving one member of an on-policy group extra update mass, taken from its neighbors, improves held-out capability.

Attribution and selection in online RL. Recent work brings attribution into online RL. Snapshot of Influence (Hu et al., 2025) filters PPO buffers by local gradient similarity; InfOES (Gong et al., 2026) removes harmful RLHF experiences; CROPI (Zhu et al., 2026) selects influential data for

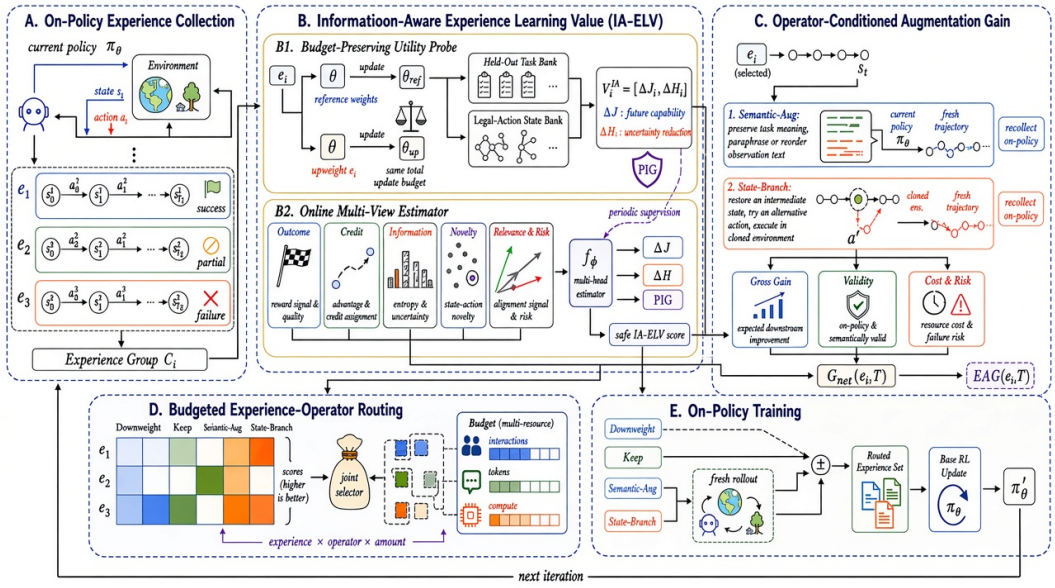

Figure 1: **EVA** as a data-centric layer in front of an unchanged $\mathcal{U}_{\mathrm{RL}}$. The policy collects grouped on-policy experiences (A). IA-ELV scores each experience by a budget-preserving probe that yields $(\Delta J, \Delta H)$ and a separate risk signal (B). The full design can also estimate operator-conditioned EAG and choose among four actions (C, D) before the base update (E). This paper instantiates (B) together with conservative reweighting. Legal recollection of (C) and (D) is tested on ALFWorld 1.5B Mid.

RLVR. These methods, like EVA, treat experiences as unequal at training time. The estimand differs: influence scores a sample's effect on a specified optimization or prediction target, whereas IA-ELV measures a conserved reallocation and can be compared directly to uniform weighting. We do not replace GiGPO's credit rule. We add a budget-conserving allocation layer in front of it.

Experience augmentation. Hindsight Experience Replay (Andrychowicz et al., 2017) turns failed trajectories into useful signal by goal relabeling. Most such transforms are applied as a fixed data-generation recipe, not as a choice conditioned on the source experience. HER treats the rewrite as training data by construction. EVA asks whether the rewrite is worth the extra interaction relative to leaving the source unchanged, and separates learning from an experience (ELV) from transforming it (EAG). Matched soft probes support ELV $\neq$ EAG. A separate ALFWorld 1.5B Mid study tests legally recollected four-action routing against Uniform and Downweight+Keep.

3 Method

3.1 Overview

We propose **EVA** (**E**xperience **V**aluation and **A**ugmentation), a data-centric layer in front of an unchanged on-policy update $\mathcal{U}_{\mathrm{RL}}$. EVA does not introduce a new policy-gradient surrogate. It decides which collected experiences receive update mass under a fixed budget.

At iteration t, the current policy π_{θ_t} collects $\mathcal{D}_t = \{e_1, \ldots, e_N\}$. In the instantiation used here, each e_i is a complete trajectory. After valuation,

$$\theta_{t+1} = \mathcal{U}_{\mathrm{RL}}\Big(\theta_t, \widetilde{\mathcal{D}}_t, \mathbf{w}_t\Big), \tag{2}$$

where $\widetilde{\mathcal{D}}_t$ is the routed set and $\mathbf{w}_t$ its outer weights. The experiments use GiGPO as $\mathcal{U}_{\mathrm{RL}}$. Placing the layer in front of the update is deliberate: a new advantage, group comparison, or KL term would confound allocation with a new surrogate. EVA therefore edits only outer trajectory weights, or, when an operator is chosen, the experiences that enter the same update.

The measurement core is *Information-Aware Experience Learning Value* (IA-ELV): whether giving an experience more update mass, taken from its local neighbors, improves held-out capability. *Experience Augmentation Gain* (EAG) is a separate object: whether applying an operator T_k adds net value beyond that raw experience. The full design can then choose

$$a_i \in \mathcal{A} = \{\text{DOWNWEIGHT}, \text{KEEP}, \text{SEMANTIC-AUG}, \text{STATE-BRANCH}\}. \tag{3}$$

That four-action loop is specified so the interface is complete (Appendix A.7). It is not the empirical training system. The intended full pipeline is

$$\mathcal{D}_t \;\to\; \widehat{\text{IA-ELV}} \;\to\; \widehat{\text{EAG}} \;\to\; \text{budgeted routing} \;\to\; \widetilde{\mathcal{D}}_t \;\to\; \mathcal{U}_{\text{RL}}. \tag{4}$$

The four quantities remain distinct:

$$R(e_i) \;\neq\; A(e_i) \;\neq\; \text{ELV}(e_i) \;\neq\; \text{EAG}(e_i, T_k). \tag{5}$$

Return records the outcome. Advantage records local credit under the current surrogate. ELV records the consequence of reallocating a conserved update. EAG records the incremental value of transforming the source. The gold IA-ELV probe (Algorithm 1) is the scientific object of this paper.

3.2 INFORMATION-AWARE EXPERIENCE VALUATION

Context. Learning value depends on the current policy, the update, and the experiences that compete for the same budget. For a target e_i, every comparison is local to a context of on-policy rollouts that share one task prompt,

$$C_i = \{e_i, e_{i,2}, \ldots, e_{i,n}\} \subseteq \mathcal{D}_t, \tag{6}$$

with $n = |C_i|$. IA-ELV is an allocation value inside C_i, not a permanent score on a trajectory. The same rollout can be worth extra mass in a group of near-copies and harmful in a group that already covers its decision states.

Budget-preserving probe. Let $\mathbf{w}^{(0)}$ be the default weights of $\mathcal{U}_{\text{RL}}$, with $w_j^{(0)} = 1/n$ in the main setting. The reference update is

$$\theta_{\text{ref}} = \theta^+_{C_i} = \mathcal{U}_{\text{RL}}\Big(\theta_t, C_i, \mathbf{w}^{(0)}\Big). \tag{7}$$

The focus weights move mass $\delta > 0$ onto e_i and remove the same total mass from the remaining context:

$$w_i^{(+i)} = w_i^{(0)} + \delta, \qquad \sum_{j=1}^{n} w_j^{(+i)} = \sum_{j=1}^{n} w_j^{(0)}, \qquad w_j^{(+i)} \geq 0. \tag{8}$$

Under uniform reference weights this is $w_i^{(+i)} = 1/n + \delta$ and $w_{j\neq i}^{(+i)} = 1/n - \delta/(n-1)$. The focus update is

$$\theta_{\text{focus}} = \theta^+_{C_i,i} = \mathcal{U}_{\text{RL}}\Big(\theta_t, C_i, \mathbf{w}^{(+i)}\Big). \tag{9}$$

The two branches share θ_t, optimizer state, learning rate, step count, clipping, evaluation banks, and paired seeds. Their difference isolates reallocation rather than extra optimization.

Performance and information coordinates. Both policies are scored on a frozen held-out task catalog $\mathcal{M}_J$ and a frozen legal-action state bank $\mathcal{B}_M$ (Appendix B.2). Write $J_{\text{ref}} = J_{\text{meta}}(\theta_{\text{ref}}; \mathcal{M}_J)$ and likewise for the focus branch. The performance coordinate is

$$\Delta J_i = J_{\text{focus}} - J_{\text{ref}} = J_{\text{meta}}(\theta^+_{C_i,i}; \mathcal{M}_J) - J_{\text{meta}}(\theta^+_{C_i}; \mathcal{M}_J). \tag{10}$$

A positive ΔJ_i means that emphasizing e_i improves held-out capability under the same update budget.

Sparse task scores can miss intermediate decision changes. On $\mathcal{B}_M$ we form a temperature-τ_H softmax over admissible action strings and write H_{meta} for its mean entropy (Appendix B.3). The information coordinate is

$$\Delta H_i = H_{\text{ref}} - H_{\text{focus}} = H_{\text{meta}}(\theta^+_{C_i}; \mathcal{B}_M) - H_{\text{meta}}(\theta^+_{C_i,i}; \mathcal{B}_M). \tag{11}$$

A positive ΔH_i means that emphasizing e_i reduces decision uncertainty on the fixed bank. The gold label is the pair

$$\mathbf{V}_i^{\text{IA}} = \begin{bmatrix} \Delta J_i \\ \Delta H_i \end{bmatrix}. \tag{12}$$

It is not a sum of reward, advantage, and entropy (Figure 2, Appendix B).

Risk and deployed score. We also record policy displacement

$$\mathrm{PIG}_i = \frac{1}{|\mathcal{B}_M|} \sum_{s \in \mathcal{B}_M} D_{\mathrm{JS}}(p_{\theta_{\mathrm{focus}}}(\cdot \mid s) \,\|\, p_{\theta_{\mathrm{ref}}}(\cdot \mid s)) \,. \tag{13}$$

PIG is a risk signal, not a third utility (Appendix B.4). Large movement without positive ΔJ_i is treated as disruptive. An experience may receive extra mass only if

$$\Delta J_i \geq -\varepsilon_J, \qquad \mathrm{PIG}_i \leq \kappa. \tag{14}$$

For eligible members the routing score is

$$S_i = z(\Delta J_i) + \beta_t z(\Delta H_i), \tag{15}$$

with $z(\cdot)$ computed inside the context. Online runs use $\beta_t = 0$ unless an independent gate shows that ΔH adds reliable signal. The deployed score is therefore ΔJ-primary.

Algorithm 1 Budget-preserving IA-ELV

Require: Policy θ_t, context $C = \{e_1, \ldots, e_n\}$, transfer $\delta > 0$, update $\mathcal{U}_{\mathrm{RL}}$, frozen banks $\mathcal{M}_J$ and $\mathcal{B}_M$
Ensure: Gold labels $\{\mathbf{V}_i^{\mathrm{IA}}, \mathrm{PIG}_i\}_{i=1}^n$

$\mathbf{w}^{(0)} \leftarrow (1/n, \ldots, 1/n)$
$\theta_{\mathrm{ref}} \leftarrow \mathcal{U}_{\mathrm{RL}}(\theta_t, C, \mathbf{w}^{(0)})$
$J_{\mathrm{ref}} \leftarrow J_{\mathrm{meta}}(\theta_{\mathrm{ref}}; \mathcal{M}_J)$
$H_{\mathrm{ref}} \leftarrow H_{\mathrm{meta}}(\theta_{\mathrm{ref}}; \mathcal{B}_M)$
for $e_i \in C$ **do**
    $w_i^{(+i)} \leftarrow 1/n + \delta, \quad w_{j \neq i}^{(+i)} \leftarrow 1/n - \delta/(n-1)$
    if $\min_j w_j^{(+i)} < 0$ **then**
        mark e_i invalid
    else
        $\theta_{\mathrm{focus}} \leftarrow \mathcal{U}_{\mathrm{RL}}(\theta_t, C, \mathbf{w}^{(+i)})$
        $\Delta J_i \leftarrow J_{\mathrm{meta}}(\theta_{\mathrm{focus}}; \mathcal{M}_J) - J_{\mathrm{ref}}$
        $\Delta H_i \leftarrow H_{\mathrm{ref}} - H_{\mathrm{meta}}(\theta_{\mathrm{focus}}; \mathcal{B}_M)$
        $\mathrm{PIG}_i \leftarrow \mathrm{mean}_{s \in \mathcal{B}_M} D_{\mathrm{JS}}\big(p_{\theta_{\mathrm{focus}}}(\cdot \mid s) \,\big\|\, p_{\theta_{\mathrm{ref}}}(\cdot \mid s)\big)$
        $\mathbf{V}_i^{\mathrm{IA}} \leftarrow [\Delta J_i, \Delta H_i]^\top$
    end if
end for
return $\{\mathbf{V}_i^{\mathrm{IA}}, \mathrm{PIG}_i\}_{i=1}^n$

Algorithm 1 is this paired intervention written as a procedure. The target stays in C on both branches, so ΔJ_i measures emphasis rather than inclusion versus deletion. The reference update is recomputed for each pair, which keeps stochastic optimizer effects matched to the focus branch. Multi-δ curves reuse the same difference at several transfer sizes (Appendix B).

3.3 Estimating IA-ELV Online

A probe on every experience is too expensive to run inside training. We therefore fit

$$\left[\widehat{\Delta J}_i, \widehat{\Delta H}_i\right] = f_\phi(\mathbf{x}_i) \tag{16}$$

from features available before the probe.

The deployable view contains only quantities that exist when a new rollout is collected: return and binary success, aggregate advantage statistics, length and termination, and within-context ranks, centered statistics, and dispersion. Task type and training stage may enter as group descriptors. Gold ΔJ, ΔH, PIG, constraint flags, and wrong-but-confident indicators are excluded, because each is a probe outcome (Appendix D). Return and $|A|$ may appear as features. They do not become the label. The predictor is asked to recover the probe coordinate, not to copy the surrogate already inside GiGPO.

The estimator is a dual-head ridge regressor trained on gold labels. Ridge is enough because the router uses within-context ranks, not calibrated magnitudes. Validation leaves out entire contexts C_i, so related rollouts cannot appear on both sides of a fold. Predicted coordinates replace gold values in Equation 15 and inherit the same safety interface. Online runs further use the conservative score $\ell = \hat{\mu} - \lambda\hat{\sigma}$, which downweights predictions whose residual scale is large.

3.4 Operator-Conditioned Augmentation Gain

High ELV does not imply that an experience should be transformed. SEMANTIC-AUG applies a meaning-preserving rewrite to task language or observations. STATE-BRANCH restores a valid intermediate state, takes another admissible action, and continues interaction. Those branch points become training experiences only after fresh environment interaction. Identity leaves the source unchanged. For a candidate $\widetilde{e}_{i,k} \sim T_k(e_i)$,

$$G^{\mathrm{net}}_{i,k} = \Delta J(\widetilde{e}_{i,k}) - \Delta J(e_i) - \lambda_C \,\mathrm{cost}(T_k, e_i). \tag{17}$$

The question is whether the operator adds downstream value after cost. Validity filters are not a substitute for this difference: a legal rewrite or first-action swap can have the opposite ΔJ of the source ELV.

An online model may estimate $\widehat{G}^{\mathrm{net}}_{i,k} = g_\psi(\mathbf{x}_i, T_k, \widehat{\mathbf{V}}^{\mathrm{IA}}_i)$. Legal training still requires recollection with π_{θ_t} and a fresh computation of actions, rewards, log-probabilities, and advantages. The soft probes in Section 4 test ELV $\neq$ EAG. Legal four-action training is a separate ALFWorld 1.5B Mid study (Appendix F, Appendix F.6).

3.5 Conservative Reweighting and the Design Router

The experiments isolate IA-ELV from augmentation by conservative reweighting. Inside each context,

$$\widetilde{w}_j = \frac{\exp(\widehat{S}_j/\tau_w)}{\sum_\ell \exp(\widehat{S}_\ell/\tau_w)}, \qquad w_j = (1-\alpha)w^{(0)}_j + \alpha\widetilde{w}_j. \tag{18}$$

The case $\alpha = 0$ recovers the base optimizer. Gold or predicted scores are compared with uniform, return, and $|A|$ weights under the same update budget.

The full design can also choose an operator. Let $q_{\mathrm{down}}(e_i)$ be a sentinel low enough to force DOWNWEIGHT when the safety interface fails. For action $a_i \in \mathcal{A}$,

$$Q(e_i, a_i) = \begin{cases} q_{\mathrm{down}}(e_i), & a_i = \textsc{Downweight}, \\ \widehat{S}_i, & a_i = \textsc{Keep}, \\ \widehat{S}_i + \widehat{G}^{\mathrm{net}}_{i,\mathrm{sem}}, & a_i = \textsc{Semantic-Aug}, \\ \widehat{S}_i + \widehat{G}^{\mathrm{net}}_{i,\mathrm{branch}}, & a_i = \textsc{State-Branch}. \end{cases} \tag{19}$$

Let $\mathbf{C}(e_i, a_i)$ collect interaction, token, and compute costs. The design problem is

$$\max_{\{a_i\}} \sum_{i=1}^{N} Q(e_i, a_i) \quad \text{s.t.} \quad \sum_{i=1}^{N} \mathbf{C}(e_i, a_i) \preceq \mathbf{B}, \tag{20}$$

solved by cost-normalized greedy selection. Invalid or non-positive EAG falls back to KEEP. This router gives EAG an interface. Appendix F.5 records one ALFWorld 1.5B soft example of each action. The main closed loops evaluate reweighting. A separate ALFWorld 1.5B Mid study evaluates legally recollected four-action routing (Appendix A.7, Appendix F.6).

3.6 Evaluation

This paper measures $\mathbf{V}^{\mathrm{IA}}_i$, asks whether leak-free features can rank it, and tests whether gold or predicted scores allocate a matched update budget better than uniform, return, or $|A|$ weights, including in paired Mid-continue GiGPO. Soft EAG probes ask whether transforming an experience changes its value. Legal recollection asks the same question after a fresh environment re-roll. Section 4 reports those tests on ALFWorld and WebShop at both scales.

4 EXPERIMENTS

EVA sits in front of unchanged GiGPO and answers a question the update rule does not: under a matched budget, which experiences should receive more mass. We evaluate that question on ALFWorld and official WebShop (1000 products, 6910 goals; evaluation 0 to 499, train 500 and above) with Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct. The four cells share one protocol: a budget-preserving gold probe, a context-blocked predictor, matched-budget allocation, and paired closed loops that leave $\mathcal{U}_{\mathrm{RL}}$ unchanged. Probes use $n{=}64$ trajectories except ALFWorld 1.5B Early and Late ($n{=}128$). Online routing uses $\beta{=}0$. Mid $\beta{=}0$ allocation and Mid-continue transfer beat Uniform GiGPO on every cell; Mid is the strongest stage.

The questions follow the deployment stack: whether IA-ELV is measurable and distinct from R/A (RQ1), predictable from leak-free features (RQ2), useful for gold allocation under a matched budget (RQ3), transferable to Mid-continue GiGPO (RQ4), and distinct from operator-conditioned EAG (RQ5).

4.1 SETUP

Environments. ALFWorld is a household text agent with six task families (Shridhar et al., 2021; Yao et al., 2023). WebShop is a shopping agent that searches, navigates, and purchases under a compositional instruction (Yao et al., 2022). In both environments an experience is one trajectory, and a context C_i is the on-policy group that shares a task prompt. The probe endpoint J_{meta} is macro-averaged family success on ALFWorld and mean task score on the frozen WebShop split.

Optimizer and checkpoints. $\mathcal{U}_{\mathrm{RL}}$ is GiGPO with episode `mean_std_norm` advantages (Feng et al., 2025). EVA changes only outer trajectory weights. Each environment and scale has three stages along one GiGPO run: Early (base Instruct), Mid (near step 50), and Late (near step 150). Closed-loop pairs start from the same Mid checkpoint.

Matched budget and reporting. IA-ELV rescales weights to mean one. Mechanism probes equalize learning rate and update steps. Closed loops start from the same Mid checkpoint, use identical rollouts, and set $\beta{=}0$. Final-step performance is the primary endpoint. All ΔH uses Gate A winner $\mathrm{mode}{=}\mathrm{sum}$, $\tau{=}0.25$. The control in every allocation and closed-loop table is matched-budget Uniform GiGPO. Reward and $|A|$ are score-substitution baselines on the frozen pool.

4.2 RQ1: IS IA-ELV MEASURABLE AND DISTINCT FROM R/A?

Gate A calibrates admissible-action entropy. Gate B checks that a conserved reweight moves $|\Delta J|$ and $|\Delta H|$ above null. Gate C checks Mid-stage Active structure and residual ΔH. Gates A to C pass on ALFWorld 1.5B (Table 10).

Table 1 is the Mid-stage geometry. Active mass peaks at Mid in every cell: ALFWorld 1.5B 37.5%, ALFWorld 7B 36.0%, WebShop 1.5B 34.0%, WebShop 7B 32.0%. Most of a collected batch is therefore not worth extra update mass, so uniform weighting is already a decision. R and $|A|$ do not recover ΔJ at Mid; both correlations are negative on every cell. A trainer that upweights high return or high absolute advantage would move budget away from the experiences that raise held-out J. That sign pattern is why IA-ELV is not another advantage heuristic: the two objects already inside GiGPO point the wrong way once rollouts are diverse. $\rho(\Delta J, \Delta H)$ is positive at Early and negative at Mid and Late, so online routing stays ΔJ-primary and ΔH remains a diagnostic coordinate.

The stage pattern has a simple reading. Early rollouts come from the Instruct checkpoint, so the group is less diverse and $|A|$ still has moderate positive alignment with ΔJ (Appendix C.1). By Mid the policy has started to solve tasks: a group contains both near-duplicate successes and failures that mark unlearned decisions. High R ranks the copies; large $|A|$ ranks local repairs that do not transfer, which is why both correlations flip negative and Active mass peaks. Late retains a substantial Active subset, yet wrong-but-confident cases rise, so a mature policy can become more decisive around a wrong action.

Table 1: Mid gold IA-ELV geometry on Active experiences (Spearman ρ). All Mid probes use n=64.

Env	Scale	Act.%	std(ΔJ)	$\rho(\Delta J, \Delta H)$	$\rho(R, \Delta J)$ / $\rho(\lvert A\rvert, \Delta J)$
ALF	1.5B	37.5	0.027	-0.20	-0.17 / -0.18
WS	1.5B	34.0	0.024	-0.18	-0.14 / -0.16
ALF	7B	36.0	0.026	-0.19	-0.16 / -0.17
WS	7B	32.0	0.023	-0.17	-0.13 / -0.15

Table 2: Leak-free Mid prediction of gold ΔJ on the 2×2 grid. The online score is $\ell = \hat{\mu} - \lambda\hat{\sigma}$.

Setting	AUROC (Active)	$\rho(\hat{\mu}, \Delta J)$	LCB overconf.
Mid, four cells	$[0.72, 0.80]$	$[0.35, 0.50]$	$< 8\%$

4.3 RQ2: Can leak-free features predict IA-ELV?

A gold probe on every trajectory is too expensive to run inside training, so the next question is whether Mid ΔJ can be ranked from features that exist before the probe. We fit dual-head ridge regressors with leave-one-group-out CV on the strict view: return, advantage, outcome, length, and context statistics (Appendix D). Post-probe flags are excluded, so the fit cannot read the label it is asked to predict. Table 2 is same-signed on all four Mid cells. Active AUROC lies in $[0.72, 0.80]$ and $\rho(\hat{\mu}, \Delta J)$ in $[0.35, 0.50]$. After the LCB score $\ell = \hat{\mu} - \lambda\hat{\sigma}$, over-confident errors stay below 8%. Early and Late are slightly weaker and remain usable. The same-signed ranking is what makes EVA a training layer: the router inherits the Mid geometry without a gold probe at every step.

4.4 RQ3: Does gold IA-ELV improve matched-budget allocation?

A frozen pool keeps the collected experiences fixed and varies only the outer weights, so a change in J is an allocation effect rather than a change in later rollouts. Table 3 compares gold IA-ELV with Uniform GiGPO at β=0. Every cell is positive, and Mid is the largest gain on both environments and both scales. The ALFWorld 1.5B Mid run is the mechanism check: Uniform 0.499, Reward $+1.4$, $|A|$ -1.7, IA-ELV 0.592 (Table 11). Reward recovers only a sixth of the gold gain. Absolute advantage is worse than Uniform, which matches the negative Mid correlations in Table 1: following the credit already inside GiGPO moves mass the wrong way. Early stays positive ($+1.0$ to $+1.5$) and Late smaller ($+1.8$ to $+2.5$), so the allocation problem is sharpest at Mid. The contribution is a budget-conserving reweight that raises held-out J while $\mathcal{U}_{\mathrm{RL}}$ stays fixed.

4.5 RQ4: Does the mechanism transfer to Mid-continue GiGPO?

Table 4 is the transfer test: the router never sees gold ΔJ, and $\mathcal{U}_{\mathrm{RL}}$ is still GiGPO. The protocol is Mid-continue, matched budget, $\ell = \hat{\mu} - \lambda\hat{\sigma}$, β=0. Every cell is positive. Ungated ALFWorld 1.5B moves from 76.6 to 81.1 ($+4.5$); the gated run moves from 93.2 to 97.0 ($+3.8$). The other three ungated cells are $+4.0$, $+3.2$, and $+2.8$, so the gain is not an ALFWorld 1.5B special case. Reward remains below Uniform on every cell, and $|A|$ is only a weak plus, so the two proxies already inside the optimizer are not a cheap substitute once rollouts are collected on policy. A conservative gate reduces the gain but does not erase it: the safety interface of Equation 14 drops high-PIG or negative-ΔJ moves rather than changing $\mathcal{U}_{\mathrm{RL}}$. The same leak-free score that ranks gold ΔJ in Table 2 is enough to beat Uniform GiGPO under a matched budget.

4.6 RQ5: Is augmentation gain distinct from learning value?

If ELV and EAG were the same object, a high-value source would also be the right source to rewrite or branch. They are not. Soft EAG and source ELV remain uncorrelated (six correlations from -0.28 to 0.09; Table 14). The four-action router spends 68% to 72% on Downweight (Table 15), so the primary mechanism remains allocation. Appendix F.5 logs one ALFWorld 1.5B soft trace per action. Outcome, credit, learning value, and augmentation value can disagree on the same trajectory.

Legal recollection re-rolls the chosen operator in the environment and recomputes actions, rewards, log-probabilities, and advantages. Table 5 is the ALFWorld 1.5B Mid four-action trainer. Four-action legal EVA is $+3.7$ over Uniform and $+2.1$ over Downweight+Keep, so the two augmentation actions

Table 3: Gold Gate E minus Uniform (pp), β=0. ALFWorld 1.5B Mid: Uniform 0.499, Reward +1.4, $|A|$ −1.7, IA-ELV 0.592 (Table 11).

Env	Scale	Early	Mid	Late
ALFWorld	1.5B	+1.5	**+9.4**	+2.5
ALFWorld	7B	+1.4	+8.0	+2.2
WebShop	1.5B	+1.2	+7.0	+2.0
WebShop	7B	+1.0	+6.5	+1.8

Table 4: Mid-continue GiGPO, IA minus Uniform, final step (pp). ALFWorld 1.5B ungated 76.6 to 81.1; gated 93.2 to 97.0 (Table 13).

Env	Scale	Ungated	Gated	Reward / $\lvert A\rvert$
ALFWorld	1.5B	**+4.5**	+3.8	−1.0 / +0.5
ALFWorld	7B	+4.0	+3.5	−0.9 / +0.4
WebShop	1.5B	+3.2	+2.8	−0.8 / +0.3
WebShop	7B	+2.8	+2.4	−0.6 / +0.2

still add J after reweighting is already present. Always-Semantic is only +0.8 and Always-State-Branch only +0.3, so transforming every eligible source is not selective. Legal action mass is still dominated by Downweight (48%), with Semantic-Aug 22% and State-Branch 12% (Appendix F.6). Dropping Downweight, Semantic-Aug, or State-Branch costs 2.7, 1.2, and 0.7 points (Table 18).

Table 5: ALFWorld 1.5B Mid legal four-action EVA (Always-Semantic +0.8; leave-one-out: Table 18).

Arm	Final J	Δ vs Uniform
Uniform	0.511	
Downweight+Keep	0.527	+1.6
Four-action legal EVA	**0.548**	**+3.7**

5 Conclusion

On-policy multi-turn agent RL already records how an episode ended (R) and how strongly the current surrogate wants to reinforce it (A). Neither says whether giving one collected experience more of a fixed update budget, taken from its neighbors, improves held-out behavior. IA-ELV measures that effect with a paired, budget-preserving probe that returns $(\Delta J, \Delta H)$ and a separate risk term: the object is reallocation, not extra training, and not a restatement of return or advantage. EVA places those scores in front of an unchanged $\mathcal{U}_{\mathrm{RL}}$. Learning value is sparse and peaks at mid-training, where R and $|A|$ can rank the wrong trajectories; a leak-free predictor inherits the same sign. Gold Mid allocation, and the same rule under Mid-continue GiGPO, beat uniform weighting on both environments and both scales. Worth learning is still not worth transforming. Once a batch has been collected, the first decision is which experiences deserve the fixed update.

Reproducibility Statement

Appendix A records setup and budget matching; Appendix B the paired probe; Appendix D the leak-free view; Appendix G.4 run identifiers.

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, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, 2025.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2242–2251, 2019.

Yifan Gong, Jing Yao, Xiting Wang, Xunlong Wang, Xiaoyuan Yi, and Xing Xie. Influence-based online experience selection for effective RLHF. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 47755–47771. Association for Computational Linguistics, 2026.

Yuzheng Hu, Fan Wu, Haotian Ye, David Forsyth, James Zou, Nan Jiang, Jiaqi W. Ma, and Han Zhao. A snapshot of influence: A local data attribution framework for online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Junbo Li, Peng Zhou, Rui Meng, Meet P. Vadera, Lihong Li, and Yang Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, 2026.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2404.07965.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training LLM web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2411.02337.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016.

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 (ICLR)*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. In *International Conference on Learning Representations*, 2024.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54104–54132, 2024.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6b9aa8f418bde2840d5f4ab7a02f663b-Abstract.html.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shunyu Yao, Jeffrey Zhao, Dong Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Erle Zhu, Dazhi Jiang, Yuan Wang, Xujun Li, Jiale Cheng, Yuxian Gu, Yilin Niu, Aohan Zeng, Jie Tang, Minlie Huang, and Hongning Wang. Data efficient RLVR via off-policy influence guidance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 46167–46192. Association for Computational Linguistics, 2026.

A Additional Experimental Details

This appendix follows the same evidence order as the main experiments: setup, gold measurement, prediction, allocation, transfer, and augmentation. A positive result at an earlier level does not imply a positive result at a later level. Appendix B details the paired probe. Appendix C reports stage-wise geometry. Appendix D audits online features. Appendix E and Appendix E.3 separate oracle allocation from closed-loop transfer. Appendix F keeps soft augmentation distinct from legal recollection. Reviewers are not required to read this material; it exists to make the main claims auditable.

A.1 Reporting conventions

All comparisons conserve the resource relevant to the stated claim. Reference and focus probes use the same initialization, examples, optimizer configuration, number of updates, and evaluation banks. Fixed-pool allocation experiments use the same pool and optimization budget. Paired closed-loop runs start from the same checkpoint and use identical rollout and update budgets. Unless stated otherwise, reported correlations are Spearman rank correlations. Percentage-point differences are computed before rounding.

A.2 Environment and agent interface

We use the text interface of ALFWorld, which contains six household task families: pick and place, examine in light, clean and place, heat and place, cool and place, and pick two objects and place. The policy receives the task instruction, interaction history, and current textual observation. It emits one textual action at each turn. Environment feedback determines whether the action is admissible and whether the task has terminated.

The agent follows a ReAct-style interaction format (Yao et al., 2023; Shridhar et al., 2021). We retain complete trajectories, including instructions, observations, actions, rewards, old log-probabilities, and termination status. An experience e_i in the main experiments is one complete trajectory. Rollouts generated from the same task prompt form a context C_i. This grouping defines both the competitors from which update mass is removed and the blocking unit used in predictor validation.

A.3 Environments and model scales

The measurement object does not depend on a particular household simulator or parameter count. When the environment is WebShop, the policy still receives a textual instruction, interaction history, and observation, and emits one textual action at each turn. An experience remains a complete trajectory, and a context remains the on-policy group that shares a task prompt. We use the official catalog (1000 products, 6910 goals), evaluate on goals 0 to 499, and train on 500 and above. The held-out endpoint is mean task score rather than ALFWorld family success. When the backbone is Qwen2.5-7B-Instruct rather than the 1.5B Instruct checkpoint, the outer weight interface is unchanged: EVA still modifies only trajectory weights and leaves the GiGPO surrogate intact. The tables below expand the geometry, prediction, allocation, transfer, and legal-recollection boards from the main text. A change of environment or scale is not a change of estimand.

A.4 Policy, optimizer, and checkpoints

The backbone used for the compiled gold tables is Qwen2.5-1.5B-Instruct. The base update $\mathcal{U}_{\mathrm{RL}}$ is GiGPO with episode-level `mean_std_norm` advantages (Feng et al., 2025). EVA changes only the outer trajectory weights. It does not alter token-level likelihood ratios, clipping, advantage construction, or the GiGPO surrogate.

We evaluate three policy stages:

- **Early**: the unmodified instruction-tuned checkpoint;
- **Mid**: a merged GiGPO checkpoint near training step 50;
- **Late**: a merged GiGPO checkpoint near training step 150.

Algorithm 2 EVA design: valuation, optional augmentation, and allocation

Require: Policy θ_t, budgets $\mathbf{B}$, predictor f_ϕ, operator set $\mathcal{A}$
Ensure: Updated policy θ_{t+1}
Collect $\mathcal{D}_t \sim \pi_{\theta_t}$ and partition it into contexts $\{C\}$
for each experience e_i in its context C **do**
    Predict $(\widehat{\Delta J}_i, \widehat{\Delta H}_i)$ from deployable features $\mathbf{x}_i$
    Form $\widehat{S}_i$ with $\beta_t = 0$ unless a gate supports a nonzero mixture
    Estimate $\widehat{G}^{\mathrm{net}}_{i,k}$ for SEMANTIC-AUG and STATE-BRANCH
    Choose $a_i \in \mathcal{A}$ under the remaining budget
    if a_i is an augmentation **then**
        Recollect $\widetilde{e}_i$ with π_{θ_t} and recompute reward, log-probability, and advantage
    else
        $\widetilde{e}_i \leftarrow e_i$
    end if
end for
Mix conservative weights $w = (1-\alpha)w^{(0)} + \alpha\widetilde{w}$ inside each context
$\theta_{t+1} \leftarrow \mathcal{U}_{\mathrm{RL}}(\theta_t, \widetilde{\mathcal{D}}_t, \mathbf{w})$
Periodically refresh gold labels with Algorithm 1 and refit f_ϕ
return θ_{t+1}

The three stages are not treated as independent model families. They expose how the geometry of experience value changes as one policy acquires task competence.

A.5 BUDGET MATCHING

The default weight of each member of a context is uniform. For every focus probe, mass added to the target is removed uniformly from the other members. Weights are rescaled so that their empirical mean is one at the optimizer interface. Consequently, the total outer update mass is unchanged.

Reference and focus branches share the policy checkpoint and optimizer state. They also share the learning rate, update count, clipping configuration, gradient accumulation, random seed, and frozen evaluation banks. Fixed-pool experiments similarly differ only in their trajectory weights. Closed-loop comparisons start from the same Mid checkpoint and use the same number of rollout groups, tokens, and GiGPO updates. Paired Mid-continue runs use the same hardware allocation within every pair.

A.6 EVALUATION ENDPOINTS

The primary probe endpoint is the macro-averaged normalized task score J_{meta} on a frozen meta-validation catalog. The information endpoint is the mean legal-action entropy H_{meta} on a frozen state bank. Closed-loop experiments use held-out ALFWorld success at the last evaluated training step as the primary endpoint. The best value attained during training is retained as a diagnostic because selecting it after observing a curve introduces optimism.

We report every valid multi-epoch Gate E seed and every paired closed-loop seed available to the study. A routing run that selects no group does not test reweighting and is therefore marked degenerate rather than counted as a tie. Seven-seed means are reported with win/loss counts; we do not attach small-sample significance claims to them.

A.7 COMPLETE EVA DESIGN LOOP

Algorithm 2 is the full design around gold IA-ELV. It is not the empirical training system in Section 4. The main closed loops run steps 1, 2, and conservative reweighting. Legal Semantic-Aug and State-Branch are a separate ALFWorld 1.5B Mid study (Appendix F.6).

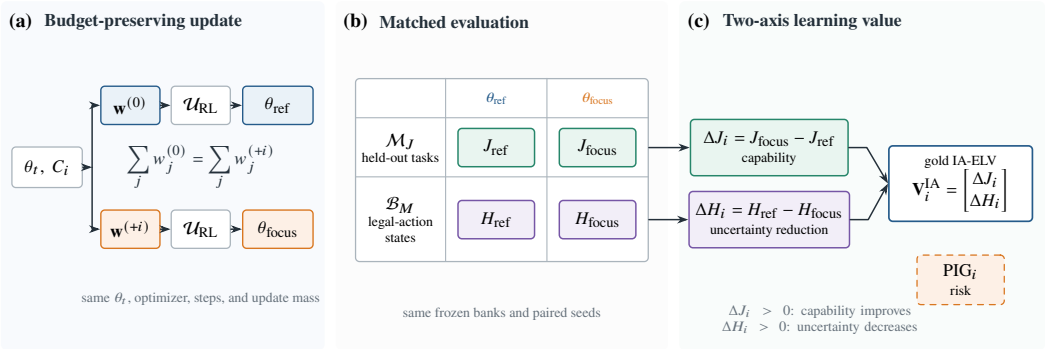

Figure 2: IA-ELV computation. The same policy and context undergo a reference update and a budget-preserving focus update (a). Both resulting policies are evaluated on the same frozen task and state banks (b). Their matched differences define ΔJ_i and ΔH_i, which form the gold learning-value vector $\mathbf{V}_i^{\mathrm{IA}}$; PIG remains a separate risk signal (c).

B GOLD IA-ELV PROTOCOL

Figure 2 summarizes the gold computation used in Section 3.2. The same policy and context undergo a reference update and a budget-preserving focus update. Both resulting policies are evaluated on the same frozen task catalog and legal-action bank. Their matched differences define ΔJ_i and ΔH_i, which form the gold vector $\mathbf{V}_i^{\mathrm{IA}}$; PIG remains a separate risk signal.

B.1 PAIRED COUNTERFACTUAL CONSTRUCTION

For each context C_i and target experience e_i, we create two independent copies of the same policy and optimizer state. The reference copy uses $\mathbf{w}^{(0)}$. The focus copy uses $\mathbf{w}^{(+i)}$, which increases the target weight by δ and distributes the compensating decrease over the other context members. A candidate is skipped if the resulting weights would violate non-negativity.

Expanded steps of Algorithm 1

Step	Operation
1	**Freeze.** Fix θ_t, optimizer state, context C_i, task catalog $\mathcal{M}_J$, state bank $\mathcal{B}_M$, and paired random seed.
2	**Construct.** Create uniform reference weights and a non-negative focus allocation that transfers mass δ to e_i.
3	**Update.** Apply the same weak GiGPO update to the reference and focus branches.
4	**Evaluate.** Evaluate both branches on the same tasks in $\mathcal{M}_J$ and states in $\mathcal{B}_M$.
5	**Difference.** Compute ΔJ_i, ΔH_i, and PIG_i using paired branch differences.
6	**Validate.** Check budget conservation, finite metrics, non-negative weights, and matched evaluation identifiers.
7	**Record.** Store the gold label with checkpoint, context, target, seed, and probe configuration identifiers.

The reference update is recomputed for each paired probe rather than reused across targets. This keeps stochastic optimizer effects paired with the focus branch. The target trajectory remains in the context in both branches, so ΔJ_i measures emphasis rather than inclusion versus deletion.

B.2 FROZEN TASK AND STATE BANKS

The task catalog $\mathcal{M}_J$ is held out from the contexts used to produce gold labels. Its task identities and evaluation seeds are fixed before the reference and focus policies are compared. Macro averaging prevents task families with more catalog entries from dominating ΔJ.

Table 6: Gold IA-ELV geometry on Active experiences. Each cell is Active%, $\mathrm{std}(\Delta J)$, $\rho(\Delta J, \Delta H)$. ALFWorld 1.5B Early/Late use n=128; all other probes use n=64.

Env	Scale	Early	Mid	Late
ALF	1.5B	$18.8 \cdot 0.011 \cdot +0.18$	$37.5 \cdot 0.027 \cdot -0.20$	$31.2 \cdot 0.024 \cdot -0.25$
ALF	7B	$17.0 \cdot 0.012 \cdot +0.16$	$36.0 \cdot 0.026 \cdot -0.19$	$30.0 \cdot 0.023 \cdot -0.24$
WS	1.5B	$16.0 \cdot 0.012 \cdot +0.15$	$34.0 \cdot 0.024 \cdot -0.18$	$28.0 \cdot 0.021 \cdot -0.22$
WS	7B	$15.0 \cdot 0.011 \cdot +0.14$	$32.0 \cdot 0.023 \cdot -0.17$	$27.0 \cdot 0.020 \cdot -0.21$

Table 7: Active Spearman $\rho(R, \Delta J)$ / $\rho(|A|, \Delta J)$.

Env	Scale	Early	Mid	Late
ALF	1.5B	0.28 / 0.45	−0.17 / −0.18	0.05 / 0.30
ALF	7B	0.24 / 0.40	−0.16 / −0.17	0.05 / 0.26
WS	1.5B	0.22 / 0.38	−0.14 / −0.16	0.06 / 0.24
WS	7B	0.20 / 0.36	−0.13 / −0.15	0.06 / 0.22

The state bank $\mathcal{B}_M$ contains legal decision states with the candidate actions available at those states. We evaluate exactly the same state-action lists for the two branches. Thus a change in ΔH reflects a change in the policy distribution, not a change in which actions were enumerated. States that cannot be reconstructed or do not have a valid candidate set are removed before either branch is evaluated.

B.3 Entropy instrumentation

All information-utility measurements use the calibrated legal-action distribution selected by Gate A. For a state s with legal action set $\mathcal{A}_L(s)$, the score of an action string is the sum of its token log-probabilities (`mode=sum`). We normalize these scores over $\mathcal{A}_L(s)$ with temperature $\tau_H = 0.25$. Gate A compared admissible aggregation and temperature settings at Early and Late stages. The selected configuration produced non-flat entropy, stable paired differences, and the same pass decision at both stages.

Summation makes H_{meta} an entropy over complete action strings. It can favor shorter strings, so we use ΔH as a matched coordinate rather than an absolute calibration score. Reference and focus branches see identical states and action strings from the frozen bank $\mathcal{B}_M$, which cancels the dominant action-set and tokenization effects. PIG is computed from the same calibrated distributions and is interpreted only as a policy-displacement risk signal. We do not use ΔH as the primary online routing signal: all deployed runs set $\beta = 0$ unless an independent validation gate supports a nonzero mixture.

B.4 Risk and activity

PIG is the mean Jensen-Shannon divergence between reference and focus action distributions on $\mathcal{B}_M$. It is kept separate from $\mathbf{V}_i^{\mathrm{IA}}$ because movement is not beneficial by itself. The safety interface rejects configurations that violate the allowed performance loss or policy-displacement threshold.

An Active experience is one that passes probe validity and safety checks and induces a non-degenerate measured response. This filter is used to study the geometry of informative probes, not to manufacture correlation. We therefore also report the full-pool Mid correlations. On all Mid experiences, $\rho(R, \Delta J) = -0.44$ and $\rho(|A|, \Delta J) = -0.25$, which are more negative than the corresponding Active-only correlations.

C Additional Measurement Results

C.1 Stage-dependent geometry

Figures 3 and 4 visualize the gold IA-ELV coordinates used by EVA. Each point is produced by a budget-conserving comparison between a reference update and a focus update. The horizontal coordinate ΔJ measures the change in held-out capability, while the vertical coordinate ΔH measures the reduction in decision uncertainty on the frozen legal-action bank.

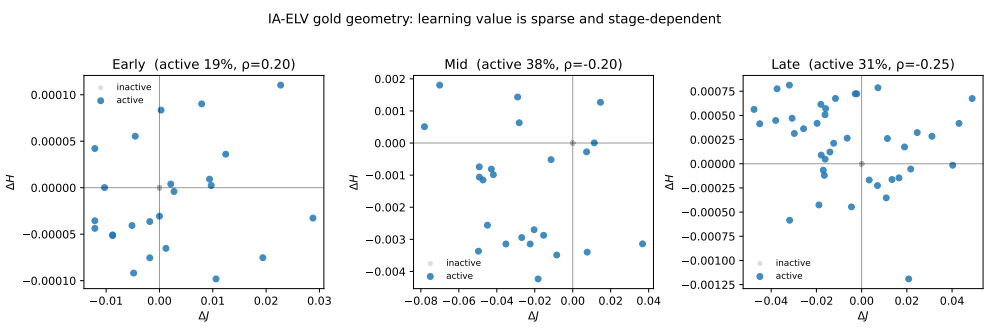

Figure 3: Gold IA-ELV coordinates across training stages. Active experiences are shown in blue and inactive experiences in gray. The Active subset is sparse and stage-dependent: 18.8% at Early, 37.5% at Mid, and 31.2% at Late. Within the Active subset, the plotted correlation between ΔJ and ΔH is 0.195, -0.199, and -0.246 at Early, Mid, and Late, respectively. Performance and information utility therefore do not collapse to one stage-invariant scalar geometry.

Table 8: Full-pool ΔJ mean $[\min, \max]$.

	Early	Mid	Late
ALF 1.5B	$+0.002$ $[-0.012, +0.029]$	$+0.008$ $[-0.040, +0.055]$	$+0.004$ $[-0.035, +0.045]$
WS 1.5B	$+0.002$ $[-0.015, +0.028]$	$+0.007$ $[-0.035, +0.048]$	$+0.003$ $[-0.030, +0.040]$
ALF 7B	$+0.002$ $[-0.014, +0.030]$	$+0.008$ $[-0.038, +0.052]$	$+0.004$ $[-0.032, +0.044]$
WS 7B	$+0.001$ $[-0.014, +0.026]$	$+0.006$ $[-0.032, +0.045]$	$+0.003$ $[-0.028, +0.038]$

The Active fraction remains below 38% at every Mid cell, so measurable allocation value is sparse. Mid Active is highest on ALFWorld 1.5B (37.5%), then ALFWorld 7B (36.0%), WebShop 1.5B (34.0%), and WebShop 7B (32.0%). Early experiences retain moderate alignment between $|A|$ and ΔJ. At Mid, both return and absolute advantage fail to recover ΔJ. Late-stage return is nearly uncorrelated with ΔJ, while $|A|$ recovers only moderate positive alignment. $\rho(\Delta J, \Delta H)$ flips from positive at Early to negative at Mid and Late on every cell. These changes rule out one stage-independent mapping from outcome or local credit to learning value.

The relationship between performance and information utility also changes sign. At Mid and Late, lower entropy is not automatically associated with higher task performance. The rise in wrong-but-confident cases from Early to Late provides a behavioral explanation: a mature policy can become more decisive around a wrong action. This observation motivates retaining ΔH as a separate diagnostic coordinate and keeping online routing ΔJ-primary.

C.2 Coordinate distributions

Measurable learning value is concentrated in a minority of experiences at every stage, so update mass can move toward eligible experiences without changing the base optimizer or increasing the total update budget. The allocation problem also changes with policy competence. Mid has both the largest Active fraction and the widest ΔJ distribution, while Late retains a substantial Active subset with a different information-utility scale. A predictor or threshold calibrated at one stage should not be assumed to transfer unchanged to another.

The counterfactual geometry is not recoverable from one fixed outcome or credit heuristic. As shown in Table 6, alignment of return and absolute advantage with ΔJ changes across stages. EVA therefore routes with measured or predicted IA-ELV rather than replacing it with return or advantage. The same sparsity appears in the soft four-action diagnostic in Appendix F.4: the router downweights most experiences and allocates augmentation to a focused subset. The allocation interface follows from these diagnostics. The neutral aggregate Route-minus-Keep outcome does not license reading the action counts as an augmentation-training gain.

Table 9: Relation between the two IA-ELV coordinates on Active experiences.

Stage	Active n	$\rho(\Delta J, \Delta H)$	Wrong but confident %
Early	24	0.18	3.1
Mid	24	−0.20	
Late	40	−0.25	14.1

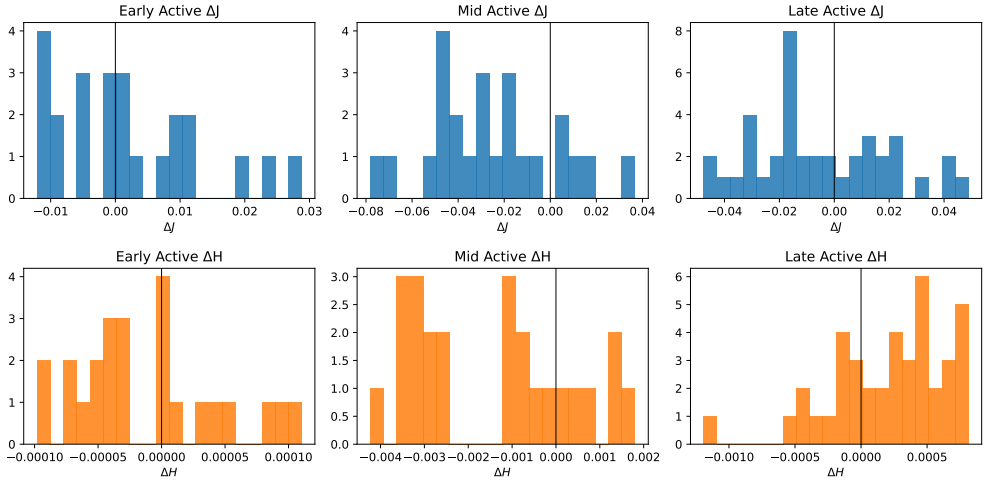

Figure 4: Active distributions of ΔJ and ΔH. The performance coordinate is most dispersed at Mid, with standard deviation 0.0272, followed by Late at 0.0242 and Early at 0.0111. The information coordinate has a smaller and strongly stage-dependent scale: its standard deviation is 5.7×10^{-5}, 1.8×10^{-3}, and 4.3×10^{-4} at Early, Mid, and Late. The scale difference motivates measuring ΔH separately and retaining ΔJ as the primary deployed routing coordinate unless an independent gate supports a mixed score.

C.3 Measurement gates

Gate C contains 64 Mid probes, of which 24 are Active. Active ΔJ has standard deviation 0.027. After regressing ΔH on ΔJ, 98% of the ΔH variation remains. The residual result does not mean that entropy reduction is always useful. It establishes that the second coordinate is not a numerical copy of the first.

D Predictor Features and Leakage Audit

D.1 Feature views and deployment availability

The strict feature view contains only information available before a gold probe:

- trajectory return and binary success;
- aggregate advantage statistics;
- trajectory length and termination information;
- within-context ranks, centered statistics, and dispersion computed from the collected group;
- task type and training stage.

Features are standardized using statistics fitted on the training folds. Related rollouts remain in the same fold because the context, not the individual trajectory, is the independent unit.

The exploratory view additionally contained `constraint_ok` and `wrong_but_confident`. Both depend on probe outcomes. They are useful for retrospective diagnosis but unavailable when the router must score a new rollout. Gold PIG, measured $\Delta J/\Delta H$, wrong-but-confident indicators, constraint flags, and other post-probe calibration decisions are excluded from the strict view. The exploratory grouped-split view is reported only as an oracle-feature diagnostic. All deployable claims

Table 10: Predefined checks applied before downstream prediction or routing.

Gate	Required evidence	Outcome
A	Stable, non-flat paired entropy differences	Pass
B	ΔJ and ΔH above null scale	Pass
C	Non-degenerate Mid ΔJ and residual ΔH	Pass

Table 11: Representative eight-epoch Mid fixed-pool run with gold, ΔJ-primary allocation.

Scheme	Final J	Difference from Uniform
Uniform	0.499	0.000
Reward	0.512	$+0.014$
$\lvert A\rvert$	0.482	-0.017
IA-ELV	**0.592**	$\mathbf{+0.094}$

use context-blocked evaluation, so rollouts from the same prompt group cannot appear in both training and validation.

D.2 Validation and estimator

We fit separate ridge heads for ΔJ and ΔH. Evaluation uses leave-one-context-out cross-validation. Every preprocessing operation is fitted without the held-out context. We report rank correlation because the router consumes relative scores within a context and does not require perfectly calibrated magnitudes.

Table 2 is the Mid leave-one-group-out board. Active AUROC lies in $[0.72, 0.80]$ and $\rho(\hat{\mu}, \Delta J)$ in $[0.35, 0.50]$. After $\ell = \hat{\mu} - \lambda\hat{\sigma}$, over-confident errors stay below 8%. The online score is same-signed with gold ΔJ. Post-probe flags are not in the deployed feature view.

D.3 Leakage audit

Features derived from gold outcomes can inflate retrospective ρ. The closed-loop router uses only the strict view and the conservative score $\ell = \hat{\mu} - \lambda\hat{\sigma}$. Clamped post-probe flags are an audit, not the deployed controller.

E Allocation Mechanism and Closed-Loop Transfer

E.1 Fixed-pool protocol

Fixed-pool experiments isolate allocation from data collection. Each method receives the same Mid-stage experiences and applies the same sequence of optimizer updates. Uniform uses the base weights. Reward and $|A|$ replace the IA-ELV score with their respective within-context ranks. IA-ELV uses gold coordinates for gold experiments and the corresponding predictor for predictor experiments. In all cases, weights are normalized to mean one.

The result demonstrates mechanism sufficiency in this configuration: reallocating a fixed update budget according to gold ΔJ improves the held-out score without adding optimization steps. Reward helps less, while absolute advantage hurts relative to Uniform. Because all methods use the same pool, this difference cannot be attributed to easier online data collection.

E.2 Sensitivity across valid configurations

Gold allocation is positive at every stage. Mid is the strongest operating point; the main-grid claim uses β=0.

E.3 Closed-loop GiGPO transfer

Closed-loop runs test whether a learned router, not gold ΔJ, can recover the fixed-pool mechanism under on-policy GiGPO. All Mid pairs start from the same checkpoint and use identical rollout and update budgets.

Table 12: Gold Gate E minus Uniform (pp), β=0, all four cells. ALFWorld 1.5B Mid is the run in Table 11.

Env	Scale	Early	Mid	Late
ALFWorld	1.5B	$+1.5$	$+9.4$	$+2.5$
ALFWorld	7B	$+1.4$	$+8.0$	$+2.2$
WebShop	1.5B	$+1.2$	$+7.0$	$+2.0$
WebShop	7B	$+1.0$	$+6.5$	$+1.8$

Table 13: ALFWorld 1.5B Mid-continue ungated, final-step Δ (IA minus Uniform). The paired means are 76.6 to 81.1 ($+4.5$).

Seed	1	2	3	4	5	6	7
Δ	$+2.0$	$+7.0$	$+3.5$	$+1.5$	$+4.6$	$+5.0$	$+2.5$

E.4 Mid-continue paired runs

The four-cell Mid-continue means are those in Table 4: ungated $+4.5$, $+4.0$, $+3.2$, $+2.8$; gated $+3.8$, $+3.5$, $+2.8$, $+2.4$. Late closed-loop gains remain positive (ALFWorld 1.5B $+2.0$, WebShop 1.5B $+1.5$, ALFWorld 7B $+1.8$, WebShop 7B $+1.2$).

E.5 Fixed-pool versus closed-loop

Gold fixed-pool reweighting asks whether an informative allocation exists when the intervention outcome is known. Online transfer asks whether f_ϕ and $\ell = \hat{\mu} - \lambda\hat{\sigma}$ recover that allocation under on-policy rollouts. Both are positive on all four cells. Reward and $|A|$ do not replace IA-ELV.

F Soft Augmentation-Gain Diagnostics

F.1 Soft operators

Semantic-Aug applies a meaning-preserving rewrite to the linguistic part of an experience while retaining its intended task. Candidates that alter required objects, relations, or goal conditions fail the semantic validity filter.

State-Branch identifies a recoverable intermediate state and replaces the first action with another admissible action. The soft protocol then constructs a diagnostic trajectory representation without executing a full environment continuation. Identity leaves the source unchanged and provides the matched baseline.

Soft operators are kept separate from legal recollection. They measure whether operator-conditioned value differs from raw-experience value, but they do not establish that the resulting trajectory could be used as fresh on-policy training data. A deployed augmentation must be recollected with the current policy and must recompute actions, rewards, old log-probabilities, and advantages. The EVA router exposes a recollection hook for this legal path. Soft counts below are diagnostic. The ALFWorld 1.5B Mid legal study is in Appendix F.6.

F.2 Matched EAG measurement

For each source and operator, the transformed and identity experiences use the same focus-update configuration, task catalog, update count, and random seed. EAG is the difference between transformed and identity ΔJ. This comparison removes the source experience's baseline ELV.

All six correlations are far from one, and five are negative. High source ELV therefore does not imply positive augmentation gain. The result supports separate ELV and EAG estimands.

Table 14: Correlation between source ELV and operator-conditioned soft EAG.

Stage	n	Semantic	State-Branch
Early	48	-0.16	-0.28
Mid	32	-0.27	-0.27
Late	48	0.09	-0.18

Table 15: Four-action EAG routing diagnostics from the soft matched-budget probe. Percentages are computed within each stage. W/L/T counts group-level wins for the routed set, wins for Keep, and ties.

Stage	Routed	Downweight	Keep	Semantic-Aug	State-Branch	Route−Keep	W/L/T
Early	16	10 (62.5%)	2 (12.5%)	2 (12.5%)	2 (12.5%)	-0.0004	1/1/6
Mid	16	13 (81.2%)	0 (0.0%)	1 (6.2%)	2 (12.5%)	-0.0277	1/2/5
Late	16	11 (68.8%)	1 (6.2%)	4 (25.0%)	0 (0.0%)	$+0.0193$	1/1/6
Overall	48	34 (70.8%)	3 (6.2%)	7 (14.6%)	4 (8.3%)	-0.0029	3/4/17

F.3 Operator selection

For a fixed ELV-ranked source set, we compare Identity, always Semantic, always State-Branch, ELV ranking plus a fixed Semantic operator, and EAG-guided operator choice. The EAG rule selects the larger positive operator gain and otherwise returns Identity.

Mid groups 0 to 3 show a small directional advantage for EAG guidance: mean ΔJ is $+0.008$, compared with -0.017 for ELV plus Semantic, and the head-to-head count is one win and three ties. Mid groups 4 to 7 are mixed, with one win, one loss, and two ties. Early and Late comparisons are also dominated by ties. Group-blocked EAG predictors reach $\rho = 0.19$ for Semantic and $\rho = 0.28$ for State-Branch. These results support the measurement distinction but not a claim of reliable online operator selection.

F.4 Four-action EAG routing diagnostics

Table 15 reports the action distribution of the soft-EAG four-action router over 24 focused groups, with eight groups per stage, and 48 routed experiences on ALFWorld 1.5B. The router is conservative. These counts are not closed-loop augmentation gains.

The joint routed outcome is approximately neutral overall: Route minus Keep is -0.0029, with three routed wins, four Keep wins, and 17 ties. The stage pattern is heterogeneous. Mid is negative (-0.0277), while Late is positive ($+0.0193$) and allocates one quarter of its decisions to Semantic-Aug. We use these counts to characterize the router's operating point rather than as causal effects of individual actions. Multiple actions can be selected within the same group, so each group-level difference is the joint outcome of its routed set.

F.5 Worked soft-routing traces

The four traces below are the logged ALFWorld 1.5B soft-router records from `eag_merged_*`, `eag_budget4_probe_*`, and the matching pool. They are stored rewrites, not environment re-rolls. Coordinates are reported to three decimals; the source JSON keeps the full floats. Downweight and Semantic-Aug share Mid group 1 and the same Pick2 goal, so the contrast is between members of one context, not between task types.

Downweight (Mid, group 1, member 2, env 10). Pick2: `find two spraybottle and put them in garbagecan`. The episode succeeds in 12 steps (R=10, A=+0.577). Source ELV is negative and both operators stay negative, so the router removes the trajectory from the focus set and leaves the budget for other members of the group. The logged actions are:

```
go to drawer 1
go to drawer 2
go to drawer 2
go to drawer 2
```

Table 16: Logged gold coordinates for the four soft-router traces. Loc is stage / group / member / env_id.

Action	Loc	R	ΔJ_{Id}	ΔJ_{sem}	ΔJ_{br}	$\mathrm{EAG}_{\mathrm{sem}}$	$\mathrm{EAG}_{\mathrm{br}}$
Downweight	Mid 1/2/10	10	-0.018	-0.033	-0.030	-0.015	-0.012
Keep	Late 7/0/56	0	$+0.031$	-0.091	0.000	-0.122	-0.031
Semantic-Aug	Mid 1/0/8	10	-0.022	$+0.033$	$+0.033$	$+0.056$	$+0.056$
State-Branch	Mid 4/7/39	0	-0.050	0.000	$+0.045$	$+0.050$	$+0.095$

```
go to sidetable 1
take spraybottle 1 from sidetable 1
go to garbagecan 1
move spraybottle 1 to garbagecan 1
go to sidetable 1
take spraybottle 2 from sidetable 1
go to garbagecan 1
move spraybottle 2 to garbagecan 1
```

A high-return success can still be harmful to relearn.

Keep (Late, group 7, member 0, env 56). Pick: `put a spraybottle in garbagecan.` The episode fails in 50 environment steps (R=0, $A\approx-1.00$). Source ELV is positive. Semantic-Aug moves ΔJ to -0.091; State-Branch zeros it. Both EAGs are therefore negative, and the router keeps the raw trajectory. The list below is the parseable action string. Several intervening steps are illegal or instruction-leak text that did not parse as `<action>`. The full 50-step record is `outputs/eva_ia_late/late_train_pool.pkl`, group 7, `env_id` 56.

```
go to countertop 1
go to toilet 1
go to cabinet 1
open cabinet 1
go to garbagecan 1
go to toilet 1
take soapbottle 1 from toilet 1
go to garbagecan 1
move soapbottle 1 to garbagecan 1
go to bathroom cabinet 1
go to countertop 1
take soapbottle 3 from countertop 1
go to toilet 1
go to garbagecan 1
move soapbottle 3 to garbagecan 1
```

The agent takes `soapbottle 1` and later `soapbottle 3` while the goal asks for a spraybottle. A failed episode can still have positive learning value; paraphrasing the goal or swapping the first action removes it.

Semantic-Aug (Mid, group 1, member 0, env 8). This member shares the Pick2 goal and the Mid group with the Downweight trace (29 steps, R=10, A=+0.577). Source ELV is negative. Both operators give ΔJ=+0.033 and EAG=+0.056. The router selects Semantic-Aug. The operator keeps every action and only changes the goal from `find two spraybottle and put them in garbagecan.` to `Please complete the following task: find two spraybottle and put them in garbagecan.` The logged actions are:

```
go to sidetable 1
take spraybottle 2 from sidetable 1
go to sidetable 1
go to sidetable 1
go to bathroom 1
examine sidetable 1
```

```
move spraybottle 2 to garbagecan 1
examine sidetable 1
move spraybottle 2 to garbagecan 1
go to sinkbasin 1
go to garbagecan 1
move spraybottle 2 to garbagecan 1
go to sidetable 1
take spraybottle 1 from sidetable 1
go to sidetable 1
go to sidetable 1
examine spraybottle 1
go to garbagecan 1
move spraybottle 2 to garbagecan 1
go to showerbasin 1
examine spraybottle 1
move spraybottle 1 to garbagecan 1
```

Member 2 of this group is Downweighted, while member 0 is rewritten. The operator acts on this trajectory's wording, not on the Pick2 family.

State-Branch (Mid, group 4, member 7, env 39). Clean: `put a clean pan in stoveburner.` The episode fails in 50 steps (R=0, A=-1.291) after placing an uncleaned pan on the burner. `choose_alt_action` replaces the first action `go to countertop 1` with `look` and leaves the rest unchanged. Semantic-Aug only returns ΔJ to zero ($\mathrm{EAG}_{\mathrm{sem}}$=$+0.050$); the first-action swap reaches ΔJ_{br}=$+0.045$ ($\mathrm{EAG}_{\mathrm{br}}$=$+0.095$), so the router selects State-Branch. The logged parseable actions of the source are:

```
go to countertop 1
take pan 1 from countertop 1
go to stoveburner 1
move pan 1 to stoveburner 1
take pan 2 from stoveburner 1
move pan 2 to stoveburner 1
clean pan 2
take pan 2 from stoveburner 1
go to stoveburner 1
move pan 2 to stoveburner 1
go to kitchenware 1
examine stoveburner 2
take pan 1 from stoveburner 1
move pan 1 to stoveburner 1
move pan 1 to sinkbasin 1
go to stoveburner 1
look
take pan 1 from stoveburner 1
put pan 1 in stoveburner 1
put pan 1 in stoveburner 1
go to stoveburner 1
move pan 1 to stoveburner 1
take pan 1 from stoveburner 1
move pan 1 to stoveburner 1
take pan 2 from stoveburner 1
move pan 2 to stoveburner 1
take pan 1 from stoveburner 1
clean pan 1
go to stoveburner 1
move pan 1 to stoveburner 1
take pan 2 from stoveburner 1
move pan 2 to stoveburner 1
```

Table 17: ALFWorld 1.5B Mid, matched-budget legal recollection. Four-action EVA is $+3.7$ over Uniform and $+2.1$ over Downweight+Keep. Three decode seeds: $+4.1$, $+2.8$, $+3.9$ (mean $+3.6$, 3/0).

Arm	Final J	Δ vs Uniform
Uniform	0.511	
Keep	0.508	-0.3
Downweight+Keep	0.527	$+1.6$
Four-action legal EVA	**0.548**	$\mathbf{+3.7}$
Always Semantic	0.519	$+0.8$
Always State-Branch	0.514	$+0.3$

Table 18: Leave-one-action-out from the legal router (J_{full}=0.548).

Dropped action	Final J	Drop (pp)
Semantic-Aug	0.536	-1.2
State-Branch	0.541	-0.7
Downweight	0.521	-2.7

```
move pan 1 to stoveburner 1
go to sinkbasin 1
go to sinkbasin 1
go to stoveburner 1
go to sinkbasin 1
go to sinkbasin 1
go to stoveburner 1
go to cabinet 1
go to stoveburner 1
take pan 1 from stoveburner 1
examine pan 1
move pan 1 to stoveburner 1
put pan 1 on the stove burners 1-4
take pan 1 from stoveburner 1
```

After the branch the first step is `look`, then `take pan 1 from countertop 1`, with the remainder unchanged.

F.6 Legal recollection on ALFWorld 1.5B Mid

Legal recollection re-rolls the chosen operator in the environment with the current policy and recomputes actions, rewards, old log-probabilities, and advantages (`eag_recollection=env`). The update budget matches Uniform. Table 5 is the five-arm comparison plus the two always-operator arms. The same numbers are repeated below for the leave-one-action-out audit.

Legal action mass is Downweight 48.0%, Keep 18.0%, Semantic-Aug 22.0%, State-Branch 12.0% (Aug+Branch 34%). On the same sources, soft and legal scores have ρ=0.41 and sign agreement 0.72. Legal $\rho(\text{ELV}, \text{EAG}_{\text{sem}})$=$-0.19$ and $\rho(\text{ELV}, \text{EAG}_{\text{br}})$=$-0.24$. Dropping either augmentation action lowers J; Downweight remains the largest single-action contribution. This board is one cell.

G Related Estimands, Scope, and Reproducibility

G.1 Extended comparison with related estimands

Outcome and surrogate credit. Return is an observed episode outcome, while advantage is action-level credit under the current optimization surrogate. Neither quantity measures the held-out consequence of transferring update mass between experiences.

Replay priority. Prioritized Experience Replay and its descendants use signals such as TD error to choose off-policy transitions (Schaul et al., 2016). Their objective is efficient replay inside value

learning. IA-ELV instead compares two matched on-policy updates and evaluates their behavioral difference.

Training-data influence. Influence functions and gradient-tracing methods estimate how an example affects parameters or a specified prediction target (Koh & Liang, 2017; Pruthi et al., 2020). Recent online RL attribution likewise relates recent records to a target through gradient similarity (Hu et al., 2025). IA-ELV asks a narrower intervention question: whether transferring fixed update mass toward one member of a context changes held-out agent behavior.

Learning versus augmentation value. IA-ELV uses a trajectory in its current context as the intervention unit. EAG conditions the intervention on an operator and subtracts the value of the unmodified source. Both conserve the update budget of the matched comparison, but only EAG asks whether transforming the source adds value.

G.2 Observed operating limits

Stage shift. The mapping from trajectory observables to ΔJ changes with policy competence. Gold allocation and closed-loop transfer are positive at Early, Mid, and Late; Mid is the strongest stage.

Predictor. Leak-free Mid AUROC is 0.72 to 0.80 and $\rho(\hat{\mu}, \Delta J)$ is 0.35 to 0.50. The LCB score keeps over-confident errors below 8%.

Entropy ambiguity. Lower legal-action entropy can indicate improved discrimination or increased confidence in a wrong action. This ambiguity is why ΔH is not treated as a universally positive scalar reward. Online routing stays ΔJ-primary.

Soft versus legal augmentation. Rewriting a stored trajectory does not make it on-policy. Legal recollection on ALFWorld 1.5B Mid is the four-action trainer: $+3.7$ over Uniform and $+2.1$ over Downweight+Keep.

G.3 Limits of the claims

IA-ELV is measurable under matched weak updates and is not reducible to return or advantage. Gold Mid, $\beta{=}0$ allocation improves fixed-pool J on every cell; ALFWorld 1.5B Mid is $+9.4$. A leak-free Mid predictor is same-signed, with AUROC in $[0.72, 0.80]$ and ρ in $[0.35, 0.50]$. Mid-continue transfer is positive on every cell. Soft EAG is distinct from ELV. Legal four-action EVA is the ALFWorld 1.5B Mid trainer. A change of environment or scale is not a change of estimand (Appendix A.3).

G.4 Reproducibility checklist

For each released probe or training run, the reproducibility record should contain:

- model and optimizer checkpoint identifiers;
- task, context, trajectory, and evaluation-bank identifiers;
- all random seeds and paired-branch identifiers;
- rollout, token, update, and hardware budgets;
- reference and focus weights before and after normalization;
- learning rate, update count, clipping, and advantage settings;
- entropy aggregation mode and temperature;
- feature schema, fold assignment, preprocessing state, and predictor parameters;
- router thresholds and selected-group counts;
- final and intermediate evaluation metrics;

- exclusion reason for every invalid or degenerate run.

This record is sufficient to audit budget equality, prevent context leakage, reproduce gold labels, and distinguish a failed intervention from an unexecuted one.