VISION-GROUNDED ENERGY CRITICS FOR PHYSICALLY FEASIBLE VISION-AND-LANGUAGE ACTION POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Vision-Dominated Energy Learning (VEL), a lightweight, backbone-agnostic framework that augments any frozen or fine-tuned vision-and-language action (VLA) policy with a compact energy head. The energy head maps the hidden representations of the VLA and a candidate action to a scalar energy value, where lower energy indicates higher compatibility between the action and the current scene. Training uses a contrastive objective on demonstration data, and inference applies a single gradient-based residual correction that shifts the proposed action toward a lower-energy alternative. This procedure requires no environment rollouts, no additional data collection, and no modification of the VLA backbone. At the population optimum, the contrastive objective recovers the log density ratio between the conditional expert action distribution and the action marginal, and a single gradient step is guaranteed to reduce the learned energy under local smoothness conditions. Over five training seeds of the energy head, VEL raises the LIBERO-Long success rate of OpenVLA-OFT from 94.4% to 95.8% and that of $\pi_{0.5}$ from 92.4% to 93.6%, which removes 24% and 16% of the residual failures, while the shorter suites show smaller changes of mixed sign. Controlled comparisons show that search over the same energy, a direct correction head of matched capacity, and a score-matching energy head do not reproduce the long-horizon gain. On a physical screw-insertion task with a $\pi_{0.5}$ backbone, VEL raises full-task success from 35% to 50%, and the gain concentrates in the precision-demanding insertion stage. The energy head introduces fewer than 4.0M additional parameters and 4.61 GFLOPs per correction step, an increase of 0.4% over OpenVLA-OFT and 5.2% over $\pi_{0.5}$.

1 INTRODUCTION

Vision-and-language action (VLA) models trained through behavior cloning (BC) have become a dominant paradigm for mapping visual observations and language instructions to continuous robot actions (Zhong et al., 2025; Kim et al., 2025b;a; Ghosh et al., 2024). The standard BC objective learns a forward mapping that reproduces expert actions from demonstrated state–action pairs, and the policy is deployed without further refinement. This paradigm gives the policy no internal signal for assessing the physical plausibility of its own proposals. At deployment, the model therefore cannot distinguish a feasible action from one that would collide with a cabinet door, compromise grasp stability, or violate a task precondition such as an object placed in a pose that blocks subsequent manipulation. The absence of such a feasibility check is especially damaging in long-horizon manipulation, where a single locally suboptimal action, such as a bowl placed one centimeter too close to a microwave door, can invalidate every subsequent step and cause complete task failure. These failure modes expose a structural gap in the standard VLA pipeline: the lack of a mechanism that evaluates a candidate action against the current scene configuration and revises it before execution.

Reinforcement learning (RL) offers one route toward corrective capability by fine-tuning a BC policy through reward-driven interaction (Durante et al., 2025; Li et al., 2025; Lu et al., 2025). In instruction-conditioned manipulation with realistic visual observations, however, RL demands extensive online rollouts, careful reward specification, and a safe exploration environment, requirements that limit its practical adoption (Dulac-Arnold et al., 2021; Li, 2022). An alternative line of work applies energy-

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.

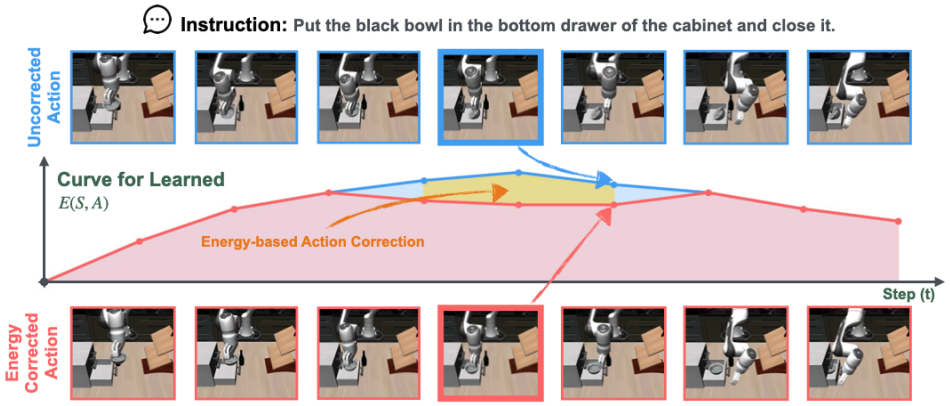


Figure 1: Energy-based action correction. The energy head scores each state–action pair, and one local descent step toward lower energy corrects the action so the bowl no longer blocks the cabinet.

based models (EBMs) to learn a compatibility function over state–action pairs (Du & Mordatch, 2019; Liu et al., 2021; Boney et al., 2020). Implicit Behavior Cloning (Florence et al., 2022) and diffusion-based action generators (Chi et al., 2025) sample actions from a learned energy landscape and can capture multi-modal distributions more faithfully than regression. These models, however, replace the action-generation pipeline entirely; their architectures are coupled to a specific decoder design and cannot be transferred to an arbitrary VLA backbone as a post-hoc correction module. NEAR (Diwan et al., 2025) instead learns an energy from demonstrations and optimizes a policy against it with reinforcement learning. The middle ground, in which an already-produced action is evaluated and refined without retraining the base policy or rolling out new data, remains open.

The signal required for such a critic is already present in the representations a modern VLA computes for every forward pass. Token-level analysis of representative VLA architectures reveals that vision tokens dominate both the input sequence and the cumulative attention mass of the final decoder layer (Remark 1), while language tokens carry compact task intent. The visual stream therefore encodes the information that distinguishes a feasible action from a subtly flawed one: object poses, contact geometry, and the spatial relations that determine whether a placement blocks a door or destabilizes a grasp (Fig. 1). Demonstration data also provides an implicit ranking, since an expert action paired with its own scene is, in a statistical sense, more compatible with that scene than an action drawn from another scene, so a compatibility score is learnable from demonstrations alone.

Building on this observation, we introduce Vision-Dominated Energy Learning (VEL), a lightweight framework that augments any frozen or fine-tuned VLA with a compact energy head trained through a contrastive objective and corrects actions through a single gradient-based residual update at inference. The energy head attaches to the frozen backbone as an external module, keeps the original action-generation pipeline intact, and therefore composes with both regression-based and diffusion-based VLAs without rollouts, reward engineering, or retraining. The correction is a single conservative step near the base proposal, since further descent or search over the learned energy degrades success (Section 4.3). Over five training seeds of the energy head, VEL improves the average success rate of OpenVLA-OFT (Kim et al., 2025a) and $\pi_{0.5}$ (Physical Intelligence et al., 2025) on LIBERO (Liu et al., 2023), with significant gains on the long-horizon suite, and on a physical screw-insertion task it raises the full-task success of $\pi_{0.5}$ from 35% to 50%. The main contributions of this work are:

- An inference-only, energy-based correction paradigm for VLA policies: a compact energy head attaches to any frozen or fine-tuned VLA backbone, and a single gradient step on the learned energy refines BC actions without rollouts, reward signals, or changes to the base model.
- A characterization of the contrastive critic: the InfoNCE optimum recovers the log density ratio between the conditional expert action distribution and the action marginal, so its action gradient equals the expert score minus the marginal score (Proposition 1), and a single step reduces the learned energy under local smoothness (Proposition 2).
- Five-seed LIBERO results with significant gains on LIBERO-Long for both backbones, controls showing that search over the same energy, a matched-capacity correction head,

and a score-matching energy head do not reproduce the long-horizon gain, and a real-robot evaluation in which the correction improves the precision-demanding insertion stage.

Section A reviews related work and relates the contrastive objective to score matching.

2 Preliminary: Behavior Cloning (BC) in VLA Models

A typical VLA receives a visual observation o_t (camera images and a language instruction) and predicts a continuous action $a_t \in [-1, 1]^D$ controlling the robot's end-effector, where D is the action dimensionality. The policy π_θ is realized as a large vision-language model whose parameters θ are fine-tuned on a dataset of expert demonstrations through a BC objective. In the regression-based variant adopted by OpenVLA-OFT (Kim et al., 2025a), the model predicts an action chunk $A_{t:t+K-1}$ of K consecutive actions and minimizes the mean absolute error against the ground-truth chunk $\tilde{A}_{t:t+K-1}$:

$$\mathcal{L}_{\text{BC}} = \frac{1}{KD} \sum_{k=0}^{K-1} \sum_{d=1}^{D} \left| A_{t+k,d} - \tilde{A}_{t+k,d} \right|.$$

The chunk is then executed open-loop before the next observation is queried. Diffusion-based VLAs such as $\pi_{0.5}$ (Physical Intelligence et al., 2025) replace the regression head with a denoising process, and the outer training loop remains imitation-driven. Either pipeline produces a point estimate for each action without any signal of whether it fits the physical configuration of the scene.

Energy-based models (EBMs) offer a principled framework for supplying such a signal. An EBM defines a scalar energy $\mathcal{E}(x)$ and induces the Boltzmann distribution $p(x) \propto \exp(-\mathcal{E}(x)/\tau)$, where the temperature τ controls its sharpness. For action evaluation, x is a state–action pair and the energy is a compatibility score, which contrastive objectives such as InfoNCE (van den Oord et al., 2018) train by pushing the energy of matched pairs below that of mismatched ones.

3 Vision-Dominated Energy Learning

3.1 Problem Formulation

Let π_θ denote a frozen or fine-tuned VLA policy and let $h_t^{(-1)} \in \mathbb{R}^{S \times d_h}$ be the last-layer hidden representation that the backbone produces at timestep t, where S is the token count and d_h is the hidden width. The representation $h_t^{(-1)}$ encodes the visual observation, the language instruction, and the internal computation of the backbone, and the BC head maps it to a predicted action $a_{\text{BC}} \in [-1, 1]^D$. The core limitation is that a_{BC} is accepted unconditionally: no scalar signal rates how compatible a_{BC} is with the physical scene encoded in $h_t^{(-1)}$.

VEL addresses this gap by introducing an energy function $\mathcal{E}_w \colon \mathbb{R}^{S \times d_h} \times [-1, 1]^D \to \mathbb{R}$ parameterized by a compact set of weights w. The energy function maps a state–action pair to a scalar, $e_{a_t} \triangleq \mathcal{E}_w\big(h_t^{(-1)}, a_t\big) \in \mathbb{R}$, where lower energy indicates higher action–scene compatibility. Given this function, inference-time correction reduces to local energy minimization, $a^{\text{opt}} = \arg\min_{a \in [-1,1]^D} \mathcal{E}_w(h_t^{(-1)}, a)$. The question then splits into two sub-problems: (i) how to train $\mathcal{E}_w$ so that it assigns low energy to expert-like actions and high energy to suboptimal ones, and (ii) how to approximate the minimizer efficiently at inference without expensive search.

3.2 Theoretical Foundations

An effective energy function must satisfy three properties: *compatibility* (the energy e_{a_t} is small when a_t matches the expert action under $h_t^{(-1)}$), *discriminability* (the energy usefully ranks candidate actions under the same state), and *differentiability* (the gradient $\nabla_a \mathcal{E}_w(h_t^{(-1)}, a_t)$ exists and is continuous). The following analysis characterizes the ideal contrastive score and a local energy-descent step; the experiments assess candidate ranking and task-level correction.

Proposition 1 (Density-Ratio Recovery of Contrastive Energy)**.** *Let $p_{\text{data}}(a \mid h)$ denote the expert action distribution conditioned on state h, and let $p(a)$ denote the negative-sampling action marginal, which is positive on the relevant expert support. Assume sufficient critic capacity and population*

optimization. For at least one independent marginal negative per positive, the population minimizer $\mathcal{E}_w^{\mathrm{opt}}$ *of the InfoNCE objective satisfies*

$$-\frac{\mathcal{E}_w^{\mathrm{opt}}(h,\,a)}{\tau} \;=\; \log \frac{p_{\mathrm{data}}(a \mid h)}{p(a)} \;+\; c(h),$$

where $c(h)$ *is a state-dependent constant. The quantity the objective recovers is therefore the log density ratio between the conditional expert distribution and the action marginal, and the conditional expert distribution is recovered only with the action marginal as a reference measure:*

$$p_{\mathrm{data}}(a \mid h) = \frac{p(a)\exp\!\big(-\mathcal{E}_w^{\mathrm{opt}}(h,a)/\tau\big)}{\int p(a')\exp\!\big(-\mathcal{E}_w^{\mathrm{opt}}(h,a')/\tau\big)\,\mathrm{d}a'}.$$

Normalizing the exponential energy alone gives a normalized density ratio, not the conditional expert density, unless the action marginal is uniform on the relevant support.

The derivation is provided in Section C.1 and follows interpretations of InfoNCE (van den Oord et al., 2018). Proposition 1 characterizes the ranking at the population optimum; finite batches, critic approximation, and false negatives (one action may suit several states) affect the learned one.

The correction follows the action gradient of the energy. Where the densities are positive and differentiable, the ideal identity gives $-\nabla_a \mathcal{E}_w^{\mathrm{opt}}(h,a)/\tau = \nabla_a \log p_{\mathrm{data}}(a \mid h) - \nabla_a \log p(a)$: the field is the conditional expert score minus the action-marginal score. It measures state-specific compatibility relative to how frequent an action is in the demonstrations, whereas denoising score matching (Vincent, 2011) estimates the score of a noise-smoothed expert distribution (compared in Section 4.3).

Proposition 2 (Local Energy Descent)**.** *Fix a state representation* $h_t^{(-1)}$ *and suppose the energy gradient* $\nabla_a \mathcal{E}_w(h_t^{(-1)}, \cdot)$ *is* L*-Lipschitz continuous on a neighborhood of the BC action that contains the update segment. For the residual correction with step size* $0 < \alpha \le 1/L$ *and before clipping, the energy of the corrected action satisfies*

$$\mathcal{E}_w\big(h_t^{(-1)},\, a_{\mathrm{BC}} - \alpha\, g\big) \;\le\; \mathcal{E}_w\big(h_t^{(-1)},\, a_{\mathrm{BC}}\big) \;-\; \frac{\alpha}{2}\,\|g\|_2^2, \quad \textit{where} \quad g = \nabla_a \mathcal{E}_w\big(h_t^{(-1)},\, a_{\mathrm{BC}}\big).$$

The proof (Section C.2) follows from the descent lemma. The guarantee is local to this neighborhood and concerns the learned energy only; it implies nothing about task success. The bound quantifies descent through the gradient norm at the proposed action, so the correction is largest where the local energy gradient is steep. This quantity is not the distance to the energy minimum, and the two coincide only for a well-conditioned, bowl-shaped landscape; the relationship between energy and the ℓ_1 distance to the expert action is instead examined empirically in Section 4.5. The smoothness assumption is reasonable near the base action, since the energy head comprises linear projections, softmax attention, and SiLU activations, whose composition preserves gradient regularity.

The single-step design is a conservative choice supported by the geometry of the learned landscape rather than a consequence of Proposition 2. The local smoothness that licenses the descent bound cannot be assumed to persist along an iterated descent path, and the learned energy is a proxy calibrated only near the demonstration manifold, so descending it further can reduce energy without improving the action. The k-step ablation and the collapse of unconstrained search over the same energy exhibit this failure empirically (Sections 4.3 and E.5).

Remark 1 (Vision-Token Dominance in VLA Representations)**.** *Token-level analysis of two representative VLA architectures reveals that vision tokens account for over 85% of the total token count in* $h_t^{(-1)}$ *and receive over 65% of the cumulative attention mass in the final decoder layer. Language tokens, by contrast, encode compact task-level intent and occupy a small fraction of the sequence.*

Remark 1 motivates the use of the hidden sequence rich in visual tokens. These tokens carry multimodal context, and language remains relevant to task compatibility (Section E.5).

3.3 Framework and Energy Head Architecture

VEL instantiates the energy critic as a three-stage pipeline attached to a frozen VLA backbone (Fig. 2): the backbone produces $h_t^{(-1)}$ and a_{BC} through its standard forward pass, the energy head $\mathcal{E}_w$

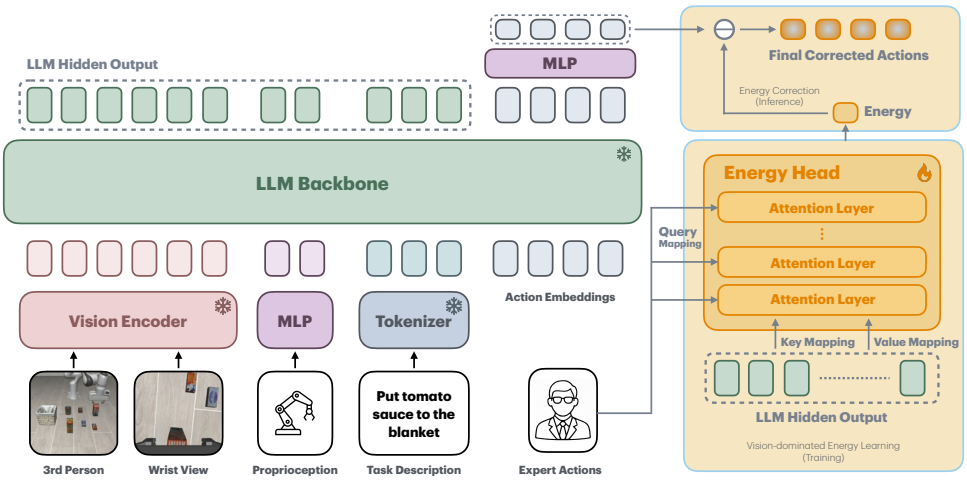


Figure 2: Overview of VEL. The frozen backbone yields hidden features $h_t^{(-1)}$ (action-slot tokens $\mathbf{e}$ masked) and a BC action a_{BC}. The action embedding queries the state tokens through MHA, a SiLU feed-forward readout gives a scalar energy, and its gradient corrects a_{BC} in one residual step.

fuses them through cross-attention into a scalar energy, and a single residual correction shifts a_{BC} toward a lower-energy alternative. The backbone parameters θ stay frozen throughout. In a shared space of dimension d_e, a projector P_h maps the hidden features to a token sequence, a projector P_a maps the action to a single embedding, and the action embedding attends to the state tokens through multi-head attention (MHA):

$$Z_h = P_h\big(h_t^{(-1)}\big) \in \mathbb{R}^{S\times d_e}, \quad z_a = P_a\big(a_t\big) \in \mathbb{R}^{d_e}, \quad \tilde{z}_a = \mathrm{MHA}\big(Q = z_a, K = Z_h, V = Z_h\big).$$

The MHA mechanism allows the action representation to selectively attend to $h_t^{(-1)}$, so the energy computation is grounded primarily in visual evidence of the scene configuration. Using the action as query and the state tokens as keys and values lets the action embedding attend to task-relevant spatial regions, whereas the reverse assignment or simple concatenation would compress the state sequence into a bottleneck before it interacts with the action. A feed-forward block with SiLU (Elfwing et al., 2018) activation and residual connections further processes $\tilde{z}_a$, and a final linear readout maps the fused representation to a scalar energy, $e_{a_t} = W_{\mathrm{out}}^{\top}\,\mathrm{SiLU}\big(\mathrm{FFN}(\tilde{z}_a)\big) + b_{\mathrm{out}}$. Here $\mathbf{e}$ denotes the action-slot tokens, the positions in the backbone's input sequence that hold the placeholder embeddings from which the BC head decodes its action prediction (Fig. 2). The tokens corresponding to $\mathbf{e}$ in $h_t^{(-1)}$ are masked when constructing Z_h, which ensures that the energy head evaluates the candidate action against the scene representation rather than against the backbone's own prediction.

3.4 Contrastive Training Objective

The energy head is trained with a contrastive objective that encourages the compatibility and discriminability properties described in Section 3.2. Let $\{(h_b, a_b^+)\}_{b=1}^B$ be a mini-batch of state representations and their paired expert actions. An in-batch InfoNCE (van den Oord et al., 2018) loss treats a_b^+ as the positive for state h_b and all remaining actions $\{a_j\}_{j\neq b}$ as negatives:

$$\mathcal{L}_{\mathrm{VEL}} = \frac{1}{B}\sum_{b=1}^{B}\left[-\log\frac{\exp\big(-\mathcal{E}_w(h_b,\, a_b^+)/\tau\big)}{\sum_{j=1}^{B}\exp\big(-\mathcal{E}_w(h_b,\, a_j)/\tau\big)}\right],$$

where $\tau > 0$ is a temperature hyperparameter. The in-batch design supplies negatives without a separate mining procedure. Cross-state actions can be inappropriate despite geometric similarity, but some remain valid in more than one state, so the objective encourages state-dependent compatibility while admitting such false negatives. Gradients are stopped before the backbone, so only w is updated, with standard weight decay, and the critic cannot alter the policy's representation.

3.5 Energy-Based Residual Action Correction

At inference, VEL avoids exhaustive search over the continuous action space and refines the BC action through one gradient step toward lower energy. With the energy gradient $g \triangleq \nabla_a \mathcal{E}_w(h_t^{(-1)}, a_{\mathrm{BC}})$ at

Table 1: Success rates (%) on LIBERO (Liu et al., 2023). ‡: no wrist camera. VEL: mean$_{\pm\text{std}}$ over five energy-head seeds; others: one run of a released or frozen checkpoint. Bold: best per block.

Method	Success rate (%) ↑				
	Spatial	Object	Goal	Long	Average
Diffusion Policy (Chi et al., 2025) (scratch)‡	78.3	92.5	68.3	50.5	72.4
Octo (Ghosh et al., 2024) (fine-tuned)‡	78.9	85.7	84.6	51.1	75.1
DiT Policy (Hou et al., 2024) (fine-tuned)‡	84.2	96.3	85.4	63.8	82.4
OpenVLA (Kim et al., 2025b) (fine-tuned)‡	84.7	88.4	79.2	53.7	76.5
OpenVLA-OFT (Kim et al., 2025a)‡	96.2	98.3	96.2	90.7	95.3
Multimodal Diffusion Transformer (Reuss et al., 2024)	78.5	87.5	73.5	64.8	76.0
π_0 (Black et al., 2024) + FAST (Pertsch et al., 2025)	96.4	96.8	88.6	60.2	85.5
π_0 (Black et al., 2024)	96.8	98.8	95.8	85.2	94.2
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	98.8	98.2	98.0	92.4	96.8
$\pi_{0.5}$ + VEL (ours)	$98.2_{\pm 0.4}$	$97.6_{\pm 0.8}$	$98.3_{\pm 0.5}$	$\mathbf{93.6}_{\pm 0.6}$	$\mathbf{97.0}_{\pm 0.2}$
OpenVLA-OFT (Kim et al., 2025a)	97.4	98.4	96.8	94.4	96.7
OpenVLA-OFT + VEL (ours)	$97.9_{\pm 0.9}$	$98.2_{\pm 1.7}$	$97.5_{\pm 0.8}$	$\mathbf{95.8}_{\pm 0.9}$	$\mathbf{97.3}_{\pm 0.3}$

the proposed action, the corrected action is obtained through a residual update:

$$\tilde{a} \;=\; a_{\mathrm{BC}} - \alpha\, g, \qquad a_{\mathrm{out}} \;=\; \mathrm{clip}\big(\tilde{a},\; -\mathbf{1},\; \mathbf{1}\big), \tag{1}$$

where $\alpha > 0$ is a small step size and the clipping enforces the action bounds. A gradient-norm clipping threshold γ is applied before the update, $g \leftarrow g \cdot \min(1, \gamma/(\|g\|_2 + \varepsilon))$, to prevent overshooting in regions where the energy landscape is steep. Proposition 2 applies to the unmodified gradient before action projection. For the implemented update, let $c = \min(1, \gamma/(\|g\|_2 + \varepsilon))$ and $a_{\mathrm{out}} = \Pi_{[-1,1]^D}(a_{\mathrm{BC}} - \alpha c\, g)$, with g the original gradient and Π the Euclidean projection. Under the same smoothness condition on the projected update segment and $\alpha c \le 1/L$, the displacement-based bound $\mathcal{E}_w(h_t^{(-1)}, a_{\mathrm{out}}) \le \mathcal{E}_w(h_t^{(-1)}, a_{\mathrm{BC}}) - \|a_{\mathrm{out}} - a_{\mathrm{BC}}\|_2^2/(2\alpha c)$ holds (Section C.3). Neither bound guarantees task success. A fresh hidden representation is computed at each new observation, which need not occur at every control step. For VLAs with action chunking, such as OpenVLA-OFT, VEL corrects each action in the chunk independently before open-loop execution; for diffusion-based VLAs such as $\pi_{0.5}$, the correction is applied to the final denoised action sample.

4 Experimental Results

4.1 Experimental Setup

Benchmark and Metric. Our experiments were conducted on the LIBERO (Liu et al., 2023) benchmark, which employs a simulated Franka Emika Panda arm and provides demonstrations with multiple camera images, robot state readings, task annotations, and delta end-effector pose actions. Its four task suites (Spatial, Object, Goal, and Long) each contain 500 expert demonstrations across ten tasks and evaluate generalization over spatial configurations, object variations, goal changes, and long-horizon manipulation. For the LIBERO adaptation, unsuccessful demonstrations were filtered out and models pre-trained with VLM objectives were fine-tuned by LoRA (Hu et al., 2022). The metric is the Success Rate (SR), the percentage of successful episodes.

Baselines. To test the backbone-agnostic claim, VEL is evaluated on (i) *OpenVLA-OFT* (Kim et al., 2025a), which uses parallel decoding, a continuous action head, and an ℓ_1 regression loss, and (ii) $\pi_{0.5}$ (Physical Intelligence et al., 2025), a diffusion-based VLA co-trained on heterogeneous data.

Evaluation Protocol. Each VEL result reports the mean and standard deviation over five independent training seeds of the energy head, with 500 evaluation episodes per suite per seed. The base-policy rows evaluate the frozen checkpoint once under deterministic decoding as a fixed reference. Significance is assessed per suite by a two-sided one-sample t-test on the five seed-level differences from that reference ($\mathrm{df} = 4$). All success rates use the single-step correction at $k = 1$ and $\alpha = 0.1$.

4.2 Quantitative Results

Overall Performance. Table 1 summarizes SR across all four LIBERO suites. VEL improves the average SR for both backbones, and the improvement concentrates on LIBERO-Long, where it lifts

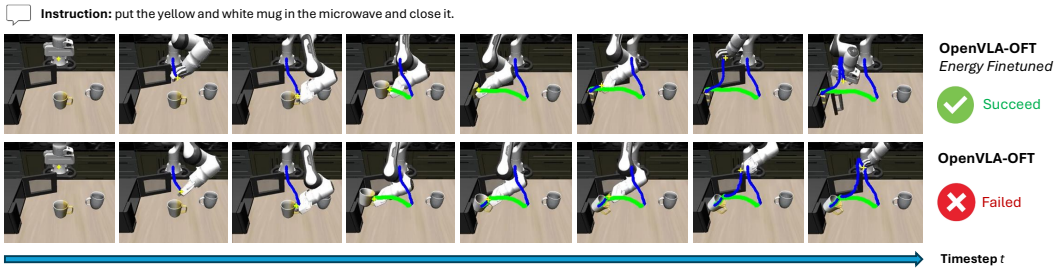


Figure 3: LIBERO-Long task index 8. **Top**: OpenVLA-OFT with VEL succeeds. **Bottom**: OpenVLA-OFT alone fails. Blue/green: closed/open gripper; yellow crosses: end effector.

Table 2: Refinement mechanisms on OpenVLA-OFT (%). Learned heads use the pooled demonstrations of all suites and the five-seed protocol of Table 1; gradient heads use $k = 1$, $\alpha = 0.1$.

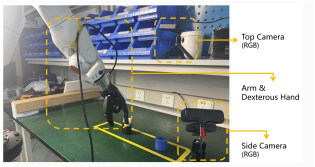
Method	Spatial	Object	Goal	Long	Average
OpenVLA-OFT (Kim et al., 2025a) (base)	97.4	98.4	96.8	94.4	96.7
Search over the trained energy					
+ Best-of-N (noise injection)	$97.3_{\pm 2.7}$	$97.8_{\pm 0.8}$	$93.9_{\pm 2.7}$	$92.6_{\pm 3.4}$	$95.4_{\pm 0.4}$
+ Best-of-N (policy resampling)	$82.2_{\pm 4.2}$	$84.4_{\pm 1.1}$	$68.8_{\pm 2.6}$	$55.8_{\pm 5.5}$	$72.8_{\pm 2.4}$
+ IBC-style sampling (CEM, anchored)	$96.8_{\pm 1.5}$	$97.8_{\pm 1.1}$	$91.6_{\pm 2.2}$	$91.8_{\pm 2.8}$	$94.5_{\pm 0.7}$
Learned heads of matched capacity					
+ Direct correction head	$88.0_{\pm 1.5}$	$95.4_{\pm 0.3}$	$86.7_{\pm 1.6}$	$69.6_{\pm 0.5}$	$84.9_{\pm 0.4}$
+ Score-matching energy head	$98.1_{\pm 0.5}$	$97.6_{\pm 0.5}$	$98.1_{\pm 0.3}$	$92.8_{\pm 1.0}$	$96.7_{\pm 0.5}$
+ VEL (ours)	$97.9_{\pm 0.9}$	$98.2_{\pm 1.7}$	$97.5_{\pm 0.8}$	$\mathbf{95.8}_{\pm 0.9}$	$\mathbf{97.3}_{\pm 0.3}$

OpenVLA-OFT from 94.4% to $95.8 \pm 0.9\%$ and $\pi_{0.5}$ from 92.4% to $93.6 \pm 0.6\%$. Both gains are significant, at $t(4) = 3.30$, $p = 0.030$ and $t(4) = 4.45$, $p = 0.011$ respectively, and the mean gain exceeds the cross-seed standard deviation on this suite for both backbones. All five $\pi_{0.5}$ seeds and four of the five OpenVLA-OFT seeds improve on the base policy. Read as a reduction in failure rate, the correction removes 24% of the residual failures of OpenVLA-OFT on LIBERO-Long, from 5.6% to 4.24%, and 16% of those of $\pi_{0.5}$, from 7.6% to 6.36%, which is the quantity a per-step feasibility filter is expected to move. The three shorter suites operate between 96.8% and 98.8% with smaller changes of mixed sign. The concentration on LIBERO-Long is consistent with local correction reducing errors across an extended action sequence, although high baseline scores and task composition also shape this comparison. Fig. 3 illustrates the mechanism on LIBERO-Long task index 8: VEL adjusts the end-effector pose during the initial mug grasp, so the mug clears the microwave door edge instead of becoming caught on it (Section E.6).

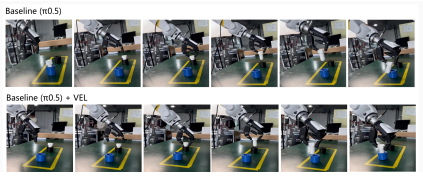
4.3 Refinement Mechanisms and Controls

The improvement could instead come from any test-time refinement, from the parameters the head adds, or from energy learning in general rather than from the field the contrastive objective recovers. Table 2 separates these explanations under a shared backbone, data, step size, and protocol, so the learned heads differ only in their output parameterization and training objective.

Single-step correction outperforms sampled search. Best-of-N with noise injection perturbs the base proposal and keeps the candidate the energy ranks lowest, Best-of-N with policy resampling draws candidates from the base policy, and the IBC-style baseline runs a cross-entropy method (CEM) over the same energy anchored to the base action (Section D). Every one of these mechanisms falls below the base policy, and the damage grows with the horizon: policy resampling loses 38.6 points on LIBERO-Long, and even the anchored search stays below the baseline despite 384 energy evaluations per correction step, against the single forward and backward pass of the gradient step. An unconstrained CEM variant collapsed within the first episodes. A candidate that replaces the base action discards the kinematics the base policy established, and the resulting deviation compounds over the episode. Implicit Behavior Cloning obtains an energy valid over the whole action space through counter-example mining and iterative sampling at both training and inference (Florence et al., 2022); the critic here is trained and used as a local ranking signal, so its information is confined to a narrow neighborhood of the base action, where the method exploits its gradient.



(a) Arm, dexterous hand, and top and side RGB cameras.



(b) $\pi_{0.5}$ (top) approaches off-axis and fails; with VEL (bottom) it inserts.

Policy	Grasp	Insert	Full
$\pi_{0.5}$	82	44	35
+ VEL	83	**60**	**50**

(c) Success (%), 100 runs; Full = both stages.

Figure 4: Real-robot evaluation of VEL on screw insertion with a $\pi_{0.5}$ backbone.

Matched-capacity correction. The direct correction head matches the VEL head in capacity and regresses a residual correction instead of descending an energy. It falls below the base policy on every suite, at 84.9% on average and 69.6% on LIBERO-Long, so added capacity alone does not explain the VEL gains. A regression head always emits a correction and must average over multimodal expert deviations, whereas the energy gradient adapts its magnitude to the local landscape.

Contrastive correction retains its benefit on long-horizon tasks. The score-matching head parameterizes a bounded energy trained with denoising score matching (Vincent, 2011), the objective of NEAR (Diwan et al., 2025). It reaches 98.1% on Spatial and Goal, but on LIBERO-Long it falls below the base policy, at 92.8% against 94.4%, while the contrastive head retains its largest margin. The absolute score contains the marginal component $\nabla_a \log p(a)$, which points toward globally frequent actions irrespective of the state, while the contrastive field subtracts it (Section 3.2). Over a long horizon, the accumulation of this state-independent pull plausibly works against corrections that must remain specific to the state, and the fixed noise scale of score matching may also contribute; we state these as mechanisms consistent with the data rather than as established causes.

4.4 Real-Robot Evaluation

To test whether the correction matters under real contact, the physical study uses a Rokae xMate arm with torque sensing in all seven joints, an Inspire RH56 hand with six actuated degrees of freedom over twelve joints, and side and overhead RGB cameras whose complementary views supply depth cues (Fig. 4a). The task is to pick up a white 3D-printed screw from a black base and insert it into a blue base; insertion tolerates little positional error, so a locally unsuitable action is punished where the critic is meant to act. A $\pi_{0.5}$ (Physical Intelligence et al., 2025) backbone is fine-tuned for five epochs on 300 teleoperated demonstrations, the energy head on the same demonstrations with the backbone frozen, and the correction uses one step at $\alpha = 0.01$, smaller than in simulation since a real end-effector cannot absorb the overshoot a simulated one tolerates. Each policy runs 100 episodes.

Fig. 4c places the effect where the simulation predicted it. Grasping is almost unchanged, at 82 against 83 percent, because the base policy already reaches the screw reliably. Insertion rises from 44 to 60 percent and the full task from 35 to 50 percent, and the two gains coincide, so the episodes VEL recovers are those that grasped correctly and would have failed on the precise placement that follows. As on LIBERO-Long, the correction acts on the stage where a small pose error has consequences and leaves competent behavior alone (Fig. 4b). The energy during execution (Section E.1) has a narrower band and a lower mean under correction, with the largest reduction in the insertion stage.

4.5 Evaluation of the Learned Energy Head

Energy Landscape. Fig. 5 visualizes the energy surface in the neighborhood of expert actions. The plots reveal a pronounced valley with the expert action at the minimum and non-uniform curvature within the surrounding valley. This local geometry motivates a bounded single step, with descent conditions in Proposition 2, and the k-step ablation (Section E.5) confirms that repeated corrections degrade performance. The energy also correctly orders actions from random proposals to ground-truth actions, and its magnitude scales with the ℓ_1 distance to the expert action (Section E.3).

Action Correction On-the-Fly. Consistent with Proposition 2, the corrected action has lower energy than the original proposal at most inference timesteps, although not at all of them, and the two energies remain close because α is small. The contribution of VEL is most visible at failure-prone moments, such as a phase in which the baseline policy drives the end-effector into a moka-pot handle and the correction redirects the motion (Section E.2).

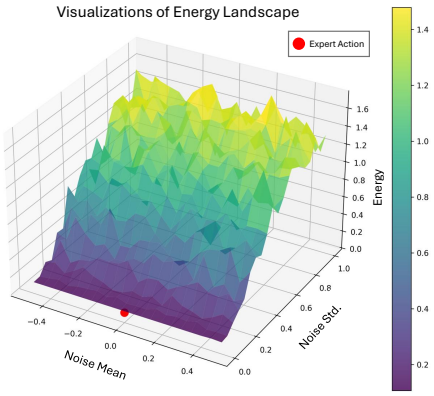


Figure 5: Learned energy landscape; the red dot marks the expert action.

Table 3: Per-step inference cost (mean of five steps, one GPU); the lower block decomposes the remote $\pi_{0.5}$ deployment.

Component	Latency (ms) ↓	GFLOPs ↓
OpenVLA-OFT	103.3	1178.83
+ VEL	106.4	1183.44
$\pi_{0.5}$	80.2	88.92
+ VEL	83.3	93.53
Decomposition for $\pi_{0.5}$		
Energy head forward pass	0.7	1.53
Energy gradient (backward)	2.4	3.08
Client-server transfer	92.6	–
Total, remote deployment	175.9	93.53
Total, co-located	83.3	93.53

Energy, Task Outcome, and Robustness. Over 100 LIBERO-Long episodes of an earlier energy-head checkpoint, episode-level mean energy separates successful from failed episodes with an AUROC of 0.749, and the cumulative energy reduction correlates with task success at 0.511 ($p = 5.48 \times 10^{-8}$). Failed episodes receive more than three times the cumulative correction, which is consistent either with episodes already in difficulty eliciting sustained correction or with the corrections themselves disturbing execution; no failure of that checkpoint was classified as a collision with scene geometry. Under paired object perturbation on LIBERO-Spatial, VEL improves success by 3.4 points ($p = 0.043$) with 25 rescues and 8 losses over 500 episodes, whereas camera perturbation gives -0.4 points ($p = 0.37$) with 14 rescues and 16 losses. Section E.4 reports these analyses in full, together with the failure-mode breakdown.

4.6 Inference Cost

VEL is designed as an external critic for different backbones, so its deployment cost is part of the evaluation, and the measurement has to separate the method from the setup in which it runs. A direct profile of the energy head gives 1.53 GFLOPs for a forward pass and 3.08 GFLOPs for the gradient, hence 4.61 GFLOPs per correction step on top of the backbone forward pass. Against the 1178.83 GFLOPs of OpenVLA-OFT this is an increase of 0.4%, and against the 88.92 GFLOPs of $\pi_{0.5}$ it is 5.2%. The corresponding latency is 0.7 ms for the forward pass and 2.4 ms for the gradient. Table 3 decomposes the per-step cost for $\pi_{0.5}$, whose policy runs on a remote server while the simulation runs on the client. The total in that deployment is 175.9 ms against a base of 80.2 ms, and 92.6 ms of the difference is serialization and client-server communication. Co-locating the policy removes that confound and leaves 83.3 ms, within a few milliseconds of the base model. The near doubling of latency is therefore a property of the deployment rather than of the correction, and the efficiency profile is consistent across backbones once the two are separated.

5 Conclusion

This paper introduced Vision-Dominated Energy Learning (VEL), a backbone-agnostic energy critic that refines behavior-cloned actions through a single gradient-based residual correction at inference. Its contrastive objective estimates a conditional-to-marginal density ratio at the population optimum, and the local descent analysis characterizes the energy reduction from a bounded gradient step. The learned energy landscape exhibits an interpretable valley centered on expert actions, and energy magnitudes correlate with the ℓ_1 error of the action. Over five energy-head training seeds, VEL removes approximately 24 and 16 percent of the residual LIBERO-Long failures of OpenVLA-OFT and $\pi_{0.5}$, while the shorter suites show smaller, mixed changes. Comparisons with search, regression, and score-matching heads support the contrastive single-step design on long-horizon tasks, and the gain holds under object perturbation but not under camera perturbation. On a physical screw-insertion task VEL raises full-task success from 35 to 50 percent, with the gain concentrated in the precision-demanding insertion stage. With under 4.0M parameters and a small fraction of backbone compute, VEL is a practical add-on for existing VLA pipelines; Section B discusses its limitations.

Reproducibility Statement

Section 3 specifies the energy head, the contrastive objective, and the residual correction. Section D lists all hyperparameters. Section 4.1 describes the evaluation protocol with the number of seeds, the number of episodes, and the statistical test, and Section 4.4 describes the real-robot setting. Section C states the assumptions of both propositions and gives complete proofs. The source code of VEL is provided in the supplementary material, and its evaluation script uses $k = 1$ and $\alpha = 0.1$ as defaults.

AI Use Statement

In this work, we used generative AI tools to draft sections of the paper and to edit the text for grammar, spelling, and word choice. We have reviewed all AI-assisted work and take responsibility for the final content, which covers any text, claims, or artifacts produced with the aid of generative AI.

References

Kevin Black, Noah Brown, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint*, 2024.

Rinu Boney, Juho Kannala, and Alexander Ilin. Regularizing model-based planning with energy-based models. In *Conference on Robot Learning (CoRL)*, 2020.

Kang Chen et al. π_{RL}: Online RL fine-tuning for flow-based vision-language-action models. *arXiv preprint*, 2025.

Cheng Chi, Zhenjia Xu, et al. Diffusion policy: Visuomotor policy learning via action diffusion. *International Journal of Robotics Research*, 2025.

Yung-Sung Chuang et al. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Anish Abhijit Diwan, Julen Urain, Jens Kober, and Jan Peters. Noise-conditioned energy-based annealed rewards (NEAR): A generative framework for imitation learning from observation. In *International Conference on Learning Representations (ICLR)*, 2025.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2019.

Gabriel Dulac-Arnold et al. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 2021.

Zane Durante, Ran Gong, et al. An interactive agent foundation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 2018.

Pete Florence et al. Implicit behavioral cloning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

Dibya Ghosh et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

Yanjiang Guo et al. Improving vision-language-action model with online reinforcement learning. *arXiv preprint*, 2025.

Tuomas Haarnoja et al. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning (ICML)*. PMLR, 2017.

Zhi Hou, Tianyi Zhang, Yuwen Xiong, Hengjun Pu, Chengyang Zhao, Ronglei Tong, Yu Qiao, Jifeng Dai, and Yuntao Chen. Diffusion transformer policy. *arXiv preprint arXiv:2410.15959*, 2024.

Edward J Hu, Yelong Shen, et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

Jiangyong Huang et al. An embodied generalist agent in 3D world. *arXiv preprint*, 2023.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*, 2024.

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

Moo Jin Kim, Karl Pertsch, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2025b.

Ilya Kostrikov et al. Offline reinforcement learning with Fisher divergence critic regularization. In *International Conference on Machine Learning (ICML)*, 2021.

Haozhan Li et al. SimpleVLA-RL: Scaling VLA training via reinforcement learning. *arXiv preprint*, 2025.

Xinghang Li and Minghuan others Liu. Vision-language foundation models as effective robot imitators. *arXiv preprint*, 2023.

Yuxi Li. Reinforcement learning in practice: Opportunities and challenges. *arXiv preprint*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, et al. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Minghuan Liu, Tairan He, Minkai Xu, and Weinan Zhang. Energy-based imitation learning. In *International Conference on Autonomous Agents and MultiAgent Systems*, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, et al. RDT-1B: A diffusion foundation model for bimanual manipulation. *arXiv preprint*, 2024.

Guanxing Lu et al. VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning. *arXiv preprint*, 2025.

Tim Pearce, Tabish Rashid, et al. Imitating human behaviour with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2023.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Physical Intelligence et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint*, 2025.

Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *Robotics: Science and Systems (RSS)*, 2024.

Tal Schuster, Adam Fisch, et al. Confident adaptive language modeling. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Nur Muhammad Shafiullah et al. Behavior transformers: Cloning k modes with one stone. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:22955–22968, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. BranchyNet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint*, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Quan Vuong, Sergey Levine, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *Conference on Robot Learning (CoRL)*, 2023.

Junjie Wen, Yichen Zhu, et al. TinyVLA: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters (RAL)*, 2025.

Hongzhi Zang et al. RLinf-VLA: A unified and efficient framework for VLA+RL training. *arXiv preprint arXiv:2510.06710*, 2025.

Simon Zhai et al. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tony Z Zhao et al. Learning fine-grained bimanual manipulation with low-cost hardware. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Haoyu Zhen, Xiaowen Qiu, et al. 3D-VLA: A 3D vision-language-action generative world model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Yifan Zhong et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint*, 2025.

Brianna Zitkovich, Tianhe Yu, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, pp. 2165–2183. PMLR, 2023.

A Related Work

Fine-Tuning VLMs for Robotics. The majority of VLA models use pre-trained VLMs as a backbone to obtain strong vision–language representations and then fine-tune these models through behavior cloning to produce policies that map visual observations and natural language instructions to continuous actions. During fine-tuning, many works project encoded features to actions through a linear action head (Zhao et al., 2023; Shafiullah et al., 2022; Kim et al., 2025b; Zitkovich et al., 2023; Vuong et al., 2023; Li & Liu, 2023; Durante et al., 2025; Huang et al., 2023; Zhen et al., 2024; Wen et al., 2025). Motivated by the improved multimodal modeling of diffusion methods, several approaches instead decode features with a diffusion head (Black et al., 2024; Physical Intelligence et al., 2025; Liu et al., 2024; Ghosh et al., 2024; Pearce et al., 2023; Chi et al., 2025; Kim et al., 2025a). For example, OpenVLA (Kim et al., 2025b) fine-tuned pre-trained VLMs (Karamcheti et al., 2024) on large-scale robot trajectory data, and OpenVLA-OFT (Kim et al., 2025a) further optimized the fine-tuning pipeline through parallel decoding and action chunking. These approaches retain the efficiency and simplicity of behavior cloning, but they lack a mechanism that distinguishes suboptimal actions before errors accumulate in unseen scenes. In contrast, VEL ranks candidate actions against their context and can identify and correct a suboptimal proposal.

Several works fine-tune VLMs with reinforcement learning to improve long-horizon behavior (Zhai et al., 2024; Chen et al., 2025; Li et al., 2025; Zang et al., 2025; Lu et al., 2025; Guo et al., 2025). SimpleVLA-RL (Li et al., 2025) builds on OpenVLA-OFT and fine-tunes the policy using GRPO (Shao et al., 2024), while VLA-RL (Lu et al., 2025) proposes a task-specific robot process reward to enhance the data processing pipeline for RL fine-tuning. These RL-based methods can improve long-horizon planning by leveraging interaction for credit assignment and recovery from mistakes. Reinforcement learning remains costly in practice due to sample inefficiency, reward design burden, and safety risks during exploration. VEL occupies a complementary niche: rather than optimizing the policy through environment interaction, it corrects individual actions at inference through a learned energy critic. The two approaches are not mutually exclusive; an RL-fine-tuned policy can still benefit from energy-based correction at deployment.

Energy-Based Action Policies. Energy-based models (EBMs) (Du & Mordatch, 2019) have been applied to both reinforcement learning (Haarnoja et al., 2017) and imitation learning (Boney et al., 2020; Liu et al., 2021; Kostrikov et al., 2021; Florence et al., 2022). Implicit Behavior Cloning (Florence et al., 2022) learns an energy function over state–action pairs and samples actions from the resulting landscape at inference, while Diffusion Policy (Chi et al., 2025) extends this idea with iterative denoising. Both approaches use the energy function as the action generator: the energy landscape *replaces* the policy's decoder and is tightly coupled to a specific architecture. VEL differs in a fundamental way: the energy head is an *external critic* that evaluates and corrects an action already produced by an unmodified VLA backbone. This separation means VEL composes with any existing policy without retraining, and the correction is a single gradient step rather than an iterative sampling procedure. The external-critic perspective also connects to inference-time verification strategies in language modeling, where a separately trained reward model reranks or refines candidate outputs; VEL adapts this principle to continuous action spaces with gradient-based refinement instead of discrete reranking.

Energy Learning Objectives. The closest energy-learning objective is that of NEAR (Diwan et al., 2025), which learns noise-conditioned energy functions from demonstrations through annealed score matching and then trains a policy with reinforcement learning to minimize the learned energy. The two objectives estimate different fields. Differentiating Proposition 1 with respect to the action shows that the InfoNCE (van den Oord et al., 2018) gradient field is the difference between the state-conditioned expert score and the score of the action marginal,

$$\nabla_a \mathcal{E}^{\mathrm{opt}}(h, a) = -\tau\big(\nabla_a \log p_{\mathrm{data}}(a \mid h) - \nabla_a \log p(a)\big),$$

whereas denoising score matching (Vincent, 2011), as instantiated in NEAR, estimates the absolute score of the noise-smoothed data distribution at a chosen noise scale. The contrastive field subtracts the marginal score, a term shared across states, and requires no noise scale or annealing schedule. This relationship yields a testable prediction: the two objectives should behave alike where a generic pull toward frequent actions is harmless, and should separate where the correction must be precisely conditioned on the state. Section 4.3 evaluates this prediction with the same architecture and data, and VEL uses the resulting energy to edit an already produced action through one gradient step rather than as a reward for policy optimization.

B LIMITATIONS

VEL corrects proposals near the support of the demonstration data. Larger deviations require recovery or replanning beyond this local operation, and repeated descent can leave the region where the energy was supervised, as the multi-step ablation illustrates. Adaptive step control and training against spurious minima could support larger corrections. The critic also adds computation at each controlled step. A second boundary concerns the representation supplied to the critic. Section E.4 reports an improvement under object perturbation and no measurable gain under camera perturbation, so the dependence of the critic on the features of the base model may limit correction under perceptual shift. The physical evidence comes from one insertion task on one platform with a single backbone, so it establishes that the correction transfers to real contact rather than how far it generalizes across tasks, embodiments, and sensing conditions.

C THEORETICAL PROOFS

C.1 PROOF OF PROPOSITION 1 (DENSITY-RATIO RECOVERY OF CONTRASTIVE ENERGY)

The derivation follows from the mutual-information interpretation of InfoNCE (van den Oord et al., 2018). Consider the InfoNCE objective with B samples. For a given state h, the positive pair is (h, a^+) drawn from $p_{\text{data}}(a \mid h)$, and the $B-1$ negatives $\{a_j\}_{j=2}^{B}$ are drawn i.i.d. from the marginal $p(a)$. The InfoNCE loss for this sample is

$$\mathcal{L}(h) = -\log \frac{\exp(-\mathcal{E}_w(h, a^+)/\tau)}{\exp(-\mathcal{E}_w(h, a^+)/\tau) + \sum_{j=2}^{B} \exp(-\mathcal{E}_w(h, a_j)/\tau)}.$$

van den Oord et al. (2018) show that minimization of this objective over the critic function maximizes a lower bound on the mutual information $I(h; a)$. Defining the log-critic as $f_w(h, a) \triangleq -\mathcal{E}_w(h, a)/\tau$, the optimal population critic, under adequate capacity and independent marginal negatives, satisfies

$$f_w^{\text{opt}}(h, a) \;=\; \log \frac{p_{\text{data}}(a \mid h)}{p(a)} \;+\; c(h),$$

where $c(h)$ is a constant that depends only on h and leaves the contrastive ranking unchanged at fixed h. Substituting back $f_w^{\text{opt}} = -\mathcal{E}_w^{\text{opt}}/\tau$ yields

$$-\frac{\mathcal{E}_w^{\text{opt}}(h, a)}{\tau} \;=\; \log \frac{p_{\text{data}}(a \mid h)}{p(a)} \;+\; c(h).$$

Recovery of the conditional expert density requires the action marginal as a reference measure:

$$\frac{p(a) \exp\big(-\mathcal{E}_w^{\text{opt}}(h, a)/\tau\big)}{\int p(a') \exp\big(-\mathcal{E}_w^{\text{opt}}(h, a')/\tau\big)\, \mathrm{d}a'} = p_{\text{data}}(a \mid h). \qquad \square$$

C.2 PROOF OF PROPOSITION 2 (LOCAL ENERGY DESCENT)

The result is a direct application of the descent lemma for L-smooth functions.

Setup. Fix a state representation $h = h_t^{(-1)}$ and define $f(a) \triangleq \mathcal{E}_w(h, a)$. By assumption, $\nabla_a f$ is L-Lipschitz continuous on a neighborhood $\mathcal{U}$ containing the update segment:

$$\|\nabla_a f(a) - \nabla_a f(a')\|_2 \le L\|a - a'\|_2 \quad \forall\, a, a' \in \mathcal{U}.$$

Descent lemma. For any L-smooth function f and any $a, a' \in \mathcal{U}$ whose connecting segment lies in $\mathcal{U}$,

$$f(a') \le f(a) + \langle \nabla_a f(a),\, a' - a \rangle + \frac{L}{2}\|a' - a\|_2^2.$$

Application. Set $a = a_{\mathrm{BC}}$ and $a' = a_{\mathrm{BC}} - \alpha g$, where $g = \nabla_a f(a_{\mathrm{BC}})$ and $\alpha \le 1/L$. Then $a' - a = -\alpha g$, and

$$\begin{aligned} f(a_{\mathrm{BC}} - \alpha g) &\le f(a_{\mathrm{BC}}) + \langle g,\, -\alpha g\rangle + \frac{L}{2}\|\alpha g\|_2^2 \\ &= f(a_{\mathrm{BC}}) - \alpha\|g\|_2^2 + \frac{L\alpha^2}{2}\|g\|_2^2 \\ &= f(a_{\mathrm{BC}}) - \alpha\left(1 - \frac{L\alpha}{2}\right)\|g\|_2^2. \end{aligned}$$

Since $\alpha \le 1/L$, we have $L\alpha \le 1$ and thus $1 - L\alpha/2 \ge 1/2$. The bound therefore reduces to

$$\mathcal{E}_w(h,\, a_{\mathrm{BC}} - \alpha g) \;\le\; \mathcal{E}_w(h,\, a_{\mathrm{BC}}) \;-\; \frac{\alpha}{2}\|g\|_2^2. \qquad \square$$

Remark 2. *The proposition applies to the unconstrained update $\tilde{a} = a_{\mathrm{BC}} - \alpha g$. The subsequent clipping to $[-1, 1]^D$ is a projection onto a convex set and may further alter the energy. The projected-update result below gives a displacement-based bound under local smoothness. The empirical energy curves (Section 4.5) show energy reduction at most, but not all, timesteps.*

C.3 Projected and Gradient-Clipped Correction

Let a lie in the convex action set $\mathcal{A} = [-1, 1]^D$, let $g = \nabla_a f(a)$, and let $c = \min(1, \gamma/(\|g\|_2 + \varepsilon)) > 0$. Define $\eta = \alpha c$ and $a^+ = \Pi_{\mathcal{A}}(a - \eta g)$. The projection optimality condition, with a as a feasible comparison point, gives

$$\langle g,\, a^+ - a\rangle \;\le\; -\frac{1}{\eta}\,\|a^+ - a\|_2^2.$$

If f is L-smooth along this update segment and $0 < \eta \le 1/L$, the descent lemma gives a displacement-based bound:

$$f(a^+) \;\le\; f(a) - \Big(\frac{1}{\eta} - \frac{L}{2}\Big)\|a^+ - a\|_2^2 \;\le\; f(a) - \frac{1}{2\eta}\,\|a^+ - a\|_2^2. \qquad \square$$

D Training Details

Our experiments were conducted by first fine-tuning the backbone models until convergence. In the subsequent phase, we froze the backbone and trained the energy head independently. Table 4 reports the complete configuration of the energy head and the action correction. Positive state–action pairs come from demonstrations, and negative actions come from other states in the mini-batch. The parameters reported in the table refer solely to VEL and remain separate from those used during backbone training. Dropout is applied only to the positional encodings (PE) of the LLM feature sequence. The matched-capacity direct correction head in Section 4.3 uses the same hidden width of 512. The real-robot study in Section 4.4 uses a correction step size of $\alpha = 0.01$.

Search baselines. The IBC-style baseline in Table 2 runs CEM over the trained energy around the BC action, with 128 samples per step, three iterations, an elite fraction of 0.1, an initial standard deviation of 0.05, a mean update anchored to the base action, and the lowest-energy elite sample returned. The unconstrained variant uses an initial standard deviation of 0.15 with an elite-mean update and collapsed within the first episodes.

E Additional Experimental Results

E.1 Energy During Real-Robot Execution

Fig. 6 reports the energy the critic assigns during execution on the screw-insertion task of Section 4.4, separated into the grasp and insertion stages and averaged over ten episodes per policy. Three regularities appear. The band around the corrected policy is narrower than the band around the base policy in both stages, so the correction reduces the spread of executed actions across episodes rather than shifting them uniformly. The corrected policy also runs at a lower mean energy throughout, the

Table 4: Energy-head training and action-correction parameters.

Hyperparameter	**Value**
Optimizer	AdamW
Base learning rate	5.0×10^{-4}
Learning-rate scheduler	Linear
Batch size (global)	64
Total training steps	200,000
Dropout rate	0.2 (PE)
Action correction step size α	0.1
Number of layers in the energy head	4
Number of attention heads in the energy head	8
Action dimension	7
Hidden embedding size	512
Gradient clipping norm	1.0
Mixed-precision training	False
Random seed	Same as OpenVLA-OFT (Kim et al., 2025a)

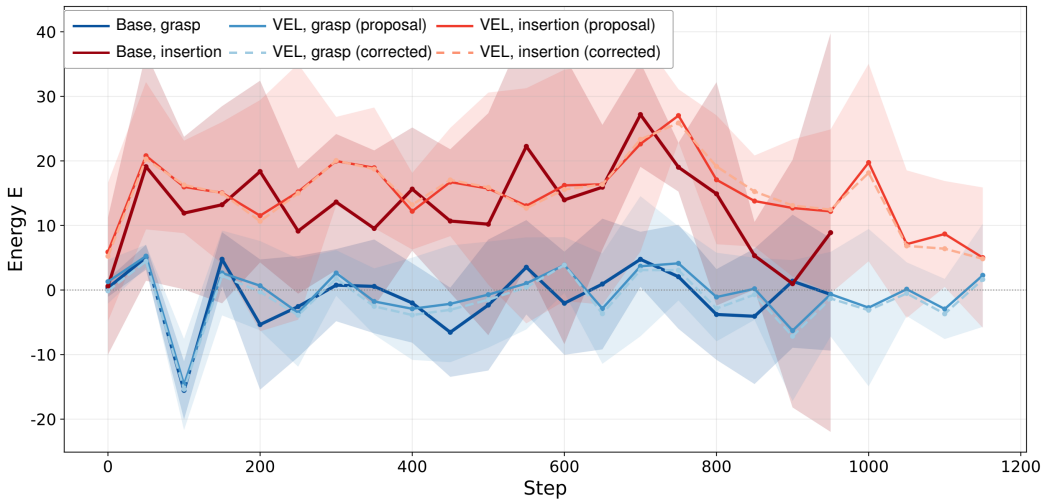


Figure 6: Energy during real-robot execution, for the grasp stage (blue) and the insertion stage (red), with and without VEL. Curves show the mean and shaded bands one standard deviation over ten episodes per policy. The base policy has no active head, so its executed action equals its proposal; the bands of the corrected VEL actions are omitted because they overlap those of the proposals. The corrected policy has a narrower band and a lower mean energy in both stages, and the reduction is largest in the insertion stage.

online signature of a critic that revises the proposal at every step rather than at isolated moments. The separation between the two stages is itself informative: grasp-stage energy stays near zero for both policies while insertion-stage energy is markedly higher and is where VEL lowers it most, so the critic concentrates its correction on the stage that Fig. 4c identifies as the one that fails.

E.2 Online Action Correction

Fig. 7 plots the energy of the uncorrected and corrected actions at every inference timestep. Consistent with the descent guarantee in Proposition 2, the corrected action achieves lower energy than the original proposal at most timesteps. The two curves remain close because α is small. At a subset of timesteps the corrected energy exceeds the original value, so empirical energy changes vary across decisions; the descent bound applies only when the update satisfies the local conditions of Proposition 2. The contribution of VEL is most visible at failure-prone moments. For example, the green region in Fig. 7 marks a phase in which the baseline policy drives the end-effector into the

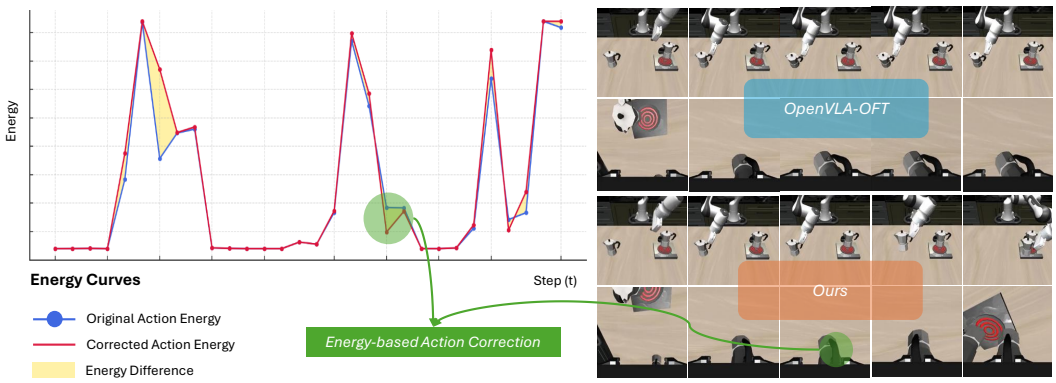


Figure 7: Illustration of real-time energy curves. Given the instruction "put both moka pots on the stove", the left plot compares the energy of the original and corrected actions of OpenVLA-OFT over an episode. The right side shows both observations during the task. The green region marks the action correction.

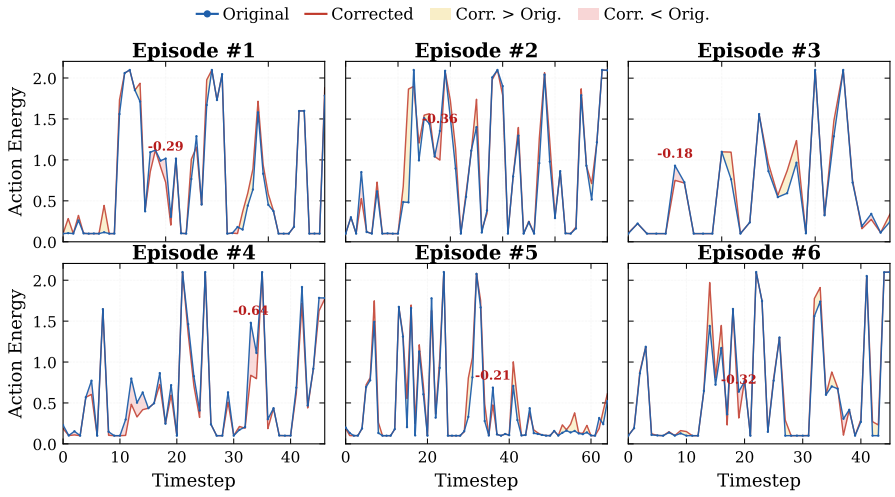


Figure 8: Additional real-time energy curves for the original and VEL-corrected actions.

moka-pot handle, and the energy correction redirects the motion and avoids the collision. Fig. 8 presents further real-time energy curves for the original and VEL-corrected actions.

E.3 Energy Ranking Diagnostics

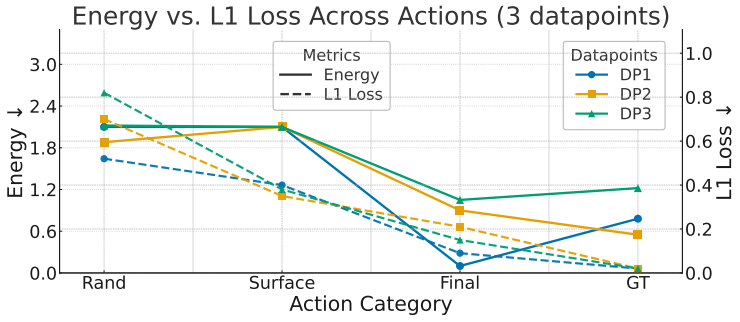


Figure 9: Energy value and ℓ_1 distance across three data points. Dotted lines indicate the ℓ_1 distance to the ground-truth action, and solid lines represent the energy values.

Fig. 9 analyzes energy values across several action types to test the ranking capability of the energy head. Actions span a spectrum of incorrectness: (i) *Rand.* denotes random actions; (ii) *Surface.* denotes actions read from intermediate layers of the backbone through early exit (Chuang et al., 2024; Teerapittayanon et al., 2016; Schuster et al., 2022); (iii) *Final.* denotes actions predicted by OpenVLA-OFT; and (iv) *GT.* denotes the ground-truth action. The ordering of energies follows the expected pattern. Random actions yield the highest energies. Surface actions contain partially suboptimal components and therefore produce high energies, although lower than those of random actions. OpenVLA-OFT predictions fall further down the energy scale, and ground-truth actions occupy the lowest energies. In addition, the energy values scale with the ℓ_1 error, which confirms that the critic supplies a meaningful signal for local action correction.

E.4 Energy and Task Outcome, Failure Modes, and Perturbation Robustness

Table 5: Energy as a predictor of task outcome on LIBERO-Long (Liu et al., 2023), over 100 episodes of an earlier energy-head checkpoint chosen so that enough failures occur for the comparison to be informative.

Quantity	**Successful episodes**	**Failed episodes**
Mean executed-action energy	-11.40 ± 2.08	-9.05 ± 2.70
Peak executed-action energy	4.06 ± 3.74	4.79 ± 3.93
Cumulative energy reduction by VEL	-21.06 ± 15.41	-66.93 ± 59.79

Energy and task outcome. The action-space analysis above links energy to distance from the expert action. Manipulation success, however, depends on contacts, object states, and long-horizon consequences that a per-step ℓ_1 distance does not express. Table 5 extends the analysis to task outcomes over 100 LIBERO-Long episodes. To obtain enough outcome variance for the comparison to be informative, these episodes come from an earlier energy-head checkpoint that fails more often than the final model. Episode-level mean energy separates successful from failed episodes with an AUROC of 0.749, and the point-biserial correlation between the cumulative energy reduction the critic applies and task success is 0.511 at $p = 5.48 \times 10^{-8}$. The energy therefore carries task-relevant signal beyond the action-level correlation. The association admits two readings, and the difference between them matters for the claim the method makes. Failed episodes receive more than three times the cumulative correction, which is consistent either with episodes already in difficulty eliciting sustained correction or with the corrections themselves disturbing execution. Two observations favor the first reading without settling it. Each correction is bounded by construction, through the step size and the gradient-norm clipping threshold, and enabling the correction raises success rates overall, which a systematically harmful intervention would not do. The paired perturbation analysis below provides a separate measure of benefit and harm on identical episodes.

Table 6: Failure modes of the corrected policy on LIBERO-Long (Liu et al., 2023), as a percentage of the 31 failed episodes of the checkpoint in Table 5.

Failure mode	**Share of failures (%)**
Early-stage misplacement that blocks a later subgoal	58.1 (18/31)
Timeout without an irrecoverable error	22.6 (7/31)
Grasp instability or dropped object	9.7 (3/31)
Other	9.7 (3/31)
Collision with scene geometry	0.0 (0/31)

Failure modes on long-horizon episodes. Table 6 categorizes every failed episode of the corrected policy on LIBERO-Long for the same earlier energy-head checkpoint, so that enough failures exist to classify. No failure in this sample was classified as a collision with scene geometry. This observation concerns 31 failures from an earlier corrected-policy checkpoint. Early misplacement dominates the remaining failures, followed by timeouts and grasp instability. The distribution is compatible with a local critic that cannot resolve all consequences of earlier poses. Task index 8 in Table 10 provides a qualitative example: the corrected grasp allows the mug to clear the microwave door edge. A matched baseline classification would be needed to quantify which failure mechanisms are reduced.

Table 7: Out-of-distribution evaluation on LIBERO-Spatial (Liu et al., 2023) (%), mean and standard deviation over five evaluation seeds. The base policy and the corrected policy are evaluated on identical episodes within each seed.

Method	Object perturbation	Camera perturbation
OpenVLA-OFT (Kim et al., 2025a)	70.6 ± 4.0	60.8 ± 7.6
OpenVLA-OFT + VEL	74.0 ± 3.9	60.4 ± 8.2
Difference	$+3.4$ $(p = 0.043)$	-0.4 $(p = 0.37)$

Robustness to scene and sensor perturbation. A critic that reads the hidden states of the base policy occupies an unusual position with respect to distribution shift, because it observes the world only through the representation whose reliability is in question. Table 7 separates two kinds of shift on LIBERO-Spatial at perturbation magnitude 0.05, with 100 paired episodes per seed over five evaluation seeds and 500 episodes per condition. Under object perturbation the arrangement of the scene changes while the perceptual grounding remains valid. Under camera perturbation the visual representation itself moves. Under object perturbation, correction improves success by 3.4 percentage points, with positive differences across all five seeds and a paired statistic of $t(4) = 2.92$. Camera perturbation gives a difference of -0.4 points, with mixed signs and $t(4) = -1.00$. VEL therefore improves success under the tested object perturbations, while camera perturbations remain a challenge for correction based on the backbone representation. The paired episodes provide a separate measure of intervention benefit and harm. Under object perturbation, VEL rescues 25 episodes that the base policy fails and loses 8 that the base policy solves, for a net gain of 17 over 500 episodes. The black-bowl task improves from 66.0 to 86.0 percent with ten rescues and no losses, and the top-drawer task gains four rescues without losses. Tasks with baseline success of at least 80% are unchanged in this evaluation. These results support a net benefit concentrated in difficult tasks, but VEL still disrupts some otherwise successful episodes. Under camera perturbation, 14 rescues and 16 losses indicate no net improvement.

E.5 Ablations

Table 8: VEL ablations on OpenVLA-OFT. Block (A): effect of gradient steps k (with $\alpha = 0.1$). Block (B): effect of step size α (with $k = 1$) and the vision-only critic input. Entries are success rates (%) from single evaluation runs of each configuration, so the $k = 1$, $\alpha = 0.1$ row is not the five-seed mean of Table 1.

(A) k ablation (baseline $\alpha = 0.1$)

k	Spatial	Object	Goal	Long
1	**97.8**	**98.6**	96.6	**96.6**
2	97.6	97.4	**97.8**	95.6
3	96.8	97.2	97.0	94.6
4	96.8	96.5	95.0	91.8

(B) α ablation (baseline $k = 1$) and vision-only setting

Setting	Spatial	Object	Goal	Long
$\alpha = 1.0$	96.2	97.0	80.4	79.2
$\alpha = 0.5$	97.0	**99.0**	92.0	91.4
$\alpha = 0.1$ (baseline)	**97.8**	98.6	**96.6**	**96.6**
Vision-only ($\alpha = 0.1$)	97.0	98.1	94.2	90.5

Table 8 ablates the two correction hyperparameters and the critic input. **Correction steps k.** Repeating the residual update degrades performance, most visibly on long-horizon tasks. The learned energy is a proxy calibrated near the demonstration manifold, so an iterated descent continues to reduce energy while it leaves the region in which that energy was supervised, and the action it returns need not be better than the one it started from. **Step size α.** Larger step sizes reduce success in this ablation. They

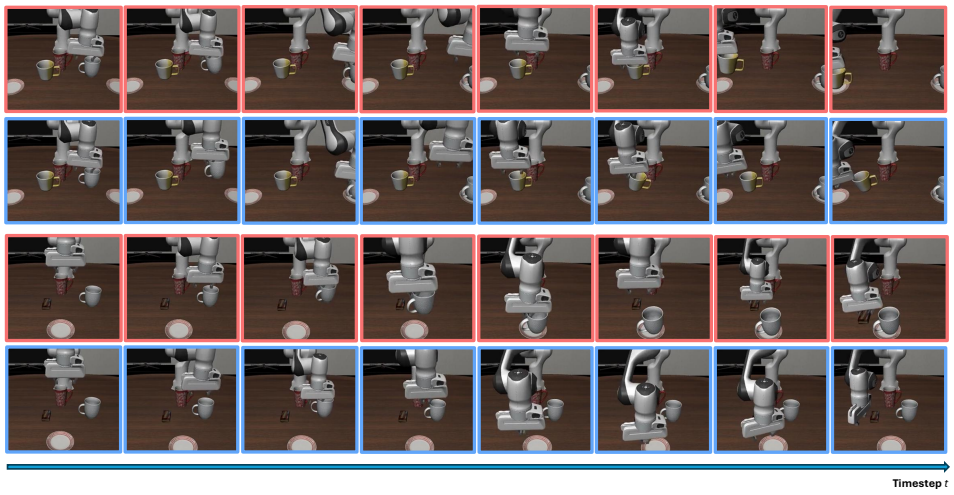


Figure 10: Trajectories produced by the baseline method with and without VEL correction. Baseline trajectories are shown in blue; VEL-corrected trajectories are shown in red.

can move a proposal beyond the locally useful region of the energy landscape or push the unprojected update outside $[-1, 1]^D$. Output clipping enforces these bounds but does not ensure that the resulting action is suitable for the task. The baseline $\alpha = 0.1$ provides the strongest overall performance among the tested step sizes. **Instruction tokens.** Training the energy head on vision tokens alone degrades performance: as a task-oriented critic, VEL must rank candidate actions conditioned on the task specification, and without instruction information the energy head cannot reliably identify the task goal.

E.6 LIBERO-Long Tasks and Per-Task Results

Table 9: Task descriptions for the ten LIBERO-Long subtasks.

Index	Description
0	put both the alphabet soup and the cream cheese box in the basket
1	put both the alphabet soup and the tomato sauce in the basket
2	put both the cream cheese box and the butter in the basket
3	put the white mug on the left plate and put the yellow and white mug on the right plate
4	put the white mug on the plate and put the chocolate pudding to the right of the plate
5	turn on the stove and put the moka pot on it
6	put the black bowl in the bottom drawer of the cabinet and close it
7	put both moka pots on the stove
8	put the yellow and white mug in the microwave and close it
9	pick up the book and place it in the back compartment of the caddy

Table 9 lists all ten LIBERO-Long subtasks and the indices used throughout the paper. These tasks require coordinated multi-step manipulation in which earlier object placements directly constrain later actions. Table 10 reports per-subtask success rates of VEL compared to the baseline. VEL produces the largest improvements on tasks whose final success is highly sensitive to earlier object poses, because it mitigates error accumulation so that objects placed early in a sequence remain in favorable poses. For example, task index 8 requires the agent to "put the yellow and white mug in the microwave and close it". VEL adjusts the end-effector pose during the initial cup grasp, which ensures that the mug can be placed safely into the microwave and prevents it from becoming caught on the door edge after a suboptimal grasp. Fig. 3 visualizes this behavior.

Table 10: LIBERO-Long success rates (%) per subtask under the single-seed protocol of the original evaluation. The five-seed suite-level evaluation is reported in Table 1.

Method	**index 9**	**index 8**	**index 7**	**index 6**	**index 5**
OpenVLA-OFT	88.0	84.0	98.0	88.0	98.0
OpenVLA-OFT + VEL	**88.0**	**96.0**	**98.0**	**92.0**	**100.0**
Method	**index 4**	**index 3**	**index 2**	**index 1**	**index 0**
OpenVLA-OFT	94.0	88.0	**100.0**	98.0	**98.0**
OpenVLA-OFT + VEL	**96.0**	**98.0**	98.0	**98.0**	96.0

E.7 Additional Qualitative Results

Additional qualitative results appear in Fig. 10. The top two rows depict the episode associated with instruction index 3 in Table 9, and the bottom two rows depict the episode associated with instruction index 4. In both episodes, VEL adjusts the original motion so that the end-effector gripper reaches the precise location required to grasp the cup, which ensures stable motion during the subsequent placement on the plate.