When Robots Obey the Patch: Universal Transferable Patch Attacks on Vision-Language-Action Models

Supplementary Material

A. Proof for Proposition 1

Proof. By Assumption 1, there exists a matrix $A^\star \in \mathbb{R}^{d\times d}$ and a residual term $e(\cdot)$ such that, for every $\mathbf{x}$,

$$f_\pi(\mathbf{x}) = f_{\hat{\pi}}(\mathbf{x})A^\star + e(\mathbf{x}), \tag{21}$$

and for all pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ under consideration the residual difference is uniformly bounded:

$$\big\|e(\tilde{\mathbf{x}}) - e(\mathbf{x})\big\|_2 \;\le\; \varepsilon_E. \tag{22}$$

Step 1: Expressing the target deviation. For a fixed pair $(\mathbf{x}_i, \tilde{\mathbf{x}}_i)$, denote the residual difference by

$$\Delta\mathbf{e}_i := e(\tilde{\mathbf{x}}_i) - e(\mathbf{x}_i).$$

Using (21),

$$\begin{aligned}\Delta\mathbf{g}_i &= f_\pi(\tilde{\mathbf{x}}_i) - f_\pi(\mathbf{x}_i)\\ &= \big(f_{\hat{\pi}}(\tilde{\mathbf{x}}_i)A^\star + e(\tilde{\mathbf{x}}_i)\big) - \big(f_{\hat{\pi}}(\mathbf{x}_i)A^\star + e(\mathbf{x}_i)\big)\\ &= \big(f_{\hat{\pi}}(\tilde{\mathbf{x}}_i) - f_{\hat{\pi}}(\mathbf{x}_i)\big)A^\star + \big(e(\tilde{\mathbf{x}}_i) - e(\mathbf{x}_i)\big)\\ &= \Delta\mathbf{z}_i A^\star + \Delta\mathbf{e}_i.\end{aligned}$$

Step 2: Lower-bounding the ℓ_2 norm. Applying the reverse triangle inequality to $\Delta\mathbf{g}_i$ gives

$$\|\Delta\mathbf{g}_i\|_2 = \|\Delta\mathbf{z}_i A^\star + \Delta\mathbf{e}_i\|_2 \;\ge\; \|\Delta\mathbf{z}_i A^\star\|_2 - \|\Delta\mathbf{e}_i\|_2. \tag{23}$$

By the residual bound (22), we have $\|\Delta\mathbf{e}_i\|_2 \le \varepsilon_E$.

Next, recall the standard singular value inequality: for any $A^\star \in \mathbb{R}^{d\times d}$ and any vector $\mathbf{v} \in \mathbb{R}^d$,

$$\|\mathbf{v}A^\star\|_2 \;\ge\; \sigma_{\min}(A^\star)\,\|\mathbf{v}\|_2, \tag{24}$$

where $\sigma_{\min}(A^\star)$ is the smallest singular value of $A^\star$. Applying (24) with $\mathbf{v} = \Delta\mathbf{z}_i$,

$$\|\Delta\mathbf{z}_i A^\star\|_2 \;\ge\; \sigma_{\min}(A^\star)\,\|\Delta\mathbf{z}_i\|_2.$$

Combining this with (23) yields

$$\|\Delta\mathbf{g}_i\|_2 \;\ge\; \sigma_{\min}(A^\star)\,\|\Delta\mathbf{z}_i\|_2 - \varepsilon_E,$$

which is exactly (7).

Step 3: From ℓ_2 to ℓ_1 norms. We now derive a corresponding bound in ℓ_1. First, note that for any $\mathbf{v} \in \mathbb{R}^d$,

$$\|\mathbf{v}\|_2 \;\le\; \|\mathbf{v}\|_1, \tag{25}$$

and Hölder's inequality gives

$$\|\mathbf{v}\|_1 \;\le\; \sqrt{d}\,\|\mathbf{v}\|_2 \quad\Longrightarrow\quad \|\mathbf{v}\|_2 \;\ge\; \frac{1}{\sqrt{d}}\,\|\mathbf{v}\|_1. \tag{26}$$

Starting from (7) and using (25) on the left and (26) on the right, we obtain

$$\begin{aligned}\|\Delta\mathbf{g}_i\|_1 &\ge \|\Delta\mathbf{g}_i\|_2\\ &\ge \sigma_{\min}(A^\star)\,\|\Delta\mathbf{z}_i\|_2 - \varepsilon_E\\ &\ge \sigma_{\min}(A^\star)\,\frac{1}{\sqrt{d}}\,\|\Delta\mathbf{z}_i\|_1 - \varepsilon_E.\end{aligned}$$

This is precisely the claimed inequality (8).

Thus both bounds (7) and (8) hold, completing the proof.

B. Implementation Details

Implementation details. In all experiments, we optimize a square noise patch of size 50×50 pixels placed on RGB observations of size 224×224. The batch size is fixed to 2. For the perturbation-augmentation stage, we set the budget on the sample-wise noise to $\epsilon_\sigma = 2/255$, and adopt a nested optimization with $I = 8$ inner steps and $K = 50$ outer steps. The step sizes are $\eta_\sigma = 1/510$ for the sample-wise perturbations and $\eta_\delta = 1 \times 10^{-3}$ for the universal patch. For different values of ϵ_σ, the step size η_σ is set such that $I \times \eta_\sigma = 2 \times \epsilon_\sigma$. The three loss components are weighted by $\lambda_{\text{con}} = 10$, $\lambda_{\text{PAD}} = 1$, and $\lambda_{\text{PSM}} = 0.5$, respectively. We run the optimization for 2000 iterations in all settings and report the performance at the final iteration.

For the InfoNCE loss, we use a temperature $\tau = 0.07$. For the Patch Attention Dominance (PAD) term, we aggregate the last two text→vision attention layers, apply a non-patch weight of $\lambda_{\text{non}} = 0.8$, and restrict the attention reweighting to the top-$\rho = 0.3$ text tokens ranked by their clean attention mass. We further enforce a margin constraint such that the patch-induced attention increment exceeds the strongest non-patch increment by at least $m = 0.1$. For the Patch Semantic Misalignment (PSM) loss, we set $\alpha = 1.0$, $\beta = 0.5$, and use temperature $\tau = 0.3$ in the soft alignment terms. The sensitivity of our method to these hyperparameters is analyzed in Appendix E.

Table 4. We report the success rate (SR) on LIBERO simulation in a white-box setup. * marks an in-domain dataset matching the patch-training data, and $^{\triangle}$ marks a transfer evaluation on a different victim dataset.

Objective	Simulated					Physical				
	Spatial$^{\triangle}$	Object$^{\triangle}$	Goal$^{\triangle}$	Long*	Avg.	Spatial$^{\triangle}$	Object$^{\triangle}$	Goal$^{\triangle}$	Long$^{\triangle}$	Avg.
Benign	84.7 ± 10.2	88.4 ± 10.0	79.2 ± 12.0	53.7 ± 18.6	76.5	84.7 ± 10.2	88.4 ± 10.0	79.2 ± 12.0	53.7 ± 18.6	76.5
Random Noise	71.2 ± 24.2	85.2 ± 7.9	79.0 ± 15.5	51.6 ± 14.8	71.7	71.2 ± 24.2	85.2 ± 7.9	79.0 ± 15.5	51.6 ± 14.8	71.7
UMA_{1}	0.0 ± 0.0	1.0 ± 3.0	0.0 ± 0.0	0.0 ± 0.0	0.2	10.4 ± 15.3	39.2 ± 16.8	35.2 ± 26.8	20.0 ± 20.8	26.2
UMA_{1-3}	0.0 ± 0.0	1.2 ± 2.6	1.0 ± 2.4	0.0 ± 0.0	0.5	3.4 ± 6.2	43.6 ± 20.4	20.0 ± 20.5	18.0 ± 17.9	21.2
UADA_{1}	0.0 ± 0.0	0.8 ± 2.4	0.0 ± 0.0	0.0 ± 0.0	0.2	0.8 ± 1.8	1.2 ± 2.4	7.4 ± 15.2	3.4 ± 4.8	3.2
UADA_{1-3}	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0
UPA	3.8 ± 11.4	22.2 ± 12.5	12.0 ± 10.4	3.2 ± 3.0	10.3	4.4 ± 9.5	43.0 ± 23.6	42.8 ± 24.4	27.2 ± 19.7	29.3
TMA (Avg.)	0.8	13.6	11.1	2.0	6.9	9.9	48.9	41.3	21.1	30.3
Our	0.0 ± 0.0	0.0 ± 0.0	2.0 ± 0.0	0.0 ± 0.0	0.5	0.0 ± 0.0	9.0 ± 0.0	0.0 ± 0.0	2.0 ± 0.0	2.75

C. Main Results under White-box Setting

Tab. 4 evaluates task success rates in the LIBERO simulator under a white-box setup. Although our objective is explicitly designed for black-box transfer, it still shows competitive white-box performance. In the simulated setting, our patch almost completely disables the policy, driving success rates to near zero across all suites with an average of only 0.5%, on par with the strongest UMA/UADA variants and far below UPA and TMA (10.3% and 6.9% on average). In the physical setting, our method again reduces success to almost zero (2.75% on average), ranking second only to UADA_{1-3} and clearly outperforming UPA and TMA. These results indicate that the proposed universal patch retains strong white-box attack capability while being tailored for transfer.

D. Main Results on π_0

In the main text we reported transfer results from *OpenVLA-7B* [23] to *OpenVLA-oft-w* [24] and *OpenVLA-oft*. Tab. 5 complements these experiments by showing transfer to the π_0 [5]. This transfer is substantially harder, since π_0 differs from OpenVLA along almost every axis, including model architecture, pretraining pipeline, training data, and action head design, making cross-model transfer particularly challenging.

Even under this large surrogate to victim gap, our universal patch still achieves the strongest degradation in task success. In the simulated setting, the benign policy succeeds on 92.0% of tasks on average, whereas our method reduces the success rate to 86.0%, which is 2.5% percentage points lower than the best baseline (UADA_{1}, 88.5%). The advantage becomes even clearer in the physical setting: our average success rate of 83.50% is 5.50% points below the strongest baseline (89.0%), while other objectives stay closer to the benign performance. These results indicate that our feature and attention level design remains effective even when transferring from OpenVLA-7B to a structurally and procedurally very different VLA model, and highlight our superior transferability under the challenging physical to simulation cross-setting transfer.

E. Detailed Ablation Study

Impact of Patch Size. Tab. 6 ablates the patch area, varying it from 3% to 10% of the input image. We observe a clear monotonic trend: larger patches yield stronger attacks. A very small 3% patch already degrades performance compared to the baseline methods, but still leaves a high average success rate of 79.75%, indicating limited capacity to transfer attack. Increasing the size to 5%, our default choice, substantially strengthens the attack, reducing the average success rate to 61.50% while keeping the patch relatively compact and unobtrusive. When the patch occupies 7% or 10% of the image, the policy is almost completely disabled (39.00% and 20.75% on average), with object-centric success even dropping to 6% at 10%. This suggests that once the patch area is large enough to consistently intersect action-relevant regions, our feature and attention based objectives can fully dominate the visual stream. In practice, 5% offers a favorable trade-off between visual footprint and attack strength, while larger patches mainly amplify the effect rather than changing the attack behavior.

Impact of λ_{con}. Tab. 7 ablates the weight λ_{con} that balances the feature-space ℓ_1 term and the contrastive loss in our objective (both coefficients are rescaled by a factor of 0.1 during optimization for numerical stability). We observe that increasing λ_{con} from 1 to 10 steadily strengthens the attack, with the average success rate dropping from 63.75% to 61.50% and saturating once $\lambda_{\text{con}} \geq 5$. This trend is consistent with Tab. 2 and our theory that $\mathcal{L}_{\text{con}}$ primarily controls the *direction* of feature displacement, while $\mathcal{L}_1$ controls its magnitude: when λ_{con} is too small, the ℓ_1 term dominates and the patch mainly enlarges deviations without steering them into transferable directions; giving the contrastive term comparable or larger weight leads to more aligned, high-CCA feature shifts and thus better cross-

Table 5. Task success rate (%) when transfer from the surrogate OpenVLA-7B to the victim π_0 on LIBERO.

objective	Simulated					Physical				
	spatial	object	goal	long	avg.	spatial	object	goal	long	avg.
Benign	96	98	95	79	92.00	96	98	95	79	92.00
UMA_1	100	**94**	91	**72**	89.25	98	99	98	79	93.50
UMA_{1-3}	99	97	95	77	92.00	97	97	90	72	89.00
UADA_1	93	96	90	75	88.50	**93**	94	96	73	89.00
UADA_{1-3}	95	96	96	79	91.50	96	96	94	70	89.00
DOF_1	98	97	90	72	89.25	96	97	94	78	91.25
DOF_7	98	97	94	75	91.00	97	99	86	79	90.25
Our	**91**	96	**85**	**72**	**86.00**	**93**	**92**	**82**	**67**	**83.50**

Table 6. Ablation on patch size for transfer to openvla-oft in the physical setting.

Patch size	spatial	object	goal	long	avg.
3%	79	86	88	66	79.75
5%	69	74	76	27	61.50
7%	28	78	35	15	39.00
10%	41	6	35	1	20.75

Table 7. Ablation on λ_{con} for transfer to openvla-oft in the physical setting.

objective	spatial	object	goal	long	avg.
$\lambda_{\text{con}} = 1$	68	77	75	35	63.75
$\lambda_{\text{con}} = 2$	72	73	76	28	62.25
$\lambda_{\text{con}} = 5$	70	77	67	33	61.75
$\lambda_{\text{con}} = 10$	69	74	76	27	61.50

Table 8. Ablation on ϵ in RUPA for transfer to openvla-oft in the physical setting.

objective	spatial	object	goal	long	avg.
$\epsilon = 1/255$	73	73	77	28	62.75
$\epsilon = 2/255$	69	74	76	27	61.50
$\epsilon = 4/255$	66	71	62	33	58.00
$\epsilon = 8/255$	72	62	70	38	60.50
$\epsilon = 16/255$	75	67	68	36	61.50

model transfer. At the same time, the plateau between $\lambda_{\text{con}} = 5$ and 10 indicates that our method is not overly sensitive once the contrastive component is sufficiently emphasized.

Impact of ϵ in RUPA. Tab. 8 ablates the perturbation bound ϵ used for sample-wise inner minimization in Phase 1 of RUPA. Recall that these per-sample perturbations act as on-the-fly "hard" augmenters around each patched input. We observe that moderate noise levels yield the strongest transfer: increasing ϵ from 1 to 4 steadily lowers the average success rate from 62.75% to 58.00%, while further enlarging ϵ to 8 or 16 degrades performance again (60.5% and 61.5%).

This pattern suggests that RUPA behaves like a localized adversarial training loop around the universal patch. When ϵ is too small, the inner minimization explores only a narrow neighborhood and fails to expose the patch to sufficiently challenging geometric and appearance variations, limiting robustness. A moderate ϵ ($\epsilon = 4/255$) encourages the patch to align with features that remain effective within a realistic but nontrivial perturbation ball, leading to better transfer. However, overly large ϵ pushes samples far from the natural data manifold; the inner loop then overfits to unrealistic, heavily corrupted views, which weakens the invariances shared between surrogate and victim and ultimately harms black-box performance.

F. Real-world Performance

Beyond digital simulation, we qualitatively assess our adversarial patches in a physical robot setup under a black-box setting. We run repeated trials across three distinct tasks, including object grasping, placement, and manipulation, 3 times. As shown in Fig. 3, the patch reliably steers the robot to fail all tested executions. In the real world, each task failure represents a successful transfer attack on the black-box VLA model, highlighting the strong real-world transferability of our method. Detailed recordings are provided as videos in the supplementary material. From the videos, we observe that the attack is insensitive to patch location: across three qualitative trials, patches placed at different positions consistently cause the tasks to fail.

G. Training Video Visualisation

Figure 4 illustrates the training videos used for our universal patch optimization. The top row shows eight frames

Figure 3. **Qualitative real-world results.** The top row displays benign executions, while the bottom row shows their adversarial counterparts.

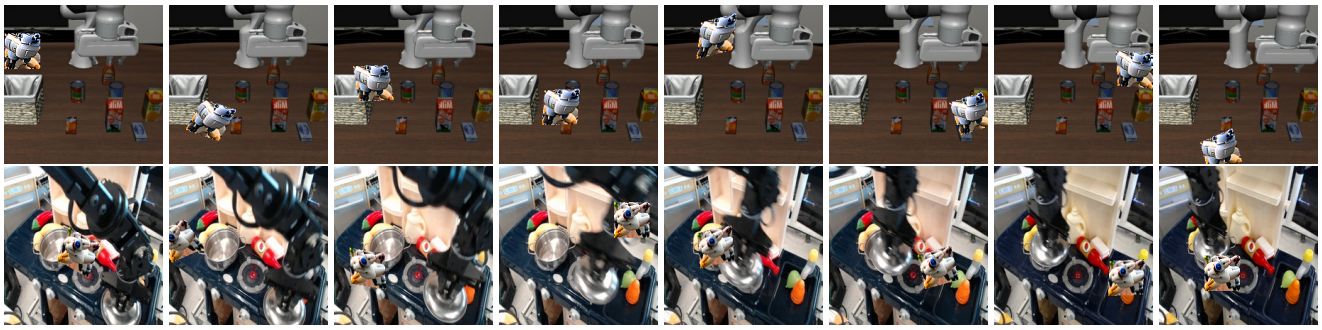

Figure 4. **Training videos from simulated and physical settings.** The top row shows eight frames sampled from a simulated training video, while the bottom row shows eight frames from a physical training video.

from a simulated setting, and the bottom row shows eight frames from a physical setting. In both rows, frames include sample-wise perturbations and patch geometric transformations (random position, skew, and rotation). The sample-wise perturbations are bounded by $\epsilon = 2/255$, making them imperceptible to the human eye and thus unlikely to affect real-world test performance. The patch geometric transformations follow the implementation of RoboticAttack [57]. Additional qualitative comparisons between our patch and the baseline patch on LIBERO are provided as videos in the supplementary material.

Why is physical transfer harder? Fig. 4 highlights a pronounced gap between simulated and physical training videos: the physical scenes exhibit richer clutter, stronger noise and motion blur, and more severe perspective distortions, leading to a much broader and more complex perceptual distribution. In simulation, actions are almost directly driven by visual tokens, so misguiding them quickly causes failure. Whereas on the real robot, trajectory smoothing and mechanical redundancy can partially compensate for perturbed decisions. These factors together make cross-setting transfer substantially harder and explain the larger performance gap between simulated and physical attacks.