2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SVP: Improving Vision-Language-Action Models with Dual Stochastic Visual Prompting

Zhide Zhong*[1], Haodong Yan*[1], Tianran Zhang[1], Lujia Wang[1], Jin Wu[2], Jun Ma[1], Xinhu Zheng[1], Haoang Li[1]

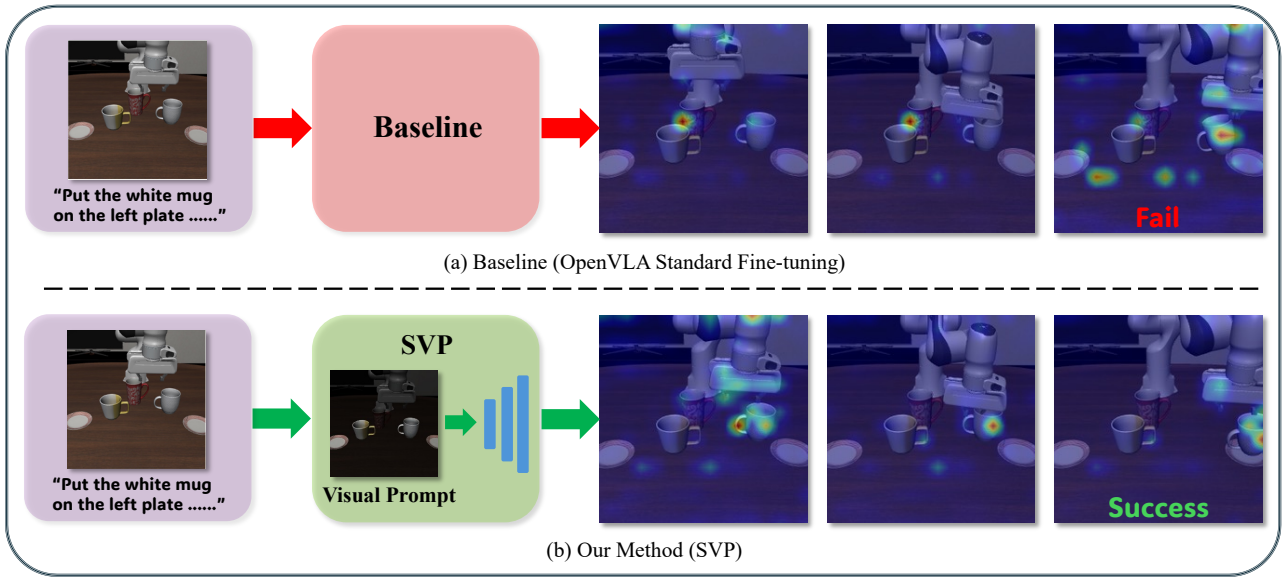


Fig. 1: **Overview of SVP and its effect on model attention.** (a) The baseline model, trained with standard behavior cloning, directly maps observations and instructions to actions. As the **attention visualizations** show, this often leads to shortcut learning that learns spurious correlations with background textures or wrong objects, resulting in failure. (b) In contrast, our SVP model is trained with an additional guidance signal: **stochastic visual prompts**. These prompts function as a **training-only** "visual scaffold", compelling the model to internalize a robust, instruction-grounded attention policy. Consequently, at inference time on original unprompted images, our model's **attention** remains accurate and consistent, successfully performing the task.

***Abstract*— Vision-Language-Action (VLA) models, such as OpenVLA, hold the promise of generalist robots, yet their performance is often impaired by distracted attention, which we identify as a manifestation of shortcut learning. We posit that the solution lies not in architectural modifications, but in a new training paradigm centered on visual prompts that provide explicit visual guidance to the model. We introduce Dual Stochastic Visual Prompting (SVP) as a concrete realization of this paradigm. SVP functions as a training-only "visual scaffold", a non-invasive mechanism that requires no architectural modifications. Our work demonstrates that this data-centric training paradigm is a highly effective strategy for mitigating distracted attention, enabling the learning of more robust and capable policies without architectural overhead. SVP yields substantial gains on the challenging LIBERO benchmark and real robot experiments. It improves the absolute success rate of the standard OpenVLA by 8.2% on long-horizon tasks and enhances the performance of the highly optimized OpenVLA-OFT. These improvements are validated on a real robot, where our model consistently outperforms baselines across a variety of manipulation tasks.**

This work was supported in part by the Natural Science Foundation of China under Grant 62403401, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2026A1515012323 and Grant 2024A1515011992, in part by the Guangdong Provincial Project under Grant 2024QN11X127, and in part by the AI Research and Learning Base of Urban Culture under Grant 2023WZJD008.

*Zhide Zhong and Haodong Yan contributed equally to this work.
Corresponding Author: Haoang Li (haoangli@hkust-gz.edu.cn)
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]University of Science and Technology Beijing

I. Introduction

The rise of large-scale pre-trained models has led to a major shift in robotics, marked by the emergence of Vision-Language-Action (VLA) models [1–5]. These models form the policies for a new class of generalist agents capable of understanding natural language and executing complex tasks in diverse, unstructured environments. This approach holds great promise for improving data efficiency and generalization, suggesting a path beyond narrow, single-task systems.

However, a critical vulnerability often emerges during fine-tuning, which we call "distracted attention". We argue that this is not random behavior but a direct result of **shortcut learning** [6]. Lacking explicit visual supervision, the model learns incorrect associations from spurious correlations in the training data. For example, as shown in Figure 1(a), rather than tracking command-specified objects, the model's attention may focus on a visually prominent feature like the "wood grain of the table", having wrongly associated this texture with general manipulation actions. This attention is visualized as a heatmap generated by computing the

cross-attention from **action tokens to image tokens** and upsampling the result to the image resolution. This unreliable focus frequently leads to policy failure, making the resulting policies a major barrier to their real-world deployment.

We believe the solution lies not in more complex architectures, but in a new training paradigm that leverages visual prompting as a form of explicit yet temporary guidance. To this end, we introduce **Dual Stochastic Visual Prompting (SVP)**, a method that introduces randomness along two dimensions: (1) the decision of whether to apply a prompt to a training sample, and (2) the visual intensity of the prompt when it is applied. This dual randomness prevents the model from over-relying on the prompts and forces it to learn robust, generalizable features. Consequently, SVP acts as a non-invasive, training-only "visual scaffold" without requiring any architectural changes. Unlike standard augmentation, which modifies visuals irrespective of the language instruction, SVP is fundamentally instruction-dependent. It resolves attentional ambiguity by programmatically highlighting the object specified in the language goal. By stochastically applying this guidance, we make spurious background cues statistically unreliable. This forces the model to abandon these shortcuts and instead learn the invariant causal link between the language instruction and an object's intrinsic visual features. The result is an internalized, robust attention policy that functions without external guidance at inference time.

To validate our approach, we integrate SVP into the training process of leading VLA models and evaluate them on the challenging LIBERO benchmark [7] and real-world manipulation tasks. Our experiments show substantial gains. When applied to the standard OpenVLA model [1], SVP improves the overall success rate from 74.0% to 77.7%, with the largest impact observed where attention failures are most common. In the complex LIBERO-10 suite, which contains the longest and most difficult tasks, SVP delivered a remarkable **8.2% absolute gain** in the success rate (Table I). Importantly, to test the general applicability of our paradigm, we also applied SVP to the highly optimized OpenVLA-OFT [4], confirming that our method enhances its performance as well. Grounding these simulation results, we conducted further experiments on a real robot. These real-world evaluations corroborate our findings, showing that the SVP-enhanced model consistently outperforms strong baseline policies across a variety of manipulation tasks.

Our work makes the following contributions:

- We identify attention instability in VLA fine-tuning as a key manifestation of shortcut learning. To address this fundamental issue, we propose a novel training paradigm centered on dynamic data intervention, shifting the focus from architectural change to data-centric solutions.
- As a concrete implementation of this paradigm, we introduce SVP, a simple and non-invasive method. Through SVP, we provide the community with both a practical tool and a new perspective, demonstrating that substantial improvements can be unlocked not by designing more complex architectures, but by fundamentally improving the training process itself.
- We validate our approach through extensive experiments, demonstrating that SVP yields significant performance gains on a standard baseline (OpenVLA) and a highly optimized variant (OpenVLA-OFT), which showed particular improvement on challenging long-horizon tasks, as well as in real-world manipulation tasks.

II. Related Work

A. Generalist Robot Models

Recent progress in generalist robot models has largely followed two paths. One line of work, exemplified by RT-1 [8] and BridgeData [9], focuses on scaling up large, multi-task, multi-domain datasets to train high-capacity Transformer models for broad generalization. A complementary direction, seen in works like ALOHA [10] and Diffusion Policy [11], prioritizes data efficiency and architectural novelty, using techniques like semantic augmentation, action chunking, and diffusion-based representations to learn robust skills from smaller datasets. While these approaches have greatly advanced robot capabilities, our research places a stronger emphasis on scenarios involving detailed, open-ended language instructions to guide and condition behavior.

The paradigm of Vision-Language-Action (VLA) models [1–3, 5, 12–19] has become central to building generalist policies. By fine-tuning pre-trained Vision-Language Models (VLMs) on large-scale robot data, this approach enables policies to inherit a rich semantic understanding and commonsense reasoning from web-scale pre-training. Leading works have firmly established this direction: RT-2 [2] demonstrated the surprising effectiveness of transferring web knowledge to control, while subsequent models like OpenVLA [1], Octo [3], and $\pi 0$ [17] have solidified this approach by scaling models and data. Consequently, a primary research focus has been on model-centric refinements, such as designing more efficient architectures or improving action decoding schemes, from autoregressive generation to parallel decoding for faster inference. While these efforts have pushed performance boundaries, our work diverges from this trend. Instead of focusing on architectural changes, we introduce Stochastic Visual Prompting (SVP), a training-time regularization technique designed to address the fundamental problem of attention instability and improve language grounding.

III. Methodology

Our methodology is developed based on the hypothesis that the observed attention instability in VLA models is a direct consequence of visually rooted shortcut learning. **The core logic of our approach, including (1) identifying the problem, (2) proposing a solution, and (3) achieving the desired result, is visually summarized in Figure 2.** This section first outlines the standard VLA fine-tuning framework to establish the context. We then present the motivation for our approach by formally analyzing how the sparsity of behavioral cloning supervision contributes to

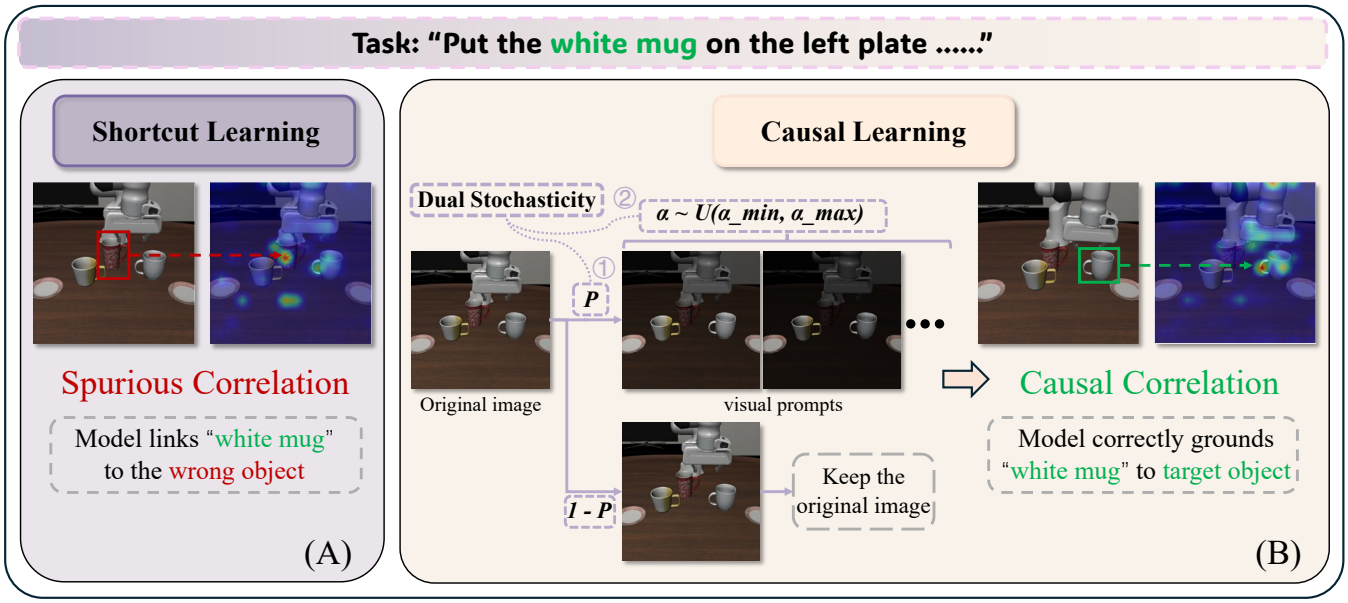


Fig. 2: **Overview of our Dual Stochastic Visual Prompting (SVP) methodology.** Our approach transforms a model from a shortcut learner into a causal learner. **(A) The Problem: Shortcut Learning.** Standard fine-tuning can lead to *spurious correlations*. For the task "Put the white mug...", a baseline model may erroneously focus on a more visually salient but incorrect object (e.g., the red mug), failing to correctly ground the language instruction. **(B) Our Solution: Causal Learning via SVP.** We introduce a training-only prompting with *Dual Stochasticity*: 1) A visual prompt is probabilistically applied to the target object with probability P. 2) The prompt's intensity, α, is randomized by sampling from a uniform distribution $U(\alpha_{\min}, \alpha_{\max})$. This process breaks spurious correlations and compels the model to learn the true *causal correlation* between the text "white mug" and its corresponding visual features. Consequently, at inference time on a new unseen image, the SVP-trained model correctly grounds the instruction to the target object, demonstrating robust and accurate attention.

the model's reliance on spurious visual correlations. Based on this analysis, we introduce our proposed solution, Dual Stochastic Visual Prompting (SVP), a training-only paradigm designed to mitigate these shortcuts by acting as a "visual scaffold".

A. Preliminaries: VLA Framework

A Vision-Language-Action (VLA) model, denoted as a policy f_θ parameterized by θ, predicts a sequence of actions $a = \{a_1, \ldots, a_T\}$ based on a natural language instruction l and a history of visual observations $I = \{I_1, \ldots, I_T\}$. The action at each timestep t is thus given by $a_t = f_\theta(l, I_t)$.

Fine-tuning such models on a downstream task dataset $\mathcal{D}$ requires adapting the standard behavioral cloning (BC) objective to the model's language-centric architecture. Each sample in $\mathcal{D}$ is a triplet $(l, I, a_{\text{expert}})$. Following the approach of modern VLAs like OpenVLA [1], the continuous expert actions a_{expert} are first discretized. Each dimension of the continuous action vector is quantized into a finite set of bins (e.g., 256), converting the action into a sequence of discrete "action tokens". The model's parameters θ are then updated via BC loss. This is formulated as minimizing the negative log-likelihood of the expert's discretized action trajectory, which corresponds to a cross-entropy loss objective:

$$\mathcal{L}_{\text{BC}} = \mathbb{E}_{(l,I,a_{\text{expert}})\sim\mathcal{D}}[-\log \pi_\theta(a_{\text{expert}}|l, I)], \qquad (1)$$

where $\pi_\theta(a_{\text{expert}}|l, I)$ is the likelihood of the expert action token sequence, typically factorized over timesteps and action dimensions. Despite its effectiveness for action generation, this token-level supervision does not offer direct guidance on where the model should look to make its decision.

B. Motivation: Shortcut Learning as the Root of Instability

The supervision signal from the BC loss is inherently indirect for visual representation learning. It provides feedback only at the action output level and offers no explicit guidance on the model's intermediate reasoning, specifically on which visual features its attention should focus. This lack of direct visual supervision creates an under-constrained optimization problem, where multiple visual-to-action mappings can yield a low loss. As identified in recent work [6], this condition allows the model to converge to statistically simple but non-causal solutions by exploiting **spurious correlations**—a phenomenon known as shortcut learning. Following their framework, we decompose a visual observation into task-relevant causal factors (e.g., the target object) and task-irrelevant, non-causal factors (e.g., background, lighting). Shortcut learning is thus defined as the acquisition of a policy that develops a dependency on non-causal factors.

This form of learning directly manifests as the **attention instability** we observe. Without explicit constraints in the loss function, the model's attention mechanism is not incentivized to align with the causal factors. Instead, its attention weights often converge to features in non-causal factors that are statistically prominent but non-robust. **This is visualized in Panel (A) of Figure 2**, where the attention distribution is diffuse or incorrectly allocated to task-irrelevant regions, rather than being concentrated on the critical objects. We posit that this instability is a fundamental

bottleneck preventing the model from learning a generalizable policy. Therefore, our key insight is that a robust policy requires a direct intervention that provides an explicit visual signal to **disrupt the learned correlation between non-causal factors and the action**, while simultaneously guiding attention towards causal factors.

C. Dual Stochastic Visual Prompting (SVP)

To directly combat this visually-rooted shortcut learning, we introduce Dual Stochastic Visual Prompting (SVP). Conceptually, SVP functions as a **visual scaffold**: a temporary, auxiliary structure present only during training to guide the model's learning process. **The mechanism, which we term a "visual scaffold", is illustrated in Panel (B) of Figure 2.** This scaffold provides a rich, targeted supervisory signal for attention, yet it introduces zero architectural changes or inference overhead. To implement this, SVP is composed of two key components: focus-enhancing visual prompts and a dual stochasticity mechanism.

a) Focus-Enhancing Visual Prompts: The structure of our visual scaffold is a "spotlight" effect that highlights the task-relevant object. For each training image $\mathbf{I}_t$, we first obtain the segmentation mask $\mathbf{M}_{obj}$ of the target object, which is readily available in simulated environments or from standard segmentation models. The prompt is then generated by keeping the pixels within the mask $\mathbf{M}_{obj}$ unchanged while dimming the pixels in the background. This creates a strong saliency signal, making the target object visually "pop out" and providing an unambiguous guide for the model's attention, **effectively disrupting the information content of potential shortcuts**.

b) Dual Stochasticity Mechanism: A key challenge when using any form of training-time guidance is to prevent the model from becoming dependent on this "scaffolding". The goal is for the model to *internalize* the learned policy so it can stand on its own after the scaffold is removed. The Dual Stochasticity Mechanism is critical to achieving this. It incorporates two complementary forms of randomness.

First, *Randomized Background Dimming*. The intensity of the scaffold is not fixed. For each application of SVP, we sample a dimming factor α from a uniform distribution $\mathcal{U}(\alpha_{\min}, \alpha_{\max})$. The augmented image $\mathbf{I'}_t$ is then generated as:

$$\mathbf{I'}_t(x,y) = \begin{cases} \mathbf{I}_t(x,y) & \text{if } (x,y) \in \mathbf{M}_{obj} \\ \alpha \cdot \mathbf{I}_t(x,y) & \text{if } (x,y) \notin \mathbf{M}_{obj} \end{cases} \quad (2)$$

This randomization forces the model to learn features that are robust to varying contrast and lighting conditions, rather than simply learning a bright-versus-dark detector.

Second, *Probabilistic Application*. The entire visual scaffold is applied probabilistically. At each training step, we apply the prompt with a probability p (e.g., $p = 0.5$). For the remaining $1 - p$ of the time, the model is trained on the original, unmodified image. This is crucial for "weaning" the model off the scaffold. By frequently exposing the model to the standard visual input, we ensure it does not become dependent on the prompts. This forces the model to internalize the attention policy learned from the scaffolded examples into a general capability that functions robustly when the scaffold is fully absent at inference time.

The complete training procedure for our SVP, is outlined in Algorithm 1.

Algorithm 1 SVP Training Paradigm

Input: Pre-trained VLA model f_θ, Dataset $\mathcal{D}$, application probability p, dimming range $[\alpha_{min}, \alpha_{max}]$.
for each training sample $(l, \{\mathbf{I}_t\}_{t=1}^T, \{\mathbf{a}_{t,\text{expert}}\}_{t=1}^T)$ in $\mathcal{D}$ **do**
  Initialize prompted image sequence $\{\mathbf{I'}_t\}_{t=1}^T$
  for $t = 1$ to T **do**
    Draw a random number $r \sim \text{Uniform}(0, 1)$.
    if $r < p$ **then**
      Get object mask $\mathbf{M}_{obj,t}$ for the task-relevant object in $\mathbf{I}_t$.
      Sample dimming factor $\alpha \sim \mathcal{U}(\alpha_{\min}, \alpha_{\max})$.
      Generate prompted image $\mathbf{I'}_t$ using Eq. (2).
    else
      $\mathbf{I'}_t \leftarrow \mathbf{I}_t$
    end if
  end for
  Compute loss $\mathcal{L}_{\text{BC}}$ using the prompted sequence $\{\mathbf{I'}_t\}$ and Eq. (1).
  Update parameters θ using gradient descent on $\mathcal{L}_{\text{BC}}$.
end for
Note: At inference, use the original, unmodified images $\{\mathbf{I}_t\}$.

IV. Experiments

Our experiments are designed to answer three key questions:

1) Does our SVP paradigm significantly improve the performance of VLA models compared to standard fine-tuning, both on simulated benchmarks and in real-world manipulation tasks?
2) Are the core components of SVP, particularly the dual stochasticity mechanism, essential for its success?
3) Does SVP function by stabilizing the model's attention mechanism as hypothesized?

A. Experimental Setup

a) Base Models: Our experiments are conducted on both a standard VLA model and its state-of-the-art variant to test the generalizability of our method. The models are:

- **OpenVLA** [1]: A standard and widely-adopted 7B parameter Vision-Language-Action model used as our baseline.
- **OpenVLA-OFT** [4]: A state-of-the-art enhancement that applies an Optimized Fine-Tuning (OFT) recipe to the base OpenVLA, incorporating parallel decoding and an L_1 regression objective.

TABLE I: Effectiveness of our Stochastic Visual Prompting (SVP). SVP is applied to both the standard OpenVLA and the optimized OpenVLA-OFT, showing consistent improvements. We report the average success rate (%) over 500 rollouts.

Method	LIBERO-Spatial	LIBERO-Object	LIBERO-Goal	LIBERO-10	Average
Diffusion Policy [11]	78.3	92.5	68.3	50.5	72.4
Octo [3]	78.9	85.7	84.6	51.1	75.1
TraceVLA [20]	84.6	85.2	75.1	54.1	74.8
SpatialVLA [21]	88.2	89.9	78.6	55.5	78.1
WorldVLA [22]	85.6	89.0	82.6	59.0	79.1
CoT-VLA [23]	87.5	91.6	87.6	69.0	81.1
OpenVLA [1]	84.7	87.2	74.4	49.8	74.0
+ SVP (ours)	87.8	87.3	77.8	58.0	77.7
Improvement (Δ)	***+3.1***	***+0.1***	***+3.4***	***+8.2***	***+3.7***
OpenVLA-OFT [4]	96.2	98.3	96.2	90.7	95.3
+ SVP (ours)	97.2	98.0	96.2	93.2	96.2
Improvement (Δ)	***+1.0***	***-0.3***	***+0.0***	***+2.5***	***+0.9***

TABLE II: Ablation study on the LIBERO-10 task suite. Success rates (%) show the contribution of each SVP component and the effect of stochasticity on internalizing the guidance.

Method	Success Rate (%)
OpenVLA Vanilla Fine-tuning (Baseline)	49.8
Ablations	
Deterministic Prompt (Train) / No Prompt (Test)	50.0
Deterministic Prompt (Train & Test)†	59.0
Always-On SVP ($p = 1.0$, $\alpha \sim \mathcal{U}(0.3, 0.5)$)	52.1
Fixed Dimming ($p = 0.5$, $\alpha = 0.4$)	56.3
OpenVLA + SVP (Ours, No Prompt at Test)	**58.0**

†Tested with prompts, providing a performance upper bound.

b) Benchmark and Metrics: We evaluate our method on the **LIBERO benchmark** [7] and use the task success rate as our primary metric. LIBERO is a suite of challenging, long-horizon manipulation tasks ideal for testing generalization. We report the average success rate over 500 evaluation rollouts across its four main suites: LIBERO-Spatial (testing spatial relationships), LIBERO-Object (varying object identities), LIBERO-Goal (altering task objectives), and the particularly difficult LIBERO-10 (combining all challenges in complex, multi-stage scenarios). This benchmark is a fitting testbed for our work, as its compositional nature and long-horizon tasks create ample opportunities for models to form spurious correlations, making it highly sensitive to the attention instability we aim to solve.

c) Implementation Details: Across all experiments, we use the AdamW optimizer with a learning rate of 5e-4 and train models until convergence. Our method, Stochastic Visual Prompting (SVP), is applied to two distinct baseline configurations. For the standard **OpenVLA**, we follow its original autoregressive procedure with a batch size of 16, and apply SVP with an application probability of $p = 0.5$. For **OpenVLA-OFT**, we adopt its official recipe featuring parallel decoding and L1 regression, using a batch size of 64. Unless otherwise specified, the dimming factor α is sampled from $\mathcal{U}(0.3, 0.5)$. This range was empirically chosen to strike a balance between providing a strong attentional cue and preserving scene context. As illustrated in Figure 3, an overly aggressive factor (e.g., $\alpha = 0.2$) obscures crucial spatial references, while a conservative factor (e.g., $\alpha = 0.6$) fails to create sufficient contrast.

Instruction: "Pick up the black bowl between the plate and the ramekin and place it on the plate."

(a) Original Image. (b) $\alpha = 0.2$. (c) $\alpha = 0.6$.

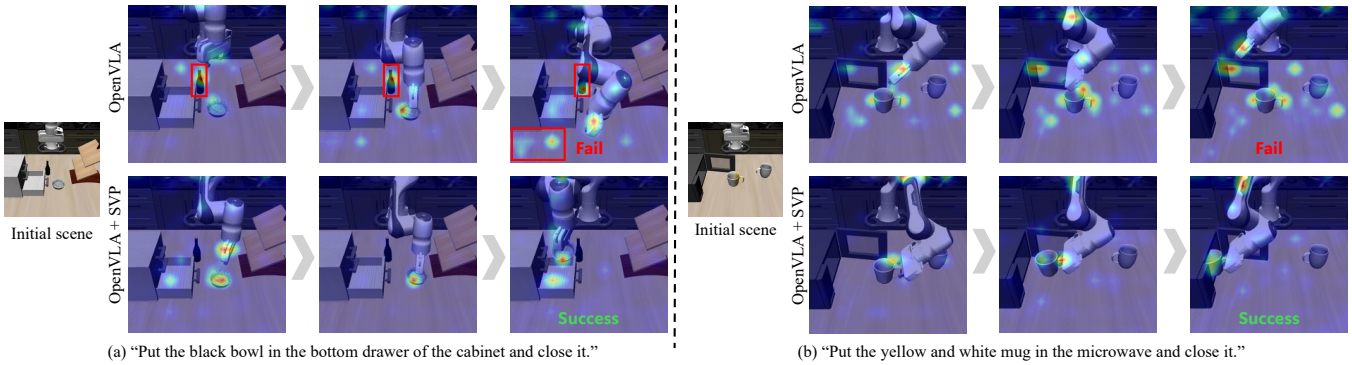
Fig. 3: **Visual justification for the choice of dimming factor α.** We compare the original image (a) with the effects of two extreme α values. An overly aggressive factor (b) obscures vital spatial references, while a conservative factor (c) provides an insufficient attentional cue. Our chosen range, $\alpha \sim \mathcal{U}(0.3, 0.5)$, strikes a balance between these extremes, effectively highlighting the target while preserving essential scene context.

B. Main Results: Effectiveness of SVP

We analyze the effectiveness of our Stochastic Visual Prompting (SVP) by applying it to two distinct foundational models, as shown in Table I.

First, we evaluate SVP on the standard OpenVLA baseline. Our method provides a clear performance advantage, boosting the average success rate from 74.0% to 77.7%, a notable **3.7% absolute improvement**. The impact is most evident on the most challenging suite, **LIBERO-10**, where SVP increases the success rate from 49.8% to 58.0%—an impressive **8.2% gain** (a 16.5% relative improvement). This highlights SVP's ability to enhance robustness in complex scenarios where attention is prone to fail, while maintaining performance on simpler tasks like LIBERO-Object, where the baseline is already stable.

To further demonstrate the generalizability of our method, we apply SVP to the highly optimized OpenVLA-OFT, which already achieves a very strong average success rate of 95.3%. Even on this powerful baseline, SVP delivers notable improvements, especially on challenging suites. For instance, on the difficult LIBERO-10 suite, SVP again provides a significant boost, raising performance from **90.7%**



(a) "Put the black bowl in the bottom drawer of the cabinet and close it."

(b) "Put the yellow and white mug in the microwave and close it."

Fig. 4: **Qualitative comparison of attention maps for OpenVLA and our OpenVLA + SVP.** Each task (a) and (b) compares a failing trajectory from the baseline (top) with a successful one from our model (bottom). The baseline's attention is unstable and distracted, leading to failure. Conversely, our model maintains a tight, consistent focus on the task-relevant objects, demonstrating that SVP effectively regularizes the model to ground language instructions to visual elements.

to **93.2%**. On the simpler LIBERO-Object suite, we observe a marginal decrease of 0.3%. We hypothesize this is because the suite's less complex tasks allow the baseline to achieve near-optimal results simply by overfitting demonstrations, a strategy that succeeds without the robust causal reasoning that SVP promotes and thereby diminishes the impact of our method's attention guidance. Nevertheless, the pattern of significant gains on complex suites is the dominant trend, culminating in our final model, OpenVLA-OFT + SVP, which sets a new state-of-the-art performance on this benchmark. This second experiment confirms that SVP is not merely a fix for a specific model's weakness but a versatile regularization technique that enhances the stability of modern VLAs, particularly when facing complex, long-horizon tasks.

C. Ablation Study: Importance of SVP Components

To understand how SVP achieves robust, guidance-free performance, we conduct a detailed ablation study on the challenging LIBERO-10 task suite. Using OpenVLA as our baseline, we analyze the respective roles of the visual prompt and the dual stochasticity, SVP's core components, by comparing our full method against several key variants. The results are summarized in Table II.

a) The Power and Peril of a Deterministic Scaffold: First, we establish the potential of the visual prompt itself. When a deterministic prompt (always on, $p = 1.0$, with a fixed dimming factor, $\alpha = 0.4$) is used during both training and testing, the success rate reaches an impressive **59.0%**. This result serves as a practical upper bound, demonstrating that the "spotlight" provides a powerful and effective learning signal when consistently available.

However, this reliance on a constant scaffold reveals a critical vulnerability. When the same deterministically trained model is tested *without* the prompt—mimicking real-world deployment—its performance collapses to 50.0%, barely above the baseline (49.8%). This drastic drop highlights a classic case of overfitting: the model did not learn to *focus*, but rather learned to *depend on the prompt* as a shortcut. It failed to internalize the attention policy.

b) Stochasticity as the Bridge to Internalization: This is where stochasticity becomes essential. Its role is not merely to provide guidance, but to systematically break the model's dependency on it. We analyze each stochastic mechanism individually:

- **Varying the Prompt's Appearance:** By introducing randomness only to the dimming factor ($\alpha \sim \mathcal{U}(0.3, 0.5)$) while keeping the prompt always on ($p = 1.0$), the scaffold-free performance rises to 52.1%. This suggests that preventing the model from over-relying on a specific visual cue is beneficial.
- **Making the Prompt Unreliable:** More strikingly, making the prompt's presence itself unreliable ($p = 0.5$) yields a massive performance leap to **56.3%**. By forcing the model to perform correctly even on the many samples where no prompt is given, this mechanism directly compels it to learn the invariant relationship between the command and the object's intrinsic features.

c) Achieving Full Internalization with Dual Stochasticity: Finally, our full OpenVLA + SVP method, which combines both sources of randomness, achieves the highest success rate of **58.0%** in a scaffold-free test environment. This result is remarkable: the model has successfully **internalized** the guidance, retaining nearly all the performance benefits of the "always-on" prompt (58.0% vs. the 59.0% upper bound) without needing it at deployment.

In conclusion, this study reveals that SVP's success stems from a sophisticated mechanism. The visual prompt provides a strong initial signal, while the dual stochasticity is the critical engine that forces the internalization of this signal, transforming a dependency into a robust, self-sufficient policy.

D. Qualitative Analysis: Attention Visualization

To provide mechanistic insight into why SVP improves performance, we visualize and compare the internal attention maps of our OpenVLA + SVP model against the OpenVLA baseline. Figure 4 presents this side-by-side comparison across two challenging, long-horizon tasks, where high-attention areas are indicated in red.

A clear pattern of failure emerges for the baseline model: its attention is unstable and fails to lock onto the correct

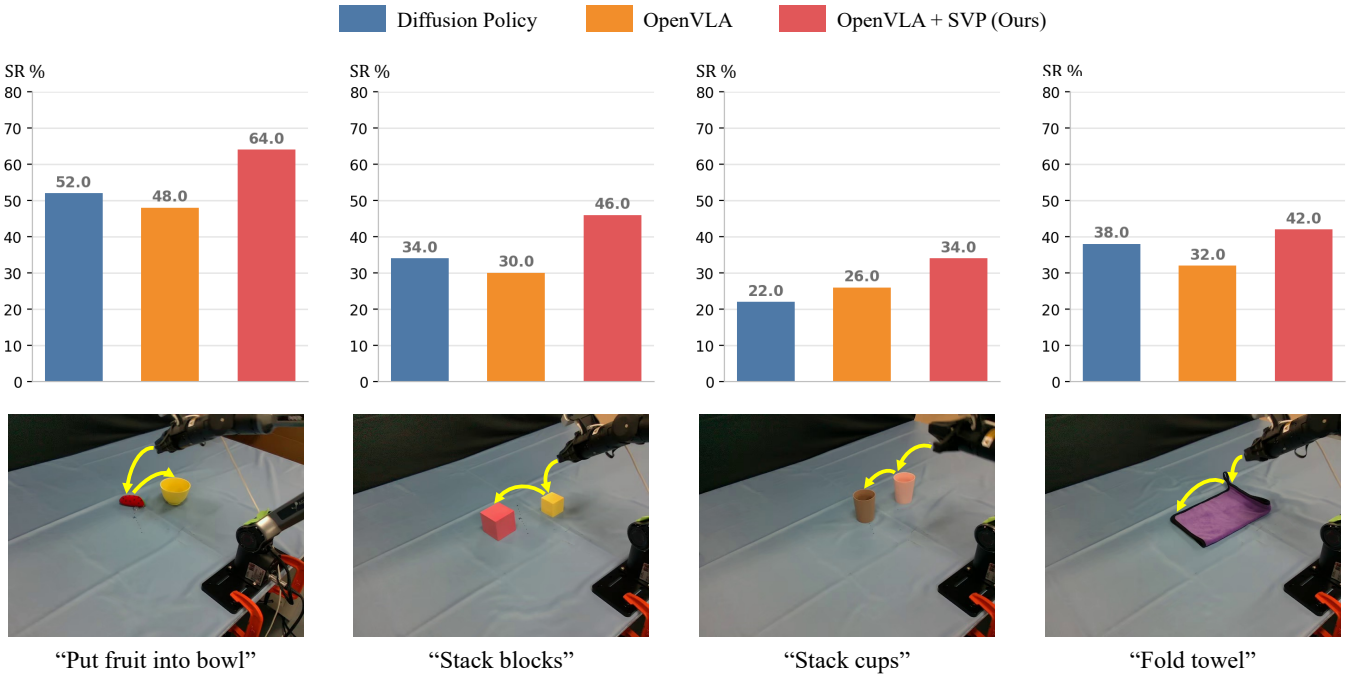


Fig. 5: **Real-world robot experiments.** We propose four different tasks: "Put fruit into bowl", "Stack blocks", "Stack cups", and "Fold towel", which evaluate a range of capabilities from basic pick-and-place to more complex dexterous manipulation. Our method consistently achieves the highest performance across all four scenarios.

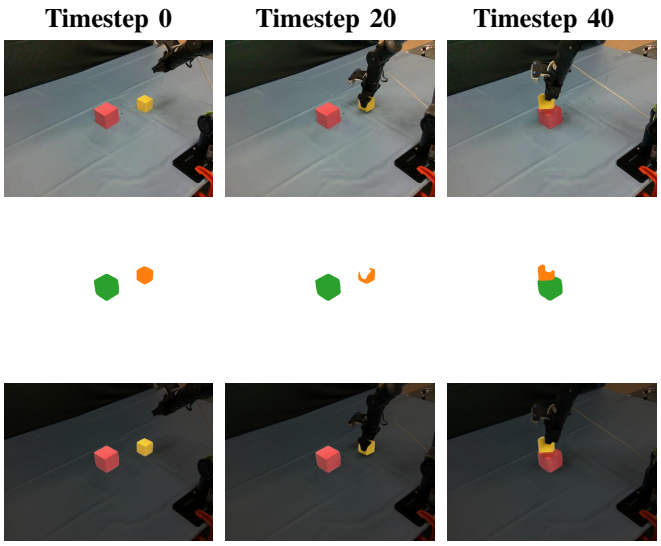


Fig. 6: **SVP data generation and application over a trajectory.** This grid visualizes our data processing pipeline across three distinct timesteps of a single task. **Columns (left to right):** Show the scene at an early ($t = 0$), middle ($t = 20$), and late ($t = 40$) stage of the trajectory. **Rows (top to bottom):** Correspond to (1) the original camera views, (2) the object masks extracted via tracking, and (3) the images after applying the visual prompt.

target from the outset. For instance, in task (a), the baseline's attention is divided between the target bowl and a salient but task-irrelevant distractor—the wine bottle. Throughout the trajectory, the model never resolves this ambiguity, and during the grasp attempt, the focus incorrectly intensifies on the bottle. This fundamental failure to disambiguate the target from the distractor directly causes the grasp to fail. Similarly, in task (b), the baseline's attention is erroneously scattered across the scene from the beginning, with significant activation on the background and the destination (the microwave) rather than the specified mug. This demonstrates an inability to ground the language instruction, again resulting in failure.

In stark contrast, our OpenVLA + SVP model maintains a remarkably stable and precise attention focus. In task (a), attention remains sharply locked onto the target "black bowl" throughout the entire trajectory, enabling a successful grasp and placement. For the more complex task (b), our model correctly grounds the "yellow and white" attributes, focusing exclusively on the correct mug from the outset and ignoring distractors. This sustained focus is directly correlated with its successful execution of the task.

These comparative visualizations provide compelling qualitative evidence that SVP acts as an effective regularizer. By guiding the model to develop a robust attention mechanism that is less prone to distraction, the SVP ensures that language instructions are accurately and consistently grounded in the correct visual elements over time. This stability in attention is a key factor behind the observed improvements in task performance.

E. Real-Robot Experiments

a) Real-Robot Setup: We conduct our real-world experiments with a 7-DoF Piper arm from AgileX Robotics and a single third-person Orbecc DABAI RGB camera. As shown in Figure 5, we design a suite of manipulation tasks (e.g., "stack blocks") requiring precise control. For each task, we collect 20-100 human-teleoperated demonstrations for fine-tuning. To generate the visual prompts required by our SVP method, we employ a practical semi-automated pipeline:

after manually providing a single keypoint for the target object in the initial frame, we leverage SAM2 [24] to automatically propagate the segmentation mask throughout the entire trajectory, as shown in Figure 6. This approach avoids tedious per-frame manual annotation. Finally, each trained policy is evaluated for 50 trials to validate its effectiveness in a real-world setting.

b) Results: The results of our real-robot evaluation, presented in Figure 5, demonstrate the practical benefits of our method. We compare our final model, OpenVLA + SVP (red), against two strong baselines: an enhanced OpenVLA variant (orange) with action chunking and parallel decoding, and Diffusion Policy (blue). The primary finding is that applying SVP consistently improves the performance of the OpenVLA model across all tasks. On the challenging deformable object task, "Fold towel", SVP increases the success rate from 32.0% for OpenVLA to 42.0%. Similarly, for "Stack blocks", performance jumps from 30.0% to 46.0%. Furthermore, our OpenVLA + SVP model outperforms not only its direct OpenVLA baseline but also the Diffusion Policy baseline in every tested scenario. These results validate the effectiveness of SVP in a real-world setting, demonstrating its ability to enhance policy robustness even with limited demonstration data.

V. Conclusion

In this work, we introduce Dual Stochastic Visual Prompting (SVP), a training paradigm that addresses attention instability in VLA models by operating as a temporary "visual scaffold." Requiring no architectural changes or inference overhead, SVP yields an 8.2% absolute improvement for OpenVLA on LIBERO's most challenging tasks, with its effectiveness further validated in real-world experiments. This success is attributed to the robust, intrinsic attention policy that the model internalizes from the stochastic guidance. More broadly, our findings suggest that the path to more capable robots lies not only in architectural innovation but also in the development of sophisticated training methodologies. SVP offers a conceptual blueprint for how temporary structured guidance can unlock a model's latent capabilities, underscoring a crucial principle: teaching a model *how to learn* is as vital as designing the model itself.

References

[1] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *in Conference on Robot Learning*, 2024.

[2] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[3] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[4] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[5] Y. Wang, X. Li, W. Wang, J. Zhang, Y. Li, Y. Chen, X. Wang, and Z. Zhang, "Unified vision-language-action model," *arXiv preprint arXiv:2506.19850*, 2025.

[6] Y. Xing, X. Luo, J. Xie, L. Gao, H. Shen, and J. Song, "Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation," *arXiv preprint arXiv:2508.06426*, 2025.

[7] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 776–44 791, 2023.

[8] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *Robotics: Science and Systems*, 2023.

[9] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," *Robotics: Science and Systems*, 2022.

[10] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *Robotics: Science and Systems*, 2023.

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

[12] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," in *ICLR*, 2024.

[13] C.-L. Cheang, G. Chen, Y. Jing, T. Kong, H. Li, Y. Li, Y. Liu, H. Wu, J. Xu, Y. Yang, H. Zhang, and M. Zhu, "Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[14] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," in *The Thirteenth International Conference on Learning Representations*.

[15] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "Univla: Learning to act anywhere with task-centric latent actions," *arXiv preprint arXiv:2505.06111*, 2025.

[16] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *Robotics: Science and Systems*, 2025.

[17] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164

[18] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," in *8th Annual Conference on Robot Learning*.

[19] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen, "Video prediction policy: A generalist robot policy with predictive visual representations," in *Forty-second International Conference on Machine Learning*.

[20] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang, "Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies," in *The Thirteenth International Conference on Learning Representations*.

[21] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *Robotics: Science and Systems*, 2025.

[22] J. Cen, C. Yu, H. Yuan, Y. Jiang, S. Huang, J. Guo, X. Li, Y. Song, H. Luo, F. Wang, *et al.*, "Worldvla: Towards autoregressive action world model," *arXiv preprint arXiv:2506.21539*, 2025.

[23] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1702–1713.

[24] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," in *The Thirteenth International Conference on Learning Representations*.