VGAS: Value-Guided Action-Chunk Selection for Few-Shot Vision-Language-Action Adaptation

Changhua Xu , En Yu , Junyu Xuan , Jie Lu
Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia
changhua.xu@student.uts.edu.au; {en.yu-1, junyu.xuan, jie.lu}@uts.edu.au

Abstract

Vision–Language–Action (VLA) models bridge multimodal reasoning with physical control, but adapting them to new tasks with scarce demonstrations remains unreliable. While fine-tuned VLA policies often produce semantically plausible trajectories, failures often arise from unresolved geometric ambiguities, where near-miss actions lead to divergent execution outcomes under limited supervision. We study few-shot VLA adaptation from a *generation–selection* perspective and propose a novel framework **VGAS** (**V**alue-**G**uided **A**ction-chunk **S**election). It performs inference-time best-of-N selection to identify action chunks that are both semantically faithful and geometrically precise. Specifically, **VGAS** employs a fine-tuned VLA as a high-recall proposal generator and introduces the Q-Chunk-Former, a geometrically grounded Transformer critic to resolve fine-grained geometric ambiguities. In addition, we propose *Explicit Geometric Regularization* (EGR), which shapes a discriminative value landscape to preserve action ranking resolution among near-miss candidates while mitigating value instability under scarce supervision. Experiments and theoretical analysis demonstrate that **VGAS** consistently improves success rates and robustness under limited demonstrations and distribution shifts. Our code is available at https://github.com/Jyugo-15/VGAS.

1 Introduction

Vision-Language-Action (VLA) models have emerged as a transformative paradigm for embodied AI, bridging multimodal reasoning with physical control [Brohan *et al.*, 2022; Zitkovich *et al.*, 2023; He *et al.*, 2026]. By pretraining on vast robotic datasets, these generalist policies learn to map complex visual observations and linguistic instructions directly into executable actions [Black *et al.*, 2024; Kim *et al.*, 2024; Team *et al.*, 2024; Intelligence *et al.*, 2025]. However, the reliability of VLAs in downstream applications remains heavily bottlenecked by the prevailing Supervised Fine-Tuning (SFT) paradigm [Kim *et al.*, 2025; Zhang *et al.*, 2025]. SFT-based adaptation demands a high volume of expert demonstrations to bridge the gap between generalist priors and task-specific requirements. This is often challenging in real-world environments where high-quality robotic data collection is costly, unscalable, and prone to out-of-distribution (OOD) uncertainties [Dass *et al.*, 2022; Xin *et al.*, 2024; Sapkota *et al.*, 2025; Yu *et al.*, 2026]. Consequently, under data-scarce regimes, VLA policies often exhibit brittle performance, failing to generalize across even minor distribution shifts [Liu *et al.*, 2025; Li *et al.*, 2025a; Yu *et al.*, 2025; Tan *et al.*, ; Guo *et al.*, 2025].

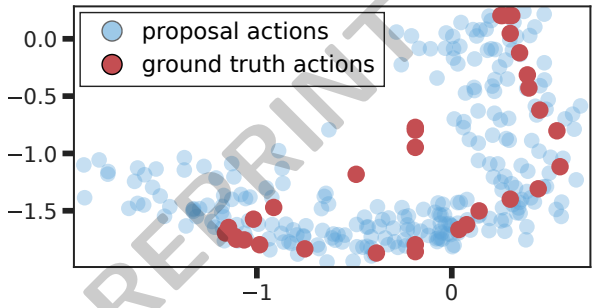


Figure 1: Illustration of near-miss action distribution under 5-shot VLA fine-tuning.

To study this data-scarce regime, we simulate a realistic few-shot adaptation setting by fine-tuning a pretrained VLA policy with only five demonstrations per task, and evaluating it on held-out demonstrations with different initial spatial configurations (e.g., object positions). As shown in Figure 1, the fine-tuned policy generally preserves the task semantics, producing actions toward the correct object. However, due to the limited state–action coverage, even slight variations in the initial spatial configuration can make the policy's action predictions less concentrated, yielding dispersed near-miss predictions around the ground-truth actions. Although semantically reasonable, these predictions are often geometrically imprecise and can cause failures such as inaccurate grasps, end-effector pose deviations, or joint-angle overshoot [Kumar *et al.*, 2022; Zhao *et al.*, 2023]. This suggests that few-shot VLA adaptation is primarily limited by geometric precision under sparse supervision, rather than semantic understanding alone. Motivated by this, we reformulate adaptation as a value-guided selection problem: instead of requiring a generative policy to jointly acquire semantic reasoning and fine-

grained geometric control end-to-end, we decouple adaptation into high-recall proposal generation and high-precision value-based selection, prioritizing candidates with the highest likelihood of long-horizon success.

Offline Reinforcement Learning (ORL) provides a natural framework for this selection objective, as it learns an outcome-aware critic that maps long-horizon success into a scalar value signal [Sutton *et al.*, 1998]—perfectly suited for ranking candidate proposals [Ghasemipour *et al.*, 2021; Janner *et al.*, 2022; Luo *et al.*,]. However, applying existing ORL methodologies to modern VLA policies exposes two fundamental limitations: *1) Structural and observational mismatch:* Standard RL assumes per-step atomic actions, whereas modern VLAs output temporally extended action chunks, inducing an SMDP structure that complicates value learning and temporal credit assignment [Sutton *et al.*, 1999]. Moreover, most offline RL is evaluated with compact, near-Markovian state inputs; in contrast, VLA relies on high-dimensional vision–language observations with geometric grounding, making value estimation substantially harder and comparatively under-explored [Fu *et al.*, 2020; Lu *et al.*, 2022]. *2) Ranking resolution vs. conservatism:* Offline RL often controls extrapolation to low-support actions via conservative objectives or behavior-regularized extraction [Kumar *et al.*, 2020; Kostrikov *et al.*, 2021]. In few-shot sparse-reward settings, however, such regularization can compress value gaps among proposal-supported near-miss candidates, yielding low-contrast gradients and weakening inference-time Best-of-N selection [Lyu *et al.*, 2022]. These limitations raise two major research questions: ***RQ1.*** *What critic architecture can robustly ground high-dimensional VLA observations into precise value estimates for temporally extended action chunks?* and ***RQ2.*** *How can a value function be trained under scarce demonstrations to maintain high ranking resolution among near-miss action chunks?*

To address these questions, we propose **VGAS** (**V**alue-**G**uided **A**ction-chunk **S**election) for VLA adaptation via generation–selection decoupling. For ***RQ1***, we introduce Q-Chunk-Former, a geometrically grounded critic architecture built on a Transformer backbone. By leveraging the Transformer's sequence modeling capability, our design naturally captures temporal dependencies within action chunks. Crucially, our architecture preserves fine-grained, token-level features, allowing attention to explicitly focus on geometric cues that are critical for precise value estimation. For ***RQ2***, we propose a hybrid offline RL objective that *anchors* temporal consistency via a proposal-constrained Bellman backup, augmented by Explicit Geometric Regularization (EGR). Unlike traditional conservative methods that indiscriminately penalize out-of-distribution actions, EGR injects dense geometric supervision, shaping the value landscape into a smooth funnel anchored at expert demonstrations. This allows the critic to maintain high ranking resolution among near-miss candidates even under scarce supervision. Our contributions are summarized as follows:

- We reformulate few-shot VLA adaptation as a value-guided selection problem and propose the novel **VGAS** method, shifting the paradigm from likelihood-based generation to outcome-aware ranking.
- We propose Q-Chunk-Former, which enables precise geometric grounding for action-chunk evaluation. We introduce EGR, a regularization technique that injects dense geometric priors into offline RL to maintain high ranking resolution under data-scarce regimes.
- We provide theoretical guarantees for the convergence of our chunk-level value operator and demonstrate through extensive experiments on the LIBERO benchmark that VGAS consistently outperforms SFT and standard ORL baselines, particularly in terms of success rate and robustness under distribution shifts.

2 Preliminary

Offline Reinforcement Learning. We consider a Markov Decision Process (MDP) defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma)$, where $s \in \mathcal{S}$, $a \in \mathcal{A}$, $\mathcal{P}(s' \mid s, a)$ is the transition kernel, $r(s,a) \in \mathbb{R}$ is the reward function, $\rho_0 \in \Delta(\mathcal{S})$ is the initial-state distribution, and $\gamma \in [0,1)$ is the discount factor. The objective is to learn a policy π maximizing the expected discounted return $J(\pi) = \mathbb{E}_\pi\Big[\sum_{t\geq 0} \gamma^t r(s_t, a_t)\Big]$.

Value-based methods estimate the optimal action-value function $Q^\star$ as the fixed point of the Bellman optimality operator [Fujimoto *et al.*, 2019; Levine *et al.*, 2020]:

$$(\mathcal{T}^\star Q)(s,a) = r(s,a) + \gamma\, \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right]. \quad (1)$$

In offline RL, the agent learns from a fixed dataset $\mathcal{D}$ collected by a behavior policy π_β, without additional environment interaction [Bacchiocchi *et al.*, 2024]. A key challenge is extrapolation error [Zhang and Tan, 2023]: the maximization over a' may select actions outside the data support, leading to overestimation and instability. Many offline RL methods [Shin and Kim, 2023] mitigate this via conservative regularization [Kumar *et al.*, 2020], which discourages high values on out-of-distribution actions.

Action Chunking in VLAs. Modern VLAs condition on multimodal inputs, which we denote by the state $s_t = (I_t, L_t, p_t)$, comprising visual tokens I_t, language instruction L_t, and robot proprioception p_t [Zitkovich *et al.*, 2023]. With a slight abuse of notation, we treat this policy input as the MDP state. Instead of per-step control, these models often output a temporally extended *action chunk* with horizon h:

$$\mathrm{A}_t := (a_t, \ldots, a_{t+h-1}) \in \mathcal{A}^h, \qquad \mathrm{A}_t \sim \pi_\mu(\cdot \mid s_t). \quad (2)$$

Few-shot VLA Adaptation Objective. Given a few-shot expert dataset $\mathcal{D} = \{\tau_i\}_{i=1}^K$, we first obtain a task-aligned base chunk policy $\pi_\mu(\mathrm{A} \mid s)$, for example via SFT on $\mathcal{D}$. In this regime, SFT can capture task semantics yet may struggle to resolve fine-grained geometric ambiguities. We therefore treat π_μ as a high-recall proposal distribution and assume non-trivial local support around each demonstrated chunk:

$$\pi_\mu(\{\mathrm{A} : \|\mathrm{A} - \mathrm{A}_t\|_2 \leq \varepsilon\} \mid s_t) \geq p_0, \forall (s_t, \mathrm{A}_t) \in \mathcal{D}, \quad (3)$$

for some $\varepsilon > 0$ and $p_0 > 0$. Intuitively, p_0 ensures the proposal covers valid near-miss candidates. Our goal is to improve upon π_μ by learning an offline-adapted chunk policy within a proposal-constrained class Π_μ:

$$\pi^\star := \arg\max_{\pi\in\Pi_\mu} J(\pi), \quad (4)$$

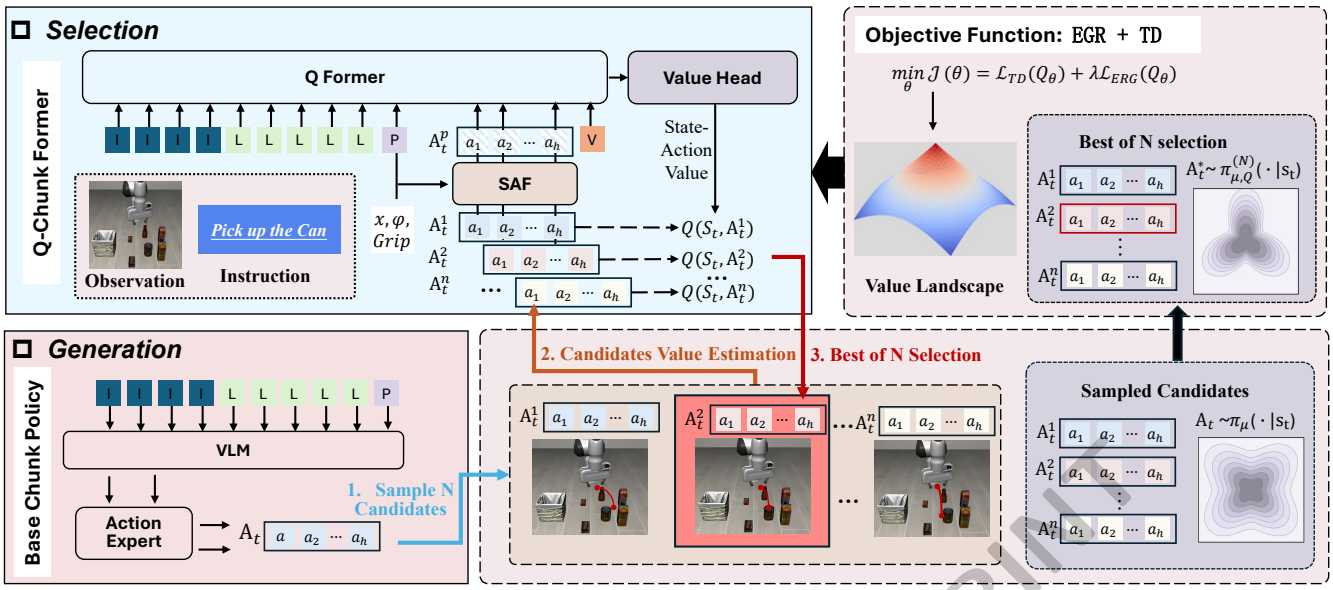


Figure 2: The overall framework of **VGAS**. **Generation:** A fine-tuned VLA policy proposes N candidate action chunks from multimodal inputs. **Selection:** Q-Chunk-Former learns a scoring function Q via the EGR+TD objective. Best-of-N selection defines the induced policy $\pi^{(N)}_{\mu,Q}$ by maximizing over a discriminative value landscape shaped by EGR, prioritizing expert-aligned candidates and thereby mitigating geometric drift.

where Π_μ denotes policies supported by (or centered at) the proposal π_μ. Under chunked execution, the return is

$$J(\pi) := \mathbb{E}_{\tau\sim\pi}\left[\sum_{k=0}^{\infty}(\gamma^h)^k\, R_h(s_{t_k}, \mathrm{A}_{t_k})\right], \qquad (5)$$

where $t_k = kh$ denotes the start time of the k-th chunk, $\mathrm{A}_{t_k} \sim \pi(\cdot \mid s_{t_k})$ is the action chunk, and $R_h(s_{t_k}, \mathrm{A}_{t_k}) = \sum_{j=0}^{h-1}\gamma^j\, r(s_{t_k+j}, a_{t_k+j})$ is the discounted cumulative reward over the chunk.

3 Methodology

Framework Overview. **VGAS** reformulates few-shot VLA adaptation as a generate-then-select process. As illustrated in Figure 2, our pipeline decouples the policy into two components: a *high-recall generator* and a *high-precision critic*. First, we utilize a supervised fine-tuned (SFT) VLA model as the base policy $\pi_\mu(\mathrm{A}_t \mid s_t)$ to provide a proposal distribution covering plausible action chunks. At inference time, we sample N candidates $\{\mathrm{A}_t^{(i)}\}_{i=1}^N \sim \pi_\mu(\cdot \mid s_t)$ and employ a learned critic Q_θ to execute Best-of-N selection, $\mathrm{A}_t^\star = \arg\max_i Q_\theta(s_t, \mathrm{A}_t^{(i)})$. This strategy approximates policy improvement within the support of π_μ, prioritizing geometric precision without requiring online exploration.

To realize this selection mechanism effectively, VGAS addresses two core challenges: *representation* and *optimization*. First, regarding critic representation (Sec.3.1), we introduce Q-Chunk-Former, a Transformer-based architecture tailored for VLA inputs. To prevent high-dimensional visual tokens from overwhelming physical cues, we design a State-Action Fusion (SAF) module that explicitly grounds action chunks in proprioceptive states before multimodal integration. Second, for critic optimization (Sec.3.2), we propose a hybrid learning objective that combines temporal and spatial supervision. We stabilize offline training using a *Proposal-Constrained Chunked Expected-Max* backup, which enforces temporal consistency across action chunks. We further augment this with *Explicit Geometric Regularization* (EGR), a dense supervision signal that directly shapes the value landscape based on geometric proximity to expert demonstrations, enabling the critic to reliably distinguish near-miss actions from failures even under sparse task rewards.

3.1 Q-Chunk-Former

To optimize the few-shot adaptation objective in Eq. (4), the critic network must accurately estimate the long-horizon value of a temporally extended action chunk A_t given state s_t. This imposes two key requirements, (i) Chunk-level evaluation with temporal structure: the critic must assign a single long-horizon value to an entire action chunk A_t for Best-of-N selection (Eq. (9)), while preserving the within-chunk temporal ordering of actions; (ii) Multimodal fusion without geometric collapse: unlike classical critics operating on compact state vectors, VLA conditioning involves heterogeneous inputs such as vision, language, and proprioception, where naive compression can discard geometry-critical cues needed for feasibility-aware value estimation.

Motivated by token-level multimodal modeling [Marafioti *et al.*, 2025], we introduce a Transformer-based critic. A naive design treats all modalities as a single concatenated token sequence:

$$Q_\theta(s_t, \mathrm{A}_t) = \mathrm{Transformer}_\theta([I_t,\, L_t,\, p_t,\, \mathrm{A}_t]), \qquad (6)$$

where $[\cdot]$ denotes concatenation. However, in practice, the

self-attention mechanism tends to be dominated by the abundant visual and linguistic tokens, so the single proprioceptive token p_t receives insufficient attention. This is detrimental because value estimation for manipulation requires joint reasoning over the external world context (from I_t and L_t) and the internal robot state (embodiment and configuration encoded by p_t). If p_t is under-utilized, the critic becomes less sensitive to geometric feasibility, weakening feasibility-aware value estimation.

To address this, we introduce a lightweight State-Action Fusion (SAF) module that conditions the raw chunk A_t on p_t prior to mixing with high-dimensional perceptions. The SAF module produces proprioception-grounded action tokens A_t^p,

$$\mathrm{A}_t^p = \mathrm{SAF}(\mathrm{A}_t, p_t) = \mathrm{W}_{\text{fuse}}\left([\mathrm{W}_a \mathrm{A}_t \parallel \mathrm{W}_p p_t]\right), \quad (7)$$

where W_a and W_p are learnable projections mapping inputs to a shared latent space and W_{fuse} aggregates the concatenated features. This design enforces a high-fidelity interaction between action tokens and the proprioceptive state, ensuring that feasibility cues are embedded directly into the chunk representation. The grounded action tokens A_t^p are then integrated with the frozen perceptual tokens[1] and a learnable [VALUE] token v via a Transformer decoder, denoted as the Q-Former (QF). Finally, a Value Head (VH) maps the output embedding of v to a scalar Q-value:

$$Q_\theta(s_t, \mathrm{A}_t) = \mathrm{VH}(\mathrm{QF}(I_t,\, L_t,\, p_t,\, \mathrm{A}_t^p,\, \mathrm{v})). \quad (8)$$

In summary, Q-Chunk-Former comprises SAF, QF, and VH modules. This architecture mitigates attention imbalance and ensures that value estimation is strictly grounded in geometric reality, providing a reliable signal for selection.

3.2 Optimization Objective

Our goal is to learn a critic that supports *value-guided selection* over a fixed proposal distribution. We assume access to a task-adapted base chunk policy $\pi_\mu(\mathrm{A} \mid s)$ that generates semantically plausible action chunks. At inference time, we draw N i.i.d. candidates from π_μ and select the best one according to a scoring rule Q:

$$\mathrm{A}_Q^\star(s) \;:=\; \arg\max_{i\in[N]} Q\Big(s, \mathrm{A}^{(i)}\Big). \quad (9)$$

Although the maximization in Eq. (9) is deterministic conditioned on the sampled set, proposal sampling induces a stochastic *selection policy*. We denote this induced policy by $\pi_{\mu,Q}^{(N)}(\mathrm{A} \mid s)$, defined as the distribution of $\mathrm{A}_Q^\star(s)$.

To learn Q from a fixed offline dataset $\mathcal{D}$, we train the critic as a proxy for the long-horizon return of the induced policy $\pi_{\mu,Q}^{(N)}$ in the chunk-induced SMDP. A reliable critic in the few-shot regime must satisfy two coupled desiderata: (i) **temporal consistency**, i.e., Bellman-style alignment with the expected *max* return under best-of-N selection over π_μ samples; and (ii) **spatial consistency**, i.e., preserving fine-grained geometric ranking among near-miss proposals. We instantiate these principles with the following hybrid objective:

$$\min_\theta \; \mathcal{J}(\theta) \;=\; \mathcal{L}_{\texttt{TD}}(Q_\theta) \;+\; \lambda\, \mathcal{L}_{\texttt{EGR}}(Q_\theta). \quad (10)$$

[1]Perceptual tokens I_t and L_t are extracted from the pre-trained VLM encoder of the base policy to ensure feature alignment and computational efficiency.

Temporal Consistency. The primary goal of $\mathcal{L}_{\texttt{TD}}$ is to align critic learning with our inference-time execution, i.e., to learn the value function induced by Best-of-N selection, $\pi_{\mu,Q}^{(N)}$. This alignment requires two ingredients. First, since execution commits to a length-h action chunk, the critic must evaluate the return of an entire chunk. We therefore adopt the chunk-level TD formulation from Q-Chunking [Li *et al.*, 2025b] and learn $Q(s_t, \mathrm{A}_t)$ that conditions on the full action chunk.

Second, we must ensure that the *Bellman backup* matches the same Best-of-N rule used at inference. To this end, inspired by Expected-Max Q-learning (EMaQ) [Ghasemipour *et al.*, 2021], the key insight of EMaQ is to construct a Bellman backup that replaces the standard expectation under a policy with an expected maximization over a set of sampled proposals. By targeting this maximum, the objective ensures that the critic trained with this operator is guaranteed to converge to the value function of the corresponding Best-of-N selection policy $\pi_{\mu,Q}^{(N)}$. Thus, we define the proposal-constrained *Chunked Expected–Max* backup operator $\mathcal{T}_\mu^N$ as

$$\begin{aligned}&(\mathcal{T}_\mu^N Q)(s, \mathrm{A}) := R_h(s, \mathrm{A}) + \\ &\gamma^h \mathbb{E}_{s'\sim\mathcal{P}_h(\cdot|s,\mathrm{A})}\Big[\mathbb{E}_{\mathrm{A}'_{1:N}\overset{\text{i.i.d.}}{\sim}\pi_\mu(\cdot|s')}\big[\max_{i\in[N]} Q(s', \mathrm{A}'_i)\big]\Big],\end{aligned} \quad (11)$$

where R_h is the discounted cumulative reward over the chunk and $\mathcal{P}_h$ is the h-step transition kernel. The inner maximization corresponds exactly to selecting $\mathrm{A}_Q^\star(s')$ from Eq. (9), thereby enforcing strict train–test consistency.

This operator formulation offers rigorous theoretical guarantees for offline adaptation, which we summarize below.

Proposition 1 (Chunked Expected–Max in tabular SMDPs)**.** *In the tabular chunk-induced SMDP, assume bounded rewards and $\gamma^h \in (0,1)$. Then $\mathcal{T}_\mu^N$ in Eq. (11) is a γ^h-contraction under $\|\cdot\|_\infty$ and has a unique fixed point Q_μ^N. Let $\pi_\mu^{(N)} := \pi_{\mu,Q_\mu^N}^{(N)}$ be the induced Best-of-N policy. Then $Q_\mu^N = Q^{\pi_\mu^{(N)}}$. Monotonicity in N and the $N\to\infty$ limit are given in Prop. 3 and Thm. 2 (App. B).*

Proposition 1 shows that the proposed backup is well-defined in the tabular induced SMDP and that its unique fixed point corresponds exactly to the value function of the induced Best-of-N selection policy. A detailed proof of these properties in the tabular SMDP setting is provided in Appendix B.

Based on this operator, we instantiate the final temporal-difference loss with a standard target network $Q_{\bar\theta}$ for stability:

$$\begin{aligned}&\mathcal{L}_{\texttt{TD}}(\theta) = \mathbb{E}_{(s_t,\mathrm{A}_t,s_{t+h})\sim\mathcal{D}}\Big[\big(Q_\theta(s_t,\mathrm{A}_t) - y_t\big)^2\Big],\\ &\text{where } y_t = R_h(s_t,\mathrm{A}_t) + \gamma^h \max_{i\in[N]} Q_{\bar\theta}(s_{t+h}, \mathrm{A}'^i_{t+h}).\end{aligned} \quad (12)$$

Here, the proposals $\{\mathrm{A}'^i_{t+h}\}_{i=1}^N$ are sampled from the frozen proposal distribution $\pi_\mu(\cdot \mid s_{t+h})$ at the next decision state. This loss provides a stable, proposal-constrained temporal anchor for our critic, paving the way for the spatial regularization described next.

Spatial Consistency. While the proposal-constrained TD loss provides a stable temporal anchor, offline critic learning remains vulnerable to extrapolation error, where OOD (off-demo) candidates deviating from the training data distribution receive spuriously high values. Standard conservative methods, such as CQL [Kumar *et al.*, 2020], mitigate this by indiscriminately suppressing values for all low-support actions. However, in the few-shot regime, this uniform penalty is overly aggressive, because it compresses the value dynamic range among proposal-supported candidates and reduces the fine-grained ranking resolution needed to distinguish plausible near-miss candidates from catastrophic failures. This value collapse directly undermines inference-time Best-of-N selection.

To address this, we introduce *Explicit Geometric Regularization (EGR)*. Instead of uniformly suppressing off-demo actions, EGR serves as a structural regularizer. During training, we regularize the critic with off-demo candidates sampled from $\rho(\cdot \mid s_t)$, a proposal-centered mixture detailed in Appendix D.1. At inference time, Best-of-N selects only among proposal samples from $\pi_\mu(\cdot \mid s_t)$. Accordingly, EGR targets two desiderata: **(i)** *TD-Anchored Calibration:* anchor the value scale to the TD target y_t to *mitigate* overestimation on off-demo candidates; **(ii)** *Geometric Discriminability:* within the inference-time proposal set, preserve a graded preference for geometric proximity, providing a smooth signal to separate recoverable near-misses from divergence.

We formalize EGR as the following weighted combination of *Anchoring* and *Ranking* losses.

$$\mathcal{L}_{\texttt{EGR}}(\theta) = \mathcal{L}_{\texttt{anchor}}(\theta) + \eta\, \mathcal{L}_{\texttt{rank}}(\theta), \tag{13}$$

where $\eta \geq 0$ balances the *absolute scale* (Anchoring) and the *local ordering* (Ranking).

a. Geometric anchoring ($\mathcal{L}_{\text{anchor}}$). For any off-demo action chunk $\hat{\mathrm{A}}_t \sim \rho(\cdot \mid s_t)$, we define

$$\mathcal{Y}(s_t, \hat{\mathrm{A}}_t) \coloneqq \mathrm{sg}(y_t) - \beta||\hat{\mathrm{A}}_t - \mathrm{A}_t||^2_{\mathcal{W}}, \tag{14}$$

where $\mathrm{sg}(\cdot)$ stops gradients and y_t is the TD target from Eq. (12). Crucially, we do not posit Euclidean distance as a global ground-truth metric; rather, we employ this surface as a *structural inductive bias* to shape the value landscape in regions where task supervision is absent. In our setting, we employ a weighted metric $\|\hat{\mathrm{A}}_t - \mathrm{A}_t\|^2_{\mathcal{W}}$ to prioritize critical kinematic dimensions (e.g., end-effector position) within the normalized control space. This serves as a robust local proxy for geometric proximity: under smooth dynamics, small weighted action deviations tend to induce small short-horizon trajectory deviations. (See Appendix D.3 for details).

To satisfy TD-anchored calibration, we align the critic's estimates on candidate chunks with the reference surface:

$$\mathcal{L}_{\text{anchor}}(\theta; s_t, \mathrm{A}_t) = \mathbb{E}_{\hat{\mathrm{A}}_t \sim \rho(\cdot|s_t)}\Big[\ell\big(Q_\theta(s_t, \hat{\mathrm{A}}_t),\, \mathcal{Y}(s_t, \hat{\mathrm{A}}_t)\big)\Big], \tag{15}$$

where $\ell(\cdot, \cdot)$ is the squared error.

b. Geometric ranking ($\mathcal{L}_{\text{rank}}$). While anchoring constrains the absolute scale, it does not guarantee robust *local* discrimination. To enforce geometric discriminability, we introduce a pairwise ranking loss. For any pair $(\hat{\mathrm{A}}_t^i, \hat{\mathrm{A}}_t^j) \sim \rho(\cdot \mid s_t)$, define squared distances to the expert:

$$d_i \coloneqq \left\|\hat{\mathrm{A}}_t^i - \mathrm{A}_t\right\|^2_{\mathcal{W}}, \qquad d_j \coloneqq \left\|\hat{\mathrm{A}}_t^j - \mathrm{A}_t\right\|^2_{\mathcal{W}}. \tag{16}$$

By Eq. (14), the reference differential value satisfies $\mathcal{Y}(s_t, \hat{\mathrm{A}}_t^i) - \mathcal{Y}(s_t, \hat{\mathrm{A}}_t^j) = \beta(d_j - d_i)$. We thus encourage the critic to preserve the same relative differences among candidate chunks:

$$\begin{aligned}\mathcal{L}_{\text{rank}}(\theta; s_t, \mathrm{A}_t) = \mathbb{E}_{\hat{\mathrm{A}}_t^i,\, \hat{\mathrm{A}}_t^j \sim \rho(\cdot|s_t)}\Big[\ell\Big(&Q_\theta(s_t, \hat{\mathrm{A}}_t^i) \\ &- Q_\theta(s_t, \hat{\mathrm{A}}_t^j), \beta(d_j - d_i)\Big)\Big].\end{aligned} \tag{17}$$

The Closed Loop of Spatio-Temporal Consistency. The TD and EGR form a mutually reinforcing loop for safe and effective value learning. The Expected-Max TD objective provides a foundational safety layer by inherently operating within the support of the proposal distribution π_μ. However, this implicit constraint alone can still be vulnerable to selecting outlier candidates that are accidentally overestimated. EGR reinforces this safety by explicitly shaping the OOD value landscape into a geometric funnel. This structure actively biases the Best-of-N maximization toward candidates that remain close to expert behavior, providing a tighter and more reliable bound on the TD target. This enhanced safety guarantee is formalized by the following proposition:

Proposition 2 (Best-of-N bound under an EGR anchoring envelope)**.** *Fix a demonstration pair $(s_t, \mathrm{A}_t) \sim \mathcal{D}$ and a candidate distribution $\pi_\mu(\cdot \mid s_t)$. Define the EGR reference surface $\mathcal{Y}(s_t, \mathrm{A}'_t) = \mathrm{sg}(y_t) - \beta\|\mathrm{A}'_t - \mathrm{A}_t\|^2_{\mathcal{W}}$ (Eq. (14)) and the* anchoring residual $\delta_\theta(s_t, \mathrm{A}'_t) \coloneqq Q_\theta(s_t, \mathrm{A}'_t) - \mathcal{Y}(s_t, \mathrm{A}'_t)$. *Assume it is uniformly upper-bounded on the candidate set:*

$$\sup_{\mathrm{A}'_t \in \mathrm{supp}(\pi_\mu(\cdot|s_t))} \delta_\theta(s_t, \mathrm{A}'_t) \leq \varepsilon.$$

Draw N candidates $\{\mathrm{A}'^i_t\}_{i=1}^N \sim \rho(\cdot \mid s_t)$. Then

$$\max_{i \in [N]} Q_\theta(s_t, \mathrm{A}'^i_t) \leq \mathrm{sg}(y_t) + \varepsilon. \tag{18}$$

See Appendix B.2 for the full statement and proof.

TD Anchors and Calibrates EGR. Conversely, the TD objective provides an essential grounding signal for EGR. The TD target y_t sets the absolute scale for the EGR reference surface, preventing the geometric shaping from degenerating into an uncalibrated ranking function. This synergy allows the TD-EGR loop to progressively correct value estimates: TD provides the return-aware scale, while EGR provides the fine-grained geometric structure, jointly enabling robust ranking.

4 Experiment

In our experiments, we aim to answer the following research questions: *RQ1: Does **VGAS** benefit from a Transformer-based chunk critic (Q-Chunk-Former) for modeling fine-grained geometric dependencies under multimodal inputs? RQ2: Does Explicit Geometric Regularization (EGR) improve value calibration and ranking over state–action-chunk candidates for Best-of-N selection?*

Type	Method	LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long		Average	
		SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)
BC-Only	SmolVLA	46.0	4	45.8	4	52.0	3	15.5	4	39.8	4
BC + RL	QC-M	46.0	4	42.1	5	49.2	5	14.4	5	37.9	5
	QC-M+CQL	47.8	2	50.3	2	54.2	2	16.5	3	42.2	2
	QC-T+CQL	47.7	3	50.1	3	51.9	4	17.5	2	41.8	3
	VGAS (Ours)	**56.2**	**1**	**59.0**	**1**	**60.8**	**1**	**20.0**	**1**	**49.0**	**1**

Table 1: Results on LIBERO (5-shot per task). SR denotes success rate (%).

4.1 Experiment Settings

Benchmark and Architecture. We evaluate **VGAS** on the widely used simulation benchmark, LIBERO [Liu *et al.*, 2023]. LIBERO is a lifelong learning benchmark focused on language-guided manipulation tasks across diverse object types, task specifications, and environments scenarios. Specifically, it includes 4 suites: Goal, Spatial, Object, and Long. Each suite is designed to evaluate a specific aspect of object manipulation and containing 10 distinct tasks. We use SmolVLA-0.5B [Shukor *et al.*, 2025] as the base chunk policy. We initialize Q-Chunk-Former with the first two decoder layers of the pre-trained SmolVLM2 [Marafioti *et al.*, 2025], keeping the critic's token space aligned with the policy. Comprehensive architectural details are provided in Appendix D.

Baselines. We compare **VGAS** against baselines covering VLA fine-tuning and offline value-based improvement. **BC-only** fine-tunes SmolVLA with behavior cloning. **CQL** [Kumar *et al.*, 2020] serves as a representative conservative offline RL objective and is widely used in VLA settings [Song *et al.*, 2025; Chebotar *et al.*, 2023; Huang *et al.*, 2025; Nakamoto *et al.*, 2025; Chen *et al.*, 2025]. **QC-M** follows Q-Chunking [Li *et al.*, 2025b], training an MLP critic with standard TD learning over action chunks. **QC-M+CQL** augments **QC-M** with the CQL regularizer. Finally, **QC-T+CQL** retains the same objective but replaces the MLP with our Q-Chunk-Former, enabling a controlled comparison of critic architectures under identical conservative constraints. Implementation details are provided in Appendix D.2.

4.2 Main Results

Table 1 reports LIBERO success rates. The unregularized Q-Chunking baseline (**QC-M**) underperforms BC (37.9% vs. 39.8%), reflecting a maximization issue in offline RL. With sparse demonstrations, the critic can overestimate slightly off-demonstration near-miss chunks due to function approximation error, and Best-of-N selection then preferentially picks these overvalued outliers, inducing compounding drift as rollouts enter poorly covered state–action regions [Kumar *et al.*, 2020; Mark *et al.*, 2024].

Adding CQL alleviates this failure, improving QC-M from 37.9% to 42.2% (**QC-M+CQL**), but the gain over BC remains modest. Replacing the MLP with our Transformer backbone yields a similar result (41.8% for **QC-T+CQL**), suggesting that indiscriminate conservative suppression compresses value gaps among proposal-supported candidates, leaving Best-of-N with insufficient ranking resolution.

Finally, **VGAS** achieves 49.0%, outperforming **QC-T+CQL** by a clear margin. Since both methods share the same Q-Chunk-Former backbone, this improvement is primarily attributed to Explicit Geometric Regularization (EGR). See Sec. 4.4 for further comparative analysis.

4.3 Ablation Studies

Table 2 isolates the contribution of each component. The dominant gain arises from *Explicit Geometric Regularization* ($\mathcal{L}_{\text{EGR}}$). Removing $\mathcal{L}_{\text{EGR}}$ leads to a sharp drop in success rate from 49.0% to 36.4%, effectively collapsing performance back to the unregularized QC-M baseline (37.9%). This regression suggests that, without geometric shaping, the critic is susceptible to a canonical offline RL failure mode: it assigns spuriously high values to off-demo candidates, and the subsequent Best-of-N maximization amplifies these overestimations. These results indicate that EGR plays a dual role: beyond suppressing erroneous high values on off-demo actions, it explicitly structures the local value landscape around demonstrations, preserving the ranking resolution required for reliable Best-of-N selection. In addition, ablating the temporal consistency term ($\mathcal{L}_{\text{TD}}$) reduces performance to 45.5%, showing that while EGR provides strong spatial guidance, a TD-based anchor remains necessary to stabilize long-horizon value estimates.

Variants	Spatial	Object	Goal	LONG	Avg
w/o SAF	51.8	54.2	56.8	17.4	45.1
w/o QCF	53.4	55	55.9	16.8	45.3
w/o $\mathcal{L}_{\text{EGR}}$	42.6	42.3	48.2	12.5	36.4
w/o $\mathcal{L}_{\text{TD}}$	55.2	54.5	56.0	16.4	45.5
VGAS	56.2	59.0	60.8	20.0	49

Table 2: Ablation study on LIBERO (success rates (%)).

Architectural choices prove equally critical. We denote our Transformer-based critic as QCF (Q-Chunk-Former). Replacing QCF with a standard MLP backbone (w/o QCF) yields 45.3%, validating the necessity of the Transformer's attention mechanism for modeling complex multimodal dependencies. Notably, ablating the State-Action Fusion mod-

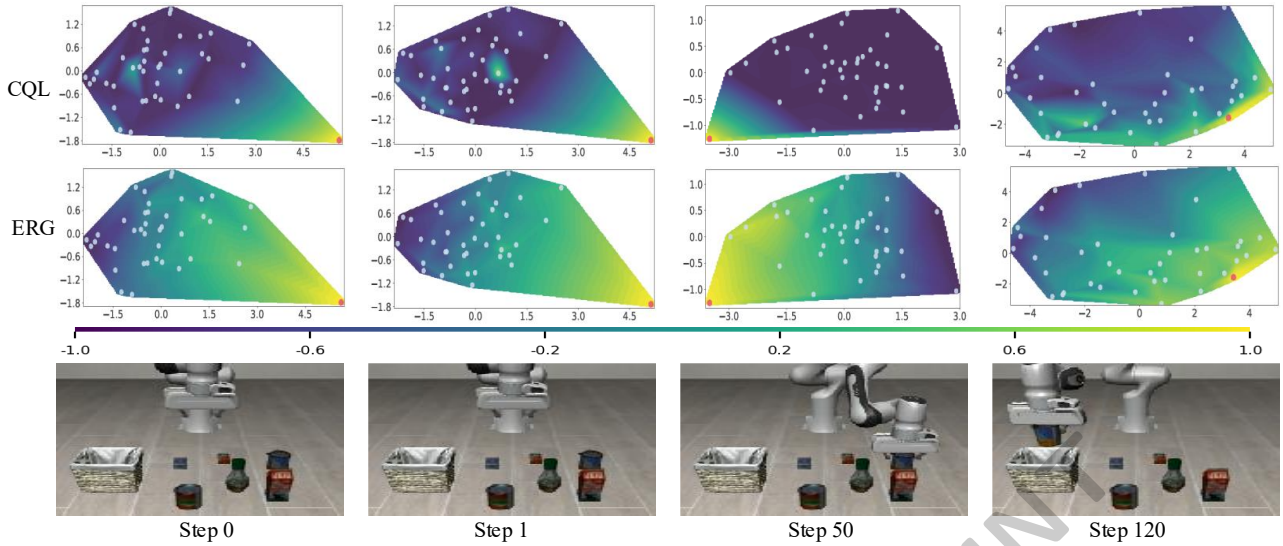
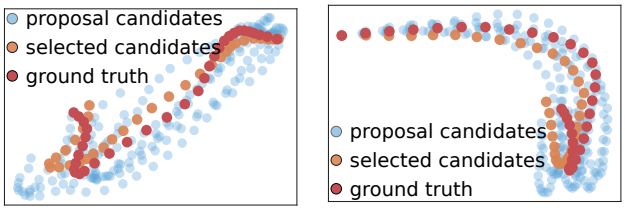


Figure 3: Visualization of the Proposal-Candidate Value Landscape: CQL vs. EGR (Ours)



(a) Projection on X-Z Plane (b) Projection on Y-Z Plane

Figure 4: Multi-view Spatial Rollouts of Action Chunks and **VGAS** Selection. Trajectories are reconstructed via temporal integration in orthogonal views. **VGAS** identifies the trajectory aligning with the expert across 3D space.

ule (w/o SAF) further lowers performance to 45.1%. This suggests that without explicit grounding, the critic struggles to resolve fine-grained geometric ambiguities amidst high-dimensional visual features.

4.4 Visualization Analysis

Figure 3 visualizes the landscape of estimated action-chunk values projected onto a 2D plane, with Q-values normalized to $[-1, 1]$ for comparison. As shown in the top row, standard conservative regularization (CQL) results in a collapsed value landscape. It indiscriminately suppresses the Q-values of all proposal candidates (white dots) to a uniformly low level. Consequently, the critic loses the resolution to differentiate recoverable near-miss proposals from failures, rendering the selection process ineffective. In contrast, the bottom row demonstrates that VGAS successfully restores fine-grained ranking resolution. Instead of uniform suppression, the value signal exhibits a graded geometric preference that decays smoothly as candidates deviate from the expert actions (red dot). This structure ensures that the critic can meaningfully rank candidates based on their physical proximity to the optimal solution. The detailed analysis is in Appendix C.

To assess execution precision, we visualize the spatial rollouts induced by action chunks on held-out task instances. Although these instances remain semantically aligned with the training demonstrations, they contain subtle yet consequential spatial variations. The SFT baseline (blue) exhibits pronounced dispersion, producing a "cloud" of candidates that frequently drifts away from the target. This indicates poor geometric generalization under few-shot supervision: the policy tends to memorize demonstration-specific trajectories rather than adapt its execution to instance-level spatial configurations. In contrast, **VGAS** acts as a geometric stabilizer. By enforcing structural consistency rather than exact path memorization, the learned critic suppresses this variance and selects the candidate (orange) that best matches the current spatial arrangement. Consequently, **VGAS** corrects execution drift induced by rigid imitation in the base policy. Additional experiments on other VLA baselines, inference-budget sensitivity, demonstration-budget scaling, and hyperparameter sensitivity are provided in Appendix E.

5 Conclusion and Limitations

We propose **VGAS**, which reformulates few-shot VLA adaptation as value-guided selection by decoupling high-recall proposal generation from a high-precision geometric critic. Two components jointly address the near-miss failure mode: Q-Chunk-Former grounds multimodal observations into chunk-level value estimates, and EGR prevents value-landscape collapse to preserve ranking resolution. On LIBERO, **VGAS** consistently outperforms SFT and offline RL baselines. Despite these improvements, two practical limitations remain. First, inference-time Best-of-N selection incurs computational latency, limiting applicability in high-frequency control. Second, extending **VGAS** to real-world platforms remains an important direction for future work.

Acknowledgments

The work was supported by the Australian Research Council (ARC) under Laureate project FL190100149.

References

[Bacchiocchi *et al.*, 2024] Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Papini, Alberto Maria Metelli, Nicola Gatti, et al. Online learning with off-policy feedback in adversarial mdps. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 3697–3705, 2024.

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

[Brohan *et al.*, 2022] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[Chebotar *et al.*, 2023] Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, et al. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pages 3909–3928. PMLR, 2023.

[Chen *et al.*, 2025] Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *arXiv preprint arXiv:2502.05450*, 2025.

[Dass *et al.*, 2022] Shivin Dass, Karl Pertsch, Hejia Zhang, Youngwoon Lee, Joseph J Lim, and Stefanos Nikolaidis. Pato: Policy assisted teleoperation for scalable robot data collection. *arXiv preprint arXiv:2212.04708*, 2022.

[Fu *et al.*, 2020] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[Fujimoto *et al.*, 2019] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

[Ghasemipour *et al.*, 2021] Seyed Kamyar Seyed Ghasemipour, Dale Schuurmans, and Shixiang Shane Gu. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *International Conference on Machine Learning*, pages 3682–3691. PMLR, 2021.

[Guo *et al.*, 2025] Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. *arXiv preprint arXiv:2501.16664*, 2025.

[He *et al.*, 2026] Keji He, Yan Huang, Ya Jing, Qi Wu, and Liang Wang. Fine-grained alignment supervision matters in vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6525–6540, 2026.

[Huang *et al.*, 2025] Dongchi Huang, Zhirui Fang, Tianle Zhang, Yihang Li, Lin Zhao, and Chunhe Xia. Co-rft: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning. *arXiv preprint arXiv:2508.02219*, 2025.

[Intelligence *et al.*, 2025] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[Janner *et al.*, 2022] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[Kim *et al.*, 2024] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[Kim *et al.*, 2025] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[Kostrikov *et al.*, 2021] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[Kumar *et al.*, 2020] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Kumar *et al.*, 2022] Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials. *arXiv preprint arXiv:2210.05178*, 2022.

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[Li *et al.*, 2025a] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. Simplevla-rl: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025.

[Li *et al.*, 2025b] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. *arXiv preprint arXiv:2507.07969*, 2025.

[Liu *et al.*, 2023] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero:

Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[Liu *et al.*, 2025] Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can rl bring to vla generalization? an empirical study. *arXiv preprint arXiv:2505.19789*, 2025.

[Lu *et al.*, 2022] Cong Lu, Philip J Ball, Tim GJ Rudner, Jack Parker-Holder, Michael A Osborne, and Yee Whye Teh. Challenges and opportunities in offline reinforcement learning from visual observations. *arXiv preprint arXiv:2206.04779*, 2022.

[Luo *et al.*,] Kairong Luo, CAIWEI XIAO, Zhiao Huang, Zhan Ling, Yunhao Fang, and Hao Su. Dreamfuser: Value-guided diffusion policy for offline reinforcement learning.

[Lyu *et al.*, 2022] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1711–1724, 2022.

[Marafioti *et al.*, 2025] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[Mark *et al.*, 2024] Max Sobol Mark, Tian Gao, Georgia Gabriela Sampaio, Mohan Kumar Srirama, Archit Sharma, Chelsea Finn, and Aviral Kumar. Policy agnostic rl: Offline rl and online rl fine-tuning of any class and backbone. *arXiv preprint arXiv:2412.06685*, 2024.

[Nakamoto *et al.*, 2025] Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. In *Conference on Robot Learning*, pages 4996–5013. PMLR, 2025.

[Sapkota *et al.*, 2025] Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

[Shin and Kim, 2023] Wonchul Shin and Yusung Kim. Guide to control: Offline hierarchical reinforcement learning using subgoal generation for long-horizon and sparse-reward tasks. In *IJCAI*, pages 4217–4225, 2023.

[Shukor *et al.*, 2025] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

[Song *et al.*, 2025] Haoming Song, Delin Qu, Yuanqi Yao, Qizhi Chen, Qi Lv, Yiwen Tang, Modi Shi, Guanghui Ren, Maoqing Yao, Bin Zhao, et al. Hume: Introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*, 2025.

[Sutton *et al.*, 1998] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[Sutton *et al.*, 1999] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[Tan *et al.*,] Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models (2025). *arXiv preprint arXiv:2505.17016*.

[Team *et al.*, 2024] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[Xin *et al.*, 2024] Jimmy Xin, Linus Zheng, Kia Rahmani, Jiayi Wei, Jarrett Holtz, Isil Dillig, and Joydeep Biswas. Programmatic imitation learning from unlabeled and noisy demonstrations. *IEEE Robotics and Automation Letters*, 9(6):4894–4901, 2024.

[Yu *et al.*, 2025] En Yu, Jie Lu, Xiaoyu Yang, Guangquan Zhang, and Zhen Fang. Learning robust spectral dynamics for temporal domain generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Yu *et al.*, 2026] En Yu, Jie Lu, and Guangquan Zhang. Generalized incremental learning under concept drift across evolving data streams. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 3905–3916, 2026.

[Zhang and Tan, 2023] Zhe Zhang and Xiaoyang Tan. Adaptive reward shifting based on behavior proximity for offline reinforcement learning. In *IJCAI*, pages 4620–4628, 2023.

[Zhang *et al.*, 2025] Dapeng Zhang, Jing Sun, Chenghui Hu, Xiaoyan Wu, Zhenlong Yuan, Rui Zhou, Fei Shen, and Qingguo Zhou. Pure vision language action (vla) models: A comprehensive survey. *arXiv preprint arXiv:2509.19012*, 2025.

[Zhao *et al.*, 2023] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost arms. In *Robotics: Science and Systems (RSS)*, 2023.

[Zitkovich *et al.*, 2023] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.