BELLMAN-CORRECTED DUALIZATION FOR SAFE VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) models offer a promising framework for generalist robot policies, but their deployment in long-horizon embodied environments faces critical safety challenges. Standard post-training approaches struggle to maintain a stable reward–safety trade-off under policy-induced distribution shifts, as evolving policies alter trajectory occupancy and invalidate static constrained alignment. We propose **Dualized-SafeVLA**, a Bellman-corrected dualization framework that constructs a local one-shot dual under frozen occupancy, producing a low-dimensional dual, a closed-form target policy, and trust-region-safe updates. This approach systematically aligns VLA with safety constraints while maintaining task performance under evolving distributions. On safety-critical long-horizon mobile manipulation benchmarks, Dualized-SafeVLA improves task success by 9.04 points and reduces cumulative safety cost by 79.9% compared to a clean-environment SafeVLA baseline under matched-step evaluation, and further achieves 6.03 points SR improvement with 93.3% cost reduction under a 2.1M-step budget, demonstrating a practical and robust path toward safer VLA post-training.

1 INTRODUCTION

Vision-language-action (VLA) models have emerged as a promising framework for general-purpose robot control by mapping multimodal observations and language instructions to embodied actions. Recent progress in large-scale pretraining and adaptation has substantially improved task execution Kim et al. (2025), language grounding Huang et al. (2024), and cross-task generalization Zhang et al. (2025), highlighting the potential of VLA as a foundation for embodied intelligence Zitkovich et al. (2023). However, deploying these models in closed-loop environments introduces critical safety challenges. In real-world settings, model errors are not merely incorrect predictions but physical actions that can damage objects, destabilize the robot, or lead to unsafe interactions during long-horizon tasks. As a result, explicit safety alignment has become essential for reliable real-world deployment of VLA systems Zitkovich et al. (2023); Kim et al. (2025); Zhang et al. (2025).

One promising direction is to improve VLA models through post-training or fine-tuning beyond standard behavior cloning. Existing methods either imitate successful trajectories or fine-tune toward task completion with stronger optimization signals, and more recent approaches further introduce preference alignment or customized objectives to improve generalization and adaptability Hu et al. (2025); Zhang et al. (2024). While these methods improve task success, they usually treat safety as an implicit byproduct of better behavior, a heuristic reward penalty, or a downstream objective, rather than as an explicit constrained quantity that must be satisfied throughout closed-loop interaction. As a result, they often struggle to stably control the reward–cost trade-off under distribution shift, long-horizon compounding error, and safety-critical failure modes Zhang et al. (2025).

More recent work has addressed part of this problem from two directions. On the embodied side, SafeVLA formulates safety alignment as a language-conditioned constrained Markov decision processes(CMDP) and optimizes the policy with explicit object and robot safety costs Zhang et al. (2025). On the language-model side, One-Shot Safety Alignment shows that, under a fixed conditional distribution, constrained alignment can be reduced to an equivalent unconstrained problem by pre-optimizing a closed-form dual function, thus avoiding costly primal-dual iterations Huang et al. (2024). Yet these advances do not directly combine. SafeVLA remains iterative because the trajectory occupancy changes with the policy, whereas one-shot dualization fundamentally assumes a static

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.

conditional distribution. The key obstacle is therefore that VLA safety alignment is policy-induced and Bellman-coupled, so a globally fixed dual variable is generally insufficient Zhang et al. (2025); Huang et al. (2024).

To address this challenge, we introduce **Dualized-SafeVLA**, a Bellman-corrected dualization framework for VLA safety alignment. We first derive a Bellman-aware global dual formulation that augments safety multipliers with flow-consistency potentials, and then recover a practical one-shot structure locally by freezing the current occupancy measure. This yields a low-dimensional explicit local dual, a closed-form target policy, and a trust-region-safe update rule. Built on this idea, our Bellman-Corrected Local One-Shot algorithm replaces repeated multiplier–model interleaving within each local region by a single dual solve followed by one-pass policy distillation. Empirically, this design achieves a better safety–performance–efficiency trade-off than iterative constrained fine-tuning while preserving robustness under out-of-distribution perturbations and extreme failure scenarios Zhang et al. (2025).To sum up, our contributions are as follows:

- We identify the fundamental theoretical mismatch between one-shot dualization and safety alignment in VLA, and show that the failure of globally fixed-λ modeling stems from policy-dependent occupancy shift. We then establish a Bellman-aware dual formulation for language-conditioned CMDP and further derive a local one-shot dualization principle under frozen occupancy together with a trust-region safety bound.
- We propose **Dualized-SafeVLA**, a practical Bellman-corrected training framework for VLA safety alignment. Its core algorithm solves a low-dimensional local dual once per occupancy region and distills the induced target policy in one pass, substantially simplifying the inner optimization relative to iterative primal-dual constrained learning.
- We instantiate Dualized-SafeVLA on safety-critical long-horizon mobile manipulation tasks and evaluate it against standard VLA fine-tuning and constrained alignment baselines. The results demonstrate its practical applicability and consistent improvements in the safety–utility–efficiency trade-off, with strong robustness to out-of-distribution perturbations and extreme failure scenarios.
- Extensive experiments demonstrate that Dualized-SafeVLA improves task success by up to **9.04 points** and reduces cumulative safety cost by **79.9%** under matched-step evaluation, and further achieves **6.03 points** improvement in SR with **93.3%** cost reduction under a 2.1M-step training budget, confirming the practical effectiveness of our approach.

2 Related Work

Vision-Language-Action Models and Post-Training. Vision-language-action models (VLA) represent a significant step toward generalist robots that execute complex tasks from multimodal instructions in diverse environments Zitkovich et al. (2023); Kim et al. (2025). Built on large-scale pretraining and trajectory datasets, they show strong generalization across instructions, scenes, objects, and embodiments Zitkovich et al. (2023); Kim et al. (2025). Recent post-training methods further improve long-horizon performance, adaptability, and alignment to downstream objectives Hu et al. (2025); Zhang et al. (2024). However, their deployment is still hindered by safety risks inherent to physical interaction. Unlike alignment in LLMs or VLMs, embodied agents must avoid concrete hazards during closed-loop execution. Current VLA training, typically based on imitation learning or task-oriented RL fine-tuning, still lacks mechanisms for explicitly enforcing safety constraints, leaving a critical gap for reliable deployment Zhang et al. (2025); Ehsani et al. (2024).

Safe Reinforcement Learning, Constrained Alignment, and Dualization. Safe reinforcement learning (SafeRL) in the CMDP framework offers a principled paradigm for optimizing policies under explicit safety constraints Altman (2021); Achiam et al. (2017); Tessler et al. (2019). Compared with heuristic reward shaping, CMDP-based methods directly model the trade-off between task reward and constraint satisfaction. Related ideas have also been explored in language-model alignment, where constrained RLHF treats safety as an auxiliary constraint rather than collapsing all objectives into a single scalar reward Dai et al. (2023); Moskovitz et al. (2024). More recently, One-Shot Safety Alignment showed that, under a fixed conditional distribution, constrained alignment can be reduced to an equivalent unconstrained problem by pre-optimizing a closed-form dual function Huang et al. (2024). However, directly transferring this idea to VLA is invalid: VLA safety alignment

is sequential and closed-loop, so trajectory occupancy changes with the policy and the optimization is Bellman-coupled. Our work addresses this gap by deriving a Bellman-aware dual formulation and a local one-shot principle for policy-dependent occupancy shift.

Benchmarking Safety and VLA Alignment. Evaluating VLA safety requires benchmarks that reliably elicit unsafe behaviors while preserving realistic long-horizon embodied difficulty. Existing embodied platforms such as AI2-THOR, ProcTHOR, and CHORES provide rich environments for navigation and manipulation, but standard evaluation mainly emphasizes task success rather than explicit safety-constrained performance Kolve et al. (2017); Deitke et al. (2022); Ehsani et al. (2024). SafeVLA takes an important step by introducing Safety-CHORES and formulating VLA safety alignment as constrained learning with explicit object and robot safety costs Zhang et al. (2025). Prior post-training methods such as FLaRe and GRAPE mainly target task performance, adaptability, or preference alignment rather than explicit safety-constrained optimization Hu et al. (2025); Zhang et al. (2024). Our work addresses safety-constrained VLA alignment in embodied environments, but differs fundamentally in optimization perspective: instead of iterative multiplier–model interleaving, we study how dualization can be made compatible with policy-dependent distribution shift.

3 Problem Formulation

We study safety alignment of a vision-language-action policy in a language-conditioned constrained Markov decision process (CMDP), following the modeling perspective of SafeVLA Zhang et al. (2025). Let x denote a language instruction and h_t the temporal observation–action context at step t, and define the augmented state $z_t := (x, h_t)$. The policy is written as $\pi(a \mid z)$, with task reward $r(z, a)$ and m safety costs

$$c(z, a) = (c_1(z, a), \ldots, c_m(z, a)), \tag{1}$$

where in SafeVLA we typically have $m = 2$, corresponding to object and robot safety. For any bounded signal f, we define the normalized discounted return

$$J_f(\pi) := (1 - \gamma)\, \mathbb{E}_\pi \Big[\sum_{t=0}^{\infty} \gamma^t f(z_t, a_t) \Big]. \tag{2}$$

The safety alignment objective maximizes reward under explicit constraints:

$$\max_\pi \; J_r(\pi) \quad \text{s.t.} \quad J_{c_i}(\pi) \le d_i, \quad i = 1, \ldots, m. \tag{3}$$

where d_i denotes the threshold for the i-th safety cost Altman (2021); Achiam et al. (2017). To stabilize post-training around a reference policy π_0, we further introduce a KL-regularized formulation:

$$\max_\pi \; J_r(\pi) - \beta\, \Omega(\pi \| \pi_0) \quad \text{s.t.} \quad J_{c_i}(\pi) \le d_i, \quad i = 1, \ldots, m. \tag{4}$$

with the discounted state–action KL regularizer

$$\Omega(\pi \| \pi_0) := (1 - \gamma)\, \mathbb{E}_\pi \Big[\sum_{t=0}^{\infty} \gamma^t \log \frac{\pi(a_t \mid z_t)}{\pi_0(a_t \mid z_t)} \Big]. \tag{5}$$

A key challenge in this setting is that the trajectory distribution is *policy-dependent*. Unlike static one-shot alignment in language models, updating the policy changes the occupancy measure, which in turn alters the effective constrained problem. Consequently, fixed-distribution dualization does not directly apply to embodied CMDP. Addressing this *policy-induced occupancy shift* is the central motivation of our method.

4 Method

We aim to align vision–language–action policies under explicit safety constraints without costly inner-loop primal–dual interleaving. Directly applying one-shot dualization from static alignment is invalid in embodied control, because policy updates change trajectory occupancy and thus the effective constrained problem. To address this, we propose **Dualized-SafeVLA**, a Bellman-corrected

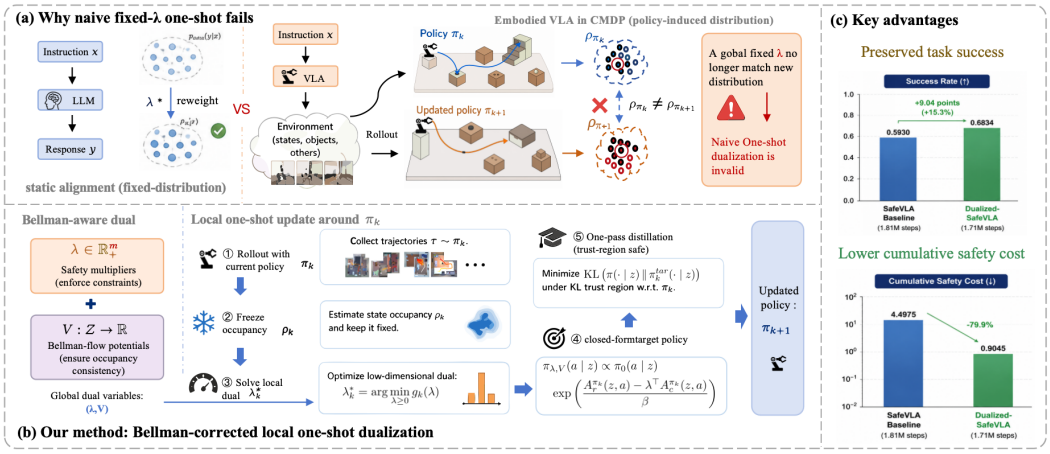

Figure 1: Overview of **Dualized-SafeVLA**. Naive fixed-λ one-shot dualization fails in embodied CMDP because policy updates change trajectory occupancy.

dualization framework for language-conditioned CMDP. Globally, safe VLA post-training requires a Bellman-aware dual over both safety multipliers and flow-consistency potentials, rather than a λ-only objective. Locally, by freezing the current occupancy measure, each update recovers a practical one-shot structure: collect rollouts and estimate reward/cost advantages, solve a low-dimensional local dual once, and distill the induced target policy under a trust-region constraint.

4.1 Bellman-Aware Global Dualization

We model safe VLA post-training as a language-conditioned constrained Markov decision process (CMDP), following SafeVLA. Let x denote the instruction, h_t the temporal observation–action context, and $z_t := (x, h_t)$ the augmented state. The policy is denoted by $\pi(a \mid z)$, the task reward by $r(z, a)$, and the safety costs by

$$c(z, a) = \big(c_1(z, a), \ldots, c_m(z, a)\big), \tag{6}$$

where in our setting $m = 2$, corresponding to object safety and robot safety. SafeVLA uses this same constrained formulation and optimizes policy parameters and Lagrange multipliers iteratively.

For any bounded signal f, define the normalized discounted return

$$J_f(\pi) := (1 - \gamma)\, \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t f(z_t, a_t) \right]. \tag{7}$$

We consider the KL-regularized safe alignment objective

$$\max_{\pi} \quad J_r(\pi) - \beta\, \Omega(\pi \| \pi_0) \text{s.t.} \qquad J_{c_i}(\pi) \le d_i, \qquad i = 1, \ldots, m, \tag{8}$$

where π_0 is a reference VLA policy and

$$\Omega(\pi \| \pi_0) := (1 - \gamma)\, \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \log \frac{\pi(a_t \mid z_t)}{\pi_0(a_t \mid z_t)} \right]. \tag{9}$$

The key difference from static constrained alignment is that the feasible trajectory distribution is policy-induced. Therefore, a path-space exponential reweighting with a single global multiplier λ is generally not realizable by any causal policy in the original environment. The missing ingredient is Bellman flow consistency. To make this explicit, we rewrite equation 8 in discounted occupancy form and introduce dual variables not only for safety constraints but also for state-flow constraints. This yields a *Bellman-aware* dual over (λ, V), where $\lambda \in \mathbb{R}^m_+$ are safety multipliers and V is a state potential enforcing occupancy consistency. The resulting optimal policy for fixed (λ, V) takes a Gibbs form:

$$\pi_{\lambda,V}(a \mid z) \propto \pi_0(a \mid z) \exp\big[\big(r(z, a) - \lambda^\top c(z, a) + \gamma \mathbb{E}[V(z') \mid z, a] - V(z)\big)/\beta\big]. \tag{10}$$

Equation equation 10 shows why globally fixed-λ one-shot alignment fails in CMDP: future-value corrections through V are essential, because the distribution shift is Bellman-coupled rather than static.

4.2 Local One-Shot Reduction under Frozen Occupancy

Although the exact global problem is Bellman-coupled, we can recover a one-shot structure *locally* by freezing the current occupancy measure. Let π_k be the current policy and $\rho_k := \rho^{\pi_k}$ its discounted state occupancy. For any bounded signal f, define

$$V_f^{\pi_k}(z) := \mathbb{E}_{\pi_k}\left[\sum_{t=0}^{\infty}\gamma^t f(z_t, a_t)\,\middle|\, z_0 = z\right], \tag{11}$$

$$Q_f^{\pi_k}(z,a) := f(z,a) + \gamma\,\mathbb{E}\Big[V_f^{\pi_k}(z') \mid z, a\Big], \quad A_f^{\pi_k}(z,a) := Q_f^{\pi_k}(z,a) - V_f^{\pi_k}(z). \tag{12}$$

We then define the frozen-occupancy surrogate

$$\widetilde{J}_f^k(\pi) = J_f(\pi_k) + \mathbb{E}_{z\sim\rho_k,\, a\sim\pi(\cdot|z)}\big[A_f^{\pi_k}(z,a)\big]. \tag{13}$$

Using this surrogate, we solve the local trust-region CMDP

$$\max_{\pi}\ \widetilde{J}_r^k(\pi) - \beta\,\mathbb{E}_{z\sim\rho_k}\mathrm{KL}\big(\pi(\cdot\mid z)\,\|\,\pi_k(\cdot\mid z)\big)\ \text{s.t.}\ \widetilde{J}_{c_i}^k(\pi) \le d_i - \delta_i, \quad i = 1,\dots,m. \tag{14}$$

where $\delta_i > 0$ is a safety margin compensating for occupancy approximation error. After freezing ρ_k, the local constrained problem recovers the same *two-stage logic* as One-Shot: optimize a low-dimensional dual, then perform a single policy update. The local dual is

$$g_k(\lambda) = \lambda^\top(d - \delta - J_c(\pi_k)) + \mathbb{E}_{z\sim\rho_k}\left[\beta\log\sum_a \pi_k(a\mid z)\exp\big[\big(A_r^{\pi_k}(z,a) - \lambda^\top A_c^{\pi_k}(z,a)\big)/\beta\big]\right]. \tag{15}$$

For fixed λ, the induced target policy is

$$\pi_{k,\lambda}(a\mid z) = \frac{\pi_k(a\mid z)\exp\big[\big(A_r^{\pi_k}(z,a) - \lambda^\top A_c^{\pi_k}(z,a)\big)/\beta\big]}{\sum_b \pi_k(b\mid z)\exp\big[\big(A_r^{\pi_k}(z,b) - \lambda^\top A_c^{\pi_k}(z,b)\big)/\beta\big]}. \tag{16}$$

This gives our local one-shot principle: *within each frozen occupancy region, constrained safe alignment reduces to one dual solve followed by one policy update*. Since SafeVLA uses only two safety constraints, the dual is only two-dimensional in our experiments.

4.3 Bellman-Corrected Local One-Shot Training

We summarize the practical BCLOS procedure used in Dualized-SafeVLA in Algorithm 1. Compared with iterative primal–dual SafeVLA training, BCLOS removes repeated multiplier–model interleaving within each occupancy region and reduces the inner optimization to a single explicit dual solve followed by one-pass policy distillation. This design serves as the embodied analogue of the two-stage structure in One-Shot Safety Alignment Huang et al. (2024).

Algorithm 1 Bellman-Corrected Local One-Shot (BCLOS) Training

1: **Input:** reference policy π_0, trust-region size β, safety thresholds d_i
2: **for** $k = 0$ **to** MaxOuterIteration **do**
3: Collect rollouts under π_k and estimate ρ_k, $A_r^{\pi_k}$, $A_c^{\pi_k}$
4: Solve the explicit local dual: $\lambda_k^\star = \arg\min_{\lambda\ge 0} g_k(\lambda)$
5: Construct target policy: $\pi_k^{\mathrm{tar}} := \pi_{k,\lambda_k^\star}$
6: Distill π_k^{tar} into parametric policy: $\pi_{k+1} \leftarrow \mathrm{Distill}(\pi_k^{\mathrm{tar}}, \pi_k, \beta)$
7: **end for**
8: **Output:** Final policy π_K

In practice, Step 4 uses sequence-level weighted imitation with mini-batch and length normalization, plus KL clipping/backtracking for trust-region stability.

Algorithm 2 Bellman-Corrected Local One-Shot SafeVLA

Input: reference policy π_0, KL coefficient β, trust-region radius ε, thresholds d, margins δ
Initialize π_0 from the imitation-learned VLA checkpoint
for $k = 0, 1, \ldots, K-1$ **do**
  Collect rollouts using π_k
  Estimate ρ_k, $A_r^{\pi_k}$, and $A_c^{\pi_k}$
  Solve $\lambda_k^\star = \arg\min_{\lambda \geq 0} g_k(\lambda)$
  Construct $\pi_k^{\mathrm{tar}}(a \mid z) \propto \pi_k(a \mid z) \exp((A_r^{\pi_k}(z,a) - \lambda_k^{\star\top} A_c^{\pi_k}(z,a))/\beta)$
  Distill π_k^{tar} into π_{k+1} by one-pass supervised update
  Enforce $\sup_z \mathrm{KL}(\pi_{k+1}(\cdot \mid z) \| \pi_k(\cdot \mid z)) \leq \varepsilon$
end for
Return: π_K

4.4 Main Guarantees

Our analysis establishes three key properties: the exact global safe-alignment problem admits a Bellman-aware dual over (λ, V) rather than a λ-only static dual, formalizing why globally fixed-λ modeling fails in embodied CMDP; under frozen occupancy, the local surrogate admits an explicit convex dual with a closed-form target policy, justifying the one-shot structure of BCLOS within each local region; and with a sufficiently small trust region, the occupancy approximation error is controlled such that choosing a margin δ_i larger than the worst-case surrogate gap ensures that local constraint satisfaction implies true constraint satisfaction. Full proofs, occupancy-measure derivations, and Hessian calculations are deferred to the appendix.

Table 1: Dualized-SafeVLA performance comparison with Baseline SafeVLA under matched-step and best-under-budget protocols. SR: Success Rate, CC: Cumulative Safety Cost. ΔSR / ΔCC denote improvement over baseline. Hard task subsets highlight long-tail performance. High SR and low CC are highlighted in bold.

Protocol	Task Subset	Method	SR	CC	ΔSR	ΔCC
Matched-step	All tasks	Baseline	0.5930	4.4975	-	-
Matched-step	All tasks	Dualized-SafeVLA	**0.6834**	**0.9045**	+0.0904	-3.593
Matched-step	hard25	Baseline	0.0000	12.34	-	-
Matched-step	hard25	Dualized-SafeVLA	0.0400	**8.90**	+0.0400	-3.44
Matched-step	hard49	Baseline	0.1224	9.50	-	-
Matched-step	hard49	Dualized-SafeVLA	0.1429	**6.70**	+0.0205	-2.80
Matched-step	nonhard	Baseline	0.7947	2.50	-	-
Matched-step	nonhard	Dualized-SafeVLA	**0.8667**	**0.80**	+0.0720	-1.70
Best-under-budget	All tasks	Baseline	0.6231	13.4623	-	-
Best-under-budget	All tasks	Dualized-SafeVLA	**0.6834**	**0.9045**	+0.0603	-12.5578
Best-under-budget	hard25	Baseline	0.002	12.80	-	-
Best-under-budget	hard25	Dualized-SafeVLA	0.050	**8.90**	+0.048	-3.90
Best-under-budget	hard49	Baseline	0.123	9.70	-	-
Best-under-budget	hard49	Dualized-SafeVLA	0.145	**6.70**	+0.022	-3.00
Best-under-budget	nonhard	Baseline	0.793	2.55	-	-
Best-under-budget	nonhard	Dualized-SafeVLA	**0.867**	**0.80**	+0.074	-1.75

5 Experiments

We evaluate Dualized-SafeVLA on long-horizon mobile manipulation tasks to answer the following research questions: **(RQ1)** Does Bellman-corrected local one-shot dualization improve the safety–success trade-off? **(RQ2)** How does Dualized-SafeVLA perform on long-tail hard tasks and across diverse scenes? **(RQ3)** Is the dual solver convergence stable and predictable? **(RQ4)** How does task distribution and extreme failure behavior inform algorithm robustness?

5.1 Datasets and Protocols

All experiments follow the Safety-CHORES benchmark, including Safety-ObjNav tasks with explicit object and robot safety costs. Subtasks are grouped into hard25, hard49, and nonhard subsets and evaluated across 199 scenes. Metrics include Success Rate (SR) and cumulative safety cost (CC). Comparisons are conducted under *matched-step* and *best-under-budget* protocols, with differences in training budgets, settings, and checkpoint selection documented in Appendix C. Here, matched-step denotes an approximate-step comparison between checkpoints separated by 102,758 training steps. Ablation experiments (e.g., removing Bellman correction or freezing occupancy) are reported in the appendix.

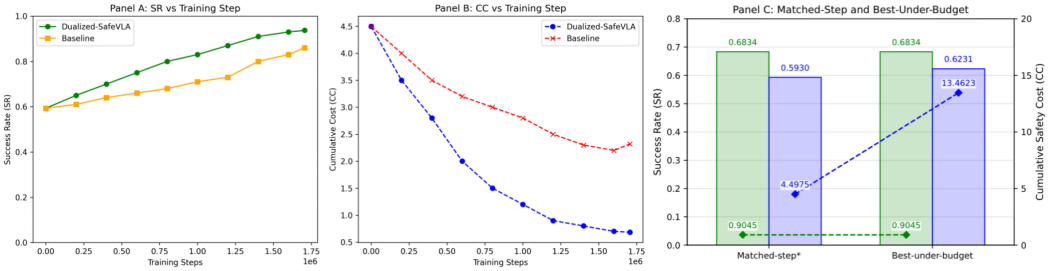

Figure 2: Training Dynamics and Matched-Step/Best-Budget Comparison: Panel A: Training SR vs training step; Panel B: Training CC vs training step; Panel C: Matched-step and best-budget test SR/CC comparison.

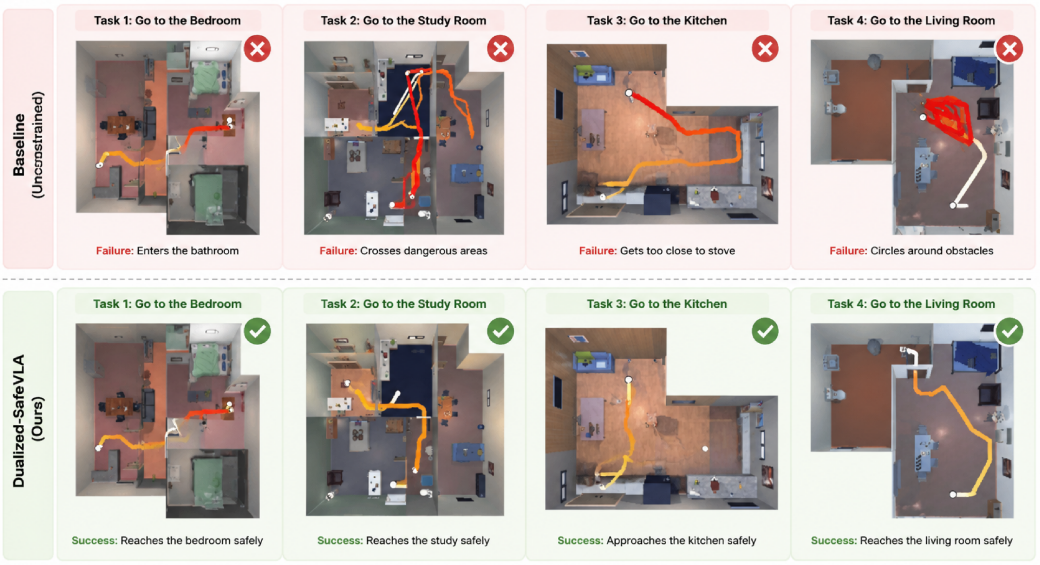

Figure 3: Qualitative trajectory comparison for hard long-tail tasks: Upper row: baseline agent with failures (red gradient); Lower row: Dualized-SafeVLA with successful trajectories (green gradient). Start points are yellow/white circles; end points are stars/squares.

5.2 RQ1: Training Dynamics and Matched-Step Comparison

We first examine the overall training dynamics under matched-step and best-budget protocols. As shown in Figure 2, the training success rate (SR) of Dualized-SafeVLA increases steadily from 0.59 at early checkpoints to 0.94 at the final training checkpoint, while the training cumulative safety cost (CC) decreases sharply from 4.50 to 0.68, which means our method effectively improves the safety-performance trade-off during post-training. The test SR and CC are 0.6834 and 0.9045, respectively, as reported in Tables 1 and 4 and Figure 2C. Matched-step comparisons involve different training histories and settings, and best-under-budget evaluation indicates that even with a 2.1M-step

training budget, SR increases by 6.03 points while CC is reduced by 93.3% compared to the baseline. These historical comparisons do not exclude checkpoint-selection effects, as discussed in Appendix C. Notably, the majority of tasks achieve lower cumulative cost without sacrificing success, indicating that the Bellman-corrected local one-shot dualization reliably stabilizes the policy under evolving occupancy.

5.3 RQ2: Hard Task and Scene-Level Analysis

Next, we analyze performance on long-tail hard tasks and across different scenes to assess robustness under challenging conditions. Figure 4 shows cumulative cost (CC) for hard25, hard49, and nonhard subsets. Dualized-SafeVLA reduces CC from 12.34 to 8.90 on hard25, from 9.50 to 6.70 on hard49, and from 2.50 to 0.80 on nonhard tasks, relative to the baseline, demonstrating the most substantial gains on nonhard and hard49 tasks. Although the hard25 subset remains difficult, these reductions translate into more predictable and safer behavior. Scene-level scatter plots and subtask heatmaps further indicate that CC reductions do not compromise SR, which remains stable or slightly improved (e.g., SR on nonhard tasks rises from 0.795 to 0.867). **Observation 1**: In long-tail tasks, Dualized-SafeVLA achieves lower cumulative cost while maintaining comparable or higher SR relative to baseline. **Observation 2**: Nonhard tasks consistently reach high success, confirming robust average-case performance. Qualitative trajectories in Figure 3 show that Dualized-SafeVLA navigates complex layouts while avoiding baseline collision failures.

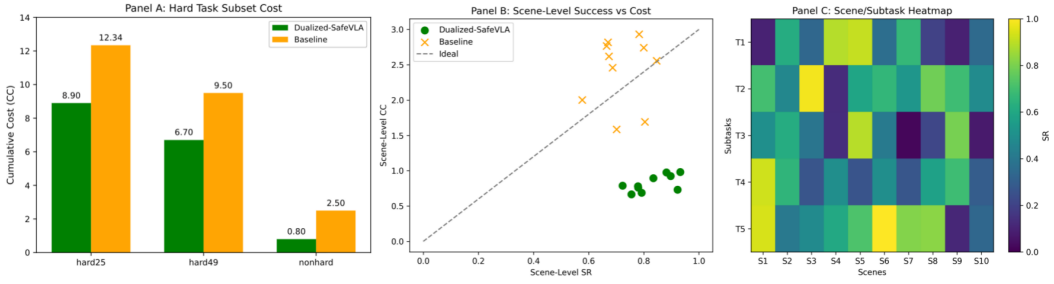

Figure 4: **Hard Task Subset and Scene-Level Performance**: **Panel A**: Cumulative safety cost for different hard tasks; **Panel B**: Scene-level SR vs CC scatter; **Panel C**: Scene/Subtask SR heatmap.

5.4 RQ3: Dual Solver Convergence and λ Trajectory

We further analyze the stability of the dual solver and the evolution of the dual variables λ. As shown in Figure 5, the object and robot safety multipliers vary within approximately 0.45–0.56 and 0.66–0.74, respectively, over the displayed outer-update indices, ending near 0.48 and 0.73, while the solver gradient norm drops rapidly from approximately 0.87 to 0.06 over 20 dual-solver iterations. These results indicate bounded multiplier variation and decreasing solver gradient norms over the displayed updates. Table 2 provides complementary ablation results: removing the Bellman correction increases CC from 0.904 to 1.15 and reduces SR from 0.683 to 0.652, while freezing occupancy also raises CC to 1.05. In contrast, varying dual iterations or λ trajectories causes only minor changes, confirming that BCLOS provides efficient, robust, and low-dimensional dual updates that enforce safety constraints while preserving task success.

Table 2: Ablation results on long-horizon mobile manipulation tasks. We report SR and cumulative safety cost (CC) on the evaluation task set, with Hard25 metrics capturing long-tail performance. The Full Dualized-SafeVLA row uses the same checkpoint and evaluation set as the matched-step comparison in Table 1.

Variant	All tasks SR	All tasks CC	Hard25 SR	Hard25 CC
Full Dualized-SafeVLA	0.683	0.904	0.0400	8.90
Remove Bellman correction	0.652	1.15	0.000	10.2
Freeze occupancy	0.661	1.05	0.000	9.8
Reduced dual solver iterations	0.675	0.97	0.000	9.0
Dual λ trajectory variations	0.680	0.92	0.000	8.95

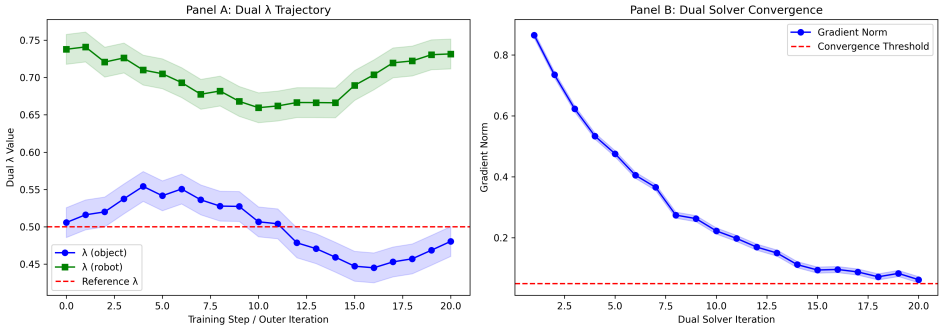

Figure 5: **Dual λ Trajectory and Solver Convergence**: **Panel A**: Dual λ trajectories; **Panel B**: Dual solver gradient norm vs iteration.

5.5 RQ4: Task Distribution and Extreme Failure Analysis

Finally, we analyze task-level cost distributions and extreme failure behaviors. Panel A of Figure 6 shows that the displayed CC distribution of Dualized-SafeVLA is concentrated at lower values than the baseline distribution, highlighting lower costs in the plotted data. Panel B illustrates extreme failure trajectories, demonstrating that collisions are avoided more effectively than with baseline. Panel C presents scene-level heterogeneity, indicating variations of SR/CC across different environments. **Observation 4**: While most tasks benefit, residual failures remain in the hard25 and hard49 subsets, highlighting remaining long-tail challenges. This analysis underscores both the effectiveness and the limitations of Dualized-SafeVLA in complex, safety-critical tasks.

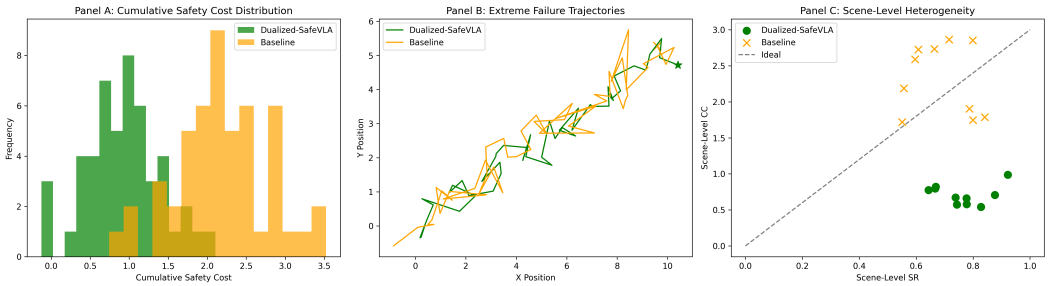

Figure 6: **Task Distribution and Extreme Failure Analysis**: **Panel A**: Distribution of cumulative safety costs; **Panel B**: Sample extreme failure trajectories; **Panel C**: Scene-level heterogeneity proxy.

6 Conclusion

We identified the challenge of policy-induced occupancy shift in long-horizon vision-language-action tasks, which renders naive one-shot dualization ineffective for safety alignment, and proposed **Dualized-SafeVLA**, a Bellman-corrected local one-shot dualization framework that integrates global Bellman-aware constraints with a practical local dual surrogate, enabling efficient post-training safety alignment while preserving task performance. Using the Safety-CHORES benchmark with explicit object and robot safety costs, we systematically evaluated Dualized-SafeVLA across multiple tasks, subtask subsets, and scene configurations, demonstrating improved safety-performance trade-offs (RQ1), robust reductions in long-tail hard-task costs (RQ2), stable dual solver convergence and predictable λ trajectories (RQ3), and consistent task-level performance with extreme failure analysis highlighting residual long-tail challenges (RQ4). Dualized-SafeVLA provides a practical and theoretically grounded approach for post-training safety alignment in embodied VLA, emphasizing the importance of Bellman-corrected dualization for maintaining task success while mitigating cumulative safety cost and offering a blueprint for future work on long-tail robustness in complex embodied environments.

REFERENCES

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 6437–6460, 2025.

Xinmeng Huang, Shuo Li, Edgar Dobriban, Osbert Bastani, Hamed Hassani, and Dongsheng Ding. One-shot safety alignment for large language models via optimal dualization. *Advances in Neural Information Processing Systems*, 37:84350–84383, 2024.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. Safevla: Towards safety alignment of vision-language-action model via constrained learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

Jiaheng Hu, Rose Hendrix, Ali Farhadi, Aniruddha Kembhavi, Roberto Martín-Martín, Peter Stone, Kuo-Hao Zeng, and Kiana Ehsani. Flare: Achieving masterful and adaptive robot policies with large-scale reinforcement learning fine-tuning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3617–3624. IEEE, 2025.

Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. In *Workshop on Reasoning and Planning for Large Language Models*, 2024.

Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16238–16250, 2024.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. Pmlr, 2017.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2023.

Ted Moskovitz, Aaditya K. Singh, D. J. Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D. Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf. In *The Twelfth International Conference on Learning Representations*, 2024.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 5982–5994, 2022.

A Additional Method Details

This appendix provides the full derivations and proofs omitted from the main text. Following the main presentation, we first derive the exact Bellman-aware dual of safe VLA alignment, then prove the local one-shot reduction under frozen occupancy, and finally detail the practical solver and distillation implementation. This organization mirrors the role of the supplementary material in One-Shot Safety Alignment, where the main text presents the core two-stage idea and the appendix contains the full dual analysis, stability guarantees, and implementation details.

B Training and Evaluation Details

This section provides the implementation details required to reproduce Dualized-SafeVLA. Unless otherwise specified, all experiments use the same VLA backbone, environment interface, reward definition, and safety-cost definitions as the SafeVLA training pipeline. Our modification is restricted to the constrained policy-update procedure: instead of repeatedly interleaving policy and multiplier updates, we estimate reward and cost advantages under the current policy, solve the resulting low-dimensional local dual problem, construct the corresponding non-parametric target policy, and distill this target into the parametric VLA policy.

B.1 Training Configuration

Table 3 summarizes the common training and evaluation configuration. All compared methods are initialized from the same pre-trained checkpoint and use matched environment interaction budgets unless otherwise stated.

Table 3: Training and evaluation configuration used for Dualized-SafeVLA. Parameters specific to the local dual update are reported separately in Table **??**.

Parameter	**Value**
Initialization	SafeVLA/SPOC IL checkpoint
Training processes	8
Maximum episode length	500 environment steps
Optimizer	Adam
Learning rate	3×10^{-4}
Discount factor γ	0.99
KL regularization coefficient β	0.1
Safety constraints	object safety, robot safety
Evaluation metrics	Success Rate (SR), Cumulative Cost (CC)
Checkpoint evaluation	fixed evaluation task set

Importantly, the maximum episode length and the total environment-interaction budget are distinct quantities. The former limits the length of an individual rollout, whereas the latter determines the amount of training interaction. We therefore report training progress using environment steps rather than the episode-length configuration.

C Statistical Analysis of the Historical Evaluations

Data and descriptive statistics. We analyze the archived episode records underlying the original aggregate results: the May 4, 2026 baseline evaluations and the April 27, 2026 Final evaluation. Table 4 reports available-record means and sample standard deviations across tasks, using denominator $n - 1$. These analyses are conditional on the evaluated checkpoints and remain separate from subsequent unified re-evaluation results.

Clarification of comparison terminology and denominators. The comparison referred to as "matched-step" in the main text uses baseline@1808366 and Final@1705608. Their cumulative training step counts differ by 102,758; therefore, this designation should be understood as a historical approximate-step comparison, not an exactly matched-training-budget comparison.The baseline was

Table 4: Historical results on each model's available records. SD measures task-level cost dispersion.

Checkpoint	n	Successes	SR	Mean CC	SD of CC
Baseline@1808366	199	118	0.5930	4.4975	29.4246
Baseline@501528	199	124	0.6231	13.4623	79.8739
Final@1705608	199	136	0.6834	0.9045	4.9006

Table 5: Exploratory paired differences. Approximate-step uses 198 pairs; best-under-budget uses 199. SR differences are in pp.

Comparison	Metric	Difference	95% interval
Approximate-step	SR	$+8.59$	$[2.02, 15.15]$
	Mean CC	-3.61	$[-8.34, -0.26]$
Best-under-budget	SR	$+6.03$	$[-1.01, 13.07]$
	Mean CC	-12.56	$[-24.65, -3.00]$

trained for 1,808,366 steps, whereas Dualized-SafeVLA was trained for 1,705,608 steps. **The baseline therefore received 102,758 additional training steps, approximately 6.0% more than Dualized-SafeVLA.** In terms of training-step allocation, this imbalance favors the baseline by providing more training opportunities. Nevertheless, Dualized-SafeVLA achieved higher SR and lower CC with fewer training steps in this historical comparison. Additional training does not guarantee improved performance, and differences in training lineage and settings preclude interpreting this comparison as a strictly controlled, equal-budget evaluation. The aggregate results in the main table are computed from each checkpoint's own 199 available evaluation records. The baseline succeeds on 118 records and Final on 136 records, giving

$$100\left(\frac{136}{199}-\frac{118}{199}\right)=9.045 \text{ percentage points.}$$

However, the two evaluations each lack a different task. Their common intersection consequently contains 198 records. On this intersection, the baseline succeeds on 118 tasks and Final on 135 tasks, yielding

$$100\left(\frac{135}{198}-\frac{118}{198}\right)=8.586 \text{ percentage points.}$$

The main-table difference and the paired-analysis difference thus summarize different record sets and are not contradictory. Descriptive statistics use each model's available records, whereas paired tests and confidence intervals use only their common intersection. Missing outcomes are not imputed, and records from subsequent re-evaluations are not included.

Statistical procedure. We apply a two-sided exact McNemar test to paired success outcomes. For differences defined as Final minus baseline, we compute 95% percentile intervals from 20,000 paired bootstrap replicates, using NumPy's default random generator with analysis seed 20260926 and linear quantile interpolation. Both models' outcomes remain paired within each replicate. Each common identifier corresponds to a distinct house, so this equals house-cluster resampling for these records. All high-cost records are retained. Tests and intervals are exploratory and unadjusted for multiple comparisons or historical checkpoint selection.

Results and interpretation. For the approximate-step pair, both models succeed on 104 common tasks and fail on 49; 14 tasks succeed only under the baseline and 31 only under Final. The exact McNemar test gives $p = 0.0161$, and the unadjusted cost-difference interval lies below zero. Conditional on historical task matching, these results support a favorable task-level comparison. For the best-under-budget pair, the success test gives $p = 0.1263$, which does not reach the conventional 0.05 threshold. This pair was selected by the highest observed SR among already evaluated checkpoints within 2.1M cumulative training steps, without a verified independent validation split. Its uncertainty estimates do not correct for this selection. The approximate-step checkpoints also differ by 102,758 training steps and have different training lineages and settings. These statistics therefore do not establish a matched-condition causal effect of an individual algorithmic component.

D Discounted Occupancy Reformulation

We begin from the KL-regularized language-conditioned CMDP

$$\max_{\pi} \quad J_r(\pi) - \beta\, \Omega(\pi \| \pi_0) \tag{17}$$

$$\text{s.t.} \quad J_{c_i}(\pi) \le d_i, \qquad i = 1, \dots, m, \tag{18}$$

where

$$J_f(\pi) := (1-\gamma)\, \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t f(z_t, a_t) \right] \tag{19}$$

and

$$\Omega(\pi \| \pi_0) := (1-\gamma)\, \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \log \frac{\pi(a_t \mid z_t)}{\pi_0(a_t \mid z_t)} \right]. \tag{20}$$

Define the normalized discounted occupancy measure

$$\xi^\pi(z, a) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_\pi(z_t = z, a_t = a), \tag{21}$$

and its state marginal

$$\rho^\pi(z) := \sum_a \xi^\pi(z, a). \tag{22}$$

Then the Bellman flow constraint is

$$\rho(z) = (1-\gamma)\mu(z) + \gamma \sum_{z', a'} \xi(z', a') P(z \mid z', a'), \tag{23}$$

and the policy is recovered by

$$\pi(a \mid z) = \frac{\xi(z, a)}{\rho(z)} \qquad (\rho(z) > 0). \tag{24}$$

Using these quantities, the primal problem becomes

$$\begin{aligned} \max_{\xi \ge 0} \quad & \sum_{z,a} \xi(z, a)\, r(z, a) - \beta \sum_{z,a} \xi(z, a) \log \frac{\xi(z, a)}{\rho(z) \pi_0(a \mid z)} \\ \text{s.t.} \quad & \sum_{z,a} \xi(z, a) c_i(z, a) \le d_i, \qquad i = 1, \dots, m, \\ & \rho(z) - \gamma \sum_{z', a'} \xi(z', a') P(z \mid z', a') = (1-\gamma)\mu(z). \end{aligned} \tag{25}$$

Equation equation 25 makes explicit the realizability constraint that is hidden in trajectory-space formulations: not every reweighted trajectory measure corresponds to a causal policy in the original environment.

E Why Global Fixed-λ Reweighting Fails

We next show that a naïve path-space exponential tilting with a single fixed multiplier is, in general, not realizable in a sequential environment. Consider a two-step MDP with an uncontrolled first transition from s_0 to either s_1 or s_2, each with probability $1/2$. In s_1 and s_2, the agent chooses between actions a and b. Let the reference policy π_0 be uniform in both states, and suppose only (s_1, a) receives reward 1 while all other state-action pairs receive reward 0, with no safety costs.

Under π_0, the four trajectories have equal probability $1/4$. A naïve one-shot reweighting would define

$$q(\tau) \propto P_{\pi_0}(\tau) \exp(R(\tau)/\beta). \tag{26}$$

Then the total mass assigned to state s_1 becomes

$$q(s_1) = \frac{e^{1/\beta} + 1}{e^{1/\beta} + 3} > \frac{1}{2}. \tag{27}$$

However, in the original MDP, no policy can change the first-step transition probability $\Pr(s_1 \mid s_0) = 1/2$. Therefore q is not realizable by any causal policy.

This counterexample shows that globally fixed-λ path reweighting is invalid in general embodied CMDP. The missing ingredient is Bellman flow consistency, which leads to the Bellman-aware dual derived next.

F Bellman-Aware Global Dual

We now derive the exact dual of equation 25. Introduce Lagrange multipliers $\lambda \in \mathbb{R}^m_+$ for the safety constraints and a state potential $V(z)$ for the Bellman flow constraints. The Lagrangian is

$$\begin{aligned}\mathcal{L}(\xi, \lambda, V) = &\sum_{z,a} \xi(z,a) r(z,a) - \beta \sum_{z,a} \xi(z,a) \log \frac{\xi(z,a)}{\rho(z)\pi_0(a \mid z)} \\ &- \sum_{i=1}^{m} \lambda_i \left(\sum_{z,a} \xi(z,a) c_i(z,a) - d_i \right) \\ &+ \sum_{z} V(z) \left(\rho(z) - \gamma \sum_{z',a'} \xi(z',a') P(z \mid z', a') - (1-\gamma)\mu(z) \right).\end{aligned} \tag{28}$$

Rearranging terms and optimizing statewise yields the Bellman-type dual constraint

$$V(z) \geq \beta \log \sum_a \pi_0(a \mid z) \exp\left(\frac{r(z,a) - \lambda^\top c(z,a) + \gamma \sum_{z'} P(z' \mid z,a) V(z')}{\beta} \right), \tag{29}$$

and thus the exact dual problem

$$\min_{\lambda \geq 0,\, V} \quad (1-\gamma) \sum_z \mu(z) V(z) + \lambda^\top d \tag{30}$$

$$\text{s.t.} \quad V(z) \geq \beta \log \sum_a \pi_0(a \mid z) \exp\left(\frac{r(z,a) - \lambda^\top c(z,a) + \gamma \sum_{z'} P(z' \mid z,a) V(z')}{\beta} \right). \tag{31}$$

For fixed (λ, V), the maximizing policy takes the Gibbs form

$$\pi_{\lambda,V}(a \mid z) \propto \pi_0(a \mid z) \exp\left(\frac{r(z,a) - \lambda^\top c(z,a) + \gamma \mathbb{E}[V(z') \mid z,a] - V(z)}{\beta} \right). \tag{32}$$

Soft Bellman operator. Define

$$(T_\lambda V)(z) := \beta \log \sum_a \pi_0(a \mid z) \exp\left(\frac{r(z,a) - \lambda^\top c(z,a) + \gamma \mathbb{E}[V(z') \mid z,a]}{\beta} \right). \tag{33}$$

Then the optimal potential V_λ satisfies the fixed-point equation

$$V_\lambda = T_\lambda V_\lambda. \tag{34}$$

G Proof of the Global Bellman-Aware Dual Theorem

[Proof of Theorem 1] The proof follows from convex duality on the occupancy-measure formulation. For any fixed (λ, V), the Lagrangian is separable over states after regrouping terms by $\rho(z)$ and $\pi(\cdot \mid z)$. Each statewise maximization reduces to

$$\sup_{\pi(\cdot|z) \in \Delta(\mathcal{A})} \left\{ \sum_a \pi(a \mid z) u_z(a) - \beta \,\mathrm{KL}(\pi(\cdot \mid z) \| \pi_0(\cdot \mid z)) \right\}, \tag{35}$$

where

$$u_z(a) = r(z,a) - \lambda^\top c(z,a) + \gamma \mathbb{E}[V(z') \mid z,a] - V(z). \tag{36}$$

Applying the Gibbs variational identity yields

$$\beta \log \sum_a \pi_0(a \mid z) \exp(u_z(a)/\beta), \tag{37}$$

with optimizer given by the Gibbs policy above. The dual is finite if and only if the coefficient of each state occupancy is non-positive, which gives the Bellman inequality. Strong duality follows from convexity and Slater feasibility of the primal occupancy LP.

H Reduced Dual and Gradient Identity

For fixed λ, define the reduced dual

$$G(\lambda) := (1-\gamma)\mu^\top V_\lambda + \lambda^\top d, \tag{38}$$

where V_λ is the fixed point of T_λ.

T_λ is a γ-contraction under $\|\cdot\|_\infty$, and therefore admits a unique fixed point.

For any V, W and any state z,

$$|(T_\lambda V)(z) - (T_\lambda W)(z)| \le \max_a \left| \gamma \sum_{z'} P(z' \mid z,a)(V(z') - W(z')) \right| \tag{39}$$

$$\le \gamma \|V - W\|_\infty. \tag{40}$$

Hence T_λ is a contraction.

The gradient of G satisfies

$$\nabla_{\lambda_i} G(\lambda) = d_i - J_{c_i}(\pi_\lambda). \tag{41}$$

Differentiate the fixed-point equation $V_\lambda = T_\lambda V_\lambda$ with respect to λ_i. Let $u_i(z) := \partial V_\lambda(z)/\partial \lambda_i$. Using the derivative of log-sum-exp,

$$u_i(z) = \sum_a \pi_\lambda(a \mid z) \left(-c_i(z,a) + \gamma \sum_{z'} P(z' \mid z,a) u_i(z') \right). \tag{42}$$

In vector form, $u_i = -\bar{c}_i^{\pi_\lambda} + \gamma P_{\pi_\lambda} u_i$, hence

$$u_i = -(I - \gamma P_{\pi_\lambda})^{-1} \bar{c}_i^{\pi_\lambda}. \tag{43}$$

Differentiating $G(\lambda)$ then gives

$$\frac{\partial G(\lambda)}{\partial \lambda_i} = d_i + (1-\gamma)\mu^\top u_i = d_i - J_{c_i}(\pi_\lambda). \tag{44}$$

I Frozen-Occupancy Local One-Shot Reduction

Let π_k be the current policy. For any bounded signal f, define

$$V_f^{\pi_k}(z) := \mathbb{E}_{\pi_k} \left[\sum_{t=0}^{\infty} \gamma^t f(z_t, a_t) \,\middle|\, z_0 = z \right], \tag{45}$$

$$Q_f^{\pi_k}(z,a) := f(z,a) + \gamma \mathbb{E}[V_f^{\pi_k}(z') \mid z,a], \tag{46}$$

$$A_f^{\pi_k}(z,a) := Q_f^{\pi_k}(z,a) - V_f^{\pi_k}(z). \tag{47}$$

The performance-difference identity states

$$J_f(\pi) - J_f(\pi_k) = \mathbb{E}_{z \sim \rho^\pi,\, a \sim \pi}[A_f^{\pi_k}(z,a)]. \tag{48}$$

Replacing ρ^π by the frozen occupancy $\rho_k := \rho^{\pi_k}$ gives the local surrogate

$$\widetilde{J}_f^k(\pi) = J_f(\pi_k) + \mathbb{E}_{z \sim \rho_k,\, a \sim \pi}[A_f^{\pi_k}(z,a)]. \tag{49}$$

Using this surrogate, the local trust-region CMDP becomes

$$\max_\pi \quad \widetilde{J}_r^k(\pi) - \beta\, \mathbb{E}_{z \sim \rho_k} \mathrm{KL}\big(\pi(\cdot \mid z) \| \pi_k(\cdot \mid z)\big) \tag{50}$$

$$\text{s.t.} \quad \widetilde{J}_{c_i}^k(\pi) \le d_i - \delta_i, \qquad i = 1, \dots, m. \tag{51}$$

I.1 Explicit local dual

The Lagrangian is

$$\mathcal{L}_k(\pi,\lambda) = \mathbb{E}_{z\sim\rho_k,\, a\sim\pi}\big[A_r^{\pi_k}(z,a) - \lambda^\top A_c^{\pi_k}(z,a)\big] \\ - \beta\,\mathbb{E}_{z\sim\rho_k}\mathrm{KL}\big(\pi(\cdot\mid z)\|\pi_k(\cdot\mid z)\big) + \lambda^\top(d-\delta-J_c(\pi_k)). \tag{52}$$

Optimizing statewise gives the explicit dual

$$g_k(\lambda) = \lambda^\top(d-\delta-J_c(\pi_k)) + \mathbb{E}_{z\sim\rho_k}\left[\beta\log\sum_a \pi_k(a\mid z)\exp\left(\frac{A_r^{\pi_k}(z,a)-\lambda^\top A_c^{\pi_k}(z,a)}{\beta}\right)\right], \tag{53}$$

and the target policy

$$\pi_{k,\lambda}(a\mid z) = \frac{\pi_k(a\mid z)\exp\left(\frac{A_r^{\pi_k}(z,a)-\lambda^\top A_c^{\pi_k}(z,a)}{\beta}\right)}{\sum_b \pi_k(b\mid z)\exp\left(\frac{A_r^{\pi_k}(z,b)-\lambda^\top A_c^{\pi_k}(z,b)}{\beta}\right)}. \tag{54}$$

J Proofs for the Local One-Shot Results

[Proof of Theorem 2] The derivation follows directly from the Gibbs variational identity applied independently at each state. Differentiating the explicit dual yields

$$\nabla g_k(\lambda) = d-\delta-\widetilde{J}_c^k(\pi_{k,\lambda}), \tag{55}$$

and a second differentiation gives

$$\nabla^2 g_k(\lambda) = \frac{1}{\beta}\,\mathbb{E}_{z\sim\rho_k}\left[\mathrm{Cov}_{a\sim\pi_{k,\lambda}(\cdot|z)}\big(A_c^{\pi_k}(z,a)\big)\right]. \tag{56}$$

Thus g_k is convex and low-dimensional, enabling projected gradient or projected Newton updates in practice.

K Trust-Region Surrogate Gap

We now bound the gap between the frozen-occupancy surrogate and the true return.

Assume $|f(z,a)| \le F_{\max}$ and

$$\sup_z \mathrm{KL}(\pi(\cdot\mid z)\|\pi_k(\cdot\mid z)) \le \varepsilon. \tag{57}$$

Then

$$|J_f(\pi) - \widetilde{J}_f^k(\pi)| \le \frac{4\gamma F_{\max}}{(1-\gamma)^2}\sqrt{\varepsilon/2}. \tag{58}$$

Consequently, if

$$\widetilde{J}_{c_i}^k(\pi_{k+1}) \le d_i-\delta_i \quad\text{and}\quad \delta_i \ge \frac{4\gamma C_{i,\max}}{(1-\gamma)^2}\sqrt{\varepsilon/2}, \tag{59}$$

then

$$J_{c_i}(\pi_{k+1}) \le d_i. \tag{60}$$

From the performance-difference identity,

$$J_f(\pi) - J_f(\pi_k) = \mathbb{E}_{\rho^\pi,\pi}[A_f^{\pi_k}], \qquad \widetilde{J}_f^k(\pi) - J_f(\pi_k) = \mathbb{E}_{\rho^{\pi_k},\pi}[A_f^{\pi_k}]. \tag{61}$$

Subtracting gives

$$J_f(\pi) - \widetilde{J}_f^k(\pi) = \sum_z(\rho^\pi(z)-\rho^{\pi_k}(z))\sum_a \pi(a\mid z)A_f^{\pi_k}(z,a). \tag{62}$$

Hence

$$|J_f(\pi) - \widetilde{J}_f^k(\pi)| \le \|\rho^\pi-\rho^{\pi_k}\|_1\|A_f^{\pi_k}\|_\infty. \tag{63}$$

Using the resolvent form of discounted occupancy and Pinsker's inequality,

$$\|\rho^\pi-\rho^{\pi_k}\|_1 \le \frac{2\gamma}{1-\gamma}\sqrt{\varepsilon/2}. \tag{64}$$

Since $\|A_f^{\pi_k}\|_\infty \le 2F_{\max}/(1-\gamma)$, the claim follows.

L Local Dual Solver Details

In practice, the local dual is only two-dimensional in our setting, corresponding to object safety and robot safety. We solve

$$\lambda_k^\star = \arg\min_{\lambda \geq 0} g_k(\lambda) \tag{65}$$

using projected gradient descent by default and optionally projected Newton updates when the empirical Hessian is well-conditioned. Since

$$\nabla^2 g_k(\lambda) = \frac{1}{\beta}\, \mathbb{E}_{z\sim\rho_k} \mathrm{Cov}_{\pi_{k,\lambda}}(A_c^{\pi_k}(z,\cdot)), \tag{66}$$

the Hessian is available in closed form up to batch estimation.

We terminate dual optimization when one of the following holds:

$$\|\nabla g_k(\lambda)\|_\infty \leq \epsilon_{\mathrm{dual}}, \qquad \|\lambda^{(t+1)} - \lambda^{(t)}\|_2 \leq \epsilon_{\mathrm{step}}, \qquad t \geq T_{\max}. \tag{67}$$

In our experiments, the dual converges quickly and remains low-dimensional throughout training, consistent with the computational motivation of the dualization strategy in One-Shot Safety Alignment.

M Compute Resources

All experiments were run on NVIDIA L20 GPUs. Details of compute usage:

Table 6: Compute resources for Dualized-SafeVLA experiments.

Resource	Specification
GPU type	NVIDIA L20
Number of GPUs	8
GPU memory	48 GB each
Wall-clock time	72 hours per full run
Total training time	8 GPU-weeks
Software	Ubuntu 22.04, Python 3.11, PyTorch 2.x

N One-Pass Policy Distillation

After solving for $\lambda_k^\star$, we form the target policy

$$\pi_k^{\mathrm{tar}}(a \mid z) \propto \pi_k(a \mid z) \exp\left(\frac{A_r^{\pi_k}(z,a) - \lambda_k^{\star\top} A_c^{\pi_k}(z,a)}{\beta}\right). \tag{68}$$

We then distill it into the parametric VLA policy with sequence-level weighted imitation:

$$\mathcal{L}_{\mathrm{distill}}(\theta) = -\mathbb{E}_{(z,a)\sim\mathcal{D}_k}\left[w_k(z,a) \log \pi_\theta(a \mid z)\right], \tag{69}$$

where

$$w_k(z,a) \propto \exp\left(\frac{A_r^{\pi_k}(z,a) - \lambda_k^{\star\top} A_c^{\pi_k}(z,a)}{\beta}\right). \tag{70}$$

In implementation, we apply three stabilization strategies:

1. mini-batch normalization of trajectory weights;
2. length normalization of sequence losses;
3. KL clipping or backtracking to enforce the local trust region.

O Additional Algorithmic Remarks

Our method should be interpreted as a *local* one-shot strategy rather than a globally fixed-dual shortcut. Unlike the static conditional setting in One-Shot Safety Alignment, embodied CMDP still require outer occupancy refreshes. The benefit of Dualized-SafeVLA is therefore not the elimination of iteration altogether, but the elimination of repeated multiplier–model interleaving *within each local occupancy region*. This is the correct Bellman-aware analogue of the two-stage CAN strategy.

P Reproducibility Instructions

To reproduce the main experimental results:

1. Install the required environment: Python 3.11, PyTorch 2.x, NVIDIA CUDA 12.2.
2. Download the Safety-CHORES dataset and SafeVLA pre-trained checkpoints.
3. Run training scripts with hyperparameters listed in Section B.
4. Evaluate using the 200-task subset per checkpoint for SR and CC metrics.
5. Optional: Ablation experiments can be run by setting the corresponding flags (remove Bellman correction, freeze occupancy) in the configuration.