

Article

Building an All-Shot Expected-Score Distribution Model from Real-Match Curling Boards: Shot-Wise Accuracy and Plausibility Analysis

Rintaro Chiba ¹, Yasumasa Tamura ^{2,3,*} , Shimpei Aihara ³  and Masahito Yamamoto ^{2,3} 

¹ Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan; chiba.rintaro.f5@elms.hokudai.ac.jp

² Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan; masahito@ist.hokudai.ac.jp

³ Department of Sport Science, Japan Institute of Sports Sciences, Tokyo 115-0056, Japan; shimpei.aihara@jpnssport.go.jp

* Correspondence: ytamura@ist.hokudai.ac.jp

Abstract

Curling is a strategic sport in which shot decisions involve both expected rewards and inherent risks; expected-score distributions (ESDs)—probability distributions over possible final scores—capture this uncertainty as a risk-aware strategic indicator. Although unified frameworks predicting ESDs across all shots have been proposed, their internal behavior and learned output characteristics have not been systematically examined. We construct an all-shot ESD prediction framework grounded entirely in real-match board configurations drawn from World Curling Federation championship events and conduct a shot-wise analysis along two complementary axes: *accuracy*, the model's reproduction of its training targets, and *plausibility*, the validity of those targets at aggregate and in-match scales. Accuracy degrades monotonically with shot number; only shot 16 admits comparison with an independent reference, while intermediate-shot labels are bootstrap rollouts of the next-shot model. An independent ground-truth probe at the first backward step (shot 15,500 boards $\times$ 95 contexts) bounds the consequence of this bootstrap structure: the resulting deviation is a small selection-induced hammer-underestimation bias common to both FiLM and concatenation chains, whose magnitude is more than an order of magnitude larger than the FiLM chain versus concatenation chain stage difference, and full multi-step verification beyond the first step is structurally out of reach. Within this scope the aggregate ESD shares the gross shape of the empirical end-score distribution and the qualitative hammer/non-hammer ordering, with a hammer-favorable offset attributed—after disentangling intent from execution on the shot-percentage 100% subset—to a label execution-noise envelope that is tighter than the realized play of top-tier matches. Within real matches the chain responds smoothly to each delivered stone and, at the directly validated terminal shot, assigns mean probability ≈ 0.71 – 0.72 to the realized end score under intended execution (against 0.28 for a marginal-frequency baseline), transferring from senior to junior populations. The framework is best read as an internally coherent chain anchored to a directly validated final-shot calibration, statistically plausible under intended execution with a clear boundary at execution failure. Two structural limitations remain: rare high-magnitude outcomes are scarce in real data and produce a heavy upper tail of accuracy errors, and a single fixed execution-noise envelope that is tighter than top-tier realized play accounts for the aggregate hammer-side offset and motivates recalibration against real-match execution statistics.



Academic Editor: Rui Araújo

Received: 31 May 2026

Revised: 30 June 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: curling; digital curling; expected-score distribution; shot-wise analysis; real-match data; artificial intelligence; transformer; FiLM

1. Introduction

Curling is a strategic sport with stochastic outcomes. Even with the same intended shot, variations in ice conditions and player skill lead to variability in results. Furthermore, physical factors such as stone-to-stone collisions and the effects of sweeping increase outcome uncertainty.

A curling match is divided into a fixed number of *ends* (typically ten), each end being an independent scoring round analogous to an inning in baseball; the team with the larger cumulative score after all the ends have been played wins the match. Within a single end, only the team with the stone closest to the button (the center of the house, the concentric target circles) scores, earning one point for each of its stones lying closer to the button than the opponent's nearest stone; at most one team scores per end. Because each team delivers eight stones in an end, an end yields between 0 and 8 points. Throughout this paper we express the end score from the hammer team's perspective as a signed value in $[-8, +8]$: positive denotes points scored by the hammer team, negative a steal by the opponent, and zero a blank end (no score), which lets the hammer team retain the last-stone advantage.

Each end comprises 16 shots, with the eight stones of each team delivered alternately. Strategic decision-making occurs at two levels. Within each end, tactical choices involve whether to place a stone in an advantageous position in the house, remove an opponent's stone, or position a guard to protect other stones. At the match level, score management is key. The team that scores in an end becomes the first-throw team in the next end, relinquishing the hammer, the last-stone advantage, to the opponent. Since the hammer confers a tactical advantage, strategies exist where teams intentionally choose to blank an end to retain it depending on the score situation. Thus, the combination of stochastic outcome variability and strategic complexity involving score differential, remaining ends, and hammer possession requires a probabilistic approach that considers the full distribution of possible outcomes and prioritizes final win probability rather than single-point estimates. Modeling outcomes as full distributions rather than as point estimates has become standard practice across sports analytics and game-playing AI: run-scoring distributions have been derived from Markov-chain models of baseball innings [1], score distributions in billiards from shot-outcome probabilities [2], and learned position evaluations for stochastic board games such as backgammon [3]; expected-goals (xG) models in football estimate the probability that a shot becomes a goal and underpin distributional descriptions of how a possession is likely to end [4]; and distributional reinforcement learning has shown that representing the full distribution of returns rather than only their mean improves both stability and policy quality [5,6].

Digital Curling [7] is a physics-based simulation platform that computationally reproduces curling rules and dynamics. It has been established as a standard experimental platform for tactical AI development in curling research.

Within this context, our AI system is based on expected-score distributions (ESDs) [8]—probability distributions over possible final scores (-8 to $+8$) at the end of an end (Figure 1). Unlike single-point estimates such as expected score, ESDs capture the full range of possible outcomes and their associated probabilities, enabling risk-aware strategic evaluation. By integrating ESDs with win-probability tables [9]—which provide empirical win rates from the perspective of the team currently holding the hammer, indexed by score differential and remaining ends—decision-making can prioritize final win probability rather than merely

maximizing expected score (the explicit construction of the per-score win-probability vector is given in Section 2.1.1).

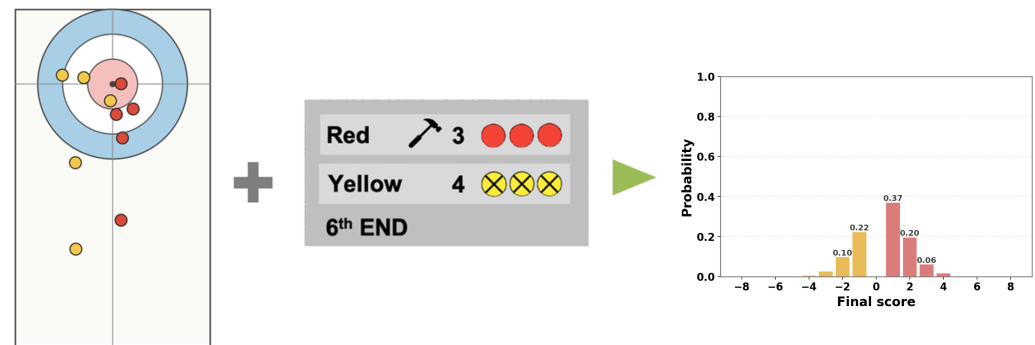


Figure 1. Example of an expected-score distribution (ESD): a stone configuration together with the game-state context (here the hammer team trailing by one point in the 6th end) is mapped to a probability distribution over possible final end scores from -8 to $+8$. The distribution captures both likely outcomes and tail risks, enabling strategic evaluation beyond single-point estimates.

This approach enables quantitative evaluation of tactical trade-offs and has demonstrated effectiveness in competitive tournaments. Despite this practical traction, the internal behavior and learned output characteristics of ESD prediction models have not been systematically examined: prior work has focused on driving down a scalar accuracy metric on a single shot, while the way the predictions evolve across shots, the shape of the predicted distributions, and the factors that dominate their variation are still largely uncharted. The present paper closes this gap by constructing an all-shot ESD model trained entirely on real-match board configurations from World Curling Federation championship events and then performing a comprehensive shot-wise analysis of it along two complementary axes.

ESD prediction in Digital Curling has evolved through two pressures. On the modeling side, the original framework [8] consisted of separate shot-specific models—which we will refer to throughout as the *shot- n model*, the model that takes the board configuration immediately before shot n and outputs the predicted ESD of the end—with no game-context input; pseudo-label generation was therefore forced to commit to a context-free objective (expected score), and the framework could learn no strategy beyond expected-score maximization. Yet the strategic value of a final score depends on its contribution to win probability, which varies with the score differential, the number of remaining ends, and hammer possession: the same board may therefore call for very different play in different game states. For example, the hammer team near a tied score often prefers either a multi-point score or a deliberate blank to retain the hammer, whereas a one-point lead late in the game can make a blank-or-better outcome preferable. On the simulator side, a separate trigger came from an update to Digital Curling’s underlying physics engine, which rendered the original models inconsistent with the new dynamics and made re-training a prerequisite for any subsequent analysis. Responding to both pressures, Iwasaki et al. first incorporated game-context conditioning—score differential and remaining ends—into a curling evaluation function [10] and subsequently proposed a transformer-based [11] ESD framework conditioning on a one-hot game context [12]; our previous work [13] then refined this design with win-probability-based weight vectors, coordinate normalization, and a rollout-based pseudo-label pipeline, significantly improving final-shot prediction accuracy over the predecessor framework and establishing the unified shot- n formulation we adopt here.

Two ingredients of that formulation were not, however, carried through to a single end-to-end all-shot framework. First, the terminal-shot labels in [13] were produced by

large-scale random simulation, decoupling the training corpus at the very shot that anchors the rollback chain from the real-match configuration distribution; an earlier attempt to train on real-game data [14] did anchor the terminal shot to real boards but covered only shot 16 and a restricted game-context grid, leaving the all-shot full-context regime open. Second, game-context conditioning was implemented by sequence concatenation, which multiplies the encoder cost by the number of game contexts at every (board, candidate) pair and makes the backward chain across all fifteen earlier shots computationally prohibitive. Closing both of these gaps end to end—and turning the resulting framework into the substrate for the systematic shot-wise analysis the literature has not previously supported—constitutes the contribution of this paper. Concretely:

- **Real-match grounding at every shot, end to end.** Every training board, at every shot from 2 to 16, is drawn from real-match configurations extracted from World Curling Federation championship events. In particular, the shot-16 labels are no longer produced from synthetic random boards but from the same real-match corpus as the earlier shots, and the rollback chain of [13] is extended to all sixteen shots of an end (Section 2.2). Beyond consistency with the current physics engine, this grounding preserves *human players' tactical distributions and configurational priors*: the boards on which the model is trained and evaluated are the boards that competitive players actually produce, and the framework's value as a downstream strategic indicator depends on the resulting predictions feeling *commensurate with players' own sense of the position*—a coherence that random-board training, by construction, cannot supply.
- **FiLM-based game-context conditioning for end-to-end tractability.** We replace the sequence-concatenation conditioning of [13] with feature-wise linear modulation (FiLM) [15], in which the game context is applied as an inexpensive element-wise affine modulation of a pooled board representation rather than as additional transformer tokens, yielding a $74.8\times$ speedup at the per-board all-context labeling step (each board paired with all $K = 95$ game contexts) and bringing the end-to-end rollback chain into a realistic wall-clock budget for the first time (Section 2.3 and Appendix A).
- **Systematic shot-wise analysis along two complementary axes, including a bounded ground-truth probe at the first backward step.** We dissect the resulting framework along the *accuracy* axis—the model's reproduction of its training targets, measured by the Kullback–Leibler divergence (Section 3.1)—and the *plausibility* axis—the validity of the targets themselves, examined at both aggregate and in-match scales (Sections 3.2, 3.3, 4.1 and 4.2), together with a bounded ground-truth probe at shot 15 (Section 4.3) that directly observes the existence, direction, and near-architecture-invariance of the backward-propagating bias the in-match validation alone cannot rule out. This directly closes the gap identified above: prior all-shot frameworks were never systematically probed for the internal behavior and learned output characteristics on which their strategic value depends.

2. Materials and Methods

2.1. Base Model Architecture

Building on our previous work [13], we employ a transformer-based architecture in which the game context is incorporated through feature-wise linear modulation (FiLM) [15]. The overall model architecture is illustrated in Figure 2. The model receives stone features together with an attention mask and a game-context vector. Each of the (up to) 15 stones is embedded into a 256-dimensional space and processed as a sequence of 15 tokens by a 4-layer transformer encoder (with 8 heads per layer), capturing inter-stone dependencies through self-attention. The resulting stone representations are mean-pooled over the valid stones into a single 256-dimensional board vector, which is then modulated by the game

context via FiLM. Finally, the modulated vector is passed through fully connected layers to output a 17-dimensional probability distribution over final scores ranging from -8 to $+8$. The accuracy advantage of this transformer-based encoder over simpler predictors was established in prior work [12,13]; we inherit that architecture and do not re-evaluate it here, focusing instead on its shot-wise behavior.

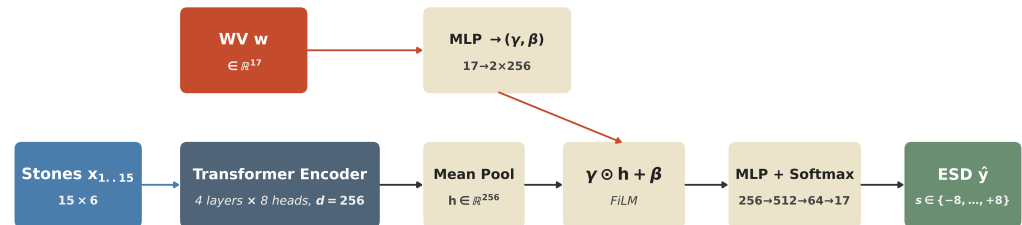


Figure 2. Architecture of the FiLM-conditioned transformer model for ESD prediction. The win-probability vector (WV) is mapped by a small MLP to FiLM parameters (γ, β) that modulate the pooled board representation from the transformer encoder; the modulated vector passes through a feed-forward head producing the predicted ESD.

2.1.1. Input Representation

Stone Features: Each of the (up to) 15 stones is encoded as a 6-dimensional vector consisting of its (x, y) coordinates, distance from the button (center of the house), an in-house flag, a team ID, and a score flag. The (x, y) coordinates are normalized to the range $[-1, 1]$ based on the valid curling play area. The score flag marks the stones that currently count toward the score, namely the in-house stones of the team whose nearest in-house stone is closest to the button that themselves lie closer to the button than the opponent’s nearest in-house stone.

Game Context: The game context is represented as a 17-dimensional vector $\mathbf{w} = (w_{-8}, \dots, w_{+8}) \in [0, 1]^{17}$, hereafter the *win-probability vector* (WV). We construct $\mathbf{w}$ from a precomputed win-probability table WT [9], compiled from a corpus of historical World Curling Federation match records as the relative frequency with which the team currently holding the hammer goes on to win the match from each pre-end state. Because every input to our model is expressed from the hammer team’s perspective (the sign convention introduced in Section 1), the current hammer state is fixed by construction and need not be carried as a separate index. The table is therefore indexed only by score differential d and remaining ends e ,

$$\text{WT}(d, e) = \Pr[\text{the team currently holding the hammer wins the match} \mid d, e], \quad (1)$$

and absorbs hammer rotation into its empirical structure (i.e., the post-end state with score s is summarized by $(d + s, e - 1)$ regardless of who holds the hammer in the next end). Given the current state (d, e) , the s -th element of $\mathbf{w}$ is the value of this table at the post-end state reached when this end is hypothetically settled with score s :

$$w_s = \text{WT}(d + s, e - 1), \quad s = -8, \dots, +8. \quad (2)$$

Each w_s is therefore a win probability in $[0, 1]$, *not* a probability mass over scores: $\mathbf{w}$ is not normalized to sum to one. This vector is used directly, without any nonlinear transformation, as the conditioning input from which the FiLM modulation parameters are generated.

Attention Mask: The attention mask is a 15-dimensional Boolean vector whose elements correspond to the stone inputs and mask out non-existent stones. Because the game

context conditions the model through FiLM rather than as a sequence token, no additional mask element is required.

2.1.2. Encoding and Processing

Each 6-dimensional stone vector is embedded into a 256-dimensional feature space and processed as a sequence of 15 tokens by a 4-layer transformer encoder [11,16], which captures dependencies among stones using multi-head self-attention mechanisms. Because the stones in a board carry no inherent ordering, the encoder is operated without positional encoding and its outputs are mean-pooled over the valid stones to form a single 256-dimensional board representation $\mathbf{h}$, yielding a permutation-invariant set encoder of the kind studied in the deep-sets literature [17,18]. The game context modulates this representation via FiLM: a multilayer perceptron maps the 17-dimensional weight vector to scale and shift parameters $(\gamma, \beta) \in \mathbb{R}^{256}$, and the representation is transformed as $\mathbf{h}' = \gamma \odot \mathbf{h} + \beta$.

2.1.3. Output

The FiLM-modulated board representation $\mathbf{h}'$ is passed through fully connected layers ($256 \rightarrow 512 \rightarrow 64 \rightarrow 17$) with rectified linear unit (ReLU) activations [19] and He initialization [20] and dropout regularization [21] between each pair of layers and transformed into a 17-dimensional probability distribution via a softmax function.

2.2. Dataset

Datasets are constructed for shots 2–16. The empty pre-shot-1 board is implicitly covered by applying the shot-2 model to the resulting post-shot-1 position, so 15 datasets suffice to span every position in an end. In contrast to our previous work [13], where the shot-16 dataset was produced by large-scale random simulation, all shots here share a common board source: real-match boards extracted from official competition records.

Board source. We use a curling match database constructed by applying the sheet-image stone-position extraction system of Suzumura et al. [22] to World Curling Federation championship events from 2022 to 2026, comprising 32 senior events (1752 games) across men's and women's World, European, Asia-Pacific, Olympic, and Olympic Qualification competitions. Junior events are excluded so that they remain available as a held-out distinct-skill validation population in the analysis of Section 4.2. From each game we extract the on-sheet stone configuration immediately before every shot and pair each configuration with the same $K = 95$ game contexts—the reachable combinations of score differential $d \in \{-5, -4, \dots, +5\}$ and remaining ends $e \in \{0, 1, \dots, 9\}$ —generating one ESD label per (board, context) pair.

Split and mirror augmentation. For each shot the boards are partitioned into training, validation, and test sets in an 80/10/10 ratio, stratified by stone count to keep the validation and test sets representative of the full stone-count distribution—which varies widely within a single shot number since take-outs remove stones and dense configurations are otherwise prone to under-representation in a small split. Horizontal mirror augmentation is then applied to the training split only: because the curling sheet is exactly symmetric about its center line, the mirror image of each board carries an identical ESD label by physical necessity [23], so mirroring the training partition roughly doubles its size while the validation and test sets remain mirror-disjoint and free of label leakage through this augmentation. The resulting per-shot dataset comprises approximately 24,700 training boards (mirrors included), 1500 validation boards, and 1500 mirror-disjoint test boards; each is expanded by the $K = 95$ game contexts, yielding about 2.35 million training samples per shot.

Label Generation

The ESD label for each board–context pair is the predicted-score distribution of the candidate shot that maximizes win probability. A naive procedure would, at every label, simulate many noisy releases of every candidate to absorb the shot-execution uncertainty; this multiplies the labeling cost by the noise sample size on every shot and is prohibitively expensive at our scale. We therefore decouple the two: the *noiseless* outcome of each candidate is obtained once per label, while the shot-execution uncertainty is folded in by a single convolution with a candidate-to-candidate transition-probability matrix that is constructed only once, from tens of thousands of perturbed releases per candidate, and reused across all subsequent labels. The transition matrix thereby amortizes the cost of the noisy simulations across the entire labeling pipeline. The noiseless outcome itself is obtained differently at the terminal and non-terminal shots. For shot 16, the noiseless outcome of each candidate is obtained by **direct simulation**: the deterministic stop position is simulated and its final score is read off. For shots 2–15, where play continues afterwards, the noiseless outcome is supplied instead by a **one-step rollout of the next-shot model** [24]: the candidate’s noiseless post-shot board is evaluated by the trained shot- $(N+1)$ model, yielding the candidate’s noiseless ESD. This rollout-based pseudo-labeling follows the pipeline of our previous work [13] and is structurally analogous to the bootstrapped value targets used in tabular dynamic programming and reinforcement learning [6]—a deliberate design decision rather than a convenience. Producing an independent ground-truth label at any intermediate shot by direct simulation would require recursively expanding the full candidate tree: roughly 7700 candidates per shot multiplied at every continuation, a cost that is multiplicative in shot depth and computationally infeasible even one step away from the terminus. The next-shot model is therefore the only practical source of intermediate labels, and the cost of bootstrap—error propagation along the chain, addressed at population scale in Section 4.2—is accepted in exchange for tractability. Physically implausible scores are masked out before normalization.

Shot-execution uncertainty. Every shot specifies a deterministic intended trajectory, but a real release deviates from this intent: the rock leaves the hand with a perturbed speed and release angle. We adopt the noise model of the Digital Curling simulator [7], which adds a Gaussian random vector to the initial velocity at the moment of release, with independent standard deviations $\sigma_v = 0.0076$ m/s on the shot speed and $\sigma_\theta = 0.0018$ rad on the release angle. Because the realized stop distribution of the stone is shaped non-trivially by stone–stone collisions and sweeping dynamics, it has no closed form, and any quantity that depends on it must be obtained by Monte Carlo simulation.

Candidate shot set and transition matrix. The space of physically reachable shots is continuous; for tractable enumeration we discretize it into a finite set of representative candidates spanning the play area (Figure 3). Two families are constructed: *draws*, targeting rest positions on a uniform grid over the play area with both release spins (3762 selectable draw targets within the play area; 4958 once marginal transition targets are included), and *take-outs*, aimed at a far-side target of $y = 100$ m to deliver *peel weight*—enough release speed to clear the released rock off the playing surface together with the target stone (144 selectable take-out lines; 2746 with neighboring lines). Peel weight is essential for blank-end strategy: without it the hammer team is forced to score even when retaining the hammer is tactically preferable. The full candidate set comprises 7704 shots (3906 selectable). To express the shot-execution uncertainty as transitions *within* this discrete set, we sample tens of thousands of perturbed trajectories from each candidate, snap each realized stop position to its nearest candidate, and tally the frequencies into a square stochastic matrix—the transition-probability matrix invoked in the rollout procedure above. The discretization

grid is fine enough that the associated approximation error is negligible relative to the trends reported below.

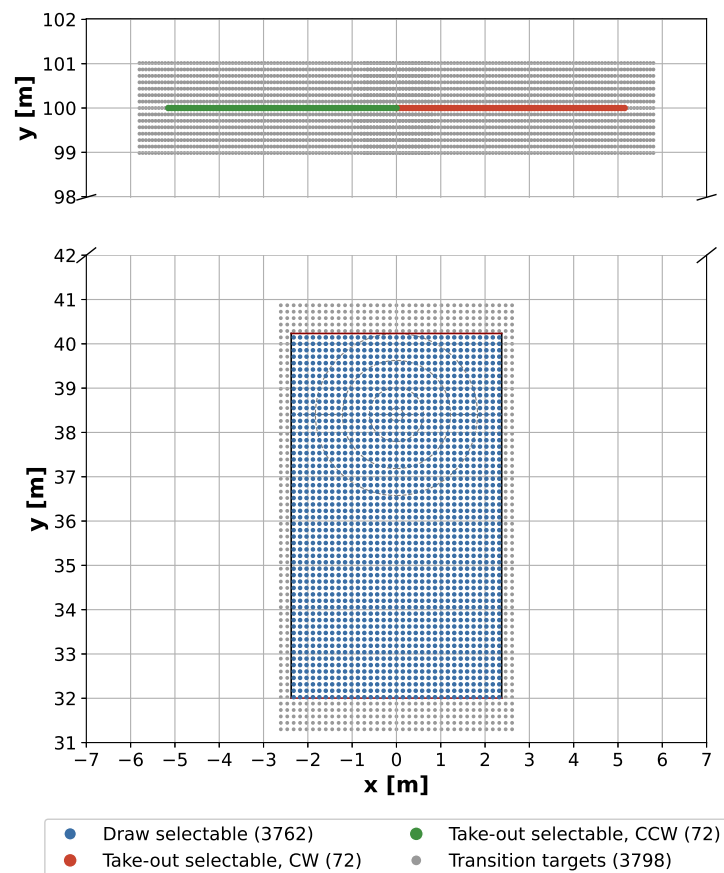


Figure 3. Spatial layout of the discretized candidate shot set. Top: take-out candidates aimed at a far-side $y = 100$ m target (peel weight), with selectable clockwise (red) and counter-clockwise (green) release lines and transition targets (gray) covering stops just outside the selectable lines. Bottom: draw candidates targeting rest positions on a uniform grid over the play area (blue) with their transition margin (gray). The house outline (dashed concentric circles near $y = 38.4$ m) and the sheet/hog/back-line geometry are drawn for orientation; the wavy break omits the unused longitudinal region between the take-out target altitude and the play area.

Selection criterion and win-probability saturation. Among the candidates, the target shot is selected to maximize win probability. A naive criterion is pure expected-win-probability (EW) maximization, $\arg \max_c \sum_s P_c(s) w_s$, where P_c is candidate c 's ESD and w_s is the same win probability introduced in Equation (2). However, the win-probability table saturates at the extremes: when one team holds a large lead, a whole range of favorable scores all yield $w_s \approx 1$ (a certain win) and symmetrically $w_s \approx 0$ for large deficits. In such contexts EW maximization is degenerate: candidates whose outcomes fall within the saturated region have nearly identical EW, so the choice is governed by numerically negligible differences in w_s and becomes effectively arbitrary.

A natural-sounding remedy would be to break this tie by mixing in the expected score over *all* outcomes, but doing so would corrupt the criterion precisely where it should be most discerning. In the non-saturated closely contested situations where EW genuinely separates candidates, an unconditional expected-score term would systematically reward taking extra points even when doing so is strategically counterproductive. The most important casualty is the *blank end* (final score 0), which lets the team holding the hammer retain the last-stone advantage for the next end: a blank can be the strictly preferred

outcome in many tactical configurations, yet any global expected-score contribution would always rate it below an alternative that scores even a single point. Pure expected-score maximization, or a globally mixed criterion of the same kind, therefore fails as a label objective: it cannot represent the score-versus-blank trade-offs that constitute much of curling strategy.

To preserve win-probability supremacy in the non-saturated regime while still breaking ties in the saturated regime, we adopt the *sat_outcomes* criterion, which augments EW with an expected-score tiebreaker restricted to the saturated outcomes:

$$c^* = \arg \max_c \left[(1 - \alpha) \text{EW}_c + \alpha \frac{1}{8} \sum_{s=-8}^8 P_c(s) s m_s \right], \quad m_s = \mathbb{I}[w_s < \epsilon \text{ or } w_s > 1 - \epsilon], \quad (3)$$

with $\epsilon = 0.01$ and $\alpha = 0.05$; both are kept deliberately small. Because the score term is masked to the saturated bins only, in non-saturated contexts, $m_s \approx 0$ everywhere that EW differs meaningfully across candidates, so the criterion reduces to pure EW maximization and leaves blank-versus-score and similar tight-game trade-offs entirely to the win-probability table. In saturated contexts, where EW is uninformative, the tiebreaker selects the candidate with the highest expected score among the win-equivalent (or loss-equivalent) outcomes, recovering the intuitive strategy of scoring as many points as possible once the win is secured. Supplementary Section S1 verifies empirically that this design preserves blank-versus-score arbitration relative to a strict-lexicographic baseline, in contrast to a naive globally mixed criterion.

2.3. Training Strategy

Our trained framework consists of 15 shot models, one for each shot $n = 2, \dots, 16$. Following the naming convention introduced in Section 1, the shot- n model takes as input the board configuration immediately before shot n is played and outputs the predicted ESD of the end. The 15 models therefore cover every pre-shot board state except the empty pre-shot-1 board, which contains no stone information and which we cover implicitly by applying the shot-2 model to the resulting post-shot-1 position; this is why 15 models are sufficient to support a shot-wise analysis spanning all 16 shots of an end. All 15 shot models are implemented in PyTorch 2.1 [25] and NumPy [26] and trained independently, each from random initialization, by minimizing the Kullback–Leibler (KL) divergence [27] between the predicted and target ESDs. We use the AdamW optimizer [28,29] with a learning rate of 3×10^{-4} , a batch size of 2048, and at most 100 epochs with early stopping [30,31] on the validation KL (patience of 5 epochs). Predicting a discrete probability distribution over outcomes rather than a single point estimate follows the distributional-prediction perspective established in distributional reinforcement learning [5].

The models are trained in backward order, from shot 16 down to shot 2. Because the pseudo-labels for shot N are produced by rolling out the already-trained shot- $(N+1)$ model (as described in the label-generation procedure above), each model must be trained before the shot that precedes it. Knowledge about future play therefore propagates backward through the rollout-generated labels at training time.

The cost of running this rollback chain end to end depends sensitively on how the game context is mixed with the stone tokens. Pseudo-label generation for each board requires evaluating roughly 7700 simulated candidate shots against $K = 95$ game contexts. Under the concatenation architecture of our previous work [13], in which the game context is appended as an additional token to the stone sequence, each (candidate, context) pair triggers an independent full encoder pass, and the cost on an NVIDIA RTX 4080 is approximately 10.5 s per board; pseudo-label generation for a single shot ($\sim 15,400$ boards) then takes about 45 h, and the full backward chain over shots 15–2 amounts to roughly

28 days. The FiLM architecture adopted here instead runs the encoder once per (board, candidate) pair and applies the context as an inexpensive element-wise affine modulation of the pooled board representation independently of K . In a controlled benchmark on the same hardware, with identical preprocessing and encoder configuration and with mixed precision and graph compilation disabled for both architectures, this brings pseudo-label generation to ~ 140 ms per board—a $74.8\times$ speedup—and shrinks the full backward chain to roughly 9 h. The margin is structural: per-pass encoder cost is comparable (~ 85 ms for FiLM versus ~ 110 ms for concatenation), and the speedup arises because concatenation must repeat the encoder for each of the K contexts while FiLM encodes once; it therefore widens with K .

To quantify the accuracy cost of this switch, we compared FiLM with a matched concatenation model at shot 16—the only shot with ground-truth labels—over five seeds on the same mirror-disjoint test set; the concatenation reference is in the same architectural family as the shot-16 model of our previous study [13] and currently provides the highest shot-16 accuracy on this task. FiLM’s mean test KL is modestly higher (0.266 ± 0.009 versus 0.230 ± 0.007 ; paired mean $\Delta\text{KL} = +0.036$, 95% CI $[0.029, 0.043]$ by board-level bootstrap), with the gap confined to the heavy upper tail of dense high-magnitude boards (paired median $\Delta\text{KL} = +0.005$; same regime as Section 4.4). By Pinsker’s inequality [32] $\text{TVD} \leq \sqrt{\Delta\text{KL}/2}$, this is at most a 5.0% probability-mass shift on a typical board (95% CI $[4.5\%, 5.4\%]$) and 13.4% on average ($[12.0\%, 14.7\%]$); the two architectures correspondingly agree on $\sim 81\%$ of modal scores and differ by 0.04 in integrated $E[s]$. We accept this small typical-board shift as the price of a tractable backward chain and treat the resulting final-shot accuracy as suitable for practical tactical-decision support, with the high-magnitude tail retained as a deliberate target of future work. Because all shot-2–15 labels are generated by the resulting FiLM model, a corresponding small per-shot error may in principle propagate backward through the chain; we cannot quantify that propagation independently, but the population-level coherence checks of Section 4.2 provide no evidence of accumulated drift along realized match trajectories. Per-seed results, the matched-condition timing breakdown, the derivation of the TVD bounds quoted above, and an additional comparison of the absolute shot-16 KL against a configuration-blind statistical baseline are given in Appendix A.

2.4. Analysis Methodology

We analyze the system along three axes. The first axis, *prediction accuracy*, evaluates the trained model against its training targets. The remaining two axes, *distributional characteristics* and *dominant factors*, instead characterize the training targets themselves—the labels the model is trained to predict—computed on the labels rather than on model outputs so that they describe the targets the model learns rather than its own prediction error. The caveat noted in Section 2.2 applies throughout: for shots 2–15 the targets are bootstrap rollouts of the next-shot model rather than independent ground truth, so the trends below describe the learned targets.

2.4.1. Prediction Accuracy Analysis

To quantify the prediction accuracy for each shot, we use the Kullback–Leibler (KL) divergence, consistent with the training objective. It measures the divergence of the predicted distribution $\hat{p}$ from the target distribution p (the label):

$$\text{KL}(p \parallel \hat{p}) = \frac{1}{N} \sum_{i=1}^N \sum_{s=-8}^8 p_i(s) \log \frac{p_i(s)}{\hat{p}_i(s)} \quad (4)$$

where N is the number of test samples and s represents the final score.

2.4.2. Distributional Characteristics Analysis

We characterize the training targets from three perspectives. Throughout, p_i denotes the i -th label.

Average Score Probabilities: We visualize the average label probability of each score for each shot as the mean ESD $\bar{p}$:

$$\bar{p}(s) = \frac{1}{N} \sum_{i=1}^N p_i(s), \quad s \in \{-8, \dots, 8\} \quad (5)$$

This captures the overall likelihood of each score for that shot.

Uncertainty: We quantify the uncertainty of the label ESDs from two perspectives:

- **Entropy of Mean ESD:** The entropy of the mean ESD $\bar{p}$ over all samples, indicating how broad the range of possible scores is for that shot overall:

$$H(\bar{p}) = - \sum_{s=-8}^8 \bar{p}(s) \log_2 \bar{p}(s) \quad (6)$$

- **Average Sample Entropy:** The entropy of each individual label ESD, averaged over all samples:

$$\text{Avg } H = \frac{1}{N} \sum_{i=1}^N H(p_i) \quad (7)$$

This indicates the average outcome uncertainty of a single situation.

Distributional Diversity: To quantify how much the shape of the label ESDs varies across samples, we compute the total probability variance across all score positions:

$$\text{Total Probability Variance} = \sum_{s=-8}^8 \frac{1}{N} \sum_{i=1}^N (p_i(s) - \bar{p}(s))^2 \quad (8)$$

Higher values indicate that the shape of the label ESDs varies substantially across samples. The difference $H(\bar{p}) - \text{Avg } H$ also provides an information-theoretic measure of this diversity.

2.4.3. Dominant-Factor Analysis

To quantify the relative contributions of stone configuration and game situation, we perform a variance decomposition of the expected score computed from the label ESDs.

Due to the dataset structure, each stone configuration is paired with 95 different game situations. Leveraging this structure, we decompose the total variance of expected scores into two components:

- **Between-configuration variance:** The variance of expected scores across different stone configurations. This reflects the influence of stone configuration.
- **Within-configuration variance:** The variance of expected scores when different game situations are applied to the same stone configuration. This reflects the influence of game situation.

The relative influence of game situation is quantified by the variance ratio:

$$\text{Game Situation Influence} = \frac{\text{Within-configuration variance}}{\text{Total variance}} \quad (9)$$

A higher ratio indicates that game situation is the dominant factor.

3. Results

3.1. Accuracy: Reproduction of the Training Targets

Reading guide for this subsection. The Kullback–Leibler divergence reported below measures reproduction of the training targets, not calibration against an independent physical reference. For shots 2–15 those targets are themselves bootstrap rollouts of the next-shot model (Section 2.2), so the KL on those shots quantifies fidelity to a learned objective; only at shot 16, whose labels come from direct simulation, does the KL measure agreement with an externally generated reference. This distinction is structural and is repeatedly invoked in the discussion below; the independent validity of the labels themselves is examined separately in the plausibility analysis (Sections 3.2, 3.3, 4.1 and 4.2).

We measure *accuracy* as how closely each shot-*n* model reproduces its training targets, quantified by the Kullback–Leibler divergence between the predicted ESD and the label ESD on the held-out test set.

Table 1 reports the per-shot mean and median KL divergence between the model prediction and the label. Accuracy degrades monotonically as the shot number increases: the mean KL rises from 3.1×10^{-4} at shot 2 to 2.7×10^{-1} at shot 16, a roughly 900-fold (nearly three orders of magnitude) increase, with no oscillation between odd and even shots. This monotonic trend mirrors the growth of the prediction target’s own dispersion (Section 3.2): later shots admit a wider and more configuration-dependent range of outcomes that are intrinsically harder to predict. The KL therefore conflates a genuine increase in task difficulty with the growth in the targets’ intrinsic dispersion, and the two cannot be fully separated here.

Shot 16 is distinctive. Although its *mean* KL is the largest, its *median* (5.1×10^{-2}) is in fact lower than that of shot 15 (6.7×10^{-2}), indicating a heavy-tailed error distribution: the great majority of final-shot boards are still predicted well, while a small set of difficult boards inflate the mean. The structure of these hard cases is revisited in the Discussion (Section 4.4).

Table 1. Per-shot prediction accuracy on the test set, measured by the KL divergence between the model output and the target label ($\approx 1.5 \times 10^5$ test samples per shot, i.e., test boards $\times$ 95 contexts; mirror augmentation is applied to the training split only). Accuracy degrades monotonically with shot number, and the error distribution becomes heavy-tailed: at shot 16 the median (5.1×10^{-2}) is below that of shot 15, indicating a small number of densely populated boards that are very hard to predict.

Shot	Mean	Median	Shot	Mean	Median
2	3.1×10^{-4}	7.6×10^{-5}	10	1.3×10^{-2}	5.9×10^{-3}
3	8.0×10^{-4}	1.9×10^{-4}	11	1.8×10^{-2}	7.2×10^{-3}
4	1.4×10^{-3}	3.7×10^{-4}	12	2.4×10^{-2}	1.1×10^{-2}
5	2.9×10^{-3}	7.8×10^{-4}	13	3.4×10^{-2}	1.7×10^{-2}
6	4.4×10^{-3}	1.4×10^{-3}	14	6.7×10^{-2}	3.9×10^{-2}
7	7.3×10^{-3}	2.0×10^{-3}	15	1.4×10^{-1}	6.7×10^{-2}
8	7.9×10^{-3}	2.9×10^{-3}	16	2.7×10^{-1}	5.1×10^{-2}
9	1.0×10^{-2}	3.5×10^{-3}			

For interpretation: KL values at or below the median (5×10^{-2}) correspond to predictions whose modal score coincides with the label’s and whose bars nearly overlap, whereas the heavy upper tail corresponds to predictions whose modal score disagrees with the label’s—a regime that is examined further in Section 4.

3.2. Distributional Characteristics of the Training Targets

We now characterize the training targets—the labels the model is trained to predict—to assess whether they behave sensibly. All the quantities in this subsection are computed on the labels (recall Section 2.2: for shots 2–15 these are bootstrap rollouts of the next-shot model).

3.2.1. Mean Expected-Score Distributions

Figure 4 shows the mean ESD $\bar{p}$ for every shot in the end. In contrast to a point estimate, the mean ESD retains substantial spread at *every* shot: the entropy $H(\bar{p})$ stays in a narrow band, ranging from 2.28 bits at shot 13 to 2.42 bits at shot 6, with a typical value of ≈ 2.35 bits across the chain. Every distribution is centered near a final score of +1, with most of the mass on 0, +1, and +2 and lighter tails toward larger margins and toward negative scores. The shape is already broad at the very first shots, as expected when many stones remain to be thrown and the outcome is far from settled. We emphasize that a broad *mean* ESD does not imply broad *individual* targets: as the entropy decomposition below shows, at later shots the mean is broad chiefly because it averages many sharp but mutually differing distributions, not because any single target is diffuse.

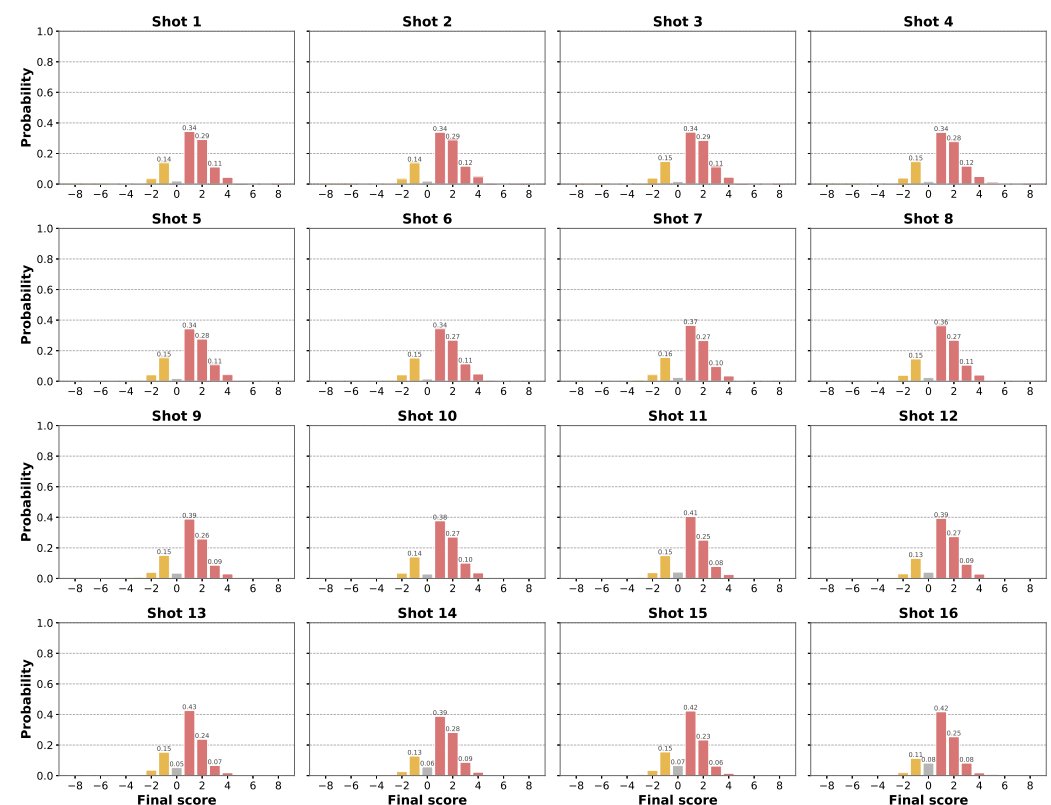


Figure 4. Mean ESD of the training targets for all 16 shots. The distribution is broad and centered near +1 at every shot ($H(\bar{p}) \approx 2.3$ bits, $E[s] \approx +1$), with probability mass migrating from the central modes toward the bars of larger absolute score as the shot number decreases. Bar colors encode the score sign: hammer-side scoring (rose), blank end (gray), and opponent steal (amber).

A second pattern is visible when comparing early and late shots side by side. As one traces the chain backwards from shot 16 to shot 1, mass progressively migrates from the three central modes (0, +1, +2) toward the bars of larger absolute score: high hammer-side outcomes ($s \geq +3$) and opponent steals ($s \leq -1$) all carry slightly more probability at early shots than at late ones. This is exactly what curling intuition predicts. At the start of an end almost any outcome remains reachable, so the targets must allocate non-negligible

mass to big ends and to steals; as stones accumulate the realized configuration commits to one regime, and the targets concentrate on the three commonplace results. The mean ESD therefore reflects a chain in which early-end uncertainty is expressed as broader tails over extreme scores rather than as a uniform inflation of the entropy.

3.2.2. Entropy and Diversity

Although the mean ESD stays broad throughout, the *reasons* for that breadth change completely over the course of an end. Figure 5 decomposes it using the entropy of the mean ESD $H(\bar{p})$, the average per-sample entropy $\text{Avg } H$, and their difference.

The average per-sample entropy $\text{Avg } H$ decreases monotonically, from 2.19 bits at shot 2 to 0.61 bits at shot 16. This matches the natural expectation that the outcome of a single situation is highly uncertain early in the end—many stones remain and almost any score is still reachable—and progressively resolves toward a determined result as stones are played.

Conversely, the diversity $H(\bar{p}) - \text{Avg } H$ grows monotonically, from 0.07 bits at shot 2 to 1.66 bits at shot 16. This too is expected: as play proceeds, the set of reachable game states branches combinatorially, so the population of boards—and hence of their ESDs—becomes increasingly diverse. Early on, all boards look alike and their ESDs are broad but mutually similar ($\text{Avg } H \approx H(\bar{p})$); late in the end each board's ESD is sharp but the boards differ greatly, so a broad mean is assembled from many distinct sharp distributions.

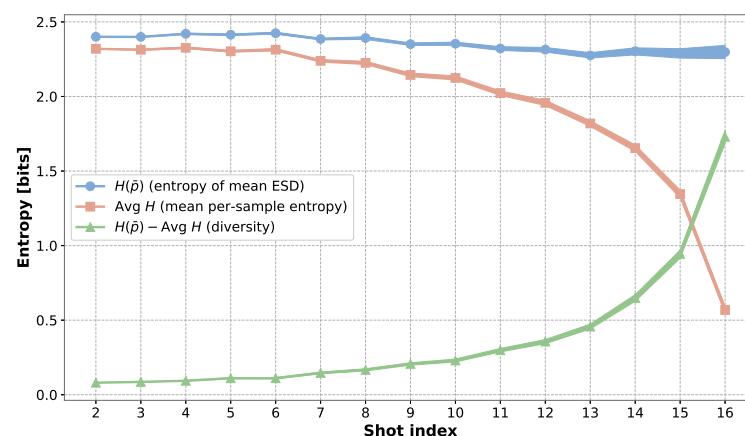


Figure 5. Entropy of the target ESD across shots: entropy of the mean ESD $H(\bar{p})$ (nearly constant), average per-sample entropy $\text{Avg } H$ (monotonically decreasing as outcomes resolve), and their difference (diversity, monotonically increasing).

3.2.3. Expected Score and Hammer Asymmetry

Figure 6 shows the expected end score $E[s]$ of the mean ESD. It stays close to +1 throughout, consistent with the well-known one-point character of a typical curling end. A clear asymmetry between the two teams is superimposed on this level: even shots, taken by the team holding the hammer, have $E[s] > 1$ (rising to +1.12 at shot 16), whereas odd shots, taken by the team without the hammer, have $E[s] < 1$ (down to +0.85 at shot 15). The gap widens toward the end of the end, reflecting the growing realized value of last-stone advantage.

3.3. Dominant Factors in the Training Targets

To identify what governs the ESD at each shot, we decompose the variance of the expected score (computed from the labels) into a within-configuration component (variation across the 95 game contexts applied to a fixed board) and a between-configuration component (variation across boards). The game-situation variance ratio—the within-configuration

fraction of the total—measures how much the game situation, as opposed to the stone configuration, drives the outcome.

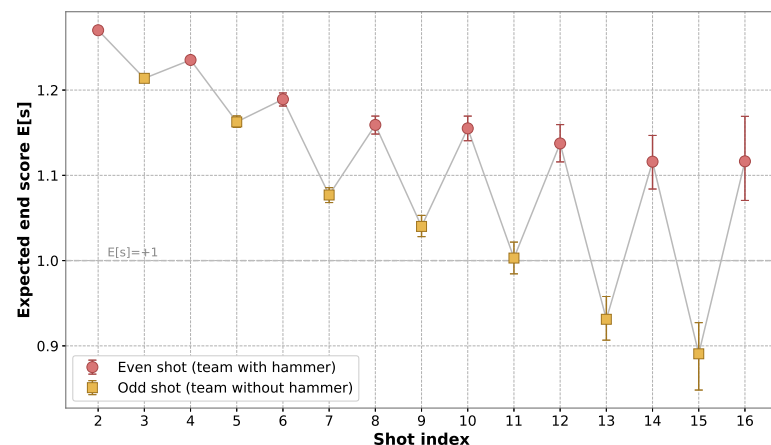


Figure 6. Expected end score $E[s]$ of the target ESD. The value remains near +1 for all shots; even shots (with hammer) lie above +1 and odd shots (without hammer) below, with the gap widening toward shot 16.

Figure 7 shows this ratio decreasing monotonically with shot number, from 99.6% at shot 2 to 12.3% at shot 16. Early in the end, when few stones are on the sheet, the eventual score is determined almost entirely by the game situation (score differential and remaining ends) rather than by the sparse configuration, and the ratio is near 100%. As stones accumulate, the concrete configuration increasingly dictates the outcome, and, by the final shot, the stone configuration is dominant. A faint even/odd ripple is visible in mid-game—the team with the hammer retaining slightly more situation dependence—but it is minor compared with the overall monotonic transition. These ratios are conditioned on the fixed set of 95 game contexts used throughout—a designed grid of feasible score-differential and remaining-end combinations rather than a sample from the empirical context distribution—so the within-context variance, and hence the absolute ratio, depends on the spread of this set; the qualitative shift from situation- to configuration-dominance, however, is expected to hold for any reasonable context set.

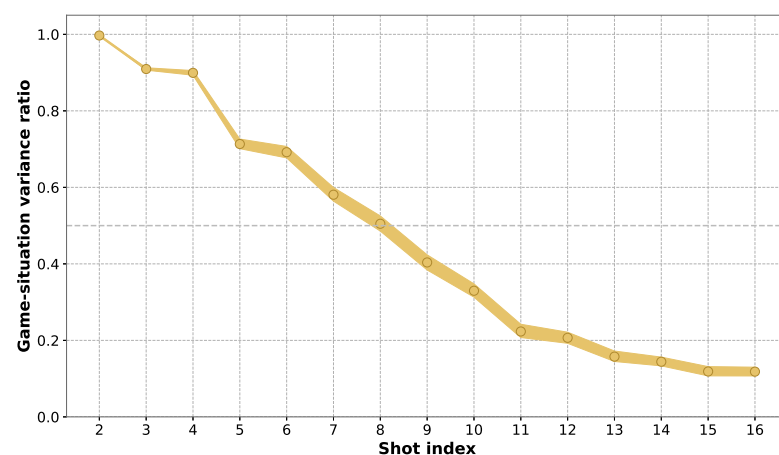


Figure 7. Game-situation variance ratio across shots, computed from the labels. The ratio falls monotonically from $\approx 100\%$ to $\approx 12\%$, indicating a smooth transition of dominance from the game situation to the stone configuration.

4. Discussion

4.1. Plausibility of the Aggregate Distribution

We first verify the ESD as a population-level quantity by inspecting how the targets are distributed over the dataset and how that aggregate compares with the empirical end-score distribution of real matches; the underlying physics is guaranteed by the simulator, so what we verify here is statistical plausibility rather than physical correctness.

The statistical evolution of the targets across shots matches a priori expectations about how an end unfolds: per-sample entropy decreases monotonically as outcomes resolve, distributional diversity grows monotonically as reachable game states branch combinatorially, the dominant factor shifts smoothly from the game situation in the sparse early end to the stone configuration in the dense late end, and the mean ESD's extreme-tail mass—large hammer-side scores and steals—progressively contracts into the commonplace one- and two-point modes (Figures 4, 5 and 7). None of these trends is imposed by construction, and their agreement with intuition supports the ESD as a meaningful indicator.

The mean ESD also recovers the familiar hammer-side advantage (Figure 6): even shots (with hammer) lie above +1 and odd shots (without hammer) below, with the gap widening toward shot 16. This asymmetry is consistent with the input encoding—the per-score win-probability vector together with the hammer-side sign convention (Equation (2)) carries the hammer state directly into the conditioning input—so it should be read as confirmation that the encoding is propagated coherently through the model rather than as independent evidence that the predictor has discovered the advantage from data. The asymmetry does not appear in the prediction-accuracy or entropy metrics, corroborating this reading.

More directly, the aggregate ESD recovers the overall shape of outcomes actually observed in real matches. Figure 8 overlays the context-matched mean ESD—each shot-16 board is assigned the single `sat_outcomes`-selected label corresponding to its own match-context win-probability vector $w_{(d, e_{\text{left}}, \text{false})}$, then averaged over the senior corpus ($n = 14,165$)—on the empirical distribution of end scores in our match database. The two share the same hammer-side mode and decay at a comparable rate on either side. Quantitatively, the total variation distance between the two distributions is 0.192 (95% bootstrap CI [0.172, 0.211]); the Earth Mover's distance [33], which respects the ordinal structure of the score axis, is 0.321 ([0.292, 0.352]); and the Kullback–Leibler divergence [27] from the empirical distribution to the context-matched mean ESD is 0.044 ([0.039, 0.052]). The mean-score offset is hammer-favorable: the context-matched target sits at $E[s] = +1.128$ against an empirical $+0.807$, a $+0.321$ gap.

The offset itself merits a closer look. The labels model the score distribution under optimal play with a fixed simulator-noise envelope, whereas the empirical aggregate reflects realized play in which the players' intent (not necessarily optimal) and the realized execution uncertainty are entangled. Interpreting the offset therefore requires disentangling intent from execution. We do so on a subset where intent can be read independently. For each shot-16 board in the senior corpus we recompute the `sat_outcomes`-selected stone with the context-matched win-probability vector and record both its noise-free deterministic outcome (the “label intent”) and the realized end score from the WCF record. Restricting to the shot % = 100 subset ($n = 9828$)—where the called shot and the executed shot coincide, so the realized end score reflects the players' intent without execution noise—label intent and the realized end-score distribution coincide to total variation distance 0.024 and Kullback–Leibler divergence 0.006 ($E[s] = +1.473$ versus $+1.478$; Figure 9). On this subset where intent can be read directly, the players' intent matches the `sat_outcomes` criterion that builds the labels. We proceed under the natural assumption that this intent alignment carries over to the full population: execution failures (shot % < 100) are taken to

reflect noise-side departures from an otherwise sat_outcomes-consistent intent rather than a different intent altogether.

Intent alignment having been established, the remaining axis—the labels’ execution-noise envelope—must account for the aggregate $+0.321$ offset. The hammer-favorable direction of the offset diagnoses which way the envelope mismatch points. A broader-than-realized envelope would scatter outcomes across the score range, including substantial mass on the failure tail, pulling the labels’ aggregate $E[s]$ below the realized $+0.807$; instead the labels’ aggregate $E[s] = +1.128$ stays close to the high-score intent. The offset is therefore the signature of an envelope that is *tighter* than the realized play of top-tier (WCF senior) matches as a population: the labels posit an idealized execution precision that exceeds championship-level players’ realized dispersion across boards, under-representing the execution-failure tail those matches in fact realized. A second distinct contribution to the hammer-favorable direction would arise from a winner’s-curse mechanism on the labels’ own selection: the sat_outcomes arg max over thousands of noisy candidate continuations preferentially picks those whose noise has, by chance, pushed the hammer-side win probability upward (the structural mirror image of the shot-15 arg min bias documented in Section 4.3). Each candidate’s noise estimate is built from a large Monte Carlo sample, so the per-candidate noise variance is expected to be small and this contribution presumably minor relative to the envelope-tightness one; we cannot, however, rule it out as a residual term.

As a further consistency check on the rollback chain itself, the per-board expected win probability $EW = \sum_s P(s) w_s$ (with w_s as in Equation (2)), averaged over boards within each shot, stays close to 0.55 across all shots with bootstrap 95% confidence intervals [34] tight to a fraction of one percent. The chain neither drifts nor accumulates over- or under-confidence as the rollout progresses. We caveat that EW is a per-shot aggregate and therefore averages out shot-to-shot fluctuations of opposing sign within and across trajectories; backward error propagation in the chain cannot be ruled out by this check alone, and we revisit the within-trajectory question in Section 4.2.

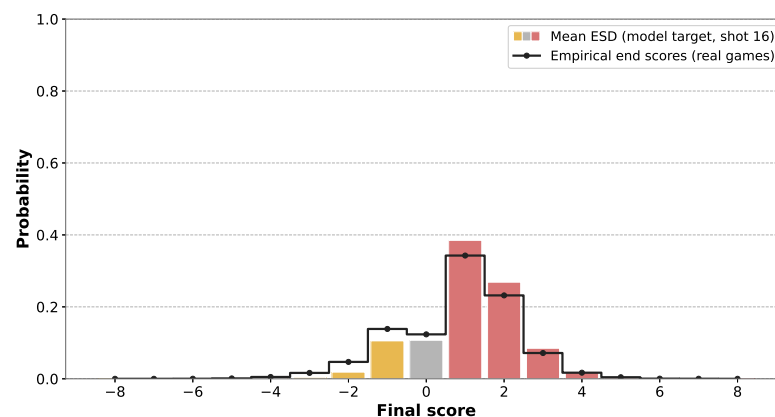


Figure 8. Context-matched mean ESD (the prediction target at the final shot with score-colored bars) overlaid on the empirical distribution of end scores in real matches (black step). Bin-wise mass is normalized to sum to 1. The two share the same hammer-side mode, with the labels slightly more concentrated near the high-score intent.

Aggregate-level agreement, however, does not by itself guarantee that the chain behaves coherently *within* a real match. A sequence of predictions that is internally consistent on average can still respond abruptly or implausibly to particular shot sequences. We therefore turn to plausibility along actual match trajectories.

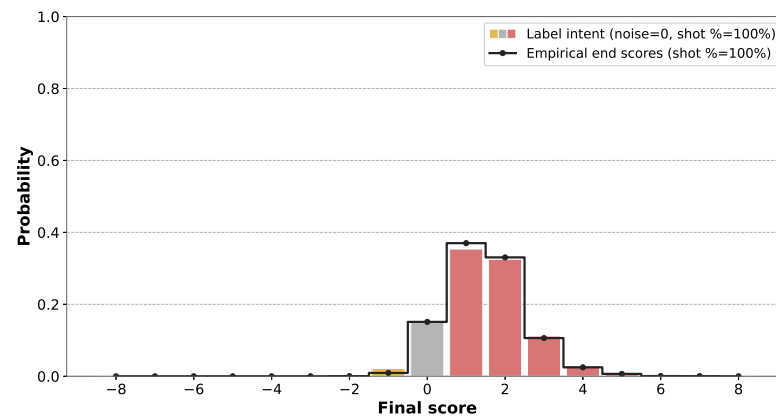


Figure 9. Label intent (noise-free outcome of the sat_outcomes-selected stone, score-colored bars) versus the empirical end-score distribution on the shot % = 100 subset (black step). The two nearly coincide (TVD 0.024), confirming intent alignment.

4.2. Plausibility Within Real Matches

Applying the chain to the actual shot sequence of a recorded end produces, at every pre-shot board, a predicted ESD; the trajectory of these distributions is the natural object on which to ask two questions: (i) does the chain respond *coherently* to each delivered stone, neither freezing nor jumping anomalously; and (ii) when the intended final shot is in fact realized, does the predicted ESD recover the realized end score? We address (i) through frame-by-frame qualitative examples and a population-level smoothness statistic and (ii) through a direct stratification by execution skill.

Qualitative spine: frame-by-frame on held-out matches.

Junior events were excluded from training and therefore form a natural held-out test set, useful here because we can illustrate the chain's response on data it has never seen. We applied the full chain to every shot of every end of a complete ten-end match from the 2022 World Junior Curling Championships (men's bracket, Germany vs Canada, final score 8–5 in Germany's favor). Across the ten ends, the predicted $E[s]$ at the final shot lay within ± 0.1 of the realized end score in six cases and agreed in sign in eight. Routine hammer-side scoring ends—such as end 2, played from a tied score—behave as expected: the chain concentrates mass on +1 and +2 throughout and aligns with the realized +2. Two more demanding ends from the same match expose how the chain handles unusual dynamics.

Figure 10 shows the first such end (end 5, played from a four-point hammer-side deficit) in which the hammer-holding team ultimately scored two—a sharp reversal of the cumulative trend. The early shots of the end carry $E[s]$ near +1; as the configuration develops favorably for the hammer team, the predicted distribution progressively shifts toward +2, tracking the realized outcome despite the unfamiliar scoreline and junior shot tendencies.

Figure 11 shows the second (end 4, played from a two-point hammer-side deficit) that was lost by the hammer team for two points (a −2 steal). The predicted ESD remains centered near +1 throughout yet places non-trivial mass on negative outcomes—approximately 12% on −2 alone—rather than committing entirely to the typical hammer-side modes. The actual reversal to −2 arose from a contingent execution failure: the final hammer-side stone, a hit-and-roll, is recorded with a shot success of 0% in the WCF database, whereas the preceding opponent shot was executed cleanly. As the static board configurations themselves remained consistent with a one- or two-point hammer-side outcome, the predicted ESDs are interpretable as the model's evaluation of the realized configurations rather than a failure

to anticipate the steal as a configurational property; the chain still allocates measurable probability to the steal outcome that ultimately materialized.

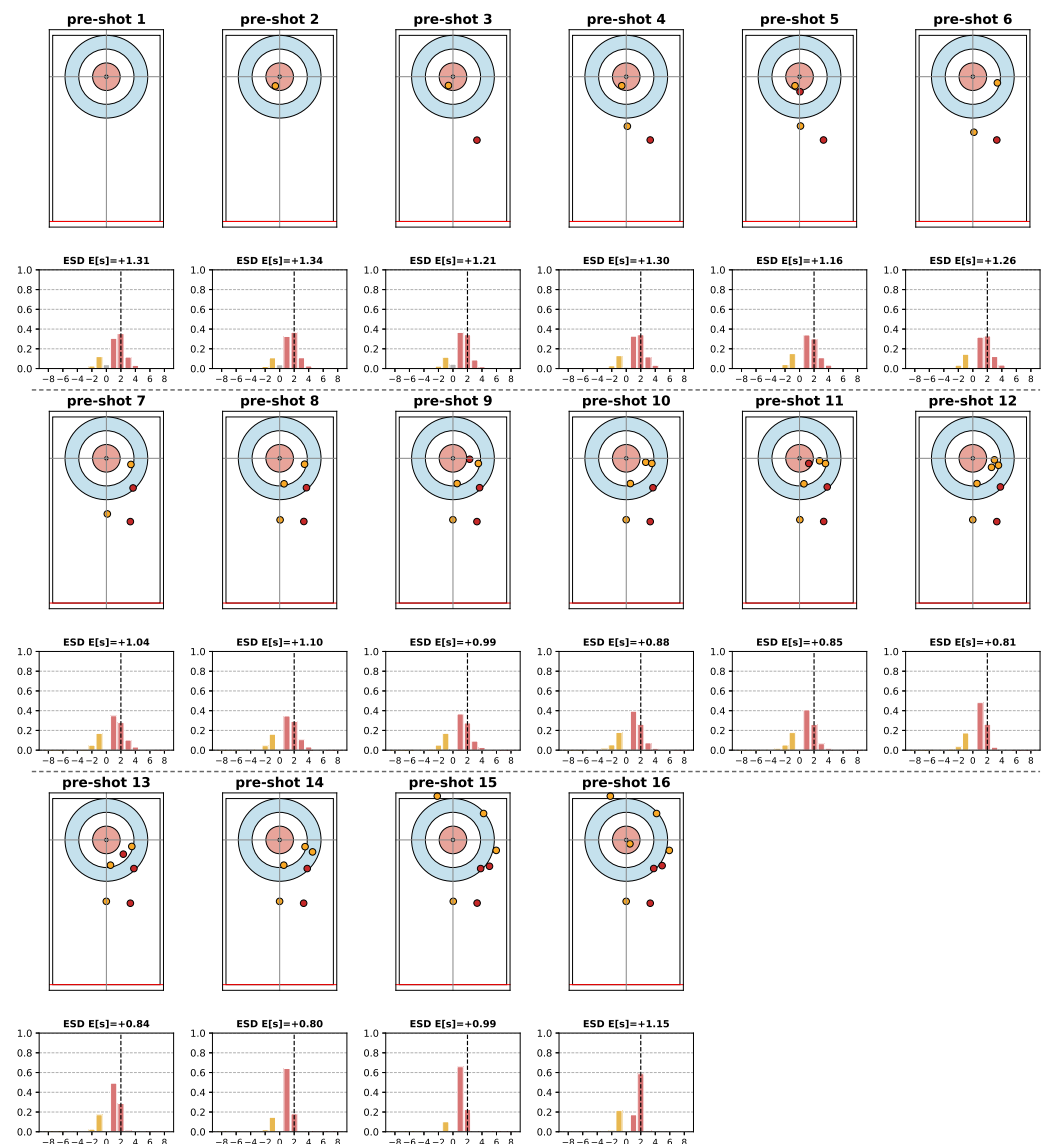


Figure 10. End 5, played from a four-point hammer-side deficit; the hammer team ultimately scored +2. Each column corresponds to one shot, with the board state immediately before the shot (**top**) and the predicted ESD (**bottom**); the black dashed line marks the realized final score. Despite the unfamiliar scoreline, the predicted ESD progressively shifts toward the realized +2 outcome as the end unfolds.

The two ends together suggest a pair of population-level claims. End 5 exemplifies the chain responding smoothly and proportionately to stones that progressively commit the configuration; end 4 exemplifies the limiting case in which the actual outcome is dominated by execution rather than configuration, and the steal is realized despite a board that the model rightly reads as a one- or two-point hammer-side end. We now verify each of these readings beyond the anecdotal scale.

Quantitative pillar I: trajectory smoothness.

For every recorded end we trace the expected end score $m_k = E_{\text{ESD}_k}[s]$ through the actual shot sequence and tabulate the shot-to-shot increment $\Delta_k = m_{k+1} - m_k$. If the chain responds proportionately to each delivered stone and never jumps anomalously, $\text{Var}(\Delta_k)$

should grow smoothly with shot number and contain no within-curve discontinuities. Figure 12 shows this variance on a sample of senior matches ($n_{\text{ends}} = 965$) and on the junior held-out population ($n_{\text{ends}} = 1538$). Both curves rise smoothly by roughly two orders of magnitude from shot 2 to shot 15, with no jumps, and senior and junior remain nearly indistinguishable at every shot. The first observation confirms the trajectory smoothness inferred from end 5; the co-trajectory of the two populations indicates that the dynamics carry over to a distinct skill level—a quantitative population statement rather than an anecdote.

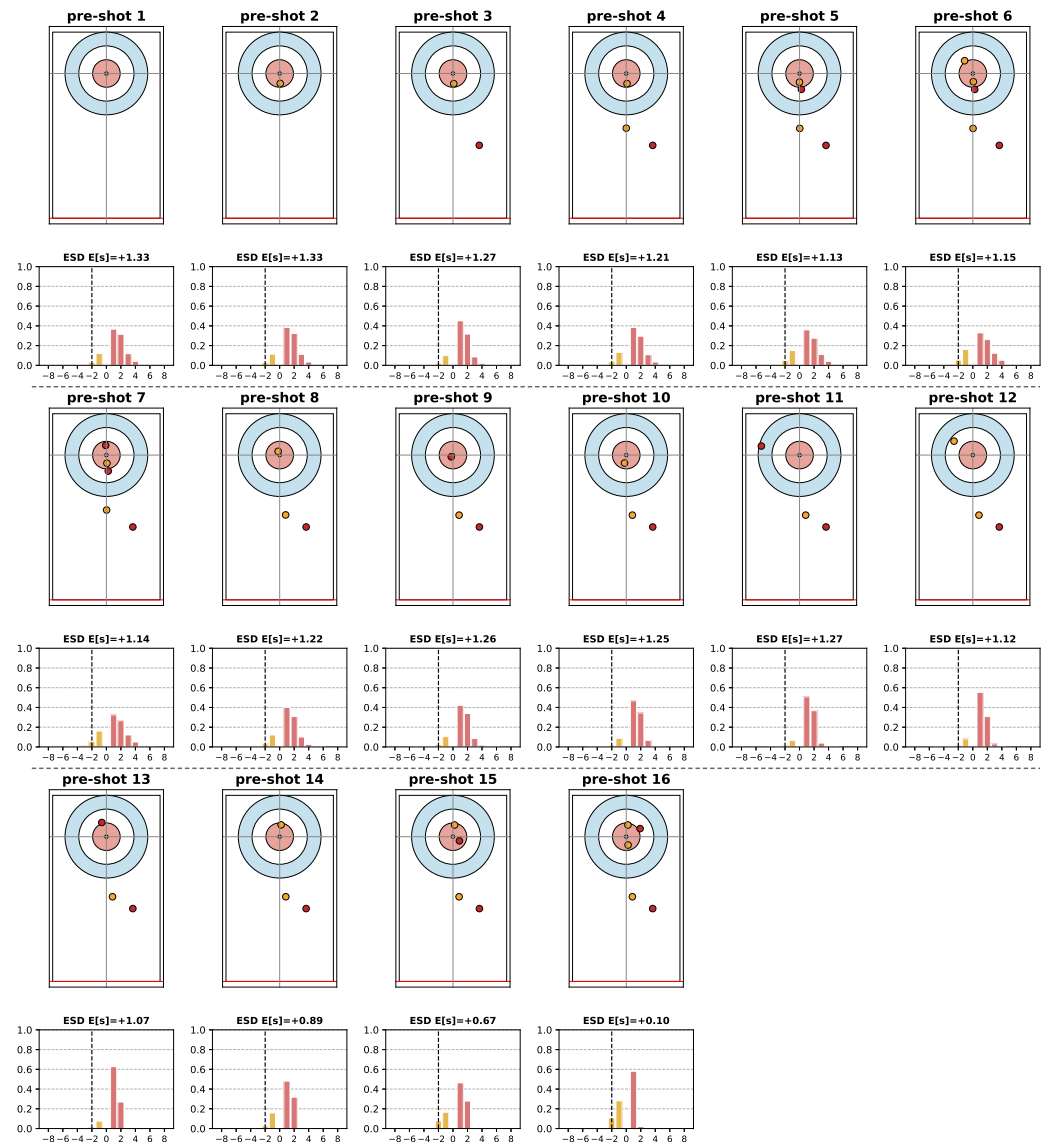


Figure 11. End 4, played from a two-point hammer-side deficit; the hammer team conceded a -2 steal (same layout as Figure 10). The predicted ESD stays centered near $+1$ but assigns measurable mass to steals ($\approx 12\%$ on -2). The reversal traces to a contingent execution failure on the final hammer-side stone (a hit-and-roll recorded at 0% success in the WCF database) rather than to a configurational property the static model should have recognized.

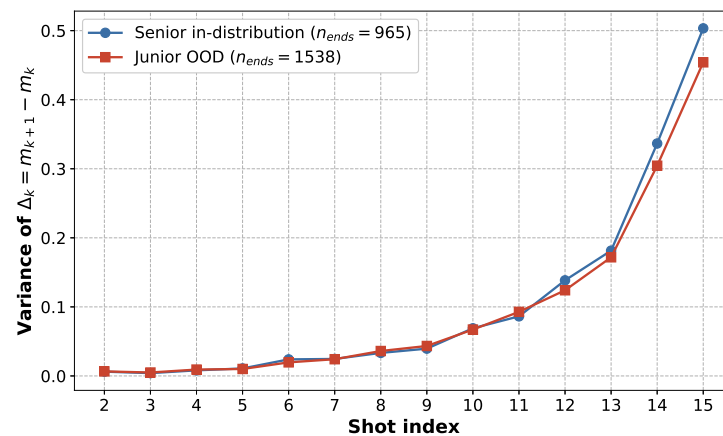


Figure 12. Variance of the per-shot increment $\Delta_k = m_{k+1} - m_k$ in the expected end score along the actual shot sequence of every recorded end, plotted against shot index k . Senior in-distribution ($n_{\text{ends}} = 965$, blue) and junior out-of-distribution ($n_{\text{ends}} = 1538$, red) populations are overlaid. The two curves are nearly identical at every shot, smoothly rising by roughly two orders of magnitude from shot 2 to shot 15, with no discontinuities within either curve.

Quantitative pillar II: execution conditioning.

The end 4 reading—that the steal is realized despite a configurationally favorable board as a contingent execution failure—is a single-end claim. To check whether the underlying relationship holds at population scale, we exploit a property of the predicted ESD itself: a candidate’s ESD describes the score distribution *under (near-)intended execution*, so, when the intended shot is in fact realized, the ESD should be a high-fidelity predictor of the realized end score. The World Curling Federation publishes, for every delivered stone, a *shot percentage*: a stone-by-stone judgment of how close the realized stone landed to the called shot, scored on a five-point ordinal scale (0, 25, 50, 75, or 100 in quarter increments, where 100 denotes a fully intended execution). We restrict attention to the final stone (shot 16) so that the percentage clearly reflects the very execution the predicted ESD assumes, stratify shot-16 ESDs by that value, and compute, on each subset, the probability the predicted ESD assigns to the realized end score. Figure 13 shows the result for $n = 4000$ senior ends (left, in-distribution) and $n = 1464$ junior ends (right, out-of-distribution). On both populations the assigned probability rises monotonically with shot percentage and reaches ≈ 0.71 (senior) and ≈ 0.72 (junior) at full success ($=100$); the predicted ESD therefore does recover the realized score distribution when the intended shot is realized, directly validating the optimal-play assumption that defines the prediction. To set this number in context, two information-free baselines computed on the same $= 100$ subset assign far less mass to the realized score: a constant-modal predictor (always predicting the empirical mode, $+1$) yields $P(\text{actual}) = 0.37$ (senior) and 0.36 (junior) and a marginal-frequency predictor (always predicting the empirical end-score distribution of the subset) yields 0.28 for both. The ESD therefore exceeds the better baseline by roughly a factor of two, indicating that the assigned mass reflects board-specific information rather than the unconditional ease of the subset. The same subset is, however, somewhat enriched in easier boards: the mean predicted-ESD entropy at $=100$ is modestly lower than the population average (0.67 versus 0.80 bits in the senior subset; 0.67 versus 0.82 in the junior), so part of the elevation in $P(\text{actual})$ reflects this enrichment, but the residual gain over the marginal baseline ($\approx +0.43$) is too large to be explained by it alone. At low success (<25) the same probability falls to ≈ 0.20 – 0.22 —the expected mirror image as the ESD is not designed to track execution failures and recovers the end 4 reading at population scale. Here the difficulty proxy points in the opposite direction (mean entropy $1.11/1.12$ bits, higher than the population mean), so the low-success degradation conflates execution-

driven and configuration-driven contributions; cleanly separating the two would require finer covariates on board difficulty than the per-shot percentage alone can provide. The senior and junior curves remain co-trajectorial at every value of the shot percentage, so this per-percentage calibration generalizes beyond the training distribution to a distinct skill level.

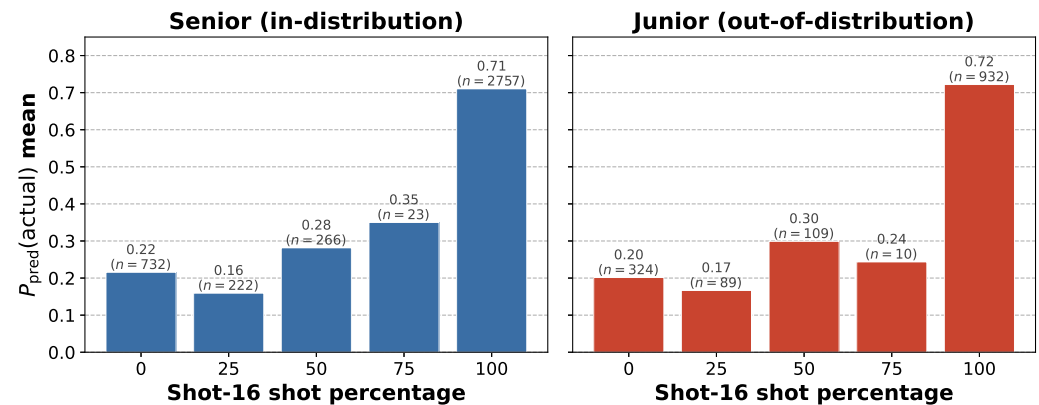


Figure 13. Direct verification of the optimal-play assumption that defines the ESD, stratified by the WCF shot percentage of the executed final stone (five-point ordinal scale; 100 = fully intended execution). Left: senior in-distribution ($n = 4000$ ends, blue). Right: junior out-of-distribution ($n = 1464$ ends, red). Each bar shows the mean probability that the predicted ESD assigns to the realized end score. At full success (=100), the ESD assigns ≈ 0.71 – 0.72 of its mass to the realized score, confirming that the predicted distribution describes the outcome under (near-)intended execution; the monotone increase with shot percentage and the co-trajectory of senior and junior together extend this verification to a distinct skill level.

Scope of the in-match validation.

Taken together, the two pillars anchor the chain to a directly validated terminus at shot 16 (pillar II, full success) and demonstrate smooth anomaly-free propagation backward through to shot 2 (pillar I). We emphasize the precise scope of this validation: the intermediate ESDs are checked for coherence with the validated terminus, not individually calibrated against an independent ground truth. The latter is a structural boundary of the approach rather than a remediable gap. Producing an independent ground-truth label at any intermediate shot would require recursively expanding the candidate tree at every continuation (Section 2.2), which is computationally infeasible. Pillar I rules out within-trajectory jumps and discontinuities but does not rule out a smooth slowly varying bias that propagates backward and only vanishes at the terminus; pillar II calibrates the ESD against an independent reference only at shot 16, where execution stratification is meaningful. Within these bounds the chain is internally coherent on real-match trajectories, terminates in a directly validated final-shot calibration, and behaves indistinguishably on a distinct skill population.

4.3. Direct Ground-Truth Probe of the First Backward Step

This subsection adds a single-step *direct* check to the chain-coherence-based validation of Section 4.2, complementing rather than replacing it. The scope paragraph of Section 4.2 states that producing an independent ground-truth label at an intermediate shot would require recursively expanding the candidate tree at every continuation, a cost that grows multiplicatively in shot depth and is computationally infeasible across the full chain. A bounded restriction of this validation is, however, tractable: at the first backward step (shot 15), with the board sample limited to several hundred boards, an independent ground truth obtained by direct physical simulation lies within reach across all $K = 95$ game contexts.

We use this bounded probe to examine the existence and direction of the smooth backward-propagating bias that Section 4.2 could not rule out by trajectory smoothness alone.

For the shot-15 ground truth we generate, on each pre-shot-15 board and for every context, the full sat_outcomes-selected label is obtained by direct simulation of the candidate shots, convolving with the same transition matrix used at shot 16 and selecting with the non-hammer-side arg min. The evaluation is carried out on $N = 500$ boards and all $K = 95$ win-probability vectors, yielding 47,500 independent (board, context) pairs. For comparison, the same boards and contexts are also labeled by two backward-rollout chains: (i) the FiLM model used throughout the paper (B_aftersplit) and (ii) the concatenation baseline (concat_raw_s42) referenced in Table A1 (Appendix A).

Table 2 reports the three-way comparison. The two chain labels sit at comparable distances from the ground truth, with a median total variation distance (TVD) of 0.150 (FiLM) and 0.140 (concatenation), modal-score agreement of 75.5% and 75.8%, respectively, and a common hammer-side underestimation bias of approximately -0.087 on $E[s]$. The FiLM-versus-concatenation distance is consistently smaller than either chain's distance to the ground truth (TVD median 0.101, modal agreement 80.1%, and $E[s]$ bias $+0.002$), so the architecture contribution at the chain-label stage is below the selection-induced floor common to both chains. The hammer-underestimation bias direction is shared by 95.8% of contexts under the FiLM chain and by 98.9% under the concatenation chain, well above what would be expected by chance and consistent with the selection mechanism analyzed below.

Table 2. Three-way comparison of shot-15 labels on $N = 500$ boards across all $K = 95$ contexts (47,500 pairs per row). GT denotes the directly simulated ground truth; FiLM chain and Concat chain denote backward-rollout chains using the FiLM and concatenation shot-16 models, respectively.

Pair	TVD Median	TVD p90	Modal Agreement	$E[s]$ Bias
GT vs. FiLM chain	0.150	0.424	75.5%	-0.087
GT vs. Concat chain	0.140	0.402	75.8%	-0.085
FiLM chain vs. Concat chain	0.101	0.328	80.1%	$+0.002$

Comparing the per-(board, context) KL of the shot-16 model output against the shot-16 ground truth with that of the shot-15 FiLM chain label against the shot-15 ground truth (Table 3; the two columns measure different referents and strict numerical matching is not the claim) shows an asymmetric per-percentile transformation: the chain label is moderately higher than the shot-16 model at the median (roughly a factor of two) but markedly closer at p90 and beyond (the p90 figure compresses by roughly 30% and the mean by roughly 8%). Four mechanisms superpose to compress the chain-stage error relative to a hypothetical single-stage propagation of the shot-16 model's error: (i) selection bleaching of the random-noise component, (ii) decreasing inter-board target diversity as the chain runs backward, (iii) re-application of the physical-reachability score mask, and (iv) sign-balancing across the alternating arg max / arg min chain. Concretely:

- (i) **Selection bleaching.** The chain's arg min over thousands of candidate continuations averages down large shot-16 errors that appear in random sign orientations across candidates, depressing the upper-percentile chain-label error below shot-16 single-stage error.
- (ii) **Decreasing inter-board diversity backwards from shot 16.** Figure 5 shows that the inter-board diversity of the targets grows monotonically with shot number, so shot-15 labels are mutually more similar across boards than shot-16 labels, and any single label sits less often far from the typical one.

- (iii) **Re-application of the score mask** (Section 2.2). The chain pipeline zeroes out physically unreachable scores on each shot-15 post-candidate board and renormalizes, trimming probability mass that the shot-16 model had placed on physically implausible scores and removing the most extreme part of the heavy tail.
- (iv) **Alternating-selection sign-balancing**. The chain composes a hammer-side arg max (in the shot-16 sub-label produced on each shot-15 candidate's post-board) with the shot-15 non-hammer-side arg min. Any systematic component carried by the trained shot-16 model is amplified in opposite directions at the two stages, so the consecutive opposite-sign selections do not coherently amplify it as a single-direction arg chain would; the residual random component is processed by mechanism (i) above.

The residual median increase is a small selection-induced bias on typical-case outputs, examined in the next paragraph.

Table 3. Per-(board, context) KL of the shot-16 model output against the shot-16 ground truth versus the shot-15 FiLM chain label against the shot-15 ground truth. The two columns measure different referents: the shot-16 column quantifies model-prediction error on the $1540 \times 95 = 146,300$ test (board, context) pairs, while the shot-15 column quantifies chain-label deviation on the $500 \times 95 = 47,500$ probe of this section. These are diagnostic per-percentile values for trend comparison; absolute matching across columns is not the claim and would not be meaningful given the differing referents.

Percentile	Shot 16: Model Output vs. GT-16	Shot 15: Chain Label vs. GT-15
p50 (median)	0.051	0.105
p75	0.288	0.269
p90	0.823	0.589
p99	2.579	2.524
Mean	0.272	0.250

The common hammer-side underestimation of -0.09 on $E[s]$ in both chains arises from a winner's-curse mechanism. Shot 15 selects the candidate that minimizes the hammer-side win probability (arg min) over thousands of noisy shot-16 sub-labels supplied by a trained model; this preferentially picks continuations where the noise has, by chance, pushed the hammer-side win probability downward, biasing the chain label in the hammer-underestimation direction. The structural mirror image of the same mechanism was already invoked at shot 16, as the residual winner's-curse contribution noted in Section 4.1 (arg max over candidates with simulator-noise envelope), where the envelope-tightness effect dominates the hammer-favorable offset of Figure 8. The two together identify the selection-noise interaction as a property of the labeling criterion rather than of any particular architecture.

Two consequences follow. First, the architecture contribution to chain-label fidelity is structurally sub-dominant to this selection-induced floor: the FiLM-versus-concatenation difference of 0.006 on $E[s]$ is an order of magnitude below the common bias of 0.09 , and the paired ΔKL of $+0.036$ that distinguishes the two architectures at the shot-16 prediction stage (Table A1) is therefore not visible at the chain-label stage. Second, the probe is bounded (a single context, a hundred-board sample, and a single backward step). Extending it further is rapidly cost-prohibitive: this shot-15 probe at $N = 100$ boards on a single context already requires several hours of single-machine compute, and an analogous shot-14 probe would multiply by the ~ 7700 -fold candidate-tree expansion to roughly twenty days per board, placing a statistically meaningful sample ($N \geq 20$) beyond a year of single-machine wall clock. The bias-propagation question must therefore be addressed by structural rather than empirical means: as long as the chain alternates arg max and arg min, the same bleaching and the same small bias should continue to apply. This bounded probe nonetheless

directly observes the existence and direction of the smooth backward-propagating bias that Section 4.2 could not rule out by trajectory smoothness alone.

4.4. Limitations

Two structural limitations follow from the analysis above.

Lim A: under-representation of rare high-magnitude outcomes in the accuracy tail. The heavy upper tail of the final-shot error (Table 1) is concentrated on densely populated boards from which the team with the hammer can, by playing an optimal final stone, secure a large score. On such boards the model recognizes the high-magnitude regime but tends to shift its modal score down by one or two points relative to the label. Each individual mismatch is mild, but analogous shifts accumulated over the rare configurations that admit high scores contribute appreciably to the heavy upper tail of the final-shot KL. This is a shortfall in reproducing the labels on rare configurations (the accuracy axis), distinct from the outcome calibration under the intended execution established in Figure 13. We attribute it to the composition of the dataset: because every board is taken from real World Curling Federation matches, ends that finish with a large score—or, symmetrically, a large steal—are rare. From the hammer team’s perspective the empirical distribution of end scores in these matches is concentrated on the four most common outcomes—a one-point steal through a two-point score—and its tails decay steeply: beyond a two-point score the frequency falls by roughly a factor of three to four for each additional point (+3, +4, and +5 ends occur in about 7%, 1.7%, and 0.5% of cases), and steals decay at a comparable rate. The model accordingly encounters very few such examples during training and systematically under-represents the extremes of the score range, which is exactly where the large-error boards concentrate.

Lim B: a single skill-independent execution-noise envelope. Whereas the boards and the win-probability vectors that condition the model are drawn from real-match data, execution uncertainty is not: by construction the framework assumes a single execution-noise model, with fixed Gaussian standard deviations $\sigma_v = 0.0076$ m/s and $\sigma_\theta = 0.0018$ rad (Section 2.2), applied uniformly to senior and junior populations alike. This is the one component of the framework that the real-data grounding has not yet reached. Two consequences follow. First, as analyzed in Section 4.1, the labels’ execution-noise envelope is tighter than the realized play of top-tier WCF senior matches as a population, producing the +0.321 aggregate hammer-side offset of Figure 8 as its signature (with a presumably minor residual winner’s-curse contribution from arg max over noisy candidates). Second, while the percentage *outcome* calibration transfers cleanly from senior to junior populations (Figure 13, indistinguishable curves at every shot percentage)—as expected when intended-execution physics is essentially skill-independent within the WCF-grade resolution—neither the noise envelope itself nor the skill-dependent marginal distribution of execution quality has been independently verified, and behavior at a genuinely different skill level (e.g., recreational or club play) is not tested by the present held-out data. We note that the bounded ground-truth probe of Section 4.3 cannot detect this envelope mismatch by construction: the shot-15 simulation ground truth inherits the same execution-noise envelope as the chain label, so its first-order contribution cancels in the comparison and only the selection-noise residue surfaces. The throwing-side bias signature of this mismatch is documented at shot 16 in Section 4.1 and verified in mirror at shot 15 in Section 4.3.

4.5. Future Work

The two structural limitations above motivate a 1:1 set of future directions. FW-A, addressing Lim A, is to extend coverage of rare high-magnitude outcomes through carefully validated data augmentation, pursued with the constraint that augmentation

must improve rare–extreme coverage without compromising the in-distribution realism that anchors the present approach. FW-B addresses Lim B in light of the offset analysis in Section 4.1, where the labels’ fixed execution-noise envelope is shown to under-represent the failure tail of top-tier realized play and produce the aggregate hammer-side offset as a consequence. Two prongs follow: (i) recalibrating the execution-noise envelope (σ_v, σ_θ)—or, more generally, the conditional distribution of execution outcomes given an intended shot—against the realized-play statistics of top-tier (WCF senior) matches themselves in the same data-grounded spirit that anchors the boards and the win-probability vectors to real play (this recalibration is itself non-trivial since realized play entangles execution and intent and disentangling the two is part of the work) and (ii) validating the chain at a genuinely different skill level (e.g., recreational or club play, where shot percentages differ substantially from the championship distribution sampled here) to characterize how the envelope changes with skill. Finally, the dominant-factor decomposition presented here suggests a route to practical decision support, and validation against expert judgment is planned.

5. Conclusions

Working from real-match boards rather than from large-scale random simulation, we constructed an all-shot expected-score distribution (ESD) prediction framework covering every position in an end and dissected its behavior shot by shot. Monotone evolution with shot number is the unifying signature on both axes: within-end uncertainty resolves, the diversity of distributions across boards grows, and the dominant factor shifts from the game situation to the stone configuration as an end resolves (Sections 3.2 and 3.3); the per-shot KL degrades in step, tracking the growing dispersion of those targets.

Along the two axes the analysis evaluates, the verdict is asymmetric and precise. Accuracy is bounded by construction: only shot 16 admits comparison with an independent reference, and the intermediate-shot ESDs are checked for coherence with this validated terminus (Sections 2.2 and 4.2). We complement this with an independent ground-truth probe at the first backward step (shot 15, single context, hundred-board sample, Section 4.3), which shows the resulting deviation to be a small selection-induced bias common to both FiLM and concatenation chains rather than an architecture-specific drift; full multi-step verification beyond the first step remains structurally out of reach. Plausibility is the stronger conclusion within this scope. The aggregate ESD shares the gross shape of the empirical end-score distribution of real matches and the qualitative hammer/non-hammer ordering consistent with the input encoding—with a non-trivial systematic offset attributed, after disentangling intent from execution on the shot-percentage 100% subset, to a label execution-noise envelope that is tighter than the realized play of top-tier matches—and the chain stays internally coherent across the rollback. The bounded ground-truth probe of Section 4.3 corroborates this attribution: the simulation ground truth inherits the same execution-noise envelope as the chain label, so the envelope contribution cancels in that comparison and the residual hammer-underestimation that surfaces is the structural mirror image of the shot-16 hammer-favorable offset, both driven by selection over noisy candidates with opposite alternating signs. Within real matches the chain responds smoothly to each delivered stone, and, at the directly validated terminal shot, the predicted ESD becomes a high-fidelity predictor of the realized end score (mean assigned probability ≈ 0.71 – 0.72 , against 0.28 for a marginal-frequency baseline on the same subset) when the called shot is in fact executed. These trajectory- and execution-conditioned properties transfer from senior to junior populations, so the framework is best understood as an internally coherent chain anchored to a directly validated final-shot calibration, statistically

plausible as a predictor of the score distribution under intended execution, with a clear and quantified boundary at execution failure.

The two structural limitations (Section 4.4) are of opposite kinds. Lim A is the price of the real-board grounding: rare high-magnitude outcomes are scarce in real play and produce the heavy upper tail of the final-shot KL, addressable by validated augmentation. Lim B is the one place the grounding does not yet reach: execution uncertainty enters the labels through a single fixed simulator-noise envelope, and the natural remedy is to recalibrate it against real-match statistics in the same spirit that grounded the boards and the win-probability vectors. The dominant-factor decomposition presented here additionally indicates where strategic information becomes informative as an end unfolds, suggesting a route to practical decision support built on this real-data foundation.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16146943/s1>, Supplementary Section S1: Robustness of the *sat_outcomes* Selection Criterion—Table S1: Selection-criterion comparison at $\alpha = 0.05$; Table S2: ϵ sweep for *sat_outcomes* at $\alpha = 0.05$.

Author Contributions: R.C. conducted the research, developed the methodology, performed all experiments and analysis, and wrote the manuscript. Y.T. and M.Y. supervised the work. S.A. provided research support. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by “The Enhancement of HPSC Infrastructure through Technology Innovation Project” of Japan Sports Agency.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The stone-position dataset generated in this study and the code used for model training and inference are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Code (Anthropic, powered by the Claude Opus 4 model) for English-language editing and proofreading, for translation assistance, and as a coding aid. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CI	Confidence Interval
ESD	Expected-Score Distribution
EW	Expected Win Probability
FiLM	Feature-Wise Linear Modulation
GT	Ground Truth
KL	Kullback–Leibler (divergence)
TVD	Total Variation Distance
WCF	World Curling Federation
WT	Win-Probability Table
WV	Win-Probability Vector

Appendix A. Architecture Comparison Details

This appendix provides the per-seed predictive-accuracy and matched-condition timing tables that underlie the FiLM-versus-concatenation comparison in Section 2.3. Both architectures share identical preprocessing, encoder configuration ($d=256$, 4 transformer layers, 8 heads, and mean pooling), optimizer, batch size, learning rate, and epoch count; five seeds were used per architecture; test KL is on the same mirror-disjoint test set. Paired statistics in Table A1 are computed at the (board, context) level, $\Delta\text{KL} = \text{KL}_{\text{FiLM}} - \text{KL}_{\text{Concat}}$, with a board-level bootstrap ($K = 95$ contexts per board) for the 95% CI. Timings in Table A2 were measured on an NVIDIA RTX 4080 with mixed-precision training [35] and graph compilation disabled for both architectures and a chunk size of 16,384 at $K = 95$; the per-pass components shared between the two—candidate simulation (~ 30 ms), feature construction (~ 10 ms), and T -matrix evaluation (~ 10 ms)—account for the bulk of fixed cost.

Table A1. Per-seed shot-16 test KL for the concatenation and FiLM architectures. Five-seed mean $\pm$ std for per-seed summaries; paired Δ statistics aggregate the five-seed run.

Quantity	Concatenation	FiLM
Mean KL (five-seed mean $\pm$ std)	0.2296 ± 0.0071	0.2656 ± 0.0090
Median KL (five-seed mean $\pm$ std)	0.0598 ± 0.0207	0.0761 ± 0.0209
Modal-score agreement with label (%)	82.13 ± 0.81	80.65 ± 0.47
Integrated $E[s]$ over the test set	+1.124	+1.080
Paired ΔKL mean (95% CI)	$+0.0360$ ($[+0.0286, +0.0434]$)	
Paired ΔKL median (95% CI)	$+0.0050$ ($[+0.0041, +0.0059]$)	
Fraction of pairs with $\text{KL}_{\text{FiLM}} > \text{KL}_{\text{Concat}}$	65.4%	

Table A2. Matched-condition per-board timing at $K = 95$ over ten boards on the same model and hardware.

Quantity	Concatenation	FiLM
Encoder cost per pass (ms)	~ 110	~ 85
Per-board mean (ms)	10,497	140
Per-board median (ms)	10,553	143
Speedup (FiLM over concatenation)	—	$74.8\times$ (mean) / $73.6\times$ (median)

Appendix A.1. Practical-Significance Anchor Against a Configuration-Blind Baseline

This subsection supports the FiLM-versus-concatenation trade-off summarized in Section 2.3 by sizing it along two axes: (i) the TVD-equivalent of the paired ΔKL against the concatenation reference of Table A1, which inherits the shot-16 design of our previous shot-specific study [13] and currently provides the highest shot-16 accuracy on this task, and (ii) the absolute shot-16 KL against a fully configuration-blind statistical baseline.

The concatenation reference and the proposed FiLM model share the encoder configuration ($d = 256$, 4 layers, 8 heads, and mean pooling), the preprocessing pipeline, and the game-context injection mechanism of [13]; the only departures from the prior shot-16 model are the use of real-match boards in place of large-scale random simulation and the mirror-disjoint train/val/test splits introduced here to remove symmetric-board leakage. As a configuration-blind lower bound we additionally evaluate $\hat{p}_{\text{stat}}$, the empirical end-score distribution of the senior-grade real matches—the black step in Figure 8, normalized over 15,565 end-score observations—used unchanged at every (board, context) pair.

Table A3. Shot-16 test KL of three predictors: the configuration-blind statistical baseline $\hat{p}_{\text{stat}}$ (no seed dependence), the concatenation architecture [13]—the highest-accuracy reference of Table A1—and the proposed FiLM model. Per-architecture entries reproduce the five-seed mean $\pm$ std of Table A1. The total variation distance (TVD) columns report the Pinsker upper bound $\text{TVD} \leq \sqrt{\text{KL}/2}$ evaluated at the mean and median KL of each row, an interpretable upper bound on the probability-mass disagreement between prediction and label.

Predictor	Mean KL	Median KL	TVD (Mean)	TVD (Median)
$\hat{p}_{\text{stat}}$	1.240	1.071	0.787	0.732
Concatenation	0.2296 ± 0.0071	0.0598 ± 0.0207	0.339	0.173
FiLM	0.2656 ± 0.0090	0.0761 ± 0.0209	0.364	0.195

Table A3 places the FiLM trade-off on the lower-bound axis: against $\hat{p}_{\text{stat}}$, FiLM reduces the mean KL by $4.67\times$ ($1.240 \rightarrow 0.266$) and the median KL by $14.1\times$ ($1.071 \rightarrow 0.076$), shrinking the Pinsker TVD bound from $\sim 79\%$ to $\sim 36\%$ at the mean and from $\sim 73\%$ to $\sim 20\%$ at the median. Combined with the 5.0% typical-board/13.4% average TVD bounds against the concatenation reference quoted in Section 2.3, this places FiLM well inside the operating envelope between the configuration-blind floor and the highest-accuracy reference, with the residual gap concentrated on the rare high-magnitude tail discussed as Limitation A in Section 4.4 ($+3 \approx 7\%$, $+4 \approx 1.7\%$, $+5 \approx 0.5\%$ in the senior corpus).

References

1. Bukiet, B.; Harold, E.R.; Palacios, J.L. A Markov Chain Approach to Baseball. *Oper. Res.* **1997**, *45*, 14–23.
2. Bouville, M. Position Play in Carom Billiards as a Markov Process. *J. Quant. Anal. Sports* **2007**, *3*, 4.
3. Tesauro, G. Temporal Difference Learning and TD-Gammon. *Commun. ACM* **1995**, *38*, 58–68.
4. Spearman, W. Beyond Expected Goals. In Proceedings of the 12th MIT Sloan Sports Analytics Conference, Boston, MA, USA, 23–24 February 2018.
5. Bellemare, M.G.; Dabney, W.; Munos, R. A Distributional Perspective on Reinforcement Learning. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 6–11 August 2017; pp. 449–458.
6. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
7. Ito, T.; Kitasei, Y. Proposal and implementation of “digital curling”. In Proceedings of the 2015 IEEE Conference on Computational Intelligence and Games (CIG), Tainan, Taiwan, 31 August–2 September 2015; pp. 469–473.
8. Yamamoto, M.; Katoh, S.; Iizuka, H. Learning of Expected Scores Distribution for Positions of Digital Curling. In Proceedings of the Workshop on Curling Informatics (WCI 2018), Kitami, Japan, 2–3 August 2018; pp. 8–9.
9. Yamamoto, M.; Noguchi, W.; Aihara, S. Data Analysis of Curling Using a Win-Probability Table. In Proceedings of the Symposium on Sports and Human Dynamics 2023, Nagoya, Japan, 10–12 November 2023; The Japan Society of Mechanical Engineers; Paper B-4-4. (In Japanese)
10. Iwasaki, T.; Noguchi, W.; Aihara, S.; Yamamoto, M. Learning of evaluation function and prediction of winning probability of curling based on score difference and ends remaining. In Proceedings of the Symposium on Sports and Human Dynamics 2023, Nagoya, Japan, 10–12 November 2023; The Japan Society of Mechanical Engineers; Paper A-1-3.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
12. Iwasaki, T. Training Method to Improve the Accuracy of Prediction Model of Curling Scores. Master’s Thesis, Hokkaido University, Sapporo, Japan, 2024.
13. Chiba, R.; Tamura, Y.; Aihara, S.; Yamamoto, M. Input Encoding and Data Construction Strategies for Accurate Expected Score Distribution Prediction in Digital Curling. In Proceedings of the 4th IEEE International Workshop on Sport Technology and Research (IEEE-STAR 2025), Trento, Italy, 29–31 October 2025; pp. 132–136.
14. Iwasaki, T.; Noguchi, W.; Tamura, Y.; Aihara, S.; Yamamoto, M. Creation of Training Data and Training for Prediction Model of Curling Scores Using Real Game Data. In Proceedings of the 12th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2024), Porto, Portugal, 21–22 November 2024; pp. 168–179.
15. Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018; pp. 3942–3951.

16. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
17. Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Póczos, B.; Salakhutdinov, R.; Smola, A. Deep Sets. In *Advances in Neural Information Processing Systems (NIPS 2017)*; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 3391–3401.
18. Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; Teh, Y.W. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 3744–3753.
19. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, Israel, 21–24 June 2010; pp. 807–814.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
21. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
22. Suzumura, K.; Tamura, Y.; Aihara, S.; Yamamoto, M. Extraction of Stone Positions from a Sheet Image for Curling Match Database Construction. *Appl. Sci.* **2026**, *16*, 3453. <https://doi.org/10.3390/app16073453>.
23. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60.
24. Lee, D.-H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML Workshop on Challenges in Representation Learning*, Atlanta, GA, USA, 16–21 June 2013.
25. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
26. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362.
27. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
28. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
29. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
30. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
31. Prechelt, L. Early Stopping—But When? In *Neural Networks: Tricks of the Trade*, 2nd ed.; Montavon, G., Orr, G.B., Müller, K.-R., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 53–67.
32. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: New York, NY, USA, 2009.
33. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover’s Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121.
34. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall/CRC: Boca Raton, FL, USA, 1993.
35. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. Mixed Precision Training. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.