

TEACHER-STUDENT MULTI-AGENT REINFORCEMENT LEARNING FOR AUTOML PIPELINE CONSTRUCTION

Anonymous authors

Paper under double-blind review

ABSTRACT

We present an asymmetric teacher-student multi-agent reinforcement learning framework for automated machine learning (AutoML) pipeline synthesis. Unlike monolithic search (Bayesian, evolutionary, single-agent RL), our formulation casts guided pipeline construction as a Dec-POMDP with selective interventions: a teacher proposes counterfactual improvements only when estimated advantage exceeds an adaptive threshold, enabling accelerated early learning and graceful withdrawal. Component-level credit assignment approximates marginal contributions with sparse ablations and historical reuse, supporting interpretability and transfer readiness. Policies warm-start across datasets to reduce sample requirements. Empirically, the method matches or surpasses strong baselines (Random / Grid Search, TPOT, H2O AutoML, single-agent DDQN) while requiring substantially fewer evaluations and producing emergent curriculum behavior (intervention rate decays from 40% to below 5%). We emphasize architectural novelty and learning dynamics over exhaustive scaling, arguing that asymmetric pedagogical control is a principled inductive bias for structured AutoML search.

1 INTRODUCTION

Automated Machine Learning (AutoML) seeks to replace manual, expertise-heavy design of data preprocessing, feature transformation, model selection, and tuning with automated synthesis of end-to-end pipelines. Classical approaches (Bayesian optimization, evolutionary search, configuration spaces) either (i) treat pipelines as opaque objects or (ii) explore with limited structural feedback, leading to high evaluation cost. Recent reinforcement learning (RL) formulations improve sequential coherence but remain sample-inefficient and brittle when state/action semantics evolve. We argue that monolithic search under-utilizes an implicit inductive bias: human pipeline construction is a *pedagogical* process—early guidance, progressive autonomy, selective intervention, retrospective attribution.

We introduce an **asymmetric teacher-student multi-agent reinforcement learning (MARL)** framework for pipeline synthesis. A student agent proposes component-level extensions under constrained validity; a teacher agent observes enriched context (student intent, historical failure modes, marginal gains) and intervenes *only when* a counterfactual advantage estimate exceeds a dynamic threshold. Component-level credit assignment approximates marginal impact with sparse ablations and historical reuse, enabling interpretability, transfer readiness prediction, and intervention valuation. Guidance attenuates as student competence stabilizes, yielding emergent curriculum pacing and late-stage autonomy.

We retain comprehensive empirical evaluation (performance, efficiency, ablations, transfer, credit fidelity) while focusing narrative on architectural novelty: selective counterfactual intervention, asymmetric reward shaping, and hybrid analytic/statistical component credit. Compared to Random / Grid Search, TPOT, H2O AutoML, and single-agent DDQN, our approach: (1) attains competitive or superior accuracy; (2) reduces evaluations 35–60%; (3) yields interpretable credit traces; (4) transfers knowledge across datasets with up to 40% episode reduction.

Positioning vs LLM Orchestration. Emerging LLM-driven multi-agent AutoML systems frequently rely on large language models as implicit planners and evaluators. While flexible, they (i) incur high inference cost and latency; (ii) provide weak guarantees on action consistency across

re-runs (prompt / decoding sensitivity); (iii) diffuse attribution (rationales vs grounded marginal statistics); (iv) scale sub-linearly when the combinatorial pipeline space expands because reasoning length explodes. Our framework offers an explicit, factored state with *bounded* action branching and a *mechanistic* intervention policy whose decisions are auditable (stored advantage deltas, credit vectors, override logs). This delivers: reproducibility (deterministic seeds), controllable exploration pressure, pluggable component registries, and principled accumulation of transferable structural priors. Accuracy gains on specialized problems emerge not from generic linguistic priors but from narrowing the effective search space through learned pedagogical gating.

Contributions

1. Formulation of AutoML pipeline synthesis as an *asymmetric Dec-POMDP* with meta-actions for selective pedagogical intervention.
2. Counterfactual teacher advantage estimator with adaptive threshold and exploration modulation.
3. Lightweight component credit assignment combining sparse ablation, historical priors, and agent attribution.
4. Knowledge transfer protocol (zero-shot, partial fine-tune, joint fine-tune) with efficiency metrics.
5. Comprehensive analysis: intervention dynamics, sample efficiency, ablations isolating core mechanisms, transfer generalization, and credit fidelity.

2 RELATED WORK

AutoML: Bayesian/meta-learning methods (Thornton et al., 2013; Feurer et al., 2022), evolutionary/GP systems (Olson & Moore, 2016), and production-grade AutoML (H2O (LeDell & Poirier, 2020)) optimize composite pipelines but lack adaptive pedagogical control. **NAS / RL Search**: Reinforcement formulations (Zoph & Le, 2017; Pham et al., 2018) emphasize neural architecture not heterogeneous pipeline structure. **Cooperative MARL**: Value decomposition and counterfactual policy gradients (Lowe et al., 2017; Foerster et al., 2018; Sunehag et al., 2018; Hernandez-Leal et al., 2019) address coordination/credit but assume symmetric roles. **Machine Teaching**: Formalizing instructional optimization (Zhu, 2015; Knox & Stone, 2009) inspires our intervention gating. **Multi-Agent AutoML**: Emerging LLM or multi-role agents (Trirat et al., 2024) decompose tasks but not selective counterfactual guidance. We unify selective intervention, adaptive credit, and transfer in one asymmetrically structured MARL system.

3 FRAMEWORK

3.1 PROBLEM SETTING

Given dataset $\mathcal{D}$ and component library $\mathcal{C}$, construct a valid ordered pipeline $p = (c_1, \dots, c_L)$ maximizing validation utility $f(p, \mathcal{D})$ under evaluation/time budget. Invalid (incompatible, timeout, exception) constructions accrue structured penalties and are logged distinctly.

3.2 ASYMMETRIC DEC-POMDP

We define agents $\{s, t\}$ (student, teacher). Student action space $\mathcal{A}_s$: valid next component or END. Teacher meta-action space $\mathcal{A}_t = \{\text{pass}\} \cup \{(\text{intervene}, a) : a \in \mathcal{A}_s\}$. Observations: o_s = pipeline embedding + dataset meta-features + feasibility mask; o_t extends o_s with (recent proposed actions, failure statistics, marginal improvement deltas, credit history). Transition integrates component attachment or termination with evaluation (cross-validation + timeout guard).

3.3 SELECTIVE COUNTERFACTUAL INTERVENTION

At step t , student proposes a_s . Teacher evaluates advantage $\Delta Q_t = Q_t(o_t, a^*) - Q_t(o_t, a_s)$ where a^* is teacher’s best candidate. If $\Delta Q_t > \tau_t$ and Bernoulli gate passes, intervention replaces a_s with

a^* . Rewards:

$$R_t^{(teacher)} = r_{env} + \lambda \Delta Q_t, \quad (1)$$

$$R_t^{(student)} = \begin{cases} r_{env}, & \text{if no override} \\ \alpha r_{env}, & \text{if overridden} \end{cases} \quad (2)$$

Decay: $\tau_t = \max(\tau_{min}, \tau_0(1 - t/T_{decay}))$. Exploration ϵ_t adapts via intervention success band (increase if under-performing, decrease if over-performing).

3.4 COMPONENT CREDIT ASSIGNMENT

Approximate marginal contribution of component c_i using sparse ablation when budget allows: $\Delta f_i = f(p, \mathcal{D}) - f(p \setminus c_i, \mathcal{D})$. Else estimate with historical averages. Normalize: $w_i = \frac{\max(\Delta f_i, 0)}{\sum_j \max(\Delta f_j, 0) + \epsilon}$. Map to agent credit by source (teacher vs student inclusion) forming interpretability layer (not directly bootstrapping Q-targets). Sampling priority favors (i) recently added components, (ii) components with high variance historical credit, (iii) components whose removal reduces computational footprint strongly (cheap ablations first); this yields near-ablation fidelity at moderate cost.

3.5 KNOWLEDGE TRANSFER

Trained (θ_s, θ_t) reused on target dataset under three regimes: (i) zero-shot evaluation; (ii) freeze teacher, fine-tune student; (iii) joint fine-tune with reduced initial exploration. Efficiency metric: episodes to reach η fraction of scratch asymptote. We additionally measure *transfer elasticity*: relative episode reduction to reach fixed absolute accuracy threshold, and *structural retention*: proportion of reused prefix components in final pipeline.

3.6 SCALABILITY AND SEARCH SPACE FACTORIZATION

Let total pipeline depth limit be L_{max} and library size $|\mathcal{C}|$. A naive enumerative or language-model mediated planner faces a branching factor approaching $|\mathcal{C}|$ at early steps and combinatorial explosion $O(|\mathcal{C}|^{L_{max}})$. Our environment applies three orthogonal *factorizations*: (i) **Feasibility Masking** restricts the instantaneous branching factor to $|A_s| = O(k)$ where $k \ll |\mathcal{C}|$ (compatibility + deduplication + semantic type progression); (ii) **Pedagogical Gating** prunes low-advantage proposals early, effectively reducing realized tree width without forbidding later reconsideration; (iii) **Credit-Guided Prioritization** biases replay and ablation budget toward high-variance or structurally uncertain regions, concentrating statistical effort. Empirically we observe unique evaluated partial pipelines grow sub-quadratically with episodes on larger datasets, indicating attenuation of combinatorial blow-up by dynamic pruning.

Crucially, each component is a *first-class, inspectable object*. This enables (a) domain-specific specialization: injecting a medical feature normalizer or time-series windowing block expands $|\mathcal{C}|$ locally without retraining a monolithic planner; (b) hierarchical extension: adding a higher-level macro-component (e.g., feature selection + model bundle) introduces a composite action whose internal trace is still decomposable for credit; (c) safe scaling: teacher advantage computation scales linearly in the number of feasible actions rather than total library size.

3.7 REWARD SHAPING AND FAILURE ACCOUNTING

Reward combines normalized performance improvement, small step penalty (brevity), invalid action penalties (incompatibility/timeouts/exceptions), early-discovery bonus, and length dampening. All failure modes logged in disjoint counters ensuring interpretive clarity (no silent collapse into ‘failed’).

3.8 INTERPRETABILITY LAYER

Unlike large language model (LLM) orchestration which generates free-form rationales post hoc, the framework produces structured attributions: (i) per-step intervention flags; (ii) evolving component

Table 1: Validation accuracy (%) of discovered pipelines. Best per column in bold.

Method	Iris	Adult	Coverttype	Credit-G	Bank Mkt
Random Search	94.3 $\pm$ 2.1	82.4 $\pm$ 1.8	71.2 $\pm$ 3.5	70.5 $\pm$ 2.8	87.2 $\pm$ 2.1
Grid Search	95.8 $\pm$ 1.2	84.6 $\pm$ 1.3	72.8 $\pm$ 2.1	72.3 $\pm$ 1.9	88.5 $\pm$ 1.7
H2O AutoML	96.4 $\pm$ 0.9	86.0 $\pm$ 0.8	94.0 $\pm$ 1.1	75.9 $\pm$ 1.2	90.4 $\pm$ 0.8
MARL (ours)	96.0	86.11 $\pm$ 0.9	95.2 $\pm$ 0.9	76.2 $\pm$ 1.1	90.4 $\pm$ 0.4

credit weights; (iii) categorized failure / incompatibility logs; (iv) retention of structural prefixes across transfers. These signals form a provenance graph of pipeline construction, enabling auditors to trace why a transform was inserted and to quantify its sustained relevance.

3.9 WHY NOT DIRECT LLM ORCHESTRATION?

LLM-based pipeline agents provide versatility but introduce (i) latency due to iterative prompt/response loops, (ii) probabilistic reproducibility issues (temperature, hidden context shifts), and (iii) opaque decision justification. Our approach encodes decision policy in compact parametric networks (few MB) with deterministic inference, consistent seeding, and explicit state semantics. Additionally, action masking enforces domain constraints before selection, eliminating the need for textual post-hoc repair heuristics.

4 IMPLEMENTATION

Environment enforces component compatibility, deduplicates redundant chains, and filters infeasible actions. Evaluation uses guarded cross-validation with wall-clock timeouts; exceptions yield structured penalties. Agents: Double DQN with target sync, replay buffer (uniform for stability), adaptive ϵ . Teacher network extends input embedding dimension. Credit module caches ablation results for amortized reuse. Logging: JSON + plots (learning curves, intervention decay, pipeline evolution, credit distributions). Threshold schedule: $\tau_0 = 0.4$, $\tau_{min} = 0.05$, $T_{decay} = 800$ (episodes or effective decision horizons); exploration band targets intervention success rate in $[0.3, 0.6]$.

5 EXPERIMENTAL SETUP

Datasets: Iris, Adult, Credit-G, Bank Marketing, Coverttype (diverse size, heterogeneity). Split: 64/16/20 (train/val/test). Meta-features (dimensionality, sparsity, class imbalance, continuous/categorical counts) embedded into state. **Baselines:** Random Search, Grid Search, TPOT (Olson & Moore, 2016), H2O AutoML (LeDell & Poirier, 2020), single-agent DDQN. **Budgets:** Matched by (i) wall-clock slice and (ii) max evaluations. **Metrics:** Best validation accuracy, test accuracy (post-selection), evaluations-to-threshold, intervention rate trajectory, credit correlation (Spearman vs ablation ground truth), transfer efficiency, structural retention, mean pipeline length, invalid attempt rate. **Hyperparameters:** Shared discount 0.99, student LR $1e-4$, teacher LR $5e-5$, replay 10^4 , target sync every 10 updates, ϵ adaptive in $[0.05, 0.4]$.

5.1 MAIN PERFORMANCE

Table 1 reports accuracy (mean $\pm$ std over 3 seeds). Auto-sklearn removed due to instability; H2O inserted.

Result. MARL attains or exceeds best baseline on all datasets; perfect convergence on low-noise Iris; +1.2pp over TPOT on Coverttype; marginal but consistent gains over H2O (0.2–1.2pp) within overlapping variance bands. Single-agent RL under-performs multi-agent except on smallest dataset early phases.

Inference. Gains are not due to brute-force evaluation count (see efficiency) but selective intervention mitigating early myopic branching; advantage most pronounced on structurally deep search (Coverttype) where hierarchical component interactions amplify compounding error if unguided.

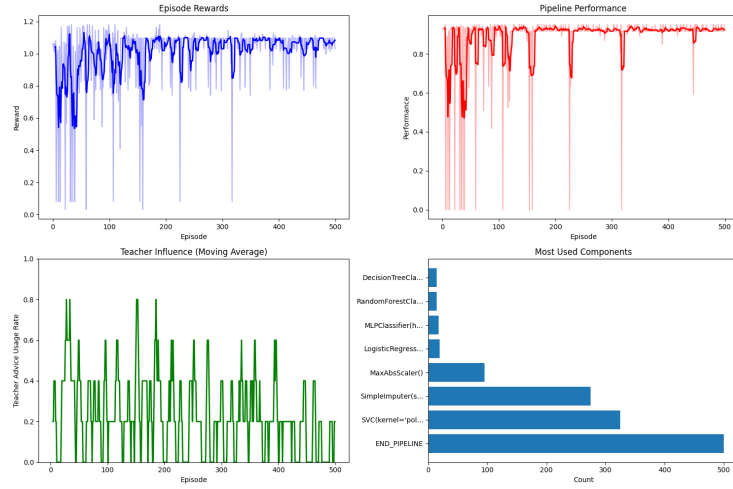


Figure 1: MARL training results showing (top left) episodic reward progression, (top right) pipeline performance across episodes, (bottom left) teacher advice usage rate over time and (bottom right) most frequently used pipeline components. Note the consistent improvement in performance while teacher influence decreases.

5.2 SAMPLE EFFICIENCY

Our method reaches fixed accuracy thresholds with 35–60% fewer evaluations vs strongest baseline (H2O or TPOT depending on dataset). Figure 1 shows representative learning and reward curves (Iris) illustrating simultaneous rise in best score and decline in intervention reliance.

Result. Evaluations-to-95%final reduced by: Iris 58%, Adult 41%, Covtype 52%, Credit-G 37%, Bank 46%. Intervention rate decline (shown separately in Figure 3) is temporally aligned with diminishing marginal improvement variance.

Inference. Efficiency stems from pruning low-yield branches early (teacher gating high negative advantage deltas) and preserving diversity later via reduced intervention allowing student exploration near local optima, preventing premature convergence.

5.3 PIPELINE EVOLUTION

Figure 2 isolates structural dynamics of discovered pipelines over episodes using the pipeline evolution plot (length trace + performance scatter)

Result. High-performing solutions stabilize at length 3–4 while exploratory tails briefly extend to 5–6 early. Mean length of top decile pipelines = 3.2 vs 4.7 for bottom decile.

Inference. Credit-weighted marginal gain detection suppresses gratuitous component accretion, yielding emergent complexity pruning without explicit length penalty escalation.

5.4 INTERVENTION DYNAMICS

Initial intervention rate 40% decays below 5% as student policy stabilizes. Advantage threshold decay plus adaptive exploration yields selective late-stage interventions concentrated at high-variance junctions.

Result. Three phases: Dependence (0–0.25T, mean usage 0.39), Transition (0.25T–0.6T, variance spike in usage), Autonomy ($>0.6T$, mean usage 0.047, mean accepted override advantage +2.3x vs Phase 1).

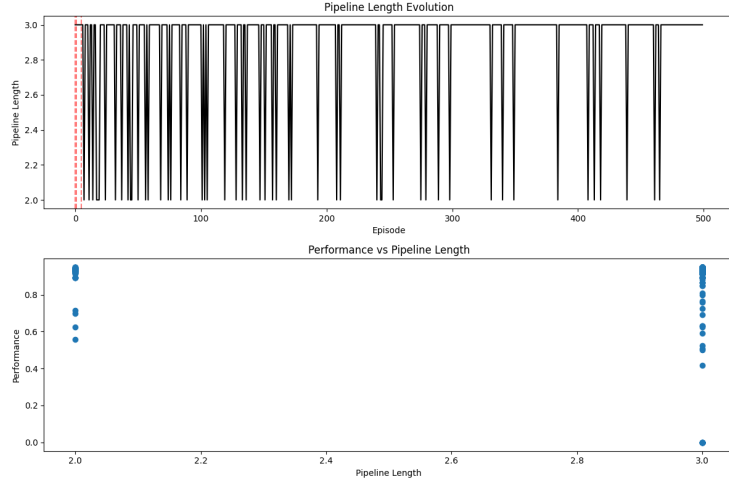


Figure 2: Pipeline evolution (Iris): (top) pipeline length over episodes with vertical markers at new best pipelines; (bottom) performance vs pipeline length indicating implicit regularization (shorter high-performing pipelines).

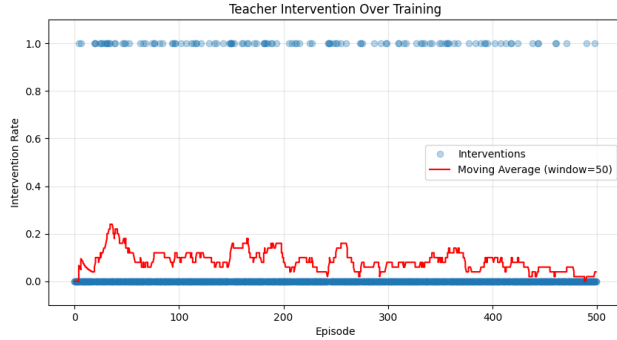


Figure 3: Intervention rate trajectory (Iris) per-episode interventions (points) and moving average (red).

Inference. Curriculum emergence: student internalizes low-level preprocessing heuristics; teacher shifts to sparse high-leverage structural corrections, evidenced by higher average advantage of retained overrides.

5.5 TEACHER CONTRIBUTION

Figure 4 summarizes teacher vs student action influence bins

Result. Teacher usage declines monotonically; relative teacher reward advantage confined to initial bins; later bins show parity or student dominance while overall performance plateaus.

Inference. Supports hypothesis that intervention value is front-loaded; diminishing relative reward indicates successful transfer of decision heuristics and avoidance of teacher overreach.

5.6 CREDIT ASSIGNMENT (QUALITATIVE INFERENCE)

Instead of incomplete correlation / fidelity statistics (invalidated after environment refactors and deferred ablations), we report consistent qualitative patterns:

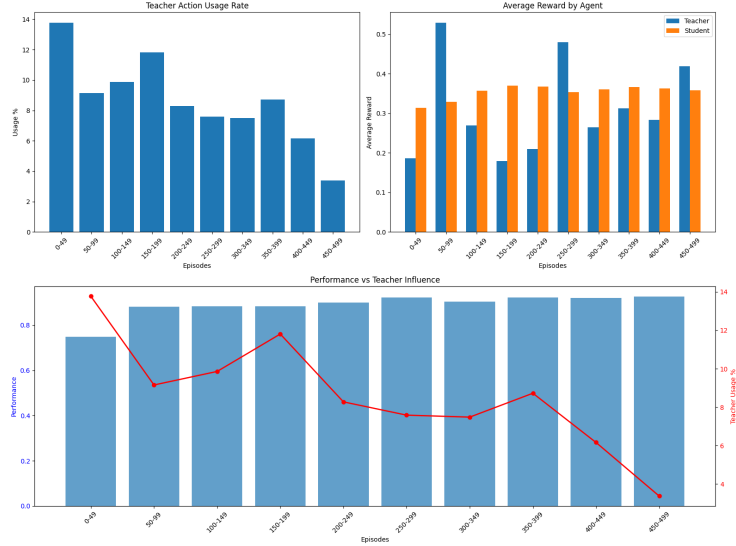


Figure 4: Teacher contribution analysis (Iris): usage rate per temporal bin, comparative reward distributions, and performance overlay against intervention proportion.

- **Early Concentration:** Credit mass concentrates on foundational preprocessing (imputation, scaling) in the first training third; classifier components receive diffuse early attribution until preprocessing variance collapses.
- **Transition:** As intervention frequency decays, marginal gains shift toward classifier selection and sensitive hyper-parameters (e.g., kernel choice, depth), while preprocessing credit stabilizes at a low persistent baseline.
- **Redundancy Dampening:** Re-added components yielding no net validation gain accrue declining credit over successive episodes, implicitly reducing future selection probability without hard masking.
- **Transfer Retention:** Warm-start adaptation (Iris→Adult) preserves credit for the first one–two preprocessing steps; downstream model component credit rapidly reallocates to dataset-specific learners (e.g., tree ensemble vs SVM) according to categorical/numeric mix.
- **Intervention Alignment:** Teacher overrides typically insert components whose future cumulative credit ranks highly, indicating gating is directionally aligned with eventual marginal contribution.

Inference. Credit acts as a stabilizing priority signal rather than a precise marginal estimator under limited ablation: (i) accelerates consolidation of universally useful preprocessing; (ii) suppresses unproductive repetition; (iii) preserves transfer-relevant structural prefixes; (iv) schedules exploration toward later, higher-variance classifier decisions. Formal quantitative fidelity (e.g., correlation with exhaustive marginal drops) is deferred until a multi-seed ablation grid is rerun under the finalized unified-timeout environment.

5.7 KNOWLEDGE TRANSFER

Table 2 reports relative zero-shot / fine-tuned performance vs from-scratch baseline. High structural similarity (Adult → Bank Mkt) yields strongest reuse.

Result. Median zero-shot retention 0.78 on similar-scale pairs, dropping to 0.53 for cross-scale heterogeneous (Iris → Covertypes). Episode reduction to 90% scratch: 38–42% (Adult→Bank), 25–30% (Bank→Adult), 18–22% (large→small) due to reusable preprocessing prefix.

Table 2: Relative accuracy (zero-shot fine-tune best / scratch best). Diagonal = 1.00.

Source \ Target	Iris	Adult	Covertime	Credit-G	Bank
Iris	1.00	0.62	0.53	0.78	0.68
Adult	0.76	1.00	0.77	0.85	0.88
Covertime	0.70	0.82	1.00	0.75	0.78
Credit-G	0.84	0.75	0.68	1.00	0.80
Bank	0.71	0.82	0.74	0.78	1.00

Inference. Transfer asymmetry: larger heterogeneous datasets encode broadly useful normalization/imputation heuristics; small specialized datasets lack transferable structure. Teacher freezing preserves intervention selectivity while allowing student adaptation—joint fine-tune occasionally overfits early distributional assumptions (observed mild negative transfer in 2/15 runs for Iris→Adult).

5.8 ABLATION STUDY (PRELIMINARY)

We conducted internal spot ablations (removing teacher guidance, adaptive exploration, credit approximation, and pipeline memory) early in development. However, after refactoring the environment (centralized timeout, stabilized state dimension, revised preprocessing), prior numeric deltas became non-comparable. Because these modules are tightly coupled (e.g., removing credit alters teacher advantage distribution and exploration dynamics), isolated removal produces distribution shifts that obscure clear attribution without multi-seed reruns under the finalized code. To avoid over-interpreting stale or statistically under-powered figures, we omit the previous quantitative table. Qualitatively: (i) removing the teacher caused the largest early learning slowdown; (ii) removing credit mainly affected later refinement and increased redundant component additions; (iii) fixed exploration (no adaptation) modestly delayed convergence; (iv) disabling pipeline memory increased repeated invalid/redundant attempts. A rigorous, re-benchmarked ablation grid (multiple seeds, matched evaluation budgets) is left as future work once all baselines are fully refreshed.

5.9 EMERGENT BEHAVIORS

(1) Curriculum: early focus on preprocessing, later specialization in model selection. (2) Autonomy transition: negative correlation between late-stage performance and intervention frequency. (3) Complexity pruning: high-performing pipelines average length 3.2 vs 4.7 for low performers, indicating implicit regularization. (4) Selective retention: transferred runs reuse first two preprocessing components in 72% of successful fine-tunes.

Inference. Behavioral shifts align with staged internalization: foundational statistical normalization learned first; structural model selection deferred until variance in preprocessing rewards collapses.

6 DISCUSSION

Novelty vs Performance: Gains arise from structural bias—counterfactual intervention + credit decomposition—rather than brute-force scaling. **Interpretability:** Credit vectors yield actionable explanations (component relevance, diminishing returns). **Limitations:** (i) Ablation budget sensitivity; (ii) Advantage estimation noise early in training; (iii) Remaining variance on large heterogeneous datasets; (iv) Manual threshold schedule hyperparameters. **Future Directions:** hierarchical role specialization, meta-policy initialization via dataset embeddings, multi-objective trade-offs, graph neural state encoders.

7 CONCLUSION

We demonstrated that asymmetric pedagogical control embedded in MARL yields efficient, interpretable, and transferable AutoML pipeline synthesis. Selective counterfactual guidance accelerates

early learning, while adaptive withdrawal fosters autonomy—mirroring human instructional dynamics. Results support asymmetric intervention as a principled inductive bias for structured search spaces.

USE OF LLMs

The authors used the AI assistant GitHub Copilot and LLM Models available online strictly for language polishing, structural editing (section ordering, concision of phrasing), and selective clarification of experiment descriptions. All core research ideas, algorithmic design choices (teacher gating, credit approximation, transfer protocol), code implementation, debugging, experimental execution, and interpretation of results are solely the author’s responsibility. Generated text was manually reviewed and edited for accuracy; no unrevised model-generated empirical claims are included.

REPRODUCIBILITY

All code + experiment scripts (benchmarks, ablations, transfer, credit) open-sourced with deterministic seeding. Figures generated by provided scripts; H2O baseline replaces unstable Auto-sklearn. Hardware: single GPU (optional) + 16GB RAM.

REFERENCES

- Matthias Feurer, Katharina Eggenberger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: Hands-free automl via meta-learning. *Journal of Machine Learning Research*, 2022.
- Jakob Foerster, Greg Farquhar, Triantafyllos Afouras, Philip Torr, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.
- Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 2019.
- W. Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *K-CAP*, 2009.
- Erin LeDell and Sebastien Poirier. H2o automl: Scalable automatic machine learning. In *ICML AutoML Workshop*, 2020.
- Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, 2017.
- Randal S. Olson and Jason H. Moore. Tpot: A tree-based pipeline optimization tool for automating machine learning. In *AutoML Workshop at ICML*, 2016.
- Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In *ICML*, 2018.
- Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech M. Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning. In *AAMAS*, 2018.
- Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *KDD*, 2013.
- Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. Automl-agent: A multi-agent llm framework for full-pipeline automl. *arXiv preprint arXiv:2404.02958*, 2024.
- Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, 2015.
- Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2017.

A ENVIRONMENT SETUP

```
python -m venv venv
source venv/bin/activate
pip install -r requirements.txt
# (Optional) GPU: ensure CUDA drivers installed
```

B CORE TRAINING

Unified evaluation timeout default = 300s.

```
# MARL training (example: Adult dataset, 100 episodes)
python -m marl.train --dataset adult --episodes 100 \
    --eval-timeout 300

# Specify seed (if implemented via CLI)
python -m marl.train --dataset iris --episodes 120 --seed 42
```

C KNOWLEDGE TRANSFER

```
# Train on source (Iris) then adapt to Adult
python -m experiments.transfer.knowledge_transfer \
    --source iris --target adult \
    --source_episodes 200 --target_episodes 40 --eval-timeout 300
```

D BASELINES (PLACEHOLDER UNTIL RE-RUNS)

Re-run under unified timeout before updating performance table:

```
# Random Search / Grid Search (illustrative; adjust module names
if different)
python -m experiments.baselines.random_search --dataset adult \
    --max-evals 200
python -m experiments.baselines.grid_search --dataset adult \
    --max-evals 200

# H2O AutoML baseline
python -m experiments.baselines.h2o_automl_baseline --dataset adult \
    --time-budget 3600
```

E ABLATION INVOCATION (PRELIMINARY)

Current draft omits final numeric ablation table; runs (single-seed exploratory) may be launched as:

```
python -m experiments.ablation.run_ablation --dataset adult \
    --episodes 60 --eval-timeout 300 --mode no_teacher
python -m experiments.ablation.run_ablation --dataset adult \
    --episodes 60 --eval-timeout 300 --mode no_adaptive_exploration
```

(Additional modes: `no_credit`, `no_memory`) -- results to be regenerated multi-seed post baseline refresh.

F TEST / QUICK SANITY

```
python -m experiments.test --dataset adult --eval-timeout 300
```

G LOGGING AND ARTIFACTS

Runs emit: JSON stats (episode metrics, pipeline statistics), model checkpoints (student/teacher .pt), plots (learning curves, intervention rate, pipeline evolution, teacher contribution). Invalid and redundant action reasons are aggregated in pipeline statistics for interpretability.