Extended Abstract

Extended Abstract

Behavioral cloning provides a straightforward framework for imitation learning by training policies to mimic expert demonstrations. However, its vulnerability to compounding errors—arising from deviations that accumulate over time—limits its reliability in stochastic or dynamically changing environments. Recent methods have addressed this by employing action chunking, wherein the policy predicts temporally extended action sequences rather than single-step actions. While this enhances temporal consistency and captures higher-order behavior patterns, it introduces a tradeoff: longer chunks reduce reactivity, especially under uncertain dynamics.

Bidirectional Decoding (BID) has been proposed as a solution to this tradeoff, selecting action chunks based on a weighted combination of forward contrast and backward coherence. However, BID applies these weights uniformly across timesteps, ignoring environmental context and test-time uncertainty. In high-stochasticity regimes, this leads to suboptimal chunk selection.

We propose TV-BID, a test-time adaptive extension of BID that modulates its forward-backward loss terms using a measure of stochasticity derived from Total Variation Distance (TVD) between consecutive action chunk distributions. TVD offers a bounded, symmetric, and efficiently computable proxy for test-time distributional drift, allowing the agent to adjust its planning strategy based on observed temporal variability. We introduce both pairwise and sliding-window variants of TVD estimation, enabling fine-grained control over how past action predictions influence current loss weighting.

Our implementation discretizes the action space into $b \times b$ histograms and compares chunk-wise distributions over overlapping timesteps. We evaluate our approach in both closed-loop (AH $= 1$) and open-loop (AH $= 3$) control regimes on the Push-T manipulation task using the VQ-BET policy. Across a range of noise levels, TV-BID consistently outperforms baseline BID, particularly in high-stochasticity settings. Notably, our open-loop policy—augmented with TVD-based adaptation—achieves higher success rates than its closed-loop counterpart, demonstrating that adaptive chunk selection can mitigate the reactivity limitations of chunked policies.

These results highlight the effectiveness of simple distributional statistics as signals for test-time adaptation in imitation learning. By incorporating stochasticity-aware loss modulation, TV-BID bridges the gap between stable long-horizon planning and reactive short-horizon correction without requiring additional supervion or retraining.

Test-Time Stochasticity Estimation for Adaptive Action Chunk Selection

Sarosh Khan
Department of Computer Science
Stanford University
skhan44@stanford.edu

Ellie Tanimura
Department of Electrical Engineering
Stanford University
etanim@stanford.edu

Abstract

Behavioral cloning (BC) is a widely used imitation learning method, but it suffers from compounding errors caused by distributional drift at test time. Action chunking—predicting temporally extended sequences—offers improved consistency but reduces responsiveness in dynamic environments. Bidirectional Decoding (BID) aims to balance consistency and reactivity by selecting action chunks based on forward contrast and backward coherence scores. However, BID statically weights these scores, ignoring test-time uncertainty. We propose TV-BID, a simple extension that adapts this weighting based on the Total Variation Distance (TVD) between consecutive action chunk distributions. TVD acts as a proxy for environmental stochasticity, allowing the policy to dynamically prioritize reactivity or stability. We introduce both pairwise and sliding-window TVD estimation schemes and evaluate them under closed-loop and open-loop execution settings on the Push-T manipulation task using the VQ-BET policy. Our method improves robustness in high-noise conditions without requiring retraining, demonstrating that lightweight test-time adaptation can significantly enhance chunked behavior policies.

1 Introduction

Behavioral Cloning (BC) offers a simple and effective paradigm for learning control policies from demonstration. However, a long-standing limitation of BC lies in its sensitivity to covariate shift: small deviations from expert trajectories can accumulate over time, leading to catastrophic distribution drift and compounding error Ross et al. (2011).

Recent approaches address this issue by leveraging *action chunking*, in which the policy predicts temporally extended sequences $(a_t, a_{t+1}, \ldots, a_{t+\ell})$ instead of stepwise actions. Chunking improves temporal consistency and encodes latent structure across actions—such as subgoals or stylistic behavior—but sacrifices short-horizon adaptability. In particular, long chunks may become unreliable under stochastic transitions or perceptual noise, where frequent replanning is crucial.

Bidirectional Decoding (BID) Liu et al. (2024) was proposed to navigate this consistency-reactivity tradeoff. At each timestep, BID samples multiple candidate chunks and selects the optimal one via a weighted combination of two scores: *Backward Coherence*, which encourages alignment with the previously executed chunk, and *Forward Contrast*, which favors distinguishability from weaker decoders. However, BID applies these weights uniformly, regardless of context or uncertainty. This potentially leads to overly rigid plans in volatile settings, or overly reactive plans in stable ones.

This paper introduces a simple yet effective extension to BID: adaptively mixing the backward and forward terms based on the *Total Variation Distance* (TVD) between consecutive action distributions. The key idea is to treat TVD as a proxy for test-time stochasticity—when predicted distributions are stable across time, TVD is low, and the loss encourages backward continuity; when distributions shift significantly, TVD rises, and the loss instead promotes contrastive replanning.

Stanford CS224R 2025 Final Report

To compute TVD, we discretize the continuous action space into $b \times b$ histograms and compare distributions across overlapping chunk predictions. We consider both pairwise and sliding-window variants, and evaluate under both closed-loop (AH $= 1$) and open-loop (AH $= 3$) policies.

Empirically, our TVD-augmented decoder outperforms baseline BID in stochastic environments, showing improved robustness without retraining. Our results suggest that temporal distributional shift is a strong signal for adaptive planning in chunked policies — and that even simple statistical measures like TVD can yield powerful inductive biases when applied at test time.

2 Related Work

2.1 Behavioral Cloning and Compounding Errors

Behavioral cloning has been widely used for robot learning from human demonstrations due to its simplicity and ability to leverage large-scale human-collected datasets (Atkeson and Schaal (1997)). However, a fundamental challenge with BC is the phenomenon of compounding errors: small prediction mistakes in a trajectory can lead to states outside the training distribution, causing subsequent predictions to degrade further (Ross et al. (2011)). Several works have proposed strategies to mitigate compounding errors, including data aggregation through expert interventions (Ross et al. (2011); Mendonca et al. (2021)) and noise injection during training (Laskey et al. (2017)). Nevertheless, these solutions require additional supervision or assumptions during training and do not directly address test-time recovery.

2.2 Action Chunking and Temporal Dependencies

To improve temporal consistency and robustness to human demonstration variability, recent methods have turned to action chunking (Zhao et al. (2023); Chi et al. (2024)). Rather than predicting a single action at each timestep, the policy predicts a sequence of future actions, capturing long-term latent strategies such as multi-step planning and style preferences. Action chunking has been shown to mitigate short-term noise sensitivity and better model idle behaviors and latent subgoals in human demonstrations (Chi et al. (2024); Lee et al. (2024)). However, by committing to longer action horizons, action chunking also reduces the policy's ability to react quickly to unexpected environmental changes.

2.3 Bidirectional Decoding for Closed-Loop Operations

BID (Liu et al. (2024)) was proposed to address the trade-off between temporal consistency and reactivity in action-chunked policies. Instead of committing to a single predicted action chunk, BID samples multiple candidate chunks at each timestep and selects the best one based on two factors:

Backward Coherence: This parameter encourages continuity between successive action chunks by favoring candidates that align closely with the previously executed chunk. It is computed as a weighted sum of Euclidean distances over the overlapping steps:

$$\mathcal{L}_B(a) = \sum_{\tau=0}^{\ell-1} \rho^\tau \|a_{t+\tau}^{(t)} - a_{t+\tau}^{(t-1)}\|_2$$

where:$a_{t+\tau}^{(t)}$ is the action predicted at time $t+\tau$ from the current chunk sampled at timestep t, $a_{t+\tau}^{(t-1)}$ is the action at the corresponding timestep from the previously selected chunk, ℓ is number of predicted future steps, or the length of the chunk, $\rho^\tau \in (0, 1]$ is a temporal decay factor that lowers the importance of similar actions later in the action chunk.

Forward Contrast: This parameter promotes selecting chunks that are more similar to predictions made by a stronger policy and dissimilar to those from a weaker policy from earlier points in training. It is computed as:

$$\mathcal{L}_F(a) = \frac{1}{N}\left(\sum_{a^+ \in \mathcal{A}^+}\sum_{\tau=0}^{\ell-1} \|a_{t+\tau}^{(t)} - a_{t+\tau}^{+}\|_2 - \sum_{a^- \in \mathcal{A}^-}\sum_{\tau=0}^{\ell-1} \|a_{t+\tau}^{(t)} - a_{t+\tau}^{-}\|_2\right)$$

where: $\mathcal{A}^+$ is the set of action chunks sampled from a strong policy π, $\mathcal{A}^-$ is the set of action chunks sampled from a weak policy π', and N is the total number of sampled action chunks used to compute the loss.

BID applies a uniform balancing of these two losses and selects action chunks that minimize a combination of $\mathcal{L}_B$ and $\mathcal{L}_F$. However, this method does not dynamically adapt based on environment stochasticity. As such, BID especially struggles during closed loop operation in high stochasticity environments.

2.4 Environment-Aware Decision-Making and Test-Time Adaptation

Several works have explored adjusting behavior at test time. These works utilize techniques based on learned uncertainty estimators or other external guidance signals (Dhariwal and Nichol (2021); Meister et al. (2023)). Recent work in value-guided sampling (Nakamoto et al. (2025)) shows that conditioning sampling on reward estimates can improve robustness. However, there remains limited exploration into dynamically adapting decoding strategies based on environment stochasticity during execution. Our proposed method builds on these ideas by designing a stochasticity estimator that modulates the BID loss dynamically, allowing the policy to intelligently trade off between consistency and reactivity at each timestep based on situational awareness.

3 Method

Since BID's rigidity is problematic in environments with varying levels of stochasticity, we propose TV-BID and Sliding Window BID, both of which use Total Variation Distance (TVD) to estimate the environmental stochasticity from action distribution shifts and adjust the ratio between forward contrast and backward coherence.

Total Variation Distance. Total Variation Distance is a fundamental divergence measure between probability distributions:

$$\text{TVD}(P, Q) = \tfrac{1}{2} \sum_i |P_i - Q_i|$$

We chose TVD because it satisfies several key properties that make it particularly well-suited for our test-time policy adaptation framework. First, it is both bounded and interpretable: it lies in the interval $[0, 1]$, and directly measures the maximal possible shift in probability mass between two distributions, such that a value of 1 indicates total disjointness, and a value of 0 implies identity. This makes the function smooth, predictable, and easy to tune, unlike alternatives such as KL divergence, which can diverge or produce very large values under distribution mismatch. Furthermore, TVD is symmetric, satisfying $\text{TVD}(P, Q) = \text{TVD}(Q, P)$. This symmetry allows us to compare past and current action distributions in an unbiased manner, which is especially important in test-time settings where neither distribution can be treated as the definitive reference. Finally, the computation of TVD is efficient and scalable. It reduces to an ℓ_1-norm over discretized histograms and requires no additional model inference or gradient tracking. This makes it practical for real-time adaptation, where we must compute divergence estimates from sampled actions under strict latency constraints.

Taken together, these properties make TVD not only theoretically principled but also empirically stable.

TV-BID. At each timestep t, the model samples N action chunks:

$$A_t^{(i)} = \left\{ a_t^{(i)}, a_{t+1}^{(i)}, \ldots, a_{t+\text{PH}-1}^{(i)} \right\}, \quad i = 1, \ldots, N$$

where PH is the prediction horizon and $a_{t+\tau}^{(i)} \in \mathbb{R}^d$ is a continuous action vector. The action horizon $\text{AH} \leq \text{PH}$ specifies how many of the predicted actions are actually executed before resampling the next chunk. We evaluate both closed-loop ($\text{AH} = 1$) and open-loop ($\text{AH} = 3$) policies.

To quantify temporal variation in predicted behavior, we convert each timestep's N predicted actions into a discrete distribution. Specifically, we bin the continuous actions into a $b \times b$ histogram over the 2D action space and normalize to form a distribution:

$$P_{t+\tau} \in \Delta^{b^2}, \quad \text{where } \Delta^{b^2} \text{ is the probability simplex over } b^2 \text{ bins.}$$

To detect distributional shifts between consecutive action chunks, we compute the Total Variation Distance (TVD) between histograms at overlapping timesteps. Let A_t^{prev} and $A_{t+\text{AH}}^{\text{curr}}$ be two chunks sampled at times t and $t + \text{AH}$, respectively. Since each chunk spans PH future steps, these two chunks overlap for $\text{PH} - \text{AH}$ timesteps, specifically from $t + \text{AH}$ to $t + \text{PH} - 1$.

We compute the TVD at each overlapping timestep $t_i = t + \text{AH} + x - 1$, for $x = 1, \ldots, \text{PH} - \text{AH}$:

$$\text{TVD}_{t_i} = \tfrac{1}{2} \sum_j \left| P_{t_i}^{\text{curr}}[j] - P_{t_i}^{\text{prev}}[j] \right|.$$

Finally, we compute a temporally weighted average of the per-timestep TVDs:

$$\overline{\text{TVD}} = \sum_{x=1}^{\text{PH}-\text{AH}} \rho_x \cdot \text{TVD}_{t+x+\text{AH}-1}, \quad \rho_x = \frac{\gamma^x}{\sum_{j=1}^{\text{PH}-\text{AH}} \gamma^j},$$

where $\gamma \in (0, 1]$ is a decay factor that emphasizes more recent discrepancies.

This formulation generalizes to both frequent resampling (closed-loop) and temporally extended execution (open-loop), enabling TV-BID to dynamically adapt to action distribution shifts regardless of the policy horizon.

Sliding Window BID (Only for AH $= 1$). In this second variant, we attempt to smooth out short-term noise and determine a more stable estimate of environmental stochasticity by comparing current action distributions to a sliding weighted average over the past N timesteps, where N is the prediction horizon.

At each timestep t in our chunk, we consider previously sampled action chunks that contain predictions for timestep t. Each of these past chunks yields a distribution P_{t-k} that aligns with the current timestep t, where $k = 1, 2, \ldots, \text{PH} - 1$ represents how far back the chunk was sampled. We compute the Total Variation Distance (TVD) between the current distribution P_t and each of these overlapping past distributions:

$$\text{TVD}_t(k) = \tfrac{1}{2} \sum_j \left| P_t[j] - P_{t-k}[j] \right|.$$

To emphasize more recent chunks, we apply exponentially decaying weights:

$$\rho_k = \frac{\gamma^k}{\sum_{j=1}^{\text{PH}-1} \gamma^j}, \quad \text{and} \quad \overline{\text{TVD}}_t = \sum_{k=1}^{\text{PH}-1} \rho_k \cdot \text{TVD}_t(k).$$

For example, to compute $\overline{\text{TVD}}_{t_5}$, we compare the current distribution at t_5 to distributions at the same timestep generated from previous chunks sampled at t_1 through t_4. If we denote these past chunk-aligned distributions as A, B, C, D, and the current one as E, then:

$$\overline{\text{TVD}}_{t_5} = \rho_1 \cdot \text{TVD}(D, E) + \rho_2 \cdot \text{TVD}(C, E) + \rho_3 \cdot \text{TVD}(B, E) + \rho_4 \cdot \text{TVD}(A, E).$$

This sliding window approach should help smooth out short-term noise and provides a more stable estimate of temporal change.

Loss Mixing

Given the $\overline{\text{TVD}}_t$ from these approaches, we compute a nonlinear mixing coefficient:

$$\sigma_t = \tanh\left(s \cdot \overline{\text{TVD}}_t\right)$$

The final loss becomes:

$$\mathcal{L}_t = \sigma_t \cdot \mathcal{L}_F + (1 - \sigma_t) \cdot \mathcal{L}_B$$

where $\mathcal{L}_F$ is the forward contrastive loss and $\mathcal{L}_B$ is the backward coherence loss.

4 Experimental Setup

4.1 Task and Policy

We evaluate TVD-BID and Sliding Window BID on the **Push-T** task from Diffusion Policy Chi et al. (2024), where the robot pushes a T-shaped object to a goal pose using visual feedback. The underlying policy is **VQ-BET** Lee et al. (2024), a latent autoregressive model producing chunked actions via vector quantization. Success is measured by the Intersection-over-Union (IoU) between the final block position and the target. To simulate environmental stochasticity, we adopt the action-space noise injection scheme from Liu et al. (2024), which perturbs actions $\mathbf{a}_t$ with zero-mean, temporally correlated Gaussian noise. We evaluate:

$$\eta = 0.0 \quad \text{(no noise)}, \quad \eta = 1.0 \quad \text{(moderate)}, \quad \eta = 1.5 \quad \text{(severe)}$$

4.2 Closed-Loop and Open-Loop Modes

We evaluate our method in both closed-loop and open-loop settings. In the closed-loop regime (AH $= 1$), the model is queried at every timestep, allowing for continuous resampling and rapid responsiveness to environmental changes. In contrast, the open-loop regime (AH $= 3$) executes a fixed sequence of three actions from each sampled chunk before the next resampling, introducing temporal commitment and limiting reactivity.

4.3 Hyperparameter Sweeps

We perform a grid search for each $\{\text{AH}, \eta\}$ combination:

Parameter	Symbol	Values
TVD sharpness	s	$\{1.0, 2.0, 3.0, 5.0\}$
Sliding window length	CQ	$\{1, 2, 3, 4\}$
Histogram resolution	$b \times b$	$\{5, 8, 12, 16\}$

Each run is repeated with three random seeds and results are averaged.

5 Results and Analysis

Given the results of our hyperparameter sweep, we report the two best configurations for the closed-loop and open-loop policy with the best average performance.

5.1 Policy Performance under Closed-Loop vs. Open-Loop Execution

Table 1: Task success rates (%) under varying action horizons and noise levels for BID and TVD-BID. Bold indicates best performance for each setting.

AH	**Policy**	$\sigma = 0.0$	$\sigma = 1.0$	$\sigma = 1.5$
Closed Loop (AH $= 1$)	BID	**66.67** ± 12.06	44.67 ± 13.61	34.67 ± 11.02
	TVD-BID	64.00 ± 19.00	**50.67** ± 15.23	**39.33** ± 6.11
	Sliding Window-BID	66.0 ± 8.03	46.12 ± 13.07	38 ± 5.62
Open Loop (AH $= 3$)	BID	62.00 ± 3.46	42.00 ± 5.29	32.67 ± 3.06
	TVD-BID	**72.00** ± 10.58	**52.67** ± 8.08	**36.00** ± 7.21

5.2 Qualitative Analysis

After our hyperparameter sweep, we found that $s = 1.0, b = 12$ yielded the best results for TV-BID across both closed loop and open loop. However, open-loop performance exceeded that of the closed-loop policy almost all noise levels. This was surprising, because we expected the sliding window method to give us a more accurate estimation of our noise, thus helping us better weigh the backward

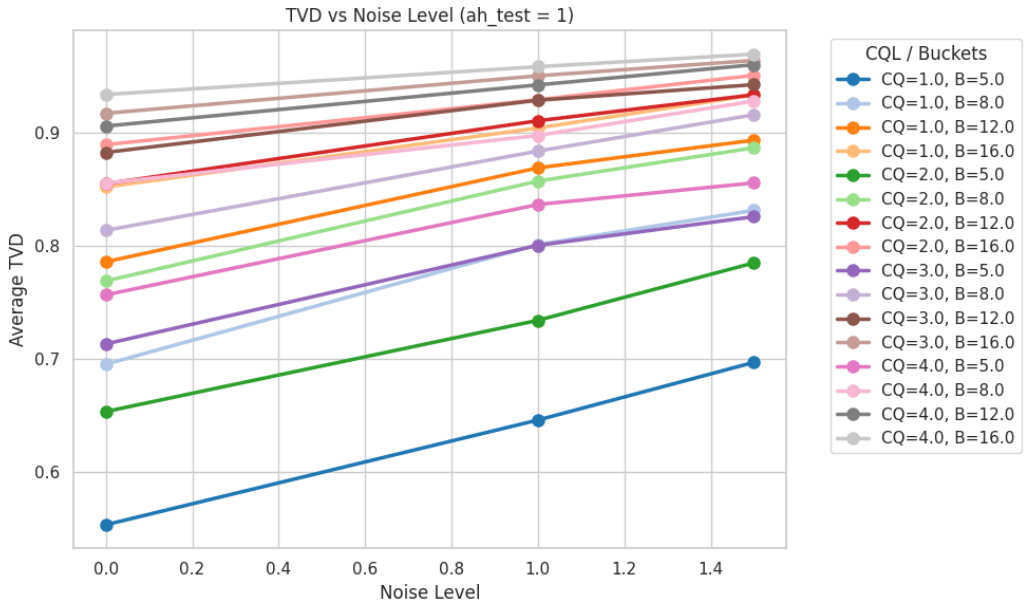

Figure 1: Average TVD vs noise for AH=1 under various grid resolutions and sliding-window lengths.

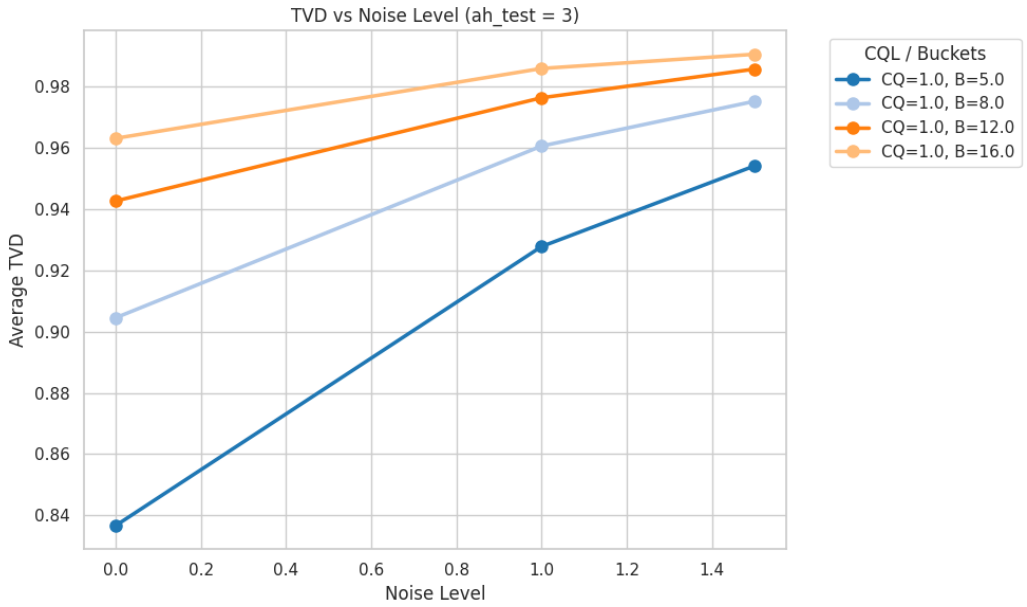

Figure 2: Average TVD vs noise for AH=3. Only histogram resolution is varied; CQ $= 1$ is fixed.

coherence and forward contrast. However, we expect that adding the extra windows might have actually decreased the accuracy of our stochasticity estimate since adding more previous timesteps dilutes the immediate stochasticity that the agent is observing. This can also be seen by the fact that with increasing chunk length in Figure 1, the delta between average noise level TVD decreases. But after plotting the average TVD across all three noise levels under various grid resolutions, we noticed that the average TVD remains more varied even at higher bucket levels for ah_test = 3. Please see Figures 1 and 2.

Additionally, the optimized closed-loop policy showed improved stability at high noise, with a smaller variance compared to the baseline. We also say improvements in the open loop policy across all noise levels compared to the baseline.

While TVD scores improve robustness under noise, they do not fully recover success rates from lower noise levels.

5.3 Quantitative Analysis

To examine these results further, we calculated the average TVD and average latent scores across the different noise levels. The average TVD score is the mean $\overline{TVD}$ score across episodes. The average latent difference is calculated by first averaging the latent embeddings across all N sampled action chunks to obtain a mean latent trajectory then computing the ℓ_2 distance between overlapping segments of consecutive timesteps, averaged across the prediction horizon. Given these insights, we propose the following conclusions:

Table 2: Average TVD, $\tanh$(TVD), and latent difference across noise levels.

ah_test	Noise	Avg.TVD	σ	Avg. Latent Diff.
1	0.0	0.6306	0.5585	0.996
	1.0	0.7022	0.6057	1.030
	1.5	0.7431	0.6310	8.760
3	0.0	0.9038	0.7177	1.365
	1.0	0.9510	0.7393	1.380
	1.5	0.9674	0.7473	1.620

TVD reflects difficulty and guides adaptive loss. TVD increases consistently with noise under both closed-loop (AH $= 1$) and open-loop (AH $= 3$) settings (Table 2), confirming its role as a proxy for control difficulty. TVD is notably higher in the open-loop regime even at $\sigma = 0.0$ (0.9038 vs. 0.6306), reflecting increased uncertainty due to delayed feedback and longer action commitment. This higher baseline suggests that adaptive loss modulation is more critical in open-loop policies.

Latent shift aligns with performance robustness. Latent differences grow with noise in both regimes, but the trend is more dramatic in the closed-loop case, where the average latent shift jumps from 1.030 to 8.760 between $\sigma = 1.0$ and $\sigma = 1.5$, indicating a breakdown in representation stability. This corresponds with the baseline BID performance drop (Table **??**), where success rate falls from 44.67% to 34.67%. In contrast, the open-loop regime shows more gradual increases in latent shift, and TVD-BID maintains relatively high performance, suggesting that adaptive loss mixing helps manage representational drift more gracefully in delayed feedback scenarios.

Tradeoffs between robustness and clean performance. While TVD-BID improves robustness under noise, it slightly underperforms baseline BID at $\sigma = 0.0$ in the closed-loop setting (64.00% vs. 66.67%) and only marginally outperforms it at higher noise. This suggests that adaptivity can dampen sharp decision-making in low-noise settings where deterministic planning is optimal. In contrast, in the open-loop regime, TVD-BID outperforms baseline BID across all noise levels—including clean settings—highlighting its greater benefit when feedback is infrequent. These results underscore the need for noise-aware tuning of the loss modulation parameter to balance adaptivity with decisiveness.

6 Discussion

Limitations. Our experimental design was heavily constrained by compute availability. We initially ran experiments on an AWS GPU instance, but frequent shutdowns interrupted progress. As a result, we migrated to Google Cloud Compute and incurred out-of-pocket costs, which limited our ability to exhaustively run all configurations across multiple seeds. While we selected the best-performing configurations for closed-loop and open-loop settings, a more comprehensive sweep could yield additional insights, particularly around hyperparameter sensitivity and robustness variance.

Broader Impacts. Our approach proposes a lightweight, distribution-aware mechanism for improving robustness in learned control policies without additional training or supervision. This has positive

implications for real-world deployment of robot policies in uncertain or dynamic environments, where retraining is impractical. However, because our method dynamically adjusts behavior at test time, care must be taken to ensure that such adaptation does not amplify instability in safety-critical contexts. Future extensions should include safety guarantees or bounded behavior constraints to ensure reliable deployment in human-centric environments.

7 Conclusion.

Our results demonstrate that adapting chunk selection using test-time distributional shift signals can significantly improve the robustness of behavior cloning policies under noisy and delayed feedback conditions. By using Total Variation Distance (TVD) as a proxy for stochasticity, TVD-BID dynamically balances forward contrast and backward coherence, outperforming baseline BID in high-noise regimes. While closed-loop execution provides stability through frequent feedback, open-loop policies benefit more from adaptive mechanisms due to their inherent exposure to drift. However, in low-noise settings, adaptivity may trade off with sharpness in decision-making, suggesting that future work should explore more nuanced modulation strategies. Overall, our findings support the promise of lightweight, distribution-aware adaptation at test time for chunked policy execution.

8 Team Contributions

- **Sarosh Khan:** Implemented action distribution TVD, ran hyperparameter sweeps
- **Ellie Tanimura:** Ran baselines, implemented latent rescoring

References

Christopher G. Atkeson and Stefan Schaal. 1997. Robot Learning From Demonstration. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 12–20.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. arXiv:2303.04137 [cs.RO] https://arxiv.org/abs/2303.04137

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG] https://arxiv.org/abs/2105.05233

Michael Laskey, Jonathan Lee, Roy Fox, Anca Dragan, and Ken Goldberg. 2017. DART: Noise Injection for Robust Imitation Learning. arXiv:1703.09327 [cs.LG] https://arxiv.org/abs/1703.09327

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. 2024. Behavior Generation with Latent Actions. arXiv:2403.03181 [cs.LG] https://arxiv.org/abs/2403.03181

Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Maximilian Du, and Chelsea Finn. 2024. Bidirectional Decoding: Improving Action Chunking via Closed-Loop Resampling. arXiv:2408.17355 [cs.RO] https://arxiv.org/abs/2408.17355

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics* 11 (01 2023), 102–121. https://doi.org/10.1162/tacl_a_00536 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00536/2067865/tacl_a_00536.pdf

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. 2021. Discovering and Achieving Goals via World Models. arXiv:2110.09514 [cs.LG] https://arxiv.org/abs/2110.09514

Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. 2025. Steering Your Generalists: Improving Robotic Foundation Models via Value Guidance. arXiv:2410.13816 [cs.RO] https://arxiv.org/abs/2410.13816

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. arXiv:1011.0686 [cs.LG] https://arxiv.org/abs/1011.0686

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. arXiv:2304.13705 [cs.RO] https://arxiv.org/abs/2304.13705

Generative AI to refine prose in accordance with the Honor Code.