GO-MLLM-BENCH: PERCEPTION, DECISIONS, AND STATE UPDATES IN 19×19 GO

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models are increasingly expected to act on what they see, but reading a scene, choosing an action from it and updating it after the action are different skills that one success rate mixes together. We introduce Go-MLLM-Bench, which asks about the same 19×19 Go positions through an image, an exact board matrix or the move history, and scores board reading, move prediction and board updates separately. We will release nested training cohorts from disjoint games with board, engine and mixed targets. The main finding is simple. Models can learn to read the board almost perfectly, but this skill does not carry over to choosing moves or to updating the board after a move. A fine-tuned model that reads nearly every tracking image exactly never produces an exact update, even from the true previous board with a decoder that forces valid output. Handing the update to a rule engine, starting from the model's own reading, recovers every state. The benchmark also shows that locating a salient mark is not reading the board, that cell accuracy rewards copying a stale board, and that legal-move decoding lengthens games without finishing any. Benchmarks should measure what a model does with a state, not only whether it can report it.

1 INTRODUCTION

Multimodal large language models are evaluated on visual reasoning and broad multimodal knowledge (Johnson et al., 2017; Hudson & Manning, 2019; Liu et al., 2023b; Yue et al., 2024), on fine visual detail (Tong et al., 2024; Rahmanzadehgervi et al., 2024), and increasingly as agents that act in environments (Shridhar et al., 2021; Liu et al., 2024a;b; Yang et al., 2025). Games add verifiable rules and outcomes, and recent work studies expert supervision, visual game abilities and complete play (Ma et al., 2026; Wang et al., 2025b; Ouyang et al., 2026; Kolasani et al., 2025; Gao, 2026).

A model that reads a scene correctly may still choose badly from it, or fail to update it after an action. Reading a state, choosing an action and updating the state are separate requirements. A model may locate a conspicuous mark without recovering the rest of the scene, reconstruct the scene without using it for a decision, or copy almost every cell of a state while getting every changed cell wrong. One success rate cannot tell these failures apart.

Go-MLLM-Bench asks the same question about the same Go position in three ways and scores each skill on its own (Figure 1). Go has an exact board, recorded human moves, executable rules and a strong engine reference. Each of 300 positions from distinct human games is shown as an image, a true board matrix or its move history, so we can ask what changes when the model is simply given the correct board. Four static tasks cover board reading, last-move localization, human-move prediction and engine agreement, and separate probes cover board updates and full games. Invalid answers always count as failures.

Our central finding is that models can learn to read the board, but the skill does not transfer to choosing moves or updating the board. A fine-tuned Qwen3-VL-4B reads 65 of 66 tracking images exactly, yet produces none of 60 updates exactly, even when given the true previous board and a decoder that forces valid output. When a rule engine performs the update from the model's own readings, all 60 states are exact. On the static benchmark the same training takes exact board reading from zero to most positions, but its move agreement only equals an engine-only control matched in supervised tokens. The strongest baseline finds every last-move marker yet reconstructs almost no whole board.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

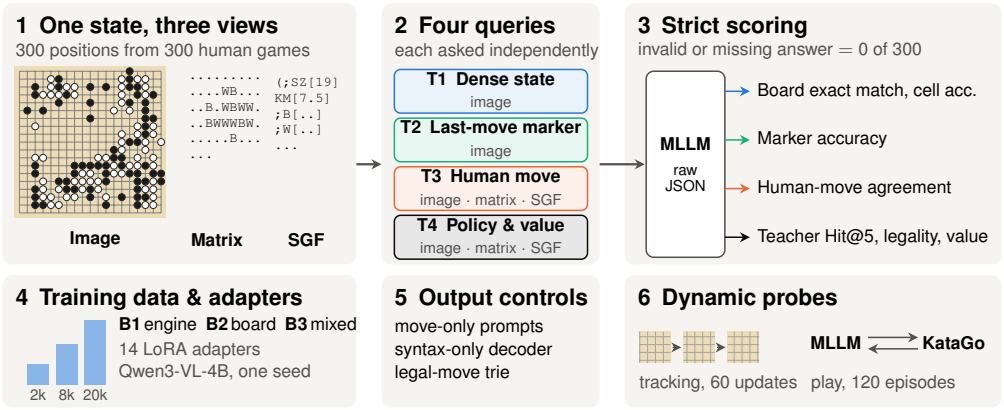


Figure 1: **Go-MLLM-Bench.** (1–3) Each of 300 positions is shown as an image, an exact occupancy matrix or an SGF prefix. Four tasks are asked as separate queries and scored against exact references, with invalid answers counted as failures. (4–6) The same scorer evaluates 14 controlled adapters, output and decoding controls, and disjoint tracking and closed-loop probes.

Scope. All claims concern observable behavior of the evaluated models under the stated prompts and interfaces. The data use one renderer and finite-search engine labels, training uses one seed, and the dynamic probes are small. The benchmark is held out from training and checkpoint selection but was seen during development, and pretraining overlap is unknown, so it is not a blind test. We do not restate these limits below.

Our contributions are threefold.

- **Benchmark.** Go-MLLM-Bench pairs visual and symbolic access to the same 300 states with exact board and legal-move checks and strict scoring.
- **Training data and a controlled study.** Nested cohorts from disjoint games with board, engine and mixed targets, and 14 adapters trained on them, test whether board reading is learnable and whether it transfers.
- **Interface controls.** Matched prompt and decoding controls separate output format from correct decisions and updates.

2 Related Work

Multimodal evaluation. CLEVR and GQA isolate visual reasoning skills (Johnson et al., 2017; Hudson & Manning, 2019), while MMBench and MMMU assess broad multimodal performance (Liu et al., 2023b; Yue et al., 2024). MMVP and BlindTest expose failures in fine visual detail (Tong et al., 2024; Rahmanzadehgervi et al., 2024). Go-MLLM-Bench instead requires every intersection of a dense state and scores it against an exact matrix.

State representations and world models. World models connect representation, dynamics and control (Ha & Schmidhuber, 2018; Schrittwieser et al., 2020), and Othello and automata studies investigate learned state representations and tracking algorithms (Li et al., 2023; Liu et al., 2023a). We measure behavior, not internal world models.

Interactive and game benchmarks. ALFWorld, AgentBench, VisualAgentBench and EmbodiedBench assess interaction across broader environments (Shridhar et al., 2021; Liu et al., 2024a;b; Yang et al., 2025). Within games, LoGos uses expert Go supervision, Playground decomposes visual game abilities, and GameWorld and LLM Chess study verifiable play (Ma et al., 2026; Wang et al., 2025b; Ouyang et al., 2026; Kolasani et al., 2025). GoBench (Gao, 2026) evaluates complete 9×9 games with symbolic boards and authoritative legal-move lists, using KataGo Elo anchors. We instead use 19×19 boards and test visual board reading and board updates separately from play.

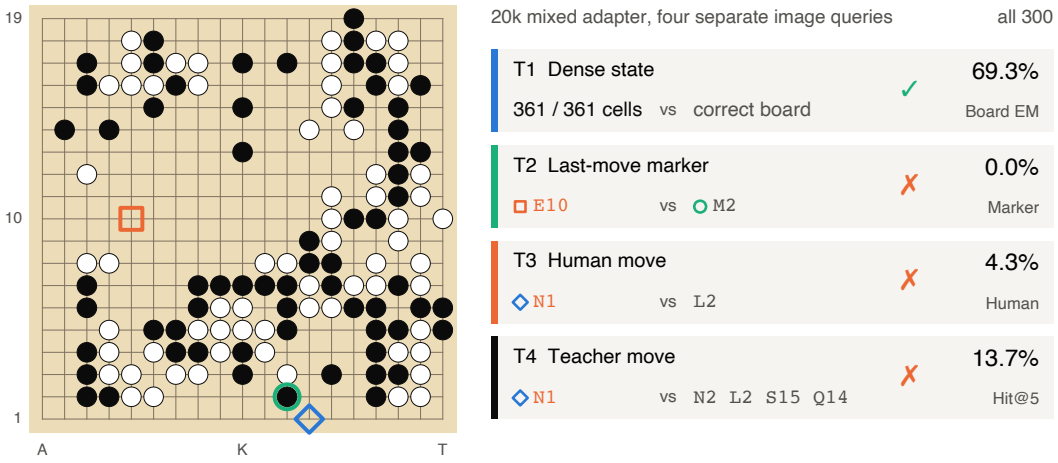


Figure 2: **One position, four separate queries.** Saved image-input answers of the 20k mixed adapter on the first benchmark position. It reproduces all 361 occupancies but names the wrong last-move marker (square, with the true last move circled) and misses the recorded human move and the teacher candidates with its move N1 (diamond). The right column gives the same adapter's scores on all 300 positions (Table 12). Symbols annotate the redrawn board and are absent from the model input.

3 GO-MLLM-BENCH

3.1 TASKS AND INPUTS

Each position can be asked about as a picture, as a matrix or as a game record. We write a Go state as $s = (B, t, H, k, r)$ with occupancy, side to play, move history, komi and rules. Every benchmark position has a 768×768 image, an exact occupancy matrix and an SGF prefix. Four tasks give eight conditions, $1 + 1 + 3 + 3$ (Figure 1), and each is a separate query, so a board reading is never fed into a later task (Figure 2).

T1 asks for the whole board as 19 strings of 19 characters over $\{\texttt{B}, \texttt{W}, \texttt{.}\}$. T2 asks for the coordinate of the last-move marker, which the prompt does not reveal. Coordinates use A–T without I and rows 1–19 from the bottom. T3 asks for the recorded human move. T4 asks for four fields from a fixed KataGo reference (Wu, 2019), namely best move, candidates, Black win probability and score lead. Human and engine best moves agree on only 136/300 positions, so T3 and T4 are different targets.

For T3 and T4 every input carries the same side, rules, komi, last move and repetition constraints. The matrix removes the need to read the image, and SGF adds the history up to the queried state. Switching from image to matrix therefore changes the whole interface, not only the vision encoder. Models answer directly, and no chain-of-thought is requested or graded.

3.2 SCORING

An answer counts only if it is exactly valid. Responses must be raw JSON matching the task schema. Prose, Markdown fences, extra keys, wrong board sizes and nonfinite numbers fail, and we never repair an answer. For a validity indicator g_i and $N = 300$,

$$\text{BoardEM} = \frac{1}{N}\sum_i g_i \mathbf{1}[\widehat{B}_i = B_i], \qquad \text{CellAcc} = \frac{1}{361N}\sum_i g_i \sum_{u=1}^{361} \mathbf{1}[\widehat{B}_{iu} = B_{iu}]. \tag{1}$$

One wrong intersection fails BoardEM, and invalid boards get no cell credit.

Teacher Hit@5 asks whether the model's `best_move` is in the engine's candidate list, which holds one to five moves (five on 182/300 positions). Legality is checked by replaying the full history with positional superko. Missing and invalid answers stay in every denominator. Value errors are mea-

Name	Meaning
Training arms (Qwen3-VL-4B, LoRA, nested 2k/8k/20k game cohorts, seed 42)	
Board-only (B2)	trained to output the board (T1)
Engine (B1-E/U/T)	trained on engine answers (T4), matched to the mixed arm in engine examples (E), optimizer updates (U) or supervised tokens (T, the *token control*, 8k/20k only)
Mixed (B3)	trained on board and engine answers; "mixed model" means 20k B3
Tracking conditions (6 games, 66 images, 60 transitions)	
Fresh reading	reconstruct the board from each new image
Neural carry	update the model's own previous board with the supplied move
True-state reset	replace the previous board with the truth every 1, 5 or 10 steps
Rule carry	a rule engine replays the moves from the model's first reading
Metrics (percent of 300 unless counts are given)	
Board EM / Cell	whole board exactly right / fraction of correct intersections
Marker / Human	last-move coordinate correct / recorded human move matched
Hit@5 / Legal	best move among the engine's candidates / best move legal
Changed points	correct intersections among the 233 whose occupancy changes

Table 1: **Names used in the results.** Four arms at 2k and five at 8k/20k give 14 adapters (Appendix B).

sured on valid T4 answers and compared with a constant guess on the same positions. A uniformly random legal move would reach 1.95% Hit@5 and 0.47% human agreement.

3.3 Data

The 300 positions come from 300 distinct public-domain human games in Joe's Go Database (JGDB, n.d.). They are balanced across opening, middle and late play (100 each) and side to play (150 each), and hold 20 to 254 stones. All use one renderer, Tromp–Taylor rules, komi 7.5 and 64-visit KataGo labels. Training games and states are disjoint from the benchmark. Appendix A documents sampling, audits and exclusions (Gebru et al., 2021).

4 Experimental Setup

Models. We test seven baselines under one fixed protocol. Four are local checkpoints from Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025a), run with greedy BF16 decoding and limits of 1,024 tokens for boards, 256 for moves and 512 for T4. Three are gateway models, whose provider compute is not equalized (Appendix A). Paired intervals resample source games.

Controlled training. The training study asks whether board reading can be taught and whether it then helps moves. Table 1 lists the 14 LoRA adapters (Hu et al., 2022). All share one checkpoint rule and keep the vision encoder frozen. No single engine arm matches all compute quantities, so we report all three.

Output controls. Three controls test whether output format explains the failures. On all 300 games the base, mixed and 20k token-control models are queried afresh with full T4, a single `best_move` field with the example, or that field without the example (8,100 responses). For tracking, all 60 true-state updates are rerun with a decoder that only forces 19 rows of 19 valid characters and knows nothing about the correct board. In play, a move-only prompt is paired with raw decoding, a trie over coordinates and PASS, or a trie over legal moves.

Dynamic probes. Tracking gives the model every intervening move across six disjoint games (Table 1). In closed-loop play, the base, mixed and 20k token-control models face 64-visit KataGo from five starts, with both colors and four access modes (vision, legality feedback, rule-maintained state and oracle state). Each of the 120 episodes allows 120 plies and 600 seconds, and full-history replay checks every accepted move.

Model	Board EM	Cell	Marker	Human	Hit@5	Legal	WR MAE
Qwen3-VL-4B	0.0	5.0	3.3	0.0	1.3	27.7	35.8
Qwen3-VL-8B	0.0	0.6	0.3	0.0	1.3	27.7	35.8
InternVL3.5-4B	0.0	3.6	6.3	0.0	0.0	26.3	34.9
InternVL3.5-8B	0.0	0.2	3.7	0.3	0.7	26.0	35.8
Gemini 3.8 Flash†	1.3	38.5	100.0	23.7	31.0	96.0	37.7
GPT-5.6 Terra	5.7	36.5	87.0	8.3	11.0	85.7	35.7
Qwen3.8-Flash	0.0	3.2	70.7	2.3	2.0	45.0	36.8

Table 2: **Image-input endpoints on 300 games.** Accuracy and legality are percentages, and missing or invalid outputs score zero. Cell accuracy is unconditional. WR MAE is conditional probability error in percentage points (lower is better), with valid coverage in Table 4. † marks one missing Gemini T4 answer. The columns are not averaged.

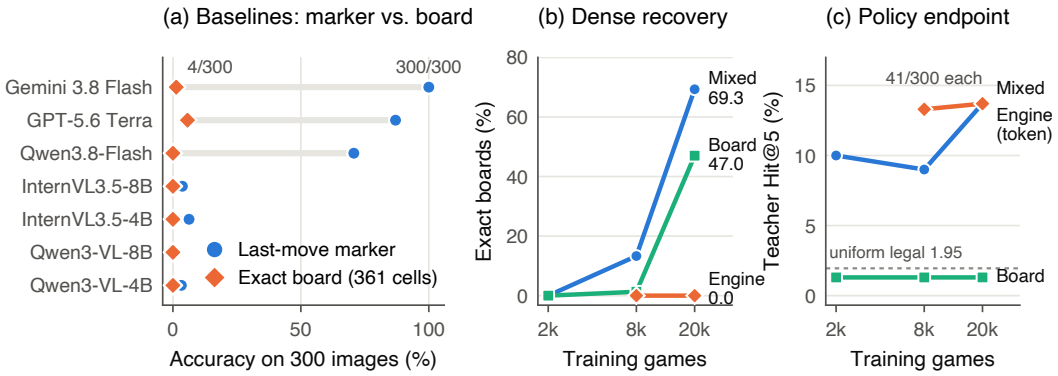


Figure 3: **Dense recovery improves without an established policy gain over the token control.** (a) Image-input baselines, marker localization versus whole-board exact match (Table 2). (b, c) Qwen3-VL-4B adapters on the same 300 images and original T4 contract, with all 14 checkpoints in Table 12.

5 Results

5.1 Finding a mark is not reading the board

The best baselines find the last move but cannot reconstruct the board (Table 2). Gemini localizes all 300 last-move markers yet reconstructs only four whole boards (Figure 3a). Terra finds 261/300 markers and 17/300 boards, and all four local baselines reconstruct none.

A valid answer is not a correct one either. Qwen3-VL-4B copies the JSON example on all 300 image T4 queries, so it is always schema-valid but hits the teacher only 4/300 times, by coincidence. Prompt-example, shuffling and processor checks are in Appendix B.4.

5.2 Board reading is learnable but does not help moves

Training teaches the model to read the board. At 20k the mixed model reconstructs 208/300 boards exactly, against 141/300 for board-only training and none for the base model (Figure 3b, Figure 4). The benchmark can therefore register real progress.

Reading improves steeply with data, and each target moves only its own endpoint (Figure 4). Exact boards for the mixed arm go from 0 to 40 to 208 across the 2k, 8k and 20k cohorts, and for board-only training from 0 to 4 to 141. Cell accuracy hides how hard this is, since the 20k mixed model gets 97.4% of cells right but only 69.3% of whole boards. Board training does not teach last-move localization, as every board-trained arm from 8k on finds at most 0.3% of markers while the engine arms find 10 to 17%. Board-only Hit@5 stays at the base model's 1.3% at every size.

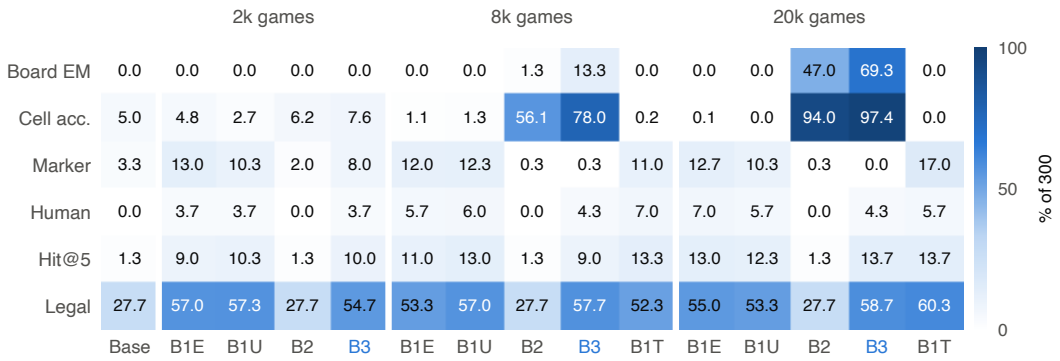


Figure 4: **Each supervision target moves a different endpoint.** Image-input scores of the base model and all 14 adapters on the 300 positions (Table 12). Board-only supervision (B2) raises reconstruction but leaves Hit@5 at the base level, engine supervision (B1) raises Hit@5 and legality without any exact board, and the mixture (B3) combines both without exceeding the token control (B1T) on Hit@5.

Reading better does not translate into better moves. The mixed model and the token control both reach 41/300 teacher hits, a paired difference of zero with 95% interval $[-4.67, 4.67]$ points (Figure 3c). On the 208 positions the mixed model reads exactly, it gains 21 teacher hits over the token control and loses 21, a descriptive split rather than a causal estimate. Feeding a fixed engine-trained policy the true board or the boards predicted by the board-only and mixed models gives 13, 12 and 12 hits, so even perfect reading helps little here (Appendix B.3).

Visual transfer is narrow. The learned reading is tied to the training renderer. On stress probes the mixed model rejects all 120 absent markers but localizes only 1/120 dots, 1/120 squares and 4/120 triangles. Switching to the gameplay renderer drops exact boards from 85/120 to 0/120 (Figure 8a, Appendix D).

5.3 Giving the true board does not improve moves

Handing a model the correct board does not make its moves better. If misreading the image were the only obstacle, the true matrix should help. For Gemini, teacher Hit@5 instead moves from 31.0% with images to 29.3% with matrices and 22.7% with SGF (Figure 5a), a paired matrix-minus-image difference of -1.7 points. Terra shows the same pattern, with 33, 16 and 36 hits and a paired difference of -5.67 points. The local baselines stay at or below 1.3% under every input, below the 1.95% random-legal anchor. A correct board still has to be interpreted, and the matrix interface is a new input format.

The paired intervals make the direction clear (Table 3). Gemini's matrix-minus-image interval, $[-6.7, +3.3]$ points, contains zero, and Terra's, $[-10.0, -1.7]$, lies entirely below zero, so for Terra the true board hurts. SGF moves every other model by at most two points but lowers Gemini to its weakest score, even though it adds the full history.

5.4 Shorter outputs do not rescue decisions

Asking for less does not reliably help. For the base model, keeping the example gives 4/300 hits under either contract (Figure 5b), and only dropping the example lifts it to 14/300, a paired gain of 3.33 points. For the mixed model the same three arms give 41, 38 and 26 hits, so the single-field prompt fixes its five invalid answers but lowers Hit@5 by 1.00 points. Among parseable full-T4 answers, ignoring errors in the other fields adds no teacher hits in any of the nine model and input cells (Appendix B.5).

Value prediction. Models also do not judge who is winning. No baseline's win-probability error is clearly below that of always answering 0.5 on the same valid positions (Table 4). Both Qwen3-VL baselines answer exactly 0.5 on all 300 positions, Gemini's error of 37.65 points is worse than the

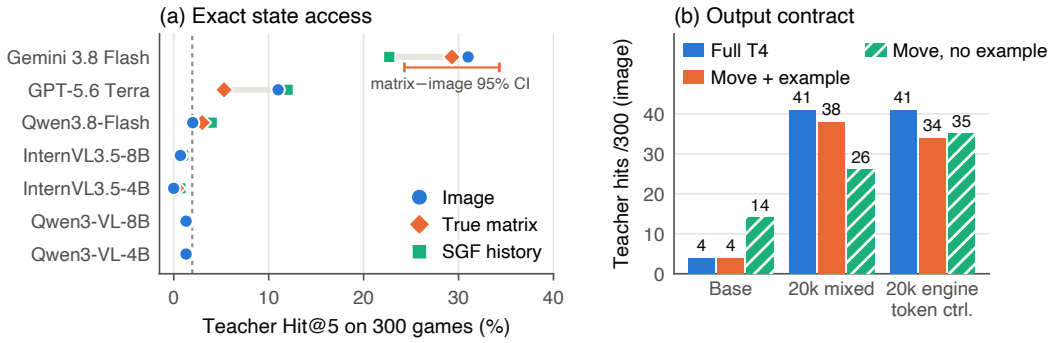


Figure 5: **Neither exact state access nor a shorter output uniformly improves decisions.** (a) Teacher Hit@5 under paired image, true-matrix and SGF access, with the 1.95% uniform-legal anchor dashed. SGF adds history, so it is not an information-matched ablation. Paired intervals are in Table 3; Terra coverage is in Appendix A.3. (b) Image-input teacher hits per newly generated arm. All have 300 valid answers except base move-only without example (293) and mixed full T4 (295) (Appendix B.5).

Model	Image	Matrix	Matrix − image [95% interval]	SGF
Qwen3-VL-4B	1.3	1.3	+0.0 [+0.0, +0.0]	1.3
Qwen3-VL-8B	1.3	1.3	+0.0 [+0.0, +0.0]	1.3
InternVL3.5-4B	0.0	0.3	+0.3 [+0.0, +1.0]	0.7
InternVL3.5-8B	0.7	0.7	+0.0 [-1.0, +1.0]	1.0
Gemini 3.8 Flash†	31.0	29.3	-1.7 [-6.7, +3.3]	22.7
GPT-5.6 Terra	11.0	5.3	-5.7 [-10.0, -1.7]	12.0
Qwen3.8-Flash	2.0	3.0	+1.0 [-1.3, +3.3]	4.0

Table 3: **Teacher Hit@5 under paired state access** (Figure 5a). Percentages and paired percentage-point differences use 300 source games. SGF adds history, so it is not an information-matched vision ablation. Missing/invalid answers score zero. Terra's paired interval uses its complete 300-game matrix; response and schema coverage are detailed in Appendix A.3.

constant's 35.55, and the best case, InternVL3.5-4B, improves on the constant by under one point. The mixed model's value error is not clearly better than a constant 0.5 guess on the same positions (paired difference -1.70 points, Appendix A.2).

5.5 Reading the board does not mean updating it

The mixed model can read the board but cannot update it. It reconstructs 65/66 fresh tracking images exactly, yet none of its 60 neural-carry updates is exact. Resetting to the true previous board at every step still gives 0/60, so accumulated error does not explain the failure.

The same split holds across all eleven models (Table 5). Fresh reading grows with board supervision, from 22/66 at 8k to 65/66 at 20k for the mixed arm and from 4/66 to 58/66 for board-only training, and engine-only adapters read no tracking image exactly. Neural-carry and true-state updates are 0/60 for every model. Rule carry tracks reading quality instead, reaching 20/60 exact states and 152/233 changed points for the 8k mixed arm and 10/60 and 81/233 for 8k board-only.

Part of the failure is formatting, but most is content. All 60 raw answers parse, yet only two carried and four reset answers satisfy the board schema. With the syntax-only decoder every answer becomes a valid board (Figure 7), and correct changed points rise from 1/233 to 11/233, but exact updates stay at zero.

Cell accuracy hides a model that does nothing. The base model copies all 60 previous boards under both decoders, which gives **98.92% cell accuracy and 0/233 correct changed points**, exactly the score $1 - 233/(60 \times 361)$ of a stale board. Only exact-match and changed-point metrics expose it.

Model	Valid/300	WR MAE	Constant MAE	Exactly 0.5
Qwen3-VL-4B	300	35.83	35.83	300
Qwen3-VL-8B	300	35.83	35.83	300
InternVL3.5-4B	300	34.89	35.83	0
InternVL3.5-8B	300	35.78	35.83	0
Gemini 3.8 Flash	292	37.65	35.55	0
Qwen3.8-Flash	297	36.76	35.82	113
GPT-5.6 Terra	297	35.74	35.74	20

Table 4: **No baseline beats a constant value guess.** Image-input teacher-value errors in percentage points. The constant predicts 0.5 and uses exactly the model's schema-valid IDs. Terra has 300 returned image-T4 answers, of which 297 are valid.

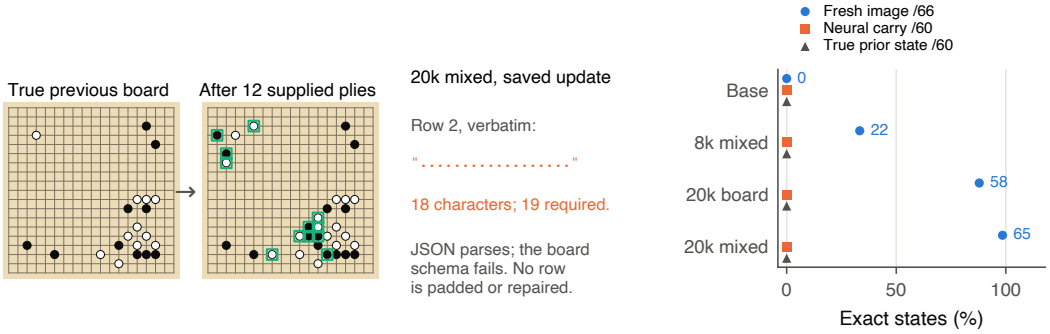


Figure 6: **Exact recognition can coexist with failed state updates.** The left boards show the first frozen transition (12 supplied plies), with changed intersections boxed. The middle text is the second row of the saved 20k mixed update, verbatim. The right panel counts exact states over six games, where neural-carry and true-state updates are 0/60 for all four checkpoints (complete panel in Table 5). The syntax control is in Figure 7.

A rule engine fixes the update. When the moves are replayed by rules from the 20k board-only and mixed models' own first readings, all 60 states and 233 changed points are exact (Appendix C). The reading is good enough, and the missing piece is the transition itself. Only three changed points are captures, so captures are lightly tested.

5.6 Legal moves do not finish games

No closed-loop episode finishes. Of the 120 episodes, 92 end on an invalid action, 26 on invalid initialization, one on an inconsistent ledger and one on wall time, and 40 direct base–mixed matches are also incomplete. We report these as execution outcomes, not losses.

Syntax is not the bottleneck, legality is (Table 19). Under raw decoding every one of the 40 base and mixed episodes ends on an invalid move within four plies on average. The coordinate trie removes malformed output but not illegal moves, so each episode still ends early after at most 4.3 plies. Supplying the oracle state instead of the image changes game length by less than one ply.

Forcing legal moves makes games longer but not complete. In the 40 legal-trie episodes illegal proposals disappear by construction, yet every episode reaches the time cap and no arm ever passes (Figure 8b, Appendix C.4). Gameplay uses its own prompts and renderer, so these outcomes are not comparable with static scores (Appendix C.3).

6 Conclusion

Go-MLLM-Bench asks about the same Go states through images, exact matrices and move histories, and scores board reading, decisions and updates separately. With training cohorts from disjoint games and matched output controls, it shows that these skills come apart. Models can learn to

Model	Fresh EM /66	Neural EM /60	Oracle EM /60	Oracle changed /233	Rule EM /60	Rule changed /233
Base	0	0	0	0	0	0
8k B1E	0	0	0	0	0	0
8k B1U	0	0	0	0	0	0
8k B1T	0	0	0	0	0	0
8k B2	4	0	0	2	10	81
8k B3	22	0	0	0	20	152
20k B1E	0	0	0	0	0	0
20k B1U	0	0	0	0	0	0
20k B1T	0	0	0	0	0	0
20k B2	58	0	0	1	60	233
20k B3	65	0	0	1	60	233

Table 5: **Reading improves with data, updating does not.** Tracking results for the base model and all 8k/20k adapters. Oracle resets the previous state at every update. Neural and rule carry start from the same first-image prediction; rule carry then applies supplied moves without reference repair.

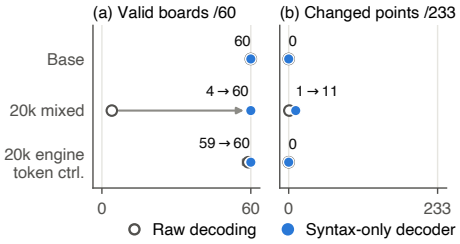


Figure 7: **Valid serialization does not recover exact updates.** All 60 transitions start from the true previous board. Exact updates are 0/60 for every model under both decoders, and invalid raw answers receive zero credit. Changed points are the 233 intersections whose occupancy changes across the six trajectories (Appendix C.2).

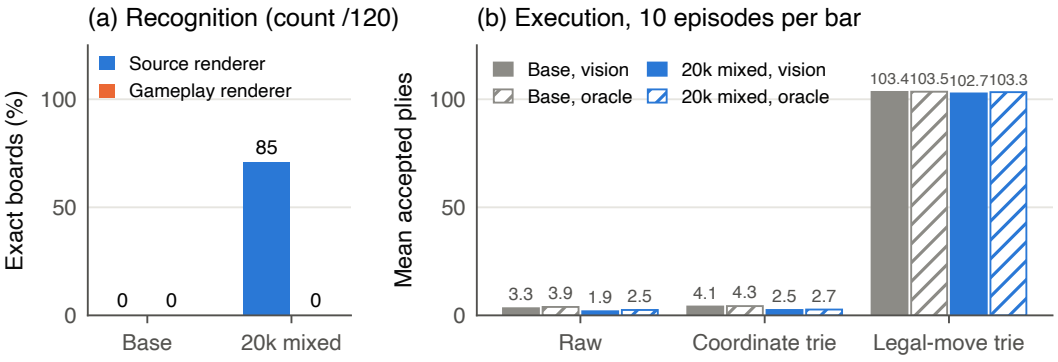


Figure 8: **Assistance lengthens play without completing games.** (a) Exact boards on 120 matched games under the source and gameplay renderers (Table 22). (b) Mean accepted plies under raw decoding, a coordinate trie and a legal-move trie (Table 19). Legal masking removes illegal proposals and runs to the time cap, but all 120 episodes remain incomplete, so the longer games are not evidence of stronger play.

read the board, but that skill neither improves their moves nor lets them update the board, while a rule engine given the same readings updates it perfectly. Locating a mark is not reading the board, cell accuracy rewards a stale board, and legal moves alone do not finish a game. Scoring each step separately, with every invalid answer counted, turns these failures into concrete targets. Wider renderers, more seeds and longer dynamic probes are the natural next steps beyond the scope stated in Section 1.

AI USE STATEMENT

AI assistants helped search for sources, and revise text and code. The authors are responsible for the claims, citations, interpretations and final manuscript.

REPRODUCIBILITY STATEMENT

We will release retained data, prompts, split manifests and scoring code with versioned results.

ETHICS STATEMENT

Human game records originate from JGDB, whose publisher releases them into the public domain (JGDB, n.d.). The study evaluates game states and moves, not player attributes; unnecessary player metadata are excluded. Board references are checked by SGF replay and image audits; policy and value labels are finite-search KataGo outputs. These automated checks are not represented as independent human image review.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Roland Gao. GoBench: Evaluating LLMs on the game of Go. https://rolandgao.com/gobench.pdf, 2026. Preprint; accessed September 26, 2026.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. URL https://arxiv.org/abs/1803.09010.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://arxiv.org/abs/1902.09506.

JGDB. Joe's go database. https://pjreddie.com/projects/jgdb/, n.d. Accessed September 26, 2026. Publisher releases the dataset into the public domain.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.06890.

Sai Kolasani, Maxim Saplin, Nicholas Crispino, Kyle Montgomery, Jared Quincy Davis, Matei Zaharia, Chi Wang, and Chenguang Wang. LLM CHESS: Benchmarking reasoning and instruction-following in LLMs through chess. *arXiv preprint arXiv:2512.01992*, 2025.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.13382.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations*, 2023a. URL https://arxiv.org/abs/2210.10749.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024a. URL https://arxiv.org/abs/2308.03688.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, et al. VisualAgentBench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024b. URL https://arxiv.org/abs/2408.06327.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023b. URL https://arxiv.org/abs/2307.06281.

Yichuan Ma, Linyang Li, Yongkang Chen, Peiji Li, Jiasheng Ye, Qipeng Guo, Dahua Lin, and Kai Chen. Mixing expert knowledge: Bring human thoughts back to the game of go. *arXiv preprint arXiv:2601.16447*, 2026.

Mingyu Ouyang, Siyuan Hu, Kevin Qinghong Lin, Hwee Tou Ng, and Mike Zheng Shou. Gameworld: Towards standardized and verifiable evaluation of multimodal game agents. *arXiv preprint arXiv:2604.07429*, 2026.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2024. URL https://arxiv.org/abs/2407.06581.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020. URL https://arxiv.org/abs/1911.08265.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.03768.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xinyu Wang, Bohan Zhuang, and Qi Wu. Are large vision language models good game players? *arXiv preprint arXiv:2503.02358*, 2025b.

David J. Wu. Accelerating self-play learning in go. *arXiv preprint arXiv:1902.10565*, 2019. URL https://arxiv.org/abs/1902.10565.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.09560.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16502.

A Data and Evaluation Protocol

A.1 Human-game source and sampling

The human-game corpora originate from Joe's Go Database (JGDB), which its publisher releases into the public domain (JGDB, n.d.). SGF records are replayed and eligible 19×19 positions are sampled with fixed random seeds, subject to move-range and validity filters. Source splits are grouped by game. Human next moves are recorded actions; policy and value references are generated by KataGo, rather than human annotations of optimal play. The retained source exports contain 100,000 static positions from 100,000 games and 34,675 trajectory positions from 698 games.

The reported benchmark is the frozen 300-position subset of these source exports. Its teacher budget is 64 visits, with Tromp–Taylor rules and komi 7.5. Earlier upstream real/self-play generation used different sampling and analysis settings; those collections do not replace the membership or labels behind the results in this paper. Some one-off local preprocessing scripts were not retained. The available manifests, fixed examples, scoring code and reported configuration define the reproducibility scope.

The final set has one position per source game, with fixed quotas of 100 opening, 100 middle and 100 late positions and 150 per side to play. Selection within quotas uses seeded hash order and validity, not model performance. Original image bytes are preserved: each 768×768 image contains coordinates and a last-move marker. SGF replay and image checks cover all 108,300 intersections. Independent human image review has not been completed.

Property	Frozen benchmark
Positions / source games	300 / 300
Move-number range	20–270
Stone count: range / median	20–254 / 117.5
Teacher-list length 1 / 2 / 3 / 4 / 5	21 / 26 / 30 / 41 / 182
Human move equals teacher best	136/300
Human / teacher-best PASS targets	0 / 0
History-repetition-forbidden positions	0

Table 6: Coverage of the evaluated cohort. Phase boundaries are moves 20–80, 81–180 and 181–280. The absence of PASS targets and repetition cases limits claims about those rule situations.

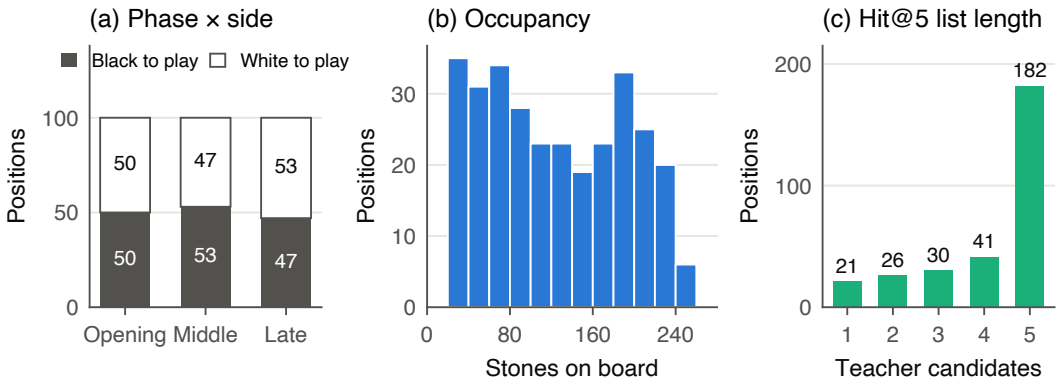


Figure 9: Coverage of the 300 benchmark positions (Table 6). (a) Phase quotas are exact; side to play is balanced overall (150/150), not within each phase. (b) Stone counts in 20-stone bins. (c) Teacher lists hold one to five candidates, so Hit@5 uses the available list. These are dataset counts, not model outcomes.

All 300 source games and D4-equivalent board-plus-side states are excluded from the original 131,307 training rows and the intervention exports. The 120-game training-validation panel, 12-game native diagnostics and six-game tracking cohort have their own recorded exclusions. The original 200 and added 100 form one primary set. The latter cannot also serve as an unopened con-

firmation set. No benchmark checkpoint selection is performed, but same-source construction and development exposure preclude a certified blind-test claim; pretraining overlap is unknown.

A.2 Output contracts and inference

Image and matrix policy inputs share side, rules, komi, last move and history-dependent repetition prohibitions. SGF inputs end at the queried position and exclude the future human move, but supply extra history. Coordinates use A–T without I and rows 1–19 from bottom to top; matrices list row 19 first. The marker task withholds its target coordinate.

Task	Required JSON fields and constraints
T1	`board`: exactly 19 strings of 19 characters from `B`, `W`, `.`.
T2	`last_move`: one coordinate or PASS.
T3	`human_next_move`: one coordinate or PASS.
T4	`best_move`, `top_moves`, `black_winrate`, `black_score_lead`. The list has 1–5 distinct moves, best first; its first move equals `best_move`. Probability is finite in $[0, 1]$; score lead is finite, with positive values favoring Black.

Table 7: Strict schemas. Additional keys, prose, fences, duplicate keys, invalid dimensions and nonfinite or overflowing numbers fail validation. Coordinate case and surrounding whitespace are normalized; boards are not repaired. T4 endpoints require all four fields to be valid.

Local inference uses native processors, BF16/SDPA and greedy direct-answer decoding, with limits of 1,024 tokens for boards, 256 for moves and 512 for engine JSON. We neither request nor evaluate chain-of-thought. Model-specific visual preprocessing and undisclosed provider computation are not matched.

Checkpoint	Revision prefix	Visual-input bounds
Qwen3-VL-4B-Instruct	`ebb281ec70b0`	200,704–802,816 pixels
Qwen3-VL-8B-Instruct	`0c351dd01ed8`	200,704–802,816 pixels
InternVL3_5-4B-HF	`6bd448740211`	1–12 patches
InternVL3_5-8B-HF	`741a7d030204`	1–12 patches

Table 8: Pinned local checkpoints. Manifests retain complete revisions, weight hashes, prompts and processor settings. Gateway names are requested service aliases, not independently certified upstream identities.

Metrics and uncertainty. Teacher Hit@5 checks the predicted *single best move* against the teacher's available list; short lists are not padded. Human accuracy uses the observed human action. Legality replays full history, including positional superko and the specified self-capture rule. Binary endpoints retain missing and schema-invalid outputs as failures. Conditional numerical errors show coverage and compare models with constants on identical valid IDs. Black-relative values do not change convention with side to play. These scores measure teacher agreement, not optimal-move regret, outcome calibration or Elo.

The primary modality contrast uses 10,000 paired game-bootstrap resamples with seed 20260906; the transfer analyses use seed 20260914. Binary-rate Wilson intervals and paired intervals are descriptive, without multiplicity adjustment or training-seed uncertainty. No returned answer is retried because it is wrong. Compatible earlier answers are reused only when IDs and model-visible inputs agree.

Value prediction needs a matched constant. Both local Qwen baselines output Black probability 0.5 on all 300 image T4 queries, matching a constant predictor's 35.83-point MAE. Gemini's variable outputs have 37.65-point MAE on 292 valid answers, versus 35.55 for constant 0.5 on those same IDs (Table 4). More variable predictions and stronger move agreement therefore do not establish more accurate numerical assessment.

The 20k mixed model has MAE 34.09 versus 35.79 for the matched constant on its 295 valid answers. Their paired difference is -1.70 points, $[-5.15, 1.77]$, which does not establish a value

advantage despite value supervision in every engine-trained arm. Valid coverage and the same-ID constant are necessary to interpret conditional errors; these remain errors relative to the fixed teacher, not game-outcome calibration.

A.3 Coverage and completed Terra acquisition

Six baseline matrices contain all 2,400 responses. Gemini has 2,399; its missing image-engine response scores zero for accuracy. Terra's final 23 transport-missing requests are completed using the same requested model alias, frozen inputs, prompts and requested token budgets. Every earlier successful response is retained, including malformed and incorrect answers; no answer is resampled on the basis of quality. Two preceding model-access failures produce no answers and are retained in the request audit. All 23 newly completed requests return the requested alias. The continuation uses Chat and Responses transports, including streamed recovery after logged failures or cancellation during the route switch. Task text, image bytes, requested reasoning effort and output-token limits are unchanged; every successful benchmark answer from the transport checks is retained. The gateway nevertheless reports reasoning and output-token use beyond requested limits on retained responses, so these settings do not establish equal inference compute. Gateway labels do not independently certify upstream checkpoint identity or stability over time. Table 9 distinguishes acquisition from schema validity and task correctness; all binary endpoints use 300 as their denominator.

Endpoint	Input	Returned	Schema-valid	Correct
Board EM	image	300/300	115/300	17/300
Marker	image	300/300	299/300	261/300
Human	image	300/300	300/300	25/300
Human	matrix	300/300	299/300	21/300
Human	sgf	300/300	294/300	18/300
Teacher Hit@5	image	300/300	297/300	33/300
Teacher Hit@5	matrix	300/300	295/300	16/300
Teacher Hit@5	sgf	300/300	281/300	36/300

Table 9: Completed Terra gateway results, with 300 positions per row. The policy endpoint is membership of the predicted single best move in the teacher's available top-five list. Invalid responses receive no credit.

B Controlled Training and Complete Adapter Results

B.1 Supervision and optimization

All adapters use Qwen3-VL-4B-Instruct and nested 2k/8k/20k training cohorts, one epoch, seed 42 and a fixed final-epoch checkpoint. The same 120 disjoint validation games are used throughout. LoRA (Hu et al., 2022) targets language attention Q/K/V/O projections with rank 32, alpha 64 and dropout 0.05; the vision encoder and merger remain frozen. Training uses BF16/SDPA, batch size 2, eight accumulation steps, learning rate 2×10^{-5}, cosine decay, 0.03 warmup and 0.01 weight decay. Image bounds are 200,704–802,816 pixels. Assistant targets and their ending are supervised; prompt, image and padding tokens are masked. Native-processor checks verify the masks and zero truncation within the 4,096-token limit. Each final adapter is independently reloaded before evaluation.

B.2 All final checkpoints on the same 300 games

Every adapter returns all 2,400 answers; the base responses are reused. Table 12 reports all image endpoints. The 20k B3 and B1-T teacher-hit difference is 0.0 percentage points with paired 95% interval $[-4.67, 4.67]$. Their equal 41/300 counts do not prove policy equivalence. On the 208 games where B3 gains exact state, it gains 21 teacher hits and loses 21. On the remaining 92 games it gains four and loses four. These post-training strata are descriptive, not causal mediation estimates.

The 20k mixed model's value error is 34.09 versus 35.79 for constant 0.5 on its 295 valid answers; the paired difference is -1.70 points, $[-5.15, 1.77]$. Every engine-supervised arm trains probability

Arm	Targets	Matched quantity / limitation
B1-E	Full engine JSON (T4)	Same distinct engine examples as B3.
B1-U	Repeated T4	Same optimizer updates as B3; fewer target tokens.
B1-T	Repeated T4 prefix	Approximately B3 target tokens; more updates.
B2	Board JSON (T1)	State supervision only.
B3	T1 and full T4	Joint board, policy and value supervision.

Table 10: Four arms at 2k and five at 8k/20k, totaling 14 adapters. B1-T is absent at 2k. No arm uses rationale targets. These controls match different quantities and do not jointly match exposure, updates and FLOPs.

Cohort	Arm	Example exposures	Steps	Target tokens
2k	B1-E	2,000	125	93,640
	B1-U	4,000	250	187,280
	B2	2,000	125	293,540
	B3	4,000	250	387,180
8k	B1-E	8,000	500	374,938
	B1-U	16,000	1,000	749,876
	B1-T	33,002	2,063	1,546,737
	B2	8,000	500	1,171,819
	B3	16,000	1,000	1,546,757
20k	B1-E	20,000	1,250	937,103
	B1-U	40,000	2,500	1,874,206
	B1-T	82,464	5,154	3,863,884
	B2	20,000	1,250	2,926,778
	B3	40,000	2,500	3,863,881

Table 11: Recorded training budgets. Exposures include repeated examples. The 20k B3/B1-T difference is three target tokens but 2,654 optimizer steps. Increasing unique training games also increases work under this recipe.

Checkpoint	Board EM	Cell	Marker	Human	Hit@5	Legal	WR MAE	Valid
Base	0.0	5.0	3.3	0.0	1.3	27.7	35.83	300/300
2k B1E	0.0	4.8	13.0	3.7	9.0	57.0	35.61	300/300
2k B1U	0.0	2.7	10.3	3.7	10.3	57.3	35.75	300/300
2k B2	0.0	6.2	2.0	0.0	1.3	27.7	35.83	300/300
2k B3	0.0	7.6	8.0	3.7	10.0	54.7	35.48	298/300
8k B1E	0.0	1.1	12.0	5.7	11.0	53.3	33.96	300/300
8k B1U	0.0	1.3	12.3	6.0	13.0	57.0	37.79	299/300
8k B2	1.3	56.1	0.3	0.0	1.3	27.7	35.83	300/300
8k B3	13.3	78.0	0.3	4.3	9.0	57.7	34.12	300/300
8k B1T	0.0	0.2	11.0	7.0	13.3	52.3	33.99	300/300
20k B1E	0.0	0.1	12.7	7.0	13.0	55.0	34.14	300/300
20k B1U	0.0	0.0	10.3	5.7	12.3	53.3	33.15	299/300
20k B2	47.0	94.0	0.3	0.0	1.3	27.7	35.83	300/300
20k B3	69.3	97.4	0.0	4.3	13.7	58.7	34.09	295/300
20k B1T	0.0	0.0	17.0	5.7	13.7	60.3	33.32	300/300

Table 12: Complete image-input results on 300 games, seed 42. Accuracy, cell accuracy and legality are percentages, with invalid answers scoring zero. WR MAE is conditional Black-probability error in percentage points; Valid is its schema-valid count. Results are not selected by test score.

and score lead, but this interval does not establish a value advantage. All intervals measure game variation, not training-seed variation.

B.3 Explicit state replacement for a fixed policy

We hold the 20k B1-T decision policy fixed and supply true occupancy or the 20k B2/B3 board predicted from the same image. Side, rules, komi, last move, system instruction and budget remain fixed. Among 600 predicted-state conditions, 349 exact prompt matches reuse saved answers, 228 require new inference and 23 fail upstream formatting. Invalid states receive no policy call and count as end-to-end failures.

State for fixed B1-T	Exact input	Available	Hit@5	Legal
True matrix	300	300	13	230
20k B2 prediction	141	284	12	215
20k B3 prediction	208	293	12	225

Table 13: Counts over the same 300 positions. No upstream board is repaired.

B3 improves the legal-action count over B2 but not teacher hits; their paired Hit@5 difference has interval $[-1.0, 1.0]$ points. The fixed policy itself is weak on true matrices (13/300 versus 41/300 with images). This limits sensitivity to state improvements. The follow-up tests an explicit interface after observing earlier results; it is neither independent confirmation nor an intervention on an internal model representation.

B.4 Prompt and processor diagnostics

Twelve additional development games exclude the benchmark and training game/D4 states. Each 208-query native diagnostic comprises 96 original-prompt, 84 example-free, 24 image-shuffled and four split-processor queries. Qwen3-VL-4B and InternVL3.5-4B produce identical tensors and answers across the split-processor checks. This compares two routes within one implementation, not independent inference engines. Removing examples changes 83/84 and 76/84 parsed answers, respectively, but produces no image teacher hits in either 12-game base-model panel.

Model	Board exact	Teacher hit	No-example hit	Shuffle changed	Gain/loss
Base	0/12	0/12	0/12	0/12	0/0
20k B1-E	0/12	1/12	0/12	12/12	0/1
20k B1-U	0/12	1/12	1/12	12/12	0/0
20k B1-T	0/12	1/12	1/12	12/12	0/0
20k B2	5/12	0/12	0/12	0/12	0/0
20k B3	9/12	1/12	1/12	11/12	1/1

Table 14: Base and all five 20k adapters on the same 12-game diagnostics. Gain/loss compares no-example and original teacher hits. Shuffle changed counts answer changes, not correct decisions on a legitimate input.

For B3, removing examples preserves 1/12 teacher hits through one gain and one regression. Its responses change on 11/12 shuffled image-policy inputs. Input sensitivity alone therefore does not show correct use of the state.

B.5 Static output contract controls

A frozen follow-up evaluates the base, 20k B3 and 20k B1-T checkpoints on the same 300 games, using images, true matrices and SGF prefixes. Each input is queried under three contracts: the original four-field T4 JSON, a sole `best_move` field retaining the D4 example, and a sole `best_move` field without an example. Only the final task instruction changes; all arms use greedy decoding and the same 512-token limit, with no decoding constraint. All 8,100 static responses are generated afresh. Strict Hit@5 requires the corresponding full contract. A separate diagnostic extracts

`best_move` from parseable original-contract responses even when their other fields fail validation; it does not replace strict scoring.

Model	Input	Full T4 (Valid, Hit@5)	Move + example (Valid, Hit@5)	Move only (Valid, Hit@5)	Full best field Hit@5
Base	Image	(300, 4)	(300, 4)	(293, 14)	4
Base	Matrix	(300, 4)	(300, 4)	(300, 10)	4
Base	SGF	(300, 4)	(300, 4)	(96, 0)	4
20k mixed	Image	(295, 41)	(300, 38)	(300, 26)	41
20k mixed	Matrix	(298, 30)	(300, 7)	(300, 26)	30
20k mixed	SGF	(299, 23)	(300, 6)	(300, 19)	23
20k engine token	Image	(300, 41)	(300, 34)	(300, 35)	41
20k engine token	Matrix	(300, 15)	(300, 8)	(300, 18)	15
20k engine token	SGF	(291, 2)	(285, 2)	(11, 0)	2

Table 15: Output contract controls; all counts use 300 games. "Move only" removes the example; the other single-field arm retains D4. The last column is a permissive diagnostic on the original T4 responses.

The full-T4 controls match all 900 historical response strings for each of the base and mixed models. For B1-T, only 89/900 complete responses match; the best-move fields agree on 278/300 image, 243/300 matrix and 280/300 SGF queries. Prompt and adapter fingerprints match, but this run uses H100 while the original uses H200. We do not identify the cause of every difference. All output-contract contrasts use the fresh controls on the same device; the original tables are retained as their own frozen experiment.

These are protocol sensitivity tests on development-exposed games, not independent confirmation. Output contract and example removal are distinct interventions, and neither isolates an internal reasoning mechanism. Paired intervals resample the 300 source games 10,000 times with seed 20260926; they do not capture training-seed uncertainty or adjust for multiplicity.

C State Tracking and Closed-Loop Evaluation

C.1 Tracking protocol and the complete model panel

Six disjoint games provide 11 observations each and every intervening SGF move. Each model makes 66 fresh-image predictions and 60 updates in each of four conditions: neural carry from its first image prediction, or true state resets every one, five or ten transitions. All 306 responses are available for B0 and the ten 8k/20k adapters (3,366 in total). Gaps span 1–17 supplied plies. Changed-cell accuracy uses the difference between consecutive true states, without repairing the model's prediction. Every invalid state scores zero. Only three of the 233 changed intersections are captured stones, so capture coverage is limited.

The rule control executes placements, captures and passes in the predicted initial occupancy. Invalid initial schemas, nonviable groups or placements conflicting with the ledger cause permanent failure for that trajectory; later observations remain in the denominator. B2/B3 at 20k initialize all six trajectories correctly and therefore reach 60/60 exact states and 233/233 changed points. Oracle-initialized rule replay yields the same counts as an implementation check, not a learned-model result.

For 20k B3, all 60 neural-carry and all 60 oracle-step outputs parse as JSON, but only 2 and 4 satisfy the board schema. Both conditions have zero exact states; oracle-step changed-cell correctness is 1/233. By contrast, B0 copies the previous board on all 60 oracle-step updates, yielding $1 - 233/(60 \times 361) = 98.92\%$ cell accuracy with no correct changed points. These are different observable failures, not evidence that the same internal mechanism fails in both models.

C.2 Separating board serialization from content

The same three checkpoints used in Appendix B.5 each receive the true previous board and supplied moves for all 60 transitions. A fresh raw arm is paired with a syntax-constrained arm using identical prompts, input tensors, greedy decoding, EOS and a 1,024-token limit. The constraint permits

any canonical sole-key JSON containing 19 strings of 19 B/W/. characters, including impossible occupancies. It has no access to the current reference board, legal actions or teacher labels.

Model	Decoder	Valid /60	Exact /60	Changed /233	Cell (%)
Base	raw	60	0	0	98.92
Base	syntax	60	0	0	98.92
20k mixed	raw	4	0	1	6.37
20k mixed	syntax	60	0	11	92.64
20k engine token	raw	59	0	0	97.27
20k engine token	syntax	60	0	0	98.90

Table 16: Matched board syntax control with a true-state reset at every transition. Invalid boards receive zero cell credit. All 360 first responses are retained; no answer is retried on the basis of correctness.

The constraint makes all 60 mixed-model outputs schema-valid, versus four in the raw arm, but exact updates remain 0/60 and changed-point correctness is 11/233 versus 1/233. The base model copies all 60 previous boards under both decoders; B1-T copies 59 under each. Thus malformed serialization does not fully explain the retained update failures. The constraint changes the decoder and does not identify the cause of the remaining content errors. Intervals resample the six source games, not 60 independent transitions. No model is retrained or selected; all three use seed-42 checkpoints where applicable. Base and mixed runs use H200, and B1-T uses H100, with identical hardware within each paired comparison; latency is not compared across them.

C.3 Common opponent and direct matches

B0, 20k B3 and 20k B1-T each face 64-visit KataGo on five fixed starts, both colors and four access modes (120 episodes). Direct B0–B3 matches use the same starts, colors and symmetric modes (40 matches). All conditions use Tromp–Taylor rules and komi 7.5.

Vision. A fresh image for each decision.

Legality feedback. The same image and one correction opportunity after a malformed or illegal proposal; no legal-move list or engine advice.

Rule-maintained state. A T1 initial-image reconstruction followed by deterministic occupancy updates. Later decisions receive the ledger matrix. Invalid initialization or inconsistent updates terminate the episode.

Oracle state. The true current matrix at every decision, without an image; this is a privileged-access control.

The prompt includes role, last move and accepted actions since the start. The adjudicator retains the preceding full SGF history for superko but does not expose it as extra model context. Models return the T4 object; all fields must pass schema validation before `best_move` is played. Other candidates never replace an illegal proposal. The dynamic renderer uses 768-pixel wood boards, 7.5% margins and a red hollow-square marker, so static-to-gameplay comparisons change both rendering and prompt context.

Budget	Common opponent	Direct match	Action repair
Accepted plies	120	120	256
Model calls	64	64 per player	128
Generated tokens	32,768	32,768 per player	8,192
Tokens per decision	512	512	64
Wall time (seconds)	600	600	1,200

Table 17: Frozen episode limits. Common/direct ledger initialization allows 1,024 tokens within the same episode budget. All model, engine and rendering time counts toward wall time; late actions are not accepted. Repair uses a separate move-only prompt and fresh matched controls.

Common-opponent policies use H100; both direct-match policies share H200. Latency is not pooled across these experiments. Only two consecutive passes complete a scored game. Invalid actions, initialization failures, inconsistent ledgers and budget exhaustion remain incomplete, not losses. Every accepted move and termination is independently replayed and checked against saved responses, prompts and checkpoint identities.

	Base			20k mixed			20k engine-only		
Mode	Complete	Plies	Invalid	Complete	Plies	Invalid	Complete	Plies	Invalid
Vision	0/10	1.7	10/16	0/10	1.5	10/15	0/10	1.1	10/13
Legality feedback	0/10	1.7	20/26	0/10	2.5	23/33	0/10	5.5	21/46
Rule-maintained state	0/10	0.0	0/0	0/10	1.3	3/9	0/10	0.0	0/0
Oracle state	0/10	4.5	10/30	0/10	2.5	10/20	0/10	4.1	10/28

Table 18: Common-opponent results, ten related episodes per model/mode. Plies count accepted actions by both players. Invalid counts include corrections; 0/0 means initialization failed before a proposal.

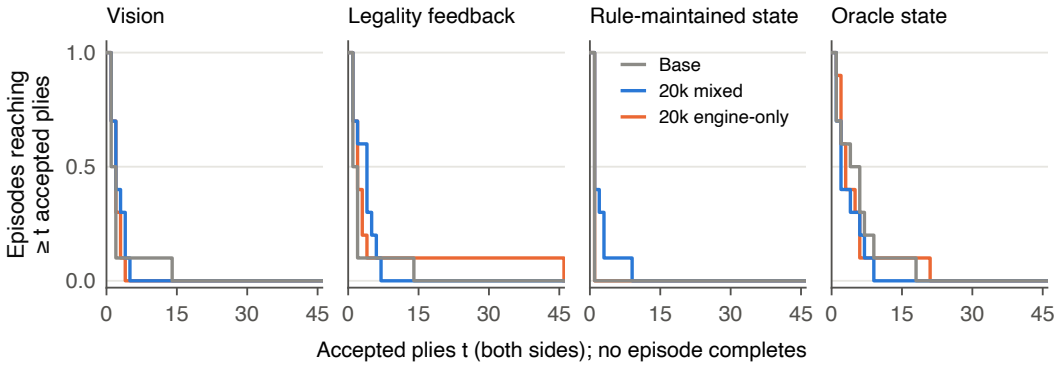


Figure 10: Episode survival in the common-opponent evaluation (Table 18): fraction of the ten episodes per model and mode that reach at least t accepted plies. Most episodes end within ten plies in every mode, including oracle state; none completes.

There are 0/120 common-opponent completions: 92 episodes end on invalid final actions, 26 on invalid initialization, one on ledger inconsistency and one on wall time. All 40 direct matches are also incomplete, providing no scored wins or win-rate estimate. The five starts are the independent units, not the repeated modes and color swaps.

C.4 Separating output syntax from action legality

Base and mixed models use a fresh sole-key `move` JSON prompt on the same five starts and two colors, with either an image or oracle state. Three matched arms use raw greedy decoding, a token trie over all 361 coordinates plus PASS, or the same trie restricted to full-history legal actions. The last arm supplies rule assistance without engine rankings. Its legality is guaranteed by the tool and is not learned rule competence.

All 120 repair episodes remain incomplete. The 40 legal-mask episodes have no illegal proposals and reach the time cap; no arm produces an accepted PASS action, and the harness forces none. Legal masking lengthens execution but does not establish stronger play (Figure 8). Recognition under the two renderers is tested separately in Appendix D.

Model	State	Constraint	Complete	Mean plies	Invalid/proposed	Wins/scored
Base	vision	raw	0/10	3.3	10/24	–
Base	vision	syntax	0/10	4.1	10/28	–
Base	vision	legal	0/10	103.4	0/519	–
Base	oracle	raw	0/10	3.9	10/27	–
Base	oracle	syntax	0/10	4.3	10/29	–
Base	oracle	legal	0/10	103.5	0/520	–
20k mixed	vision	raw	0/10	1.9	10/17	–
20k mixed	vision	syntax	0/10	2.5	10/20	–
20k mixed	vision	legal	0/10	102.7	0/516	–
20k mixed	oracle	raw	0/10	2.5	10/20	–
20k mixed	oracle	syntax	0/10	2.7	10/21	–
20k mixed	oracle	legal	0/10	103.3	0/519	–

Table 19: **Only the legal-move trie keeps games going.** Matched action controls, ten episodes per row. All trajectories are replay-verified. Unscored rows have no win-rate estimate.

D Visual Stress Tests

D.1 Marker shape and rendering perturbations

On 120 training-disjoint development games, the base, 20k mixed adapter and InternVL-4B receive paired full-board images with absent, dot, square or triangle markers. Geometry, stones and labels are held fixed. The shared prompt specifies each shape and asks for presence and coordinate; absence requires `false/null`, not PASS. This is a new schema and does not replace the original marker-localization task.

Model	Absent	Dot	Square	Triangle
Base Qwen	108/120	4/120	2/120	6/120
20k mixed	120/120	1/120	1/120	4/120
InternVL-4B	37/120	2/120	5/120	3/120

Table 20: Joint marker presence/localization correctness out of 120 games per condition, including invalid answers as failures.

A second panel uses 12 development games and 132 jointly transformed images, with the original T1/T2/T4 prompts. D4 transforms apply to both image and reference coordinates. Seven nonidentity transforms, resolution changes and monochrome rendering test transfer; inverse-output agreement is distinct from correctness. Native preprocessing can change the effective resolution, and repeated variants do not add independent games.

D.2 Source versus dynamic renderer

Base and mixed models reconstruct source and dynamic images on all 120 development games. Both styles are reacquired within each model's job with identical prompts, text tensors, image-token grids and 1,024-token output limits. This comparison changes rendering as a bundle; it does not isolate the effects of marker style, margins or individual graphic parameters.

The mixed model's 85/120 versus 0/120 exact-board counts show a substantial interface-transfer failure. These are paired development probes, not new hidden test games or an identified attention mechanism.

Model	Variant	Board EM	Marker acc.	Teacher Hit@5
Base Qwen	Control	0.0	25.0	0.0
Base Qwen	D4 mean	0.0	20.2	0.0
Base Qwen	384 px	0.0	8.3	0.0
Base Qwen	1024 px	0.0	33.3	0.0
Base Qwen	Monochrome	0.0	33.3	0.0
20k mixed	Control	0.0	0.0	0.0
20k mixed	D4 mean	1.2	0.0	6.0
20k mixed	384 px	0.0	0.0	16.7
20k mixed	1024 px	0.0	0.0	8.3
20k mixed	Monochrome	0.0	0.0	8.3
InternVL-4B	Control	0.0	8.3	0.0
InternVL-4B	D4 mean	0.0	7.1	1.2
InternVL-4B	384 px	0.0	0.0	0.0
InternVL-4B	1024 px	0.0	0.0	0.0
InternVL-4B	Monochrome	0.0	8.3	0.0

Table 21: Accuracy (%) on 12 paired games. The control is a newly rendered 768-pixel board, not an original-source image. D4 is the mean over seven nonidentity transforms. Invalid outputs remain in every denominator.

Model	Renderer	Schema valid	Exact board	Cell accuracy (%)
Base	source	20/120	0/120	4.29
Base	dynamic	17/120	0/120	3.50
20k mixed	source	118/120	85/120	98.04
20k mixed	dynamic	91/120	0/120	66.38

Table 22: Matched recognition on 120 games. Invalid boards receive zero cell credit. The mixed source-minus-dynamic BoardEM difference is 70.83 percentage points, with descriptive 95% interval [62.50,79.17].