# ACTION-STABLE HDR INTERFACES FOR FROZEN VISION-LANGUAGE-ACTION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision-language-action (VLA) policies fail when an exposure change or a bright light patch clips the camera image, even though the task is unchanged: an EV+3 shift takes OpenVLA-OFT, OpenVLA and $\pi_{0.5}$ from 82/90 to 0/90 successful episodes on LIBERO. We ask which camera interface preserves the actions of a frozen VLA and derive a minimum-intervention answer from the observation model. A common task weight cancels from the per-pixel choice of source, and clipping makes single-image recovery non-identifiable. With a second, shorter exposure under an ideal noise-free, unquantized two-level lighting model, four image operations follow: threshold the near-saturated pixels, close them to the smallest compatible lit region, dilate by one pixel to cover a one-pixel localization error, and write the recorded protected values inside. The resulting training-free interface, HDR-Guard, restores 82/90 under the EV+3 shift, where five single-image corrections reach at most 1/45. Under a sensor-aware simulation of a hard-edged light patch $8\times$ brighter than the room, it reaches 76/90 successes, against 70/90 for a soft saturation blend, 57/90 for full-frame Mertens fusion and 21/90 for the raw camera; with 500 episodes per rule, OpenVLA reaches 335/500 with HDR-Guard against 156/500 with Mertens fusion. Against the nominal frame, the lit-region error of HDR-Guard is 6.1 gray levels, against 15.7 for the soft blend; the 1.1% of the frame that the closing adds accounts for 75% of the reconstruction gain, and over 1,100 paired states, writing protected values there has an estimated +3.4-point equal-suite success effect (95% task-cluster interval $[-0.4, +7.7]$). In a simulation of trigger-free acquisition with the protected frame 50 ms old, OpenVLA reaches 362/500, against 336/500 without delay. On tripod-bracketed DSLR RAW frames, valid protected samples normalized by one calibrated global gain have a 4.8% median relative error against an independent short-exposure reference, and in 178,200 frames of real firefighting footage, 49% of frames with a person show a person with over 5% of their pixels near saturation.

## 1 INTRODUCTION

A frozen VLA policy acts on whatever image its camera delivers. When the exposure is wrong, or a patch of the workspace is lit much more brightly than the rest, the camera clips the bright pixels to a constant before the policy sees them. Clipping can remove the intensity variation of task-relevant regions such as the receptacle rim, the object boundary and the gripper contact point. On LIBERO (Liu et al., 2023), an EV+3 exposure shift takes three strong open policies, OpenVLA-OFT (Kim et al., 2025a), OpenVLA (Kim et al., 2025b) and $\pi_{0.5}$ (Black et al., 2025a), from 82/90 successful episodes to 0/90 (Figure 1), with the task, the instruction and the scene geometry unchanged. Clipping of task-relevant regions also occurs in real footage: in 297 clips (178,200 frames, 13 sessions) from an auto-exposure camera at firefighting drills, 49% of the frames that contain a person have a person with more than 5% of their pixels at or above 250/255 in the largest channel (Appendix H).

Correcting the clipped frame cannot undo this. When two scenes with different nominal actions produce the same clipped image, no function of that image can recover both actions. In our tests, five single-image corrections, including the exact inverse of the known gain, reach at most 1/45 successes. A second, shorter exposure does record the missing signal, and HDR imaging has merged bracketed exposures for decades (Debevec & Malik, 1997; Mertens et al., 2007; Hasinoff et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

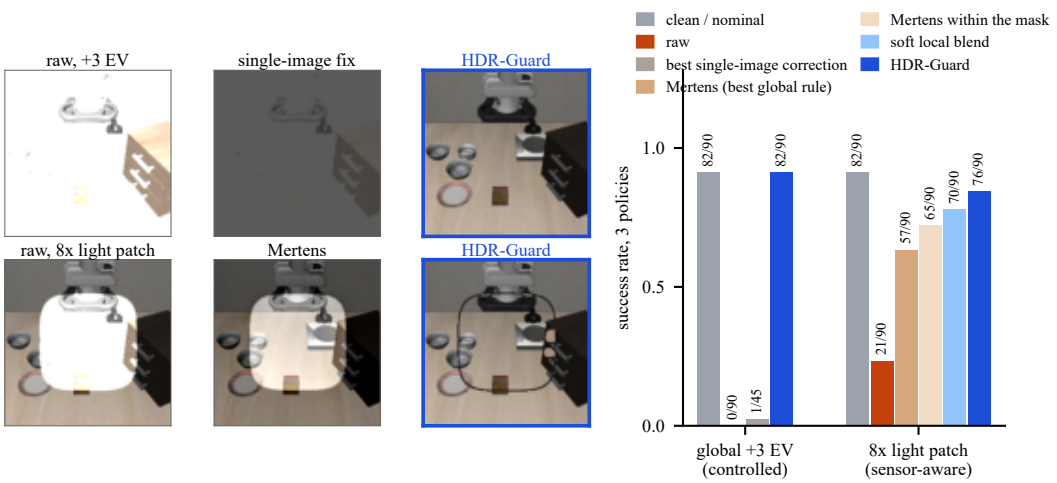


**Figure 1: Clipping removes task evidence before a frozen VLA sees the frame.** Left, top: an EV+3 exposure shift, a representative single-image correction (gamma after the known-gain inverse) and HDR-Guard. Left, bottom: a hard-edged light patch 8× brighter than the room under the sensor-aware simulation, the best global HDR rule (Mertens exposure fusion) and HDR-Guard. Right: closed-loop successes summed over OpenVLA-OFT, OpenVLA and $\pi_{0.5}$ on LIBERO-Spatial, including Mertens fusion written only inside the mask of HDR-Guard and the soft saturation blend; these local rules keep the policy stream outside a clipping-derived region.

2016). Most HDR pipelines aim at a faithful radiance map or a pleasing displayed image, and full-frame exposure fusion and tone mapping rewrite every pixel, including pixels that were already correct. A frozen policy was trained on nominally exposed frames, so its target is the frame its camera would have recorded under nominal conditions and the action it would take on it.

We therefore treat exposure robustness as an interface problem with an action-level objective: the reference is the frozen policy's own action on the nominal frame, which needs no action labels. Starting from this objective, we derive a minimum-intervention recorded-source interface from a two-level observation model and a spatial coverage constraint (Section 3). A task-relevance weight cancels from the per-pixel choice of source (Proposition 1), so the interface needs no model of the policy. No single-image repair can recover scenes that clipping makes indistinguishable (Proposition 2); a second, shorter exposure supplies the missing measurement. With both streams and a two-level lighting model, the morphological closing of the near-saturated pixels is the smallest lit region compatible with the data, its one-pixel dilation is the smallest cover robust to a one-pixel localization error, and the recorded protected values equal the nominal ones on the true lit region (Proposition 3). The result, HDR-Guard, is four image operations, and Proposition 4 bounds its residual error by four terms: bracket mismatch, noise, misregistration and protected saturation.

Our contributions are:

- **An action-level target for HDR sensing.** We formulate exposure repair for a frozen VLA as preservation of its nominal input and action rather than recovery of scene radiance or display quality, and instantiate it with HDR-Guard, a training-free interface that returns $H/g$ (the input at $g = 1$) for an empty mask and exposure selection under full clipping.
- **A conditional derivation, identifiability and an error budget.** Under the idealized assumptions of Proposition 3, the closing is the smallest observation-compatible regular region and, with constant likelihood and a prior strictly decreasing in area, the maximum a posteriori region; a one-pixel robust-coverage constraint yields the unique minimum-area dilation. When clipping maps scenes with different nominal actions to one image, no single-image repair can recover both; with a protected exposure, the remaining error splits into four terms, and a common task weight cancels from the pointwise source decision.
- **Mechanism-resolved evidence with three frozen VLAs.** On the sensor-aware hard patch, HDR-Guard reaches 76/90 successes, against 70/90 for the soft blend, 65/90 for localized Mertens, 57/90 for full-frame Mertens and 21/90 raw. Across 400 frames, the pixels the closing adds carry 75% of its reconstruction advantage over the soft blend, and

a 1,100-state, four-suite OpenVLA intervention separates their closed-loop effect (+3.4 [−0.4, +7.7] points) from source assignments elsewhere (−1.6 [−5.2, +1.9]).

## 2 Related Work

**Vision-language-action policies and their robustness.** RT-1 and RT-2 established large robot transformers and transfer from web-scale vision-language models (Brohan et al., 2023; Zitkovich et al., 2023). OpenVLA released an open 7B VLA (Kim et al., 2025b), OpenVLA-OFT added parallel decoding and action chunking (Kim et al., 2025a), and $\pi_0/\pi_{0.5}$ introduced flow-matching action experts (Black et al., 2025b;a); see also Octo and RDT (Ghosh et al., 2024; Liu et al., 2025). Robustness studies show that high clean LIBERO scores hide sensitivity to camera, lighting and texture shifts (Fei et al., 2026; Jin et al., 2026; Lu et al., 2026). The usual remedies change the policy through more data, augmentation or adaptation. HDR-Guard changes what the policy receives and leaves every weight fixed, so it composes with any of these policies.

**HDR imaging, exposure fusion and tone mapping.** Radiance recovery from bracketed exposures (Debevec & Malik, 1997; Mitsunaga & Nayar, 1999), noise-optimal weighting (Granados et al., 2010; Hasinoff et al., 2010), burst merging (Hasinoff et al., 2016) and learned deghosting (Kalantari & Ramamoorthi, 2017) target the radiance map; exposure fusion (Mertens et al., 2007) and tone mapping (Reinhard et al., 2002) target a displayable image. Algorithmically, HDR-Guard is a hard saturation-aware merge, a special case of this classical family. Its contribution is the target: the output must reproduce the nominal frame a fixed policy was trained on, scored by the policy's action rather than PSNR or perceptual quality, and Proposition 3 derives it.

**Task-driven imaging.** Camera pipelines optimized for a downstream network (Diamond et al., 2021) and learned auto-exposure (Onzon et al., 2021) and exposure fusion (Onzon et al., 2024) for HDR detection tune the imaging stack for a perception model, and UltraFusion (Chen et al., 2025) inpaints highlights from a shorter exposure for display; all train a network, while we keep the policy frozen.

## 3 Deriving the Interface from the Observation Model

We derive a minimum-intervention rule from the observation model and a spatial coverage constraint, in four steps; proofs are in Appendix A.

**Observation model.** Let $x \in [0,1]^{P\times 3}$ be the *nominal frame*: what the policy camera records under the exposure and lighting the policy was trained on. The frozen policy's reference action is $a^\star = \pi(x)$. At deployment the irradiance reaching the sensor is multiplied by a brightening $b(p) = g\,\ell(p) \geq 1$, where $g$ is the camera's exposure offset from its nominal setting (known from metadata) and $\ell(p)$ an unknown change in lighting. A global exposure shift has $g = 2^k$ and $\ell \equiv 1$; a bright light patch has $g = 1$ and $\ell(p) > 1$ on the patch. The policy stream and a protected stream exposed $r > 1$ times shorter record, per channel,

$$H = \mathrm{clip}_{[0,1]}\big(b \odot x + n_H\big), \qquad L = \mathrm{clip}_{[0,1]}\big(b \odot x/r + n_L\big), \tag{1}$$

where $n_H, n_L$ are zero-mean sensor noise with photon and read components (Foi et al., 2008), followed by 8-bit quantization. We write $C = \{(p,c) : b(p)x(p,c) \geq 1\}$ for the over-range channels of the policy stream.

**Step 1: nominal-input preservation as a per-pixel choice of recorded source.** The interface must produce $\hat{x}$ with $\pi(\hat{x}) \approx a^\star$ without access to $\pi$, so it preserves the nominal input as a policy-agnostic surrogate. With a relevance weight $W(p)$ standing in for the policy's sensitivity, each output value is either kept from the policy stream or written from the protected one.

**Proposition 1** (The task weight cancels)**.** *For a scalar channel with observed policy value $h$, let $R_K(h)$ and $R_R(h)$ be the posterior risks of keeping $H/g$ and of writing $L$ under the pixel-separable loss $W(p)\,(\hat{x}-x)^2$. For every $W(p) > 0$ the preferred source does not depend on $W(p)$. If $R_K - R_R$ is nondecreasing in $h$ and changes sign once, the preferred source is a fixed threshold on $h$.*

Within this pixel-separable two-source loss, a relevance weight known from the observation therefore cannot move the selection boundary: the question is which recorded value to pass at each pixel, $\hat{x}(p) \in \{H(p)/g,\ L(p)\}$, and under single crossing the answer is a threshold on the policy value.

**Step 2: one stream cannot make the choice.**

**Proposition 2** (Single-image repair is non-identifiable)**.** *Let two nominal frames $x, x'$ agree outside $C$ and satisfy $b\,x \geq 1$ and $b\,x' \geq 1$ on $C$. Without noise, $H(x) = H(x')$. Hence for any single-image repair $f$, deterministic policy $\pi$ and action metric $d$,*

$$\max_{s \in \{x, x'\}} d\big(\pi(f(H(s))),\, \pi(s)\big) \;\geq\; \tfrac{1}{2}\, d\big(\pi(x), \pi(x')\big),$$

*and for discrete actions the composed rule $\pi \circ f$ errs on at least one of the two scenes whenever $\pi(x) \neq \pi(x')$.*

The bound covers exposure inverses, tone curves and learned enhancers alike; with noise it becomes a posterior Bayes-risk bound (Appendix A). Clipping is a censored measurement (Tobin, 1958), so recovery cannot be guaranteed from the clipped image alone, and we acquire a second, shorter exposure.

**Step 3: with two streams, the minimum-intervention choice reduces to four image operations.** Which recorded value is right depends on the lighting only through one unknown set, and after correction by the known gain the policy stream is nominal outside the lit region, while spatial regularity identifies lit pixels even where they stay unclipped. We therefore seek the smallest intervention: replace as few pixels as possible while protecting every pixel the data certify as lit, even under a one-pixel error in localizing them. We model the brightening as two-level, $b = r$ on an unknown lit region $\Omega$ and $b = g$ elsewhere: a light patch matched by the bracket in a room at the known camera offset, with a global shift as the case $g = r$. Two assumptions make $\Omega$ estimable. *Headroom:* $g \max_c x(p, c) < \tau$ outside $\Omega$. *Regularity:* $\Omega$ has no holes or gaps narrower than a symmetric structuring element $B$, that is, $\mathrm{close}_B(\Omega) = \Omega$; every closed convex region qualifies, and both hard patches used here are exact fixed points on the pixel grid (Appendix A).

**Proposition 3** (The lighting model determines the rule)**.** *Let $S_H = \{p : \max_c H(p, c) \geq \tau\}$ and assume the two-level model, headroom and regularity, without sensor noise, quantization or misregistration. (a) Given $\Omega$, writing $L$ on $\Omega$ and $H/g$ elsewhere returns $x$ exactly; where the two recorded values differ, the other choice does not. (b) $S_H \subseteq \Omega$. For every regular region $\Omega' \supseteq S_H$ there is a scene, depending on $\Omega'$, that satisfies headroom and reproduces both streams exactly. $\mathrm{close}_B(S_H)$ is the smallest such region and the intersection of all of them, that is, the set of pixels lit under every explanation of the data. If the marginal likelihood of $(H, L)$ is constant over these regions and the region prior is a strictly decreasing function of area alone, it is the unique maximum a posteriori region. (c) Let $\hat{\Omega} = \mathrm{close}_B(S_H)$ and $N = \{-1, 0, 1\}^2$. Among replacement regions that contain every one-pixel translate $\hat{\Omega} + \delta$, $\delta \in N$, the unique smallest is $\mathrm{dilate}_{3\times3}(\hat{\Omega})$. If the translate is uniform on $N$ and the loss is pixel-separable, the same region minimizes the expected loss wherever missing a lit pixel costs at least eight times a false replacement, which squared error guarantees for pixels unclipped when lit whenever $r/g \geq 2\sqrt{2}$.*

The minimum-intervention region of (b) and (c), with the source assignment of (a), is HDR-Guard:

$$\begin{aligned} R_H &= \mathrm{dilate}_{3\times3}\big(\,\mathrm{close}_B(S_H)\big), \qquad \tau = 0.985, \\ \hat{x}(p) &= L(p) \;\text{ if } p \in R_H, \qquad \hat{x}(p) = H(p)/g \;\text{ otherwise,} \end{aligned} \tag{2}$$

with $B$ a $15 \times 15$ ellipse. Each operation is a step of the derivation. The seed takes the largest channel because the brightening is shared by a pixel's channels, so one near-saturated channel certifies the pixel. The closing adds the unclipped pixels that the seed encloses, such as a dark surface inside a light patch, which receive the same brightening. The dilation is the smallest region that still covers the certified core under a one-pixel localization error from thresholding, resizing or registration; under uniform offsets it is also expected-loss optimal wherever a miss costs at least eight false replacements, which squared error gives (64 at $r = 8$, $g = 1$) for nonzero pixels unclipped when lit. The protected value is written as recorded because it equals $x$ wherever $b = r$. Under the same model the implemented rule is exact except on lit pixels outside $R_H$, where it errs by $(r/g - 1)\,x$, and unlit pixels inside $R_H$, where it errs by $(g/r - 1)\,x$; for a global shift ($g = r$) both terms vanish. The scale of $B$ and $\tau$ are the two free parameters; the action diagnostic varies little over closing sizes of 7–21 pixels (Appendix R). When $R_H$ is empty, $\hat{x} = H/g$: an exact pass-through for $g = 1$ and the intended gain correction for a known exposure offset (isolated noise spikes above $\tau$ flag 0.01% of the pixels of a well-exposed LIBERO frame). When everything clips, $\hat{x} = L$, so the rule reduces

to exposure selection without a separate regime detector. The kernel is one channel maximum, a closing, one $3 \times 3$ max-pool and one select: 0.27 ms per $256 \times 256$ frame on an RTX 4060 Ti and 2.0 ms on one CPU core.

**Step 4: what remains is an error budget.** Outside the ideal model, four measurable terms remain.

**Proposition 4** (Error budget of the repaired frame)**.** *Let the protected stream be registered to the policy stream up to a displacement $\delta(p)$ along which $bx/r$ is continuously differentiable, and let $\epsilon_Q \le 1/510$ be the 8-bit rounding error. For each channel,*

$$|\hat{x} - x| \le \begin{cases} \underbrace{|b/r - 1|\, x}_{\text{bracket mismatch}} + \underbrace{|n_L| + \epsilon_Q}_{\text{noise}} + \underbrace{\sup_{t\in[0,1]} \|\nabla(bx/r)(p + t\delta)\|\, \|\delta\|}_{\text{misregistration}}, & p \in R_H,\ L(p,c) < 1, \\ \underbrace{|1 - x|}_{\text{protected saturation}}, & p \in R_H,\ L(p,c) = 1, \\ |\ell - 1|\, x + (|n_H| + \epsilon_Q)/g, & p \notin R_H. \end{cases}$$

*If, in addition, $\pi$ is $L_\pi$-Lipschitz for $\|e\|_W = \sum_{p,c} W(p,c)|e(p,c)|$, then $\|\pi(\hat{x}) - a^\star\| \le L_\pi \|\hat{x} - x\|_W$.*

Each term can be varied on its own (Section 4.4), and Appendix I quantifies exposure-ratio mismatch, noise and saturation in real bracketed RAW captures. Relative to the policy stream at an unclipped pixel, a protected value rescaled by $r$, as in rescaled exposure selection, has $\sqrt{r}$ times its shot-noise and $r$ times its read-noise standard deviation, motivating selective use of the protected stream.

## 4 Experiments

**Policies and tasks.** We evaluate three frozen open policies with their official LIBERO checkpoints and evaluation code: OpenVLA-OFT (third-person and wrist views, 8-step action chunks), OpenVLA (third-person view, one action per step) and $\pi_{0.5}$ (both views, 5-step replanning). Each condition runs LIBERO-Spatial tasks 0–9 with three trials per task and the benchmark's fixed initial states, giving 30 episodes per policy and 90 in total. Success is the official task predicate. Only the image transform differs from the clean run; it is applied to both camera views before each policy's own resizing (for $\pi_{0.5}$, to the $224 \times 224$ frames its client sends; Appendix C). Because the conditions share tasks and initial states, we compare rules with a paired task-cluster bootstrap: the ten tasks are resampled with replacement and the per-task success differences summed over the three policies (95% intervals from the exact empirical bootstrap distribution: all 92,378 distinct task-count vectors, each weighted by its multinomial multiplicity, which is equivalent to all $10^{10}$ ordered resamples).

**Two protocols.** The *sensor-aware simulation* (Section 4.1) exposes both streams from the same scene irradiance with photon shot noise (1000 $e^-$ full well), 2 $e^-$ read noise, clipping and 8-bit quantization (Appendix B). The *controlled identifiability test* (Section 4.3) uses a noise-free matched bracket, under which the protected stream equals the nominal frame, to measure what clipping costs.

### 4.1 Mixed illumination under a sensor-aware simulation

Mixed illumination separates spatially selective replacement from global exposure selection, which suffices for a global shift. We add a light patch $8\times$ brighter than the room, a rounded superellipse covering the center of both views (about 23% of the frame; Appendix B), with either a near-discontinuous edge or a smooth, spotlight-like falloff. The policy camera stays at its nominal exposure and the protected camera uses a 3-stop bracket, so $b = r$ on the patch interior and $b < r$ on a soft edge. The hard edge is the matched operating point of Equation (2); the soft edge, and patches $6\times$ and $12\times$ brighter with the same fixed bracket, test bracket mismatch. After the rule was fixed we added one confirmation scene that played no part in any design choice: a smaller hard-edged patch placed off center (Appendix B). We compare nine ways to build the policy's input from two noisy streams generated by the same sensor model (each rule draws its own noise realization): the policy camera as recorded (raw); *exposure selection*, which feeds the protected frame as recorded or rescaled by $r$ to nominal brightness; full-frame *Mertens exposure fusion* (Mertens et al., 2007); *HDR*

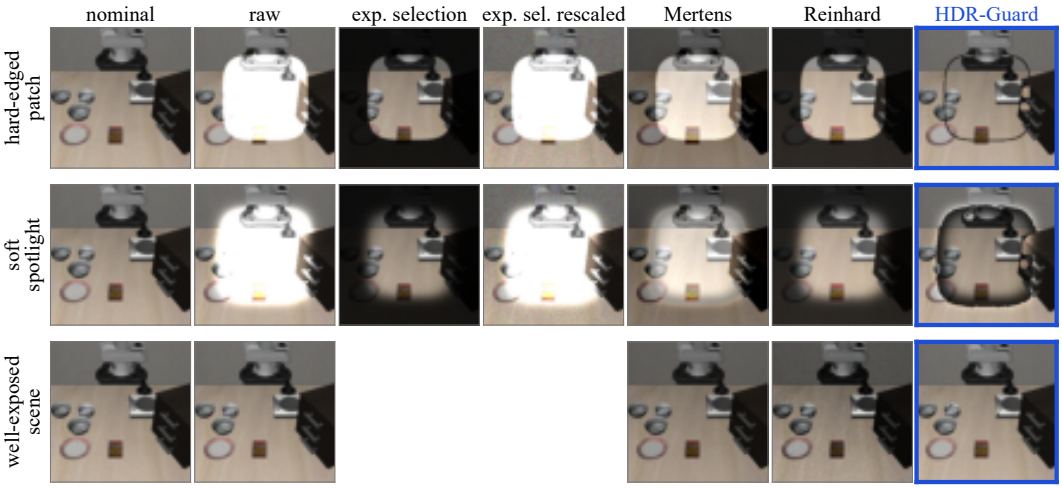


**Figure 2: What each rule feeds the frozen policy under the sensor-aware simulation** (LIBERO-Spatial task 0, third-person view). All rules receive noisy streams from the same sensor model. Exposure selection leaves the room $8\times$ too dark or, rescaled, re-clips the patch and amplifies noise; Mertens fusion and Reinhard tone mapping change every pixel, including in the well-exposed scene; HDR-Guard writes the protected value inside a mask seeded by near-saturated pixels, closed and dilated. On the soft spotlight the protected value is darker than nominal where the light falls off ($b < r$), the bracket-mismatch term of Proposition 4.

**Table 1: Sensor-aware simulation.** Closed-loop successes on LIBERO-Spatial tasks 0–9, three trials per task: per policy out of 30, totals out of 90 (sum over the three policies). $G$ is the patch brightness relative to the room and $r$ the bracket. The offset column is a smaller hard-edged patch at another position, chosen after the rule was fixed; rules not run on it are marked –. The well-exposed column uses the same bracket without a patch; its raw row is the nominal-exposure reference. The global column overexposes the policy camera by 3 stops, with the protected camera at nominal exposure. Best total per column in bold; per-policy counts are in Table 10, and the bracket-mismatch scenes in Table 11.

| | Hard-edged patch, $G = r = 8$ | | | | Offset | Well- | Global |
|---|---|---|---|---|---|---|---|
| Input to the frozen policy | OFT | OpenVLA | $\pi_{0.5}$ | Total | patch | exposed | EV+3 |
| Policy camera (raw) | 9 | 0 | 12 | 21 | 31 | 82 | 0 |
| Exposure selection, as recorded | 20 | 1 | 14 | 35 | – | – | 84 |
| Exposure selection, rescaled | 5 | 2 | 6 | 13 | – | – | – |
| Mertens exposure fusion | 28 | 10 | 19 | 57 | 69 | 83 | 82 |
| HDR merge + Reinhard | 23 | 5 | 15 | 43 | – | **87** | – |
| Localized Mertens (within $R_H$) | 27 | 15 | 23 | 65 | 69 | – | – |
| Soft saturation blend | 29 | 13 | 28 | 70 | **85** | 83 | – |
| HDR-Guard, no closing (ablation) | 29 | 14 | 26 | 69 | 81 | 82 | **85** |
| HDR-Guard | 29 | 19 | 28 | **76** | 83 | 84 | 82 |

*merge with Reinhard tone mapping* (Reinhard et al., 2002); *localized Mertens*, the same fusion composited only inside the mask $R_H$ of HDR-Guard; a *soft saturation blend* ramping from the policy to the protected stream as the policy value rises from 0.90 to $\tau$; and HDR-Guard, with and without the region closing. A well-exposed scene with the same bracket and no patch measures whether a rule harms scenes that needed no repair.

Table 1 reports closed-loop successes. On the hard-edged patch the policy camera alone gives 21/90 successes, and OpenVLA fails every episode. Exposure selection leaves the room eight times too dark (35/90), and rescaling the protected frame re-clips the patch and amplifies its noise (13/90). Mertens fusion (57/90) and Reinhard tone mapping (43/90) keep the patch visible but rewrite every pixel of the frame; enabling Mertens' well-exposedness weight, zero by default in OpenCV, gives 148/500 against 335/500 for HDR-Guard on 500 OpenVLA states (Appendix B). The rules that keep the policy stream outside a clipping-derived region all exceed the global rules: HDR-Guard reaches 76/90, the soft saturation blend 70/90, HDR-Guard without the closing 69/90 and localized Mertens 65/90. Against Mertens fusion, HDR-Guard gains +19 [+10, +28] successes (paired task-cluster bootstrap, 95% interval); against the soft blend it is +6 [+0, +13]. The closing was introduced

on this condition during development. It changes OpenVLA from 14/30 to 19/30 (24/30 at nominal exposure): without it, dark surfaces inside the patch keep their eightfold-brightened values, the kept-pixel term of Proposition 4. For the other two policies it changes at most two episodes. Under the noisy global EV+3 shift the raw camera gives 0/90, while exposure selection, Mertens fusion and HDR-Guard recover 84/90, 82/90 and 82/90.

**Reconstruction of the nominal frame.** The derivation predicts an error from lit pixels omitted by the replacement mask: in the ideal matched model, retaining $H/g$ leaves them $r/g$ times too bright, and the closing extends replacement to enclosed pixels that an intensity-only selector can miss. On 400 rendered frames from all four LIBERO suites under the hard patch, HDR-Guard's mean absolute error in the lit region is 6.1 gray levels (nominal-exposure noise reference 4.1), against 15.7 for the soft blend, 11.5 without the closing and 76.6 for full-frame Mertens fusion; the difference from the soft blend is $-9.6$ [$-11.7$, $-7.4$] (40-task cluster bootstrap). On the lit pixels the seed misses (9% of the patch), HDR-Guard reaches 11.9 against 118.9 for the soft blend, and the pixels that the closing adds, $V$ (1.1% of the frame), carry 75% of the whole-frame error reduction, with errors of 20 against 115 gray levels there (Appendix J). A $2 \times 2$ closed-loop intervention on 1,100 matched OpenVLA initial states estimates an equal-suite success effect of +3.4 [$-0.4$, +7.7] points (95% task-cluster interval) from writing protected rather than soft-blend values on $V$, positive in three of the four suites, against $-1.6$ [$-5.2$, +1.9] for the values outside $V$ (Appendix K).

**Confirmation runs.** On an offset patch fixed after the rule was locked, HDR-Guard, the soft blend and HDR-Guard without closing reach 83/90, 85/90 and 81/90, against 69/90 and 69/90 for full-frame and localized Mertens and 31/90 for the raw camera. All three local rules lead both Mertens variants (HDR-Guard minus full-frame Mertens: +14 [+7, +20]). Because each rule draws its own sensor noise, we repeated the centered patch for the three local rules with common random numbers, so that every rule sees the same noise on the same state (Appendix B): 79/90, 77/90 and 70/90. Resampling tasks with both runs summed within each task, HDR-Guard minus the soft blend is +8 [$-3$, +21] and minus its no-closing ablation +16 [+5, +28]. For OpenVLA, chosen afterwards because the local rules differ most on it, 50 trials per task (500 episodes per rule, Appendix F) give 335/500 for HDR-Guard, 305/500 for the soft blend and 289/500 without the closing (nominal exposure 422/500); on the same initial states, HDR-Guard alone succeeds on 89 and the soft blend alone on 59. On LIBERO-Object, Goal and LIBERO-10 with per-suite checkpoints, OpenVLA reaches 333/600 with HDR-Guard, 331/600 with the soft blend, 149/600 with full-frame Mertens fusion and 15/600 raw; and the difference from the soft blend is +24 [+2, +49] on Goal, +3 [$-17$, +29] on LIBERO-10 and $-25$ [$-44$, $-9$] on Object (Appendix F.1), in the order of the area the closing adds on each suite (Appendix K). To separate where a rule writes from what it writes, localized Mertens composites Mertens fusion only inside $R_H$ and keeps the policy stream elsewhere. On the matched patches it reaches 65/90 (centered) and 69/90 (offset), and HDR-Guard exceeds it by +11 [+3, +19] and +14 [+10, +18] (+15 [+5, +26] in the common-random-number run): inside the region, writing the protected exposure itself matters, not only restricting where the frame changes. Under mismatch, restricting Mertens to $R_H$ lowers it from 76/90 to 54/90 on the soft spotlight and from 73/90 to 59/90 at $G = 6$, removing much of its advantage over HDR-Guard. This is consistent with the kept-pixel term of Proposition 4 outside the mask, although the intervention also introduces a hard compositing boundary. Appendix M repeats the hard patch with $\pi_{0.5}$ on three further LIBERO suites. On the well-exposed scene all rules obtain 82–87/90 (nominal: 82/90).

**Operating regime.** Without noise, quantization or clipping, $H = bx$ and $L = bx/r$ identify the shared product $bx$, and Proposition 3 adds the two-level assumptions. The global shift is two-level exactly and the hard patch approximately. A soft spotlight brightens by $1 < b < r$ across its falloff, where the written value $bx/r$ is darker than nominal; there Mertens fusion obtains 76/90, against 48/90 for the soft blend and 45/90 for HDR-Guard (Table 11 in Appendix E). The local rule targets sharp, bracket-matched light; on the soft spotlight, full-frame Mertens leads.

### 4.2 Paired action consistency

Because an episode can succeed after a detour, we also measure the quantity HDR-Guard targets, the change in the frozen policy's action. We roll out each policy on clean inputs (LIBERO-Spatial tasks 0–9, one trial each) and, at the policy's query states, present the same state through every input rule. For the first returned action we report $D_1$, the mean over the six arm dimensions of $|a_j - a_j^\star|$ divided by the 1st–99th percentile range of the clean actions, and the fraction of queries

**Table 2: Paired action consistency**, a local diagnostic on nominal-rollout states. The frozen policy is queried on the same states through each input rule. $D_1$: normalized deviation of the first action from the clean-input action ($\times 100$, averaged within each of the ten rollouts and then across rollouts; lower is better); grip.: gripper commands flipped (%). The first row measures re-query variability, the second the deviation caused by sensor noise alone. Rollout-cluster bootstrap intervals are in Appendix P. †Lowest $D_1$ among the single-image corrections evaluated in this diagnostic (Appendix P), chosen separately for each policy.

| | OpenVLA-OFT | | OpenVLA | | $\pi_{0.5}$ | |
|---|---|---|---|---|---|---|
| Input on the same state | $D_1$ | grip. | $D_1$ | grip. | $D_1$ | grip. |
| Clean input, queried again | 0.00 | 0 | 0.00 | 0 | 2.19 | 0 |
| Nominal exposure with sensor noise | 0.60 | 0 | 3.05 | 0 | 2.60 | 0 |
| *Hard-edged patch* | | | | | | |
| Policy camera (raw) | 8.32 | 1 | 16.19 | 27 | 8.28 | 1 |
| Exposure selection, as recorded | 7.68 | 0 | 13.22 | 6 | 6.89 | 3 |
| Mertens exposure fusion | 4.81 | 1 | 9.63 | 7 | 5.66 | 0 |
| HDR merge + Reinhard | 6.80 | 1 | 12.04 | 8 | 6.44 | 2 |
| Soft saturation blend | 3.41 | 1 | 8.41 | 4 | 4.74 | 0 |
| HDR-Guard, no closing | 3.39 | 2 | 9.00 | 2 | 4.56 | 0 |
| HDR-Guard | 3.32 | 1 | 8.37 | 4 | 4.75 | 0 |
| *Well-exposed scene, same bracket* | | | | | | |
| Mertens exposure fusion | 0.59 | 0 | 2.30 | 0 | 2.66 | 0 |
| HDR merge + Reinhard | 1.03 | 1 | 3.72 | 2 | 3.14 | 1 |
| HDR-Guard | 0.72 | 0 | 3.19 | 1 | 2.59 | 1 |
| *Controlled EV+3 shift* | | | | | | |
| Policy camera (raw) | 11.67 | 3 | 19.54 | 39 | 18.01 | 16 |
| Best evaluated single-image correction† | 11.55 | 5 | 17.76 | 40 | 18.19 | 12 |
| HDR-Guard | 0.00 | 0 | 0.00 | 0 | 2.10 | 0 |

whose gripper command flips. Table 2 uses 131 OpenVLA-OFT, 110 OpenVLA and 219 $\pi_{0.5}$ query states. OpenVLA-OFT and OpenVLA are deterministic; $\pi_{0.5}$'s flow sampler draws fresh noise at every query, so re-querying its clean input sets its floor.

On the hard patch the paired metric broadly agrees with the closed-loop ordering. In the controlled EV+3 test the repaired input is the nominal input: the deviation is exactly zero for the two deterministic policies and stays at $\pi_{0.5}$'s sampling floor (2.10, against 2.19 for re-querying the clean input). The clipped frame ($D_1 = 11.67$, 19.54 and 18.01 for OpenVLA-OFT, OpenVLA and $\pi_{0.5}$) and the best single-image correction (11.55, 17.76 and 18.19) stay far above the noise floors in the first two rows. On the hard-edged patch, HDR-Guard, the soft blend and the no-closing ablation form the lowest-deviation group among the rules in this diagnostic (localized Mertens is measured in a separate pass, Appendix Q); HDR-Guard has the lowest point estimate for the two deterministic policies (3.32 and 8.37) and the no-closing ablation for $\pi_{0.5}$. Mertens fusion, the closest of the other rules, gives 4.81, 9.63 and 5.66. The raw frame flips OpenVLA's gripper command in 27% of queries; HDR-Guard in 4%. On the well-exposed scene the three rules in Table 2 stay near the sensor-noise reference. Because $\pi_{0.5}$ draws fresh sampler noise at every query, its column is descriptive. Closing sizes of 7–21 and ramp starts of 0.85–0.97 move OpenVLA-OFT's $D_1$ by at most 0.2 (Appendix R).

### 4.3 Controlled identifiability test: global exposure shift

This protocol is a controlled diagnostic rather than an HDR comparison. With a matched noise-free bracket, the protected stream equals the nominal frame and Equation (2) returns $\hat{x} = x$ exactly, for every threshold and with or without the closing and dilation. The experiment therefore measures how much clipping alone costs each frozen policy, confirms end to end that the pipeline reproduces the nominal input, and tests whether five single-image corrections of the clipped frame recover the lost behavior.

Table 3 shows that clipping alone is catastrophic for all three policies. Raw EV+3 frames give 0/90 on Spatial and 2/135 on the other three suites; the two raw successes are OpenVLA-OFT episodes on Object. Restoring the measurement recovers the deterministic policies' clean counts exactly.

**Severity and single-image repair.** EV+1 is mild, EV+2 a transition and EV+3–EV+4 complete failures (Figure 3a). Of five repairs that see only the clipped frame (Figure 3b), the known-gain

**Table 3: Controlled matched-bracket test, global EV+3 shift.** Closed-loop successes, three trials per task. By construction the repaired input equals the nominal input; the table measures the behavioral cost of clipping. Differences between the clean and repaired columns come only from $\pi_{0.5}$, whose flow sampler draws fresh noise in every run.

| | LIBERO-Spatial, tasks 0–9 | | | Object + Goal + LIBERO-10, tasks 0–4 | | |
|---|---|---|---|---|---|---|
| Policy | Clean | Raw | HDR-Guard | Clean | Raw | HDR-Guard |
| OpenVLA-OFT | 30/30 | 0/30 | 30/30 | 45/45 | 2/45 | 45/45 |
| OpenVLA | 23/30 | 0/30 | 23/30 | 30/45 | 0/45 | 30/45 |
| $\pi_{0.5}$ | 29/30 | 0/30 | 29/30 | 44/45 | 0/45 | 42/45 |
| Total | 82/90 | 0/90 | 82/90 | 119/135 | 2/135 | 117/135 |

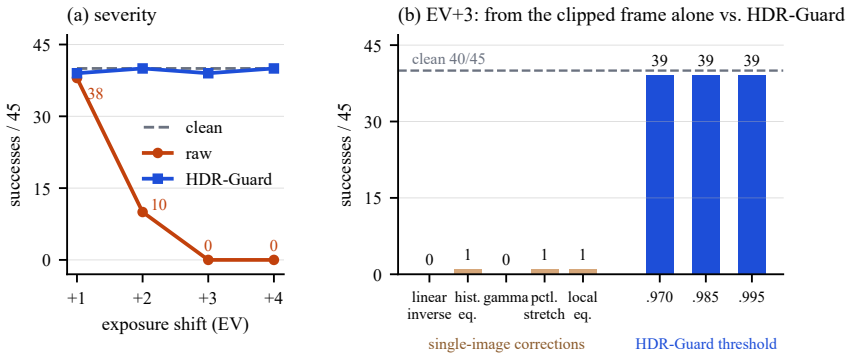


**Figure 3: Controlled test, LIBERO-Spatial tasks 0–4, 45 episodes per bar.** (a) Severity: raw success falls from 38/45 at EV+1 to 10/45 at EV+2 and 0/45 at EV+3 and EV+4; the repaired input stays at 39–40/45 (clean 40/45). (b) At EV+3, five corrections computed from the clipped frame alone reach at most 1/45, consistent with Proposition 2. With the protected measurement, every threshold $\tau \in \{0.970, 0.985, 0.995\}$ gives the same repaired image and 39/45.

linear inverse, histogram equalization, gamma 0.5 after the inverse, a 1–99th percentile stretch and adaptive histogram equalization, none exceeds 1/45, the exact-gain inverse included.

### 4.4 Error-budget stress tests

Table 4 in Appendix D varies clipped area, registration and protected-stream clipping on a solid square patch, which the closing leaves unchanged. An 8-pixel misregistration of the protected frame cost nothing and a 16-pixel shift cost 10 of 90, all OpenVLA's. With the protected stream clipped in the same region, each policy returns exactly to its raw count, consistent with the loss of the protected measurement. On tripod-bracketed RAW frames from a real DSLR, the effective exposure ratio is 16.3 rather than the recorded 15.0; after one global factor fitted on that capture, valid protected-frame samples differ from a separate short-exposure reference by a median 4.8% (90th percentile 21.7%; 3.8% predicted from the measured noise; Appendix I). In a simulation of trigger-free acquisition with the protected frame one control step (50 ms) old, HDR-Guard reaches 362/500, against 336/500 without delay (Appendix G).

## 5 Conclusion

Clipping removes visual evidence that frozen VLAs need. Under the assumptions of Proposition 3, the interface admits a conditional derivation: close the near-saturated pixels, widen the region by one pixel, and write the recorded protected values inside. In the conditions we evaluated this recovers much of the lost behavior, with stress tests consistent with the error terms of Proposition 4; on the evaluated soft spotlight, full-frame fusion is the better choice.

### REPRODUCIBILITY STATEMENT

Section 3 and Appendix B specify the repair rule, the sensor model and every image transform, including the random-number streams. Appendix C lists the checkpoints, evaluation code, task identifiers, trial counts and seeds; all three policies are used through their official LIBERO evaluation scripts with an image hook inserted before policy preprocessing. Complete proofs are in Appendix A. Closed-loop counts are in Appendices D, F, G and K, and per-task counts for the centered hard patch in Appendix L. We will release the image-condition module, the patched runners and the episode logs.

### ETHICS STATEMENT

The closed-loop experiments use simulated manipulation benchmarks and publicly released policy checkpoints. Appendix H reports only aggregate pixel statistics from recordings of firefighting drills; the paper shows no frames from these recordings. A camera-side repair can make a robot act in lighting where it would otherwise stall, so deployments should keep the policy's own safety monitors active.

### AI USE STATEMENT

We used large language models (LLMs) in four ways. (i) Writing: LLMs drafted and edited sections of the manuscript from the authors' notes, results and outlines. (ii) Retrieval: LLMs helped find related work; each reference was checked against its source. (iii) Research execution: LLMs drafted experiment-orchestration, image-condition and analysis code, which the authors ran, checked and debugged, and helped check that every reported number matches the episode logs. (iv) Mathematics: LLMs helped draft the statements and proofs of Propositions 1 to 4, which the authors checked line by line. The research question, the method, the experimental design and the claims are the authors'. All numbers come from logged runs. The authors reviewed all AI-assisted text, code and proofs and take full responsibility for the content.

## REFERENCES

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025b. doi: 10.15607/RSS.2025.XXI.010. URL https://www.roboticsproceedings.org/rss21/p010.html.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H. Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of*

*Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.025. URL https://www.roboticsproceedings.org/rss19/p025.html.

Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. UltraFusion: Ultra high dynamic imaging using exposure fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16111–16121, 2025. doi: 10.1109/CVPR52734.2025.01502.

Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97*, SIGGRAPH '97, pp. 369–378. ACM Press, 1997. doi: 10.1145/258734.258884. URL http://dx.doi.org/10.1145/258734.258884.

Steven Diamond, Vincent Sitzmann, Frank Julca-Aguilar, Stephen Boyd, Gordon Wetzstein, and Felix Heide. Dirty pixels: Towards end-to-end image processing and perception. *ACM Transactions on Graphics*, 40(3):1–15, 2021. ISSN 1557-7368. doi: 10.1145/3446918. URL http://dx.doi.org/10.1145/3446918.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: A progressive robustness benchmark for visual-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38574–38583, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fei_LIBERO-Plus_A_Progressive_Robustness_Benchmark_for_Visual-Language-Action_Models_CVPR_2026_paper.html.

A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian. Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10): 1737–1754, 2008. ISSN 1057-7149. doi: 10.1109/tip.2008.2001399. URL http://dx.doi.org/10.1109/TIP.2008.2001399.

Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.090. URL https://www.roboticsproceedings.org/rss20/p090.html.

Miguel Granados, Boris Ajdin, Michael Wand, Christian Theobalt, Hans-Peter Seidel, and Hendrik P. A. Lensch. Optimal HDR reconstruction with linear digital cameras. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 215–222. IEEE, 2010. doi: 10.1109/cvpr.2010.5540208. URL http://dx.doi.org/10.1109/CVPR.2010.5540208.

Samuel W. Hasinoff, Fredo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 553–560. IEEE, 2010. doi: 10.1109/cvpr.2010.5540167. URL http://dx.doi.org/10.1109/CVPR.2010.5540167.

Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):1–12, 2016. ISSN 1557-7368. doi: 10.1145/2980179.2980254. URL http://dx.doi.org/10.1145/2980179.2980254.

Shutong Jin, Jin Yang, Muhammad Zahid, and Florian T. Pokorny. RoboLight: A dataset with linearly composable illumination for robotic manipulation. arXiv preprint arXiv:2603.04249, 2026. URL https://arxiv.org/abs/2603.04249.

Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):1–12, July 2017. ISSN 1557-7368. doi: 10.1145/3072959.3073609. URL http://dx.doi.org/10.1145/3072959.3073609.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025a. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 44776–44791. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1939. URL http://dx.doi.org/10.52202/075280-1939.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: A diffusion foundation model for bimanual manipulation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/49f80e4d2471ad4f2edf4f5f1ab62339-Abstract-Conference.html.

Xuancun Lu, Jiaxiang Chen, Shilin Xiao, Zizhi Jin, Zhangrui Chen, Hanwen Yu, Bohan Qian, Ruochen Zhou, Xiaoyu Ji, and Wenyuan Xu. Phantom menace: Exploring and enhancing the robustness of VLA models against physical sensor attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35689–35697, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i42.40881. URL http://dx.doi.org/10.1609/aaai.v40i42.40881.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pp. 382–390. IEEE, 2007. doi: 10.1109/pg.2007.17. URL http://dx.doi.org/10.1109/PG.2007.17.

Tomoo Mitsunaga and Shree K. Nayar. Radiometric self calibration. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, CVPR-99, pp. 374–380. IEEE Comput. Soc, 1999. doi: 10.1109/cvpr.1999.786966. URL http://dx.doi.org/10.1109/CVPR.1999.786966.

Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7706–7716. IEEE, 2021. doi: 10.1109/cvpr46437.2021.00762. URL http://dx.doi.org/10.1109/CVPR46437.2021.00762.

Emmanuel Onzon, Maximilian Bömer, Fahim Mannan, and Felix Heide. Neural exposure fusion for high-dynamic range object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17564–17573, 2024. doi: 10.1109/CVPR52733.2024.01663.

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002. ISSN 1557-7368. doi: 10.1145/566654.566575. URL http://dx.doi.org/10.1145/566654.566575.

Pierre Soille. *Morphological Image Analysis: Principles and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2 edition, 2004. doi: 10.1007/978-3-662-05088-0. URL http://dx.doi.org/10.1007/978-3-662-05088-0.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958. ISSN 0012-9682. doi: 10.2307/1907382. URL http://dx.doi.org/10.2307/1907382.

Haozhe Xie, Beichen Wen, Jiarui Zheng, Zhaoxi Chen, Fangzhou Hong, Haiwen Diao, and Ziwei Liu. DynamicVLA: A vision-language-action model for dynamic object manipulation. arXiv preprint arXiv:2601.22153, 2026. URL https://arxiv.org/abs/2601.22153.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

# A PROOFS

Throughout, $x$ is the nominal frame, $b(p) = g\,\ell(p)$ the brightening, $r$ the bracket, and the streams follow Equation (1). All statements are per scalar pixel-channel $(p, c)$. $S_H$ is the set of pixels whose policy value reaches $\tau$ in some channel, $R_H$ the $3 \times 3$ dilation of its closing (the pixels HDR-Guard replaces), and $C = \{(p, c) : b(p)x(p, c) \geq 1\}$ the over-range channels of the policy stream.

**Proof of Proposition 1.** *Cancellation.* Let $z$ be the information available to the selector and $R_K(z) = \mathbb{E}[(H/g - x)^2 \mid z]$, $R_R(z) = \mathbb{E}[(L - x)^2 \mid z]$ the posterior risks of the two options. Under a common positive weight $W(p, z)$, the sign of $W(p, z)\,(R_K(z) - R_R(z))$ equals the sign of $R_K(z) - R_R(z)$, so the preferred source does not depend on the weight, even if the weight depends on the observation. What is excluded is a weight that differs between the two sources, a weight coupled to the unknown $x$, and a non-separable action loss.

*Threshold.* For a selector that uses only the policy value $h$ (or after the other information has been marginalized so that both risks are functions of $h$ alone), if $R_K - R_R$ is nondecreasing in $h$ and changes sign once, then $\{h : R_K(h) \geq R_R(h)\}$ is an interval $[\tau^\star, 1]$ and the preferred source is "replace iff $h \geq \tau^\star$". The proposition asserts the existence of such a threshold, not its value; $\tau = 0.985$ is fixed a priori near the clip level.

*An illustrative sufficient condition.* The following continuous additive-noise model illustrates when the single-crossing condition holds; it is not a proof for the Poisson–Gaussian, quantized simulator, and it ignores lower censoring, which is immaterial near the saturation threshold. Let $s = b(p)x(p, c)$ be the noise-free signal, observed as $h = \min(1, s + n)$ with $n$ additive with a positive log-concave density $f$, and let $q(h) = \Pr(s \geq 1 \mid h)$ be the posterior probability that the signal is over range (under a proper prior with positive normalizing density). This is the probability of a latent over-range signal, which coincides with the probability of an actually clipped observation only in the low-noise limit. For $h < 1$, log-concavity makes the translation kernel $K(h, s) = f(h - s)$ totally positive of order two (TP2), so the posterior of $s$ is stochastically increasing in $h$ and $q$ is nondecreasing on $[0, 1)$. For the censored observation $h = 1$ the kernel is $K(1, s) = \int_1^\infty f(y - s)\,dy$. For $s_2 > s_1$ and $h < 1$,

$$K(1,s_2)K(h,s_1) - K(1,s_1)K(h,s_2) = \int_1^\infty \big[f(y-s_2)f(h-s_1) - f(y-s_1)f(h-s_2)\big]\,dy \;\geq\; 0,$$

because $y \geq 1 > h$ and $f(\cdot - s)$ is TP2, so the censored atom is ordered above every $h < 1$ and $q$ is nondecreasing on $[0, 1]$. Now suppose each option's risk is constant within each regime: keeping costs $c_{K,U}$ on in-range and $c_{K,C}$ on over-range signals, and writing $L$ costs $c_{R,U}$ and $c_{R,C}$. Then $R_K - R_R = (c_{K,U} - c_{R,U}) + q(h)\,[(c_{K,C} - c_{R,C}) - (c_{K,U} - c_{R,U})]$ is affine and nondecreasing in $q$ whenever keeping loses more than writing on over-range signals, which is the case $c_{K,C} - c_{R,C} \geq c_{K,U} - c_{R,U}$; composed with the nondecreasing $q$, it changes sign at most once.

*What the result does not cover.* The argument is per scalar channel. HDR-Guard replaces a pixel when any of its three channels exceeds the threshold, and closes and dilates the result; that joint rule is covered by Proposition 3 in the noise-free two-level model, not by this proposition. In the controlled matched-bracket protocol both options return $x$ exactly on every in-range pixel, so the threshold sweep in Figure 3b checks the implementation rather than the proposition. □

**Proof of Proposition 2.** Without noise, $H(x)(p,c) = \min(1, b(p)x(p,c))$. On $C$ both scenes satisfy $bx \geq 1$ and $bx' \geq 1$, so $H(x)(p,c) = H(x')(p,c) = 1$; outside $C$ the scenes agree, so their observations agree. Hence $H(x) = H(x')$, any map $f$ returns one observation $\tilde{x}$ for both scenes, and a deterministic policy takes one action $\tilde{a} = \pi(\tilde{x})$. For any metric $d$, the triangle inequality gives $d(\pi(x), \pi(x')) \leq d(\tilde{a}, \pi(x)) + d(\tilde{a}, \pi(x')) \leq 2\max\{d(\tilde{a}, \pi(x)), d(\tilde{a}, \pi(x'))\}$. For discrete actions, $\tilde{a}$ cannot equal both $\pi(x)$ and $\pi(x')$ when they differ, so $\pi \circ f$ errs on at least one scene.

*With noise.* Two over-range values need not produce identical observation distributions, so the exact argument does not carry over; what remains is the posterior Bayes risk, which no rule can beat. Let the scene $X$ be random with a prior, $A = \pi(X)$ its nominal action, and $\delta(H)$ any decision rule that sees only the policy stream, including $\pi \circ f$. For discrete or quantized actions under 0–1 loss, $\Pr[\delta(H) \neq A \mid H = h] \geq 1 - \max_a \Pr(A = a \mid H = h)$, so $\Pr[\delta(H) \neq A] \geq \mathbb{E}_H[1 - \max_a \Pr(A = a \mid H)]$. For continuous actions and any loss $d$, $\mathbb{E}[d(\delta(H), A)] \geq \mathbb{E}_H\big[\inf_a \mathbb{E}[d(a, A) \mid H]\big]$. In both cases the right-hand side depends only on how much action-relevant information the clipped observation retains, not on $f$. It can be zero; it is strictly positive whenever, with positive probability, the observation leaves positive posterior probability on actions that differ under the chosen loss. For a stochastic policy the statements hold conditional on the policy's internal randomness; a distributional version would need a metric between conditional action distributions. □

**Proof of Proposition 3.** Assume $b = r\mathbb{1}_{\Omega} + g\mathbb{1}_{\Omega^c}$ with $g \geq 1$, headroom ($g\max_c x(p,c) \geq \tau \Rightarrow p \in \Omega$), regularity ($\mathrm{close}_B(\Omega) = \Omega$ for a symmetric $B$ that contains the origin), and noise-free, unquantized, registered streams $H = \min(1, bx)$, $L = \min(1, bx/r)$ with $x \in [0,1]$. *(a)* On $\Omega$, $L = \min(1, x) = x$. Off $\Omega$, headroom gives $g\,x < \tau < 1$ in every channel, so $H = gx$ and $H/g = x$. The chosen value equals $x$, so wherever $L \neq H/g$ the other value does not. *(b)* If $p \in S_H$ and $p \notin \Omega$, then $H(p,c) = g\,x(p,c) < \tau$ in every channel by headroom, a contradiction; hence $S_H \subseteq \Omega$. Let $\mathcal{A} = \{\Omega' : \mathrm{close}_B(\Omega') = \Omega',\ S_H \subseteq \Omega'\}$. For every $\Omega' \in \mathcal{A}$ there is a scene $x'$, depending on $\Omega'$, that satisfies headroom and reproduces both streams exactly. Both streams clip the same product $bx$, so $H = \min(1, rL)$ at every pixel. Set $x' = L$ on $\Omega'$; with brightening $r$ this reproduces $\min(1, rx') = H$ and $\min(1, x') = L$. Off $\Omega'$, $\max_c H < \tau$ because $S_H \subseteq \Omega'$, so $H = bx$ is unclipped and $L = H/r$; the scene $x' = H/g \leq 1$ with brightening $g$ reproduces both streams and satisfies $g\max_c x' < \tau$. The closing is extensive ($X \subseteq \mathrm{close}_B\, X$), increasing and idempotent (Soille, 2004). Idempotence and extensivity put $\mathrm{close}_B(S_H)$ in $\mathcal{A}$; monotonicity gives $\mathrm{close}_B(S_H) \subseteq \mathrm{close}_B(\Omega') = \Omega'$ for every $\Omega' \in \mathcal{A}$. Hence $\mathrm{close}_B(S_H) = \bigcap \mathcal{A}$, and it is the unique element of $\mathcal{A}$ of smallest area. Because the reproducing scene depends on the candidate region, the data restrict $\Omega$ to $\mathcal{A}$ but do not by themselves rank its elements. If the marginal likelihood $p(H, L \mid \Omega')$, with the scene integrated out, is constant on $\mathcal{A}$, the posterior restricted to $\mathcal{A}$ is proportional to the region prior. If that prior is a strictly decreasing function of area alone, its unique maximizer is the unique smallest element, $\mathrm{close}_B(S_H)$. The dilation of (c) is a separate decision, so the full rule is a plug-in construction rather than a single posterior optimum. *(c)* Let $N = \{-1, 0, 1\}^2$. *Minimum intervention.* A region $R$ is feasible if $\hat{\Omega} + \delta \subseteq R$ for every $\delta \in N$. Every feasible $R$ contains $\bigcup_{\delta \in N}(\hat{\Omega} + \delta) = \hat{\Omega} \oplus N = \mathrm{dilate}_{3\times 3}(\hat{\Omega})$, and this union is itself feasible, so it is the unique feasible region of smallest area. *Expected loss.* Let the lit region be $\hat{\Omega} + \delta$ with $\delta$ uniform on $N$. Since $N = -N$, $q(p) = \Pr(p \in \Omega) = |(p + N) \cap \hat{\Omega}|/9$, which is positive exactly on $\mathrm{dilate}_{3\times 3}(\hat{\Omega}) = \hat{\Omega} \oplus N$ and at least $1/9$ there. Let $c_M(p)$ be the loss of keeping $H/g$ at a lit pixel and $c_F(p)$ that of writing $L$ at an unlit one; both options are exact otherwise by (a). Replacing is optimal iff $q\,c_M \geq (1-q)\,c_F$, that is, $q \geq c_F/(c_M + c_F)$. If $c_M \geq 8c_F$ the right-hand side is at most $1/9$, so replacing is optimal on $\{q > 0\} = \mathrm{dilate}_{3\times 3}(\hat{\Omega})$ and keeping is optimal outside it. The choice is unique where $c_F > 0$ and either $c_M > 8c_F$ or $q > 1/9$; at $q = 1/9$ with $c_M = 8c_F$ the two options tie. The argument uses the stipulated uniform translation model; it does not show that thresholding, resizing or registration errors follow that distribution. Under squared error at a pixel with $r\,x(p,c) < 1$ in every channel, keeping errs by $(r/g - 1)\,x$ and replacing by $(1 - g/r)\,x$

per channel, so $c_M/c_F = (r/g)^2 \geq 8$ iff $r/g \geq 2\sqrt{2}$. For a global shift ($g = r$) both options are exact on unclipped pixels, and the dilation is harmless. *Convex regions are regular.* For a symmetric $B$, $p \notin \mathrm{close}_B(X)$ iff some translate of $B$ contains $p$ and misses $X$. Let $X$ be closed and convex, $B$ compact, and $p \notin X$. A separating direction $u$ gives $\langle u, p\rangle > \sup_{z\in X}\langle u, z\rangle$. Translate $B$ so that $p$ minimizes $\langle u, \cdot\rangle$ over the translate; every point of the translate then has $\langle u, \cdot\rangle \geq \langle u, p\rangle$ and lies outside $X$. Hence $\mathrm{close}_B(X) \subseteq X$, and extensivity gives equality. On the pixel grid this holds up to discretization of the boundary; the centered and offset hard patches of Section 4.1 are super-ellipses, hence convex, and at $256 \times 256$ both are exact fixed points of the $15 \times 15$ elliptical closing (checked numerically). □

**Proof of Proposition 4.** *Case $p \in R_H$, $L(p, c) < 1$.* The protected channel is not upper-clipped and the rule writes it. Write $s = (bx/r)(p + \delta) \geq 0$ and $Y = \mathrm{clip}_{[0,1]}(s + n_L)$, so that $\hat{x} = Y + e_Q$ with $|e_Q| \leq \epsilon_Q$. Since $Y < 1$, $|Y - s| \leq |n_L|$: in the interior $Y = s + n_L$, and if the lower clip is active then $n_L \leq -s$, so $|Y - s| = s \leq |n_L|$. Assume the registered signal is represented by a continuously differentiable interpolation along the displacement segment. Adding and subtracting $(bx/r)(p)$ and $x(p)$,

$$|\hat{x} - x| \leq |(bx/r)(p + \delta) - (bx/r)(p)| + |b(p)/r - 1|\, x(p) + |n_L| + \epsilon_Q,$$

and the mean-value theorem bounds the first term by $\sup_{t\in[0,1]} \|\nabla(bx/r)(p + t\delta)\|\, \|\delta\|$. The case also covers pixels added by the closing or the dilation, whose policy value is below $\tau$, since the rule writes $L$ there as well. *Case $p \in R_H$, $L(p, c) = 1$.* The protected channel sits at the digital clip level, so $\hat{x} = 1$ and $|\hat{x} - x| = |1 - x|$; this case includes samples clipped or rounded to the digital ceiling, and the written value does not identify $x$ beyond that censoring constraint, as in Proposition 2. *Case $p \notin R_H$.* No channel of the pixel reached $\tau < 1$, so the channel is not upper-clipped, and the same argument as in the first case gives $|H/g - \ell x| \leq |n_H|/g + \epsilon_Q/g$, hence the bound. The action statement follows from $\|\pi(\hat{x}) - \pi(x)\| \leq L_\pi \|\hat{x} - x\|_W$ and the triangle inequality over pixel-channels.

*Noise factor of rescaled exposure selection.* Let the full-well capacity be $F$ electrons at recorded value 1. Shot noise gives a recorded variance of $s/F$ for a recorded signal $s$, and read noise adds $\sigma_e^2/F^2$. At an unclipped pixel with $b = g = 1$, the policy stream records $s_H = x$ and the protected stream $s_L = x/r$. Multiplied by $r$, the protected value has variance $r^2(x/(rF) + \sigma_e^2/F^2) = r\,x/F + r^2\sigma_e^2/F^2$, against $x/F + \sigma_e^2/F^2$ for the kept pixel: a standard-deviation ratio of $\sqrt{r}$ when shot noise dominates and $r$ when read noise dominates. □

## B SENSOR-AWARE SIMULATION

The simulation models the sensor (photon and read noise, clipping, quantization) but not light transport: the bright region is a multiplicative field in image space, applied at the same normalized coordinates in both views, without view-dependent shading, shadows or flare.

**Sensor model.** Each stream records $y = \mathrm{clip}_{[0,1]}\big((\mathrm{Poisson}(F\,\gamma\,s) + \mathcal{N}(0, \sigma_e^2))/F\big)$, quantized to 8 bits, where $s$ is the irradiance on the nominal scale, $\gamma$ the stream's exposure relative to nominal, $F = 1000$ electrons at recorded value 1 and $\sigma_e = 2$ electrons. These values correspond to a small-pixel CMOS sensor at moderate gain (Foi et al., 2008). The nominal frame $x$ is the simulator render. The policy stream has $\gamma = 1$ for the light-patch scenes and $\gamma = 8$ for the global shift; the protected stream has $\gamma = 1/8$ and $\gamma = 1$, respectively ($r = 8$). The two streams draw independent noise. Random numbers come from a counter-based generator keyed by the condition name (which includes the rule), the camera view and a per-process call index, so every run is reproducible and each rule draws its own noise realization from the same sensor model. The common-random-number replicate of the hard patch instead seeds the generator with a hash of the scene name, the camera view and the rendered frame, so every rule sees exactly the same noisy streams whenever it visits the same state; trajectories still diverge once the rules act differently.

**Light patch.** The brightening is $\ell(u, v) = 1 + (G - 1)\,\sigma\big(k(1 - \rho(u, v))\big)$, where $\rho = \big(|u - 0.5|^4/0.25^4 + |v - 0.5|^4/0.25^4\big)^{1/4}$ is a super-ellipse radius in normalized image coordinates, $\sigma$ is the logistic function and $G = 8$. The hard edge uses $k = 200$ (the transition spans about one pixel); the soft spotlight uses $k = 14$, so $\ell - 1$ falls from 94% to 6% of $G - 1$ over about a tenth of the frame width. The bracket-mismatch scenes use the soft falloff with $G \in \{6, 12\}$ and the same fixed

bracket $r = 8$. The patch covers about 23% of the frame and is applied to the third-person and wrist views in the same normalized coordinates. The offset confirmation patch was specified after the threshold, closing and dilation were fixed: a hard edge ($k = 200$), $G = r = 8$, centered at $(u, v) = (0.40, 0.60)$ with semi-axes 0.22 and 0.17, which covers about 14% of the frame.

**Baselines.** All rules receive noisy streams generated by the same sensor model. *Raw* feeds the policy stream. *Exposure selection, as recorded* feeds the protected stream unchanged; *rescaled* multiplies it by $r$ and clips. *Mertens* applies OpenCV's implementation of exposure fusion (Mertens et al., 2007) with its default weights: contrast 1, saturation 1 and well-exposedness 0. With all three weights set to 1, OpenVLA reaches 148/500 on the 500 LIBERO-Spatial states of Appendix F, against 156/500 with the default weights and 335/500 with HDR-Guard (HDR-Guard minus this variant: +187 [+117, +252]), and 128/584 on the 584 Object, Goal and LIBERO-10 states of Appendix F.1, against 145/584 with the default weights and 320/584 with HDR-Guard (+192 [+145, +238]; task-cluster bootstrap). Enabling the well-exposedness term does not close the gap. *Localized Mertens* is the support-controlled version: with $F_{\mathrm{M}}(H, L)$ the same OpenCV output and $R_H$ the closed and dilated mask of HDR-Guard, it outputs $F_{\mathrm{M}}(H, L)(p)$ for $p \in R_H$ and $H(p)/g$ otherwise, so no Mertens value is written outside $R_H$. *HDR merge + Reinhard* averages the unclipped samples of both streams on the radiance scale (dividing each by its exposure) and applies the global operator of Reinhard et al. (2002) with key 0.18 and the white point at the maximum luminance. The *soft saturation blend* computes $w = \mathrm{clip}_{[0,1]}\left((\max_c H - 0.90)/(\tau - 0.90)\right)$ per pixel and outputs $(1 - w)H/g + wL$, a classical soft version of saturation-aware merging. HDR-Guard follows Equation (2) with $\tau = 0.985$, a $15 \times 15$ elliptical structuring element for the closing (OpenCV `MORPH_CLOSE`) and $g = 1$ (light patch) or $g = 8$ (global shift); the ablation without closing uses $R_H = \mathrm{dilate}_{3\times3}(S_H)$. In the controlled matched-bracket protocol both variants return the nominal frame exactly, because every pixel they write equals $x$.

**Well-exposed scene.** The streams use the same bracket ($\gamma = 1$ and $1/8$) with no light patch. No pixel of the noise-free LIBERO-Spatial frames reaches $\tau$; with sensor noise, isolated spikes flag 0.01% of the pixels, which the closing does not grow. HDR-Guard therefore returns the policy stream almost everywhere, while the fusion baselines rewrite every pixel.

## C Evaluation protocol

**Policies.** OpenVLA-OFT uses the official LIBERO checkpoints for each suite, both camera views, proprioception, parallel decoding and 8-step action chunks executed open loop, through the authors' evaluation script with the authors' fork of the language-model library required for bidirectional attention. OpenVLA uses the official LIBERO-finetuned checkpoints, the third-person view and one action per step. $\pi_{0.5}$ uses the official LIBERO checkpoint served over a websocket, both views and 5-step replanning. All three use the suites' default maximum episode lengths (220 steps for Spatial), 10 initial no-op steps and seed 7.

**Image hook.** The transform is applied to each raw camera frame after the benchmark's 180-degree flip and before the policy's own resizing and cropping. For OpenVLA-OFT the transform is evaluated at the steps where a new chunk is queried, which are the only frames the policy reads. The $\pi_{0.5}$ client resizes both views to $224 \times 224$ before sending them to the policy server, and the transform is applied to these resized frames.

**Trial accounting.** A run counts only if the policy produced a complete closed-loop trace under the clean condition with the same adapter. In the controlled protocol, one $\pi_{0.5}$ clean run failed during websocket warm-up before any action was produced; it was excluded and rerun, and the valid clean evaluation was 29/30.

**Compute.** The controlled-protocol runs used one 24 GB consumer GPU per policy. The sensor-aware runs used one 96 GB workstation GPU shared by up to six policy processes and three 24 GB GPUs; the nominal-exposure reference of that protocol was re-measured on the same machines and software stack. The 50-trial and cross-suite OpenVLA runs used an HPC cluster with H200 GPUs and software OpenGL rendering (OSMesa); on a reference state its first OpenVLA action matched the workstation run to every printed digit, and its nominal-exposure count is reported with those runs.

## D PER-POLICY TABLES

**Table 4: Error-budget stress rows, controlled protocol, LIBERO-Spatial tasks 0–9, 90 episodes each.** A center square of the policy frame is clipped; the protected frame is then shifted, or clipped in the same region.

| Error source probed | Setting | Raw | HDR-Guard |
|---|---|---|---|
| Clipped area | center 15% / 30% / 50% | 76 / 64 / 49 | 81 / 82 / 83 |
| Misregistration | protected frame shifted 8 px / 16 px (30% clipped) | 64 / 64 | 82 / 72 |
| Protected saturation | both streams clipped (30%) | 64 | 64 |

**Table 5:** Severity sweep, controlled protocol, LIBERO-Spatial tasks 0–4, three trials per task.

| Policy | Clean | Raw +1 | Ours +1 | Raw +2 | Ours +2 | Raw +3 | Ours +3 | Raw +4 | Ours +4 |
|---|---|---|---|---|---|---|---|---|---|
| OpenVLA-OFT | 15/15 | 15/15 | 15/15 | 10/15 | 15/15 | 0/15 | 15/15 | 0/15 | 15/15 |
| OpenVLA | 10/15 | 9/15 | 10/15 | 0/15 | 10/15 | 0/15 | 10/15 | 0/15 | 10/15 |
| $\pi_{0.5}$ | 15/15 | 14/15 | 14/15 | 0/15 | 15/15 | 0/15 | 14/15 | 0/15 | 15/15 |
| Total | 40/45 | 38/45 | 39/45 | 10/45 | 40/45 | 0/45 | 39/45 | 0/45 | 40/45 |

**Table 6:** Cross-suite breakdown of Table 3, EV+3, tasks 0–4, three trials per task.

| Policy | Suite | Clean | Raw EV+3 | HDR-Guard |
|---|---|---|---|---|
| OpenVLA-OFT | Object | 15/15 | 2/15 | 15/15 |
| OpenVLA-OFT | Goal | 15/15 | 0/15 | 15/15 |
| OpenVLA-OFT | LIBERO-10 | 15/15 | 0/15 | 15/15 |
| OpenVLA | Object | 12/15 | 0/15 | 12/15 |
| OpenVLA | Goal | 10/15 | 0/15 | 10/15 |
| OpenVLA | LIBERO-10 | 8/15 | 0/15 | 8/15 |
| $\pi_{0.5}$ | Object | 15/15 | 0/15 | 14/15 |
| $\pi_{0.5}$ | Goal | 14/15 | 0/15 | 14/15 |
| $\pi_{0.5}$ | LIBERO-10 | 15/15 | 0/15 | 14/15 |

OpenVLA on LIBERO-10 tasks 3–4 fails in the clean condition (0/6), so those episodes cannot show a repair effect; they remain in the totals.

**Table 7:** Single-image corrections and threshold sweep, EV+3, LIBERO-Spatial tasks 0–4. The corrections receive only the clipped frame.

| Policy | Linear | Hist. eq. | Gamma | Percentile | Adapt. eq. | $\tau$=0.970 | $\tau$=0.985 | $\tau$=0.995 |
|---|---|---|---|---|---|---|---|---|
| OpenVLA-OFT | 0/15 | 1/15 | 0/15 | 1/15 | 1/15 | 15/15 | 15/15 | 15/15 |
| OpenVLA | 0/15 | 0/15 | 0/15 | 0/15 | 0/15 | 10/15 | 10/15 | 10/15 |
| $\pi_{0.5}$ | 0/15 | 0/15 | 0/15 | 0/15 | 0/15 | 14/15 | 14/15 | 14/15 |
| Total | 0/45 | 1/45 | 0/45 | 1/45 | 1/45 | 39/45 | 39/45 | 39/45 |

**Table 8:** Stress rows of Table 4 per policy, LIBERO-Spatial tasks 0–9. L$n$: center $n$% of the policy frame clipped; shift and both-clipped rows use L30. The EV$-$3 columns apply the mirror rule on the dark side: the policy camera is 3 stops underexposed, and pixels whose value falls to 0.08 or below take the 3-stop-longer exposure.

| Policy | Clean | L15 | Ours | L30 | Ours | L50 | Ours | Shift 8 | Shift 16 | Both | EV$-$3 | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OpenVLA-OFT | 30/30 | 30/30 | 30/30 | 30/30 | 30/30 | 29/30 | 30/30 | 30/30 | 30/30 | 30/30 | 30/30 | 30/30 |
| OpenVLA | 23/30 | 17/30 | 23/30 | 10/30 | 23/30 | 3/30 | 23/30 | 23/30 | 13/30 | 10/30 | 16/30 | 23/30 |
| $\pi_{0.5}$ | 29/30 | 29/30 | 28/30 | 24/30 | 29/30 | 17/30 | 30/30 | 29/30 | 29/30 | 24/30 | 30/30 | 30/30 |
| Total | 82/90 | 76/90 | 81/90 | 64/90 | 82/90 | 49/90 | 83/90 | 82/90 | 72/90 | 64/90 | 76/90 | 83/90 |

**Table 9:** Wilson 95% intervals for the main aggregates.

| Row | Successes | Rate | 95% interval |
|---|---|---|---|
| Spatial clean / HDR-Guard (EV+3) | 82/90 | 0.911 | [0.834, 0.954] |
| Spatial raw EV+3 | 0/90 | 0.000 | [0.000, 0.041] |
| Cross-suite clean | 119/135 | 0.881 | [0.816, 0.926] |
| Cross-suite raw EV+3 | 2/135 | 0.015 | [0.004, 0.052] |
| Cross-suite HDR-Guard | 117/135 | 0.867 | [0.799, 0.914] |
| HDR-Guard, threshold sweep | 39/45 | 0.867 | [0.738, 0.937] |
| Best single-image correction | 1/45 | 0.022 | [0.004, 0.116] |
| HDR-Guard, L50 | 83/90 | 0.922 | [0.848, 0.962] |
| HDR-Guard, EV$-3$ | 83/90 | 0.922 | [0.848, 0.962] |

**Table 10:** Sensor-noise-aware image-space simulation, every condition completed for all three policies, LIBERO-Spatial tasks 0–9, three trials per task. Wilson 95% intervals on the pooled total are descriptive; comparisons between rules use the paired task-cluster bootstrap.

| Condition | OFT | OpenVLA | $\pi_{0.5}$ | Total | 95% interval |
|---|---|---|---|---|---|
| Nominal exposure, no patch | 29/30 | 24/30 | 29/30 | 82/90 | [0.83, 0.95] |
| Hard patch: Policy camera (raw) | 9/30 | 0/30 | 12/30 | 21/90 | [0.16, 0.33] |
| Hard patch: Exposure selection, as recorded | 20/30 | 1/30 | 14/30 | 35/90 | [0.29, 0.49] |
| Hard patch: Exposure selection, rescaled | 5/30 | 2/30 | 6/30 | 13/90 | [0.09, 0.23] |
| Hard patch: Mertens exposure fusion | 28/30 | 10/30 | 19/30 | 57/90 | [0.53, 0.73] |
| Hard patch: HDR merge + Reinhard | 23/30 | 5/30 | 15/30 | 43/90 | [0.38, 0.58] |
| Hard patch: Localized Mertens (within $R_H$) | 27/30 | 15/30 | 23/30 | 65/90 | [0.62, 0.80] |
| Hard patch: Soft saturation blend | 29/30 | 13/30 | 28/30 | 70/90 | [0.68, 0.85] |
| Hard patch: HDR-Guard, no closing (ablation) | 29/30 | 14/30 | 26/30 | 69/90 | [0.67, 0.84] |
| Hard patch: HDR-Guard | 29/30 | 19/30 | 28/30 | 76/90 | [0.76, 0.91] |
| Hard patch: HDR-Guard, no dilation | 27/30 | 13/30 | 29/30 | 69/90 | [0.67, 0.84] |
| Hard patch, common noise: Localized Mertens (within $R_H$) | 28/30 | 13/30 | 23/30 | 64/90 | [0.61, 0.79] |
| Hard patch, common noise: Soft saturation blend | 28/30 | 20/30 | 29/30 | 77/90 | [0.77, 0.91] |
| Hard patch, common noise: HDR-Guard, no closing (ablation) | 29/30 | 13/30 | 28/30 | 70/90 | [0.68, 0.85] |
| Hard patch, common noise: HDR-Guard | 28/30 | 24/30 | 27/30 | 79/90 | [0.79, 0.93] |
| Offset patch: Policy camera (raw) | 17/30 | 1/30 | 13/30 | 31/90 | [0.25, 0.45] |
| Offset patch: Mertens exposure fusion | 30/30 | 14/30 | 25/30 | 69/90 | [0.67, 0.84] |
| Offset patch: Localized Mertens (within $R_H$) | 30/30 | 14/30 | 25/30 | 69/90 | [0.67, 0.84] |
| Offset patch: Soft saturation blend | 30/30 | 26/30 | 29/30 | 85/90 | [0.88, 0.98] |
| Offset patch: HDR-Guard, no closing (ablation) | 30/30 | 22/30 | 29/30 | 81/90 | [0.82, 0.95] |
| Offset patch: HDR-Guard | 30/30 | 25/30 | 28/30 | 83/90 | [0.85, 0.96] |
| Soft spot: Policy camera (raw) | 10/30 | 2/30 | 9/30 | 21/90 | [0.16, 0.33] |
| Soft spot: Exposure selection, as recorded | 26/30 | 13/30 | 28/30 | 67/90 | [0.65, 0.82] |
| Soft spot: Mertens exposure fusion | 29/30 | 20/30 | 27/30 | 76/90 | [0.76, 0.91] |
| Soft spot: HDR merge + Reinhard | 28/30 | 12/30 | 29/30 | 69/90 | [0.67, 0.84] |
| Soft spot: Localized Mertens (within $R_H$) | 26/30 | 8/30 | 20/30 | 54/90 | [0.50, 0.70] |
| Soft spot: Soft saturation blend | 25/30 | 5/30 | 18/30 | 48/90 | [0.43, 0.63] |
| Soft spot: HDR-Guard | 22/30 | 5/30 | 18/30 | 45/90 | [0.40, 0.60] |
| Soft, $G$=6: Policy camera (raw) | 18/30 | 2/30 | 14/30 | 34/90 | [0.28, 0.48] |
| Soft, $G$=6: Mertens exposure fusion | 28/30 | 18/30 | 27/30 | 73/90 | [0.72, 0.88] |
| Soft, $G$=6: Localized Mertens (within $R_H$) | 27/30 | 11/30 | 21/30 | 59/90 | [0.55, 0.75] |
| Soft, $G$=6: Soft saturation blend | 21/30 | 4/30 | 12/30 | 37/90 | [0.32, 0.51] |
| Soft, $G$=6: HDR-Guard, no closing (ablation) | 20/30 | 3/30 | 17/30 | 40/90 | [0.35, 0.55] |
| Soft, $G$=6: HDR-Guard | 20/30 | 5/30 | 17/30 | 42/90 | [0.37, 0.57] |
| Soft, $G$=12: Policy camera (raw) | 4/30 | 2/30 | 3/30 | 9/90 | [0.05, 0.18] |
| Soft, $G$=12: Soft saturation blend | 22/30 | 5/30 | 16/30 | 43/90 | [0.38, 0.58] |
| Soft, $G$=12: HDR-Guard | 23/30 | 4/30 | 17/30 | 44/90 | [0.39, 0.59] |
| Well-exposed: Mertens exposure fusion | 30/30 | 23/30 | 30/30 | 83/90 | [0.85, 0.96] |
| Well-exposed: HDR merge + Reinhard | 29/30 | 28/30 | 30/30 | 87/90 | [0.91, 0.99] |
| Well-exposed: Soft saturation blend | 30/30 | 24/30 | 29/30 | 83/90 | [0.85, 0.96] |
| Well-exposed: HDR-Guard, no closing (ablation) | 29/30 | 23/30 | 30/30 | 82/90 | [0.83, 0.95] |
| Well-exposed: HDR-Guard | 29/30 | 25/30 | 30/30 | 84/90 | [0.86, 0.97] |
| Global EV+3: Policy camera (raw) | 0/30 | 0/30 | 0/30 | 0/90 | [0.00, 0.04] |
| Global EV+3: Exposure selection, as recorded | 30/30 | 24/30 | 30/30 | 84/90 | [0.86, 0.97] |
| Global EV+3: Mertens exposure fusion | 30/30 | 23/30 | 29/30 | 82/90 | [0.83, 0.95] |
| Global EV+3: HDR-Guard, no closing (ablation) | 29/30 | 26/30 | 30/30 | 85/90 | [0.88, 0.98] |
| Global EV+3: HDR-Guard | 29/30 | 23/30 | 30/30 | 82/90 | [0.83, 0.95] |

## E BRACKET MISMATCH

Table 11 reports the completed scenes in which the brightening departs from the bracket. On the soft spotlight, Mertens fusion reaches 76/90 and HDR-Guard 45/90; the largest share of the gap comes from OpenVLA (20/30 against 5/30). On the soft spotlight the brightening falls from $r$ to 1 across the edge, so pixels in the falloff that the closing and dilation bring into $R_H$, and falloff pixels that still clip, receive $bx/r < x$. With $G = 6$ the written patch interior is about $25\%$ too dark; with $G = 12$ it is $50\%$ too bright, and the protected stream also clips wherever $x > 2/3$. In both cases the kept pixels just outside $R_H$ carry the brightened value $\ell x$ with $\ell > 1$. These are the two bracket-mismatch terms of Proposition 4. Because both streams identify only $b\,x$, no fixed per-pixel

rule can guarantee exact recovery over all compatible $b$ and $x$; additional scene information or an illumination prior is required, of which an explicit estimate of the illumination field is one possible implementation. The paired action metric gives the same ordering on the soft spotlight: $D_1 = 3.88$ for Mertens fusion against $5.73$ for HDR-Guard with OpenVLA-OFT, and $7.39$ against $12.38$ with OpenVLA (Table 22).

We screened one fixed alternative offline before deciding not to change the rule. Blending $H/g$ and $L$ with a Gaussian-feathered version of $R_H$ ($\sigma = 6$ px) lowers the mean absolute error to the nominal frame on the soft spotlight from 16.0 to 11.5 (8-bit units, third-person views of tasks 0, 3 and 6) but raises it on the hard patch from 6.2 to 7.7, the matched operating point. The trade-off is consistent with imposing one fixed spatial smoothness prior where the exposure pair does not identify $b(p)$: softening the boundary helps where the light falls off gradually and hurts where it does not.

**Table 11:** Bracket-mismatch scenes, LIBERO-Spatial tasks 0–9, three trials per task. Soft spotlight with $G = r = 8$ (per policy out of 30, total out of 90) and soft patches with $G \in \{6, 12\}$ under the fixed bracket $r = 8$ (totals). Best total per column in bold; –: no complete run for that cell.

| | Soft spotlight, $G = r = 8$ | | | | Soft, $r$=8 | Soft, $r$=8 |
|---|---|---|---|---|---|---|
| Input to the frozen policy | OFT | OpenVLA | $\pi_{0.5}$ | Total | $G = 6$ | $G = 12$ |
| Policy camera (raw) | 10 | 2 | 9 | 21 | 34 | 9 |
| Exposure selection, as recorded | 26 | 13 | 28 | 67 | – | – |
| Exposure selection, rescaled | 12 | – | 2 | – | – | – |
| Mertens exposure fusion | 29 | 20 | 27 | **76** | **73** | – |
| HDR merge + Reinhard | 28 | 12 | 29 | 69 | – | – |
| Localized Mertens (within $R_H$) | 26 | 8 | 20 | 54 | 59 | – |
| Soft saturation blend | 25 | 5 | 18 | 48 | 37 | 43 |
| HDR-Guard, no closing (ablation) | 24 | – | 21 | – | 40 | – |
| HDR-Guard | 22 | 5 | 18 | 45 | 42 | **44** |

## F TARGETED OPENVLA STUDIES

We chose OpenVLA after the 90-episode study because the local rules differed most on it, so the runs in this section are a targeted precision study of one policy rather than an independent test of all three. They use the centered hard patch ($G = r = 8$) with common random numbers, the official per-suite checkpoints and evaluation code, and a separate cluster whose GPUs render with software OpenGL (OSMesa). We report every rule that was run, with the primary contrast HDR-Guard minus the soft blend, and we do not pool these studies with each other or with the 90-episode study. Intervals are paired task-cluster bootstraps (95%, exact multinomially weighted enumeration of task-count vectors); for the three further suites, tasks are resampled within each suite and the suite totals summed, which weights the suites equally. As limited backend checks, one reference-state first action matches the workstation run to every printed digit, and nominal-exposure success on this cluster is 422/500 on LIBERO-Spatial, 175/200 on Object, 158/200 on Goal and 114/200 on LIBERO-10, close to the published OpenVLA results of 84.7%, 88.4%, 79.2% and 53.7% (Kim et al., 2025b); we do not treat these checks as pixel-level equivalence of the two renderers.

**Fifty trials per task on LIBERO-Spatial.** Table 12 lists the totals over the benchmark's 50 initial states per task (500 episodes per rule), Wilson intervals and paired differences from HDR-Guard.

**Table 12:** OpenVLA on the centered hard patch ($G = r = 8$), LIBERO-Spatial tasks 0–9, 50 trials per task, common random numbers.

| Input to OpenVLA | Successes | Rate | 95% Wilson | HDR-Guard−rule (task bootstrap) |
|---|---|---|---|---|
| Nominal exposure, no patch | 422/500 | 0.844 | [0.810, 0.873] | – |
| HDR-Guard | 335/500 | 0.670 | [0.628, 0.710] | – |
| Soft saturation blend | 305/500 | 0.610 | [0.567, 0.652] | +30 [−30, +105] |
| HDR-Guard, no closing | 289/500 | 0.578 | [0.534, 0.621] | +46 [−13, +119] |
| Localized Mertens (within $R_H$) | 230/500 | 0.460 | [0.417, 0.504] | +105 [+45, +161] |
| Full-frame Mertens | 156/500 | 0.312 | [0.273, 0.354] | +179 [+104, +245] |
| Policy camera (raw) | 7/500 | 0.014 | [0.007, 0.029] | +328 [+250, +389] |

### F.1 THREE FURTHER LIBERO SUITES

Table 13 repeats the hard patch on tasks 0–9 of LIBERO-Object, LIBERO-Goal and LIBERO-10, 20 trials per task, and lists per suite the difference between HDR-Guard and each rule.

**Table 13:** OpenVLA under the centered hard patch ($G = r = 8$) on three further LIBERO suites, tasks 0–9, 20 trials per task (200 episodes per cell). Differences are HDR-Guard minus the rule, with suite-stratified task-cluster bootstrap intervals (95%).

| Input to OpenVLA | Object | Goal | LIBERO-10 | Total |
|---|---|---|---|---|
| Nominal exposure | 175/200 | 158/200 | 114/200 | 447/600 |
| Policy camera (raw) | 3/200 | 12/200 | 0/200 | 15/600 |
| Full-frame Mertens | 44/200 | 101/200 | 4/200 | 149/600 |
| Soft saturation blend | 152/200 | 132/200 | 47/200 | 331/600 |
| HDR-Guard | 127/200 | 156/200 | 50/200 | 333/600 |
| HDR-Guard − soft saturation blend | −25 [−44, −9] | +24 [+2, +49] | +3 [−17, +29] | +2 [−34, +40] |
| HDR-Guard − full-frame Mertens | +83 [+56, +104] | +55 [+34, +78] | +46 [+19, +73] | +184 [+140, +226] |
| HDR-Guard − policy camera (raw) | +124 [+99, +147] | +144 [+114, +171] | +50 [+23, +77] | +318 [+271, +363] |

### F.2 STATE-LEVEL DISCORDANCE

Pairing episodes that share suite, task and initial state, Table 14 counts the states that both rules solve, that only one solves, and that neither solves. The exact $p$ is a two-sided sign test on the discordant states. It treats states as independent, although states are clustered within tasks, so it is descriptive; the task-cluster intervals above are the inferential summary.

**Table 14:** State-level outcomes of HDR-Guard against a comparison rule, OpenVLA, centered hard patch.

| Suite | Comparison rule | Pairs | Both | HDR-Guard only | Rule only | Neither | Exact $p$ |
|---|---|---|---|---|---|---|---|
| Spatial | soft blend | 500 | 246 | 89 | 59 | 106 | 0.017 |
| Spatial | no closing | 500 | 239 | 96 | 50 | 115 | <0.001 |
| Object | soft blend | 200 | 105 | 22 | 47 | 26 | 0.004 |
| Goal | soft blend | 200 | 114 | 42 | 18 | 26 | 0.003 |
| LIBERO-10 | soft blend | 200 | 23 | 27 | 24 | 126 | 0.780 |

## G UNSYNCHRONIZED PROTECTED STREAM

A second camera without a hardware trigger delivers frames that are offset in time from the policy camera. We model this by writing, at control step $t$, the protected frame rendered at step $t - k$, $k \in \{0, 1, 3, 10\}$, while the policy frame and the mask stay current. LIBERO steps at 20 Hz, so the offsets are 0–500 ms; during the first $k$ steps of an episode the protected stream repeats its first frame. The robot and the objects move between the two frames, so inside the mask the written content is displaced wherever the scene changed, which is the misregistration term of Proposition 4 with a motion-dependent $\delta$. The run uses OpenVLA on the centered hard patch, LIBERO-Spatial tasks 0–9, 50 trials per task, with sensor noise seeded from each stream's own frame; the no-delay column is rerun in the same code path and is the reference for the differences. A delay of one control step did not reduce success (+26 [+4, +48]). This simulation evaluates trigger-free acquisition at controlled frame ages: at one 20 Hz control step, HDR-Guard reaches 362/500, against 336/500 at zero age. At 150 ms the change is −38 [−83, +3], an interval that still includes zero, and at 500 ms it is −228 [−328, −126]. The loss at larger delays is consistent with temporal misregistration: the robot moves inside the lit region, and the written content shows it where it was. Proposition 4 covers the part of this change that a displacement of the signal represents.

**Table 15:** OpenVLA with a delayed protected stream, centered hard patch, LIBERO-Spatial tasks 0–9, 500 episodes per cell. Differences use the paired task-cluster bootstrap (95%). The no-delay column is an independent rerun through the delayed-frame code path, not the run of Table 12, so its totals differ.

| Protected-frame delay | 0 steps (0 ms) | 1 step (50 ms) | 3 steps (150 ms) | 10 steps (500 ms) |
|---|---|---|---|---|
| HDR-Guard | 336/500 | 362/500 | 298/500 | 108/500 |
| Soft saturation blend | 290/500 | 312/500 | 265/500 | 90/500 |
| HDR-Guard, change from no delay | – | +26 [+4, +48] | −38 [−83, +3] | −228 [−328, −126] |
| HDR-Guard − soft blend | +46 [−15, +115] | +50 [−5, +123] | +33 [−17, +93] | +18 [−64, +104] |

## H CLIPPING IN REAL CAMERA FOOTAGE

To check that clipped task content is not an artifact of simulation, we measured it in real footage: 178,200 frames (297 clips of 20 s at 30 fps, 13 sessions) recorded by the auto-exposure RGB camera of a ground-level multi-camera sensor rig during firefighting drills at outdoor live-fire stations, in an indoor burn room and in a shipping-container scenario. Every frame carries automatically generated masks for persons, fire and smoke; we use only detections not flagged as unreliable, and we do not count mannequins or human-shaped cardboard targets as persons. A pixel counts as clipped when its largest channel reaches 250 or 254 of 255; the seed of HDR-Guard ($\tau = 0.985$, i.e. 252/255 on 8-bit values) lies between the two.

**Table 16:** Clipping in 178,200 frames of real auto-exposure camera footage. Percentages of frames or of annotated pixels. "All channels" counts a pixel only when all three channels reach 250 (fully white); that column uses every fifth frame (35,640 frames, 13,132 with a person).

| | Largest channel $\geq 250$ | Largest channel $\geq 254$ | All channels $\geq 250$ |
|---|---|---|---|
| Frames with $> 1\%$ of the image clipped | 87.1 | 80.7 | 41.8 |
| Frames with $> 5\%$ of the image clipped | 11.5 | 4.7 | 0.7 |
| Median clipped area of a frame | 3.2 | 2.4 | 0.9 |
| Person pixels clipped | 16.5 | 13.3 | 5.0 |
| Frames with a person in which some person has $> 5\%$ clipped | 48.9 | 45.0 | 18.5 |
| Fire pixels clipped | 29.2 | 25.1 | – |
| Smoke pixels clipped | 13.3 | 9.4 | – |

Clipping is worst where a bright opening is seen from a dark interior: inside the container, 92.8% of frames have more than 5% of the image clipped. In the indoor burn room it comes with the fire, in 9.7% of fire frames against 0.06% of no-fire frames above 1% of the image. Near-saturation persists after eroding the person masks by 2 pixels (17.5% of pixels, against 16.5% before), so boundary spill alone does not explain it, although errors of the automatic masks remain possible. The frames are stored as JPEG, which slightly understates clipping: on 65 frames re-decoded from the uncompressed recordings, the mean clipped area is 3.55% against 3.42% for the JPEG frames. Masks are generated automatically, and the statistics pool all frames of the 297 clips.

## I REAL BRACKETED EXPOSURES

To check the image-level premise on a real sensor, we used tripod-mounted RAW captures from a consumer DSLR at ISO 400 and f/8: a static indoor scene with a strong light source, 20 frames at each of 1/2 s, 1/30 s and 1/500 s. All statistics use black-subtracted linear Bayer planes without demosaicing. The saturation level (13,256 DN) and the Poisson–Gaussian noise model (variance $= a\,s + b$ for signal $s$ in DN, $a \approx 2.4$ at ISO 400) come from dark, flat and exposure-sweep calibration frames of the same camera. One 1/2 s frame serves as the policy stream $H$ and one 1/30 s frame as the protected stream $L$. The seed of Equation (2) is applied across the four Bayer planes, and $L$, scaled by the exposure ratio, is compared with an independent reference: the 20-frame mean of the 1/500 s exposure, scaled to 1/2 s, on pixels where that reference is unclipped and above 100 DN and $L$ is unclipped and above 5% of full scale. The frames are registered by construction, so the budget of Proposition 4 reduces to bracket mismatch, noise and protected saturation. Clipping fractions average all 20 frame pairs; error statistics average five of them (every fourth).

- The long exposure clips 36.4% of the frame; in 7.6% of its clipped pixel-channels the protected frame is clipped too, the protected-saturation term, whose nominal value is not identifiable from these two recorded samples alone.
- Where neither 20-frame mean clips, the pair is linear, but its effective exposure ratio is 16.3 (5–95%: 15.8–16.8) rather than the 15.0 implied by the recorded exposure times. For this camera setting and shutter pair, the median effective ratio is 9% above the ratio implied by the recorded exposure times; a single global factor, fitted on this capture, removes the median mismatch.
- After that factor, valid unclipped protected-frame samples differ from the reference by a median 4.8% per sample (90th percentile 21.7%), against 3.8% predicted from the measured noise of the two frames alone.

After subtracting the camera's 2,048-DN black offset, the usable saturation swing is about 11,200 DN; with a conversion gain of 0.41 $e^-$/DN at ISO 400, this corresponds to about 4,600 electrons. The simulator's 1,000-electron full well (Appendix B) therefore gives about $2.1\times$ greater relative shot noise at full scale.

## J RECONSTRUCTION OF THE NOMINAL FRAME

We rendered the agent view of LIBERO-Spatial, Object, Goal and LIBERO-10, tasks 0–9, at initial states 0, 5, . . . , 45 after the benchmark's ten settle steps (400 frames), and passed each frame through the sensor model of Appendix B under three scenes: the centered hard patch, the offset hard patch and the soft spotlight, all with $G = r = 8$. Table 17 lists the mean absolute error of each rule against the clean nominal frame, in 8-bit gray levels, over the lit region (normalized patch profile above one half), the lit pixels whose policy value stays below $\tau$ in every channel (the pixels the seed misses) and the whole frame. The last row is the error of a nominal-exposure frame from the same sensor, the sensor-noise reference. Differences between rules use a cluster bootstrap over the 40 (suite, task) clusters. On the centered patch, the pixels that the closing adds to the replacement region, $V = \mathrm{dilate}_{3\times3}(\mathrm{close}_B(S_H)) \setminus \mathrm{dilate}_{3\times3}(S_H)$, cover 1.1% of the frame and are 85% lit. Pooling pixels over all frames, HDR-Guard and the soft blend err there by 20 and 115 gray levels, so $V$ carries 75% of the whole-frame advantage of HDR-Guard over the soft blend; the one-pixel ring around the seed carries 19% and the near-clip band outside the replacement region $R_H$ 7% (Appendix K), and on the seed itself the two rules coincide. All variants are built from the same $H, L$ arrays, and the pixel-pooled contributions of these mutually exclusive regions sum to the whole-frame difference before rounding; the percentages are rounded independently.

**Table 17:** Mean absolute error against the nominal frame (8-bit gray levels) over 400 rendered LIBERO frames. "Lit, missed": lit pixels below $\tau$ in every channel of the policy stream; they are 9% of the centered patch, 2% of the offset patch and 10% of the soft spotlight.

| | Centered hard patch | | | Offset hard patch | | | Soft spotlight | | |
|---|---|---|---|---|---|---|---|---|---|
| Input rule | Lit | Lit, missed | All | Lit | Lit, missed | All | Lit | Lit, missed | All |
| HDR-Guard | 6.1 | 11.9 | 6.5 | 5.3 | 4.1 | 6.1 | 12.9 | 17.8 | 16.8 |
| HDR-Guard, no closing | 11.5 | 50.6 | 7.6 | 5.8 | 14.2 | 6.0 | 17.9 | 51.9 | 17.9 |
| Soft saturation blend | 15.7 | 118.9 | 7.9 | 7.1 | 93.5 | 5.4 | 22.0 | 112.6 | 18.3 |
| Localized Mertens (within $R_H$) | 77.3 | 77.5 | 21.7 | 78.6 | 65.6 | 14.9 | 73.1 | 71.7 | 24.8 |
| Full-frame Mertens | 76.6 | 73.3 | 47.5 | 78.6 | 64.8 | 36.7 | 71.9 | 64.3 | 48.8 |
| HDR merge + Reinhard | 35.2 | 20.1 | 76.3 | 42.6 | 22.4 | 75.5 | 30.0 | 14.3 | 73.0 |
| Exposure selection, as recorded | 4.6 | 2.0 | 80.8 | 5.2 | 1.9 | 87.4 | 10.8 | 4.0 | 80.0 |
| Exposure selection, rescaled | 130.1 | 137.3 | 40.4 | 118.5 | 117.1 | 27.4 | 127.7 | 127.5 | 47.9 |
| Policy camera (raw) | 130.2 | 137.7 | 34.1 | 118.6 | 117.3 | 20.4 | 127.8 | 127.7 | 42.9 |
| Nominal exposure (sensor noise only) | 4.1 | 1.8 | 4.2 | 4.5 | 1.6 | 4.2 | 4.1 | 1.9 | 4.2 |

## K WHICH PIXELS MOVE THE POLICY

The replacement region of HDR-Guard splits into the dilated seed $\mathrm{dilate}_{3\times3}(S_H)$ and the pixels $V$ that the closing adds (Appendix J). To find which pixels carry the closed-loop difference between

HDR-Guard and the soft blend, we crossed two factors on identical initial states, with common random numbers, for OpenVLA under the centered hard patch: the value written on $V$ (the protected value or the soft blend) and the values written everywhere else (HDR-Guard's, which are the protected value on the dilated seed and $H/g$ outside the replacement region $R_H$, or the soft blend's). The four cells are HDR-Guard, HDR-Guard with the soft blend on $V$, the soft blend with the protected value on $V$, and the soft blend. We ran tasks 0–9 of the four LIBERO suites with per-suite checkpoints, 50 initial states per task on LIBERO-Spatial and 20 on the other suites (1,100 paired states, Table 18). We estimate each factor effect by averaging its two matched contrasts within each initial state and weighting the four suites equally; all four outcomes of an initial state stay together when tasks are resampled within each suite, and 95% intervals use 200,000 bootstrap replicates. Table 18 also lists pooled counts.

Writing the protected value rather than the soft blend on $V$ gives an equal-suite success increase of +3.4 [−0.4, +7.7] percentage points, with suite effects of +4.2, −3.0, +7.0, +5.3 points on Spatial, Object, Goal and LIBERO-10. The point estimate is positive under either outside-$V$ assignment: +4.0 points with HDR-Guard's values outside $V$ and +2.7 with the soft blend's. Changing the outside-$V$ assignment instead gives −1.6 [−5.2, +1.9] points. The hybrid with protected values on $V$ and the soft blend elsewhere exceeds the converse hybrid by +5.0 [+0.1, +10.0] points, with positive differences in all four suites (+2.4, +6.5, +2.0, +9.0). Across the 2,200 paired $V$ comparisons, 363 favor the protected value and 284 the soft blend; treating them as independent gives a descriptive two-sided binomial $p = 0.002$, and because each initial state contributes two comparisons and states are nested within tasks, inference uses the task-cluster intervals. These estimates agree with the reconstruction result, where $V$ accounts for 75% of the error reduction (Appendix J). With HDR-Guard's values fixed outside $V$ on LIBERO-Spatial, success is 335/500 with the protected value on $V$, 297/500 with the soft blend on $V$ and 289/500 when the closing is removed and $V$ keeps $H/g$ (Appendix F).

Outside $V$ the two rules differ on the one-pixel ring $\mathrm{dilate}_{3\times3}(S_H) \setminus S_H$ (1.8% of the frame, 46% lit; errors 63 and 77) and on the near-clip band $U = \{p : 0.90 \le \max_c H(p, c) < \tau\} \setminus R_H$ (0.16% of the frame, 1% lit; errors 8 and 64). On an unlit pixel in the ideal model the soft blend introduces the error $w(g/r - 1)x$, and retaining $H/g$ avoids this partial-replacement error.

A further arm follows a path from the soft blend to HDR-Guard that changes one region at a time. It keeps $H/g$ on $U$, computed from the observed $H$, and uses the soft blend elsewhere; it was defined and launched on the same states before the extension's results were examined. Keeping $H/g$ on $U$ changes the pooled success count by $-12$ [−41, +17], then writing the protected value on the ring by $-11$ [−55, +36], and then writing it on $V$, which yields HDR-Guard, by +55 [−15, +139] (1,100 states with all five inputs, Table 18). The largest estimated increment along this path is on $V$; each increment conditions on the preceding assignments.

The advantage of HDR-Guard over the soft blend follows the area the closing adds. On the rendered frames, $V$ covers 1.66%, 1.57%, 0.73% and 0.32% of the frame on Goal, Spatial, LIBERO-10 and Object, and on the 1,100 paired states HDR-Guard minus the soft blend is +12.0, +6.0, +1.5 and −12.5 percentage points on the same suites, in the same order. Across the 40 tasks, the Spearman correlation between the area of $V$ and this difference is +0.49 (permutation $p = 0.001$; +0.49 and $p = 0.002$ for the lit pixels below $\tau$). This analysis was run after the region swaps, and the association is carried by differences between suites (within-suite permutation $p = 0.12$).

**Table 18:** Closed-loop $2 \times 2$ region swap, OpenVLA, centered hard patch ($G = r = 8$), common random numbers, identical initial states in every row. Effect rows: per-suite count differences; pooled count difference with a suite-stratified 40-cluster bootstrap; and the equal-suite effect in percentage points with a task bootstrap within suites (95%).

| Outside $V$ | On $V$ | Spatial | Object | Goal | LIBERO-10 | Pooled | Equal-suite effect, pp [95% CI] |
|---|---|---|---|---|---|---|---|
| HDR-Guard | protected | 335/500 | 127/200 | 156/200 | 50/200 | 668/1100 | |
| HDR-Guard | soft blend | 297/500 | 138/200 | 136/200 | 42/200 | 613/1100 | |
| soft blend | protected | 309/500 | 151/200 | 140/200 | 60/200 | 660/1100 | |
| soft blend | soft blend | 305/500 | 152/200 | 132/200 | 47/200 | 636/1100 | |
| Protected minus soft on $V$ (sum of two contrasts) | | +42 | −12 | +28 | +21 | +79 [−23, +215] | +3.4 [−0.4, +7.7] |
| HDR-Guard minus soft outside $V$ (sum of two contrasts) | | +18 | −38 | +20 | −15 | −15 [−95, +74] | −1.6 [−5.2, +1.9] |
| Hybrid: protected on $V$ minus converse | | +12 | +13 | +4 | +18 | +47 [−19, +120] | +5.0 [+0.1, +10.0] |
| *Path from the soft blend to HDR-Guard, one region at a time (1,100 states, pooled counts)* | | | | | | | |
| Keep $H/g$ on the near-clip band $U$ | | −18 | +2 | −8 | +12 | −12 [−41, +17] | |
| Then protected value on the ring | | +10 | −16 | +12 | −17 | −11 [−55, +36] | |
| Then protected value on $V$ | | +38 | −11 | +20 | +8 | +55 [−15, +139] | |

## L PER-TASK BREAKDOWN OF THE HARD PATCH

The closed-loop bootstrap resamples only ten tasks, so Table 19 lists them. HDR-Guard is at or above the soft blend on nine of the ten tasks, and the largest single-task gap is three episodes (task 7), so the six-episode total is spread over several tasks rather than driven by one.

**Table 19:** Centered hard patch ($G = r = 8$), successes per LIBERO-Spatial task summed over the three policies (out of 9), first run.

| Task | Raw | Mertens | Soft blend | No closing | HDR-Guard | HDR-Guard−soft |
|---|---|---|---|---|---|---|
| 0 | 5 | 7 | 8 | 9 | 9 | +1 |
| 1 | 1 | 2 | 6 | 5 | 7 | +1 |
| 2 | 0 | 7 | 8 | 8 | 9 | +1 |
| 3 | 1 | 6 | 9 | 8 | 9 | +0 |
| 4 | 1 | 4 | 7 | 6 | 7 | +0 |
| 5 | 2 | 4 | 5 | 4 | 5 | +0 |
| 6 | 0 | 7 | 9 | 9 | 8 | −1 |
| 7 | 2 | 8 | 5 | 7 | 8 | +3 |
| 8 | 5 | 6 | 7 | 7 | 8 | +1 |
| 9 | 4 | 6 | 6 | 6 | 6 | +0 |
| Total | 21 | 57 | 70 | 69 | 76 | +6 |

## M HARD PATCH ON OTHER LIBERO SUITES

To test whether the ordering of the rules under the hard patch transfers beyond LIBERO-Spatial for $\pi_{0.5}$, we ran it on tasks 0–4 of LIBERO-Object, LIBERO-Goal and LIBERO-10 under the hard-edged patch, three trials per task (Table 20). We used $\pi_{0.5}$ because one released checkpoint covers all four suites, whereas our OpenVLA-OFT and OpenVLA runs load a separate checkpoint per suite. Over the three suites the raw camera keeps 19 of 45 episodes (nominal exposure: 44). The soft blend keeps all 45, HDR-Guard without the closing 42, HDR-Guard 40 and Mertens fusion 39. For $\pi_{0.5}$ on these suites the closing does not help and the soft blend is the strongest rule, in line with LIBERO-Spatial, where the closing changed $\pi_{0.5}$ by two episodes and its main effect was on OpenVLA.

**Table 20:** $\pi_{0.5}$ under the hard-edged patch ($G = r = 8$) on three further LIBERO suites, tasks 0–4, three trials per task (15 episodes per cell).

| Input | Object | Goal | LIBERO-10 | Total |
|---|---|---|---|---|
| Nominal exposure | 14/15 | 15/15 | 15/15 | 44/45 |
| Policy camera (raw) | 10/15 | 4/15 | 5/15 | 19/45 |
| Mertens exposure fusion | 15/15 | 13/15 | 11/15 | 39/45 |
| Soft saturation blend | 15/15 | 15/15 | 15/15 | 45/45 |
| HDR-Guard, no closing | 14/15 | 15/15 | 13/15 | 42/45 |
| HDR-Guard | 13/15 | 15/15 | 12/15 | 40/45 |

## N FRAME-LEVEL PHOTOMETRIC MEASUREMENTS

To check the clipping statistics outside LIBERO, we applied the controlled shift to 312 frames per level drawn from 39 robot-camera videos rendered in Isaac Sim for DynamicVLA (Xie et al., 2026). Table 21 reports the fraction of clipped pixels, the full-frame PSNR of the known-gain linear inverse (linear PSNR), the same PSNR restricted to the clipped region, and the fraction of local texture variance that survives on clipped pixels. The protected stream clips 0.02% of pixels at every level.

**Table 21:** Frame-level clipping statistics on 312 frames per level.

| Shift | Policy clipped | Protected clipped | Linear PSNR | Clipped-region PSNR | Texture kept |
|---|---|---|---|---|---|
| EV+2 | 87.02% | 0.02% | 9.46 dB | 8.79 dB | 7.6% |
| EV+3 | 94.03% | 0.02% | 7.03 dB | 6.75 dB | 2.7% |
| EV+4 | 96.83% | 0.02% | 6.00 dB | 5.86 dB | 1.0% |

## O PERTURBATIONS OUTSIDE THE CLIPPING MODEL

HDR-Guard acts inside a mask grown from near-clipped pixels; when every observed channel stays below $\tau$, the mask is empty and HDR-Guard returns $H/g$, which equals $H$ for $g = 1$. We report two such rows from the controlled protocol for completeness. They do not evaluate HDR-Guard: because its mask is empty, we tested exploratory two-stream blends instead, a 3:1 blend of the noise-free protected frame and the noisy policy frame for ISO noise, and a 1:1 blend of the protected frame and the white-balance-corrected policy frame for the color shift. Under Gaussian ISO-6400-like noise ($\sigma = 0.08$), the raw total is 86/90 and the blended input 82/90 (OpenVLA-OFT 30/30 and 30/30, OpenVLA 27/30 and 22/30, $\pi_{0.5}$ 29/30 and 30/30). Under a warm white-balance shift, the totals are 81/90 and 82/90. Neither perturbation removes information by clipping; Proposition 4 places additive noise in the noise term, which is why the sensor-aware simulation models noise explicitly in both streams.

## P PAIRED ACTION CONSISTENCY: ALL RULES

Table 22 lists every rule evaluated on the clean-rollout states. $D_1$ is averaged within each of the ten clean rollouts and then across rollouts, and the 95% intervals resample the ten rollouts, because query states within one rollout are correlated. The normalizing 1st–99th percentile range is computed per policy and per arm dimension from that policy's clean actions. The first row re-queries the clean input: it is zero for OpenVLA-OFT and OpenVLA, which are deterministic, so every nonzero entry for them is caused by the input rule, and it gives the sampling floor of $\pi_{0.5}$.

**Table 22:** Paired action consistency for every rule: $D_1 \times 100$ (rollout-averaged mean, 95% interval from resampling rollouts) and gripper-flip rate (%). OpenVLA-OFT: 131 states; OpenVLA: 110; $\pi_{0.5}$: 219.

| | OpenVLA-OFT | | OpenVLA | | $\pi_{0.5}$ | |
|---|---|---|---|---|---|---|
| Input on the same state | $D_1$ [95% CI] | grip. | $D_1$ [95% CI] | grip. | $D_1$ [95% CI] | grip. |
| Clean input, queried again | 0.00 [0.00, 0.00] | 0 | 0.00 [0.00, 0.00] | 0 | 2.19 [1.98, 2.42] | 0 |
| Nominal exposure with sensor noise | 0.60 [0.52, 0.69] | 0 | 3.05 [2.05, 4.01] | 0 | 2.60 [2.28, 2.98] | 0 |
| Controlled EV+3: raw | 11.67 [10.79, 12.41] | 3 | 19.54 [17.14, 21.70] | 39 | 18.01 [16.44, 19.39] | 16 |
| Controlled EV+3: HDR-Guard | 0.00 [0.00, 0.00] | 0 | 0.00 [0.00, 0.00] | 0 | 2.10 [1.90, 2.34] | 0 |
| Controlled EV+3: linear inverse | 11.74 [10.90, 12.51] | 4 | 19.54 [16.74, 22.24] | 41 | 18.19 [16.30, 19.90] | 12 |
| Controlled EV+3: histogram eq. | 11.55 [10.80, 12.23] | 5 | 17.76 [13.75, 21.45] | 40 | 18.24 [16.73, 19.53] | 17 |
| Controlled EV+3: gamma 0.5 | 11.98 [10.97, 12.87] | 3 | 19.49 [16.86, 22.27] | 47 | 18.73 [16.77, 20.46] | 14 |
| Controlled EV+3: percentile stretch | 11.72 [10.81, 12.49] | 3 | 20.26 [17.26, 22.98] | 42 | – | – |
| Controlled EV+3: local hist. eq. | 11.65 [10.75, 12.40] | 3 | 20.37 [17.95, 22.47] | 40 | – | – |
| Hard patch: raw | 8.32 [7.77, 8.90] | 1 | 16.19 [13.11, 19.42] | 27 | 8.28 [7.49, 9.10] | 1 |
| Hard patch: exp. selection | 7.68 [7.16, 8.24] | 0 | 13.22 [11.08, 15.47] | 6 | 6.89 [6.12, 7.65] | 3 |
| Hard patch: exp. sel. rescaled | 8.61 [7.97, 9.22] | 1 | 14.95 [12.29, 17.73] | 31 | – | – |
| Hard patch: Mertens | 4.81 [4.40, 5.23] | 1 | 9.63 [8.03, 11.52] | 7 | 5.66 [5.04, 6.32] | 0 |
| Hard patch: Reinhard | 6.80 [6.27, 7.30] | 1 | 12.04 [10.24, 13.75] | 8 | 6.44 [5.67, 7.21] | 2 |
| Hard patch: soft blend | 3.41 [3.01, 3.83] | 1 | 8.41 [7.34, 9.62] | 4 | 4.74 [4.41, 5.10] | 0 |
| Hard patch: HDR-Guard, no closing | 3.39 [3.01, 3.78] | 2 | 9.00 [7.22, 11.27] | 2 | 4.56 [4.19, 4.92] | 0 |
| Hard patch: HDR-Guard | 3.32 [2.92, 3.75] | 1 | 8.37 [6.72, 9.98] | 4 | 4.75 [4.32, 5.20] | 0 |
| Soft spot: raw | 7.71 [6.99, 8.46] | 2 | 13.19 [10.80, 15.65] | 17 | – | – |
| Soft spot: Mertens | 3.88 [3.45, 4.34] | 2 | 7.39 [6.51, 8.31] | 3 | – | – |
| Soft spot: soft blend | 5.52 [4.94, 6.04] | 2 | 11.45 [9.18, 14.08] | 10 | – | – |
| Soft spot: HDR-Guard, no closing | 5.64 [5.02, 6.20] | 2 | 12.57 [10.78, 14.39] | 11 | – | – |
| Soft spot: HDR-Guard | 5.73 [5.09, 6.33] | 2 | 12.38 [11.08, 13.91] | 10 | – | – |
| Well-exposed: Mertens | 0.59 [0.51, 0.69] | 0 | 2.30 [1.86, 2.79] | 0 | 2.66 [2.32, 3.09] | 0 |
| Well-exposed: Reinhard | 1.03 [0.91, 1.17] | 1 | 3.72 [2.47, 4.89] | 2 | 3.14 [2.69, 3.64] | 1 |
| Well-exposed: HDR-Guard, no closing | 0.73 [0.67, 0.80] | 0 | 3.90 [2.77, 4.97] | 1 | – | – |
| Well-exposed: HDR-Guard | 0.72 [0.66, 0.80] | 0 | 3.19 [2.29, 4.05] | 1 | 2.59 [2.32, 2.87] | 1 |
| Sensor-aware EV+3: raw | 11.62 [10.79, 12.29] | 3 | 20.66 [17.99, 23.43] | 45 | – | – |
| Sensor-aware EV+3: exp. selection | 0.58 [0.52, 0.64] | 0 | 2.34 [1.56, 3.22] | 0 | – | – |
| Sensor-aware EV+3: Mertens | 0.81 [0.73, 0.89] | 1 | 3.39 [2.59, 4.18] | 0 | – | – |
| Sensor-aware EV+3: HDR-Guard, no closing | 0.56 [0.49, 0.64] | 0 | 2.93 [1.79, 4.17] | 0 | – | – |

## Q LOCALIZED MERTENS ON THE ACTION METRIC

Table 23 repeats the paired action diagnostic in a separate clean rollout of tasks 0–9 that includes localized Mertens, run on a different GPU, so its values differ slightly from Table 22. For OpenVLA-OFT the ordering matches the closed-loop results: on the two matched patches HDR-Guard and the soft blend deviate least, localized Mertens lies between them and full-frame Mertens, and under the soft spotlight the order reverses. For OpenVLA the four rules overlap on the matched patches in this pass, and under the soft spotlight full-frame Mertens again deviates least.

**Table 23:** Paired action consistency with localized Mertens ($D_1 \times 100$, rollout-cluster 95% interval), from its own clean rollout.

| Input on the same state | OpenVLA-OFT $D_1$ [95% CI] | OpenVLA $D_1$ [95% CI] |
|---|---|---|
| *Hard-edged patch* | | |
| HDR-Guard | 3.46 [3.03, 3.94] | 9.48 [8.23, 10.71] |
| Soft saturation blend | 3.51 [3.10, 3.97] | 9.70 [8.21, 11.31] |
| Localized Mertens (within $R_H$) | 4.45 [3.98, 4.95] | 9.08 [7.65, 10.74] |
| Full-frame Mertens | 4.89 [4.40, 5.42] | 9.13 [7.99, 10.33] |
| *Offset patch* | | |
| HDR-Guard | 3.00 [2.66, 3.37] | 7.66 [6.74, 8.65] |
| Soft saturation blend | 2.67 [2.30, 3.05] | 7.18 [6.10, 8.25] |
| Localized Mertens (within $R_H$) | 3.87 [3.52, 4.24] | 8.16 [7.06, 9.28] |
| Full-frame Mertens | 4.22 [3.80, 4.63] | 8.25 [7.37, 9.14] |
| *Soft spotlight* | | |
| HDR-Guard | 5.83 [5.17, 6.47] | 12.46 [10.71, 14.48] |
| Localized Mertens (within $R_H$) | 4.66 [4.21, 5.15] | 10.33 [8.86, 11.73] |
| Full-frame Mertens | 3.89 [3.45, 4.35] | 8.40 [7.42, 9.35] |

## R SENSITIVITY TO THE CLOSING SIZE AND THE SOFT-BLEND RAMP

The closing size of HDR-Guard ($15 \times 15$) and the ramp start of the soft blend (0.90) were each set once, without a sweep. Table 24 varies both on the paired action metric of Section 4.2 for OpenVLA-OFT, from a separate clean rollout of tasks 0–9, and adds the offset confirmation patch. On the hard patch, $D_1$ stays between 3.26 and 3.37 for closing sizes from $7 \times 7$ to $21 \times 21$ and between 3.34 and 3.53 for ramp starts from 0.85 to 0.97, with overlapping intervals, so neither setting sits at a sharp optimum. On the soft spotlight the closing size moves $D_1$ by less than 0.1. On the offset patch the two local rules deviate least (soft blend 2.57, HDR-Guard 2.86), ahead of Mertens fusion (4.16) and the raw frame (6.59).

**Table 24:** Paired action consistency of OpenVLA-OFT under varied closing sizes and soft-blend ramps ($D_1 \times 100$, rollout-cluster 95% interval; grip.: gripper flips, %).

| Input on the same state | $D_1$ [95% CI] | grip. |
|---|---|---|
| *Hard-edged patch* | | |
| HDR-Guard, closing $7 \times 7$ | 3.37 [2.99, 3.76] | 1 |
| HDR-Guard, closing $11 \times 11$ | 3.33 [2.95, 3.74] | 1 |
| HDR-Guard, closing $15 \times 15$ (used) | 3.32 [2.92, 3.76] | 1 |
| HDR-Guard, closing $21 \times 21$ | 3.26 [2.88, 3.68] | 1 |
| HDR-Guard, no closing | 3.39 [3.02, 3.77] | 2 |
| Soft blend, ramp from 0.85 | 3.53 [3.17, 3.90] | 1 |
| Soft blend, ramp from 0.90 (used) | 3.41 [3.01, 3.83] | 1 |
| Soft blend, ramp from 0.95 | 3.34 [2.95, 3.75] | 1 |
| Soft blend, ramp from 0.97 | 3.34 [2.94, 3.75] | 1 |
| *Soft spotlight* | | |
| HDR-Guard, closing $7 \times 7$ | 5.79 [5.24, 6.31] | 2 |
| HDR-Guard, closing $15 \times 15$ (used) | 5.73 [5.08, 6.33] | 2 |
| HDR-Guard, closing $21 \times 21$ | 5.77 [5.14, 6.35] | 2 |
| *Offset confirmation patch* | | |
| Policy camera (raw) | 6.59 [5.93, 7.38] | 3 |
| Mertens exposure fusion | 4.16 [3.82, 4.50] | 1 |
| Soft saturation blend | 2.57 [2.25, 2.90] | 1 |
| HDR-Guard, no closing | 2.90 [2.55, 3.24] | 1 |
| HDR-Guard | 2.86 [2.53, 3.21] | 1 |