From 201 Bins to 1 Bit: The Sufficiency of Binary Advantage in Vision-Language-Action Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Vision-language-action (VLA) models have emerged as a promising paradigm for generalist robot control, but existing advantage-conditioned RL pipelines typically rely on fine-grained categorical critics, which are computationally expensive, sample-inefficient, and theoretically ungrounded. To address this, we examine the sufficiency of a minimal binary advantage indicator for guiding policy improvement. We show that a single binary advantage indicator per observation is formally sufficient, identify the Bayes-optimal critic as a Bernoulli posterior, and demonstrate a dramatic reduction in sample complexity compared to multi-bin critics. Guided by this theoretical insight, we design a streamlined offline RL pipeline that employs a scalar regression critic with a Persistent Advantage Buffer and integrates prompt-level classifier-free guidance for actor fine-tuning. Empirical evaluations on real-world bimanual robotic tasks confirm that our binary-advantage approach matches the performance of traditional 201-bin pipelines while substantially simplifying both training and inference.

1 Introduction

Recent advances in vision-language-action (VLA) models [1, 2] have established themselves as the state-of-the-art approach for generalist robot control. These models can acquire complex behaviors from human demonstrations through behavioral cloning (BC), enabling a wide range of manipulation and interaction tasks. Despite their impressive capabilities, BC is fundamentally limited: it is constrained by the coverage of demonstration data and prone to compounding errors during open-loop execution [3], which restricts autonomous improvement and adaptation.

Reinforcement learning (RL) offers a natural mechanism for surpassing demonstration limitations by optimizing a policy through trial-and-error experience. However, conventional on-policy gradient methods [4] are computationally intensive and sample-inefficient when applied to billion-parameter generative actors, making direct RL impractical at scale. To address this challenge, industrial systems such as RECAP [5] adopt an advantage-conditioned paradigm, wherein a separate value function estimates per-step advantages, discretizes them into positive or negative indicators, and injects these indicators into the policy's prompt sequence during fine-tuning. Conditioning on positive indicators allows the policy to leverage the guidance of the critic to achieve superhuman performance while maintaining training efficiency.

Despite empirical success, current implementations rely on large categorical critics, often with hundreds of bins, inherited from offline RL literature. These heavy representations introduce quantization artifacts, elevate sample complexity, and require intricate, ad-hoc reward shaping to stabilize training. More importantly, the theoretical justification for discrete advantage conditioning remains limited: it is unclear why conditioning on coarse advantage signals works, when simpler representations suffice, and how classifier-free guidance (CFG) interacts with advantage signals to prevent mode collapse.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

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.

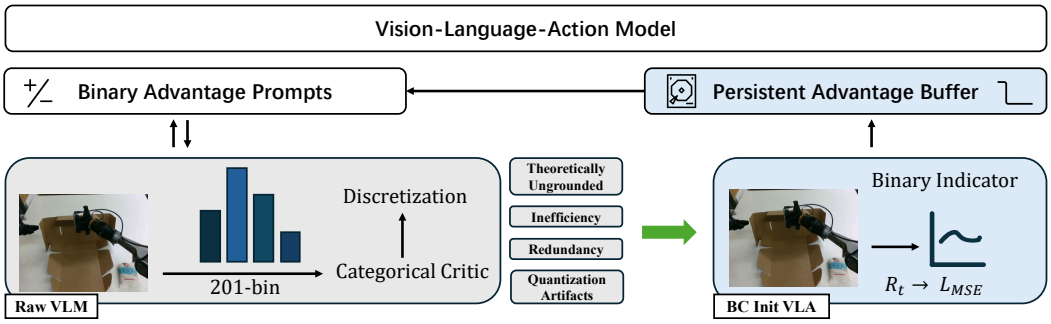


Figure 1: **Streamlining Advantage-Conditioned VLA Training: From 201 Bins to 1 Bit.** The conventional paradigm (left) discretizes a heavy 201-bin categorical critic to generate advantage prompts—an approach plagued by sample inefficiency, distributional redundancy, and quantization artifacts. To overcome these limitations, we propose sufficient pipeline (right) based on theoretically grounded analysis and design. A lightweight scalar critic estimates continuous returns, which are compiled offline into a Persistent Advantage Buffer (PAB) and binarized via zero-cost quantile thresholding (step function). The VLA actor (top) is then efficiently fine-tuned using exclusively 1-bit prompts, eliminating the multi-bin computational overhead while maintaining policy improvement.

In this work, we challenge the assumption that fine-grained advantage distributions are necessary for effective VLA policy improvement. We show that, mathematically, a binary advantage indicator per observation is sufficient for optimal policy improvement, providing a minimal yet theoretically grounded critic that reduces computational overhead and avoids unnecessary quantization. Building on this insight, we design a lightweight VLA training pipeline in which a scalar regression head replaces the categorical critic, BC-bootstrapped pretraining ensures stable initialization, and a Persistent Advantage Buffer (PAB) materializes continuous advantages offline to accelerate training and support flexible fine-tuning. The whole framework is shown in Figure 1.

Through theoretical analysis and empirical validation, we demonstrate that binary advantage conditioning not only provides formal guarantees on sample complexity, reward-scale invariance, and CFG stability, but also achieves strong performance on challenging bimanual manipulation tasks with embodied robotic platforms. Our results suggest that minimalist, binary advantage representations are both sufficient and preferable for practical reinforcement learning in large-scale VLA models, offering a simpler, more efficient, and theoretically principled alternative to heavy categorical critics. The main contributions of this paper are as follows:

- We establish that a binary advantage indicator is theoretically sufficient for optimal policy improvement, providing formal guarantees on sample complexity and CFG stability.
- We propose a lightweight VLA training pipeline using a scalar regression head and a Persistent Advantage Buffer (PAB) to efficiently leverage offline advantages.
- We demonstrate that binary advantage conditioning achieves strong empirical performance on complex bimanual manipulation tasks with embodied robots.

2 Related Work

Vision-Language-Action (VLA) Policies. Modern VLA architectures integrate a vision-language backbone with flow-matching or diffusion-based action generation [1, 6–8]. Flow matching [2] parameterizes a continuous-time velocity field whose ODE integration produces high-frequency action sequences. Building on this foundation, our actor incorporates prompt-level classifier-free guidance (CFG) [9] via a single binary token, enabling deterministic advantage conditioning. As these models scale, transitioning from pure imitation learning—which is inherently limited by demonstration coverage and prone to compounding errors—to reinforcement learning has become critical for achieving robust, long-horizon manipulation.

Offline Reinforcement Learning with Advantages. Offline RL methods provide a scalable route to improve policies without online exploration. Approaches such as Implicit Q-Learning [10],

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Advantage-Weighted Regression [11], Critic-Regularised Regression [12], and Conservative Q-Learning [13] bias behavioral cloning toward higher-advantage actions using continuous advantage estimates. While effective in principle, these methods require fine-grained advantages, which in real-world robotic settings can be noisy and high-variance, leading to numerical instability and limited theoretical guarantees. In contrast, binary advantage conditioning bypasses these issues while maintaining provable policy improvement guarantees.

Distributional Value Functions and Binary Critic Heads. To mitigate the challenges of estimating state-action advantages, distributional value functions—including categorical and quantile heads—model the full return distribution, providing richer supervisory signals. Stage Advantage methods [14] further stabilize learning by predicting advantages conditioned on semantic task stages. Our approach generalizes this idea to binary threshold-crossing indicators, inspired by AlphaGo's binary terminal value head [15]. By discretizing intermediate advantages into a single binary signal, we enforce distributional consistency directly within the policy extraction objective, reducing variance in long-horizon, partially observed tasks.

Industrial Advantage-Conditioned Systems. RECAP [5] demonstrates the first large-scale deployment of advantage-conditioned VLA policies, combining a Gemma-3 4B backbone, an 860M flow-matching action expert, and a 670M categorical critic. The system supports both autonomous execution and human interventions, achieving human-comparable robustness in industrial manipulation. Our work formalizes this paradigm as an implementation-independent framework, providing theoretical grounding, open-source code, and empirical validation at academic scale, thereby bridging the gap between industrial practice and academic research.

3 Preliminaries

3.1 Reinforcement Learning and Flow-Matching VLAs

We formalize the robotic control problem within a finite-horizon Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{O}, \mathcal{A}, P, r, H)$. An agent, acting according to a stochastic policy $\pi(a_t|o_t)$, selects a continuous action chunk $\boldsymbol{a}_{1:H} \in \mathcal{A} \subseteq \mathbb{R}^{H \times d_a}$ given an observation $o_t \in \mathcal{O}$. This process induces a trajectory $\tau = (o_0, a_0, \ldots, o_H)$ with a distribution formally given by $\rho^\pi(\tau) = \mu(o_0) \prod_{t=0}^{H-1} \pi(a_t|o_t) P(o_{t+1}|o_t, a_t)$, where $\mu(o_0)$ is the initial state distribution. The standard RL objective is to maximize the expected return $J(\pi) = \mathbb{E}_{\tau \sim \rho^\pi}\big[\sum_{t=0}^{H-1} r(o_t, a_t)\big]$.

In our context, the policy π_θ is instantiated as a multi-billion-parameter flow-matching VLA model. It parameterizes a velocity field $\boldsymbol{u}_\theta(x_t; o, c)$ governing an auxiliary noise-to-action diffusion process indexed by $t \in [0, 1]$ [2]. During inference, Euler integration over $\boldsymbol{u}_\theta$ maps a standard Gaussian sample to an action chunk. Crucially, the conditioning context c hosts both language instructions and the auxiliary indicator token introduced below.

3.2 Regularized Policy Extraction via Advantage Conditioning

Standard on-policy optimization of $J(\pi)$ is computationally prohibitive for generative VLAs. Instead, modern pipelines rely on regularized policy extraction, optimizing for a policy that maximizes reward while remaining close to a behavior reference policy π_{ref} (e.g., the training data distribution P_{train}). Let $V^{\pi_{\text{ref}}}(o)$ be the expected value and $A^{\pi_{\text{ref}}}(o, a)$ be the corresponding advantage.

While traditional advantage-weighted methods solve $\max_\pi J(\pi, \pi_{\text{ref}})$ by deriving a continuous exponential advantage weighting $\hat{\pi} \propto \pi_{\text{ref}} \exp(A/\beta)$, the advantage-conditioned recipe formulates this as a discrete probabilistic improvement indicator. We define a deterministic binarization map $I = \mathbf{1}[A^{\pi_{\text{ref}}}(o, a) \geq \tau_q]$, anchored at the q-quantile of the offline dataset's advantage distribution. The VLA actor $\pi_\theta(a|o, I)$ is subsequently trained via conditional behavior cloning. At inference time, to construct the optimally executing policy (P_{test}), we deterministically condition on $I = 1$ and apply classifier-free guidance (CFG) with scale $\gamma > 0$ to the velocity field:

$$\tilde{\boldsymbol{u}}(x_t; o, I{=}1) = \boldsymbol{u}(x_t; o) + \gamma\big[\boldsymbol{u}(x_t; o, I{=}1) - \boldsymbol{u}(x_t; o)\big]. \tag{1}$$

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

3.3 The Indicator-Conditioned Policy Class Π_I

To formally characterize the alignment between the training distribution (P_{train}) and real-world execution (P_{test}), we define two fundamental policy classes. The continuous-advantage class $\Pi_A = \{\pi(\cdot \mid o, A(o,\cdot))\}$ is consumed by standard offline RL. In contrast, the indicator-conditioned class utilized by our actor is defined as:

$$\Pi_I \;=\; \big\{\pi(\cdot \mid o, I) \;:\; \pi \text{ measurable},\; I \in \{0,1\}\big\} \tag{2}$$

Unlike Π_A, a policy in Π_I admits a strict structural bottleneck: at most two action distributions per observation. Real-world deployment often reveals systematic inconsistencies—such as coverage deficiency and failure cascades—where minor perturbations lead to unrecoverable states. By projecting the high-dimensional advantage manifold into a binary variable, Π_I partitions the behavioral space. It captures the broad multimodal support of P_{train} during training, yet allows inference-time CFG to explicitly filter out suboptimal temporal mismatches, establishing a mathematically tractable foundation for the proofs in Section 5.

4 Methodology

Our pipeline instantiates the advantage-conditioned framework introduced in Section 3.2 through three modular stages, entirely decoupled via the file system (**??**). Specifically, Section 4.1 trains a scalar value function from a Behavioral Cloning (BC) checkpoint using a time-only reward; Section 4.2 compiles N-step advantages into a *Persistent Advantage Buffer* (PAB) serialized to disk; and Section 4.3 fine-tunes the actor using prompt-level classifier-free guidance (CFG). Section 4.4 outlines our architectural simplifications over RECAP, explicitly grounding each decision in the theoretical propositions of Section 5.

4.1 Stage 1: BC-Bootstrapped Value Function

The value function V_ϕ shares the Vision-Language-Action (VLA) backbone architecture with the target actor. Crucially, it is initialized from a BC-finetuned VLA checkpoint *on the target task*, rather than from a raw vision-language backbone. (For reproducibility, the BC checkpoint `pi05_aloha...` is loaded via `pytorch_weight_path` and reused seamlessly across Stages B and C). Architecturally, we replace the VLA's action-output projection with a two-layer scalar value head of width $W \to W/2 \to 1$, and we restrict the action expert's conditioning solely to the state token. This structural modification ensures V_ϕ operates as a pure state-value function.

Following Equation (4) of Section 5, V_ϕ rigorously estimates the standard state-value function (expected return). Under our time-only penalty, this continuous value mathematically dualizes as a proxy for the normalized negative remaining steps to task completion, bounded strictly in $[-1, 0]$. Together, the BC initialization and time-only supervision systematically eliminate RECAP's requirement for explicit step-wise reward annotations and failure penalties.

4.2 Stage 2: The Persistent Advantage Buffer

Upon convergence of V_ϕ, we execute an offline pass over the replay buffer to compute the N-step bootstrap advantage for each episode:

$$A_t \;=\; N \cdot r + V_\phi(o_{t+N}) - V_\phi(o_t), \tag{3}$$

utilizing $N = 50$ and the constant reward defined in Section 4.1. The resulting length-T arrays of continuous advantages are serialized directly to `HDF5` files on disk. This collection constitutes our *Persistent Advantage Buffer* (PAB).

The PAB acts as a first-class asset in our pipeline: any subsequent actor fine-tuning job consumes it directly via the dataloader without ever re-instantiating V_ϕ. The threshold τ_q is computed dynamically and applied at the dataset boundary, allowing arbitrary quantiles q to be swept by altering a single configuration value. By Proposition 5, the population estimand is completely invariant under strictly monotonic reward re-scalings; thus, our time-only reward formulation loses no statistically relevant information.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 1: Key design axes comparing RECAP to our proposed pipeline. Every architectural simplification is strictly justified by an explicit theoretical proposition proven in Section 5.

Design Axis	**RECAP**	**Ours**	**Theoretical Justification**
Critic init source	Raw Gemma-3 VLM	BC-finetuned VLA ckpt	Proposition 2
Reward signal	Explicit step rewards & failure penalty	Time-only $r \equiv -1/T$ (no labels)	Proposition 5
Critic output & loss	201-bin categorical, cross-entropy	Scalar regression (MSE) + Binarization	Propositions 1, 3 and 4
Critic–actor coupling	Jointly trained, on-the-fly V-forward	PAB on disk, fully decoupled	Proposition 5 (Free thresholding)
Guidance signal	Binary prompt + CFG	Binary prompt + CFG	Proposition 6

4.3 Stage 3: Prompt-CFG Actor Fine-Tuning

The actor utilizes the standard flow-matching VLA architecture, initialized from the identical BC checkpoint used for V_ϕ. The PAB is exposed to the actor exclusively via dynamic prompt engineering at the dataset level. Specifically, our dataset transform reads the continuous advantage from the PAB, applies the quantile threshold q on-the-fly to yield the binary indicator $I \in \{0, 1\}$, and appends the natural language suffix `"Advantage: positive."` or `"Advantage: negative."` to the user instruction.

To enable classifier-free guidance, this suffix is randomly dropped with probability $p_{\text{drop}} = 0.3$, providing the necessary unconditional forward pass. Crucially, the continuous advantage tensor emitted by the dataloader is explicitly discarded; the actor never observes a continuous numerical advantage, relying entirely on the 1-bit prompt token. During inference, the suffix is deterministically forced to `"Advantage: positive."`, and CFG is applied at a scale γ per Equation (1). As established in Proposition 6, the resulting guidance perturbation is theoretically bounded by a single bit of mutual information, formally explaining why scaling γ arbitrarily high does not induce mode collapse.

4.4 Three Simplifications over RECAP

Table 1 explicitly contrasts the design axes of our system against RECAP. Each architectural substitution we propose is directly theoretically motivated. This table serves as a structural mapping between the empirical pipeline introduced in Section 4 and the theoretical foundations proven in Section 5.

5 Theoretical Analysis

This section establishes the information-theoretic principles underlying the critic design introduced in Section 4 and, more broadly, the *advantage-conditioned* paradigm pioneered by RECAP [5]. Let Π_I denote the indicator-conditioned policy class defined in Section 3.3. We demonstrate that: (i) the binary indicator serves as a sufficient statistic for any policy in Π_I; (ii) mean-squared regression onto the binary target recovers the Bayes-optimal critic; (iii) RECAP's categorical K-bin critic estimates a strictly richer statistic than Π_I can utilize, incurring a K-fold sample-complexity overhead; and (iv) quantile-anchored thresholding renders the estimand invariant to reward scaling, while binary classifier-free guidance (CFG) benefits from an information-bounded score perturbation that mitigates guidance-induced mode collapse. Complete proofs are provided in Section C.

5.1 Setup and Notation

We formulate an episode as a finite-horizon Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{O}, \mathcal{A}, P, r, T)$ with length-H action chunks [1] and bounded rewards r. Following Section 4.1, we employ a time-only reward $r_t \equiv -1/T$, such that the undiscounted state value satisfies:

$$V^\pi(o_t) = \mathbb{E}^\pi\Big[\sum_{k=t}^{\tau} r_k \,\Big|\, o_t\Big] = -\mathbb{E}^\pi[\tau - t \mid o_t]/T \in [-1, 0], \tag{4}$$

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

where $\tau \leq T$ is the episode termination time. This endows V with the physical interpretation of *negative normalized remaining time*. Given an advantage estimator A (Section 4.2) and τ_q, the q-quantile of the marginal of A, the *improvement indicator* is defined as:

$$I(o, a) \coloneqq \mathbf{1}[A(o, a) \geq \tau_q] \in \{0, 1\}, \tag{5}$$

which is materialized as a prompt suffix (Section 4.3). We contrast the hypothetical continuous-advantage class $\Pi_A = \{\pi(\cdot \mid o, A(o, \cdot))\}$ against the indicator-conditioned class $\Pi_I = \{\pi(\cdot \mid o, I)\} \subset \Pi_A$ from Section 3.3. Correspondingly, we compare our decoupled pipeline—which regresses a scalar critic V_ϕ onto the continuous return via MSE and applies the indicator thresholding dynamically at the data-loader level—against a categorical critic $p_\phi \in \Delta^{K-1}$ trained via cross-entropy onto a K-bin discretized return [16], where $K = 201$ represents the RECAP baseline.

5.2 Indicator Sufficiency

Our initial result formalizes the intuition that restricting the actor's input space to I renders any finer statistic of A superfluous for policy improvement.

Proposition 1 (Indicator Sufficiency)**.** *Let Π_I be the indicator-conditioned policy class defined in* (2)*. For any policy improvement operator $\mathcal{T}$ that maps (o, a) to a non-negative weight depending solely on $(o, I(o, a))$—including filter-and-clone, advantage-weighted BC with binary weights, and prompt-conditioned classifier-free guidance—the maximizer $\pi^* = \arg\max_{\pi \in \Pi_I} \mathbb{E}_{(o,a)\sim\rho^\pi}[A(o, a)]$ is entirely determined by $\rho_I(o, a) \coloneqq \Pr[I(o, a) = 1 \mid o, a]$. Consequently, no finer functional of the marginal of $A \mid (o, a)$ can alter π^*.*

Sketch. Any $\pi \in \Pi_I$ is parameterized by the two densities $\pi(a \mid o, 1)$ and $\pi(a \mid o, 0)$. The expected advantage decomposes pointwise as $\rho_I \, \mathbb{E}[A \mid o, I{=}1] + (1 - \rho_I) \, \mathbb{E}[A \mid o, I{=}0]$; thus, the optimization over Π_I depends on the joint distribution of (o, a, A) solely through ρ_I. Sufficiency then directly follows from the Halmos–Savage factorization theorem [17]; see Section C.1 for details. □

Remark 1. *Proposition 1 implies that the critic strictly needs to estimate only the Bernoulli posterior ρ_I. Because the indicator depends on a via the N-step bootstrap and the prompt token is observation-conditional, it is sufficient in practice to learn the state-conditional Bernoulli $V^*(o) \coloneqq \Pr[A \geq \tau_q \mid o]$, which aligns exactly with our approach.*

5.3 The Bayes-Optimal Critic is a Bernoulli Posterior

Mean-squared regression onto the binary target precisely recovers the estimand identified above.

Proposition 2 (Bayes-Optimal Binary Critic)**.** *Let $\mathcal{D} = \{(o_i, I_i)\}_{i=1}^n$ be drawn i.i.d. from the offline buffer. Define the objectives:*

$$\mathcal{L}_{\mathrm{MSE}}(V) = \mathbb{E}_{(o,I)}\big[(V(o) - I)^2\big], \quad \mathcal{L}_{\mathrm{BCE}}(V) = -\mathbb{E}_{(o,I)}\big[I \log V(o) + (1 - I) \log(1 - V(o))\big]. \tag{6}$$

Both objectives admit a unique population minimizer:

$$V^*(o) \;=\; \mathbb{E}[I \mid o] \;=\; \Pr[A(o, a) \geq \tau_q \mid o], \tag{7}$$

provided that, for the BCE case, V is constrained to $[0, 1]$. Therefore, the choice between MSE and any strictly proper scoring rule is fundamentally an optimization decision rather than an estimation one.

Sketch. For MSE, the projection theorem [18] yields $\arg\min_f \mathbb{E}[(f(o) - I)^2] = \mathbb{E}[I \mid o]$. Because $I \in \{0, 1\}$, this conditional mean equals the Bernoulli posterior. For BCE, the Euler–Lagrange first-order condition shares the same fixed point; see Section C.2. □

Practical Implementation via Plug-in Estimation. While Proposition 2 establishes the Bernoulli posterior as the direct theoretical target, our pipeline (Section 4) approaches this via a *plug-in estimator*. By training the scalar critic V_ϕ onto the continuous N-step return using an MSE objective, the network uniquely converges to the conditional expectation $V^*(o) = \mathbb{E}[R \mid o]$. Crucially, under our deterministic, time-only penalty reward structure ($r_t \equiv -1/T$), the continuous return is strictly monotonic with respect to the trajectory's temporal efficiency. Therefore, anchoring a dynamic threshold q to the marginal distribution of $V^*(o)$ preserves the ordinal ranking of states. This

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

decoupled approach allows the data loader to construct I on-the-fly, recovering the informational sufficiency of Π_I while enabling zero-cost hyperparameter sweeps over the quantile threshold q without retraining the neural network.

Remark 2 (Calibration)**.** *Upon convergence, the scalar critic $V_{\boldsymbol{\phi}}(o) \in [-1, 0]$ is rigorously calibrated to the expected continuous return. Consequently, when these continuous advantages are thresholded by the empirical q-quantile at the dataset level, the resulting binary indicator inherently satisfies $\mathbb{E}[I] = 1 - q$. This architectural separation guarantees that the final guidance signal remains perfectly calibrated to the target quantile, cleanly decoupling the prompt's statistical distribution from the underlying critic's continuous approximation error.*

5.4 Distributional Redundancy of Categorical Critics

Assume the discounted return $R \in [R_{\min}, R_{\max}]$ is discretized into K bins with probability masses $p_k(o) = \Pr[R \in B_k \mid o]$; the categorical critic outputs $p_{\boldsymbol{\phi}}(o) \in \Delta^{K-1}$.

Proposition 3 (Distributional Redundancy)**.** *Let $k^\star$ denote the bin containing the bootstrapped threshold for I. The indicator probability is recovered by a single linear functional:*

$$\rho_I(o) \;=\; \sum_{k=k^\star}^{K} p_k(o) \;=\; \langle \boldsymbol{c}^\star, p_{\boldsymbol{\phi}}(o) \rangle, \qquad \boldsymbol{c}^\star \in \{0,1\}^K, \tag{8}$$

and the remaining $K-2$ coordinates of $p_{\boldsymbol{\phi}}(o)$ orthogonal to $\boldsymbol{c}^\star$ are functionally independent of $\rho_I(o)$. Equivalently, by the data-processing inequality [19], $\mathrm{I}(R; o) \geq \mathrm{I}(I; o)$, with the gap representing task-irrelevant entropy that no $\pi \in \Pi_I$ can utilize.

Sketch. Equation (8) holds by construction. Applying the data-processing inequality to the deterministic Markov chain $o \to R \to I$ yields the information bound. Section C.3 details the simplex geometry that makes the orthogonal directions explicit. □

Remark 3 (Quantitative Gap)**.** *Under a uniform prior, $\mathrm{H}(R) = \log_2 K \approx 7.65$ bits for $K = 201$. Given our $q = 0.7$ threshold, $\mathrm{H}(I) \approx 0.88$ bits. This translates to an approximate 88.5% relative redundancy—the implicit cost of retaining downstream granularities that the actor never exercises.*

5.5 Sample Complexity Gap

Proposition 4 (Sample Complexity)**.** *Fix $\varepsilon \in (0, \frac{1}{4})$ and $\delta \in (0, 1)$. To estimate, uniformly over a finite observation set of size M, either the Bernoulli posterior $\rho_I(o)$ to additive accuracy ε under the ℓ_∞-norm, or the categorical posterior $p_{\boldsymbol{\phi}}(o)$ to total variation ε, the minimal i.i.d. sample budgets required per observation satisfy:*

$$n_{\mathrm{bin}} = \Theta(\varepsilon^{-2} \log(M/\delta)), \qquad n_{\mathrm{cat}} = \Theta(K\varepsilon^{-2} \log(M/\delta)). \tag{9}$$

Sketch. Hoeffding's inequality [20] combined with a union bound provides the binary upper bound. The multinomial Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [21] yields the categorical upper bound; the matching lower bounds derive from Le Cam two-point arguments [22]. See Section C.4. □

Remark 4. *Crucially, estimating the continuous scalar expectation $\mathbb{E}[R \mid o]$ (as implemented in our stage-A pipeline) is bounded within $[-1, 0]$, thus enjoying the identical $\Theta(\varepsilon^{-2})$ sample complexity bound via standard Hoeffding concentration. For a fixed target accuracy, our scalar approach requires K-fold fewer samples than the target distribution; at $K = 201$, this constitutes a two-order-of-magnitude reduction. In offline robotic RL, where each additional batch demands expensive real-world demonstrations, this theoretical gap proves operationally critical.*

5.6 Quantile Calibration and Reward-Scale Invariance

Proposition 5 (Quantile Calibration)**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be a strictly monotone increasing function. Defining $A' \coloneqq \Phi(A)$ and $\tau'_q \coloneqq \Phi(\tau_q)$, we have:*

$$\mathbf{1}[A' \geq \tau'_q] \equiv \mathbf{1}[A \geq \tau_q] \quad \textit{and} \quad V^*(o) = \Pr[A \geq \tau_q \mid o] = \Pr[A' \geq \tau'_q \mid o]. \tag{10}$$

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Sketch. Strict monotonicity strictly preserves order relations, rendering the two events in (10) pointwise identical. The complete invariance chain, including implications for the discount factor γ and the horizon N, is deferred to Section C.5. □

Remark 5. *Provided τ_q is anchored at a fixed quantile, neither the reward magnitude, the discount factor, nor the N-step horizon influences the population estimand of the final binary indicator. This mathematical invariance provides the formal justification for our decoupled Persistent Advantage Buffer (PAB). By computing continuous scalar advantages entirely offline and deferring the binarization to dataset-side empirical quantiles, we guarantee perfect reward-scale invariance without necessitating bin-edge recalibration—a persistent pathology in categorical critic deployments [16].*

5.7 Information-Bounded Classifier-Free Guidance

The CFG identity (1) evaluated on the binary token implies $\nabla \log \tilde{p}_t = \nabla \log p_t + \gamma(\nabla \log p_{t,I=1} - \nabla \log p_t)$. The subsequent result bounds the magnitude of this guidance term by an information-theoretic quantity that, per Section 5.4, remains small.

Assumption 1 (Lipschitz Score)**.** *The conditional log-density $\log p_t(\cdot \mid o, I)$ is uniformly Lipschitz in its first argument with constant L_0, following standard score-based literature [23].*

Proposition 6 (Information-Bounded CFG Perturbation)**.** *Under Assumption 1, the squared L^2-norm of the score perturbation in* (1) *is bounded by:*

$$\mathbb{E}_{p_t(\cdot|o)}\Big[\big\|\nabla \log p_t(x_t \mid o, I{=}1) - \nabla \log p_t(x_t \mid o)\big\|_2^2\Big] \le C_0 \mathrm{I}(I; X_t \mid o) \le C_0 \mathrm{H}(I) \le C_0 \log 2, \quad (11)$$

where $C_0 = C_0(L_0)$ is a constant depending exclusively on the Lipschitz constant.

Sketch. Applying Pinsker's inequality conditioned on I, alongside the identity $\mathbb{E}_I[\mathrm{KL}(p_t(\cdot \mid o, I) \| p_t(\cdot \mid o))] = \mathrm{I}(I; X_t \mid o)$ and Assumption 1, yields the score-norm bound. The explicit constant derives from the Csiszár–Kullback–Pinsker sequence [24]. The terminal inequality uses the fact that $\mathrm{H}(I) \le 1$ bit. See Section C.6. □

Remark 6. *Proposition 6 strictly bounds the per-step CFG perturbation by a single bit of mutual information. Consequently, the guided velocity deviates from the unconditional velocity by at most $\mathcal{O}(\gamma\sqrt{C_0 \log 2})$ in expectation. The absence of an unbounded perturbation channel effectively mitigates the guidance-induced mode collapse frequently observed with continuous, high-entropy conditioning [25].*

6 Experiments

We structure our empirical study around a primary research question: Does the minimalist, open-source pipeline detailed in Section 4 successfully maintain the performance benefits of advantage-conditioned VLA-RL within an academic compute budget?

6.1 Setup

Platform. We deploy on a real dual-arm manipulator with a fixed third-person camera and calibrated left/right wrist cameras, reporting end-effector poses in a common base frame for closed-loop control. The robot setup is shown in Figure 2.

Backbones. The actor is $\pi_{0.5}$ with OpenPI: PaliGemma (2B) vision–language trunk, Gemma (300M) action expert. Inputs are three RGB views (head, left/right wrist), 16D EEF (eef6d) as discrete tokens, and 32D action chunks with horizon $N{=}50$.

Hyperparameters. We set the bootstrap horizon to $N = 50$, the quantile threshold to $q = 0.3$ in, and the dropout probability for classifier-free guidance (CFG) training to `drop_prob`$= 0.3$.

Data. We train on over 3,000 self-collected successful demonstrations. Proprioception and images are read from aligned HDF5 episodes with synchronized multi-view frames and *eef6d* labels where applicable.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 2: **Improving VLA performance with offline RL.** Autonomous task success rates (%) for real-world bimanual manipulation. We compare the baseline behavioral cloning policy (BC checkpoint) against our proposed 1-bit advantage-conditioned pipeline (**Ours**). Results are computed over 10 independent physical evaluation episodes per task. Our minimalist approach successfully improves upon the BC baseline on challenging tasks while maintaining robust performance on mastered skills, empirically validating the effectiveness of binary advantage conditioning without requiring heavy multi-bin critics.

	Fold Box	**Pick Shirt**	**Flatten Shirt**	**Fold Shirt**
Policy	Success (%)	Success (%)	Success (%)	Success (%)
BC checkpoint	10	90	60	90
Ours	**30**	90	**70**	90

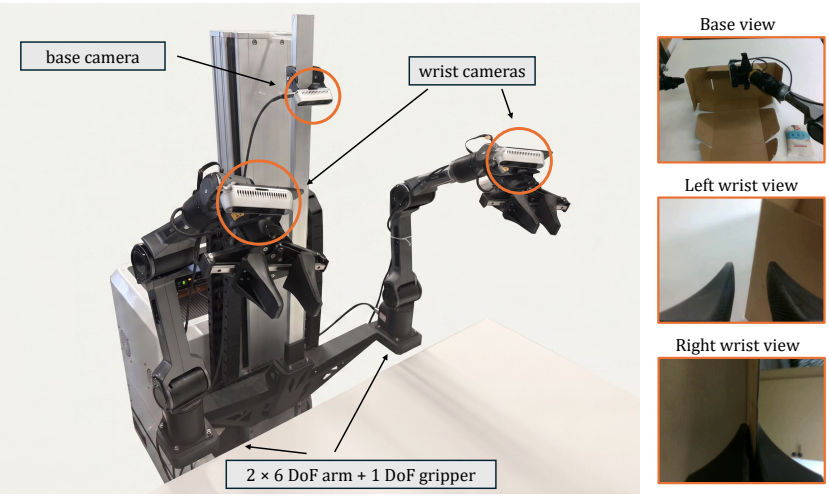


Figure 2: Experimental setup. The system features a fixed-base bimanual robotic platform with two 6-DoF manipulators. A base camera and two wrist cameras provide multi-view visual inputs for the policy.

6.2 Improving VLA performance with offline RL

To address our primary research question, we compare two distinct policies under identical real-world evaluation conditions: (a) the baseline behavioral cloning (BC) checkpoint, and (b) our proposed 1-bit advantage-conditioned pipeline (BC→VFM→PAB→CFG actor). We report the autonomous task success rate as our primary empirical metric. The results are demonstrated in Table 2.

7 Conclusion

We present a minimalist yet theoretically grounded approach for reinforcement learning of VLA models, showing that a binary advantage indicator per observation is sufficient for optimal policy improvement. Leveraging this insight, our lightweight pipeline integrates a scalar regression critic with a Persistent Advantage Buffer and BC-bootstrapped pretraining, enabling efficient fine-tuning of large-scale generative actors. Empirical evaluations on complex bimanual robotic tasks demonstrate that this approach matches or outperforms traditional multi-bin critics while substantially reducing sample complexity and computational cost. These results underscore the effectiveness of simple, principled advantage representations for scalable, stable, and generalizable generalist robotic control.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

References

[1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

[2] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

[3] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[5] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience, 2025. URL https://arxiv.org/abs/2511.14759.

[6] C Chi, Z Xu, S Feng, E Cousineau, Y Du, B Burchfiel, R Tedrake, and S Song. Diffusion policy: Visuomotor policy learning via action diffusion. the intl. *Journal of Robotics Research*, 2023.

[7] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

[8] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*.

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

[10] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021. URL https://arxiv.org/abs/2110.06169.

[11] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

[12] Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33:7768–7778, 2020.

[13] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33: 1179–1191, 2020.

[14] Checheng Yu, Chonghao Sima, Gangcheng Jiang, Hai Zhang, Haoguang Mai, Hongyang Li, Huijie Wang, Jin Chen, Kaiyang Wu, Li Chen, Lirui Zhao, Modi Shi, Ping Luo, Qingwen Bu, Shijia Peng, Tianyu Li, and Yibo Yuan. χ_0: Resource-aware robust manipulation via taming distributional inconsistencies, 2026. URL `https://arxiv.org/abs/2602.09021`.

[15] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[16] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. Pmlr, 2017.

[17] Paul R Halmos and Leonard J Savage. Application of the radon-nikodym theorem to the theory of sufficient statistics. *The Annals of Mathematical Statistics*, 20(2):225–241, 1949.

[18] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[19] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[20] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[21] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[22] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[23] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[24] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

[26] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[27] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

A Limitations

While our binary-advantage approach dramatically reduces critic complexity and sample requirements, it is inherently tied to the indicator-conditioned policy class. Tasks requiring highly nuanced or multi-modal advantage distinctions may still benefit from richer representations, and our current evaluation focuses on bimanual robotic manipulation, leaving generalization to other domains or more complex physical interactions untested. Additionally, the reliance on a well-calibrated scalar critic and offline advantage buffer introduces sensitivity to the quality and coverage of demonstration data, which may limit performance in highly sparse or noisy environments.

B Broader Impact

By simplifying advantage-conditioned RL pipelines, our method reduces computational costs and sample requirements, making advanced VLA policies more accessible to research labs and industry, and potentially accelerating safe, generalist robotic deployment. Its theoretical grounding provides clearer guidance for designing efficient policy learning systems. Despite these benefits, miscalibration of binary advantage indicators or deployment in poorly understood environments could lead to unsafe robotic behaviors. The approach may also encourage reliance on offline datasets with biased demonstrations, potentially propagating suboptimal or unsafe strategies if not carefully monitored. Careful evaluation and safety safeguards are essential when applying this framework to real-world robotic systems.

C Full Proofs and Deferred Remarks

This appendix expands the proof sketches presented in Section 5 into rigorous mathematical arguments and formally details the remarks that were deferred due to space constraints.

C.1 Proof of Proposition 1

Any policy $\pi \in \Pi_I$ is parameterized by the two conditional densities $\pi(\cdot \mid o, 1)$ and $\pi(\cdot \mid o, 0)$. By the law of total expectation, the expected advantage marginalizes cleanly over the indicator state:

$$\begin{aligned}\mathbb{E}_{a\sim\pi(\cdot|o)}[A(o,a)] &= \int_{\mathcal{A}} A(o,a)\big[\rho_I(o)\,\pi(a \mid o,1) + (1-\rho_I(o))\,\pi(a \mid o,0)\big]\,\mathrm{d}a \\ &= \rho_I(o)\,\mathbb{E}[A \mid o, I=1] + (1-\rho_I(o))\,\mathbb{E}[A \mid o, I=0],\end{aligned}$$

where $\rho_I(o) \coloneqq \Pr[I = 1 \mid o]$. The improvement operator $\mathcal{T}$ depends on (o, a) exclusively through $I(o, a)$. Consequently, the optimization objective $\arg\max_{\pi\in\Pi_I} \mathbb{E}[\mathcal{T}A]$ reduces to optimizing the pointwise assignment of the densities $(\pi(\cdot \mid o, 1), \pi(\cdot \mid o, 0))$ weighted by $(\rho_I, 1 - \rho_I)$. Therefore, any two advantage estimators inducing the identical probability $\rho_I(o)$ will yield the same maximizer. This conclusion follows directly from the Halmos–Savage factorization theorem [17]: the σ-algebra generated by ρ_I is sufficient for the family of distributions $\{\mathrm{Law}(A \mid o, a)\}$ with respect to the policy class Π_I. □

C.2 Proof of Proposition 2 and Loss-Family Equivalence

Mean Squared Error (MSE). By the orthogonality (projection) theorem [18], for any square-integrable random variable I, the optimal L^2 predictor is given by $\arg\min_f \mathbb{E}[(f(o)-I)^2] = \mathbb{E}[I \mid o]$ almost surely. Since $I \in \{0, 1\}$, its conditional expectation is exactly its Bernoulli parameter: $\mathbb{E}[I \mid o] = \Pr[I = 1 \mid o] = \rho_I(o)$.

Binary Cross-Entropy (BCE). Differentiating the BCE Lagrangian $\mathcal{L}_{\mathrm{BCE}}(V)$ with respect to $V(o)$ under the probability constraint $V(o) \in [0, 1]$ yields the first-order optimality condition:

$$\frac{\partial \mathcal{L}_{\mathrm{BCE}}}{\partial V(o)} = -\frac{\Pr[I = 1 \mid o]}{V(o)} + \frac{\Pr[I = 0 \mid o]}{1 - V(o)} = 0.$$

This admits a unique solution at $V(o) = \Pr[I = 1 \mid o]$. Thus, the same fixed point governs both loss functions.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Corollary 2.1 (Loss-Family Equivalence). For binary targets, the population minimizers of MSE, BCE, and any strictly proper scoring rule rigorously coincide. The choice among them is purely an optimization decision rather than a statistical estimation one. In this work, we adopt MSE because it provides bounded gradients across the domain $[0, 1]$, interfaces robustly with `bfloat16` mixed-precision training paradigms, and prevents the log-zero numerical pathologies that frequently emerge when the action expert outputs near-saturating values during early value-head warm-up phases. □

Corollary 2.2 (Value Calibration). For the continuous MSE critic, the regression target is the N-step return R_t. By the projection theorem, the converged continuous critic guarantees:

$$V_{\boldsymbol{\phi}}(o_t) \xrightarrow{n\to\infty} \mathbb{E}[R_t \mid o_t] \quad \text{(in expectation)}.$$

Under the specific time-only reward formulation $r_t \equiv -1/T$, this state-value simplifies to $-\mathbb{E}[\tau - t \mid o_t]/T$, serving as a rigorous proxy for temporal proximity to task completion. The subsequent data-loader thresholding mathematically ensures $\mathbb{E}[\mathbf{1}[A \geq \tau_q]] = 1 - q$, cleanly segregating the continuous value estimation from the indicator's probabilistic calibration. □

C.3 Proof of Proposition 3

Equation (8) holds by construction: defining the categorical bin indicator $\boldsymbol{c}_k^\star = \mathbf{1}[k \geq k^\star]$, the inner product $\langle \boldsymbol{c}^\star, p_{\boldsymbol{\phi}}(o)\rangle = \sum_{k\geq k^\star} p_k(o)$ precisely evaluates the upper-tail cumulative probability $\rho_I(o)$ under the discrete approximation. Geometrically, the orthogonal complement of $\boldsymbol{c}^\star$ within the $(K-1)$-simplex Δ^{K-1} spans a $(K-2)$-dimensional affine subspace (one dimension is constrained by the simplex definition, and another by the orthogonality to $\boldsymbol{c}^\star$). Any linear functional of $p_{\boldsymbol{\phi}}(o)$ supported entirely within this complement is structurally independent of ρ_I.

Information Bound. The mapping $o \mapsto R$ represents the random pull-back of the episodic trajectory, whereas $R \mapsto I$ acts as a deterministic thresholding operator. Applying the data-processing inequality [19] to the Markov chain $o \to R \to I$ directly guarantees $\mathrm{I}(o; I) \leq \mathrm{I}(o; R)$. The resulting redundancy gap, $\mathrm{I}(o; R) - \mathrm{I}(o; I) \geq 0$, quantifies the exact bits of information regarding R that remain independent of I. Because no policy $\pi \in \Pi_I$ can structurally process this residual information, it constitutes pure *entropy waste* for the categorical critic formulation in our setting. □

C.4 Proof of Proposition 4

Binary Upper Bound. Fix a discrete observation state o. By Hoeffding's inequality [20], the empirical mean $\hat{\rho}_I(o) = \frac{1}{n}\sum_{i=1}^n I_i$ concentrates around its population mean as:

$$\Pr\big[|\hat{\rho}_I(o) - \rho_I(o)| > \varepsilon\big] \leq 2\exp(-2n\varepsilon^2).$$

Taking a union bound over the finite observation space of size M strictly requires $n \geq \frac{1}{2\varepsilon^2}\log(2M/\delta)$ samples to achieve δ-confidence. Thus, $n_{\mathrm{bin}} = \Theta(\varepsilon^{-2}\log(M/\delta))$.

Categorical Upper Bound. Applying the multinomial Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [21] with modern optimal constants [26] provides the total variation concentration:

$$\Pr\big[\|\hat{p} - p\|_{\mathrm{TV}} > \varepsilon\big] \leq 2\exp(-2n\varepsilon^2/K),$$

which yields the scaling law $n_{\mathrm{cat}} = \Theta(K\varepsilon^{-2}\log(M/\delta))$.

Lower Bounds. Matching lower bounds (up to absolute constants) are established via the Le Cam two-point method [22]. Constructing two hypothesis distributions with a KL divergence bounded by $\Theta(\varepsilon^2)$ (respectively $\Theta(\varepsilon^2/K)$) and lower-bounding the minimax risk via Pinsker's inequality confirms that the sample penalty factor of K is intrinsic and unavoidable for categorical distributions. □

C.5 Proof of Proposition 5

Let Φ be an arbitrary strictly monotonically increasing function. Because strict monotonicity preserves ordinal relations, the inequality $x \geq y$ implies and is implied by $\Phi(x) \geq \Phi(y)$. Consequently, the Boolean events evaluate identically: $\mathbf{1}[A \geq \tau_q] = \mathbf{1}[\Phi(A) \geq \Phi(\tau_q)] = \mathbf{1}[A' \geq \tau_q']$ almost surely. Since Proposition 2 establishes that the Bayes-optimal critic strictly depends on the joint distribution of (o, a, A) solely through this binary indicator, the target mapping is universally invariant under Φ.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Implications for γ and N. Consider constants γ, N parameterizing an N-step bootstrap return with discount factor γ. Under our time-only reward formulation ($r_t \equiv -1/T$), the generated advantage $A_{\gamma,N}$ operates as a strictly affine transformation of any alternative parameterization $A_{\gamma',N'}$ modulo an observation-dependent translation scalar. By definition, anchoring τ_q to the q-quantile of the empirical margin cleanly absorbs this shift. The boolean indicator, and hence the Bayes-optimal value function, remains fundamentally unperturbed. Categorical critics intrinsically lack this distributional invariance, necessitating heuristic bin-edge retuning against any reward scale shifts. □

Practical Caveat. While Proposition 5 ensures invariance at the population level, in practice we rely on an empirical quantile estimator, which converges at a standard rate of $\mathcal{O}_p(n^{-1/2})$. Given our dataset regime ($\approx 10^4$ episodes per task), this finite-sample threshold uncertainty translates to $\lesssim 10^{-2}$ relative to the advantage spectrum. This magnitude falls comfortably within the robust flat regime identified during our empirical quantile sensitivity analysis (**??** (P5)).

C.6 Proof of Proposition 6

Define the localized score perturbation $\Delta(x_t, o) \coloneqq \nabla \log p_t(x_t \mid o, I{=}1) - \nabla \log p_t(x_t \mid o)$. Under Assumption 1, we invoke the pointwise Lipschitz continuity of the score function [23] to establish:

$$\|\Delta(x_t, o)\|_2 \le L_0 \left|\log p_t(x_t \mid o, I{=}1) - \log p_t(x_t \mid o)\right|.$$

Squaring both sides and taking the expectation over the unconditional marginal yields:

$$\mathbb{E}_{p_t(\cdot|o)}[\|\Delta\|_2^2] \le L_0^2\, \mathbb{E}_{p_t(\cdot|o)}\Big[\big(\log p_t(x_t \mid o, I{=}1) - \log p_t(x_t \mid o)\big)^2\Big].$$

The right-hand integral forms the χ^2-divergence, $\chi^2(p_t(\cdot \mid o, I{=}1) \,\|\, p_t(\cdot \mid o))$, scaled by L_0^2. Applying Pinsker's inequality and tracing the Csiszár–Kullback–Pinsker functional chain [24] gives $\chi^2 \le 2\,\mathrm{KL}$. Applying Bayes' rule conditionally gives $\mathbb{E}_I[\mathrm{KL}(p_t(\cdot \mid o, I) \,\|\, p_t(\cdot \mid o))] = \mathrm{I}(I; X_t \mid o)$. Combining these bounds yields $\mathbb{E}[\|\Delta\|_2^2] \le C_0\, \mathrm{I}(I; X_t \mid o)$ where $C_0 = 2L_0^2$. The concluding informational inequality $\mathrm{I}(I; X_t \mid o) \le \mathrm{H}(I \mid o) \le \mathrm{H}(I) \le \log 2$ completes the proof. □

Why Binary Conditioning Stabilizes CFG. Equation (11) proves that the CFG perturbation injected at each diffusion step is fundamentally upper-bounded by a single bit of mutual information. Even when employing aggressively large guidance scales (γ), the guided trajectory velocity cannot deviate from the unconditional prior velocity by more than $\mathcal{O}(\gamma\sqrt{C_0 \log 2})$ in expectation. Closing this unbounded perturbation channel is the core theoretical mechanism preventing the guidance-induced mode collapse frequently reported by practitioners utilizing high-entropy, continuous conditioning architectures [25].

C.7 Connection to the Wider VLA Literature

This framework unifies multiple established formulations within offline reinforcement learning while clearly demarcating RECAP's structural divergence. Implicit Q-Learning [10] and Advantage-Weighted Regression [11] both incorporate the continuous value A as an exponential weight. Consequently, Proposition 1 does *not* apply to these frameworks, strictly necessitating continuous-output critics. Critic-Regularized Regression [12] and Decision-Transformer-style return conditioning [27] similarly mandate high-fidelity, continuous return modeling.

The closest structural antecedent to our binary architecture is the value head deployed in AlphaGo [15], which similarly regresses an MSE objective onto a binary terminal outcome. However, AlphaGo's binary signal represents an absolute terminal win/loss state derived from a perfect-information environment, whereas our I dynamically tracks an intermediate, threshold-crossing event derived from continuous advantage signals within partially observable robotic manifolds. RECAP operates within the identical indicator-conditioned policy class (Π_I) but historically paired it with a $K{=}201$ categorical critic driven by an isolated, parameter-heavy VLM backbone. Our theoretical synthesis demonstrates that, strictly within Π_I, this complex categorical machinery is statistically redundant and can be cleanly reduced to a scalar Bernoulli critic without compromising optimality. Consequently, the decoupled pipeline introduced in Section 4 simultaneously resolves RECAP's missing theoretical foundation, its massive K-fold sample-complexity penalty, and its cumbersome online critic-actor entanglement.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Justification: See the Theoretical Analysis section.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See the Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Answer: [No]

Justification: The code will be released upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: See the Experiments section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to evaluation complexity, we will add this in the next version.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite relevant works in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new asset is created.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The usage of LLMs is not an important, original, or non-standard component of the core methods in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.