
Swap or Flop: Behavioral Interpretability of Switching in VGC-Bench Agents

Daniel Ju

Neural Networks Final Project

swap-or-flop

Abstract

Competitive Pokemon VGC doubles presents a useful setting for studying what learned agent policies actually internalize, as strong benchmark performance can reflect either strategically meaningful behavior or narrower, fixed-distribution shortcuts. This paper studies that distinction through switching behavior in VGC-Bench agents. It adds switch-level logging at the policy-to-environment decision boundary, introduces *Strategic Switching Score* (SSS), a 10-signal behavioral metric for voluntary switches, and *Residual Strategic Rate* (RSR), an outcome-linked residual diagnostic, and evaluates Random, Type Advantage Oracle, behavior cloning (BC), BC+self-play (BCSP-1), and a pure self-play (SP) subject across one locked training team and three locked transfer teams. The results show that some learned agents do acquire interpretable switching behavior: BC gives the clearest in-distribution SSS improvement over Random, while BCSP-1 provides the strongest overall combination of fixed-opponent performance and transfer-side behavioral evidence. At the same time, switching quality remains strongly dependent on team and opponent distribution, and pure self-play is a cautionary case, winning the fixed training matchup while exhibiting sparse and weakly persistent voluntary switching. The reported RSR analysis further supports this picture by showing that several BC and BCSP-1 switching differences remain outcome-linked after the current observable SSS features and HP context are adjusted for, while secondary analyses add consistent evidence from training-team diversity, state-matched Oracle agreement, and agent-vs-agent evaluation. Overall, the paper shows that agents do learn observable strategic structure to a degree, but that benchmark win rate, behavioral competence, and evidence of learned strategy should not be treated as interchangeable claims; RL evaluation in complex partially observed games benefits from targeted behavioral probes alongside reward-facing metrics.

1 Introduction

Competitive Pokemon Video Game Championships (VGC) doubles is a useful environment for studying learned decision making because it combines partial observability, simultaneous action selection, stochastic outcomes, and a large combinatorial team space in a strategically dense game (Angliss et al., 2026). Each player brings a team of six, selects four for a battle, and controls two active Pokemon at once. Each Pokemon contributes a type, moves, an ability, a held item, and remaining HP, so decisions depend on how those interacting attributes shape both active slots at the same time. A competent policy therefore has to act under uncertainty while reasoning about positioning, tempo, hidden information, partner interactions, and contingent future turns. Those properties make VGC doubles a strong setting for asking what a policy has learned, how that knowledge is expressed in behavior, and how stable that behavior remains when the evaluation distribution changes.

This last question matters because benchmark success alone is an incomplete account of policy quality. In reinforcement-learning settings, high reward can reflect a shortcut through a misspecified objective

rather than the intended behavior (Clark and Amodei, 2016). That concern is especially relevant in a domain like Pokemon, where a fixed team, a fixed opponent family, and repeated exposure to the same battle structures can all support superficially successful but weakly generalizable policies. Recent benchmark results already show that policy performance changes as team diversity increases (Angliss et al., 2026). What remains unclear is whether the policies that win in restricted settings do so by learning strategically interpretable behavior or by exploiting regularities that do not survive outside the original evaluation environment.

This paper studies that gap through *switching* behavior. A voluntary switch is one of the most strategically revealing actions in VGC doubles because it is costly, discrete, and usually motivated by an interaction among several factors rather than by a single immediate reward. A player gives up an attack or a support action in order to improve *board position* (the current two-on-two field configuration and its immediate strategic consequences), preserve a threatened Pokemon, improve a favorable *type matchup* (the rock-paper-scissors dynamic between the incoming and opposing Pokemon), align with field conditions like Trick Room or Tailwind (which respectively invert or accelerate speed order), create a favorable partner configuration for future turns, and much more. If a trained policy has internalized transferable strategic structure, that structure should be especially visible in how, when, and why it chooses to switch. By contrast, a policy that wins through narrow fixed-matchup routines may still achieve a strong win rate while exhibiting sparse, unstable, or distribution-bound switching behavior.

The central question of this project is therefore not whether VGC-Bench-style agents can win. That question has already been answered at the benchmark level; rather, the question here is what these agents learned in doing so. Specifically, do learned agents in VGC doubles produce strategically meaningful switching behavior that persists across locked transfer settings, or do they mainly reflect the fixed team and opponent distributions under which they were trained and evaluated? Framing the problem in this way places the contribution in behavioral interpretability rather than in model architecture or raw benchmark performance. The agents are treated as subjects of analysis, and the goal is to characterize the strategic content of their decisions.

To answer this question, the paper makes three contributions. First, it adds project-owned logging around the VGC-Bench and poke-env decision boundary so that switch actions can be analyzed with explicit battle-state context rather than only through aggregate win-loss outcomes. Second, it introduces Strategic Switching Score (SSS), a 10-signal behavioral metric for voluntary switches, together with Residual Strategic Rate (RSR), a residual outcome-linked diagnostic intended to test whether useful switch behavior remains after the observable SSS signals are accounted for. Third, it assembles an evidence base that combines the final fixed-opponent matrix for Random, Type Advantage Oracle, behavior cloning (BC), BC+self-play (BCSP-1), a pure self-play (SP) subject with locked transfer evaluation, BCSP training-team diversity, and exploratory agent-vs-agent analysis.

Taken together, the findings support a cautious middle ground. BC and BCSP-1 produce voluntary switches that score higher than Random in several key cells, which suggests that some trained policies do acquire interpretable switching behavior rather than merely matching a trivial baseline. However, that behavior is not uniformly robust: switching quality remains strongly dependent on team and opponent distribution, and win rate alone does not reliably track strategic switching. Pure self-play is the clearest cautionary case, since it can win the fixed training matchup while switching very little, indicating that competitive success does not by itself imply rich or transferable switching strategy. The residual analysis indicates that outcome-linked residual differences remain detectable in many cells after adjusting for observable switch quality and current HP position. Secondary analyses reinforce the same overall picture with appropriate caveats: broader BCSP training-team diversity is associated with higher aggregate SSS, state-matched Oracle agreement shows that many covered switch rows align with a type-based strategic comparator at the category level, and agent-vs-agent evaluation suggests that some behavioral distinctions survive a changed opponent distribution even though self-play still wins while largely avoiding voluntary switching.

2 Background and Related Work

Pokemon Battle Environments

Pokemon Showdown is an open-source Pokemon battle simulator (Smogon University et al., 2026). poke-env adds a Python control layer on top of Showdown, exposing battle state, legal actions,

and agent interfaces suitable for scripting and reinforcement-learning workflows (Sahovic, 2020). VGC-Bench extends this stack to the modern VGC doubles setting by packaging teams, baselines, checkpoints, human replay data, and evaluation conventions into a benchmark specifically aimed at learned agents in competitive doubles (Angliss et al., 2026). This ecosystem enables switch-level behavioral analysis on top of released benchmark subjects and evaluation structure without introducing a new simulator or training architecture.

RL and Imitation Learning in Pokemon

Prior work has already shown that Pokemon is a viable domain for modern learning-based agents, but the literature is much stronger on performance than on behavioral decomposition. Huang and Lee (2019) studied self-play PPO agents in Pokemon battle settings, establishing an early methodological precedent for policy-gradient training without handcrafted search. Wang (2024) later extended this direction with reinforcement learning plus search in singles Random Battles, showing that stronger competitive play can still emerge from the same general line of learning-based methods. Grigsby et al. (2025) then demonstrated that large sequence models trained on Pokemon Showdown data can achieve strong competitive performance, providing further evidence that strategic competence may emerge from large-scale behavioral training. Within this lineage, VGC-Bench is the most directly relevant predecessor as it moves from singles into official-format VGC doubles, provides standardized BC and BCSP subjects, and reports a critical result: strong policies in narrow settings do not necessarily remain strong as team diversity grows (Angliss et al., 2026). PPO remains the optimization backbone for the self-play agents in this paper as it offers a well-established clipped policy-gradient procedure with stable updates in large sequential decision problems (Schulman et al., 2017). Behavior cloning and BC-initialized self-play then provide the paper’s central contrast between policies shaped primarily by demonstration data and policies shaped primarily by reinforcement from game interaction.

Behavioral Interpretability

The interpretability problem addressed in this paper sits inside a familiar reinforcement-learning concern: high observed performance does not by itself reveal which strategic structure a policy has actually internalized. Formally, reward hacking occurs when an agent optimizes an imperfect proxy reward in ways that degrade performance on the underlying objective (Skalse et al., 2022). Clark and Amodei’s boat-racing work demonstrates a clear example: the agent learned to loop through point-scoring structures rather than complete the race as intended (Clark and Amodei, 2016). The complementary lesson from large-scale game RL is that strong play can also reflect real strategic structure that remains difficult to inspect directly. OpenAI Five showed that self-play reinforcement learning can recover coordinated, long-horizon behavior in a partially observed multi-agent game (Berner et al., 2019), and follow-up analysis of its hidden states found evidence of situational similarity and planning toward subgoals well before execution (Raiman et al., 2019). Within game-playing agents, the remaining gap is therefore less about whether strong play can emerge and more about how to characterize what strategic structure a learned policy is using.

Adjacent Pokemon work helps motivate switch-level analysis without yet closing that gap. Hu et al. (2024) identify switching failures as a visible behavioral failure mode in LLM-based agents, while Yashwanth and C (2025) show that instrumented decision traces and rationale-rich tournament logs can support richer analysis of strategic choices. Those contributions matter because they treat switching as a legitimate object of study and foreground the value of richer decision records. At the same time, they remain primarily qualitative, rationale-based, or outside the RL-style evaluation setting considered here. These gaps motivate a quantitative switch-level behavioral framework for learned agents in a complex partially observed multi-agent environment, with VGC doubles as the concrete testbed.

Strategic Structure and Behavioral Signals

Pokemon strategy is structured by several interacting decision axes rather than by type alone, and that broader structure is what motivates SSS as a multi-signal behavioral framework. Prior work includes Principal Trade-off Analysis, which recovers type clusters and type-cycle structure from game outcomes alone, showing that type interactions can be recovered from outcome data rather than only from explicit annotations (Strang et al., 2024). That result motivates including type advantage as one behavioral signal, but it does not exhaust the strategic content of switching in VGC doubles.

Doubles strategy guides discuss switching in terms of defensive and offensive repositioning, partner actions, speed and field conditions, ability and item context, and spread-move pressure (muffinhead and Cassie, 2011; youngjake93, 2014). These are not incidental details around a type-based core; together they show why a switch can be supported by multiple practical signals rather than by type matchup alone (youngjake93, 2014). The relevant lesson from prior work and domain structure is therefore not simply that type matters, but that any serious behavioral analysis of switching should be grounded in multiple strategically meaningful signals rather than a single proxy. SSS and RSR are built on this premise.

3 Methods

3.1 Agents

The analyses use five policy types. Random and Type Advantage Oracle are fixed project baselines rather than trained agents. Type Advantage Oracle is a scripted switch baseline that chooses among legal incoming Pokemon using a local type-advantage rule rather than learned policy optimization. BC and BCSP-1 are official VGC-Bench checkpoints and anchor the paper to the benchmark’s released learned subjects (Angliss et al., 2026). The pure self-play subject is the seed1/G1629 checkpoint trained to 5,013,504 timesteps and evaluated in the assembled evidence base.

Table 1: Evidence sources used in the paper-level analysis.

Agent	Source	Battles	Provenance
Random	Core matrix	1000 x 4	No checkpoint
Type Advantage Oracle	Core matrix	1000 x 4	Project scripted baseline
BC	Core matrix	1000 x 4	Official VGC-Bench BC checkpoint
BCSP-1	Core matrix	1000 x 4	Official 1-team BCSP checkpoint
SP	SP evaluation	1000 x 4	Project-trained seed1/G1629 SP

3.2 Switch Logging

The project’s primary engineering contribution is the switch-level instrumentation layer. Logging occurs where VGC-Bench and poke-env policy actions are converted into Pokemon Showdown orders, enabling observation of decisions immediately before it becomes an executed battle command. This choice matters methodologically because it captures the state in which the policy committed to the switch rather than a later reconstructed approximation of that state.

For each switch row, the logger stores the outgoing and incoming Pokemon, opponent active Pokemon, known move information, field conditions, legal switch options, partner action, switch type, HP differential at the moment of choice, and future turn timeline fields used by the RSR analyses. Primary SSS and RSR analyses are restricted to voluntary switches because these are the clearest policy-owned repositioning decisions in the logs. Forced switches and team-preview choices are partly imposed by battle mechanics or by the opponent, so mixing them with voluntary switches would confound the behavioral interpretation of why the policy chose to reposition.

3.3 Strategic Switching Score

The implemented SSS scores each voluntary switch with ten binary signals: offensive matchup gain, defensive damage mitigation, low-HP preservation, speed-control alignment, status or volatile escape, partner Protect cover, spread-move risk mitigation, negative stat reset, switch-in ability value, and predicted-KO avoidance. The total score is the number of active signals, so SSS ranges from 0 to 10.

The signals are binary because the goal is transparent behavioral description from logged state rather than a latent utility model. The aggregation is unweighted because the paper does not have a validated basis for assigning reliable cross-signal weights, and arbitrary weights would create a false sense of precision; improved weighting is therefore left to future work discussed later in the paper. SSS is therefore best read as a count of strategically interpretable reasons a switch is supported by observable evidence, not as a claim that every signal is equally important in actual play.

The purpose of SSS is not to prove that a switch is optimal. Its purpose is to provide a structured behavioral description of whether a switch aligns with recognizable strategic motives that can be verified from logged game state. This distinction is important: a single metric such as type advantage would capture only one narrow aspect of switching and would be vulnerable to false confidence if an agent repeatedly produced favorable matchups through a brittle team-specific reflex. The composite score is meant to reduce that risk by distributing evidence across multiple interpretable signals that correspond to how strong human switches are often justified in practice.

SSS is intentionally limited, as it measures observable strategic structure rather than latent reasoning, so a high score should be read as descriptive support for a switch, not as proof of optimality or full strategic understanding. A switch with $SSS = 0$ is unexplained by the implemented signals only, while a switch with $SSS \leq 2$ or $SSS \leq 3$ is low-scoring or weakly explained, not unexplained.

3.4 Residual Strategic Rate

RSR is the paper’s residual outcome-linked interpretability analysis. It fits a Random-baseline residual model using SSS and current HP differential, then compares exact $t+3$ HP-differential residuals for agent voluntary switches in the same condition. Here, exact $t+3$ HP differential means the change in side-to-side HP differential measured exactly three turns after the switch. The purpose of this analysis is to test whether useful switch behavior remains detectable after the observable switch-quality signal captured by SSS and the current board-health context are taken into account.

RSR remains post-hoc and observational; it is not causal, and it should not be interpreted as a counterfactual estimate of what would have happened had the switch not been made.

3.5 Statistical Tests

SSS distributions are bounded, ordinal, and built from unequal numbers of voluntary switches across cells, so the paper uses two-sided Mann-Whitney U (MWU) tests for within-condition agent-versus-Random comparisons (Mann and Whitney, 1947). MWU is appropriate because the question is distributional rather than purely mean-based, and normality assumptions are not justified for switch-level SSS values. These tests treat voluntary switch rows as the analysis unit; because switches can be clustered within battles, q-values are switch-row distributional diagnostics rather than battle-level confirmatory inference. Because the paper performs multiple such comparisons across agents and conditions, it controls the Benjamini-Hochberg false discovery rate (BH-FDR) at the conventional threshold $q = 0.05$ (Benjamini and Hochberg, 1995). BH-FDR is preferable here to an uncorrected testing regime because the central empirical question is not whether any single comparison can be declared significant in isolation, but which patterns remain credible after the full comparison family is considered.

The paper also reports Cliff’s δ so that the statistical analysis is not reduced to thresholded p-values (Cliff, 1993). Cliff’s δ is useful because it expresses the magnitude and direction of the distributional shift for ordinal outcomes in a way that complements MWU. Bootstrap confidence intervals are used for mean SSS summaries because mean SSS is a bounded, discrete summary of an ordinal-like switch-level score and the paper does not assume a convenient parametric sampling distribution for that descriptive mean (Efron and Tibshirani, 1993). Wilson intervals are used for win-rate summaries because win rate is a binomial proportion and Wilson’s construction is designed for binomial inference rather than an unconstrained normal-error summary (Wilson, 1927). The goal throughout is to report not only whether differences are statistically detectable, but also whether they are substantively meaningful and how much uncertainty remains around the descriptive summaries.

4 Experimental Setup

4.1 Team Conditions

Evaluations use four locked Regulation G teams, meaning teams built under the official competitive ruleset used by the benchmark artifacts. The training condition is G1629, a bulky Calyrex-Ice balance team with Trick Room, Icy Wind, and Electroweb speed control, while the transfer conditions are G969 (close transfer, Koraidon sun with a Trick Room mode), G1570 (medium transfer, Kyogre Tailwind balance), and G1562 (far transfer, Necrozma-Dusk-Mane bulky offense).

The four locked teams were selected to create an interpretable transfer ladder. G1629 is the training condition because it combines mixed physical and special offense with multiple speed-control mechanisms and enough repositioning complexity to expose switching decisions across several board-state regimes. G969, G1570, and G1562 then provide close, medium, and far transfer settings that increasingly alter weather mode, speed control, offensive profile, and board-position demands relative to that training condition. This grounds whether switching behavior persists across growing strategic distance from the training distribution rather than across arbitrary team swaps.

4.2 Evaluation Protocol

The primary evaluation is a controlled same-team mirror design, where both sides bring the same locked team against VGC-Bench’s SimpleHeuristicsPlayer, a scripted heuristic opponent. This design controls team composition on both sides and reduces roster and asymmetric-matchup confounds when comparing observed switching behavior across policies. The core fixed-opponent matrix uses 1000 battles per agent-condition cell, where a cell refers to one evaluated policy under one team condition at a fixed battle count. Random, Type Advantage Oracle, BC, BCSP-1, and SP subjects are each evaluated in this same-team mirror setting. This primary matrix provides the paper’s main controlled evidence because it holds team and opponent structure fixed while varying the policy under evaluation.

Secondary evidence expands the behavioral picture without replacing the primary controlled matrix. It consists of four analyses: a 100-battle SP checkpoint progression analysis on G1629; a BCSP training-team diversity analysis with 100-battle cells across the four locked conditions; covered-row state-matched Oracle agreement computed from existing voluntary switch logs; and an exploratory 100-battle agent-vs-agent matrix on G1629. Here, training-team diversity means the number of distinct locked teams used during BCSP training, comparing the official 1-team, 4-team, 16-team, and 64-team BCSP checkpoints. These analyses are diagnostic rather than final-quality evidence on the same footing as the 1000-battle matrix, and they test whether the main behavioral patterns persist over self-play training, under broader training exposure, under a stronger type-based comparator, and under a changed opponent distribution.

5 Results

5.1 Main Matrix

Table 2 summarizes the most relevant cells from the assembled evidence base, and Figure 1 shows the corresponding mean voluntary-switch SSS summaries. The first broad point is that strong fixed-opponent performance and strategically interpretable switching do not align reliably. BCSP-1, the official one-team BCSP checkpoint, achieves the strongest controlled training win rate, winning 0.839 of training/G1629 games against the fixed heuristic opponent. By contrast, BC has the highest training mean SSS among the benchmark-released learned checkpoints at 4.231, while SP is close to BCSP-1 in training win rate at 0.809 yet makes only 55 voluntary switches in 1000 games. This contrast motivates the later interpretation: a policy can succeed in the benchmark’s fixed-opponent setting while providing very little evidence of persistent switching behavior.

Table 2: Selected evidence-base cells. The full table is in Appendix D.

Agent	Cond.	Win	Mean SSS	Switches	Sw./battle	Source
Random	Training	0.066	3.691	1254	1.254	Core matrix
Type Oracle	Training	0.384	3.542	1872	1.872	Core matrix
BC	Training	0.701	4.231	1266	1.266	Core matrix
BCSP-1	Training	0.839	3.813	546	0.546	Core matrix
SP	Training	0.809	4.891	55	0.055	SP evaluation
SP	Close	0.063	4.348	273	0.273	SP evaluation
SP	Medium	0.132	3.839	31	0.031	SP evaluation
BCSP-1	Far	0.731	3.667	213	0.213	Core matrix
SP	Far	0.464	3.200	5	0.005	SP evaluation

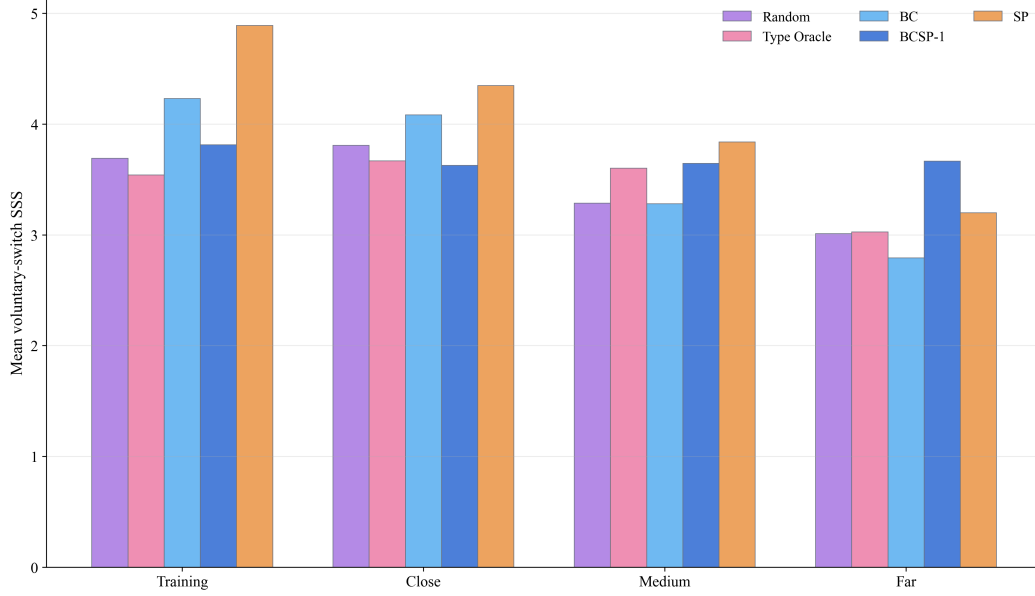


Figure 1: Mean voluntary-switch SSS by agent-condition cell. Bars summarize voluntary switches only.

5.2 SSS Effects Versus Random

Several learned-policy comparisons remain statistically distinguishable from Random after BH-FDR correction, but the pattern is not uniform across agents or conditions (Table 3 and Figure 2). That non-uniformity is informative. If switching quality simply rose with overall policy strength, the same ordering would be expected across training and transfer cells. Instead, the effects vary materially with both agent lineage and team condition, including sign changes across cells.

BC has the clearest in-distribution SSS improvement among the benchmark-released learned checkpoints. On training/G1629, BC improves on Random by $\Delta\text{SSS} = +0.541$ with Cliff's $\delta = +0.223$ and BH $q = 4.50 \times 10^{-22}$, whereas BCSP-1 is positive but smaller in the same cell at $\Delta\text{SSS} = +0.123$, $\delta = +0.069$, and $q = 0.025$. BC also remains positive on close transfer at $\Delta\text{SSS} = +0.275$ with $\delta = +0.126$ and $q = 1.04 \times 10^{-5}$, but then drops to effectively zero on medium transfer and becomes negative on far transfer at $\Delta\text{SSS} = -0.221$ with $\delta = -0.130$ and $q = 0.002$. That progression makes the distribution dependence concrete rather than rhetorical.

BCSP-1 does not dominate every transfer cell either. Its close-transfer SSS effect is slightly negative and not significant, but it turns positive again on medium transfer at $\Delta\text{SSS} = +0.358$ with $\delta = +0.154$ and $q = 0.005$, then remains clearly positive on far transfer, where the paper is most directly testing whether switching quality survives outside the training team's immediate neighborhood. In that far-transfer cell, BCSP-1 records mean SSS = 3.667 over 213 voluntary switches, with $\Delta\text{SSS} = +0.654$, Cliff's $\delta = +0.317$, and BH $q = 3.69 \times 10^{-12}$. This is stronger transfer evidence than BC provides in the same cell and is one reason BCSP-1 remains central to the paper's interpretation.

SP has the largest training SSS effect in the selected table, but the result must be interpreted alongside its switch count. A large effect computed from a sparse set of switches is evidence about the quality of those observed switches, not evidence that switching remained a frequent or central part of the learned policy. More generally, higher SSS should be read as stronger observable strategic support for a switch because it indicates that more of the paper's implemented signals fired for that decision. It should not be read as proof that the switch was globally optimal, since the score is built from observable strategic support rather than from full-game counterfactual evaluation.

Table 3: Selected voluntary-switch SSS effects versus Random. Tests are two-sided Mann-Whitney U with BH-FDR correction; Cliff’s δ is the distributional effect size. The analysis unit is the voluntary switch row.

Agent	Cond.	Δ SSS	Cliff δ	BH q	n	Sig.
BC	Training	+0.541	+0.223	4.50e-22	1266	yes
BCSP-1	Training	+0.123	+0.069	0.025	546	yes
SP	Training	+1.200	+0.545	6.42e-12	55	yes
BC	Close	+0.275	+0.126	1.04e-05	700	yes
SP	Close	+0.540	+0.264	6.42e-12	273	yes
BC	Medium	-0.004	+0.009	0.776	594	no
SP	Medium	+0.552	+0.253	0.019	31	yes
BCSP-1	Far	+0.654	+0.317	3.69e-12	213	yes
SP	Far	+0.187	+0.096	0.697	5	no

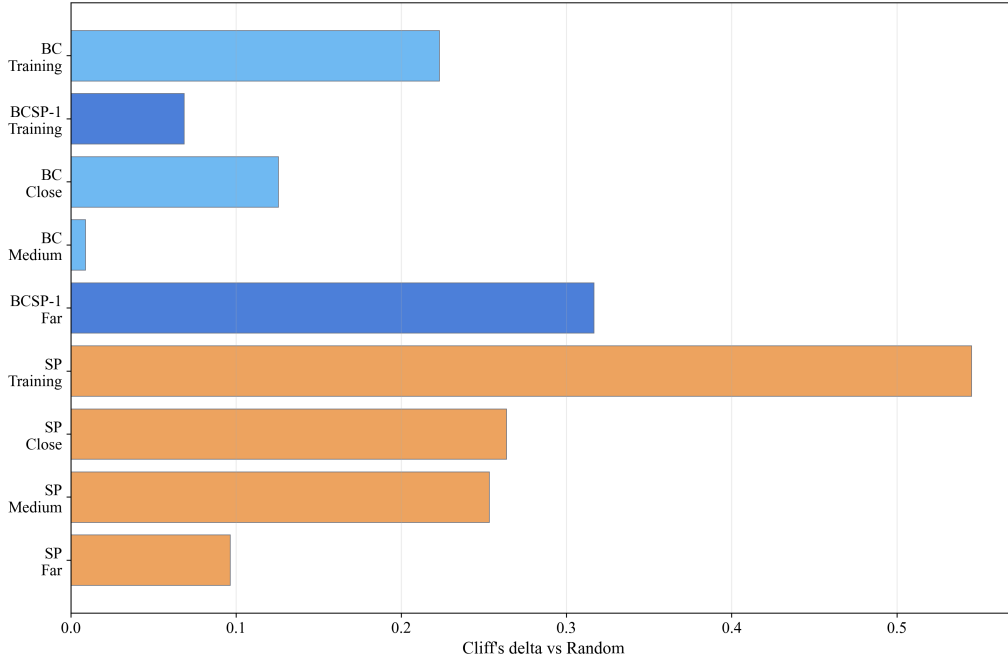


Figure 2: Selected Cliff’s δ effect sizes for voluntary-switch SSS versus Random.

5.3 Win Rate and Switching Are Not Redundant

Figure 3 illustrates the relationship between outcome success and switching behavior. SP is the most important example because it sits near the top of the win-rate axis while remaining near the bottom of the switching-frequency axis. Conversely, Random and Type Advantage Oracle, the scripted baseline that selects switch-ins by a local type-advantage rule, switch much more frequently without being the strongest performers. The implication is not that switching is unimportant: switching frequency by itself is not a proxy for strategic quality, and win rate by itself is not a proxy for interpretable strategic behavior. The two measures indicate different aspects of the learned policy, which is why this paper treats switching behavior as an interpretability target in its own right.

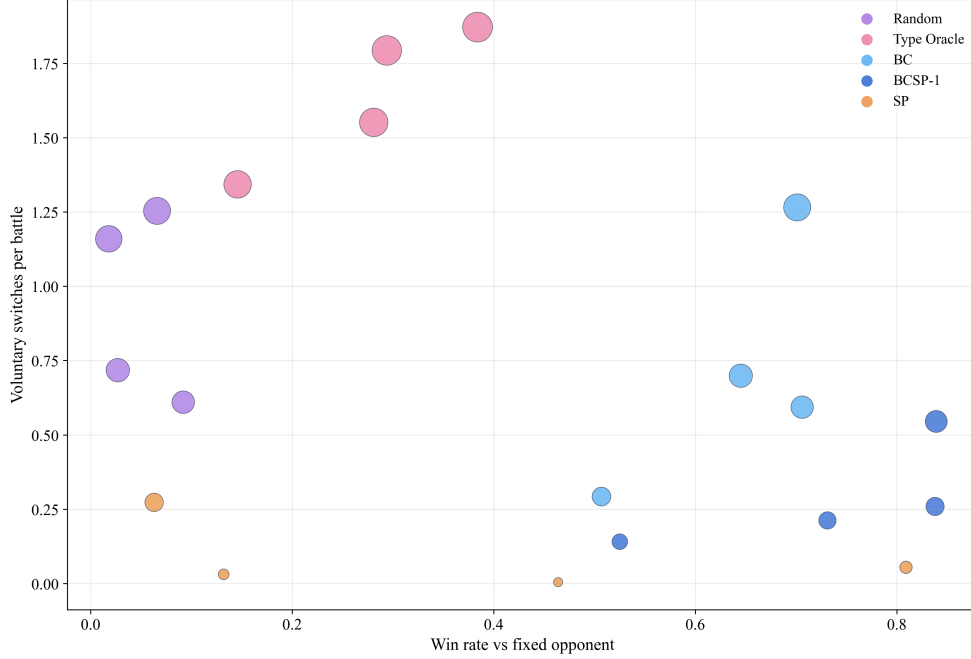


Figure 3: Win rate versus voluntary switch frequency for the evidence base. Each marker is one agent-condition cell; marker size scales with voluntary switch count. Win rate is against fixed controlled opponents.

5.4 Self-Play

The self-play subject cannot be dismissed as merely nonfunctional, because it wins the fixed training matchup at 0.809. However, the behavioral pattern attached to that win rate is narrow. On training/G1629, SP makes only 55 voluntary switches in 1000 battles, or 0.055 per battle. In close transfer it switches more often, 273 times or 0.273 per battle, but the win rate collapses to 0.063. In medium transfer it records only 31 voluntary switches and wins at 0.132, while in far transfer it records only 5 voluntary switches in 1000 battles. Far transfer also fails to produce a significant SSS effect despite a nominal win rate of 0.464. Those transfer cells therefore provide little evidence of persistent switching behavior even when some observed SSS values remain nontrivial.

The self-play checkpoint progression analysis sharpens that interpretation. As training proceeds, win rate rises dramatically while voluntary switching collapses from 1.28 switches per battle at checkpoint 0 to 0.00–0.05 after training. The exploratory agent-vs-agent analysis then shows the same pattern under a changed opponent distribution: SP wins 322 of 400 agent-vs-agent battles on G1629 while making zero voluntary switches. Table 4 summarizes the SP controlled-matrix cells. The available evidence is more consistent with a policy that deemphasized voluntary switching, or concentrated it into a very small number of contexts, than with one that learned robust, transferable switching strategy.

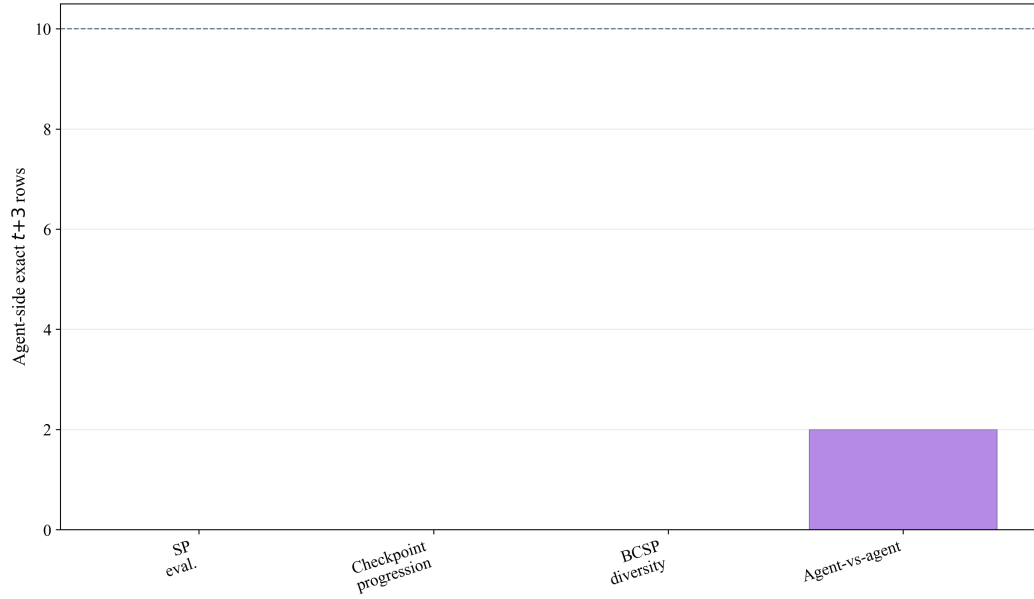
Table 4: SP versus Random baselines from the controlled matrix. Far transfer is not significant and is based on only five voluntary switches.

Cond.	Win	Switches	Sw./battle	Mean SSS	Δ SSS	BH q
Training	0.809	55	0.055	4.891	+1.200	6.42e-12
Close	0.063	273	0.273	4.348	+0.540	6.42e-12
Medium	0.132	31	0.031	3.839	+0.552	0.019
Far	0.464	5	0.005	3.200	+0.187	0.697

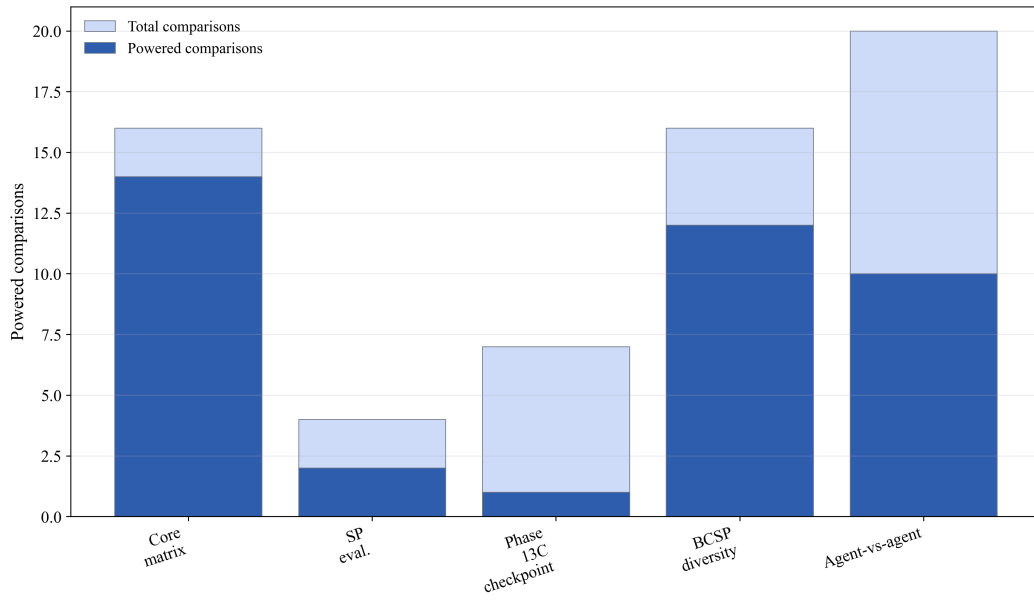
5.5 Residual Strategic Rate

The reported post-hoc RSR analysis is usable in much more of the matrix than the earlier exact $t+3$, $SSS = 0$ diagnostic (Table 5 and Figure 4). In the primary controlled analysis it contains 14,683 paired voluntary switches, with exact $t+3$ HP-differential residual comparisons powered for 14 of 16 non-random cells and good common support for 10 of those 14 powered cells. The substantive pattern is that BC and BCSP-1 remain positive relative to Random in the primary matrix. On training/G1629, BC has an exact $t+3$ residual effect of +0.492 with 95% CI [0.392, 0.596] and BCSP-1 has +0.484 with 95% CI [0.347, 0.620], whereas Type Advantage Oracle is near zero at -0.006 with 95% CI $[-0.091, 0.079]$. SP is mixed under its own seed1/G1629 evaluation: training exact $t+3$ is not clearly positive at -0.293 with 95% CI $[-0.595, 0.013]$. This pattern indicates that some outcome-linked switching differences remain after the current SSS feature set is accounted for, but the SP residual claim is not uniformly positive.

Transfer results reinforce the same interpretation, albeit unevenly. BCSP-1 remains positive in all three transfer conditions, reaching +0.702 on close transfer with 95% CI [0.551, 0.844], +0.673 on medium transfer with 95% CI [0.516, 0.858], and +0.371 on far transfer with 95% CI [0.227, 0.519], although the training, medium-transfer, and far-transfer BCSP-1 cells carry only weak common support. BC is also positive in close and medium transfer at +0.382 and +0.746, while Type Advantage Oracle becomes positive in medium and far transfer at +0.216 and +0.214. This means the RSR pattern is not just a restatement of the direct SSS table: for example, BCSP-1 is negative and non-significant on close transfer under direct SSS comparison, yet it remains strongly positive under RSR. SP is powered only for training and close transfer at exact $t+3$; its close-transfer residual effect is strongly positive at +1.517 with 95% CI [1.379, 1.653], while medium and far transfer are sparse at exact $t+3$. That close-transfer SP result suggests that the few switches SP still makes can be locally beneficial even when the overall policy performs poorly on the matchup. Because this analysis is post-hoc, these residual results should be read as outcome-linked context rather than as confirmatory evidence.



(a) Earlier exact $t+3$, $SSS = 0$ diagnostic



(b) Reported RSR analysis

Figure 4: Status comparison between the earlier exact $t+3$, $SSS = 0$ diagnostic and the reported post-hoc RSR analysis.

Table 5: RSR status by evidence source; all entries are post-hoc diagnostics.

Source	Rows	Powered exact $t+3$	Good support	Status
Core matrix	14683	14/16	10	post hoc only
SP eval.	4106	2/4	2	post hoc only
BCSP diversity	4170	12/16	5	post hoc only
Agent-vs-agent	2027	10/20	see Appendix figure	post hoc only

5.6 Secondary Analyses

Beyond the previously discussed self-play checkpoint progression, three secondary analyses extend the paper’s evidence base beyond the primary fixed-opponent matrix. The BCSP training-team diversity analysis compares the official 1-team, 4-team, 16-team, and 64-team BCSP checkpoints, where training-team diversity means the number of distinct locked teams seen during BCSP training. The synthesis in Table 6 reports weighted mean SSS values of 3.760, 4.070, 4.149, and 4.440 for BCSP 1/4/16/64 respectively, which suggests that broader training exposure is associated with higher weighted mean SSS in this exploratory summary. This pattern is useful because it connects a concrete training variable to the paper’s behavioral metric, although the cells remain 100-battle summaries and should not be treated as final-quality monotonic evidence.

The state-matched Oracle analysis strengthens the Oracle discussion by moving beyond ordinal action-order agreement. Here, Oracle means an offline replay of the same local type-advantage switch rule used by the Type Advantage Oracle baseline, not a globally optimal policy. Across about 12.4k covered evaluated-player voluntary switch rows, the diagnostic reports overall exact top-1 incoming agreement of 0.682 and type-advantage category agreement of 0.914; the compact table excludes one corrected-SP covered row. For covered learned-agent switch rows, type-advantage category agreement is high, while exact incoming agreement is more modest. This is evidence that many scored switches align with a meaningful type-based strategic lens at the category level when the same decision state is replayed. The coverage caveat remains essential: existing logs do not support a full all-action Oracle replay because non-switch decision slots lack complete state snapshots, so the result should be read as a covered-row switch diagnostic rather than as imitation or optimality evidence.

The agent-vs-agent matrix changes the opponent distribution from fixed SimpleHeuristics to agent opponents, as summarized in Table 6 and Figure 5. BC and BCSP-1 preserve meaningful aggregate switching behavior under this changed distribution, which suggests that at least some of the behavioral structure detected by SSS is not solely an artifact of the original heuristic opponent. SP again wins often without voluntary switching, reinforcing the conclusion that its fixed-matchup success should not be reinterpreted as learned switching persistence. These analyses expand the paper’s evidence base in useful ways, even though they still require the caveats attached to smaller cells, partial coverage, or changed opponent structure.

Table 6: Secondary-analysis summaries. These analyses broaden the evidence base but with caveats because they use 100-battle cells or changed opponent distributions.

(a) BCSP diversity

Checkpoint	Win	Switches	Weighted SSS
BCSP-1	0.780	96	3.760
BCSP-4	0.818	129	4.070
BCSP-16	0.782	94	4.149
BCSP-64	0.833	109	4.440

(b) Oracle agreement

Agent	Covered rows	Top-1	Type category
Random	2766	0.352	0.825
Type Advantage Oracle	6116	1.000	1.000
BC	2480	0.362	0.825
BCSP-1	1044	0.457	0.853

(c) Agent-vs-agent

Agent	Win	Switches	Mean SSS	Sw./battle
Random	0.045	375	3.664	0.938
Type Oracle	0.285	832	3.376	2.080
BC	0.585	640	4.142	1.600
BCSP-1	0.780	180	3.750	0.450
SP	0.805	0	n/a	0.000

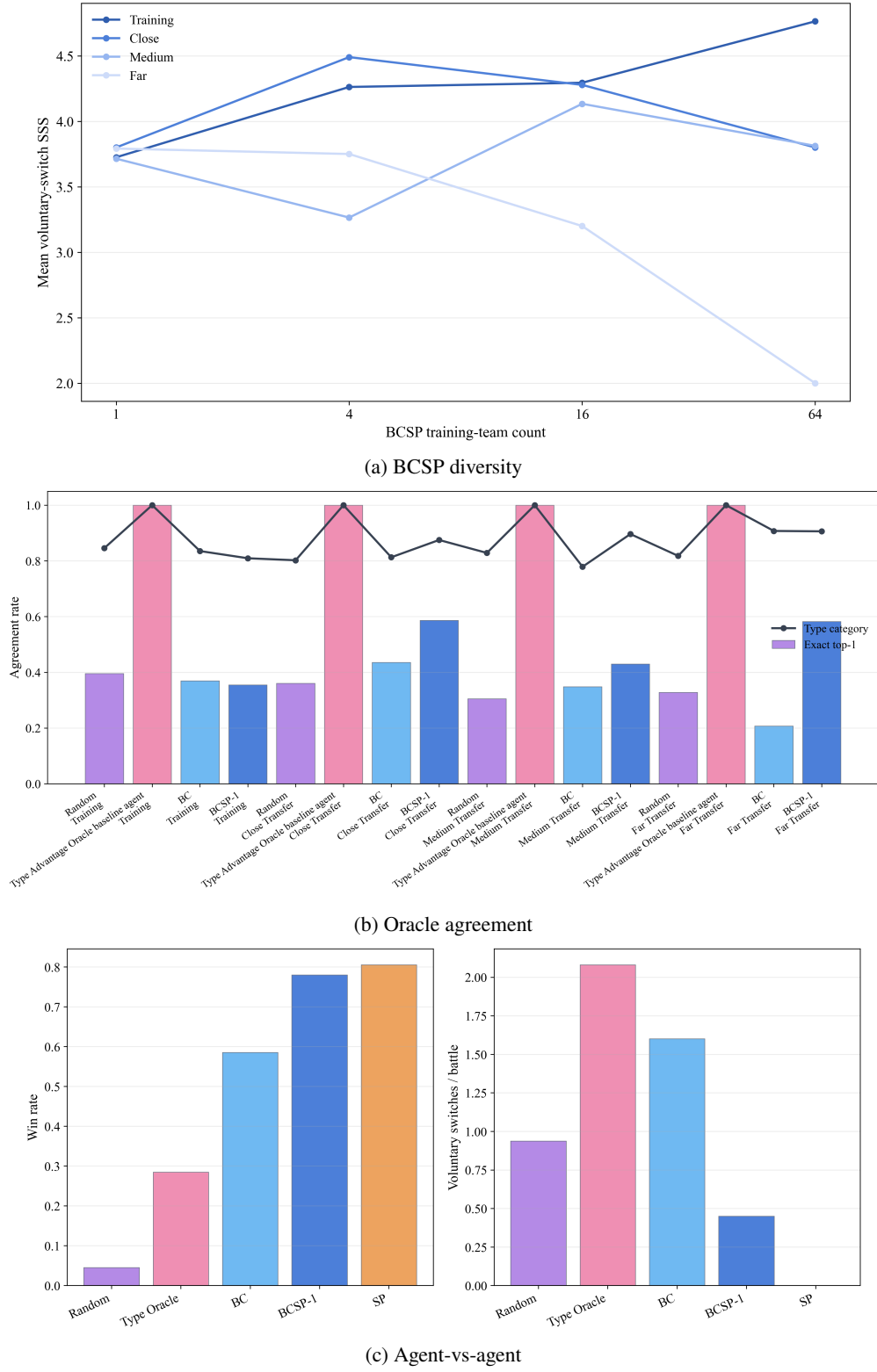


Figure 5: Secondary analyses: BCSP training-team diversity, state-matched Oracle switch-row agreement, and exploratory agent-vs-agent behavior.

6 Discussion

The evidence base supports a narrower and more defensible claim than either a pure benchmark-performance story or a strong strategic-emergence story would allow. Learned agents can produce switching behavior that is measurably different from Random under a structured behavioral score, and those differences survive multiple-comparison control in many cells. That matters because it shows that switch behavior contains interpretable signal rather than noise. At the same time, the paper does not show that high-performing agents uniformly learned robust strategic switching, nor does it show that benchmark success alone can stand in for behavioral understanding.

The most important conceptual result is that win rate and strategic behavior separate in informative ways. BCSP-1 is the strongest fixed-opponent training performer, while BC shows the clearest in-distribution SSS improvement. SP is the clearest cautionary example of why outcome success is not sufficient evidence of strategic structure. It can win the locked training matchup strongly while switching rarely, and the checkpoint-progression and agent-vs-agent diagnostics both indicate that training success was compatible with a collapse in voluntary switching rather than with the refinement of a rich switching policy. This makes the self-play result more informative, not less. These results are not saying that self-play failed; rather demonstrating that self-play can succeed in a narrow competitive sense while providing weak evidence for the behavioral competence that this paper set out to study.

In this respect, SSS is useful because it changes the object of evaluation. Instead of asking only whether the agent won, it asks whether the observable structure of the agent’s switches resembles strategically meaningful play. The answer is cautiously positive for several learned agents. BC and BCSP-1, in particular, produce switching behavior that is often distinguishable from Random, and BCSP-1’s medium- and far-transfer results suggest that some of that structure survives beyond the training condition. The residual analysis strengthens this interpretation by showing that several learned-agent differences remain detectable even after the current observable SSS features and HP context are accounted for, though only as post-hoc outcome-linked evidence and not uniformly for corrected SP. The secondary analyses broaden this picture without overturning the core matrix: diversity, state-matched Oracle, and agent-vs-agent diagnostics each point to behavioral structure that is not fully captured by win rate alone, while retaining their smaller-cell, partial-coverage, or changed-opponent caveats.

Limitations

What the paper does not support is equally important. It does not support a claim that SSS measures optimality, human likeness, or full strategic understanding. SSS is a heuristic behavioral score grounded in observable battle state. It can detect patterns involving type position, HP preservation, speed control, spread pressure, status, ability value, and partner cover, but it cannot verify opponent modeling, long-horizon sequencing, or hidden-state reasoning. Because SSS is binary and unweighted, it can also miss intensity differences and produce false positives when a locally favorable signal does not translate into globally strong play. For that reason, high SSS should be read as evidence of observable strategic support, not as proof that the agent “understands” the game in a human sense.

Likewise, the paper does not support a confirmatory residual-strategy result. The earlier preregistered residual diagnostic is underpowered, and the reported RSR analysis, while useful, remains post-hoc and observational. Its role is to provide outcome-linked context for switching behavior and to test whether residual strategic structure remains after the current SSS feature set is accounted for, not to establish a causal estimate of strategic value or to directly reveal hidden reasoning.

The evaluation setting also imposes a serious external-validity boundary. The primary evidence base is built against fixed SimpleHeuristics opponents in same-team mirror games. That design is valuable because it controls the comparison and makes the switch logs interpretable, but it also narrows the world in which the conclusions are directly valid. The exploratory agent-vs-agent evaluation partially widens this analytical world by changing the opponent distribution, yet it does so only on training/G1629 and at 100 battles per pair. For that reason, the paper should be read as a careful behavioral analysis of benchmark-trained agents under controlled conditions, not as a demonstration of unrestricted strategic generalization.

Future Work

Future work should build directly on these limitations. The most important next step would be to rerun the strongest diversity and agent-vs-agent analyses at final-quality scale so that those claims are no longer secondary. A second priority would be to preregister a stronger residual-outcome framework before evaluation, whether as a refined RSR or a different outcome-linked design, so that residual strategic claims do not depend on post-hoc rescue analyses. A third priority would be to improve the SSS aggregation itself. The current unweighted binary score is intentionally conservative, but future work could test validated signal weighting through expert elicitation, data-driven weighting, or calibration against human judgments and stronger downstream outcome criteria. A related future-work direction is to test whether richer training exposure, including broader team diversity, stronger curricula, or more varied opponent distributions, leads agents to acquire more persistent strategic switching structure. The BCSP diversity results make that possibility plausible, but the self-play checkpoint progression in this project shows that more training alone should not be assumed to produce richer switching behavior. Finally, the logging system could be extended with fuller state capture for non-switch decisions and paired with expert human ratings, allowing future work to compare SSS against richer notions of strategic quality rather than only against observable signal activation.

7 Conclusion

Swap or Flop uses VGC doubles to ask a broader RL question: when a learned policy performs well, what can actually be claimed about the strategic structure behind that performance? The answer is narrower than a pure benchmark narrative but stronger than a null result. Behavioral analysis of switching recovers meaningful differences between agents that win rate alone does not show. BC gives the clearest in-distribution evidence of strategically supported switching, BCSP-1 gives the strongest overall combination of competitive performance and transfer-side behavioral evidence, and the reported residual analysis adds limited outcome-linked support in the same direction. Consequently, these results support a real but bounded claim that some learned policies acquire observable strategic structure that is not reducible to raw benchmark success.

The self-play result shows why that distinction matters. SP can succeed in a controlled benchmark setting while providing weak evidence of persistent, transferable switching behavior in the decision class being studied. More broadly, that is the paper’s main conclusion for RL evaluation: outcome metrics and behavioral competence should not be collapsed into the same claim. In complex partially observed environments, stronger evaluation comes from pairing reward-facing benchmarks with targeted behavioral probes that test what the policy is actually doing. The immediate path forward is to validate these behavioral probes at larger scale and under broader opponent distributions.

References

- Cameron Angliss, Jiaxun Cui, Jiaheng Hu, Arrasy Rahman, and Peter Stone. VGC-Bench: Towards mastering diverse team strategies in competitive Pokémon. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026)*, Paphos, Cyprus, 2026. International Foundation for Autonomous Agents and Multiagent Systems. arXiv:2506.10326.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.
- Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Ponde de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *arXiv*, 2019. arXiv:1912.06680.
- Jack Clark and Dario Amodei. Faulty reward functions in the wild. <https://openai.com/research/faulty-reward-functions>, 2016. Accessed 2026-05-04.
- Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993. doi: 10.1037/0033-2909.114.3.494.
- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.
- Jake Grigsby, Yuqi Xie, Justin Sasek, Steven Zheng, and Yuke Zhu. Human-level competitive Pokémon via scalable offline reinforcement learning with transformers. *Reinforcement Learning Journal*, 6:2685–2719, 2025.
- Sihao Hu, Tiansheng Huang, and Ling Liu. PokeLLMon: A human-parity agent for Pokemon battles with large language models. *arXiv*, 2024. arXiv:2402.01118.
- Dan Huang and Scott Lee. A self-play policy optimization approach to battling Pokémon. In *2019 IEEE Conference on Games (CoG)*, pages 1–4, 2019. doi: 10.1109/CIG.2019.8848014.
- Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.
- muffinhead and Cassie. Fifth generation double battle guide. https://www.smogon.com/resources/competitive/bw/double_battles, 2011. Smogon University. Accessed 2026-05-04.
- Jonathan Raiman, Susan Zhang, and Filip Wolski. Long-term planning and situational awareness in OpenAI five. *arXiv*, 2019. arXiv:1912.06721.
- Haris Sahovic. poke-env: Pokémon ai in python. <https://github.com/hsahovic/poke-env>, 2020. Accessed 2026-05-04.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv*, 2017. arXiv:1707.06347.
- Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv*, 2022. arXiv:2209.13085.
- Smogon University, Guangcong Luo, and contributors. Pokémon Showdown: Pokémon battle simulator. <https://github.com/smogon/pokemon-showdown>, 2026. Accessed 2026-05-04.
- Alexander Strang, David Sewell, Alexander Kim, Kevin Alcedo, and David Rosenbluth. Principal trade-off analysis. *Information Visualization*, 23(3):258–271, 2024. doi: 10.1177/14738716241239018.
- Jett Wang. Winning at Pokémon random battles using reinforcement learning. M.eng. thesis, Massachusetts Institute of Technology, 2024.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Tadisetty Sai Yashwanth and Dhatri C. A multi-agent Pokemon tournament for evaluating strategic reasoning of large language models. *arXiv*, 2025. arXiv:2508.01623.

youngjake93. The art of switching in doubles. <https://www.smogon.com/smog/issue32/switching-doubles>, 2014. Smogon University. Accessed 2026-05-04.

A SSS Signal Definitions

SSS v1 uses ten binary signals:

1. Offensive matchup gain: incoming Pokemon has at least 1.5x type pressure into an opponent active.
2. Defensive damage mitigation: incoming Pokemon resists or is immune to a known opponent attack.
3. Low-HP preservation: switched-out Pokemon is at or below 25% HP.
4. Speed-control alignment: incoming speed is favorable under Trick Room, Tailwind, or priority context.
5. Status or volatile escape: switching removes meaningful status or harmful volatile effects.
6. Partner Protect cover: partner uses Protect or a protect-like move on the same turn.
7. Spread-move risk mitigation: incoming Pokemon reduces known spread-move pressure.
8. Negative stat reset: switching clears negative stat stages.
9. Switch-in ability value: incoming or outgoing ability has immediate contextual value.
10. Predicted-KO avoidance: outgoing Pokemon is plausibly in KO range of a known opponent attack.

B Locked Teams

Table 7: Locked team conditions.

Condition	Team	Role	Archetype
training	G1629	controlled training	Calyrex-Ice Trick Room balance
close_transfer	G969	close transfer	Koraidon sun with Trick Room mode
medium_transfer	G1570	medium transfer	Kyogre Tailwind balance
far_transfer	G1562	far transfer	Necrozma-Dusk-Mane bulky offense

C Self-Play Provenance

The final paper uses the seed1/G1629 pure self-play checkpoint as its SP subject. An earlier seed9/G1549 self-play run was later traced to the wrong team mapping and is therefore excluded from paper claims. It is retained only as project provenance in the repository handoff documents. This appendix is the only place where that methodological history is stated explicitly.

D Full Results

Table 8: Full evidence-base summary. Core matrix rows come from the benchmark-released agents, and SP rows come from the project-trained self-play evaluation.

Agent	Cond.	Win	Mean SSS	95% CI	Switches	Sw./battle	Source
Random	Training	0.066	3.691	[3.623, 3.763]	1254	1.254	Core matrix
Type Oracle	Training	0.384	3.542	[3.482, 3.601]	1872	1.872	Core matrix
BC	Training	0.701	4.231	[4.156, 4.313]	1266	1.266	Core matrix
BCSP-1	Training	0.839	3.813	[3.696, 3.934]	546	0.546	Core matrix
SP	Training	0.809	4.891	not bootstrapped	55	0.055	SP evaluation
Random	Close	0.018	3.808	[3.730, 3.883]	1160	1.160	Core matrix
Type Oracle	Close	0.294	3.668	[3.613, 3.722]	1794	1.794	Core matrix
BC	Close	0.645	4.083	[3.981, 4.180]	700	0.700	Core matrix
BCSP-1	Close	0.838	3.627	[3.465, 3.785]	260	0.260	Core matrix
SP	Close	0.063	4.348	not bootstrapped	273	0.273	SP evaluation
Random	Medium	0.027	3.287	[3.187, 3.393]	718	0.718	Core matrix
Type Oracle	Medium	0.146	3.602	[3.530, 3.672]	1343	1.343	Core matrix
BC	Medium	0.706	3.283	[3.190, 3.379]	594	0.594	Core matrix
BCSP-1	Medium	0.525	3.645	[3.411, 3.887]	141	0.141	Core matrix
SP	Medium	0.132	3.839	not bootstrapped	31	0.031	SP evaluation
Random	Far	0.092	3.013	[2.933, 3.092]	610	0.610	Core matrix
Type Oracle	Far	0.281	3.028	[2.982, 3.079]	1552	1.552	Core matrix
BC	Far	0.507	2.792	[2.676, 2.918]	293	0.293	Core matrix
BCSP-1	Far	0.731	3.667	[3.521, 3.812]	213	0.213	Core matrix
SP	Far	0.464	3.200	not bootstrapped	5	0.005	SP evaluation

Table 9: Full SSS effect table versus Random.

Agent	Cond.	Δ SSS	Cliff δ	Magnitude	BH q	n	Sig.
Type Oracle	Training	-0.149	-0.077	negligible	4.27e-04	1872	yes
BC	Training	+0.541	+0.223	small	4.50e-22	1266	yes
BCSP-1	Training	+0.123	+0.069	negligible	0.025	546	yes
SP	Training	+1.200	+0.545	large	6.42e-12	55	yes
Type Oracle	Close	-0.139	-0.058	negligible	0.009	1794	yes
BC	Close	+0.275	+0.126	negligible	1.04e-05	700	yes
BCSP-1	Close	-0.181	-0.078	negligible	0.052	260	no
SP	Close	+0.540	+0.264	small	6.42e-12	273	yes
Type Oracle	Medium	+0.315	+0.136	negligible	6.39e-07	1343	yes
BC	Medium	-0.004	+0.009	negligible	0.776	594	no
BCSP-1	Medium	+0.358	+0.154	small	0.005	141	yes
SP	Medium	+0.552	+0.253	small	0.019	31	yes
Type Oracle	Far	+0.015	+0.013	negligible	0.656	1552	no
BC	Far	-0.221	-0.130	negligible	0.002	293	yes
BCSP-1	Far	+0.654	+0.317	small	3.69e-12	213	yes
SP	Far	+0.187	+0.096	negligible	0.697	5	no

E Degenerate Diagnostics

Table 10: Selected degenerate-behavior checks. Sparse self-play switching is an interpretability result, not a preprocessing error.

Agent	Cond.	Switches	Sw./battle	Exact t+3	Forced caveat
Random	Training	1254	1.254	804	no
BC	Training	1266	1.266	924	no
BCSP-1	Training	546	0.546	263	yes
BCSP-1	Medium	141	0.141	101	yes
BCSP-1	Far	213	0.213	140	yes
SP	Training	55	0.055	16	see self-play progression analysis
SP	Close	273	0.273	146	see self-play progression analysis
SP	Medium	31	0.031	2	see self-play progression analysis
SP	Far	5	0.005	1	see self-play progression analysis

F RSR Provenance

The main text reports RSR through the post-hoc residual analysis that uses all voluntary switches. The earlier preregistered exact $t+3$, $SSS = 0$ diagnostic is retained here only as a provenance caveat because it is underpowered and non-estimable under the available data regime.

G Additional Figures

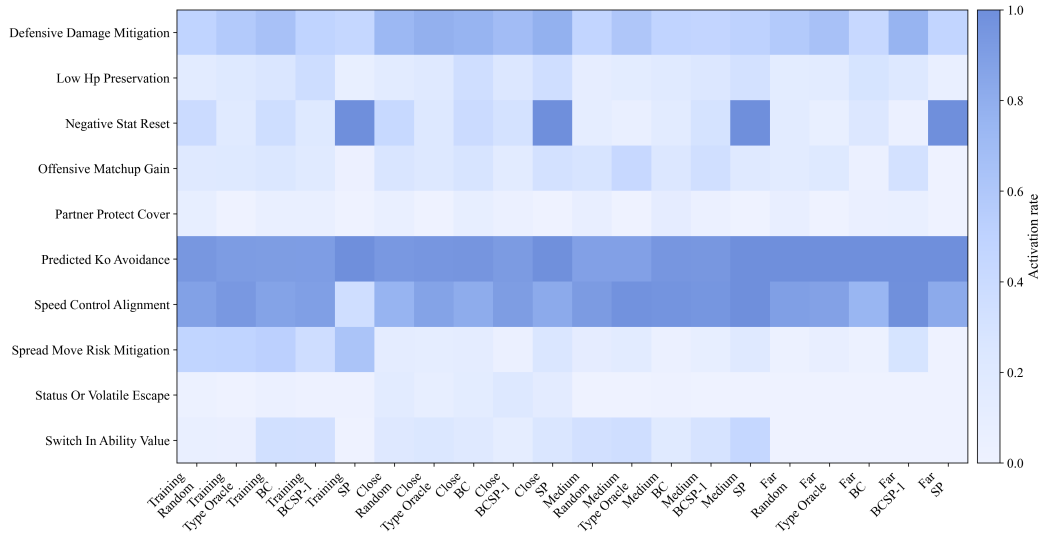


Figure 6: Signal activation heatmap for the core controlled matrix.

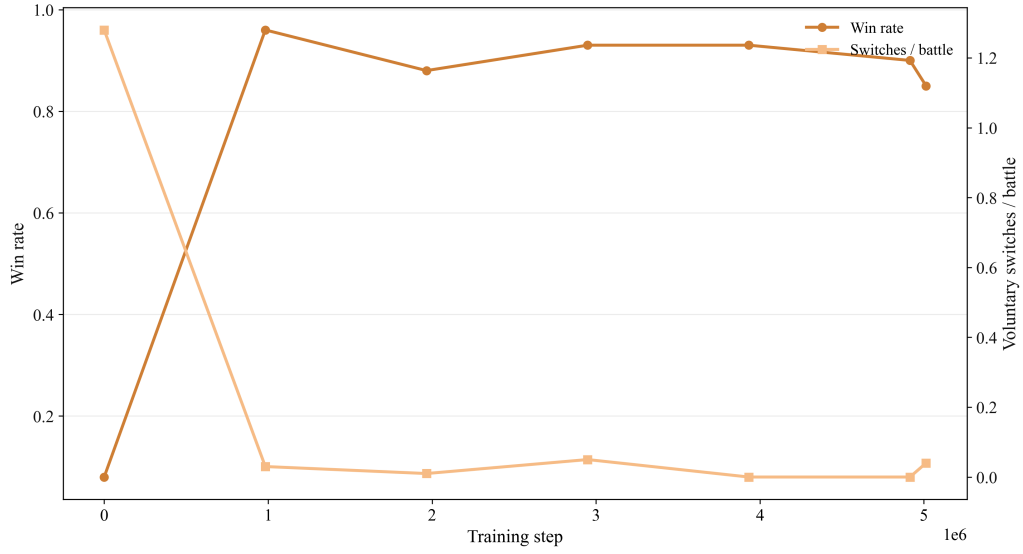


Figure 7: Self-play checkpoint progression analysis on training/G1629. This diagnostic uses 100-battle checkpoint evaluations.

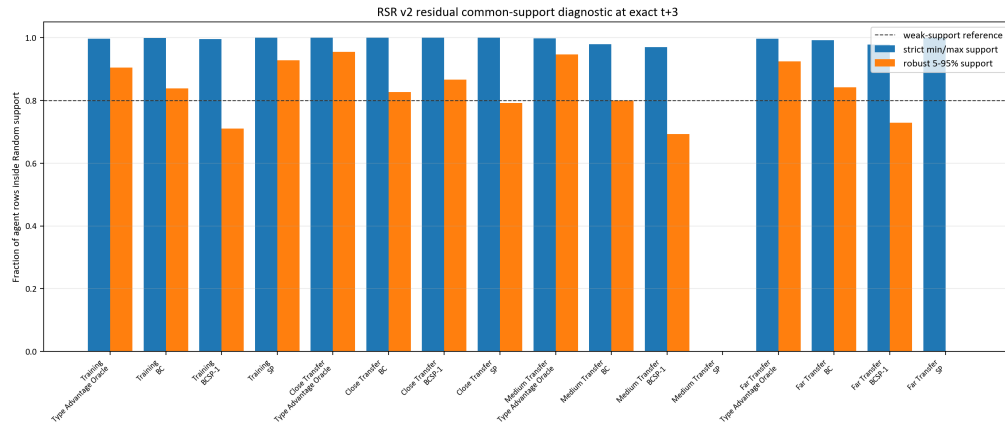


Figure 8: RSR common-support diagnostic at exact $t+3$. This figure is reported as appendix-only post-hoc support information.