V-VLAPS: Value-Guided Planning for Vision-Language-Action Models

Anonymous Authors[1]

Abstract

Vision-language-action (VLA) models provide strong action priors for robotic manipulation, but their reactive behavior can fail under distribution shift and long-horizon task structure. Recent VLA-guided planning methods improve execution by using pretrained policies to guide tree search, yet node selection still depends heavily on policy priors and visit-count exploration. Consequently, when the policy favors poor actions, the planner lacks a learned value signal to correct this bias. Prior work has shown that VLA representations encode rollout success and failure information, suggesting that they may also support value estimation during planning. We introduce Value-Guided Vision-Language-Action Planning and Search (V-VLAPS), which augments VLA-guided planning with a lightweight value head trained on offline VLA rollouts to predict Monte Carlo returns. These predictions guide Monte Carlo Tree Search toward higher-value branches. Across five LIBERO suites, V-VLAPS matches value-free planning baseline at the default search budget in aggregate, and analysis shows that many hard failures are root-level timeouts where predicted values are weakly separated. With a larger search budget, V-VLAPS improves over the baseline in all task suites with $+6$ percentage points on LIBERO-Object and $+4$ percentage points on LIBERO-10. Our results suggest that VLA representations can support not only failure prediction, but also value-guided planning when search reaches branches where value-based ranking matters.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Deploying robotic policies in open-world settings requires reliability under distribution shift. Recent advances in robot learning have been driven by large-scale Vision-Language-Action (VLA) models, transformer policies that predict action sequences from multimodal observations and language instructions. While VLA models serve as effective priors for generalist behaviors, they are fundamentally limited by their reliance on behavior cloning. As a result, they often exhibit brittle behavior when facing out-of-distribution (OOD) states.

A promising approach to address this problem is to augment the pretrained model with a planning search algorithm that explores possible outcomes in simulation. One such search algorithm is Monte Carlo Tree Search (MCTS), which constructs a search tree by simulating rollout trajectories, balancing exploration of uncertain actions with exploitation of high-likelihood ones (Świechowski et al., 2022). Following Neary et al. (2025), we can leverage a VLA model as a policy prior to guide the MCTS, using a visit-count heuristic to manage exploration. However, we observe that relying solely on the VLA prior is insufficient for robust long-horizon planning. In the formulation proposed by Neary et al. (2025), the search lacks a value function, meaning it has no grounded estimate of the expected future return. Consequently, if the VLA prior is inaccurate, assigning high probability to suboptimal actions, the planner has no mechanism to correct this bias other than exhaustive count-based exploration. In other words, the search relies on the imitation prior rather than the underlying reward structure.

The signal the search lacks may already live in the VLA's own internal features. Gu et al. (2025) show that a small probe trained on the latent features of a generalist VLA predicts task success or failure with strong cross-task generalization. If a binary success signal generalizes across tasks, the same features may also carry information useful for estimating continuous value. The natural use of such a signal in our setting is not post-hoc monitoring but active search guidance: feeding the prediction into the MCTS scoring rule to bias node selection. In this sense, our setting is an offline-to-online decision-making problem: offline rollouts of a frozen VLA are used to learn a value signal, and that signal is then used at test time to guide online MCTS

planning.

We introduce V-VLAPS, which extends the VLA-feature-probe approach of Gu et al. (2025) from passive failure detection to active value-guided planning. A three-layer MLP trained on Octo readouts predicts Monte Carlo discounted returns, and we add the prediction as the Q term in VLAPS's PUCT scoring rule. We evaluate V-VLAPS on five LIBERO suites against the plain Octo VLA and against VLAPS without a learned value, at two per-episode search budgets (600 s and 1800 s). On these suites both MCTS-based conditions outperform the reactive Octo VLA by about 27 percentage points on average, reproducing the result of Neary et al. (2025); the value head is therefore evaluated as an addition to an already-strong planning baseline.

At the 600 s budget, V-VLAPS matches VLAPS in aggregate (both at 87.4%), and a failure-mode analysis shows that many hard failures are MCTS timeouts at the root rather than clear value-head errors. Near the root, candidate branches are still far from task completion, so the value head often assigns similarly small values to many alternatives and has little discriminative power. With the budget extended to 1800 s, search can reach deeper states where branch values separate more clearly. In this regime, V-VLAPS exceeds VLAPS by $+6$pp on libero_object and $+4$pp on libero_10. These results suggest that SAFE-style probes over VLA features can act not only as monitors but also as learned planning components when search reaches states where value-based branch ranking matters.

2. Related Work

2.1. Vision–Language–Action Policies

Vision–language–action (VLA) models are designed to map visual observations and natural-language instructions directly to robot actions. Early systems showed that a single transformer policy can be trained on large collections of robot demonstrations and language-labelled tasks, and then be reused across many manipulation skills (Brohan et al., 2023b;a). In our work, we use Octo (Team et al., 2024) as a fixed generalist VLA backbone and build our method on top of its latent representation; the same method can be used with other VLA models.

Although VLA models can perform a wide range of tasks, they are typically used in a purely reactive way: at each step, the model receives the current observation and instruction and outputs the next action (or action chunk), without explicit long-horizon planning. In challenging scenes or out-of-distribution configurations, this can lead to failures that the policy does not recover from (Gu et al., 2025). In the language modeling paradigm, many methods have been introduced to mitigate similar issues by scaling test-time compute, such as chain-of-thought prompting (Wei et al., 2023), self-consistency sampling (Wang et al., 2023), and MCTS during inference (Hao et al., 2023).

Neary et al. (2025) extend these ideas to the robotics domain to address the limitations of reactive execution. They introduced Vision-Language-Action Planning and Search (VLAPS), which embeds a pre-trained VLA into a model-based search procedure and uses the VLA to define action proposals and abstractions for planning. In our work, we build on VLAPS and keep the same fixed VLA backbone, but additionally learn a value function over its latent state and incorporate the resulting value predictions into VLAPS's scoring rule to guide the search.

2.2. Monte Carlo Tree Search

A natural way to compensate for the myopic behaviour of reactive policies is to add a search procedure that can simulate possible future outcomes before committing to an action. Monte Carlo Tree Search (MCTS) is a widely used algorithm for this purpose. It incrementally builds a search tree by repeatedly simulating trajectories from the current state, and at each iteration it selects actions that balance exploring uncertain branches with exploiting branches that have produced good returns in previous simulations (Świechowski et al., 2022). Typically, each iteration proceeds through four phases: selection, expansion, simulation, and backpropagation. The resulting statistics over node visit counts and returns are then used to estimate action values at the root.

MCTS has been particularly successful in domains where a reasonably accurate simulator is available. In board games such as Go, Chess, and Shogi, MCTS combined with learned policies and values has led to systems that reach and surpass human expert performance (Silver et al., 2016; 2017a). In these settings, the search runs entirely in simulation and only the final chosen move is executed in the real environment. Neary et al. (2025) bring this style of planning to vision–language–action models in the VLAPS framework. They use MCTS over discrete action chunks, with a pre-trained VLA policy providing action priors that focus and refine the search of an otherwise completely intractable space. Their formulation combines their VLA-informed prior with visit-count-based exploration heuristics in a PUCT-style scoring rule which allows MCTS to exploit the strengths of the pre-trained policy while still exploring alternative action sequences.

2.3. Learned Value Functions for Tree Search and VLA Planning

Combining learned value functions with tree search has proved highly effective in domains where planning is possible. In AlphaGo and its successors, deep neural networks predict both a policy over moves and a scalar value estimate from a board position; Monte Carlo Tree Search uses the

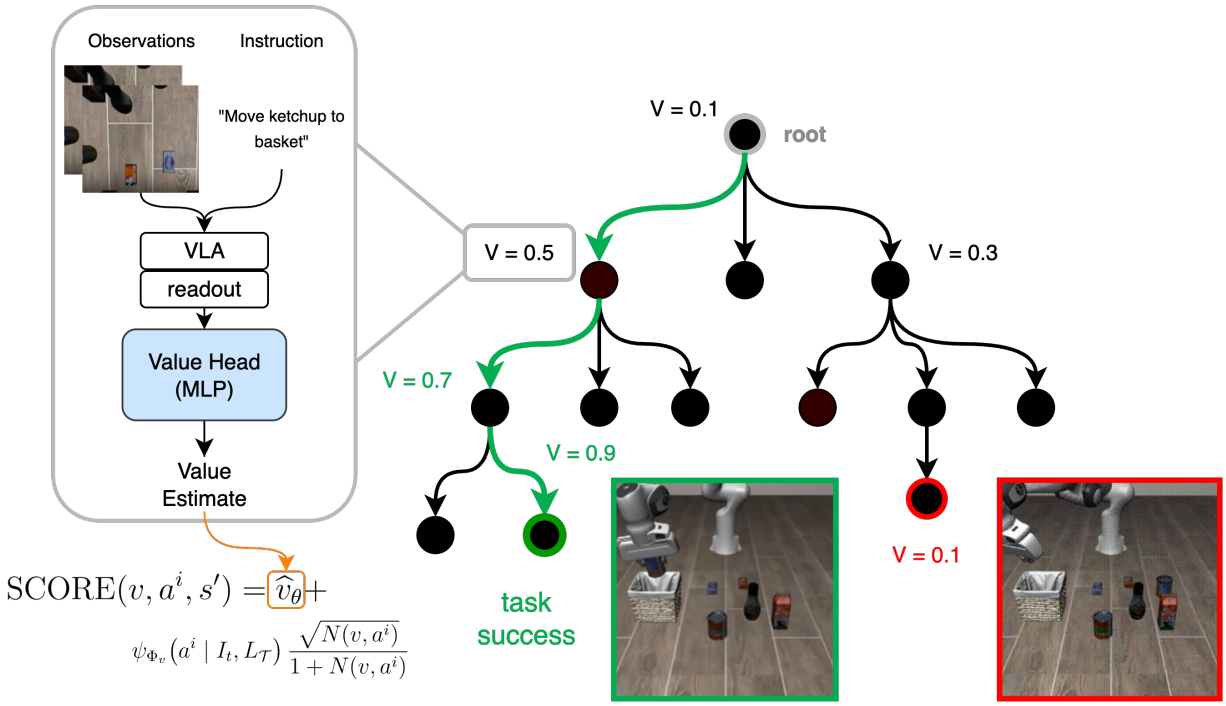

Figure 1. Overview of our value-guided VLAPS extension. At each MCTS node, the current observations and language instruction (e.g., "Move ketchup to basket") are passed through the frozen VLA backbone and our value head (MLP) to produce a scalar value estimate. This value is attached to the corresponding node and used in the VLAPS scoring rule to bias node selection. Nodes with higher predicted value (green) are selected more often and tend to lead to successful task completions, while low-value nodes (red) are down-weighted during search.

policy to prioritize promising actions and the value to evaluate leaf nodes without long rollouts (Silver et al., 2016; 2017b;a). This combination allows the search procedure to concentrate computation on moves that are both likely under the learned policy and lead to positions that the value function predicts as strong, rather than relying solely on visit counts and simulation outcomes of intractably long rollouts.

VLAPS adopts a similar strategy for vision–language–action policies, but stops short of learning an explicit value function over the VLA latent state. In Neary et al. (2025), the quality of a node is determined by search statistics and VLA-derived action priors, but there is no separate learned estimate of the state's value. In contrast, we keep the underlying VLA and search procedure fixed, and introduce a small value head which takes the VLA latent state as input. This head is trained to predict Monte Carlo returns from VLA rollouts, and its predictions are incorporated into the VLAPS selection rule. Conceptually, this brings the use of value functions in VLAPS closer to the value-guided tree search used in game-playing systems like AlphaGo, while operating in the setting of vision–language–action control.

More directly related is SAFE (Gu et al., 2025), which trains a small probe (an MLP or LSTM) on the VLA's frozen latent features and uses the resulting success/failure prediction as a post-hoc monitor. Our value head adopts the same feature source and probe family but predicts a continuous Monte Carlo discounted return, and the prediction enters the PUCT scoring rule during search rather than after the episode.

3. Method

We extend VLAPS by learning a value function over Octo latent states and using this value to guide Monte Carlo Tree Search. We first collect rollouts of the Octo model and compute Monte Carlo value targets for each state (Section 3.1). We then train a three-layer MLP (the "value head") to predict these targets from Octo's last-layer representation (readout) (Section 3.2). Finally, we integrate the learned value into VLAPS' tree search by modifying the node selection score to bias branches towards high-value nodes (Section 3.3).

3.1. Data Collection

We construct training data for the value head by rolling out a fixed pre-trained VLA policy, without any planning, on LIBERO (Liu et al., 2023) tabletop manipulation tasks. For each suite, we collect value-head training rollouts on all tasks but restrict each task to four of the ten available

initial states: indices 0, 3, 6, and 9. Evaluation uses all ten initial states per task (Section 4), so six initial states per task are unseen during value-head data collection. This gives a partial initial-state generalization test while preserving comparability with the full ten-initial-state evaluation used for all methods.

At the beginning of each episode, the environment provides an initial observation o_0 and a language instruction g. At each decision step t, the VLA (Octo (Team et al., 2024)) receives (o_t, g), computes a latent readout vector $h_t \in \mathbb{R}^d$ which acts as the summary representation of the past observations and the task. It then uses the readout to output an action chunk c_t, which consists of a sequence of low-level actions. We execute the entire chunk c_t in the environment, applying each low-level action in order, until the chunk finishes or the episode terminates. The environment then returns the next observation o_{t+1}, and we query the VLA again. This produces an episode as a sequence of decision steps $(o_0, c_0, o_1, c_1, \ldots, o_T)$.

Episodes terminate either when the task is successfully completed or when a timeout horizon is reached. We use a sparse terminal reward: if the task is successfully completed before the timeout, the episode receives a terminal reward of 1; otherwise (failure or timeout), it receives a terminal reward of 0. All intermediate rewards are zero, so each episode has a binary success/failure outcome.

For each decision step t in an episode, we define a Monte Carlo value target by propagating the terminal reward backward through time. Let $R \in \{0, 1\}$ denote the terminal reward and T the index of the final decision step. We assign each state at step t a target

$$G_t = \begin{cases} \gamma^{T-t}, & \text{if the episode ends in success } (R = 1), \\ 0, & \text{if the episode ends in failure } (R = 0). \end{cases}$$

where $\gamma \in (0, 1]$ is a discount factor; we use $\gamma = 0.99$. This target is a Monte Carlo estimate of the state value under our sparse success reward when following the VLA policy.

We pair each readout with its value target to form a training example (h_t, G_t), and aggregate these across many rollouts to build a per-suite dataset of state–value pairs. The raw target distribution is heavily skewed toward zero: failed episodes contribute only zero-valued targets, and even successful episodes have many early-step targets close to zero because γ^{T-t} is small when $T - t$ is large. A value head trained on this distribution can collapse to predicting zero almost everywhere, so we rebalance once during preprocessing: targets are bucketed into ten equal-width bins on $[0, 1]$, and the bottom bin is downsampled to match the combined size of the other nine bins.

3.2. Value Head Training

Our goal is to learn a value function that maps the VLA's latent representation at each decision step to the scalar value target defined in Section 3.1. Concretely, given the Octo readout vector $h_t \in \mathbb{R}^d$ at decision step t, we learn a function $V_\theta : \mathbb{R}^d \to \mathbb{R}$ that predicts a scalar estimate $\hat{v}_t$ of the state value: $\hat{v}_t = V_\theta(h_t)$.

We parameterize V_θ as a lightweight three-layer multilayer perceptron (MLP). It has a single hidden layer of 3072 ReLU units between a 768-dimensional input and a scalar output, for about 2.4M parameters in total. The architecture is intentionally small compared to the underlying Octo model, so that evaluating the value head adds minimal overhead during planning.

We train this value head on the dataset of readout–value pairs $\{(h_t, G_t)\}$. The training objective is to regress onto the Monte Carlo value targets G_t using mean squared error: $\mathcal{L}(\theta) = \mathbb{E}_{(h_t, G_t)}\big[(V_\theta(h_t) - G_t)^2\big]$.

In practice, we approximate this expectation by sampling mini-batches from the dataset and performing stochastic gradient descent with the Adam optimizer. During training, the parameters of the underlying Octo VLA are kept fixed; only the value head parameters θ are updated. After convergence, V_θ provides a compact estimate of the state value at each decision step, which we use to guide the tree search.

3.3. Value Integration

Motivated by the value-plus-prior structure of PUCT-style search, we integrate the learned value head into the VLAPS tree-search rule while leaving the rest of the VLAPS search procedure unchanged. VLAPS already provides the prior/exploration term through its VLA-guided action prior and visit statistics, and V-VLAPS adds the missing learned value term. At a node v, VLAPS samples a finite set of candidate action chunks $\{a^i\}$ from the VLA-induced action distribution and scores them using the VLAPS prior/exploration term

$$U_{\text{VLAPS}}(v, a^i) = \psi_{\Phi_v}\big(a^i \mid I_t, L_{\mathcal{T}}\big)\, \frac{\sqrt{N(v, a^i)}}{1 + N(v, a^i)}, \quad (1)$$

where I_t is the current observation, $L_{\mathcal{T}}$ is the task instruction, ψ_{Φ_v} is the VLAPS prior over the sampled candidate chunks, and $N(v, a^i)$ is the number of times candidate a^i has been selected from node v. We use this term exactly as in VLAPS.

V-VLAPS adds a learned successor-state value to this inherited score. For each candidate action chunk a^i, the simulator rolls the chunk forward from node v to obtain a successor state s'. We compute the Octo readout $h(s')$ for this succes-

sor state and evaluate the value head:

$$\widehat{v}_\theta(s') = V_\theta(h(s')). \tag{2}$$

We use this successor-state value as the value term for the node-action pair,

$$Q_\theta(v, a^i) = \widehat{v}_\theta(s'). \tag{3}$$

Because intermediate rewards are zero and the simulated transition for a fixed action chunk is deterministic in our LIBERO setup, the successor-state value estimates the continuation value after choosing a^i.

The final V-VLAPS selection score is

$$\mathrm{SCORE}(v, a^i) = \lambda_V Q_\theta(v, a^i) + U_{\mathrm{VLAPS}}(v, a^i). \tag{4}$$

We use a fixed value coefficient $\lambda_V = 1$ in all experiments.

Therefore, V-VLAPS differs from VLAPS only by adding the learned value term Q_θ with a fixed coefficient. The action sampler, action library, LIBERO simulator, and MCTS hyperparameters are kept fixed across the VLAPS and V-VLAPS conditions.

4. Experiments

We evaluate three conditions on five LIBERO suites: the plain Octo VLA without planning (*VLA*), VLAPS without a learned value (*VLAPS*, the formulation of Neary et al. (2025)), and V-VLAPS, which adds the learned value head from Section 3.3 to the VLAPS scoring rule. The five suites cover different task types: *libero_object* tests object identification with pick-and-place, *libero_spatial* tests spatial relations between objects and target regions, *libero_goal* tests goal-state specification, *libero_10* is ten long-horizon tasks, and *libero_90* is ninety more diverse manipulation tasks (Liu et al., 2023). The value head is trained once from offline VLA rollouts for each suite and then frozen during evaluation. All test-time adaptation comes from MCTS planning rather than parameter updates.

Each (condition, suite, budget) cell is evaluated on 100 episodes. For the four 10-task suites this is the full 10×10 grid of tasks and initial states, and for libero_90 we sample 10 tasks and use all 10 initial states per sampled task. As described in Section 3.1, the value head is trained using rollouts from only four initial states per task. The reported evaluation therefore tests both performance on the value-head data-collection states and generalization to six initial states per task that are unseen during value-head training. MCTS uses two per-episode wall-time budgets, 600 s by default and 1800 s for the extended-budget experiments. The MCTS hyperparameters are held fixed across all conditions and suites; the full configuration is in Section B.

We train one value head per suite, with a $90/10$ train/validation split (fixed seed). Per-suite balanced training-data sizes range from 7,830 examples (libero_object) to 18,792 (libero_spatial); the full breakdown is in Section A. These training sets are small relative to the value head's roughly 2.4M parameters.

5. Results

We separate the benefit of planning from the additional benefit of value guidance. Because value predictions are often weakly separated near the root but become more useful once MCTS reaches deeper branches, we analyze both the default and extended wall-time budgets.

5.1. Main results

Table 1 reports aggregate success rates for the three conditions on each of the five suites at both per-episode wall-time budgets. At 600 s, V-VLAPS and VLAPS both average 87.4% across the five suites. No per-suite difference exceeds 3pp, which sits inside the binomial standard error of $\approx$ 3pp at $p \approx 0.9$. At 1800 s the average pulls apart to 91.6% vs 88.4%, a $+3.2$pp gap. The gap is concentrated on libero_object ($+6$) and libero_10 ($+4$); the other three suites already sit at $\geq 88\%$ at 600 s in every condition and move by at most $\pm$2pp at the larger budget. Because each cell contains 100 episodes, we interpret 1–3pp differences as suggestive rather than conclusive and focus our claims on the larger gaps under the extended budget.

5.2. Wall-time scaling and a task-7 case study

The libero_object and libero_10 gains both scale with budget, but unequally. On libero_object, V-VLAPS goes from 85 at 600 s to 93 at 1800 s ($+8$pp), while VLAPS goes from 82 to 87 ($+5$pp). The V-VLAPS–VLAPS gap doubles from $+3$ to $+6$. On libero_10 the asymmetry is larger: V-VLAPS gains $+10$pp ($75 \rightarrow 85$) and VLAPS gains $+4$pp ($77 \rightarrow 81$), and the per-suite gap flips from -2 to $+4$.

This budget asymmetry is sharpest on a single hard task. To tighten the comparison on libero_object task 7, we extended the default ten initial states to twenty (indices 0–19) at 1800 s. V-VLAPS reaches $17/20$ (85%) and VLAPS reaches $13/20$ (65%). The $N = 10$ data (Appendix C) underline the same picture from the other direction: on task 7, VLAPS goes $8/10 \rightarrow 6/10$ from the larger budget while V-VLAPS goes $5/10 \rightarrow 10/10$. Plain MCTS does not get more out of the extra search time on this task; only the value-guided variant does.

libero_spatial, libero_goal, and libero_90 are already at or above 88% at 600 s, and neither the larger budget nor the

Table 1. Success rate (%) by condition and suite at the default (600 s) and extended (1800 s) per-episode search budgets. Each cell aggregates 100 episodes. **VLA** (no planning) provides context for the regime; **VLAPS** (Neary et al., 2025) is the prior-work planning baseline; **V-VLAPS** is our extension. Bold V-VLAPS cells highlight the 1800 s wall-time gains on the headroom suites (libero_object, libero_10).

		600 s budget		1800 s budget	
Suite	**VLA**	**VLAPS**	**V-VLAPS**	**VLAPS**	**V-VLAPS**
libero_object	37	82	85	87	**93**
libero_spatial	81	96	95	96	**97**
libero_goal	88	92	**93**	90	**93**
libero_10	38	77	75	81	**85**
libero_90	57	**90**	89	88	**90**
Avg.	60.2	87.4	87.4	88.4	**91.6**

value head moves them by more than $\pm$2pp.[1]

5.3. Failure-mode analysis

On libero_object tasks 6, 7, and 8 at 600 s, most failed episodes are root-level timeouts (Table 2). In the V-VLAPS condition, 13 of the 15 failed episodes contain exactly one value-head query, made at the first decision step, with wall-clocks between 605 and 611 s, just past the cap. The matched VLAPS condition shows the same signature in 14 of 18 failures. These episodes spend the entire budget expanding alternatives at the first decision and never commit to an action chunk.

Table 2. Failure modes on libero_object tasks 6–8 at the 600 s budget. We define a root-level timeout as a failed episode in which the search remains at the first decision step until the wall-time cap is reached.

Condition	**Failures**	**Root timeouts**	**Fraction**
VLAPS	18	14	77.8%
V-VLAPS	15	13	86.7%

On the timeout runs themselves the value head reports values between -0.01 and 0.02. This small range is expected for first-decision states in hard task configurations, because most candidate branches remain far from task completion, so their predicted continuation values are near zero and weakly separated. In such near-root states, the value term has little ability to distinguish among alternatives. The value head is expected to become more useful once the search reaches deeper states where different branches have more separated continuation values. This suggests that the 600 s plateau in Section 5.1 is driven largely by the search budget: in these runs, MCTS often does not reach the branches where value-based ranking could affect selection.

[1] VLAPS drops slightly on libero_goal (92 $\rightarrow$ 90) and libero_90 (90 $\rightarrow$ 88) at 1800 s. Both moves are within sampling noise.

6. Discussion and Conclusion

V-VLAPS gains over VLAPS in a specific regime: when MCTS has time to move beyond near-root states, and when the task has remaining headroom for value-guided ranking to matter. Near the root of hard tasks, candidate branches often remain far from completion and receive similarly small value estimates, so the value term provides little separation. The first condition fails at 600 s on the hardest libero_object tasks, where most failures are root-level timeouts (Section 5.3). The second fails on libero_spatial, libero_goal, and libero_90, where the prior pipeline is already near ceiling. The libero_object and libero_10 gains at 1800 s are the regime where both conditions hold: search reaches deeper branches, and those branches have enough performance headroom for the value head's ranking to change decisions.

V-VLAPS shows that the VLA-feature probe Gu et al. (2025) introduced for failure detection extends to active search guidance. With the binary failure target replaced by a Monte Carlo return and the prediction fed into the PUCT scoring rule, the same probe architecture can bias an MCTS search toward higher-return branches rather than only flag failures after the fact. We have demonstrated this for value-guided node ranking; whether the same probe family can drive other components of the planning loop, such as termination or branch pruning, is an open empirical question.

Limitations. All experiments use Octo as the VLA backbone, and we do not test whether the pattern reproduces on π_0 (Black et al., 2024) or RT-2 (Brohan et al., 2023a). The value head is trained on rollouts of the unguided VLA, so its predictions on states the search visits are off-policy with respect to V-VLAPS execution. Per-suite training sets are small, ranging from 7,830 to 18,792 examples (Appendix A), and whether more training data would widen the 1800 s gap or close the 600 s gap is untested. In addition, a controlled ablation with shuffled value targets or a randomly initialized value head would further isolate the contribution of learned value estimates, and we leave this for future work.

Future work. Two directions follow directly from the findings above. Training the value head on V-VLAPS rollouts instead of unguided VLA rollouts would address the off-policy gap, and is a prerequisite for iterative refinement of the value estimate. Using the value head to prune branches during expansion would convert the root-level timeouts of Section 5.3 into search compute spent on the branches the value head ranks highest. Two further directions are more speculative. Replacing the MSE regression objective with the HL-Gauss classification head of Farebrother et al. (2024) is a natural alternative; we attempted it and hit class-imbalance issues that we expect would ease with more per-suite training data. Aggregating the value-head input across multiple transformer layers, as Gu et al. (2025) suggest for SAFE, may expose signal the final readout averages out.

Impact Statement

This paper studies value-guided MCTS planning over a fixed pretrained vision-language-action policy, evaluated in LIBERO simulation only. Our per-episode test-time search budgets are in the hundreds to thousands of seconds, which shifts compute from training to deployment with the associated energy cost. We do not foresee impacts specific to this contribution that are not already raised by the underlying VLA, planning, and failure-detection work this paper builds on.

References

Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., Tanner, J., Vuong, Q., Walling, A., Wang, H., and Zhilinsky, U. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.-W. E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023a. URL https://arxiv.org/abs/2307.15818.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.-H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. Rt-1: Robotics transformer for real-world control at scale, 2023b. URL https://arxiv.org/abs/2212.06817.

Farebrother, J., Orbay, J., Vuong, Q., Taïga, A. A., Chebotar, Y., Xiao, T., Irpan, A., Levine, S., Castro, P. S., Faust, A., Larochelle, H., and Agarwal, R. Stop regressing: Training value functions via classification for scalable deep RL. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2403.03950.

Gu, Q., Ju, Y., Sun, S., Gilitschenski, I., Nishimura, H., Itkina, M., and Shkurti, F. Safe: Multitask failure detection for vision-language-action models, 2025. URL https://arxiv.org/abs/2506.09937.

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model, 2023. URL https://arxiv.org/abs/2305.14992.

Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.03310.

Neary, C., Younis, O. G., Kuramshin, A., Aslan, O., and Berseth, G. Improving pre-trained vision-language-action policies with model-based search, 2025. URL https://arxiv.org/abs/2508.12211.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. Mastering chess and shogi by self-play with a

general reinforcement learning algorithm, 2017a. URL https://arxiv.org/abs/1712.01815.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017b. doi: 10.1038/nature24270.

Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., Luo, J., Tan, Y. L., Chen, L. Y., Sanketi, P., Vuong, Q., Xiao, T., Sadigh, D., Finn, C., and Levine, S. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Świechowski, M., Godlewski, K., Sawicki, B., and Mańdziuk, J. Monte carlo tree search: a review of recent modifications and applications. *Artificial Intelligence Review*, 56(3):2497–2562, July 2022. ISSN 1573-7462. doi: 10.1007/s10462-022-10228-y. URL http://dx.doi.org/10.1007/s10462-022-10228-y.

A. Per-suite training-data sizes

Table 3 reports the per-suite value-head training-data sizes after rebalancing (Section 3.1). These datasets are collected only from initial states 0, 3, 6, and 9 for each task, and the main evaluation in Table 1 uses all ten initial states per task, including six initial states unseen during value-head training. For libero_10, the raw distribution was already balanced (50.0% zero), so its Balanced count equals its Raw count.

Table 3. Per-suite training-data sizes after rebalancing. **Raw** is the count collected from VLA rollouts before rebalancing. **% zero** is the fraction of raw examples with $G_t = 0$. **Balanced** is the count after downsampling the bottom bin to match the combined size of the other nine bins. **Train/Val** are the $90/10$ split sizes.

Suite	Raw	% zero	Balanced	Train	Val
libero_object	60,695	92.8	8,700	7,830	870
libero_spatial	28,580	63.5	20,880	18,792	2,088
libero_goal	34,155	70.3	20,300	18,270	2,030
libero_10	11,020	50.0	11,020	9,918	1,102
libero_90	63,595	86.8	16,810	15,129	1,681

B. Value-head training and MCTS hyperparameters

This appendix lists the hyperparameters used to train the value head (Table 4) and to run MCTS in the experiments of Section 4 (Table 5). All evaluations were run on a single NVIDIA H100 GPU; one 100-episode *(suite, condition)* cell at the 600 s budget completes in roughly 3.5–4.5 h.

Table 4. Value-head training hyperparameters (Section 3.2).

Setting	Value
Architecture (MLP)	$768 \to 3072$ (ReLU) $\to$ scalar
Parameter count	~ 2.4M
Optimizer	Adam
Initial learning rate	1×10^{-4}
Learning-rate schedule	cosine decay ($\alpha = 0.1$)
Batch size	64
Epochs	100
Train/val split	$90/10$ (fixed seed 42)
Loss	MSE on Monte Carlo returns G_t
Discount γ	0.99
Target rebalancing	10-bin equal-width on $[0, 1]$, bottom bin downsampled

Table 5. MCTS hyperparameters, shared between VLAPS and V-VLAPS (Section 3.3). The value head adds the Q term in the SCORE function for the V-VLAPS condition only.

Setting	Value
PUCT exploration weight ψ	1.0
Action samples per node	300
Expansions per node	10
Expansions per MCTS step	1
Maximum tree depth	80
Maximum simulated rollout length	300
Chunk steps simulated per expansion	4
Replan after each executed chunk	yes
Per-episode wall-time cap	600 s (default budget), 1800 s (extended budget)
Action chunk library size	2,000 medoids of all-suite VLA successes
β_Φ distribution (action sampler)	softmax over chunk distances, $\alpha = 10$, $\epsilon = 0.1$
ψ_{Φ_v} distribution (PUCT prior)	softmax over sampled chunks, $\alpha = 5$
VLA generation temperature	0.0
Terminate search on simulated success	yes

C. Per-task results on libero_object

Table 6 decomposes the libero_object row of Table 1 by task. Tasks 0–5 and 9 sit at or near the 90–100% ceiling for both conditions at both budgets; the variance that drives the suite-level differences is concentrated on tasks 6, 7, and 8. The $N = 20$ task-7 result of Section 5.2 is a tighter measurement of the same task.

Table 6. Per-task success counts on libero_object at both per-episode wall-time budgets, 10 initial states per task. Full-suite totals match the libero_object row of Table 1. The bold cell highlights the V-VLAPS $5/10 \rightarrow 10/10$ swing on task 7 at the larger budget.

	600 s budget		1800 s budget	
Task	**VLAPS**	**V-VLAPS**	**VLAPS**	**V-VLAPS**
0	10/10	10/10	10/10	10/10
1	10/10	10/10	10/10	10/10
2	10/10	10/10	10/10	10/10
3	10/10	10/10	10/10	10/10
4	8/10	9/10	9/10	9/10
5	9/10	10/10	9/10	10/10
6	5/10	3/10	6/10	5/10
7	8/10	5/10	6/10	**10/10**
8	3/10	8/10	8/10	9/10
9	9/10	10/10	9/10	10/10
Total	**82/100**	**85/100**	**87/100**	**93/100**

D. t-SNE projection of Octo readouts

Figure 2 shows a t-SNE projection of the Octo last-layer readouts on libero_object, coloured by the Monte Carlo value target G_t. The visible structure between high- and low-target states is consistent with the SAFE finding of Gu et al. (2025) that VLA latent features carry outcome-discriminative signal.

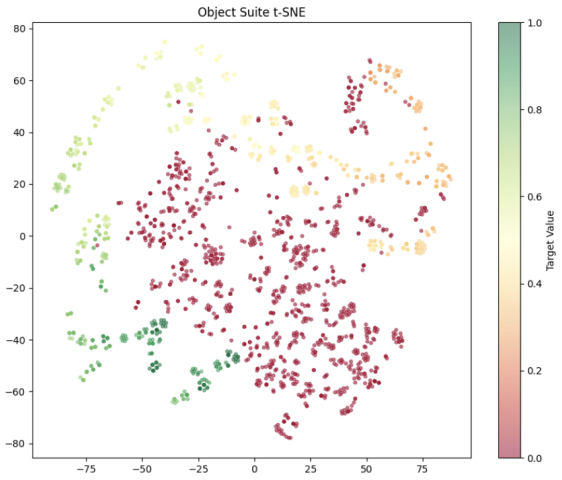

Figure 2. t-SNE projection of Octo last-layer readouts h_t on libero_object, coloured by Monte Carlo value target G_t. Successful and failed rollouts occupy visually distinct regions of the embedding.