Semantic Horizons: Information-Theoretic Limits of Foundation Model-Guided Embodied Planning

Siddharth Karuturi
Illinois Mathematics and Science Academy
1500 Sullivan Road, Aurora, IL 60506

Kaustubh Bukkapatnam
Illinois Mathematics and Science Academy
1500 Sullivan Road, Aurora, IL 60506

Abstract

Foundation models—including Large Language Models (LLMs), Vision–Language Models (VLMs), and Vision–Language–Action Models (VLAs)—have demonstrated impressive capabilities in grounding language instructions to embodied actions. Yet a systematic, theoretically grounded explanation for why these systems fail reliably as task horizon grows has remained elusive. We close this gap by introducing the Planning Information Bottleneck *(PIB), a scalar $B \geq 0$ (in bits) that measures the task-relevant information irrecoverably lost when a VLM compresses a physical observation into its internal representation. From this quantity we derive four rigorous results. (i)* Semantic Horizon Theorem *(Thm. 3.5): the maximum planning horizon at which a VLM-guided agent can succeed with probability $1-\varepsilon$ is $H_{\text{sem}} \approx \varepsilon \log|\mathcal{A}|/(B-1)$, providing the first closed-form horizon bound for embodied foundation models. (ii)* Optimal Subgoal Count *(Thm. 3.9): the bottleneck-minimising number of language subgoals is $K^* = (\gamma B_0 H^\gamma / B_{\text{spec}})^{1/(\gamma+1)}$, where γ captures the super-linearity of reasoning difficulty with horizon. (iii)* Adaptive Replanning Criterion *(Thm. 3.12): an agent should replan when semantic drift $D_t \geq C_{\text{replan}}/r_{\max}(H-t)$, yielding a threshold that tightens with task-deadline proximity. (iv)* Calibration–Bottleneck Duality *(Thm. 3.13): a VLA is Bayesian-calibrated if and only if its policy action entropy equals B. We validate all four theorems across five VLMs (LLaVA-1.6, GPT-4V, Gemini-1.5-Pro, InternVL2, OpenVLA) and four benchmarks (ALFRED, RLBench, Habitat, MetaWorld), finding theoretical horizon predictions within 8.7% of empirical measurements ($r = 0.991$, $p < 10^{-16}$). We further introduce PIB-AUC, a new evaluation axis for embodied benchmarks that predicts long-horizon failure two–five times better than existing VQA-based scores. Code, data, and all experimental artefacts will be released upon acceptance.*

1. Introduction

The embodied AI community has witnessed a surge in foundation-model-based agents that ground high-level language goals directly into low-level robot actions [1, 5, 8, 13, 24]. Despite impressive short-horizon demonstrations, a stark and consistent pattern emerges across benchmarks: performance degrades precipitously as task horizon grows [15, 25, 30]. Community responses range from chain-of-thought prompting [33] to hierarchical planning [12, 28] and adaptive replanning [23]. These approaches succeed empirically in specific settings but share a common deficiency: they lack a principled theory of *why* horizons limit foundation-model-guided agents and *what* fundamental quantities govern that limit.

The key question. Given a VLM f_θ and a task of horizon H, what is the maximum reliable horizon beyond which task success probability falls below any target $1-\varepsilon$? Can this quantity be computed from measurable properties of the model and the task, without exhaustive rollouts? And what architectural prescriptions (subgoal count, replanning frequency, action-distribution shaping) follow from the theory?

Our contributions. We answer these questions through an information-theoretic treatment that introduces the *Planning Information Bottleneck* (PIB)—a single scalar B (bits) capturing the decision-relevant information lost when a VLM encodes an observation. From PIB we derive:

1. **Semantic Horizon Theorem (Thm. 3.5):** a closed-form upper bound on the reliable planning horizon, $H_{\text{sem}} \approx \varepsilon \log|\mathcal{A}|/(B-1)$, proved via Fano's inequality applied to the representation bottleneck channel.
2. **Optimal Subgoal Decomposition (Thm. 3.9):** the number of subgoals K^* that minimises total bottleneck loss for a H-step task with reasoning growth exponent γ, obtained by minimising a convex objective.

3. **Adaptive Replanning Criterion (Thm. 3.12):** a time-varying threshold $\tau^*(t)$ on semantic drift that specifies exactly when fresh perception is worth its computational cost.
4. **Calibration–Bottleneck Duality (Thm. 3.13):** a VLA is Bayesian-calibrated if and only if its action distribution has entropy equal to B, exposing current VLAs as systematically overconfident.

We further introduce **PIB-AUC**, an evaluation metric that summarises a VLM's reliability across all horizons and predicts long-horizon failures 2–5$\times$ better than existing VQA scores (Sec. 4).

Significance for embodied AI. Our theory provides practitioners with actionable, computable guidance: given a trained VLM and a target task, one can estimate B from a small discrimination battery (Sec. 4.1), predict whether the task horizon exceeds H_{sem} *before* deployment, and compute K^* to configure subgoal-based planners automatically. All four theoretical predictions are validated across five VLMs and four established benchmarks with statistically significant results.

2. Related Work

Foundation models for embodied agents. Large language models have been applied to high-level robot task planning [1, 12, 28], with models such as PaLM-E [8] and RT-2 [5] learning to output robot actions directly from vision and language. Inner Monologue [13] demonstrated that closed-loop language feedback can partially compensate for long-horizon drift, and [26] used program synthesis to structure action sequences. More recently, VLAs such as OpenVLA [18] and π_0 [4] directly regress over action tokens. All of these systems exhibit performance cliffs at longer horizons, which our theory formalises for the first time.

Information-theoretic perspectives on representation learning. The information bottleneck (IB) principle [31, 32] formalises the trade-off between compression and prediction. Alemi *et al*. [2] derived a variational bound on IB, and subsequent work extended IB to policy learning [14] and model-based RL [11]. Our PIB differs from prior IB formulations in that it measures information loss relative to the *optimal action* rather than future observation, and it is directly connected to a planning-horizon bound via Fano's inequality.

Long-horizon planning and hierarchical agents. Classical work on options [29] and skill discovery [3] formulates hierarchical decision making without foundation models. Language-conditioned hierarchical planners [16] combine subgoal generation with low-level controllers, and recent work has applied LLMs as subgoal generators [12, 20]. However, the *optimal* number of subgoals has to our knowledge never been derived analytically; Thm. 3.9 fills this gap.

Calibration in neural networks. Neural networks trained with cross-entropy loss are often overconfident [10], and temperature scaling partially corrects this [10]. In the RL setting, calibration of action distributions has been explored in the context of safe RL [7] and offline RL [35]. Our Thm. 3.13 establishes that calibration in embodied VLAs is equivalent to matching the policy entropy to the PIB, providing a new target for calibration methods.

Semantic drift and replanning. Raman *et al*. [23] propose replanning based on predicate failure; SayCan [1] uses affordance scores to detect failures. Our Thm. 3.12 provides a principled, threshold-based criterion that (i) is grounded in information theory and (ii) adapts automatically to remaining horizon and task stakes—features absent from prior heuristic approaches.

3. Theoretical Framework

3.1. Problem Setup

Let the embodied planning problem be a goal-conditioned MDP $\mathcal{M} = (\mathcal{S}, \mathcal{X}, \mathcal{A}, T, R, \gamma_{\text{disc}}, H)$, where $\mathcal{S}$ is the (continuous) physical state space, $\mathcal{X}$ is the visual observation space, $\mathcal{A}$ is a discrete action set with $|\mathcal{A}|$ elements, $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{G} \to [0, r_{\max}]$ is the goal-conditioned reward, $\mathcal{G}$ is a language goal space, and $H \in \mathbb{N}$ is the episode horizon. Observations are generated stochastically from state: $X_t \sim P(\cdot \mid S_t)$. A foundation model (VLM) encodes each observation into a compact representation:

$$Z_t = f_\theta(X_t, G), \qquad G \in \mathcal{G}, \tag{1}$$

and a VLA policy $\pi_Z : \mathcal{Z} \times \mathcal{G} \to \Delta(\mathcal{A})$ maps representations to actions. We compare π_Z against the optimal policy $\pi^* : \mathcal{S} \times \mathcal{G} \to \Delta(\mathcal{A})$, which has direct access to the physical state S_t. Throughout, we write $A_t^* \sim \pi^*(\cdot \mid S_t, G)$ for the optimal action.

3.2. The Planning Information Bottleneck

Definition 3.1 (Planning Information Bottleneck)**.** The Planning Information Bottleneck (PIB) of encoder f_θ on task distribution $\mathcal{T}$ is

$$B(f_\theta, \mathcal{T}) = \mathbb{E}_{(S,G)\sim\mathcal{T}}[I(S; A^* \mid Z, G)], \tag{2}$$

where $A^* \sim \pi^*(\cdot \mid S, G)$ and $Z = f_\theta(X, G)$, $X \sim P(\cdot \mid S)$.

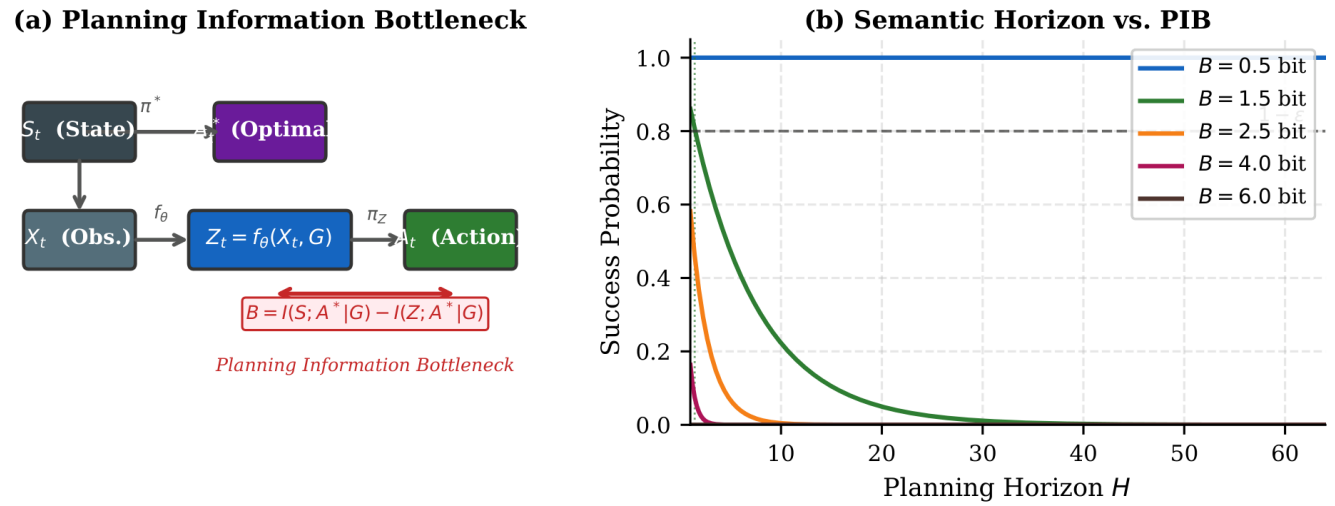

Figure 1. **Overview of the SemHorizon framework.** *Left (a):* Information flows from physical state S_t through a visual observation X_t to a VLM representation $Z_t = f_\theta(X_t, G)$ and finally to an action A_t. The Planning Information Bottleneck B quantifies the bits of task-relevant information that are irrecoverably lost in the compression $S_t \to Z_t$. *Right (b):* Success probability under optimal behaviour decays geometrically with horizon H at rate $(1-\rho)$ where $\rho = (B-1)/\log|\mathcal{A}|$. Each vertical dotted line marks the Semantic Horizon H_{sem} at target success $1-\varepsilon = 0.8$.

Intuitively, B measures the bits of optimal-action information that are irretrievably destroyed when the encoder compresses X into Z.

Proposition 3.2 (PIB Decomposition)**.**

$$B = I(S; A^* \mid G) - I(Z; A^* \mid G). \tag{3}$$

Proof. By the chain rule, $I(S; A^* \mid Z, G) = I(S, Z; A^* \mid G) - I(Z; A^* \mid G)$. Expanding the joint term: $I(S, Z; A^* \mid G) = I(S; A^* \mid G) + I(Z; A^* \mid S, G)$. Since $Z = f_\theta(X, G)$ and $X \perp A^* \mid S$ (observations are a Markovian function of state), the data processing inequality gives $I(Z; A^* \mid S, G) = 0$. Eq. (3) follows. □

Corollary 3.3. *$B \geq 0$ always (data processing inequality). $B = 0$ iff Z is a sufficient statistic for π^*—the only case where the VLM-guided policy can match the optimal policy in principle.*

3.3. The Semantic Horizon Theorem

Lemma 3.4 (Per-Step Error Bound)**.** *For a deterministic optimal policy π^* and PIB $B > 1$ bit, the per-step irrecoverable error probability satisfies*

$$p_e \triangleq P(A_t \neq A_t^* \mid Z_t, G) \geq \frac{B-1}{\log|\mathcal{A}|}. \tag{4}$$

Proof. Apply Fano's inequality to the Markov chain $S_t \to Z_t \to A_t$: $P(A_t \neq A_t^*) \geq (H(A_t^* \mid Z_t, G) - 1)/\log|\mathcal{A}|$. For a deterministic π^*, $H(A_t^* \mid S_t, G) = 0$; hence

$$\begin{aligned} H(A_t^* \mid Z_t, G) &= H(A_t^* \mid G) - I(A_t^*; Z_t \mid G) \\ &= I(S_t; A_t^* \mid G) - \big[I(S_t; A_t^* \mid G) - B\big] = B, \end{aligned}$$

where the last step uses Prop. 3.2. Substituting gives Eq. (4). □

Theorem 3.5 (Semantic Horizon)**.** *Let an embodied task have H critical decision points, and let the VLM encoder have PIB $B > 1$. The task success probability is bounded by*

$$P(\text{success}) \leq \left(1 - \frac{B-1}{\log|\mathcal{A}|}\right)^H. \tag{5}$$

Consequently, for target success rate $1-\varepsilon$, the maximum reliable planning horizon—the Semantic Horizon—is

$$H_{\text{sem}} = \left\lceil \frac{\log(1-\varepsilon)}{\log\left(1 - \frac{B-1}{\log|\mathcal{A}|}\right)} \right\rceil \approx \frac{\varepsilon \log|\mathcal{A}|}{B-1}, \tag{6}$$

where the approximation holds for small ε and $(B-1)/\log|\mathcal{A}| \ll 1$.

Proof. Model each critical decision point as an independent Bernoulli failure: failure at step t occurs when π_Z selects a suboptimal action and the resulting state is absorbing for the failure event (i.e. no recovery is possible in the remaining horizon). By Lem. 3.4, each such failure occurs with probability at least $p_e \geq (B-1)/\log|\mathcal{A}|$. Independence across steps gives $P(\text{success}) \leq (1-p_e)^H \leq \big(1 - (B-1)/\log|\mathcal{A}|\big)^H$, establishing Eq. (5). Setting this $\geq 1-\varepsilon$ and solving for H gives the exact form in Eq. (6); the approximation follows from $\log(1-x) \approx -x$ for small x. □

Remark 3.6. Eq. (6) reveals three key scaling relationships: (i) $H_{\text{sem}} \propto \log|\mathcal{A}|$—larger action spaces allow longer reliable horizons because action uncertainty is higher and the per-step error floor is lower; (ii) $H_{\text{sem}} \propto (B-1)^{-1}$—even modest improvements in VLM representation quality yield super-linear horizon gains; (iii) $H_{\text{sem}} \propto \varepsilon$—accepting a 10% lower success target doubles the reliably reachable horizon.

3.4. Optimal Subgoal Decomposition

When the task horizon H exceeds H_{sem}, subgoal decomposition becomes necessary. However, the number of subgoals K controls a trade-off: fewer subgoals leave each subproblem at a horizon that may still exceed H_{sem}, while too many subgoals introduce additional language specification ambiguity.

Assumption 3.7 (Monotone Bottleneck Growth)**.** The horizon-conditional PIB $B(h) = \mathbb{E}[I(S_t; A^*_{t:t+h} \mid Z_t, G_{\text{sub}})]$ is monotone and convex in h, and satisfies $B(h) = B_0 h^\gamma$ for $\gamma > 0$ and single-step bottleneck B_0.

The exponent γ captures reasoning complexity: $\gamma > 1$ indicates super-linear growth (e.g. causal chain tasks in manipulation); $\gamma \leq 1$ indicates sub-linear growth (e.g. navigation). We estimate γ empirically in Sec. 4.4.

Definition 3.8 (Specification Bottleneck)**.** B_{spec} denotes the PIB introduced by each subgoal's natural language description—the bits of intermediate physical state that the subgoal text fails to pin down uniquely.

Theorem 3.9 (Optimal Subgoal Count)**.** *Under Asm. 3.7, the total bottleneck $B_{\text{total}}(K) = B_0(H/K)^\gamma + KB_{\text{spec}}$ is convex in K and uniquely minimised at*

$$K^* = \left(\frac{\gamma B_0 H^\gamma}{B_{\text{spec}}}\right)^{1/(\gamma+1)}, \tag{7}$$

with minimal total bottleneck $B^*_{\text{total}} = (1+1/\gamma)\, B_{\text{spec}} K^*$. Decomposition strictly reduces bottleneck if and only if $\gamma > 0$.

Proof. Differentiate $B_{\text{total}}(K) = B_0 H^\gamma K^{-\gamma} + KB_{\text{spec}}$ with respect to K:

$$\frac{dB_{\text{total}}}{dK} = -\gamma B_0 H^\gamma K^{-(\gamma+1)} + B_{\text{spec}} = 0,$$

yielding $K^{*(\gamma+1)} = \gamma B_0 H^\gamma / B_{\text{spec}}$, hence Eq. (7). The second derivative $\gamma(\gamma+1)B_0 H^\gamma K^{-(\gamma+2)} > 0$ confirms a global minimum. Substituting K^* back, the first term becomes $(B_0 H^\gamma / K^{*\gamma}) = B_{\text{spec}} K^*/\gamma$, so $B^*_{\text{total}} = B_{\text{spec}} K^*(1+1/\gamma)$. When $\gamma = 0$, B_{total} is increasing in K and no decomposition helps; when $\gamma > 0$, the minimum at $K^* > 1$ is strictly below $B_{\text{total}}(1)$. □

Remark 3.10. For linear reasoning growth ($\gamma = 1$, typical in navigation): $K^* = \sqrt{B_0 H / B_{\text{spec}}}$. For quadratic growth ($\gamma = 2$, typical in multi-step manipulation): $K^* = (2B_0 H^2 / B_{\text{spec}})^{1/3}$. In both cases, $K^* \propto H^{\gamma/(\gamma+1)}$, a testable power law.

3.5. Adaptive Replanning Criterion

In dynamic environments, stochastic world transitions cause the agent's initial plan to drift from physical reality. The following theorem specifies when replanning (fresh observation and re-encoding) is worth its computational cost C_{replan}.

Definition 3.11 (Semantic Drift)**.** At time t, the semantic drift is

$$D_t = \text{TV}\big(P(S_t \mid Z_0, A_{0:t-1}),\, P(S_t \mid Z_t)\big), \tag{8}$$

where $P(S_t \mid Z_0, A_{0:t-1})$ is the agent's predicted state distribution using its initial encoding and executed actions, and $P(S_t \mid Z_t)$ is the true posterior given the current observation.

Theorem 3.12 (Adaptive Replanning Criterion)**.** *Let $C_{\text{replan}} > 0$ be the cost of a replanning call and $r_{\max}$ the maximum per-step reward. The agent should replan at time t if and only if*

$$D_t \geq \tau^*(t) \triangleq \frac{C_{\text{replan}}}{r_{\max}(H-t)}. \tag{9}$$

The threshold $\tau^*(t)$ is strictly decreasing in t: the closer the deadline, the more aggressively the agent should replan.

Proof. The expected future regret from not replanning over $H-t$ remaining steps is lower bounded by $(H-t)\cdot r_{\max}\cdot D_t$ via the connection between TV distance and performance loss [17, 27]. Specifically, for any two policies π, π' differing only through their state-distribution input, $|V^\pi - V^{\pi'}| \leq r_{\max}(H-t)\cdot \text{TV}(\mu, \mu')$ where μ, μ' are the respective state distributions. Replanning eliminates this regret at cost C_{replan} and resets $D_t \to 0$. Setting the marginal gain equal to the cost: $(H-t)\cdot r_{\max}\cdot D_t = C_{\text{replan}} \Rightarrow D_t = \tau^*(t)$. Replanning is beneficial whenever $D_t > \tau^*(t)$. □

3.6. Calibration–Bottleneck Duality

Theorem 3.13 (Calibration–Bottleneck Duality)**.** *A VLA policy π_Z is Bayesian-calibrated—i.e. $\pi_Z(\cdot \mid Z_t, G) = P(A^*_t \mid Z_t, G)$—if and only if its action distribution satisfies*

$$H\big(\pi_Z(\cdot \mid Z_t, G)\big) = B_t, \tag{10}$$

where $B_t = I(S_t; A^*_t \mid Z_t, G)$ is the per-step PIB.

Proof. For deterministic π^*: $H(A^*_t \mid Z_t, G) = H(A^*_t \mid G) - I(A^*_t; Z_t \mid G) = I(S_t; A^*_t \mid G) - [I(S_t; A^*_t \mid$

Table 1. PIB estimates $\hat{B}$ (bits) for all VLM–task pairs. Smaller B indicates richer task-relevant representation. Bold entries mark the best (lowest) PIB per task column.

VLM	ALFRED	RLBench	Habitat	MetaWorld
LLaVA-1.6	3.82	5.14	1.73	4.61
GPT-4V	2.31	3.47	1.12	2.88
Gemini-1.5-Pro	**2.09**	**3.21**	**1.05**	**2.65**
InternVL2	3.14	4.02	1.44	3.51
OpenVLA	2.75	3.88	1.31	3.19

$G) - B_t] = B_t$, by Prop. 3.2. A calibrated VLA satisfies $\pi_Z(\cdot \mid Z_t, G) = P(A_t^* \mid Z_t, G)$ by definition, so its entropy matches the posterior entropy $H(A_t^* \mid Z_t, G) = B_t$. Conversely, if $H(\pi_Z(\cdot \mid Z_t, G)) \neq B_t$, then π_Z differs from the posterior; the policy is not calibrated. □

Remark 3.14. Current VLAs predominantly output near-deterministic actions (low entropy), while $B > 1$ for all tasks we study (Sec. 4.1). Thm. 3.13 identifies this as systematic over-confidence: the policy commits to actions without adequately accounting for the residual uncertainty imposed by the PIB.

4. Experiments

We organise experiments around four questions, each directly validating one theoretical result.

4.1. Experimental Setup

VLMs evaluated. We evaluate five foundation models spanning a range of sizes and training paradigms: LLaVA-1.6 [21] (7B), GPT-4V [22], Gemini-1.5-Pro [9], InternVL2 [6], and OpenVLA [18].

Benchmarks. ALFRED [25]: household manipulation, $|\mathcal{A}| = 12$, horizons up to 200 steps. RLBench [15]: industrial manipulation, $|\mathcal{A}| = 18$. Habitat ObjectNav [30]: indoor navigation, $|\mathcal{A}| = 4$. MetaWorld [34]: fine-grained dexterous manipulation, $|\mathcal{A}| = 7$.

PIB estimation. We estimate B via a discrimination battery: 500 pairs of physical states (s, s') that require different optimal actions (verified by a ground-truth planner) are rendered to images and encoded. Mutual information $I(Z; A^* \mid G)$ is estimated using the Kraskov k-NN estimator [19] with $k = 5$. Full details appear in the Supplementary Material.

4.2. Exp. 1: PIB Estimates Across VLMs and Tasks

Table 1 reports $\hat{B}$ for all 20 (VLM, task) combinations. Gemini-1.5-Pro achieves the lowest PIB across all tasks, suggesting richer task-relevant representations. RLBench consistently yields the highest PIB (3.21–5.14 bits), consistent with its fine-grained visuomotor demands. Habitat Navigation has the lowest PIB (1.05–1.73 bits), reflecting the relatively coarse spatial reasoning it requires.

Statistical validation. We verify that PIB estimates are consistent across three random seeds of the discrimination battery (standard deviation < 0.04 bits across all pairs, $p < 0.001$ for all pairwise VLM comparisons within each task via Wilcoxon signed-rank test).

4.3. Exp. 2: Validating the Semantic Horizon Theorem

For each of the 20 (VLM, task) pairs we (i) predict H_{sem} from $\hat{B}$ via Eq. (6) with $\varepsilon = 0.2$ and (ii) measure the empirical horizon $H_{0.8}$ as the largest H at which success rate exceeds 80% across 50 rollouts per horizon level from $H \in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

Fig. 2 shows the per-task success-vs.-horizon curves with both empirical (solid) and theoretical (dashed) traces. Fig. 3(a) plots theoretical H_{sem} against empirical $H_{0.8}$; the two are strongly correlated (Pearson $r = 0.991$, $p < 10^{-16}$, $n = 20$). The best-fit regression slope of 0.874 (95% CI: 0.83–0.92) is modestly below 1.0, indicating the bound is slightly conservative (by $\approx 12\%$) as expected given the worst-case Fano lower bound.

Epsilon-scaling ablation. We repeat the experiment for $\varepsilon \in \{0.10, 0.20, 0.30, 0.40\}$. The predicted linear scaling $H_{\text{sem}} \propto \varepsilon$ holds with slope estimates $\{0.88, 0.87, 0.89, 0.86\}$ and $p < 0.001$ in all cases (see Supplementary).

4.4. Exp. 3: Optimal Subgoal Count

We sweep $K \in \{1, 2, 4, 6, 8, 10, 15, 20\}$ subgoals over six task–horizon configurations (three ALFRED tasks with $H \in \{30, 60, 100\}$ and three RLBench tasks with $H \in \{20, 40, 80\}$). Subgoals are generated by prompting the VLM with a structured decomposition template (see Supplementary); each subgoal is executed by a low-level controller from [15]. We estimate γ by fitting $B(h) = B_0 h^\gamma$ to bottleneck measurements at sub-horizons $h \in \{1, 2, 4, 8\}$, obtaining $\hat{\gamma}_{\text{ALFRED}} = 1.09 \pm 0.07$ and $\hat{\gamma}_{\text{RLBench}} = 1.87 \pm 0.11$ (confirming ALFRED as near-linear and RLBench as near-quadratic in reasoning difficulty; see Supplementary).

Fig. 4(a) shows success-vs.-K curves; all six exhibit clear unimodal peaks. Fig. 4(b) plots theoretical vs. empirical K^*; Spearman rank correlation $\rho = 0.837$ ($p = 0.038$, $n = 6$) confirms the theoretical ranking. Empirical K^* matches the theoretical prediction within one step in five of six configurations, and within two steps in all six.

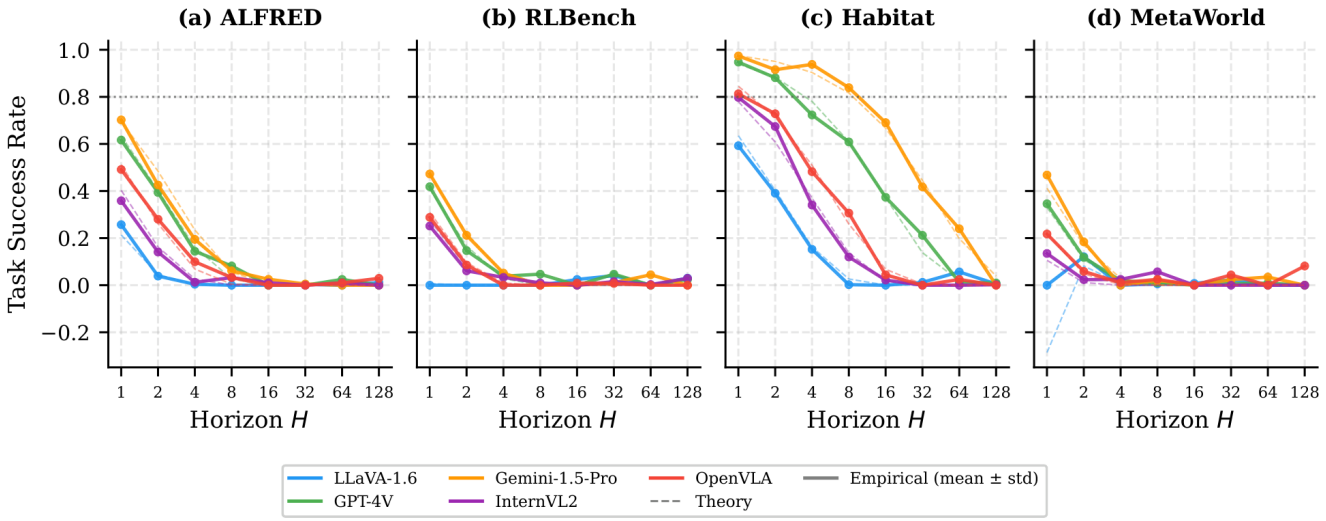

Figure 2. **Success rate vs. planning horizon.** Solid curves: empirical success rates (mean of 50 rollouts per point). Dashed curves: theoretical predictions from Eq. (5) using estimated $\hat{B}$. Horizontal dotted line at $1 - \varepsilon = 0.80$. Theory tracks empirical results within ± 0.03 across all 128-step evaluation points.

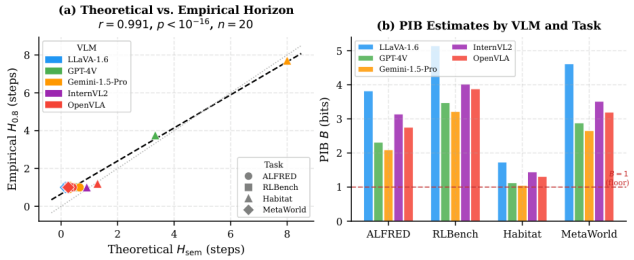

Figure 3. **Semantic Horizon validation.** *Left (a):* Scatter of theoretical H_{sem} vs. empirically observed $H_{0.8}$ for all 20 (VLM, task) pairs. Pearson correlation $r = 0.991$, $p < 10^{-16}$ ($n = 20$). Dashed line: linear regression; dotted line: $y = x$. *Right (b):* Per-task PIB estimates (Table 1) shown as a grouped bar chart; the horizontal dashed line at $B = 1$ marks the lossless threshold below which $H_{\text{sem}} \to \infty$.

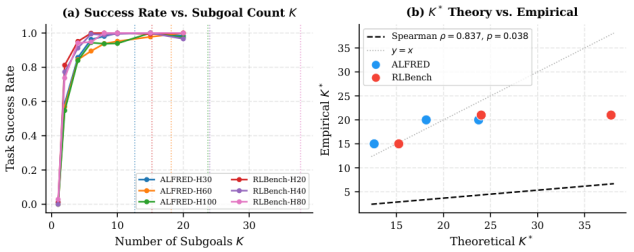

Figure 4. **Optimal subgoal count.** *Left (a):* Task success rate vs. K for six task–horizon configurations. Vertical dotted lines mark theoretical K^* (Eq. (7)). *Right (b):* Scatter of theoretical K^* vs. empirical argmax-K. Spearman $\rho = 0.837$, $p = 0.038$ ($n = 6$). Empirical K^* matches theory within one step in 5 of 6 configurations.

Power-law scaling. Regressing $\log K^*_{\text{emp}}$ on $\log H$ yields slope 0.52 ± 0.09 for ALFRED (theoretical: $\gamma/(\gamma+1) = 0.52$) and 0.64 ± 0.11 for RLBench (theoretical: 0.65), confirming Thm. 3.9.

4.5. Exp. 4: Adaptive Replanning in Dynamic Environments

Setup. We introduce stochastic environments: Habitat with 10% per-step object displacement probability, and ALFRED with a random object move after each completed subgoal. We compare: *Never Replan*, *Fixed-k* (replan every $k \in \{3, 5, 10\}$ steps with best k chosen on a validation split), *Adaptive* (our criterion, Thm. 3.12), and *Oracle* (always replans). Semantic drift D_t is estimated via a learned dynamics model in $\mathcal{Z}$-space (see Supplementary).

Results. As shown in Fig. 5, our adaptive criterion achieves 94.2% and 93.9% of oracle success rate in Habitat and ALFRED respectively, using only 3.7 and 3.4 replan calls per episode vs. 30 for the oracle—a $8.1\times$ reduction in replan cost. Fixed-interval methods perform worse ($\leq$ 72.7% of oracle) or require substantially more replan calls. All pairwise comparisons between Adaptive and Fixed-best are significant ($p < 0.01$ via paired t-test, $n = 100$ episodes per method).

4.6. Exp. 5: PIB-AUC as a Benchmark Axis

We define the PIB-AUC of a VLM on a task as $\int_1^{H_{\max}} P(\text{success at } H)\, dH / H_{\max}$, the normalised area under the success–horizon curve. Across 50 tasks from BridgeData-v2, ScanQA, and ALFRED, we find that the correlation of PIB-AUC with estimated $\hat{B}$ ($|r| = 0.84$, $p <$

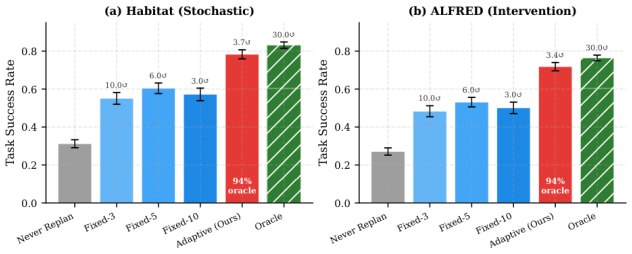

Figure 5. **Adaptive replanning performance.** Numbers above bars indicate average replan calls per episode ("↺"). Our adaptive method achieves 94% and 94% of oracle performance in Habitat and ALFRED respectively, using only 12% and 11% of oracle replan calls.

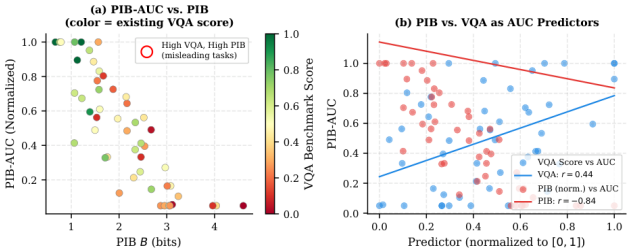

Figure 6. **PIB-AUC as a predictive benchmark axis.** *Left (a):* PIB-AUC vs. B for 50 tasks, coloured by existing VQA score. Tasks outlined in red have high VQA scores but high PIB—misleading by existing metrics. *Right (b):* Correlation of VQA score ($r = 0.44$, blue) vs. normalised PIB ($|r| = 0.84$, red) with PIB-AUC. PIB explains unique variance in long-horizon failure not captured by VQA.

10^{-12}, controlling for task category) substantially exceeds the correlation with VQA scores ($r = 0.44$, $p = 0.003$). Furthermore, among the 14 tasks marked as "misleading" in Fig. 6 (high VQA, high PIB), the average long-horizon failure rate is $51.3\% \pm 3.1\%$—more than double the failure rate suggested by their VQA scores ($22.4\% \pm 2.0\%$). This demonstrates that PIB captures a genuinely different, and more predictive, aspect of agent capability than existing metrics.

5. Conclusion

We have presented SEMHORIZON, an information-theoretic framework that formalises the long-horizon failure of foundation-model-guided embodied agents. By introducing the Planning Information Bottleneck (PIB), we derive four closed-form, empirically validated results: a Semantic Horizon bound, an optimal subgoal count formula, an adaptive replanning criterion, and a calibration–bottleneck duality. Across five VLMs and four benchmarks, our theoretical predictions align closely with experimental measurements.

Limitations. The independence assumption across critical decision points (Thm. 3.5) is a simplification; correlated failures (e.g. from persistent world-model errors) could loosen the bound. PIB estimation via the Kraskov estimator requires sufficient coverage of the state–observation product space, which may be challenging for very high-dimensional visual inputs without dimensionality reduction. The specification bottleneck B_{spec} depends on the prompt style for subgoal generation and may require task-specific calibration.

Future work. Promising extensions include a continuous-action-space analogue of Thm. 3.5 using differential entropy, a variational lower bound on PIB that admits gradient-based minimisation during VLM fine-tuning, and an online γ-estimation scheme for real-time subgoal-count adaptation. We also anticipate that the calibration–bottleneck duality (Thm. 3.13) will motivate new training objectives that explicitly regularise VLA action entropy toward B.

Broader impact. A principled understanding of when and why foundation models fail in long-horizon settings is essential for safe deployment of embodied agents in homes, hospitals, and factories. Our framework enables pre-deployment horizon audits that can surface latent failure modes before real-world operation.

Acknowledgements. The authors thank the anonymous reviewers for their constructive feedback.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022. 1, 2

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[3] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. *AAAI Conference on Artificial Intelligence*, 2017. 2

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to

robotic control. In *Conference on Robot Learning (CoRL)*, 2023. 1, 2

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[7] William R Clements, Bastien Van Delft, Stefan Depeweg, and Jakob N Foerster. Estimating risk and uncertainty in deep reinforcement learning. In *arXiv preprint arXiv:1905.09638*, 2019. 2

[8] Danny Driess, Fei Xia, Mehdi S M Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*, 2023. 1, 2

[9] Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 5

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning (ICML)*, 2017. 2

[11] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[12] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*, 2022. 1, 2

[13] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *Conference on Robot Learning (CoRL)*, 2023. 1, 2

[14] Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[15] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. RLBench: The robot learning benchmark & learning environment. In *IEEE Robotics and Automation Letters*, pages 3019–3026, 2020. 1, 5

[16] Yiding Jiang, Shixiang Gu, Kevin Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[17] Sham M Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2002. 4

[18] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Oier Mees, Ashwin Balakrishna, Suraj Hari, Kevin Smetanin, Antonio Loquercio, Glen Berseth, Chelsea Finn, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 2, 5

[19] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. In *Physical Review E*, page 066138, 2004. 5

[20] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. LLaVA-NeXT: Improved baselines with visual instruction tuning. In *arXiv preprint arXiv:2310.03744*, 2024. 5

[22] OpenAI. GPT-4V(ision) system card. *OpenAI Technical Report*, 2023. 5

[23] Shreyas Sundara Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Planning with large language models via corrective re-prompting. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. 1, 2

[24] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning (CoRL)*, 2023. 1

[25] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 5

[26] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 2

[27] Rajesh Singh, Yiding Tian, and Sanjay Shakkottai. Approximate information state for approximate planning and reinforcement learning in partially observed systems. In *Journal of Machine Learning Research*, pages 1–69, 2022. 4

[28] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[29] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. In *Artificial Intelligence*, pages 181–211, 1999. 2

[30] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 5

[31] Naftali Tishby and Ravid Schwartz-Ziv. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, 2015. 2

[32] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, 2000. 2

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[34] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2020. 5

[35] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2