EMBODIED NEURAL IMAGE SIGNAL PROCESSOR FOR ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

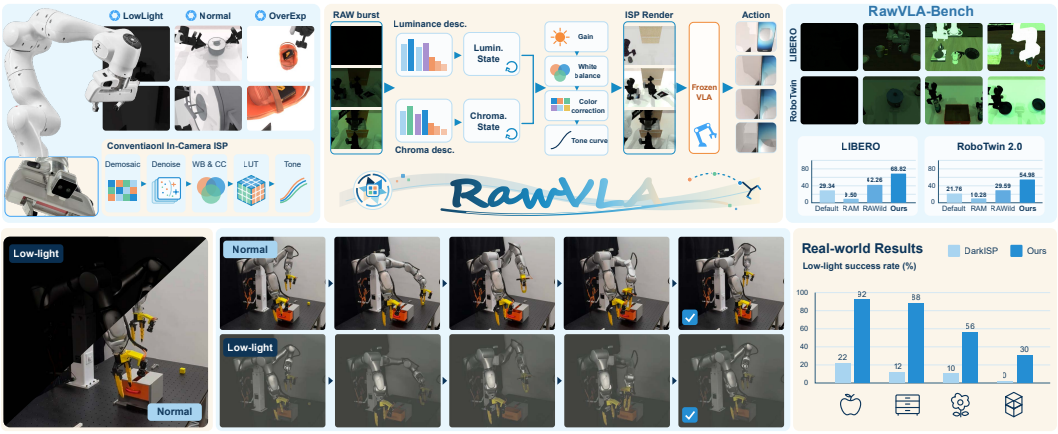


Figure 1: We propose RawVLA, an embodied and adaptive neural ISP module for VLA models, and RawVLA-Bench, a RAW-domain manipulation benchmark based on LIBERO and RoboTwin 2.0.

ABSTRACT

Vision-language-action (VLA) models typically operate on RGB images produced by a fixed camera image signal processor (ISP), leaving the imaging pipeline outside the learning and evaluation loop. We systematically examine the consequences of this overlooked design choice across five fundamental ISP dimensions: gain, sensor noise, chromatic response, tonal response, and bit depth. Our analysis reveals that RAW-to-RGB processing materially shapes both action prediction and manipulation success, with different ISP dimensions exerting substantially different effects. Guided by these findings, we introduce RawVLA, a lightweight neural ISP that adaptively renders RAW observations for frozen VLA policies while concentrating its capacity on the imaging factors most relevant to embodied behavior. We further present RawVLA-Bench, a RAW-domain manipulation benchmark built upon LIBERO and RoboTwin to expose image processing as an explicit evaluation variable across clean and adverse acquisition conditions. Experiments on RawVLA-Bench show that RawVLA preserves performance under standard conditions while substantially improving robustness under degraded imaging, establishing adaptive RAW processing as an effective interface between physical cameras and embodied policies. Code and dataset will be open-sourced.

1 INTRODUCTION

Developing agents that can perceive, reason, and act in the physical world is a central goal of embodied AI and physical intelligence (Zitkovich et al., 2023; Black et al., 2024). Vision-language-action (VLA) models offer a promising path toward this goal by unifying visual observations, language instructions, and robot actions (Zitkovich et al., 2023; Kim et al., 2025b; Black et al., 2024; Physical Intelligence et al., 2025). Recent systems further scale continuous action generation, cross-embodiment learning, and embodied reasoning through diffusion transformers and generalist robot

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.

foundation models (Hou et al., 2025; Bjorck et al., 2025; Gemini Robotics Team et al., 2025). Leveraging vision-language pretraining and robot demonstrations, they can acquire generalizable manipulation capabilities across tasks, scenes, and robot embodiments (Octo Model Team et al., 2024; Kim et al., 2025b; Li et al., 2024; Physical Intelligence et al., 2025). However, existing VLA research typically begins with camera-rendered RGB images (Zitkovich et al., 2023; Kim et al., 2025b) and abstracts the visual pathway as *RGB-to-action*. This abstraction overlooks the physical imaging pipeline that precedes model inference (Brooks et al., 2019; Wu et al., 2019). Light from a scene is captured as RAW sensor measurements and then transformed into RGB images by an image signal processor (ISP) (Brooks et al., 2019; Yu et al., 2021). Consequently, systems implicitly treat the ISP as a fixed and task-neutral preprocessing component (Diamond et al., 2021; Huang et al., 2026).

This assumption can constrain downstream robot control. ISPs adjust exposure, white balance and color, tone, denoising, and quantization (Brooks et al., 2019), potentially discarding signals needed for manipulation (Diamond et al., 2021; Huang et al., 2026). Embodied Image Compression (Li et al., 2025a) and SPARC (Kim et al., 2026a) further show that visual processing should be evaluated by closed-loop task performance rather than perceptual fidelity. Consequently, an unchanged scene can induce different VLA behaviors across ISP configurations (Wang et al., 2024a; Huang et al., 2026). This issue is especially important under low light, high dynamic range, backlighting, and unusual illumination (Guo et al., 2025; Fei et al., 2026), where clipping, noise, or aggressive tone compression can obscure task-relevant visual details (Chen et al., 2018; Watanabe et al., 2026).

RAW observations offer a natural opportunity to revisit this overlooked part of the VLA system (Cui et al., 2026; Huang et al., 2026). Compared with a single RGB image produced by a fixed ISP, RAW measurements preserve sensor information that can be weakened or irreversibly lost during RGB rendering (Chen et al., 2018; Huang et al., 2026). They also retain the flexibility to select or learn a RAW-to-RGB transformation according to the scene, task, and downstream model (Yu et al., 2021; Wang et al., 2024a). These properties suggest that the imaging pipeline should be treated as an explicit and optimizable component of a VLA system rather than as hidden camera firmware (Diamond et al., 2021; Huang et al., 2026). Nevertheless, existing task-oriented ISP studies focus mainly on static vision tasks (Cui et al., 2026; Liu et al., 2026b), leaving open how strongly ISP choices affect VLA behavior and whether task-adaptive RAW-to-RGB processing can improve manipulation performance.

In this work, we conduct a systematic analysis of VLA sensitivity to five fundamental ISP dimensions: exposure, sensor noise, chromatic response, tonal response, and bit depth (Brooks et al., 2019; Yu et al., 2021). By varying one dimension at a time while holding the underlying scene, task, and robot state fixed, we isolate the effect of RAW-to-RGB processing on downstream action prediction and task success. Our results show that ISP processing is far from task-neutral and that different ISP dimensions affect VLA models to substantially different degrees. These observations both motivate RAW-domain evaluation and provide direct guidance for designing an adaptive ISP.

We propose **RawVLA**, a lightweight neural ISP that adaptively renders RAW inputs for downstream VLA models, as illustrated in Figure 1. Unlike a conventional frozen ISP, RawVLA adapts behavior-relevant imaging factors without modifying the VLA backbone. To evaluate RAW-domain processing, we introduce **RawVLA-Bench**, a benchmark that exposes the RAW-to-RGB pipeline as an explicit degree of freedom. RawVLA-Bench spans clean and challenging acquisition conditions for evaluating robustness and RAW-domain processing headroom, complementing RGB robustness benchmarks (Fei et al., 2026; Chen et al., 2026b; Zhang et al., 2026).

We summarize our contributions below.

- We systematically analyze how the imaging pipeline affects VLA models through gain, sensor noise, chromatic response, tonal response, and bit depth. Our analysis demonstrates that ISP choices materially influence VLA action prediction.
- We construct **RawVLA-Bench**, a RAW-domain benchmark that makes image processing an explicit evaluation variable and supports controlled study of VLA robustness under clean and challenging imaging conditions.
- We propose **RawVLA**, a real-time neural ISP for VLA models. Extensive experiments on RawVLA-Bench and real-world tasks demonstrate state-of-the-art performance.



Figure 2: ISP sensitivity across simulation benchmarks. The left panel shows LIBERO results for fine-tuned VLA and WAM policies provided by StarVLA (StarVLA Community, 2026). The right panel shows Qwen3-OFT and FastWAM fine-tuned on RoboTwin 2.0 Easy and Hard (Hard adds randomized clutter, lighting, textures, and tabletop heights), which moderately improves robustness.

2 Analyzing VLA Sensitivity to ISP

To identify the imaging factors relevant to embodied actions, we conduct a controlled ISP sensitivity study on LIBERO and RoboTwin 2.0. We synthesize pseudo-RAW inputs from renderer buffers via parametric unprocessing (Brooks et al., 2019). Following standard camera-formation and ISP decompositions (Yu et al., 2021), we examine five principal factors governing signal level, acquisition noise, color, tone, and numerical precision. Holding the underlying pseudo-RAW observation R fixed, we isolate their effects through the factorized RAW-to-RGB mapping below, with detailed formulations provided in Appendices A and K:

$$Y_{e,\eta,\zeta,c_{\text{tone}},n} = \mathcal{Q}_n(\mathcal{T}_{c_{\text{tone}}}(\mathcal{C}_\zeta(\mathcal{N}_\eta(\mathcal{E}_e(R))))) \,. \tag{1}$$

Exposure. $\mathcal{E}_e$ scales the RAW signal to control shadow visibility and highlight saturation. As shown in Figure 2, exposure is the most influential axis. VLA models maintain high task performance only within a limited range around the default and collapse under severe underexposure or overexposure. World-action model (WAM) approaches have the narrowest stable range. Both model families on RoboTwin 2.0 remain vulnerable at extreme exposure settings despite fine-tuning with visual domain randomization.

Sensor Noise. $\mathcal{N}_\eta$ isolates sensor-noise sensitivity by varying noise independently of its physical coupling with illumination. VLA models tolerate mild noise after brightness restoration but degrade rapidly from moderate noise onward. WAMs degrade earlier and more sharply than VLA PI-family policies. On RoboTwin 2.0, Qwen3-OFT retains partial task success under high noise whereas FastWAM approaches complete failure.

Chromatic Response. $\mathcal{C}_\zeta$ changes chromatic rendering through absolute white-point shifts and relative hue and saturation transformations while preserving the underlying scene. VLA PI-family models benefit from visual augmentation during training and remain comparatively robust to these changes whereas WAMs exhibit stronger directional preferences. On LIBERO, the WAM agents reveal a pronounced color bias, preferring warmer renderings to cooler ones. After fine-tuning on the visually randomized RoboTwin 2.0 (Hard), Qwen3-OFT and FastWAM still exhibit white-point-dependent bias, despite remaining stable across relative-color transformations.

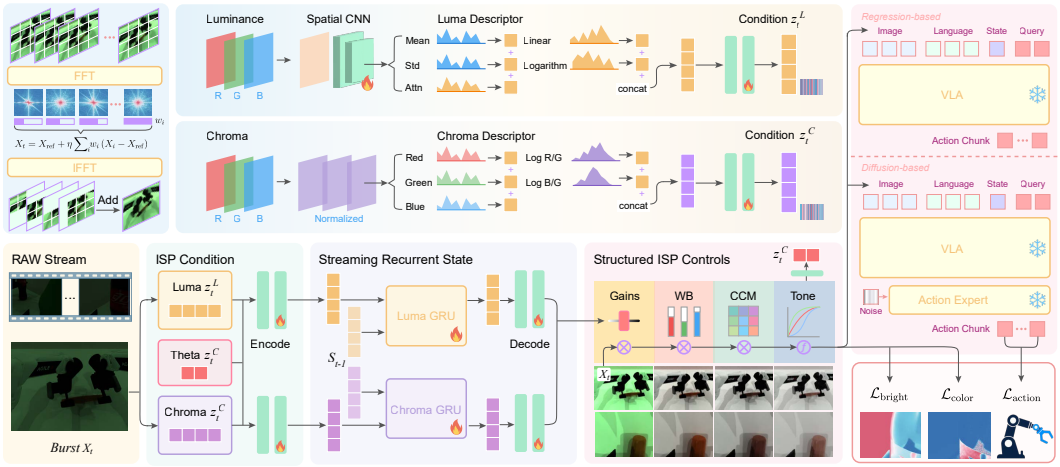


Figure 3: Overview of RawVLA. A causal linear RAW burst is summarized by factorized luminance and chroma descriptors, fused with previous ISP parameters through decoupled recurrent states, and decoded into structured exposure, white-balance, color-correction, and monotonic tone controls. The rendered image is passed to a frozen VLA policy for action prediction.

Tonal Response. $\mathcal{T}_{c_{\text{tone}}}$ controls the distribution of luminance contrast from shadows through mid-tones to highlights. VLA policies retain high success near the default response but degrade abruptly when the response becomes very flat or highly contrasted. WAMs are consistently more sensitive across the tonal range. Both model families on RoboTwin 2.0 respond more smoothly near the default setting but still lose substantial task performance at both extremes.

Bit Depth. $\mathcal{Q}_n$ quantizes the rendered output to n bits for digital image encoding. Performance remains largely stable above two bits and drops only under the most aggressive quantization. VLA models remain stable down to three bits and show a clear decline at two bits whereas WAMs begin to weaken earlier. Overall, bit depth affects both model families less than other ISP dimensions.

3 RawVLA

To account for the observed variation in sensitivity, we design RawVLA, a causal, structured ISP that selectively adapts behavior-relevant photometric controls while keeping the VLA policy frozen. We next describe its streaming formulation (Section 3.1), recurrent luminance–chroma conditioning (Section 3.2), structured decoding (Section 3.3), and action-driven optimization (Section 3.4).

3.1 Causal Streaming Formulation

RawVLA is a causal, task-driven ISP that transforms demosaiced linear RAW observations for a downstream VLA policy. It obtains the denoised RAW frame X_t from consecutive RAW frames through burst denoising while recurrently adapting factorized luminance and chroma controls. To preserve temporal context and avoid frame-to-frame rendering discontinuities, RawVLA formulates the streaming update as

$$(Y_t, S_t, \Theta_t) = \text{RawVLA}(X_t, S_{t-1}, \Theta_{t-1}), \tag{2}$$

where Y_t is the rendered observation passed to the VLA policy, S_t provides latent temporal context for illumination and color adaptation, and Θ_t is the explicit ISP parameter set controlling the current RAW-to-RGB processing.

3.2 Recurrent Photometric Conditioning

A central difficulty in learning an adaptive ISP is parameter identifiability, as global brightness can otherwise be explained by exposure, white balance, color correction, or channel-dependent tone

mapping (Brooks et al., 2019; Yu et al., 2021). RawVLA resolves this ambiguity by encoding orthogonal achromatic and chromatic conditions into decoupled recurrent states.

Luminance State. We compute an equal-channel luminance map L_t[1], from which we extract a compact spatial luminance feature g_t. We complement g_t with an absolute luminance descriptor h_t^L that summarizes intensity distributions, quantiles, and dark-clipped pixel ratios detailed in Appendix C. The logarithmic representation emphasizes low-intensity variation, whereas the linear histogram and clipping statistics capture over-exposure. The luminance recurrent state is then updated with a GRU (Cho et al., 2014) as

$$z_t^L = \phi_L(g_t, h_t^L, \Theta_{t-1}), \qquad S_t^L = \mathrm{GRU}_L(z_t^L, S_{t-1}^L). \tag{3}$$

Here, ϕ_L combines the current spatial feature g_t, the luminance descriptor h_t^L, and the previous ISP parameters Θ_{t-1}. The recurrent state S_t^L provides temporal context for predicting the current exposure and tone parameters.

Chroma State. To prevent global brightness from leaking into color estimation, the chroma pathway uses only scale-invariant channel statistics. We compute chromaticities $\boldsymbol{\chi}_t$ and log-chrominance ratios $\mathbf{q}_t$ directly from the RAW channels,

$$[\boldsymbol{\chi}_t]^k = \frac{X_t^k}{\sum_{j\in\{\mathrm{R,G,B}\}} X_t^j + \epsilon}, \quad k \in \{\mathrm{R,G,B}\}, \quad \mathbf{q}_t = \left(\log \frac{X_t^{\mathrm{R}} + \epsilon}{X_t^{\mathrm{G}} + \epsilon}, \log \frac{X_t^{\mathrm{B}} + \epsilon}{X_t^{\mathrm{G}} + \epsilon}\right). \tag{4}$$

We form the chroma descriptor h_t^C by concatenating histograms together with first- and second-order statistics, as detailed in Appendix C.4. We fuse h_t^C with the previous ISP parameters Θ_{t-1} through ϕ_C and update the chroma recurrent state as

$$z_t^C = \phi_C(h_t^C, \Theta_{t-1}), \qquad S_t^C = \mathrm{GRU}_C(z_t^C, S_{t-1}^C). \tag{5}$$

The pathway receives neither absolute luminance statistics nor RGB spatial features. Together, the luminance and chroma states form the full recurrent state $S_t = [S_t^L, S_t^C]$.

3.3 Structured Neural ISP Decoding

From the decoupled recurrent state $[S_t^L, S_t^C]$, RawVLA predicts the structured ISP parameter set $\Theta_t = \{e_t, w_t, A_t, \ell_t\}$, where e_t specifies exposure in EV, w_t encodes white balance, A_t denotes the 3×3 color-correction matrix, and $\ell_t \in \mathbb{R}^7$ parameterizes the tone curve. RawVLA first applies fixed six-frame burst denoising to linear RAW inputs, producing X_t as shown in Figure 3 (top left).

Exposure and Color Correction. We parameterize global exposure in EV space as $e_t = E_{\max}\tanh(\hat{e}_t)$, with gain 2^{e_t}, where $\hat{e}_t$ is the predicted exposure logit, following standard RAW processing conventions (Brooks et al., 2019). White balance uses zero-sum channel offsets and the corresponding diagonal gain matrix,

$$w_t = (w_t^{\mathrm{R}}, w_t^{\mathrm{G}}, w_t^{\mathrm{B}}) \quad \text{with} \sum w_t^k = 0, \qquad \text{and} \qquad W_t = \mathrm{diag}(2^{w_t^{\mathrm{R}}}, 2^{w_t^{\mathrm{G}}}, 2^{w_t^{\mathrm{B}}}), \tag{6}$$

which restricts white balance to relative channel gains while e_t controls global intensity. The two-coordinate parameterization is detailed in Appendix C.5. The color-correction matrix uses an identity-centered bounded residual, and the linear ISP output is processed as

$$A_t = \mathbb{I}_3 + \Delta A_t, \qquad Z_t = 2^{e_t} A_t W_t X_t. \tag{7}$$

ΔA_t is the predicted residual; white balance and color correction precede exposure adjustment.

Monotonic Achromatic Tone-mapping. RawVLA predicts a single tone curve shared across RGB channels, preventing the luminance pathway from introducing color casts. We use an eighth-degree Bernstein polynomial (Farouki, 2012). We set $p_0 = 0$ and construct the remaining control points directly from the predicted logits as $p_k = \sum_{j=1}^{k}[\mathrm{softmax}([\ell_t, 0])]_j$, yielding the curve

$$T_t(x) = \sum_{k=0}^{8} p_k \binom{8}{k} x^k (1-x)^{8-k} \tag{8}$$

[1]We define $L_t = \frac{1}{3}\sum_{k\in\{\mathrm{R,G,B}\}} X_t^k$ to capture achromatic intensity structure.

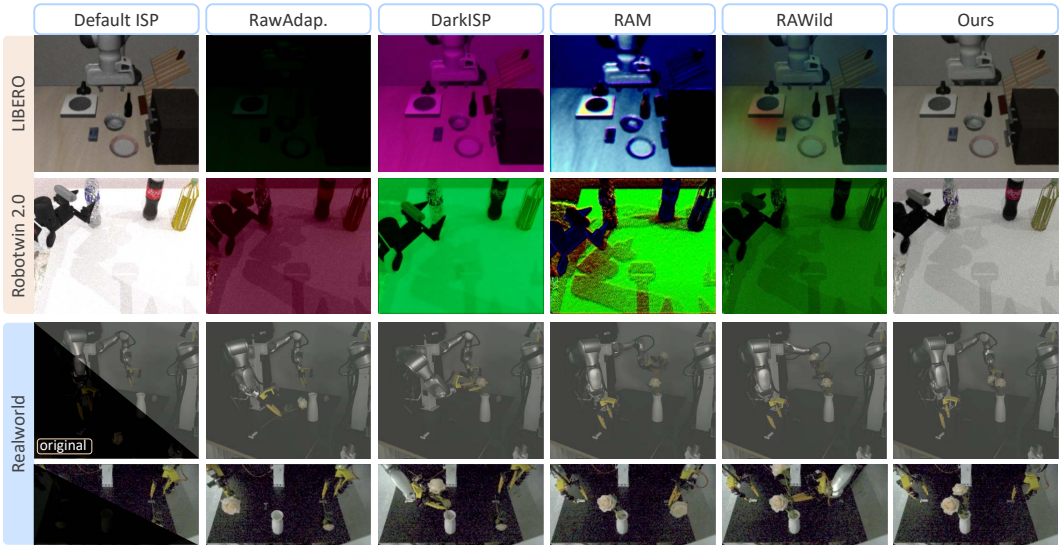


Figure 4: Qualitative results in simulation and the real world. The first and second rows compare agent-view observations rendered by the Default ISP and RawVLA on LIBERO and RoboTwin 2.0, respectively. The final row shows a real-world task completed under low-light conditions.

maps a normalized channel intensity $x \in [0, 1]$. It is monotonic and endpoint preserving, with zero logits corresponding exactly to the identity mapping. The final VLA-facing image is

$$Y_t = \mathrm{clip}\left(T_t(\mathrm{clip}(Z_t, 0, 1)), 0, 1\right). \tag{9}$$

3.4 Action-Driven Optimization

RawVLA is optimized through the native action objective of the downstream VLA policy. This objective retains the backbone's original training formulation across direct action regression (Kim et al., 2025a) and diffusion-based generation (Chi et al., 2025) and is denoted generically as

$$\mathcal{L}_{\text{action}} = \mathcal{L}_{\text{VLA}}(Y_t, \mathbf{a}_t), \tag{10}$$

where $\mathbf{a}_t$ is the ground-truth action signal. The VLA parameters remain fixed, while gradients through Y_t update RawVLA. The learned ISP is therefore trained with an action-driven objective without requiring pixel-wise supervision.

Although the action objective effectively aligns ISP rendering with downstream policy behavior, extreme illumination can hinder RawVLA from learning aggressive exposure adjustments and reliable chromatic correction early in training. We therefore introduce two pathway-specific regularizers

$$\mathcal{L}_{\text{bright}} = \frac{1}{B}\sum_{b=1}^{B}\left|\mu(Y_t^{(b)}) - 0.48\right|, \qquad \mathcal{L}_{\text{chroma}} = \frac{1}{B}\sum_{b=1}^{B}\mathrm{dist}_{\text{SL1}}\left(\mathbf{d}(Y_t^{(b)}), \Omega_C\right). \tag{11}$$

In this expression, $\mu(\cdot)$ is the mean image intensity, $\mathbf{d}(\cdot)$ denotes the output log-chroma descriptor, B is the batch size, and Ω_C is the log-chroma range estimated from RGB images randomly sampled in the simulator. The smooth distance term $\mathrm{dist}_{\text{SL1}}(\cdot)$ penalizes violations outside Ω_C. The complete objective therefore becomes

$$\mathcal{L} = \lambda_a \mathcal{L}_{\text{action}} + \lambda_b \mathcal{L}_{\text{bright}} + \lambda_{\text{chroma}} \mathcal{L}_{\text{chroma}}, \tag{12}$$

where λ_a is model-specific due to diverse action losses, while $\lambda_b = 0.001$ and $\lambda_{\text{chroma}} = 0.001$. The complete training settings are given in Appendix D.

4 RawVLA-Bench

RawVLA-Bench exposes image acquisition as an explicit evaluation variable by changing simulator illumination before RAW formation. We sample five illumination regimes spanning *ExtremeLow*,

Table 1: Suite-resolved success rates (%) of ISP modules on LIBERO. SOG averages Spatial, Object, and Goal; Long denotes LIBERO-10; Avg averages all four suites.

Illumin.	Backbone	Default			DarkISP			RAM			RAWAdapter			RAWild			**Ours**		
		SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg
Extreme Lowlight	PI-0	7.87	3.20	6.70	1.93	0.00	1.45	1.80	0.00	1.35	2.80	0.00	2.10	2.87	0.00	2.15	46.00	24.80	40.70
	PI-0.5	2.60	6.60	3.60	5.80	9.00	6.60	1.73	0.00	1.30	0.07	0.00	0.05	3.93	8.00	4.95	41.63	30.00	38.72
	Qwen3-OFT	0.00	0.00	0.00	4.13	2.40	3.70	0.00	0.00	0.00	0.00	0.00	0.00	0.87	2.60	1.30	15.20	10.60	14.05
Lowlight	PI-0	68.67	44.60	62.65	2.67	0.00	2.00	1.80	0.00	1.35	2.60	0.00	1.95	6.67	0.40	5.10	90.47	66.00	84.35
	PI-0.5	62.47	50.80	59.55	53.60	43.00	50.95	1.13	0.00	0.85	0.20	0.00	0.15	72.93	60.40	69.80	91.67	81.80	89.20
	Qwen3-OFT	36.40	19.00	32.05	47.73	29.40	43.15	0.67	0.00	0.50	0.00	0.00	0.00	74.73	45.20	67.35	87.13	64.00	81.35
Normal	PI-0	96.60	79.60	92.35	2.53	0.00	1.90	2.33	0.20	1.80	2.80	0.00	2.10	43.13	11.60	35.25	95.00	76.20	90.30
	PI-0.5	98.40	91.80	96.75	97.67	88.60	95.40	81.87	29.80	68.85	86.07	38.20	74.10	98.20	89.20	95.95	98.40	89.20	96.10
	Qwen3-OFT	98.60	92.80	97.15	98.80	91.00	96.85	69.00	23.00	57.50	97.13	84.80	94.05	98.87	91.40	97.00	98.53	92.20	96.95
OverExp	PI-0	66.00	35.40	58.35	1.60	0.00	1.20	2.53	0.00	1.90	2.93	0.00	2.20	37.13	6.20	29.40	81.93	55.40	75.30
	PI-0.5	85.20	48.20	75.95	57.80	39.60	53.25	1.67	0.20	1.30	75.73	39.00	66.55	86.67	64.40	81.10	89.00	59.00	81.50
	Qwen3-OFT	19.27	13.60	17.85	29.80	12.80	25.55	4.20	0.20	3.20	14.80	4.40	12.20	38.87	17.80	33.60	91.67	55.00	82.50
Extreme OverExp	PI-0	20.67	19.80	20.45	1.60	0.00	1.20	2.00	0.00	1.50	2.73	0.00	2.05	32.87	6.20	26.20	68.27	28.20	58.25
	PI-0.5	20.93	24.20	21.75	52.20	19.40	44.00	1.40	0.00	1.05	58.47	24.80	50.05	84.93	45.20	75.00	78.40	31.40	66.65
	Qwen3-OFT	0.00	0.00	0.00	17.13	0.00	12.85	0.00	0.00	0.00	2.33	0.00	1.75	13.07	0.00	9.80	47.80	2.00	36.35
Overall	PI-0	51.96	36.52	48.10	2.07	0.00	1.55	2.09	0.04	1.58	2.77	0.00	2.08	24.53	4.88	19.62	76.33	50.12	69.78
	PI-0.5	53.92	44.32	51.52	53.41	39.92	50.04	17.56	6.00	14.67	44.11	20.40	38.18	69.33	53.44	65.36	79.82	58.28	74.44
	Qwen3-OFT	30.85	25.08	29.41	39.52	27.12	36.42	14.77	4.64	12.24	22.85	17.84	21.60	45.28	31.40	41.81	68.07	44.76	62.24
	Avg	45.58	35.31	43.01	31.67	22.35	29.34	11.47	3.56	9.50	23.24	12.75	20.62	46.38	29.91	42.26	74.74	51.05	68.82

Low, *Normal*, *Over*, and *ExtremeOver*, with environment-specific ranges calibrated for comparable visual difficulty. To synthesize expert trajectories in the RAW domain while preserving successful expert actions and environment states, we replay demonstrations in both benchmarks. For LIBERO, we replay 2,000 official demonstrations from its four suites and retain 1,771 successful trajectories, pairing normal- and target-light observations under identical states and actions. Because MuJoCo exposes only tone-mapped RGB8, we apply fixed unprocessing (Brooks et al., 2019) to produce three-channel pseudo-RAW10 for the 256×256 agent and wrist views. For RoboTwin 2.0, we replay 650 demonstrations from 13 dual-arm tasks and obtain 632 successful trajectory pairs, synchronizing clean and target-light environments at every frame. We form RAW observations directly from SAPIEN pre-tonemap HDR buffers with heteroscedastic sensor noise, using head and dual-wrist views at 240×320. Detailed construction and evaluation settings are provided in Appendix B.

5 Experiments

Implementation Details. We train ISP modules per frozen policy for 2,000 updates on a single NVIDIA A100 80GB GPU. To alleviate temporal inconsistency in ISP baselines designed for single images, we jointly optimize them over 8-frame trajectory windows and average their corresponding losses. RawVLA instead unrolls $N_{\text{rec}} = 6$ recurrent endpoints with stride 8 for Qwen3-series and 5 for PI0 and PI0.5, and applies full BPTT. More details are provided in Appendix D.

Baseline Methods. We compare RawVLA against a fixed Default ISP and four learning-based neural ISPs. The Default ISP applies the paired deterministic RAW-to-RGB reprocessing described in Appendix A, without task adaptation. The learned baselines include DarkISP (Guo et al., 2025), RAM (Gamrian et al., 2025), RAW-Adapter (Cui et al., 2026), and RAWild (Liu et al., 2026b). We use task success rate as the evaluation metric and conduct 50 rollouts per task with distinct random seeds and initial positions. Best results are shaded as first, second, and third, respectively.

VLA Backbones. On the LIBERO split of RawVLA-Bench, the main ISP comparison uses PI-0 (Black et al., 2024), PI-0.5 (Physical Intelligence et al., 2025), and Qwen3-OFT across the four LIBERO suites and five illumination regimes; Qwen3-PI is additionally included in the ISP perturbation analysis. On RoboTwin 2.0, we use PI-0 and PI-0.5 under the same five illumination regimes. The Qwen3-series VLA models are task-finetuned with StarVLA (StarVLA Community, 2026).

Evaluation on LIBERO. Across the suite-resolved results in Table 1, RawVLA achieves the highest overall performance for every backbone, raising the cross-backbone average from 43.01% for the strongest baseline to 68.82%. This improvement is driven primarily by adverse illumination, with particularly large gains for Qwen3-OFT under overexposure, while the strong performance under

Table 2: Success rates (%) on RoboTwin 2.0 across five illumination levels. *Default* denotes the default SAPIEN tone-mapped RGB baseline.

Illumin.	Backbone	Default	DarkISP	RAM	R.Adp.	RAWild	**Ours**
Extreme Lowlight	PI-0	23.62	4.72	12.60	1.57	8.66	32.28
	PI-0.5	17.32	1.57	3.15	0.00	5.51	25.98
Lowlight	PI-0	40.94	0.79	10.24	3.15	6.30	59.06
	PI-0.5	51.18	0.00	2.36	0.00	4.72	63.78
Normal	PI-0	55.91	34.65	17.32	56.69	28.35	57.48
	PI-0.5	70.87	35.43	20.47	41.73	40.94	67.72
OverExp	PI-0	48.41	34.92	19.05	51.59	55.56	58.73
	PI-0.5	56.35	26.98	4.76	38.89	45.24	65.08
Extreme OverExp	PI-0	36.00	36.80	11.20	15.20	47.20	56.80
	PI-0.5	38.40	42.40	1.60	20.00	54.40	63.20
Overall	PI-0	40.98	22.31	14.08	25.63	29.11	52.85
	PI-0.5	46.84	21.20	6.49	20.09	30.06	57.12
	Avg	43.91	21.76	10.28	22.86	29.59	54.98

Table 3: Real-world success rates (%) on four tasks using PI-0.5. *Default* denotes the constant default ISP baseline.

Illumin.	Task	Default	DarkISP	**Ours**
Normal	*pick & place*	96.00	66.00	96.00
	drug drawer	92.00	52.00	94.00
	pick flower	72.00	44.00	66.00
	stack & cover	46.00	38.00	44.00
	Avg	76.50	50.00	75.00
Lowlight	*pick & place*	0.00	22.00	92.00
	drug drawer	0.00	12.00	88.00
	pick flower	0.00	10.00	56.00
	stack & cover	0.00	0.00	30.00
	Avg	0.00	11.00	66.50

Table 4: Inference frame-per-second.

DarkISP	RAM	R.Adp.	RAWild	**Ours**
176	972	414	320	168

normal lighting is largely preserved. Moreover, the gains hold for both the Spatial, Object, and Goal suites and the long-horizon LIBERO-10 suite, demonstrating that RawVLA benefits different task types and horizons rather than a particular benchmark subset. The first row of Figure 4 qualitatively compares the agent-view outputs of the Default ISP and RawVLA on LIBERO.

Evaluation on RoboTwin 2.0. As shown in Table 2, RawVLA achieves a cross-backbone average of 54.98% on RoboTwin 2.0, improving the success rate by 85.81% relative to the strongest neural ISP baseline. RawVLA also ranks first in every adverse illumination regime for both policies. Its largest gains occur under low light, while the improvements under overexposure confirm its effectiveness for both signal-starved and saturated observations. The second row of Figure 4 compares the corresponding agent-view outputs of the Default ISP and RawVLA on RoboTwin 2.0.

Evaluation on Real-World Dual-Arm. We evaluate RawVLA on four real-world dual-arm tasks under normal and low-light conditions with a fine-tuned PI-0.5 policy backbone. As shown in Table 3, RawVLA achieves 75.00% average success under normal illumination, comparable to the Default ISP at 76.50% and above DarkISP at 50.00%. Under low light, RawVLA maintains 66.50% success, while the Default ISP and DarkISP reach only 0.00% and 11.00%, respectively. As shown in Table 4, all evaluated ISP modules support real-time operation and run faster than the cameras' native frame rate, with RawVLA reaching 168 FPS. The final row of Figure 4 shows a representative real-world task completion sequence. Further implementation details and task execution visualizations are provided in Appendix H.

6 Ablation

Table 5 summarizes controlled ablations on two representative VLA models, Qwen3-OFT with a regression head on LIBERO and PI-0.5 with a diffusion head on RoboTwin 2.0.

Photometric Descriptors. The complete descriptor set achieves average success rates of 62.24% on LIBERO and 57.12% on RoboTwin 2.0. Removing the spatial feature g_t reduces these averages to 36.35% and 41.93%, respectively, confirming the importance of spatial luminance structure. The global luminance descriptor h_t^L is especially important on RoboTwin 2.0, whereas removing the chroma descriptor h_t^C causes the largest drop on LIBERO. These complementary trends support separate descriptors for intensity distribution, spatial structure, and color.

Recurrent Estimator. Removing either the hidden state or previous-parameter conditioning sharply degrades LIBERO performance and consistently lowers success on RoboTwin 2.0. A shared GRU also underperforms the decoupled design, particularly on RoboTwin 2.0, while either the luminance-only or chroma-only state remains inferior to their combination. Together, these results

Table 5: Ablations of RawVLA on LIBERO with Qwen3-OFT and RoboTwin 2.0 with PI-0.5. The two blocks isolate photometric descriptors and recurrent module structures.

	LIBERO / Qwen3-OFT						RoboTwin 2.0 / PI-0.5					
Variant	Ext.Low	Low	Normal	Over	Ext.Over	Avg.	Ext.Low	Low	Normal	Over	Ext.Over	Avg.
Photometric descriptors												
w/o spatial feature g_t	0.00	20.35	97.10	51.15	13.15	36.35	14.96	12.60	62.99	60.32	59.20	41.93
w/o luminance descriptor h_t^L	27.75	95.05	96.80	50.80	25.95	59.27	21.26	47.24	50.39	56.35	52.80	45.57
w/o chroma descriptor h_t^C	0.00	0.00	96.65	48.75	16.60	32.40	23.62	57.48	48.03	60.32	53.60	48.58
Recurrent estimator												
w/o hidden state $S_{t-1}^{L/C}$	0.00	0.35	97.20	51.65	22.85	34.41	22.05	57.48	46.46	57.94	53.60	47.47
w/o Θ_{t-1} conditioning	0.00	0.00	96.65	49.95	25.25	34.37	24.41	49.61	51.18	57.94	55.20	47.63
Single shared $\mathrm{GRU}_{L,C}$	28.10	91.20	96.75	55.55	35.30	61.38	22.83	55.91	40.94	63.49	56.80	47.94
Lumin. only ($S_t = S_t^L$)	28.10	94.60	96.35	51.40	29.75	60.04	21.26	61.42	66.14	67.46	62.40	55.70
Chrom. only ($S_t = S_t^C$)	28.10	93.45	96.90	50.35	25.15	58.79	22.83	51.97	58.27	61.11	64.00	51.58
Full RawVLA	14.05	81.35	96.95	82.50	36.35	62.24	25.98	63.78	67.72	65.08	63.20	57.12

show that temporal context, parameter feedback, and factorized luminance–chroma estimation each contribute to robust adaptation. Additional ablations of the training losses, individual ISP operators, and burst-denoising settings are provided in Appendix G.

7 Related Works

VLA & WAM Visual Robustness. Generalist policies span token-based and continuous-action VLAs (Kim et al., 2025b; Black et al., 2024; Physical Intelligence et al., 2025) and video-predictive WAMs (Kim et al., 2026b; Ye et al., 2026), increasingly targeting cross-embodiment control and physical reasoning (Bjorck et al., 2025; Gemini Robotics Team et al., 2025). Yet both remain vulnerable to visual changes. Robustness benchmarks use perturbations and scalable simulation to cover appearance, lighting, viewpoint, scene composition, and physical shifts (Pumacay et al., 2024; Li et al., 2025b; Wang et al., 2025; 2024b; Chen et al., 2026b; Zhang et al., 2026; Fei et al., 2026; Morgan et al., 2026; Chen et al., 2026c). Existing methods edit task-irrelevant regions, enforce action consistency, restore observations, or learn invariance to visual and illumination changes (Hancock et al., 2025; Guo et al., 2026; Zhang et al., 2025; Orjuela et al., 2026; Xie et al., 2026; Luo et al., 2026; Watanabe et al., 2026). Additional robustness can come from event, thermal, or tactile sensing, at the cost of extra sensors and cross-modal data (Zhai et al., 2026; Liu et al., 2026a; Yu et al., 2026; Huang et al., 2025). We instead operate on the existing camera signal and study robustness at the RAW and ISP processing levels.

Task-Oriented and Neural ISP. Conventional ISPs target human visual quality and may not preserve signals most useful to downstream perception. Prior work therefore co-optimizes RAW processing and recognition, builds compact machine-oriented pipelines, and searches task-specific ISP structures (Diamond et al., 2021; Wu et al., 2019; Yu et al., 2021; Shi et al., 2022). Later methods introduce content-adaptive operators and connect learnable ISP stages to downstream models for sensor- or task-conditioned processing (Sun et al., 2024; Wang et al., 2024a; Gamrian et al., 2025; Won et al., 2026; Morawski et al., 2022; Cui et al., 2026; Huang et al., 2026). Complementary work improves adverse-condition and sensor-general RAW perception through synthesis, augmentation, enhancement, adaptation, and efficient models (Punnappurath et al., 2022; Yoshimura et al., 2023; Guo et al., 2025; Hashmi et al., 2025; Chen et al., 2026a; Liu et al., 2026b; Li et al., 2026). Our work extends this direction to sequential VLA behavior and manipulation success.

8 Conclusion

We introduced RawVLA, a lightweight neural ISP for frozen VLA policies, and RawVLA-Bench, a RAW-domain manipulation benchmark. Across simulation and real-world tasks, RawVLA maintains nominal-condition performance while substantially improving robustness to adverse illumination, demonstrating that adaptive camera processing supports robust robot control.

AI USE STATEMENT

Generative AI tools assisted with writing, editing, translation, literature search, and reference verification. They were not used to formulate research ideas, design or implement the methods and experiments, generate data or figures, or analyze and interpret results. The authors reviewed all AI-assisted content and take full responsibility for the manuscript.

REFERENCES

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. Pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11028–11037. IEEE, 2019.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3291–3300. Ieee, 2018.

Kai Chen, Jin Xiao, Leheng Zhang, Kexuan Shi, and Shuhang Gu. Task-aware image signal processor for advanced visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33672–33681, 2026a.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *Proceedings of the International Conference on Machine Learning*, 2026b.

Yuhao Chen, Zhihao Zhan, Xiaoxin Lin, Zijian Song, Hao Liu, Qinhan Lyu, Yubo Zu, Xiao Chen, Zhiyuan Liu, Tao Pu, et al. Radar: Benchmarking vision-language-action generalization via real-world dynamics, spatial-physical intelligence, and autonomous evaluation. *arXiv preprint arXiv:2602.10980*, 2026c.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.

Ziteng Cui, Jianfei Yang, and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images and a benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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.

Rida T Farouki. The bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012.

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*, pp. 38574–38583, 2026.

Shani Gamrian, Hila Barel, Feiran Li, Masakazu Yoshimura, and Daisuke Iso. Beyond rgb: Adaptive parallel processing for raw object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5547–5557. IEEE, 2025.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, et al. On robustness of vision-language-action model against multi-modal perturbations. In *Proceedings of the International Conference on Learning Representations*, volume 2026, pp. 70248–70272, 2026.

Jiasheng Guo, Xin Gao, Yuxiang Yan, Guanghao Li, and Jian Pu. Dark-isp: Enhancing raw image processing for low-light object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9583–9593. IEEE, 2025.

Asher J Hancock, Allen Z Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. In *IEEE International Conference on Robotics and Automation*, pp. 9499–9506. IEEE, 2025.

Khurram Azeem Hashmi, Karthik Palyakere Suresh, Didier Stricker, and Muhammad Zeshan Afzal. Torchadapt: Towards light-agnostic real-time visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5645–5656. IEEE, 2025.

Zhi Hou, Tianyi Zhang, Yuwen Xiong, Haonan Duan, Hengjun Pu, Ronglei Tong, Chengyang Zhao, Xizhou Zhu, Yu Qiao, Jifeng Dai, et al. Dita: Scaling diffusion transformer for generalist vision-language-action policy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7686–7697. IEEE, 2025.

Jialei Huang, Shuo Wang, Fanqi Lin, Yihang Hu, Chuan Wen, and Yang Gao. Tactile-vla: unlocking vision-language-action model's physical knowledge for tactile generalization. *arXiv preprint arXiv:2507.09160*, 2025.

Wenjun Huang, Ziteng Cui, Yinqiang Zheng, Yirui He, Tatsuya Harada, and Mohsen Imani. Dr. raw: Towards general high-level vision from raw with efficient task conditioning. *Advances in Neural Information Processing Systems*, 38:22815–22842, 2026.

International Telecommunication Union. Parameter values for the hdtv standards for production and international programme exchange. Technical Report Recommendation BT.709-6, ITU-R, 2015.

Hyeonjun Kim, Jegwang Ryu, Sangbeom Ha, Junhyeok Lee, Jun-Hyuk Kim, Hyemin Ahn, and Jaeho Lee. Learned image compression for vision-language-action models. *arXiv preprint arXiv:2606.16253*, 2026a.

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*, 2025a.

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 *Proceedings of the Conference on Robot Learning*, volume 270, pp. 2679–2713. PMLR, 2025b.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026b.

Chunyi Li, Rui Qing, Jianbo Zhang, Yuan Tian, Xiangyang Zhu, Zicheng Zhang, Xiaohong Liu, Weisi Lin, and Guangtao Zhai. Embodied image compression. *arXiv preprint arXiv:2512.11612*, 2025a.

Hanxi Li, Yao Cheng, Bo Zhang, and Li Zeng. Uniisp: A unified isp framework for both human and machine vision. *arXiv preprint arXiv:2605.07359*, 2026.

Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of the Conference on Robot Learning*, 2025b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

Jiaxin Liu, Xun Xu, Zhenhao Zhang, Hanqing Wang, Ruiqi Chen, Shi Chang, Weiyu Guo, and Laurent Kneip. Event-vla: Action-conditioned event fusion for robust vision-language-action model. *arXiv preprint arXiv:2606.29384*, 2026a.

Shuhong Liu, Gengjia Chang, Jun Liu, Xuangeng Chu, Yinqiang Zheng, Tatsuya Harada, and Ziteng Cui. Rawild: Sensor-agnostic raw object detection via physics-guided curve and grid modeling. *arXiv preprint arXiv:2605.05941*, 2026b.

Jingzhou Luo, Yifan Wen, Yongjie Bai, Xinshuai Song, Yang Liu, and Liang Lin. Rovla: Multi-consistency constraints for robust vision-language-action models. *arXiv preprint arXiv:2605.19678*, 2026.

Igor Morawski, Yu-An Chen, Yu-Sheng Lin, Shusil Dangi, Kai He, and Winston H Hsu. Genisp: Neural isp for low-light machine cognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 629–638. IEEE, 2022.

Jeremy Morgan, Prajwal Vijay, Hyeonho Oh, Jincen Song, Ashvin Arora, Alina Du, Gaurav Sukhatme, Jesse Thomason, and Ishika Singh. Colosseum v2: Benchmarking generalization for vision language action models. *arXiv preprint arXiv:2605.27759*, 2026.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. 2024.

Daniel Yezid Guarnizo Orjuela, Leonardo Scappatura, Veronica Di Gennaro, Riccardo Andrea Izzo, Gianluca Bardaro, and Matteo Matteucci. Improving robustness of vision-language-action models by restoring corrupted visual inputs. *arXiv preprint arXiv:2602.01158*, 2026.

Björn Ottosson. A perceptual color space for image processing, 2020. URL https://bottosson.github.io/posts/oklab/.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. Pi0.5: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. In *Proceedings of Robotics: Science and Systems*, 2024.

Abhijith Punnappurath, Abdullah Abuolaim, Abdelrahman Abdelhamed, Alex Levinshtein, and Michael S Brown. Day-to-night image synthesis for training nighttime neural isps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10759–10768. IEEE, 2022.

Yongjie Shi, Songjiang Li, Xu Jia, and Jianzhuang Liu. Refactoring isp for high-level vision tasks. In *IEEE International Conference on Robotics and Automation*, pp. 2366–2372. IEEE, 2022.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

Xinyu Sun, Zhikun Zhao, Lili Wei, Congyan Lang, Mingxuan Cai, Longfei Han, Juan Wang, Bing Li, and Yuxuan Guo. Rl-seqisp: Reinforcement learning-based sequential optimization for image signal processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5025–5033, 2024.

Yujin Wang, Tianyi Xu, Fan Zhang, Tianfan Xue, and Jinwei Gu. Adaptiveisp: Learning an adaptive image signal processor for object detection. *Advances in Neural Information Processing Systems*, 37:112598–112623, 2024a.

Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. Ladev: A language-driven testing and evaluation platform for vision-language-action models in robotic manipulation. *arXiv preprint arXiv:2410.05191*, 2024b.

Zhijie Wang, Zhehua Zhou, Jiayang Song, Yuheng Huang, Zhan Shu, and Lei Ma. Vlatest: Testing and evaluating vision-language-action models for robotic manipulation. *Proceedings of the ACM on Software Engineering*, 2(FSE):1615–1638, 2025.

Marino Watanabe, Takami Sato, and Kentaro Yoshioka. Lights, camera, malfunction: When illumination robustness leaves vla models blind to color. *arXiv preprint arXiv:2607.14698*, 2026.

Jiyun Won, Heemin Yang, Woohyeok Kim, Jungseul Ok, and Sunghyun Cho. Pos-isp: Pipeline optimization at the sequence level for task-aware isp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, 2026.

Chyuan-Tyng Wu, Leo F Isikdogan, Sushma Rao, Bhavin Nayak, Timo Gerasimow, Aleksandar Sutic, Liron Ain-Kedem, and Gilad Michael. Visionisp: Repurposing the image signal processor for computer vision applications. In *IEEE International Conference on Image Processing*, pp. 4624–4628. IEEE, 2019.

Yuhan Xie, Yuping Yan, Yunqi Zhao, Handing Wang, and Yaochu Jin. Strong-vla: Decoupled robustness learning for vision-language-action models under multimodal perturbations. *arXiv preprint arXiv:2604.10055*, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, and Takeshi Ohashi. Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14007–14017. IEEE, 2023.

Dian Yu, Qingchuan Zhou, Bingkun Huang, Majid Khadiv, and Zewen Yang. Safe-night vla: Seeing the unseen via thermal-perceptive vision-language-action models for safety-critical manipulation. *arXiv preprint arXiv:2603.05754*, 2026.

Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, and Jinwei Gu. Reconfigisp: Reconfigurable camera image processing pipeline. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4228–4237. IEEE, 2021.

Jiajun Zhai, Hao Shi, Shangwei Guo, Kailun Yang, and Kaiwei Wang. E-vla: Event-augmented vision-language-action model for dark and blurred scenes. *arXiv preprint arXiv:2604.04834*, 2026.

Hongyin Zhang, Shuo Zhang, Junxi Jin, Qixin Zeng, Runze Li, and Donglin Wang. Robustvla: Robustness-aware reinforcement post-training for vision-language-action models. *arXiv preprint arXiv:2511.01331*, 2025.

Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, Rui Heng Yang, Yintao Ma, Amir Rasouli, Sajjad Pakdamansavoji, Yangzheng Wu, Lingfeng Zhang, Tongtong Cao, et al. Do world action models generalize better than vlas? a robustness study. *arXiv preprint arXiv:2603.22078*, 2026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the Conference on Robot Learning*, volume 229, pp. 2165–2183, 2023.

A RGB UNPROCESSING AND DEFAULT ISP

This section defines the RAW representation used throughout the paper and specifies the paired RGB-to-RAW unprocessing and default RAW-to-RGB ISP.

RAW Input Representation. We distinguish the sensor measurement from the representation processed by RawVLA. A physical camera first records a single-channel, mosaiced Bayer RAW measurement. We subtract the sensor black level, normalize by the usable white level, and demosaic the Bayer array into a three-channel linear signal. Throughout the paper, we refer to this black-level-corrected, demosaiced signal as *linear RAW*. It is the direct input representation of RawVLA. The simulated observations are three-channel *pseudo-RAW* signals created directly in the same linear RAW representation and therefore do not contain a Bayer mosaic.

We construct the simulated pseudo-RAW observations using a fixed unprocessing pipeline adapted from Brooks et al. (2019). Let $Y^{\text{render}} \in [0,1]^{3\times H\times W}$ denote the display-referred RGB observation resized to the VLA input resolution. All photometric operations are applied channel-wise. The default smoothstep tone curve is

$$T_0(z) = 3z^2 - 2z^3, \tag{13}$$

whose inverse T_0^{-1} is uniquely defined on $[0,1]$, and the standard sRGB transfer function

$$\mathcal{G}(z) = \begin{cases} 12.92z, & z \le 0.0031308, \\ 1.055z^{1/2.4} - 0.055, & z > 0.0031308. \end{cases} \tag{14}$$

Its inverse is

$$\mathcal{G}^{-1}(z) = \begin{cases} z/12.92, & z \le 0.04045, \\ \left((z+0.055)/1.055\right)^{2.4}, & z > 0.04045. \end{cases} \tag{15}$$

We first invert the default tone curve and sRGB transfer function, then remove the fixed channel-gain vector $\mathbf{m}_{\text{rgb}} = (1.8, 1.0, 1.7)$. The unprocessing operator $\mathcal{U}$ produces the simulated camera RAW observation

$$R = \mathcal{U}(Y^{\text{render}}) = \text{clip}\left(\text{diag}(\mathbf{m}_{\text{rgb}})^{-1}\mathcal{G}^{-1}\left(T_0^{-1}(Y^{\text{render}})\right), 0, 1\right). \tag{16}$$

The resulting R retains the precision available from the corresponding renderer. The complete unprocessing sequence applies inverse tone mapping, inverse sRGB transfer, inverse channel gains, and clipping. Controlled bit-depth changes use $\mathcal{Q}_n$ from Equation 30, with their environment-specific placement given in Appendix K.

Let $\mathcal{P}(\cdot, \xi)$ denote the parameterized ISP and ξ_0 its default parameter setting. The paired default operator $\mathcal{P}_0(\cdot) \equiv \mathcal{P}(\cdot, \xi_0)$ reverses the fixed photometric transformations

$$Y_0 = \mathcal{P}_0(R) = T_0(\mathcal{G}(\text{clip}(\text{diag}(\mathbf{m}_{\text{rgb}})R, 0, 1))) = Y^{\text{render}}. \tag{17}$$

This produces a three-channel pseudo-RAW signal in the linear RAW representation rather than a camera-specific Bayer measurement. The baseline unprocessing does not simulate mosaicing, sensor noise, or a camera-specific color-correction matrix. The noise axis in Appendix K introduces its sensor model separately.

B RAWVLA-BENCH CONSTRUCTION

RawVLA-Bench changes illumination inside each simulator before RAW formation rather than multiplying the final RGB image. It contains paired training caches and frozen evaluation manifests for LIBERO (Liu et al., 2023) and RoboTwin 2.0 (Chen et al., 2026b). Each training trajectory stores one target-light RAW sequence together with a normal-light RGB reference under the same actions and physical states. To avoid increasing the training cache fivefold, each successful trajectory is assigned one illumination regime. During evaluation, the same task initialization is instead expanded across all five regimes for paired comparison. In total, the training cache contains 2,403 paired trajectories and 408,075 frames, while the evaluation manifests specify 13,250 rollouts. Section B.1 describes the LIBERO (Liu et al., 2023) construction, and Section B.2 details the RoboTwin 2.0 (Chen et al., 2026b) construction.

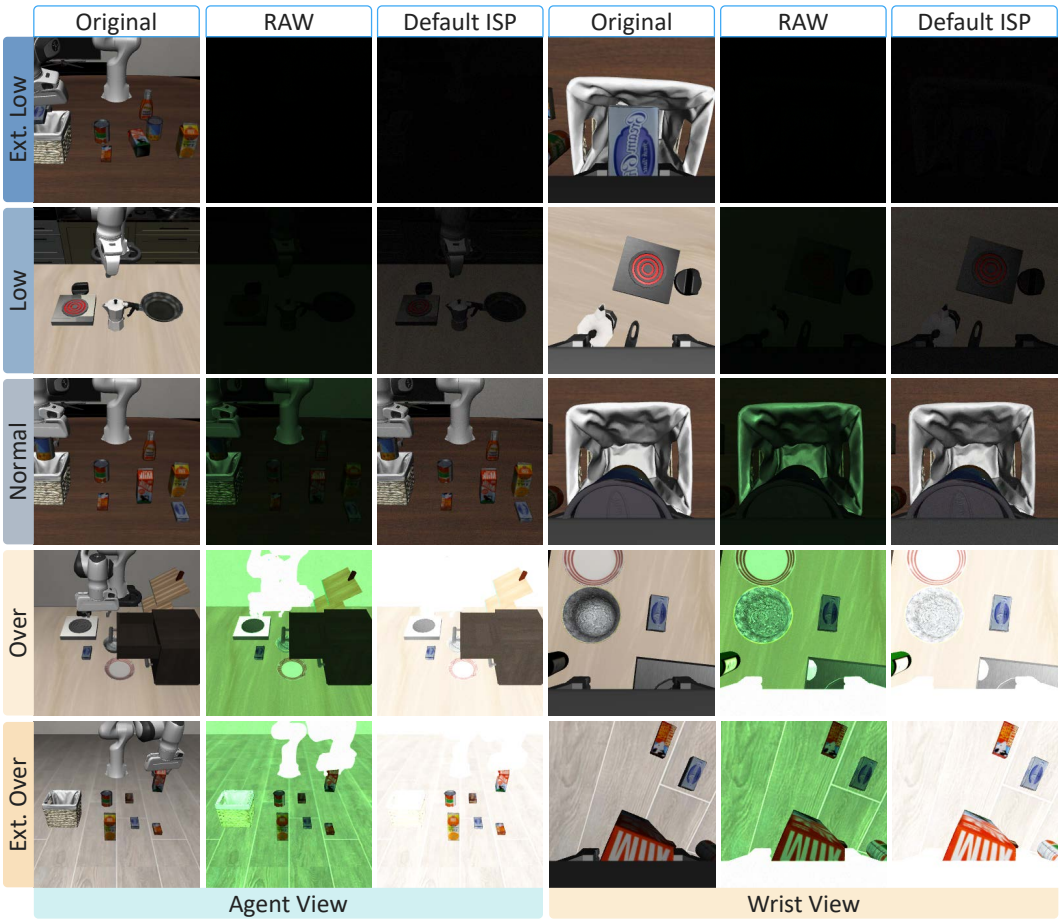


Figure 5: RawVLA-Bench observations on LIBERO (Liu et al., 2023) across five ambient-light intensity levels ranging from *extreme low light* and *low light* to *normal*, *overexposure*, and *extreme overexposure*. Each level shows the original RGB reference, synthesized RAW observation, and its default-ISP rendering. Agent and wrist views use the same unprocessing procedure.

B.1 LIBERO Construction

We use the four standard suites LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-10 from LIBERO (Liu et al., 2023), comprising 40 tasks. From at most 50 official demonstrations per task, we obtain 2,000 candidate trajectories. Each demonstration is recreated using its recorded MuJoCo XML, initial simulator state, and seven-dimensional action sequence. We replay the actions in a clean environment, remove no-op actions, and retain a trajectory only when the task-specific success predicate is reached. The same state is copied before every rendering step to a second environment whose illumination is modified, ensuring that the RGB reference and RAW observation differ only in the image-formation branch. This procedure retains 1,771 successful trajectories and 261,131 frames. These include 454 from LIBERO-Spatial, 462 from LIBERO-Object, 456 from LIBERO-Goal, and 399 from LIBERO-10 (Liu et al., 2023). Both the agent and wrist cameras render at 256×256.

MuJoCo provides only tone-mapped RGB8 for these environments. We therefore construct a three-channel pseudo-RAW signal by inverting the smoothstep tone curve and sRGB transfer function, removing fixed RGB gains, clipping, and quantizing to 10 bits. Overexposed regimes additionally apply controlled full-well saturation before clipping. We then inject heteroscedastic Gaussian noise with variance $4 \times 10^{-4}x + 10^{-5}$ using deterministic seeds that vary across trajectories, frames, and views. Figure 7 visualizes how the synthesized RAW noise becomes increasingly visible after ISP

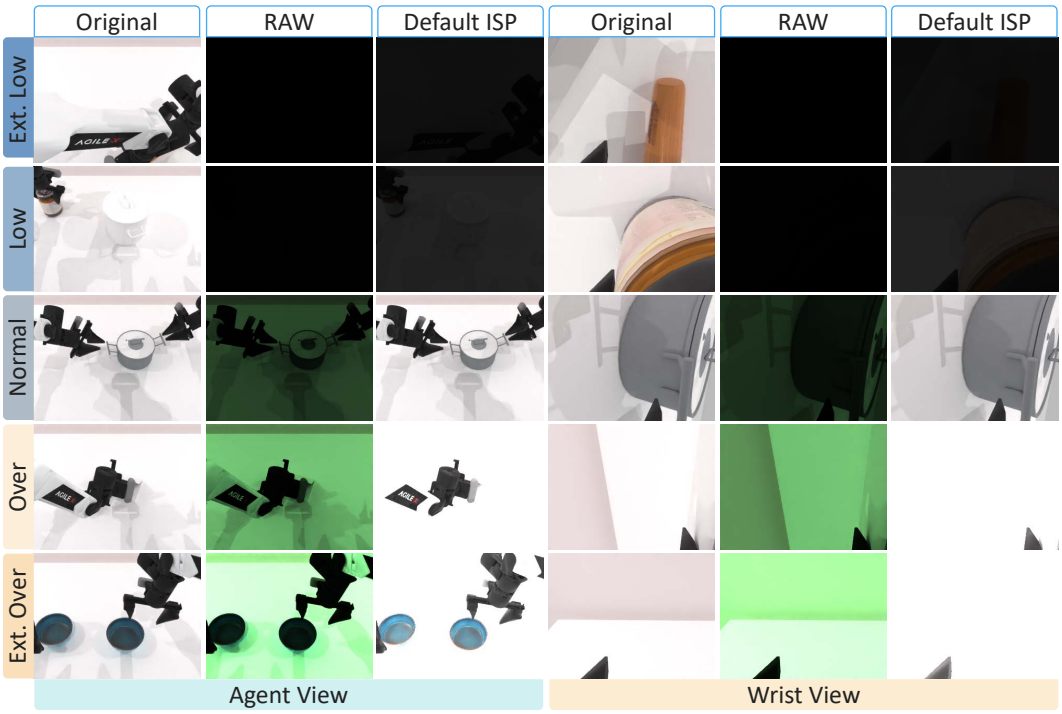


Figure 6: RawVLA-Bench observations on RoboTwin 2.0 (Chen et al., 2026b) across five ambient-light intensity levels ranging from *extreme low light* and *low light* to *normal*, *overexposure*, and *extreme overexposure*. Each level shows the original RGB reference, synthesized RAW observation, and its default-ISP rendering. Agent and wrist views use the same unprocessing procedure.

gain amplification as illumination decreases. The five regimes and EV intervals are ExtremeLow $[-5.5, -4]$, Low $[-4, -2.5]$, Normal $[-0.5, 0.5]$, Over $[1.5, 3]$, and ExtremeOver $[6, 8]$. Over and ExtremeOver also activate additional table-directed illumination and saturation pressure. These pseudo-RAW observations model the processing degrees of freedom available to the learned ISP, but are not Bayer measurements and do not contain additional renderer highlight information beyond the source RGB8.

The frozen LIBERO (Liu et al., 2023) evaluation manifest contains 50 initial states for each of the 40 tasks. Every initial state is evaluated in all five illumination regimes, producing $4 \times 10 \times 50 \times 5 = 10{,}000$ rollouts, or 2,000 rollouts per regime. Entries sharing an initial state use the same task, language instruction, and camera configuration, while only illumination and sensor-noise seeds differ.

B.2 RoboTwin 2.0 Construction

We use 13 tasks from RoboTwin 2.0 (Chen et al., 2026b) with the ALOHA-AgileX embodiment and the clean task configuration. The 650 candidates comprise 50 recorded episodes per task. We reconstruct each scene from its dataset seed and replay the stored left- and right-arm joint paths through the original controller and physics stack. An episode is retained only when both trajectory planning and the task-specific success check succeed. Of 650 candidates, 633 pass this clean replay audit. One trajectory then fails during paired regeneration because its saved right-arm path is exhausted, leaving 632 complete pairs and 146,944 frames.

To prevent contact or numerical differences from causing paired trajectories to diverge, only the clean environment executes actions. Before every camera capture, actor poses and velocities, articulation root states, and joint positions and velocities are copied to a target-light environment, which is used only for rendering. Each frame includes the head, left-wrist, and right-wrist views at 240×320.

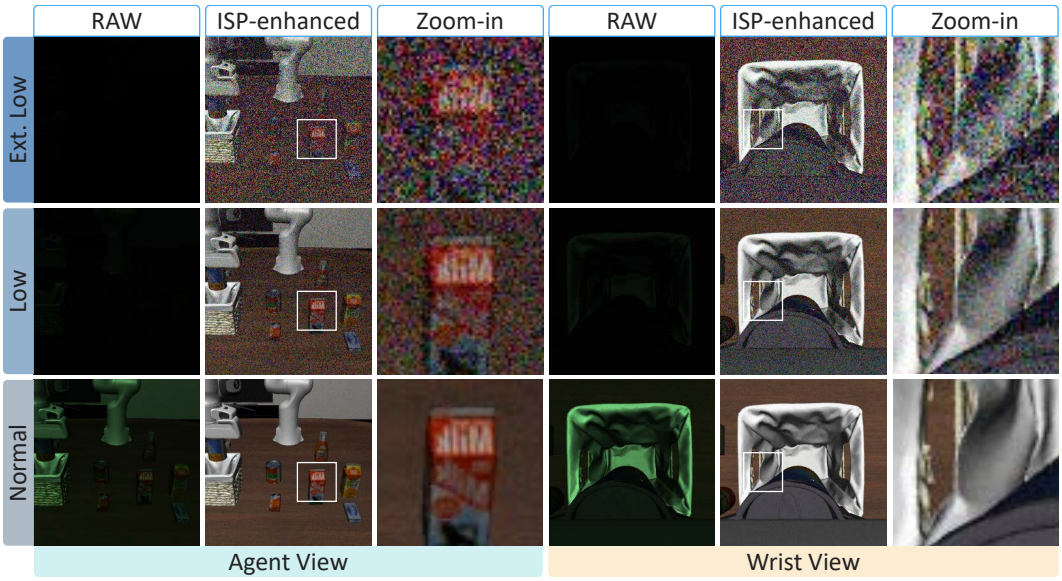


Figure 7: Synthetic sensor noise on LIBERO (Liu et al., 2023). The figure shows noisy RAW observations and their corresponding RGB renderings after ISP gain amplification and subsequent processing across progressively lower illumination levels. Noise becomes increasingly pronounced as the illumination decreases.

Unlike LIBERO (Liu et al., 2023), SAPIEN exposes a pre-tonemap HDR buffer. We normalize this linear radiance by a fixed white level, remove fixed RGB gains, and add heteroscedastic noise with variance $2.5 \times 10^{-5}x + 3.90625 \times 10^{-8}$. The resulting three-channel pseudo-RAW is stored as *float32* without 10-bit quantization.

RoboTwin 2.0 (Chen et al., 2026b) uses environment-only illumination scaling, applied from the captured baseline to ambient, directional, and point lights. Its five EV intervals are ExtremeLow $[-9, -8.5]$, Low $[-7, -6]$, Normal $[-0.5, 0.5]$, Over $[1, 1.5]$, and ExtremeOver $[2, 2.5]$. The environment-specific ranges are calibrated separately because MuJoCo and SAPIEN differ in indirect lighting, materials, HDR rendering, and tone mapping. The frozen evaluation manifest contains 50 dataset seeds for each task and expands each seed across all five regimes, yielding $13 \times 50 \times 5 = 3{,}250$ rollouts, or 650 per regime. Figures 5 and 6 show the environment-specific pipelines.

C RawVLA Architecture

This section expands the components of RawVLA. We describe burst denoising in Section C.2, the luminance and chroma descriptors in Sections C.3 and C.4, and the structured ISP parameterization in Section C.5.

C.1 Implementation Details

Our implementation receives a six-frame causal linear-RAW burst and contains 886,966 trainable parameters. The FFT burst merge has no trainable parameters. The spatial luminance encoder consists of six 3×3 convolutional blocks with channel transitions $3 \to 32$, $32 \to 32$, $32 \to 64$, $64 \to 64$, $64 \to 128$, and $128 \to 128$. Their strides are $1, 1, 2, 1, 2, 1$, respectively, and each block uses single-group normalization and a SiLU activation. The equal-channel luminance input is replicated across three channels before entering this encoder. Spatial mean, standard deviation, and learned attention-weighted mean pooling produce a 384-dimensional feature, which is encoded by a $384 \to 256 \to 128$ MLP. The attention score is predicted by a 1×1 convolution from 128 channels to one channel.

The 135-dimensional luminance descriptor is encoded by a $135 \to 128 \to 64$ MLP. We set the chroma histogram count to $B_C = 64$, giving a $5B_C + 10 = 330$ dimensional chroma descriptor that is encoded by a $330 \to 128 \to 64$ MLP. The two log-chrominance ratios are clipped to $[-4, 4]$ before histogramming. Previous luminance and chroma operating points contain nine and eight values and are encoded by $9 \to 64 \to 32$ and $8 \to 64 \to 32$ MLPs. The luminance fusion network maps $224 \to 256 \to 128$, while the chroma fusion network maps $96 \to 128 \to 128$. All MLP hidden layers use SiLU activations.

The recurrent estimator contains separate luminance and chroma GRU cells. Each cell has a 128-dimensional input and a 64-dimensional hidden state, yielding a 128-dimensional combined recurrent state. The exposure head maps $256 \to 128 \to 1$, the tone head maps $256 \to 128 \to 7$, and the chroma head maps $256 \to 128 \to 8$. The two single-layer GRU states are concatenated and carried to the next policy observation. RawVLA has no U-Net-style image decoder or spatial residual branch. The network predicts only explicit ISP controls, and the RGB output is generated analytically by burst fusion, exposure, white balance, color correction, and a shared monotonic Bernstein tone curve. The tone head predicts seven relative logits and appends one fixed-zero reference logit, producing eight monotonic intervals. Local color and local tone residuals are disabled.

We use $E_{\max} = 6$ EV, which bounds the achromatic exposure gain to $[2^{-6}, 2^6]$. White balance is represented by two independent coordinates that produce zero-sum channel offsets, with each residual bounded to $[-1, 1]$ EV. The chroma head predicts the six off-diagonal entries of the color-correction residual as $0.25 \tanh(\cdot)$. In the default parameterization, the diagonal residuals are zero, so the diagonal of the color-correction matrix remains one. The optional gray-preserving variant sets each diagonal residual to the negative sum of the off-diagonal residuals in its row, enforcing unit row sums. Neither variant applies determinant, orthogonality, non-negativity, or additional projection constraints. For burst processing, we use patch size 32, stride 16, a two-dimensional Hann window, reliability scale $\kappa = 1.8$, and temporal weights $(0.05, 0.075, 0.125, 0.1875, 0.375)$ for the five historical frames of the six-frame causal burst.

C.2 Burst Denoising

At policy step t, burst denoising uses the causal RAW sequence

$$R_{t-K+1:t} = \{R_{t-K+1}, \ldots, R_{t-1}, R_t\}, \qquad R_i \in [0, 1]^{3 \times H \times W}, \tag{18}$$

where R_t is the current input frame. We use $K = 6$, square patches of size $N = 32$, stride 16, and a separable two-dimensional Hann window. For a historical-frame spectrum F_i and the current reference spectrum F_t, let $\Delta F_i = F_i - F_t$. The frequency-wise reliability and burst merge are

$$\begin{aligned} \sigma_i^2 &= \kappa \, \mathrm{mean}_f \, |\Delta F_i|^2, & r_i(f) &= \frac{\sigma_i^2}{|\Delta F_i(f)|^2 + \sigma_i^2 + 10^{-8}}, \\ F_{\mathrm{out}} &= F_t + \gamma_{\mathrm{burst}} \sum_{i=1}^{K-1} \omega_i r_i \Delta F_i, & \gamma_{\mathrm{burst}} &= 0.5. \end{aligned} \tag{19}$$

The index f spans the frequencies in each patch. We set $\kappa = 1.8$ and use historical weights $\boldsymbol{\omega} = (0.05, 0.075, 0.125, 0.1875, 0.375)$ from the oldest frame to the most recent historical frame. The current frame remains the reference anchor with coefficient one. We fix the fusion strength to $\gamma_{\mathrm{burst}} = 0.5$ in the full model and evaluate $\gamma_{\mathrm{burst}} \in \{0.2, 0.5, 0.8\}$ in Table 11. The reliability is computed independently for each patch, channel, and frequency. Small spectral differences are fused strongly, while motion and misalignment are down-weighted. Reflection padding is used when supported by the input dimensions and replication padding otherwise. Inverse Fourier transformation and Hann-weighted overlap-add reconstruct the denoised linear RAW frame X_t before exposure and tone mapping to avoid subsequent noise amplification. Figure 8 compares the resulting observations with a single noisy frame and direct averaging of the same $K = 6$ frames. Direct averaging substantially suppresses random sensor noise, but it also blurs the moving robot arm and manipulated objects. Preserving these dynamic structures is important for accurate grasping and for estimating the current manipulation state. In contrast, the frequency-domain reliability weighting attenuates inconsistent motion and misalignment during fusion, reducing noise while retaining sharper dynamic content and lower error relative to the noise-free observation.

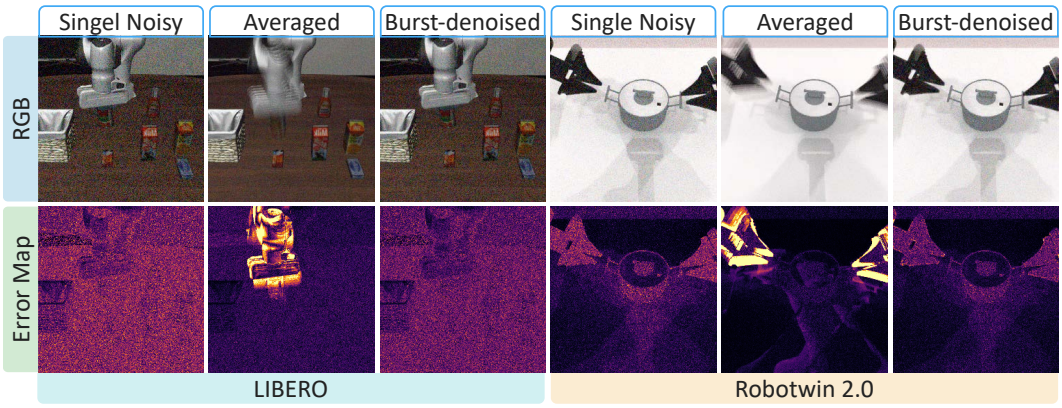


Figure 8: Burst-denoising comparison on agent-view observations from LIBERO (Liu et al., 2023) (left) and RoboTwin 2.0 (Chen et al., 2026b) (right). The top row compares a single noisy frame, direct averaging of $K = 6$ consecutive frames, and our frequency-domain burst denoising with $K = 6$. The bottom row shows the corresponding error maps relative to the noise-free clean image. Direct averaging removes substantial sensor noise but blurs moving arms and objects, whereas burst denoising better preserves dynamic scene content while suppressing noise.

C.3 Luminance Descriptor

The absolute luminance descriptor combines normalized linear- and log-luminance histograms with distribution quantiles and boundary-occupancy statistics. Specifically,

$$h_t^L = \left[\mathrm{Hist}_{64}(L_t), \mathrm{Hist}_{64}(\log(L_t + \epsilon)), \mathrm{Quant}(L_t), \rho_{\mathrm{dark}}(L_t), \rho_{\mathrm{clip}}(L_t)\right], \tag{20}$$

where each histogram contains 64 normalized bins over a fixed range and $\mathrm{Quant}(L_t) = [\pi_{0.01}, \pi_{0.05}, \pi_{0.50}, \pi_{0.95}, \pi_{0.99}]$ contains five intensity quantiles. For pixel index u, the boundary ratios are

$$\rho_{\mathrm{dark}}(L_t) = \frac{1}{HW}\sum_u \mathbf{1}[L_t(u) < 0.05], \qquad \rho_{\mathrm{clip}}(L_t) = \frac{1}{HW}\sum_u \mathbf{1}[L_t(u) > 0.98]. \tag{21}$$

$h_t^L \in \mathbb{R}^{135}$ comprises two 64-bin histograms, five quantiles, and two boundary ratios.

C.4 Chroma Descriptor

Let $\mathcal{V}_t = ([\boldsymbol{\chi}_t]^{\mathrm{R}}, [\boldsymbol{\chi}_t]^{\mathrm{G}}, [\boldsymbol{\chi}_t]^{\mathrm{B}}, [\mathbf{q}_t]_1, [\mathbf{q}_t]_2)$ contain the five component maps of $\boldsymbol{\chi}_t$ and $\mathbf{q}_t$. The chroma descriptor is

$$h_t^C = \mathrm{Concat}_{\psi \in \mathcal{V}_t}\left[\mathrm{Hist}_{B_C}(\psi), \mu(\psi), \sigma^2(\psi)\right], \tag{22}$$

where Hist_{B_C} is a normalized B_C-bin histogram over a fixed component-specific range. For pixel index u, the moments are

$$\mu(\psi) = \frac{1}{HW}\sum_u \psi(u), \qquad \sigma^2(\psi) = \frac{1}{HW}\sum_u \left(\psi(u) - \mu(\psi)\right)^2. \tag{23}$$

The resulting $h_t^C \in \mathbb{R}^{5(B_C+2)}$ summarizes the distributions of the three chromaticity maps and two log-chrominance maps.

C.5 Structured ISP Parameterization

White balance has two degrees of freedom. The network predicts the red and green offsets and constructs the zero-sum channel offsets as

$$w_t = (w_t^{\mathrm{R}}, w_t^{\mathrm{G}}, -w_t^{\mathrm{R}} - w_t^{\mathrm{G}}). \tag{24}$$

This parameterization separates relative channel gains from the global exposure parameter e_t.

D NEURAL ISP TRAINING PROTOCOL

This section specifies the shared frozen-policy training protocol in Section D.1, the baseline ISP configurations in Section D.2, and the optimization and temporal sampling used by RawVLA in Section D.3.

D.1 FROZEN-POLICY TRAINING PROTOCOL

We train an independent ISP module for every policy backbone and optimize all ISP modules for 2,000 optimizer steps. The pretrained vision or world-model backbone and action head are frozen and kept in evaluation mode. Their forward passes are not enclosed by `no_grad`, so the native action loss remains differentiable with respect to the processed image while no policy parameter is updated. Each method receives the same three-channel pseudo-RAW signal in the linear RAW representation and produces RGB observations for the policy. In the main benchmark experiments, no ISP module is trained with paired RGB reconstruction supervision. All methods use one NVIDIA A100 80GB GPU, per-device batch size one, and eight gradient-accumulation micro-steps. The shared optimizer is AdamW with learning rate 10^{-4}, $\beta = (0.9, 0.95)$, numerical epsilon 10^{-8}, 100 warmup updates, cosine decay to 10^{-6}, and gradient-norm clipping at 1.0. Weight decay is 10^{-8} for Qwen3 (StarVLA Community, 2026) and WM4A (StarVLA Community, 2026) backbones and 10^{-10} for PI0 (Black et al., 2024) and PI0.5 (Physical Intelligence et al., 2025).

D.2 BASELINE NEURAL ISP CONFIGURATIONS

We evaluate DarkISP (Guo et al., 2025), RAM (Gamrian et al., 2025), RAW-Adapter (Cui et al., 2026), and RAWild (Liu et al., 2026b) under the common 2k-step budget. Methods originally designed for Bayer inputs use deterministic channel adapters. DarkISP (Guo et al., 2025) receives $[R, G, B, G]$, while RAM (Gamrian et al., 2025) reduces its packed green channels to $[R, (G_r + G_b)/2, B]$. RAW-Adapter (Cui et al., 2026) applies its learned exposure, denoising and sharpening, white-balance, color-matrix, and optional LUT stages. RAWild (Liu et al., 2026b) uses the same pseudo-RAW observation as both its guide and apply image. These adaptations preserve a common camera signal while retaining each method's ISP parameterization.

To alleviate temporal inconsistency in methods designed for individual images, training samples trajectory-local windows of eight observations, applies the shared ISP independently to every observation, and averages the resulting action losses. This introduces no recurrence, cross-frame fusion, or explicit temporal-consistency loss. Direct-regression action heads use their native L1 objective, while diffusion and flow-matching heads retain their native velocity-prediction objective.

For LIBERO (Liu et al., 2023), the action horizon is eight for the comparison runs. For RoboTwin 2.0 (Chen et al., 2026b), the head, left-wrist, and right-wrist views are processed by the same ISP module. Qwen3-OFT (StarVLA Community, 2026) predicts a 50-step normalized action chunk, while FastWAM (StarVLA Community, 2026) uses its released 32-step flow-matching objective. The optimization budget remains 2,000 steps for both environments and all four comparison ISP modules.

D.3 RAWVLA TRAINING AND TEMPORAL SAMPLING

RawVLA follows the common optimization protocol above. Forward passes use *bfloat16* where supported, with numerically sensitive action computations promoted to *float32*.

Each sample contains $N_{\text{rec}} = 6$ recurrent endpoints and a $K = 6$ causal camera burst at every endpoint. Following the native policy observation intervals, the endpoint stride is eight environment steps for Qwen3 (StarVLA Community, 2026) and WM4A (StarVLA Community, 2026) policies and five for PI0 (Black et al., 2024) and PI0.5 (Physical Intelligence et al., 2025). The recurrent operating-point estimator reads the endpoint of each burst. The final policy image is produced from the six adjacent frames in the terminal burst. Earlier calls contribute through the recurrent state, and the complete six-update sequence is trained with backpropagation through time without detaching the hidden state or the previous ISP parameters. At episode boundaries, invalid history indices are clipped to the nearest valid frame.

Table 6: Action-loss coefficients used to train RawVLA with each frozen VLA policy.

Frozen VLA policy	λ_a
Qwen3-PI	0.0015
Qwen3-OFT	0.20
WM4A-Cosmos-GR00T	0.01
WM4A-Wan-OFT	0.02
PI0	0.15
PI0.5	0.10

Qwen3-OFT (StarVLA Community, 2026) and WM4A-Wan-OFT (StarVLA Community, 2026) use their mean-reduced action L1 losses. Qwen3-PI (StarVLA Community, 2026) and WM4A-Cosmos-GR00T (StarVLA Community, 2026) use eight independently sampled diffusion or flow-matching conditions per trajectory window. PI0 (Black et al., 2024) and PI0.5 (Physical Intelligence et al., 2025) use their native mean-reduced flow-matching losses. These stochastic repeats are Monte Carlo samples of the action objective and are distinct from the six recurrent observations and the eight optimizer-accumulation micro-steps.

The brightness and chroma weights are fixed to $\lambda_b = 0.001$ and $\lambda_{\text{chroma}} = 0.001$. The frozen VLA policies retain their native regression, diffusion, or flow-matching objectives, whose numerical scales and image-gradient magnitudes differ substantially. We therefore use the policy-specific coefficients in Table 6 to place action supervision on a comparable scale relative to the two fixed photometric regularizers.

The brightness auxiliary uses a two-sided L1 penalty with target mean 0.48. The chroma auxiliary is zero within an expanded unpaired RGB envelope and penalizes log-chroma descriptors outside it. Auxiliary renderings use straight-through hard clipping and branch-specific parameter detachment, whereas the deployed policy operator uses ordinary hard clipping. Paired default-lighting RGB is used only for no-gradient checkpoint selection and never contributes to the training loss.

E RawVLA-Bench Rendering Visualizations

Figures 9 and 10 compare the observations rendered by the Default ISP, four neural ISP baselines, and RawVLA on the LIBERO and RoboTwin 2.0 portions of RawVLA-Bench. Each comparison uses matched RAW observations from the same scene and robot state under low, normal, and overexposed illumination. Agent and wrist views are shown together to reveal whether each ISP produces a consistent multi-view representation for the downstream policy.

On LIBERO, the methods produce substantially different brightness, color, and local contrast from identical RAW observations. Several baseline renderings remain attenuated under low light or introduce strong chromatic and tonal changes, while RawVLA preserves task-relevant scene structure across both camera views.

The RoboTwin 2.0 examples show the same variation across three cameras and a different renderer. The qualitative comparison illustrates the observations received by the policy, while the closed-loop results in the main paper measure their utility for manipulation.

F Additional Robustness Evaluation under Severe Haze

An adaptive ISP can improve the visibility and task relevance of observations by adjusting their photometric rendering before they reach the policy. This capability may also benefit adverse visual conditions beyond the illumination and sensor degradations included in RawVLA-Bench. To test this broader robustness, we draw on the adverse environments introduced by LIBERO-Plus (Fei et al., 2026) and construct a substantially stronger haze condition. We then evaluate whether RawVLA can enhance these severely degraded observations and recover the closed-loop performance of a frozen PI-0 policy (Black et al., 2024).

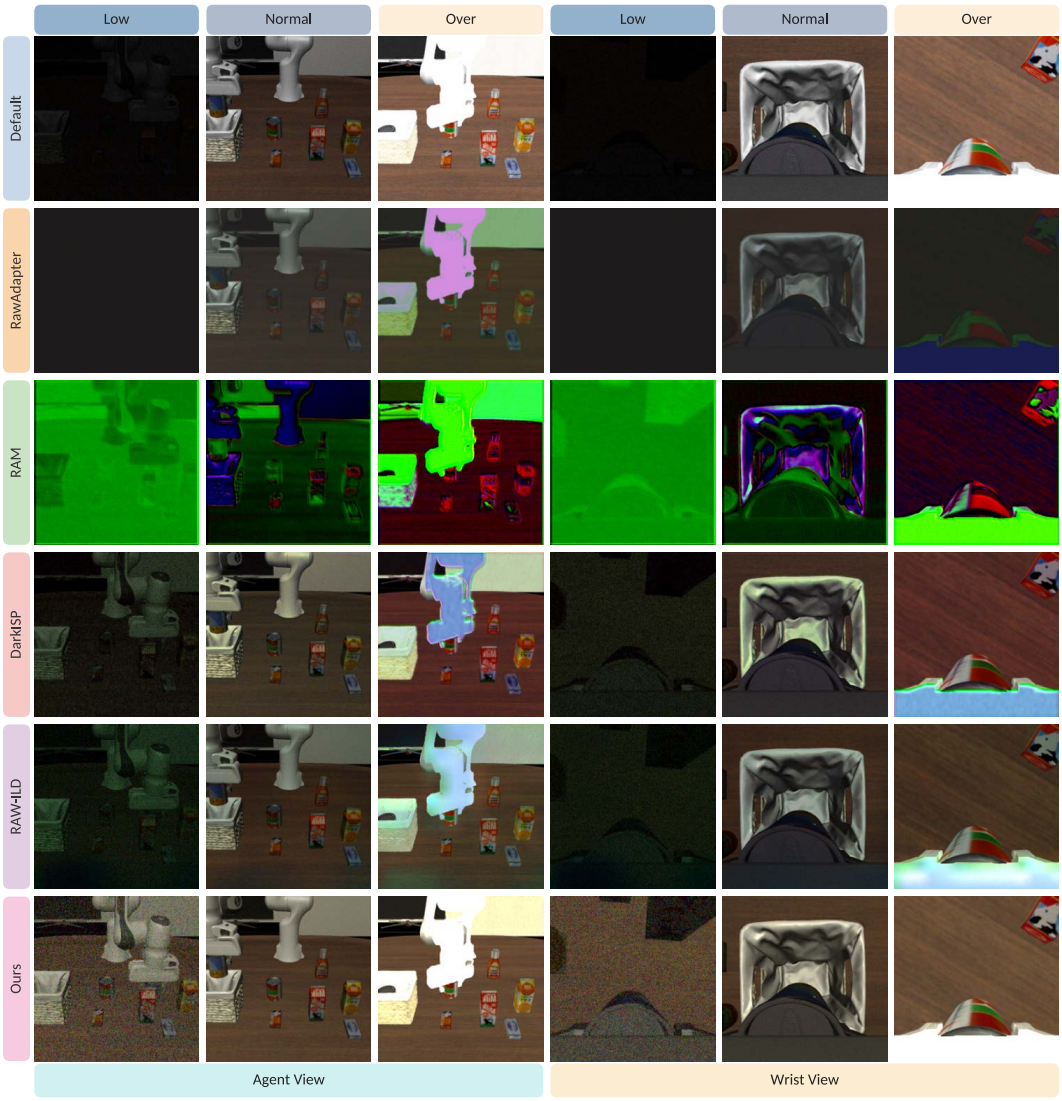


Figure 9: Processed agent-view and wrist-view observations from the LIBERO portion of RawVLA-Bench under low, normal, and overexposed illumination. The rows compare the Default ISP with neural ISP methods using Qwen3-OFT as the policy backbone.

Experimental Setting. We follow the image-space haze construction of LIBERO-Plus (Fei et al., 2026), but increase the haze coefficient to $\alpha_{\mathrm{h}} = 0.95$. For every pixel and color channel of a clean simulator RGB image $I_{\mathrm{clean}} \in [0, 255]$, we compute

$$I_{\mathrm{haze}} = \mathrm{clip}((1 - \alpha_{\mathrm{h}})I_{\mathrm{clean}} + \alpha_{\mathrm{h}} I_{\mathrm{air}}, 0, 255)\,, \qquad \alpha_{\mathrm{h}} = 0.95, \quad I_{\mathrm{air}} = 235. \tag{25}$$

The operation is performed in *float32* and the result is converted to *uint8*. Thus, only 5% of the original RGB signal is retained and 95% gray-white airlight is mixed into the observation. We add no sensor noise, blur, random image noise, or additional quantization. This experiment uses hazy RGB observations rather than sensor RAW. In this evaluation, "RAW" in RawVLA refers only to the model name.

At every simulator step, Equation 25 is applied independently to the agent and wrist views. The PI-0 baseline receives the two hazy views directly. In the enhanced condition, each view is first processed by an $\alpha_{\mathrm{h}} = 0.95$-specific RawVLA checkpoint and the resulting RGB images are passed to the same frozen PI-0 policy. Enhancement runs at the simulator resolution, before the native PI-0

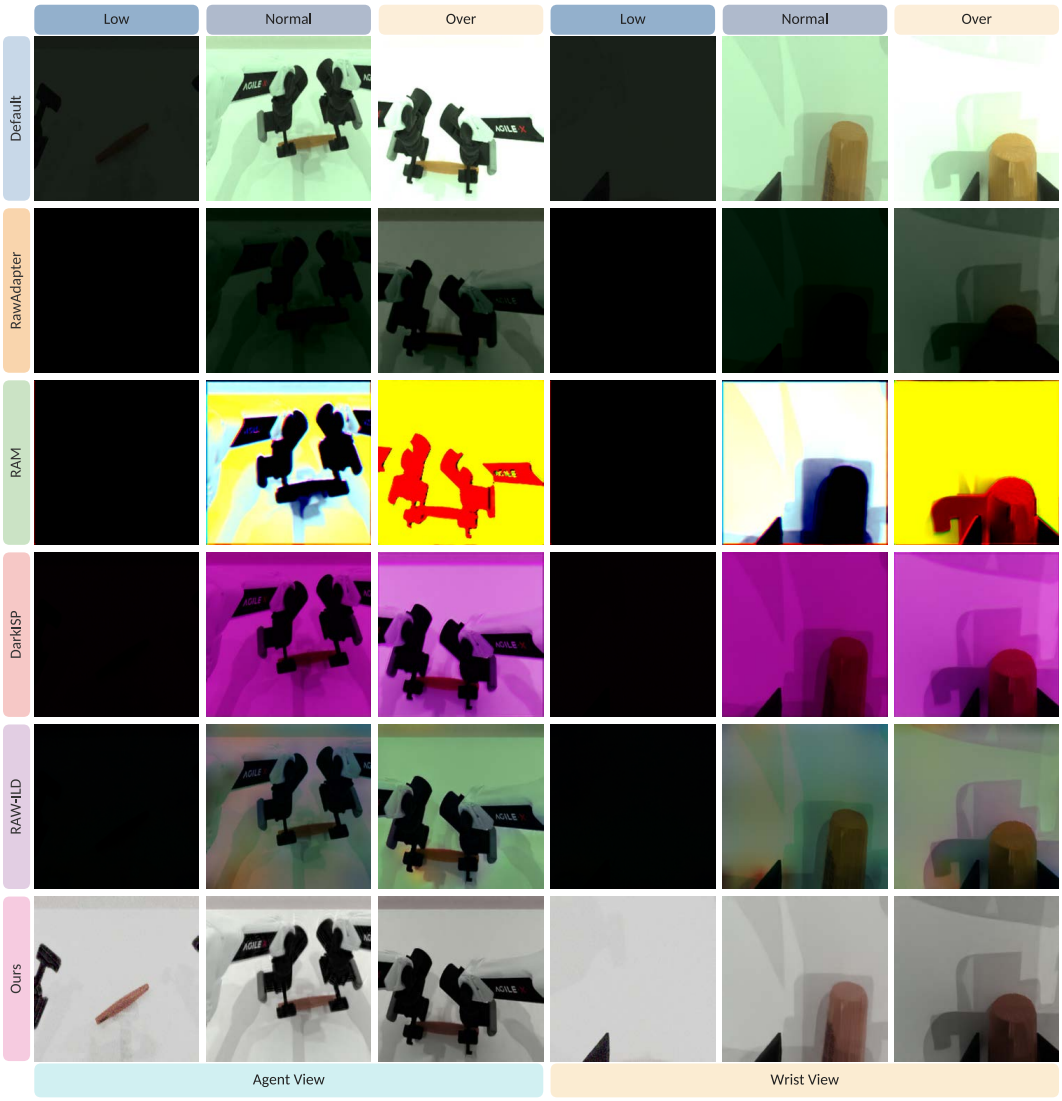


Figure 10: Processed agent-view and wrist-view observations from the RoboTwin 2.0 portion of RawVLA-Bench under low, normal, and overexposed illumination. The rows compare the Default ISP with neural ISP methods using PI-0.5 as the policy backbone.

pipeline resizes each view to 224×224. We use the standard causal RawVLA inference procedure without access to future observations or clean reference frames. Language instructions, robot states, control frequency, action horizon, and all other policy settings are held fixed. Figure 11 shows representative observations from both camera views before and after haze enhancement.

We train a haze-specific instance of the architecture described in Section 3 using paired reconstruction supervision rather than the frozen-policy objective in Section D. Training runs for 1,500 updates on paired LIBERO replay images from both camera views. Each target is a clean simulator RGB frame, and its input is generated online from that same frame using the exact $\alpha_{\mathrm{h}} = 0.95$ and $I_{\mathrm{air}} = 235$ transform, without strength jitter. The clean reconstruction targets are used only during training and are unavailable during closed-loop evaluation.

The two policy conditions use identical tasks, official initial states, and episode indices. PI-0 diffusion noise is generated deterministically from the language instruction and robot state to reduce

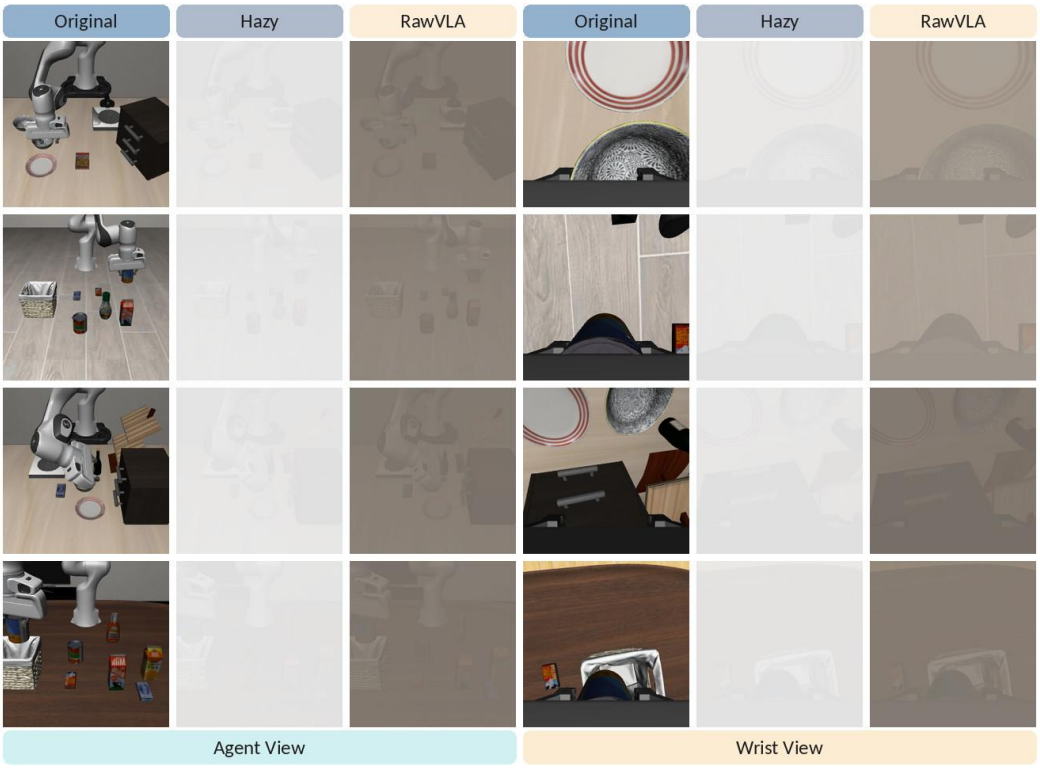


Figure 11: Representative agent-view and wrist-view observations from LIBERO under severe haze. Each triplet shows the original simulator RGB image, the hazy input generated with $\alpha_{\mathrm{h}} = 0.95$ and $I_{\mathrm{air}} = 235$, and the corresponding RGB output produced by RawVLA. The same enhanced observations are provided to the frozen PI-0 policy during closed-loop evaluation.

Table 7: PI-0 success rates (%) on LIBERO under severe haze ($\alpha_{\mathrm{h}} = 0.95$, $I_{\mathrm{air}} = 235$), evaluated with 50 rollouts per task. Avg. is the macro-average over the four suites.

Method	Spatial	Object	Goal	LIBERO-10	Avg.
PI-0	78.0	84.0	82.0	22.0	66.5
PI-0 with RawVLA	98.0	96.0	96.0	63.6	88.4
Improvement (pp)	+20.0	+12.0	+14.0	+41.6	+21.9

sampling variation between the paired visual conditions. Success is determined by the standard LIBERO task predicate. Each suite includes ten tasks with 50 distinct rollouts per task.

Evaluation on Hazy LIBERO. On held-out image pairs, enhancement reduces mean absolute error from 0.4271 to 0.1688 and raises PSNR from 6.56 to 13.60 dB. Table 7 reports closed-loop suite success rates and the macro-average that weights the four suites equally.

Across the four equally sized suites, PI-0 achieves an average success rate of 66.5%. Processing the hazy observations with RawVLA before PI-0 raises this average to 88.4%, an improvement of 21.9 percentage points. On LIBERO-10, RawVLA enables PI-0 to succeed in 230 rollouts that fail without enhancement, while 22 rollouts show the opposite outcome. The paired comparison therefore yields 208 additional successful rollouts. A two-sided exact McNemar test confirms that this improvement is statistically significant, with $p = 7.16 \times 10^{-45}$.

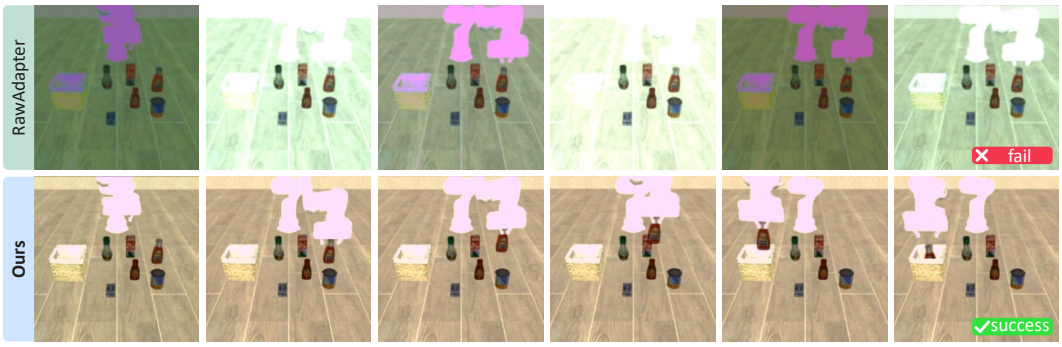


Figure 12: Temporal rendering stability on an overexposed LIBERO (Liu et al., 2023) trajectory with Qwen3-OFT (StarVLA Community, 2026). RAW-Adapter (Cui et al., 2026) exhibits pronounced frame-to-frame appearance fluctuations that destabilize the policy and lead to task failure, whereas RawVLA maintains temporally consistent observations and successfully completes the task.

Table 8: Ablation of the three training objectives. The full objective is compared with action-free auxiliary supervision and each loss used in isolation. Loss weights are fixed and are not swept.

	Training objective			LIBERO / Qwen3-OFT						RoboTwin 2.0 / PI-0.5					
Variant	$\mathcal{L}_{\text{action}}$	$\mathcal{L}_{\text{bright}}$	$\mathcal{L}_{\text{color}}$	Ext.Low	Low	Normal	Over	Ext.Over	Avg.	Ext.Low	Low	Normal	Over	Ext.Over	Avg.
Action loss only	✓			0.00	0.00	97.75	44.50	29.25	34.30	0.00	0.00	67.72	54.76	46.40	33.70
w/o action loss		✓	✓	9.50	72.75	97.50	44.25	25.50	49.90	11.81	55.12	62.20	59.52	51.20	47.94
Brightness only		✓		4.00	39.50	97.25	41.25	21.50	40.70	7.87	41.73	65.35	53.17	51.20	43.83
Color only			✓	0.00	0.00	97.50	43.50	20.00	32.20	0.00	12.60	53.54	52.38	26.40	23.73
Full RawVLA	✓	✓	✓	14.05	81.35	96.95	43.40	20.85	51.32	25.98	63.78	67.72	65.08	63.20	57.12

G Additional Ablation Studies

The following tables separate structured ISP operators from burst-length design choices. LIBERO (Liu et al., 2023) uses Qwen3-OFT (StarVLA Community, 2026), and RoboTwin 2.0 (Chen et al., 2026b) uses PI-0.5 (Physical Intelligence et al., 2025), matching the main ablation protocol.

Because exhaustive evaluation of every additional ablation is computationally expensive, we use a reduced evaluation budget for LIBERO (Liu et al., 2023) in this appendix. We retain all 40 tasks across the four LIBERO (Liu et al., 2023) suites, but evaluate each task from 10 distinct initial states, rather than the 50 initial states used for the main results. For RoboTwin 2.0 (Chen et al., 2026b), we retain the full evaluation protocol with 13 tasks and 50 distinct initial states per task.

Training Objectives. Table 8 isolates the action, brightness, and chroma objectives. The action objective alone preserves normal-light performance but fails under severe illumination changes, whereas the two photometric auxiliaries without action supervision improve robustness but do not fully align the rendered observations with the frozen policy. Combining all three objectives gives the strongest overall performance, reaching 51.32% on LIBERO (Liu et al., 2023) and 57.12% on RoboTwin 2.0 (Chen et al., 2026b). These results show that the action loss provides policy alignment, while the brightness and chroma terms stabilize adaptation in the adverse regimes where action supervision alone is difficult to optimize.

Structured ISP Operators. Table 9 evaluates exposure, white balance, color correction, and tone mapping on LIBERO (Liu et al., 2023). Removing exposure causes the largest overall degradation, reducing average success from 62.24% to 50.79%. Removing tone mapping also lowers robustness, while white balance and the color-correction matrix provide complementary gains across illumination conditions. The shared-tone control nearly matches the full model, supporting the use of a single achromatic curve across channels while leaving chromatic adaptation to white balance and color correction.

Table 9: Ablation study of RawVLA's modules on LIBERO with Qwen3-OFT (StarVLA Community, 2026). We report success rates (%) across five illumination levels and overall. Each ablated variant removes one ISP module while keeping the remaining architecture unchanged. Shared tone denotes the shared-head equivalence control.

Variant	Ext.Low			Low			Normal			Over			Ext.Over			Overall		
	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg	SOG	Long	Avg
w/o Exposure	6.87	3.60	6.05	42.80	22.00	37.60	98.53	92.20	96.95	92.34	60.00	84.25	38.47	1.00	29.10	55.80	35.76	50.79
w/o White-balance	15.87	12.60	15.05	87.46	62.00	81.10	98.53	94.20	97.45	91.67	53.00	82.00	42.80	2.00	32.60	67.27	44.76	61.64
w/o CCM	12.20	12.60	12.30	86.46	62.00	80.35	97.86	95.20	97.20	92.67	58.00	84.00	43.13	2.00	32.85	66.47	45.96	61.34
w/o Tone-mapping	7.53	5.60	7.05	73.46	46.00	66.60	97.53	91.20	95.95	90.67	54.00	81.50	45.13	2.00	34.35	62.87	39.76	57.09
⌞ Shared tone	15.20	10.60	14.05	86.80	63.00	80.85	98.53	92.20	96.95	92.00	54.00	82.50	47.47	1.00	35.85	68.00	44.16	62.04
Full RawVLA	15.20	10.60	14.05	87.13	64.00	81.35	98.53	92.20	96.95	91.67	55.00	82.50	47.80	2.00	36.35	68.07	44.76	62.24

Table 10: Inference-only ISP component ablations on RoboTwin 2.0 using PI-0.5. All ablated variants use the same frozen checkpoint and force one ISP component to identity. Full (inference-only) keeps all components enabled, while Full RawVLA reports the main benchmark result.

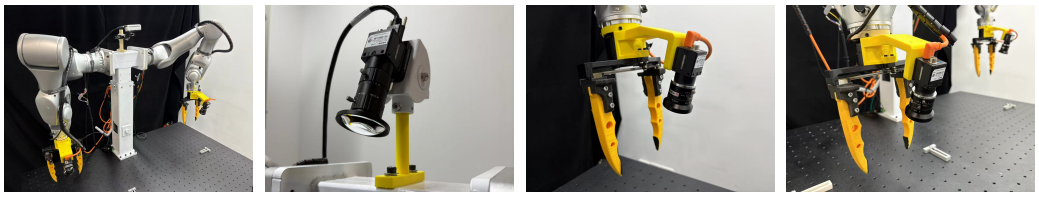
Variant	Ext.Low	Low	Normal	Over	Ext.Over	Avg.
w/o Exposure	12.60	18.11	65.35	61.11	60.00	43.35
w/o White-balance	25.98	58.27	66.14	65.08	54.40	53.96
w/o CCM	25.20	61.42	69.29	57.14	62.40	55.06
w/o Tone-mapping	22.05	50.39	64.57	65.08	60.00	52.37
Full RawVLA	25.98	63.78	67.72	65.08	63.20	57.12

Table 11: Burst-denoising ablations. We study the burst length K, global fusion weight η, and frequency-wise reliability decomposition. We compare the rendered output $\hat{I}$ against the corresponding noise-free simulation ground truth I_{GT} using PSNR (dB) and SSIM. Average is computed over the ExtremeLow, Low, and Normal illumination regimes. Underlined settings denote the default configuration.

Benchmark	Ablation	Setting	Ext.Low		Low		Normal		Average	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
LIBERO Qwen3-OFT	Burst length	$K = 1$	18.93	0.117	19.99	0.261	36.69	0.863	25.21	0.414
		$K = 3$	20.15	0.144	21.34	0.304	37.66	0.893	26.38	0.447
		$\underline{K = 6}$	21.98	0.195	23.27	0.373	37.57	0.921	27.61	0.497
	Fusion strength	$\eta = 0.2$	19.63	0.132	20.79	0.286	37.44	0.882	25.95	0.433
		$\underline{\eta = 0.5}$	21.98	0.195	23.27	0.373	37.57	0.921	27.61	0.497
		$\eta = 0.8$	20.78	0.160	22.04	0.328	37.89	0.905	26.90	0.465
	Frequency	w/o decomposition	20.54	0.152	20.43	0.295	33.21	0.908	24.73	0.452
RoboTwin 2.0 PI-0.5	Burst length	$K = 1$	18.16	0.082	27.30	0.636	48.45	0.986	31.30	0.568
		$K = 3$	18.16	0.082	27.27	0.635	46.35	0.984	30.59	0.567
		$\underline{K = 6}$	18.18	0.082	27.28	0.634	47.35	0.985	30.94	0.567
	Fusion strength	$\eta = 0.2$	18.16	0.082	27.22	0.635	45.20	0.982	30.20	0.566
		$\underline{\eta = 0.5}$	18.18	0.082	27.28	0.634	47.35	0.985	30.94	0.567
		$\eta = 0.8$	18.14	0.082	27.14	0.635	43.53	0.978	29.60	0.565
	Frequency	w/o decomposition	18.08	0.079	26.35	0.628	38.47	0.967	27.63	0.558

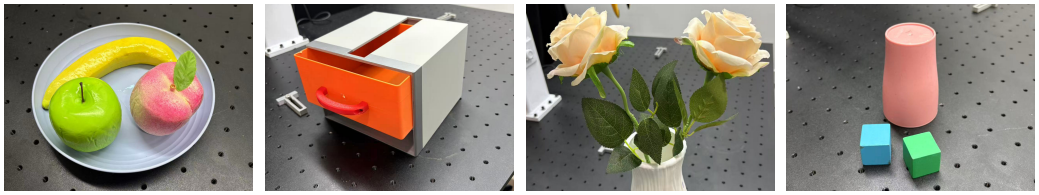
Cross-Benchmark Operator Effects. Table 10 repeats the operator analysis on RoboTwin 2.0 (Chen et al., 2026b). Exposure remains the most consequential component. Disabling it reduces average success from 57.12% to 43.35%. The remaining operator ablations also reduce the overall result, confirming that exposure, chromatic correction, and tone mapping remain complementary across environments and policy backbones.

Burst-Denoising Design. Table 11 studies burst length, fusion strength, and frequency-wise reliability decomposition using reconstruction quality relative to noise-free observations. Longer bursts substantially improve LIBERO (Liu et al., 2023), with $K = 6$ outperforming single-frame process-



(a) ROKAE AR5 (b) Agent-view camera (c) Left wrist & camera (d) Right wrist & camera

Figure 13: Real-world dual-arm platform and camera setup.



(a) Fruits (b) Drawer (c) Flowers (d) Cup & blocks

Figure 14: Objects used in the four real-world tasks.

ing, whereas RoboTwin 2.0 (Chen et al., 2026b) shows diminishing returns because its observations contain less recoverable temporal noise. A moderate fusion strength provides the best average trade-off on both benchmarks. Removing the frequency-wise decomposition produces the clearest degradation, confirming that motion-aware spectral reliability is important for suppressing noise without indiscriminately averaging dynamic content.

Temporal Stability. Figure 12 complements the quantitative ablations with an overexposed LIBERO (Liu et al., 2023) trajectory. Compared with the frame-wise RAW-Adapter (Cui et al., 2026), the recurrent formulation of RawVLA avoids abrupt rendering changes and maintains observations that support successful task execution.

H Real-World Experimental Setup

This section describes the robotic platform and cameras in Section H.1, data acquisition and calibration in Section H.2, policy training and deployment in Section H.3, and qualitative executions in Section H.4.

H.1 Robotic Platform and Camera Configuration

Our real-world platform comprises two ROKAE AR5 robot arms and three synchronized HIKROBOT MV-CS020-10UC industrial cameras, as shown in Figure 13. Each arm is equipped with a Songling gripper fitted with custom 3D-printed finger extensions to increase its usable reach for the selected manipulation tasks. The agent-view and two wrist-view streams are captured by the same camera model. The agent-view camera uses a 4-mm fixed-focal-length lens at $f/2.8$, while each wrist camera uses a 3.5-mm lens at $f/4$.

H.2 Data Acquisition and Camera Calibration

Figure 14 shows the task-specific objects used for pick-and-place, drug-drawer manipulation, flower placement, and stack-and-cover. The cameras continuously stream 12-bit mosaiced Bayer RAW measurements at 30 FPS. We subtract the camera-specific black level, normalize by the usable white level, and apply the cameras' default demosaicing to obtain the three-channel linear RAW representation defined in Appendix A. We then downsample each view to 228×171. For each task, we collect over 200 trajectories with randomized initial arm configurations, recording all three views, robot states, and expert actions at 30 FPS. The demonstrations retain the synchronized linear RAW observations and expert trajectories. This sensor-linear representation allows exposure and noise

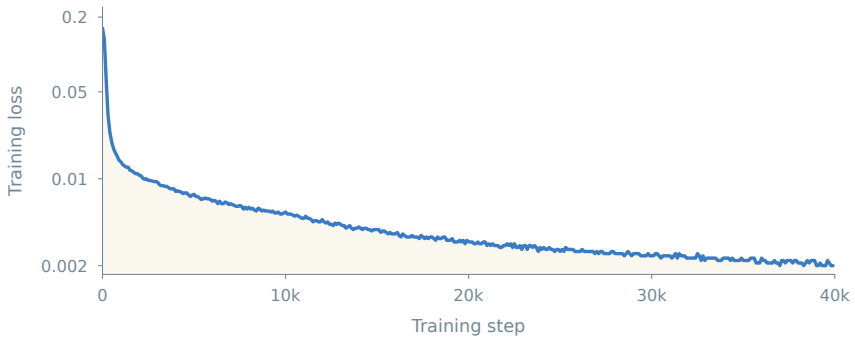


Figure 15: Action-loss curves for PI0.5 (Physical Intelligence et al., 2025) fine-tuning on the four real-world tasks, using 909 demonstrations and 263,863 frames in total.

Task	Language prompt
`pick_and_place`	Place the green apple, pink object, and yellow banana in the blue bowl.
`drag_the_drawer`	Open the drawer and place the yellow block inside.
`pick_flower`	Pick up the two flowers and place them in the white vase.
`stack_and_cover`	Stack the green and blue blocks and cover them with the pink cup.

Table 12: Language prompts for the four real-world tasks.

to be modified directly without approximately inverting the nonlinear tone mapping and display encoding already applied to RGB images.

Because the agent-view and wrist-view cameras use different lenses and apertures, their uncalibrated RAW streams exhibit different brightness levels. Before data collection, we perform a three-camera photometric calibration and increase the agent-view sensor gain so that the three RAW streams have matched illumination levels under the same scene lighting. This calibration provides a consistent multi-view input while retaining RAW-domain measurements for the learned ISP.

H.3 Policy and Neural ISP Training and Deployment

We initialize PI-0.5 (Physical Intelligence et al., 2025) from its pretrained base checkpoint and fully fine-tune the policy on 909 demonstrations comprising 263,863 frames across the four real-world tasks under normal illumination. Each training sample contains the current observations from the three cameras, the 16-dimensional robot state, and the task instruction. The target is a 16-step chunk of absolute 16-dimensional actions sampled from the 15 FPS demonstrations. We use four-task quantile statistics to normalize states and actions and apply state-text dropout with probability 0.2.

Training runs for 40,000 updates on eight NVIDIA B300 GPUs with a global batch size of 256. We use AdamW with a peak learning rate of 5×10^{-5}, a 10,000-step linear warmup, global gradient clipping at 1.0, and an exponential moving average decay of 0.999. Figure 15 shows the action-loss trajectory over the four-task fine-tuning process.

After obtaining the fine-tuned PI-0.5 policy, we freeze its parameters and train the neural ISP modules. Each learned ISP is trained on both the recorded normal-light linear RAW observations and synthetic low-light counterparts. We generate the low-light inputs directly from the recorded linear RAW images by applying an exposure offset of -6 EV and adding simulated sensor noise. The synchronized expert trajectories remain unchanged and provide supervision for the ISP training. Each module is trained for 2,000 steps, yielding one ISP that is shared across both lighting conditions.

At deployment, all evaluations use a single PI-0.5 (Physical Intelligence et al., 2025) checkpoint fine-tuned only under normal illumination. The Default ISP uses the same fixed camera pipeline under both normal and low light. DarkISP and RawVLA each deploy one learned ISP checkpoint under both conditions, without condition-specific retraining or model switching. This protocol measures cross-illumination robustness while holding the downstream policy and ISP model fixed. The policy

runs in streaming mode and predicts a 16-step action chunk from the latest multi-view observation at 4 Hz while the robot continues executing the previously committed trajectory. When a new chunk becomes available, a quadratic programming trajectory optimizer interpolates the remaining previous trajectory toward the new prediction before execution. This transition accounts for robot inertia and avoids abrupt changes between consecutive chunks, preserving smooth and continuous motion. For every ISP method and illumination condition, we conduct 50 independent trials per task. The resulting success rates are reported in Table 3, and the ISP throughput comparison is given in Table 4. We condition the policy on the task-specific language prompts listed in Table 12.

H.4 Qualitative Task Executions

Figures 16, 17, 18, and 19 visualize complete executions of *pick & place*, *drug drawer*, *pick flower*, and *stack & cover*. The normal illumination sequences demonstrate the nominal behavior of RawVLA. Under low light, the Default ISP produces nearly zero-valued observations and the policy does not initiate motion. Dark-ISP (Guo et al., 2025) recovers some scene structure, but the policy stalls during the first stage of each multi-step task. RawVLA instead preserves usable agent and wrist observations and completes every sequence. Low-light execution takes longer than execution under clear normal illumination because the policy repeatedly observes the scene and refines the end-effector pose during precise steps such as grasping the blocks and aligning a flower with the vase opening. Despite these additional adjustments, the policy completes each fine-grained operation successfully. Videos of all executions are included in the supplementary material.

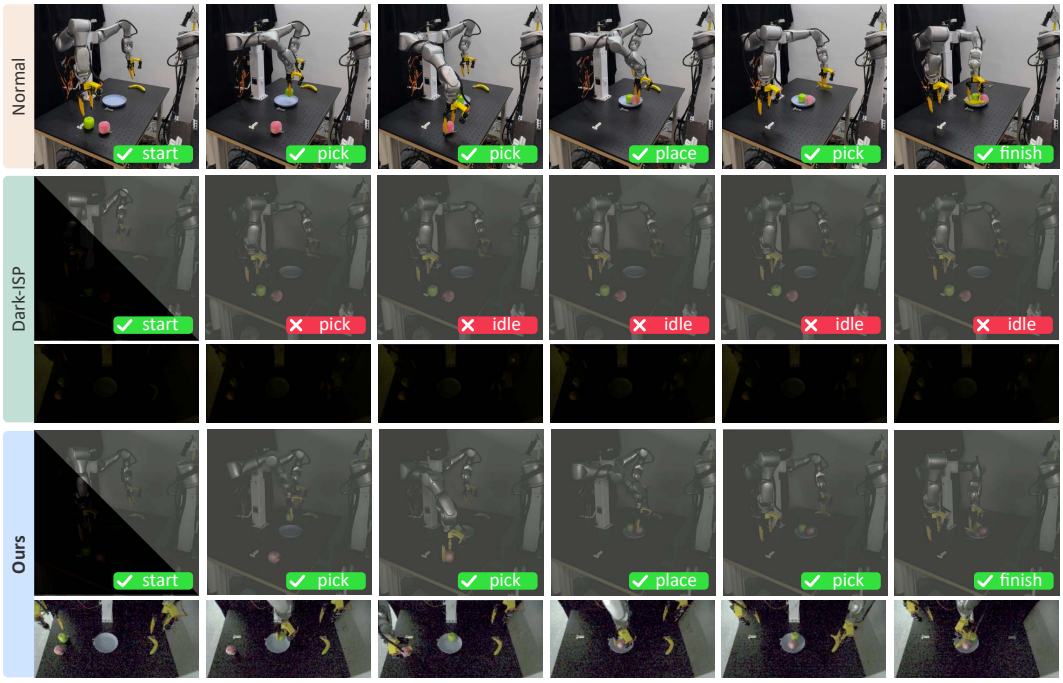


Figure 16: *Pick & place* executions under normal and low-light illumination. The third-person frames receive postprocessing enhancement only for visualization and are not provided to the policy. The lower strips show the rendered agent views produced by Dark-ISP (Guo et al., 2025) and RawVLA under low light.



Figure 17: *Drug drawer* executions under normal and low-light illumination. The third-person frames receive postprocessing enhancement only for visualization and are not provided to the policy. The lower strips show the rendered agent views produced by Dark-ISP (Guo et al., 2025) and RawVLA under low light.

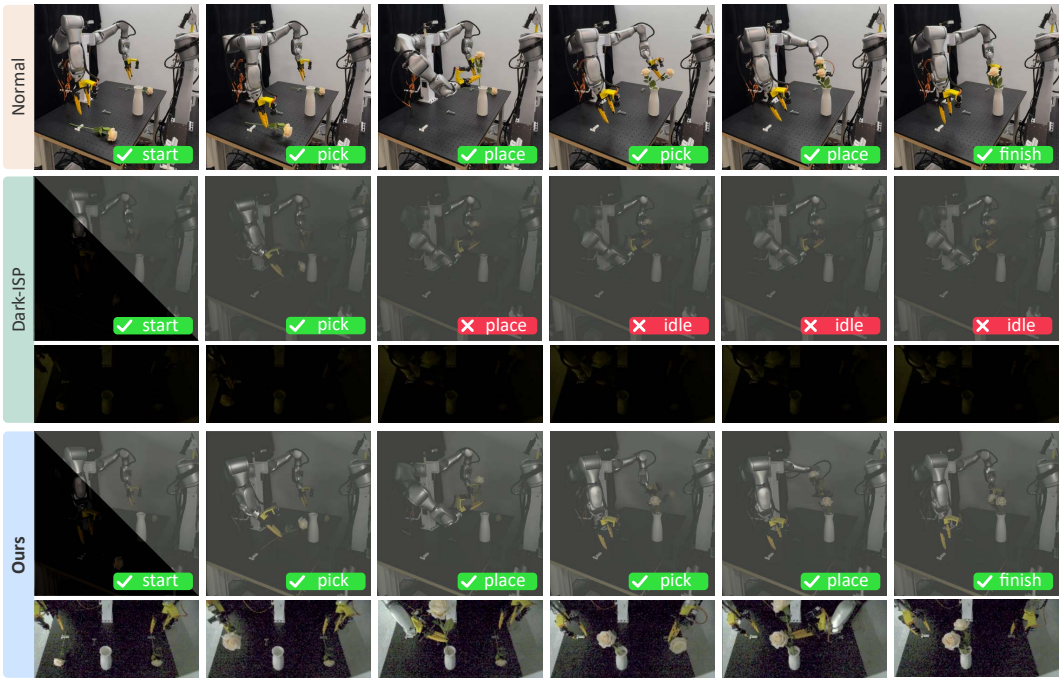


Figure 18: *Pick flower* executions under normal and low-light illumination. The third-person frames receive postprocessing enhancement only for visualization and are not provided to the policy. The lower strips show the rendered agent views produced by Dark-ISP (Guo et al., 2025) and RawVLA under low light.

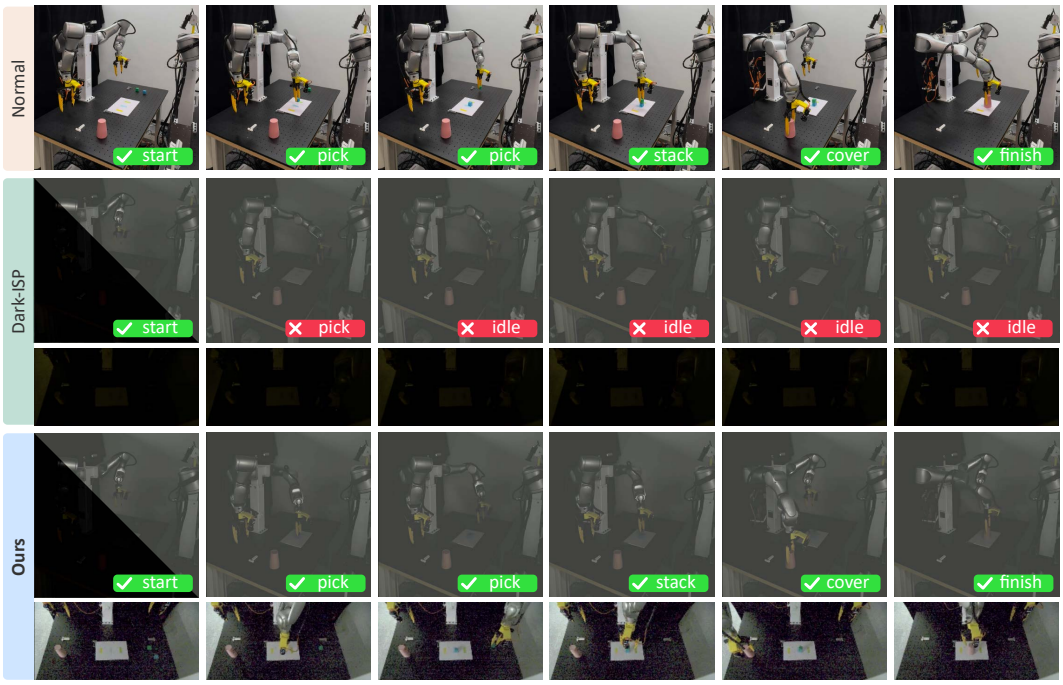


Figure 19: *Stack & cover* executions under normal and low-light illumination. The third-person frames receive postprocessing enhancement only for visualization and are not provided to the policy. The lower strips show the rendered agent views produced by Dark-ISP (Guo et al., 2025) and RawVLA under low light.

I WAM Illumination Robustness on LIBERO

We additionally evaluate RawVLA with the WM4A-Wan and WM4A-Cosmos WAM (StarVLA Community, 2026) backbones in Table 13. The results remain uniformly low, and none of the learned ISP methods produces a substantial or consistent improvement across the two backbones. Even the strongest averages differ by only 1.19 percentage points for WM4A-Wan and 0.14 percentage points for WM4A-Cosmos, which is insufficient to establish a meaningful robustness gain.

This limited effect is consistent with the sensitivity analysis in Section 2. WAM policies are highly sensitive to pixel-level appearance changes and operate within a narrower stable imaging range than the VLA backbones. Consequently, an ISP transformation that improves visibility may still move the observation away from the image distribution expected by the generative world model. Optimization is also less direct because the action objective must propagate through the VAE-based visual generation pathway before reaching the neural ISP. This long gradient path can weaken or distort the supervision available for learning task-relevant image processing. These results suggest that conventional action-driven ISP training is not sufficient for the evaluated WAM backbones. More direct supervision and WAM-specific imaging interfaces are left for future study.

Table 13: Success rates (%) of ISP modules across five illumination levels on LIBERO using WAM backbones. Avg reports the mean across illumination levels for each backbone.

Illumin.	Backbone	Default	DarkISP	RAM	RAWAdapter	RAWild	Ours
Extreme Lowlight	WM4A-Wan	0.00	0.05	0.00	0.00	0.00	0.75
	WM4A-Cosmos	0.00	0.00	0.00	0.00	0.00	0.20
Lowlight	WM4A-Wan	0.05	0.05	0.05	0.00	0.00	0.55
	WM4A-Cosmos	0.05	0.00	0.00	0.00	0.00	0.75
Normal	WM4A-Wan	76.10	76.50	0.00	38.05	75.15	78.20
	WM4A-Cosmos	85.75	0.75	1.75	2.45	4.25	83.45
OverExp	WM4A-Wan	0.50	1.45	0.00	0.00	0.55	4.55
	WM4A-Cosmos	1.90	0.05	0.00	0.00	0.35	3.50
Extreme OverExp	WM4A-Wan	0.00	0.45	0.00	0.00	0.00	0.40
	WM4A-Cosmos	0.00	0.00	0.00	0.00	0.30	0.50
Avg	WM4A-Wan	15.42	15.70	0.01	7.61	15.14	16.89
	WM4A-Cosmos	17.54	0.16	0.35	0.49	0.98	17.68

J Limitations and Future Studies

Our study has several limitations that motivate future work. The simulated experiments use three-channel linear pseudo-RAW observations rather than camera-native mosaiced measurements and therefore do not capture the full diversity of sensor spectral responses, demosaicing artifacts, lens effects, or proprietary camera pipelines. Extending RawVLA-Bench with calibrated physical sensors and multiple camera models would provide a more complete evaluation of RAW-domain manipulation across diverse robotic platforms and deployment settings.

The current training protocol learns a separate neural ISP for each downstream policy. Although our evaluation covers multiple VLA backbones, a single ISP that generalizes across policies, embodiments, and tasks remains an open challenge. Promising directions include policy-agnostic pre-training, lightweight online adaptation, and extending adaptive image processing to world-action models through dedicated objectives and imaging interfaces. The real-world experiments are also limited to a small set of indoor manipulation tasks and lighting conditions. Broader evaluations could cover outdoor scenes, spatially varying illumination, motion blur, weather effects, and longer task horizons. The severe-haze experiment further relies on synthesized RGB observations. Evaluating adverse weather with physically captured RAW data and jointly modeling multiple degradations are important directions for future study.

K CONTROLLED ISP PERTURBATION PROTOCOL

The two evaluation environments expose different renderer products. LIBERO (Liu et al., 2023) uses MuJoCo RGB8 camera observations resized to the policy input resolution, whereas RoboTwin 2.0 (Chen et al., 2026b) uses SAPIEN float RGB textures in $[0, 1]$ without an intermediate RGB8 quantization. In both cases, we apply the unprocessing in Appendix A with channel gains $(1.8, 1.0, 1.7)$ and inverse global gain 1.0, then quantize the resulting three-channel linear proxy as $R_{10} = \mathrm{round}(1023R)$. No Bayer mosaic or camera-specific color correction matrix is introduced. All policy inputs remain *float32* in $[0, 1]$, and only one perturbation axis is enabled in each evaluation. Section K.1 defines the perturbation operators and their ISP stages, while Sections K.2 and K.3 provide the environment-specific settings.

K.1 PERTURBATION OPERATORS AND PROCESSING STAGES

Exposure. Exposure acts on the linear sensor signal before nonlinear display rendering. In a physical camera, exposure time and analog gain determine the signal collected or amplified before digitization, while later digital gain adjusts brightness inside the ISP. Our perturbation uses a single RAW-domain multiplier to capture their shared effect on signal scale. Negative EV suppresses weak measurements and makes subsequent quantization more destructive. Positive EV saturates bright measurements at the sensor range. Applying the reciprocal gain can restore brightness but cannot recover samples already clipped or collapsed to the same quantization level. We therefore compare direct and recovered forms to separate reversible rescaling from irreversible information loss. For an offset e measured in exposure values (EV), we define

$$R_e = \mathrm{clip}(2^e R, 0, 1), \qquad Y_e^{\mathrm{exp}} = \mathcal{P}_0(R_e), \qquad Y_0^{\mathrm{exp}} = \mathcal{P}_0(R). \tag{26}$$

One EV doubles or halves the linear RAW signal, and clip acts elementwise.

Sensor Noise. Photon arrival follows signal-dependent counting statistics, while readout electronics add a signal-independent noise floor. Low illumination or short exposure reduces the number of collected photoelectrons and therefore lowers the signal-to-noise ratio even when an ISP later restores the mean brightness. We reproduce this process by attenuating linear RAW with a negative capture EV, sampling heteroscedastic Gaussian noise whose variance is the sum of shot and read components, and applying the matching recovery gain. The noise is added before RAW recovery and the default ISP, so recovery amplifies both the desired signal and the acquisition noise. This differs from adding a fixed image-space corruption after rendering. After forming R_e with Equation 26, we sample

$$\begin{aligned} \varepsilon_u &\sim \mathrm{Normal}\big(0, \sigma_{\mathrm{shot}}^2 [R_e]_u + \sigma_{\mathrm{read}}^2\big)\,, \\ \widetilde{R}_e &= \mathrm{clip}\big(2^{-e}\,\mathrm{clip}(R_e + \varepsilon, 0, 1), 0, 1\big)\,, \\ Y_e^{\mathrm{noise}} &= \mathcal{P}_0(\widetilde{R}_e). \end{aligned} \tag{27}$$

Here u indexes RAW samples, σ_{shot}^2 and σ_{read}^2 control shot and read noise, and $\varepsilon = \{\varepsilon_u\}_u$. Reciprocal gain restores the mean exposure while preserving the reduced signal-to-noise ratio.

Chromatic Response. Chromatic processing primarily occurs through white balance and color rendering. White balance applies channel-dependent gains so that the sensor response to an assumed illuminant becomes neutral. Moving the target white point therefore changes the absolute tint of the entire rendered observation. Color rendering then maps the balanced sensor channels into a display color space and controls relative hue and saturation. To probe this second role without changing luminance, we rotate and scale the OKLab chroma plane after the default ISP while preserving lightness. The white-point family measures sensitivity to illuminant-dependent calibration, whereas the relative-color family measures sensitivity to relationships among object and scene colors. Let $\boldsymbol{\nu}_0$ be the reference white point in the $u'v'$ plane and let $(a_{\mathrm{ok}}, b_{\mathrm{ok}})$ be the OKLab chroma coordinates (Ottosson, 2020) of Y_0. The two chromatic perturbation families are

$$\begin{aligned} \boldsymbol{\nu}_{\delta_{\mathrm{wp}},\theta} &= \boldsymbol{\nu}_0 + \delta_{\mathrm{wp}}(\cos\theta, \sin\theta), \\ Y_{\delta_{\mathrm{wp}},\theta}^{\mathrm{wp}} &= \mathcal{P}_{\delta_{\mathrm{wp}},\theta}^{\mathrm{wp}}(R), \\ \begin{bmatrix} a'_{\mathrm{ok}} & b'_{\mathrm{ok}} \end{bmatrix}^\top &= s\,\mathrm{Rot}(\theta) \begin{bmatrix} a_{\mathrm{ok}} & b_{\mathrm{ok}} \end{bmatrix}^\top. \end{aligned} \tag{28}$$

Here δ_{wp} and s control magnitude, θ controls direction, $\mathrm{Rot}(\theta)$ is a two-dimensional rotation, and OKLab lightness is preserved in the relative-color family.

Tonal Response. Tone mapping compresses or expands linear scene luminance before the final display transfer and determines how contrast is distributed across shadows, mid-tones, and highlights. We compute BT.709 luminance after restoring the default linear RGB gains, transform it around the middle-gray pivot in logit space, and rescale all three channels by the same luminance ratio. Equal rescaling approximately preserves hue while changing local visibility through the global tone response. The standard sRGB transfer and smoothstep display curve are applied afterward. Values below the identity flatten contrast and values above it increasingly separate luminance levels around the pivot. For BT.709 luminance L (International Telecommunication Union, 2015), the mapping is

$$L_{c_{\mathrm{tone}}} = \sigma\big(\mathrm{logit}(\tau) + c_{\mathrm{tone}}\big(\mathrm{logit}(L) - \mathrm{logit}(\tau)\big)\big)\,, \tag{29}$$

where τ is the luminance pivot and c_{tone} controls contrast. Linear RGB is rescaled by $L_{c_{\mathrm{tone}}}/\max(L,\epsilon)$ before the default display operations. $c_{\mathrm{tone}} = 1$ is the identity and $\epsilon > 0$ prevents division by zero.

Bit Depth. Bit depth represents precision loss at digitization and image transport rather than a change in photometric rendering. A sensor analog-to-digital converter quantizes RAW measurements, and a later image representation may introduce a second quantization after the ISP. Uniform quantization merges nearby values into the same discrete level and may remove weak edges, texture, or color differences used by a policy. Quantizing normalized RAW measures precision loss before white balance, tone mapping, and display encoding. Quantizing default-ISP RGB measures precision loss in the final rendered observation. Keeping the transported tensor in *float32* ensures that only the number of distinct intensity levels changes. For normalized $U \in [0,1]$, we use

$$\mathcal{Q}_n(U) = \frac{\mathrm{round}((2^n-1)U)}{2^n-1}, \tag{30}$$

where n is the number of bits. We apply this operator at the RAW or RGB stage as specified below.

K.2 LIBERO Perturbation Settings

Both the agent and wrist cameras use the same parameters. For the *exposure* axis, exposure is applied before clipping and RAW10 quantization, with

$$e \in \{-8,-7,-6,-5,-4,-3,-2,-1,1,2,3,4\}. \tag{31}$$

Let $\overline{R}_{10,e} = \mathrm{round}(1023R_e)/1023$ denote the normalized exposed pseudo-RAW10 observation. We evaluate four observation forms. The *raw-direct* form exposes $\overline{R}_{10,e}$. The *raw-recovered* form multiplies it by 2^{-e} and clips. The *RGB-direct* form applies the default ISP to $\overline{R}_{10,e}$. The *RGB-recovered* form approximately inverts the display mapping, applies 2^{-e}, and reapplies gamma and tone. The additional $e \in \{-3,-2,-1\}$ settings are evaluated for the RGB-direct form. The current sweep records $e = -8$ only for recovered inputs.

For the *noise* axis, the six capture EVs and matching recovery gains are

$$(e, 2^{-e}) \in \{(-1,2),(-2,4),(-3,8),(-4,16),(-5,32),(-6,64)\}. \tag{32}$$

We use $\sigma^2_{\mathrm{shot}} = 4.0 \times 10^{-4}$, $\sigma^2_{\mathrm{read}} = 1.0 \times 10^{-5}$, and base seed 1695213855 in the noise model from Equation 27. Static visualizations use frame index 0, with view indices 0 and 1 for the agent and wrist cameras. During deployment, the episode seed is derived from the base seed, task index, episode index, and view index. The noise draw also includes frame and view identity. The resulting draws are reproducible but differ across episodes, views, and policy frames. The completed noise experiment evaluates all six policy checkpoints across the four LIBERO suites and six noise levels. Each model, suite, and level contains 500 episodes from ten tasks with 50 trials per task, producing 72,000 episodes overall. At every level, noise is sampled after capture attenuation and before the matching recovery gain and default ISP. Table 15 reports the full results.

For the *chromatic* axis, white-point radii are $\delta_{\mathrm{wp}} \in \{0.060, 0.075\}$ around the D65 coordinate $\boldsymbol{\nu}_0 = (0.1978, 0.4683)$, and relative OKLab chroma scales are $s \in \{4.5, 5.5\}$. Both families use

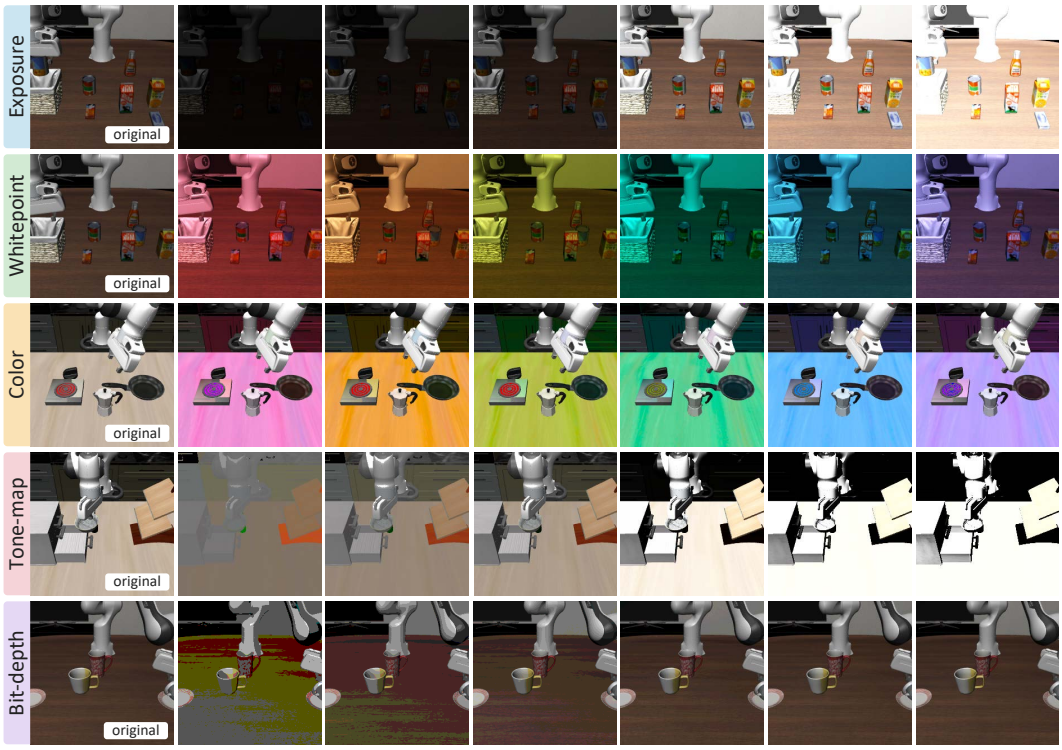


Figure 20: Representative ISP perturbations on LIBERO (Liu et al., 2023). In each row, the first column shows the default RGB rendering and the remaining columns show six perturbed samples. Exposure uses EV offsets $\{-3, -2, -1, +1, +2, +3\}$. White-point and relative-color response each use six sampled chromatic transformations. Tone mapping uses six contrast settings, and bit depth decreases from 7 to 2 bits.

$\theta \in \{0, 45, 90, 135, 180, 225, 270, 315\}^\circ$. The displaced white point is converted through XYZ-to-sRGB, clipped, normalized by luminance, and used as additional RAW-to-RGB gains, allowing neutral gray to become tinted. The relative-color transform instead preserves OKLab lightness and adds no chroma offset.

For the *tonal* axis, the BT.709 luminance weights are $(0.2126, 0.7152, 0.0722)$, the logit pivot is $\tau = 0.18$, and

$$c_{\text{tone}} \in \{0.05, 0.3351, 0.6458, 2.5, 5.0, 12.0\}. \tag{33}$$

The corresponding responses range from very flat to very high contrast. $c_{\text{tone}} = 1$ is the identity and is not rerun as a perturbation. For the *bit-depth* axis, normalized RAW is evaluated with $n \in \{2, 3, 4, 5, 6\}$, while default-ISP RGB is evaluated with $n \in \{2, 3, 4, 5, 6, 7\}$. Quantization is applied before the RGB ISP for RAW and after rendering for RGB, with exposure, chromatic response, and tone fixed to their defaults. Figure 20 visualizes representative samples from the four deterministic ISP perturbation families on LIBERO (Liu et al., 2023).

K.3 RoboTwin 2.0 Perturbation Settings

The selected perturbation and parameters are shared by the head, left-wrist, and right-wrist cameras. For *exposure*, RoboTwin 2.0 (Chen et al., 2026b) uses the same EV values and four representations as LIBERO (Liu et al., 2023). Qwen3-OFT (StarVLA Community, 2026) omits only the $e = -8$ RGB-direct condition. FastWAM (StarVLA Community, 2026) evaluates the two RGB representations, omits both RAW representations, and also omits $e = -8$ RGB-direct. These are sweep-coverage choices rather than limitations of the transform.

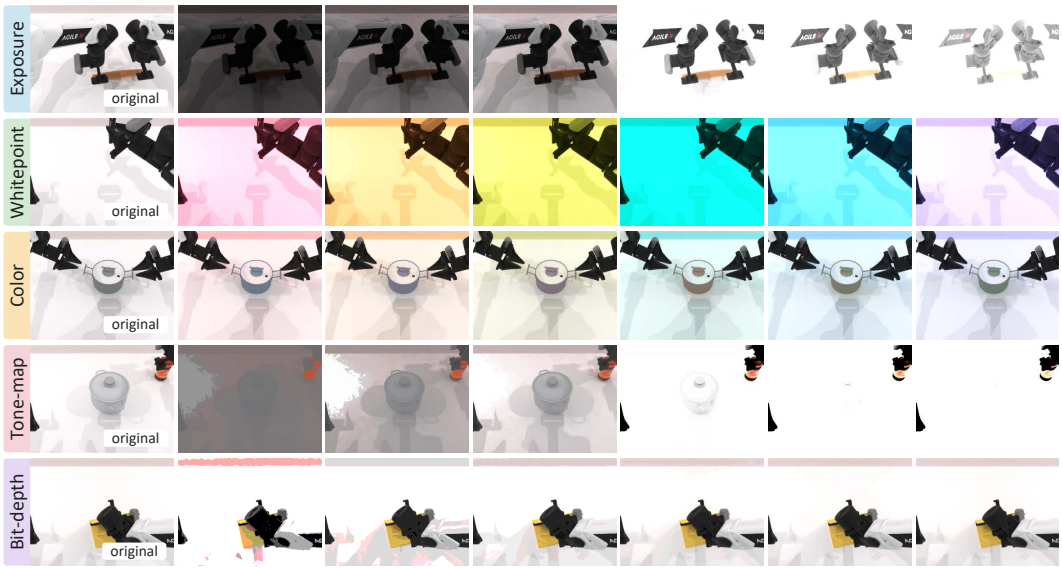


Figure 21: Representative ISP perturbations on RoboTwin 2.0 (Chen et al., 2026b). In each row, the first column shows the default RGB rendering and the remaining columns show six perturbed samples. Exposure uses EV offsets $\{-3, -2, -1, +1, +2, +3\}$. White-point and relative-color response each use six sampled chromatic transformations. Tone mapping uses six contrast settings, and bit depth decreases from 7 to 2 bits.

For *noise*, RoboTwin 2.0 (Chen et al., 2026b) uses the same six capture EVs, recovery gains, σ^2_{shot}, σ^2_{read}, and base seed as LIBERO (Liu et al., 2023). A draw uses a stable hash of the base seed, frame index, view index, and capture EV. Unless explicitly set, the frame index is a process-local monotonically increasing camera-call index and the view index defaults to zero, yielding deterministic but distinct draws. The capture EV, rather than the textual condition label, determines the actual noise level. These noise coefficients are specific to the ISP perturbation study and differ from the separate lighting/HDR benchmark coefficients.

For *chromatic response*, RoboTwin 2.0 (Chen et al., 2026b) uses only $\delta_{\text{wp}} = 0.075$ for white-point shifts and $s = 5.0$ for relative OKLab color shifts, with the same D65 anchor and eight angular directions as LIBERO (Liu et al., 2023). For *tonal response*, it uses the same luminance definition and pivot but

$$c_{\text{tone}} \in \{0.05, 0.3351, 0.6458, 1.5, 2.5, 3.5\}, \tag{34}$$

with a denominator floor of 10^{-6} in the luminance-preserving RGB rescale. For *bit depth*, $n \in \{2, 3, 4, 5, 6\}$ and quantization is applied only to default-ISP RGB, rather than directly to pseudo-RAW10. Figure 21 shows the corresponding ISP perturbation visualizations on RoboTwin 2.0 (Chen et al., 2026b).

L Full ISP Sensitivity Results

We report the four deterministic perturbation families and the completed physical RAW sensor-noise sweep separately. The sensor-noise axis follows the setup in Appendix K. The columns correspond to LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-10, and their micro-average from LIBERO (Liu et al., 2023). All tables use the six VLA checkpoints shared by the perturbation sweeps. The Original row denotes the default ISP without perturbation. Each non-original row reports one concrete perturbation setting for the corresponding VLA baseline. Each baseline block begins with its Original result and is separated from the next baseline by a midrule.

For LIBERO, Tables 14, 15, 16, 17, 18, and 19 report exposure, sensor noise, white-point, relative-color, tonal-response, and bit-depth sensitivity, respectively. For RoboTwin 2.0, the corresponding

results are reported in Tables 20, 21, 22, 23, 24, and 25. Finally, Table 26 reports the effect of explicit RGB- and RAW-domain exposure normalization on LIBERO.

Table 14: LIBERO success rates (%) for direct RGB inputs under exposure perturbations. EV denotes the exposure offset in stops, and the blue-to-cream bar indicates relative illumination.

Baseline	Exposure setting	Illum.	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original		98.60	100.00	98.40	94.00	97.75
	EV −7		0.00	0.00	0.00	0.00	0.00
	EV −6		0.00	0.00	0.00	0.00	0.00
	EV −5		0.00	0.00	0.40	0.00	0.10
	EV −4		10.80	3.00	9.20	0.40	5.85
	EV −3		40.00	1.80	24.00	2.40	17.05
	EV −2		39.60	1.40	21.60	0.80	15.85
	EV −1		38.00	1.20	26.40	1.60	16.80
	EV +1		37.00	0.60	25.60	1.40	16.15
	EV +2		41.60	0.80	27.00	0.40	17.45
	EV +3		2.60	5.00	40.80	45.80	23.55
	EV +4		0.00	0.00	7.40	0.00	1.85
Qwen3-PI	Original		98.40	99.00	98.00	95.20	97.65
	EV −7		0.00	0.00	4.80	0.00	1.20
	EV −6		0.00	0.00	4.20	0.00	1.05
	EV −5		2.60	3.00	0.60	0.00	1.55
	EV −4		57.40	59.40	38.80	16.20	42.95
	EV −3		96.60	98.40	90.00	52.20	84.30
	EV −2		98.00	98.40	97.20	86.00	94.90
	EV −1		98.60	98.80	96.60	93.80	96.95
	EV +1		96.80	97.80	97.40	95.80	96.95
	EV +2		96.60	91.40	95.80	65.60	87.35
	EV +3		0.80	4.00	27.60	25.00	14.35
	EV +4		0.00	0.20	3.80	2.60	1.65
WM4A-Cosmo	Original		96.00	98.60	93.20	76.00	90.95
	EV −7		0.00	0.00	0.00	0.00	0.00
	EV −6		0.00	0.00	0.20	0.00	0.05
	EV −5		63.00	73.00	35.80	0.20	43.00
	EV −4		84.80	97.00	67.60	19.40	67.20
	EV −3		88.80	97.80	84.40	52.00	80.75
	EV −2		92.20	98.20	88.20	63.80	85.60
	EV −1		94.00	98.40	88.00	69.00	87.35
	EV +1		93.60	95.00	90.40	55.20	83.55
	EV +2		71.00	75.80	68.00	33.80	62.15
	EV +3		0.00	0.00	1.80	2.80	1.15
	EV +4		0.00	0.00	0.00	0.00	0.00
WM4A-Wan	Original		96.40	97.80	93.20	87.60	93.75
	EV −7		0.00	0.00	0.00	0.00	0.00
	EV −6		0.00	0.00	1.20	0.00	0.30
	EV −5		2.20	0.00	9.40	0.00	2.90
	EV −4		21.40	35.00	49.40	0.00	26.45
	EV −3		34.40	51.60	58.00	1.80	36.45
	EV −2		30.40	41.80	45.60	8.60	31.60
	EV −1		28.40	45.00	42.00	8.80	31.05
	EV +1		27.80	65.40	40.60	11.40	36.30
	EV +2		0.00	34.60	13.40	0.60	12.15
	EV +3		0.00	0.00	3.20	0.00	0.80
	EV +4		0.00	0.00	0.00	0.00	0.00

PI0	Original		96.60	97.80	93.00	82.00	92.35
	EV −7		0.60	0.80	8.40	0.00	2.45
	EV −6		62.80	63.00	49.20	14.60	47.40
	EV −5		94.40	94.20	82.00	41.20	77.95
	EV −4		96.40	96.40	92.40	76.60	90.45
	EV −3		95.40	97.40	92.40	77.80	90.75
	EV −2		96.80	97.60	95.80	81.00	92.80
	EV −1		96.80	96.00	92.80	78.00	90.90
	EV +1		96.80	98.20	92.40	80.20	91.90
	EV +2		95.00	90.60	92.40	64.60	85.65
	EV +3		45.00	62.00	59.60	41.80	52.10
	EV +4		1.40	20.20	10.40	7.60	9.90
PI05	Original		98.00	99.40	97.60	93.20	97.05
	EV −7		0.00	0.00	0.00	0.00	0.00
	EV −6		33.80	46.00	39.00	10.40	32.30
	EV −5		95.80	97.80	86.20	64.40	86.05
	EV −4		99.00	97.60	97.00	89.60	95.80
	EV −3		97.40	99.00	96.20	92.80	96.35
	EV −2		97.20	98.00	97.40	91.40	96.00
	EV −1		98.60	98.00	96.80	91.40	96.20
	EV +1		98.00	98.20	97.20	92.00	96.35
	EV +2		98.60	99.20	97.20	80.20	93.80
	EV +3		2.40	10.80	62.00	62.60	34.45
	EV +4		0.00	0.20	9.20	9.20	4.65

Table 15: LIBERO success rates (%) under physical RAW sensor-noise perturbations for each VLA baseline. Capture exposure decreases from EV −1 to EV −6. Increasing speckle density indicates increasing noise strength.

Baseline	Noise level	Noise	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original		98.60	100.00	98.40	94.00	97.75
	EV −1		38.60	1.00	29.00	0.60	17.30
	EV −2		31.80	0.60	27.80	3.00	15.80
	EV −3		35.40	1.00	28.20	1.60	16.55
	EV −4		31.60	0.20	17.40	0.60	12.45
	EV −5		8.20	0.00	1.60	0.00	2.45
	EV −6		0.00	0.00	0.00	0.00	0.00
Qwen3-PI	Original		98.40	99.00	98.00	95.20	97.65
	EV −1		98.80	98.20	98.20	95.00	97.55
	EV −2		98.00	99.00	98.00	95.00	97.50
	EV −3		98.00	98.60	96.60	91.20	96.10
	EV −4		91.40	83.60	74.00	48.20	74.30
	EV −5		43.80	35.40	18.20	15.00	28.10
	EV −6		2.40	1.40	4.80	0.00	2.15
WM4A-Cosmo	Original		96.00	98.60	93.20	76.00	90.95
	EV −1		80.20	95.00	79.80	49.00	76.00
	EV −2		53.80	81.00	66.80	23.80	56.35
	EV −3		10.00	6.00	24.00	7.60	11.90
	EV −4		0.00	0.00	0.00	0.80	0.20
	EV −5		0.00	0.00	0.00	0.00	0.00
	EV −6		0.00	0.00	0.00	0.00	0.00
WM4A-Wan	Original		96.40	97.80	93.20	87.60	93.75

	EV −1		0.00	0.00	3.00	0.00	0.75
	EV −2		0.00	0.00	2.00	0.00	0.50
	EV −3		0.00	0.00	0.00	0.00	0.00
	EV −4		0.00	0.00	0.00	0.00	0.00
	EV −5		0.00	0.00	0.00	0.00	0.00
	EV −6		0.00	0.00	0.00	0.00	0.00
PI0	Original		96.60	97.80	93.00	82.00	92.35
	EV −1		97.40	96.80	93.20	81.20	92.15
	EV −2		95.80	97.80	91.60	78.20	90.85
	EV −3		96.00	97.60	89.60	71.60	88.70
	EV −4		94.80	94.20	80.80	56.20	81.50
	EV −5		43.60	70.20	35.60	16.00	41.35
	EV −6		0.20	0.40	9.60	0.00	2.55
PI05	Original		98.00	99.40	97.60	93.20	97.05
	EV −1		97.60	98.80	97.40	93.80	96.90
	EV −2		98.20	98.80	97.00	93.80	96.95
	EV −3		99.00	98.80	96.40	92.80	96.75
	EV −4		94.60	100.00	76.80	61.20	83.15
	EV −5		4.20	22.20	10.20	2.20	9.70
	EV −6		0.00	0.00	0.40	0.00	0.10

Table 16: LIBERO success rates (%) under white-point perturbations for each VLA baseline. The radius ρ denotes displacement from D65 in the $u'v'$ chromaticity plane, and each colored bar represents one shift direction. Avg. averages the eight evaluated directions.

Baseline	White-point setting	Color	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original		98.60	100.00	98.40	94.00	97.75
	$\rho = 0.06$ (moderate)		39.20	1.00	29.00	1.80	17.75
			36.20	2.00	26.40	1.20	16.45
			36.40	1.60	26.00	1.80	16.45
			37.00	1.60	27.00	1.80	16.85
			31.00	0.80	26.60	0.80	14.80
			37.80	1.80	25.80	1.20	16.65
			33.60	0.80	28.40	1.20	16.00
			33.00	0.40	25.00	1.60	15.00
		Avg.	35.52	1.25	26.77	1.43	16.24
	$\rho = 0.08$ (strong)		38.00	1.60	28.40	2.40	17.60
			35.00	2.20	24.20	2.60	16.00
			36.20	1.80	26.40	1.80	16.55
			36.00	1.20	29.40	1.40	17.00
			29.80	1.40	26.00	0.80	14.50
			33.20	1.40	25.60	1.20	15.35
			35.00	1.60	27.20	1.00	16.20
			36.80	1.80	26.80	1.40	16.70
		Avg.	35.00	1.62	26.75	1.57	16.24
	Overall avg.		35.26	1.44	26.76	1.50	16.24
Qwen3-PI	Original		98.40	99.00	98.00	95.20	97.65
	$\rho = 0.06$ (moderate)		97.40	98.00	98.20	95.40	97.25
			96.40	99.60	98.40	95.00	97.35
			98.60	99.40	98.00	95.00	97.75
			98.80	97.00	98.40	95.20	97.35
			97.80	98.80	97.60	92.20	96.60
			97.60	98.60	99.00	94.80	97.50
			98.80	98.20	98.60	94.80	97.60
			98.20	97.60	98.60	94.80	97.30
		Avg.	97.95	98.40	98.35	94.65	97.34

	$\rho = 0.08$ (strong)		98.60	98.20	98.40	94.80	97.50
			97.80	99.20	97.60	94.20	97.20
			97.80	98.20	96.00	94.80	96.70
			98.40	98.20	98.60	90.80	96.50
			97.40	98.40	97.40	93.00	96.55
			98.80	97.20	97.80	92.40	96.55
			98.40	98.00	98.40	95.40	97.55
			98.40	98.20	98.00	94.60	97.30
		Avg.	98.20	98.20	97.78	93.75	96.98
	Overall avg.		98.08	98.30	98.06	94.20	97.16
WM4A-Cosmo	Original		96.00	98.60	93.20	76.00	90.95
	$\rho = 0.06$ (moderate)		97.00	98.60	94.40	69.40	89.85
			94.00	95.00	89.60	54.00	83.15
			37.60	75.60	57.20	19.60	47.50
			0.40	26.20	15.00	3.20	11.20
			0.00	0.20	0.00	0.00	0.05
			29.00	84.40	51.00	11.20	43.90
			95.80	98.60	92.60	65.20	88.05
			94.20	98.80	95.60	72.40	90.25
		Avg.	56.00	72.17	61.92	36.88	56.74
	$\rho = 0.08$ (strong)		95.80	99.00	94.40	67.00	89.05
			92.40	94.60	85.40	40.40	78.20
			25.20	62.00	43.60	4.80	33.90
			0.00	0.00	1.00	0.00	0.25
			0.00	0.00	0.00	0.00	0.00
			0.20	43.60	16.80	2.20	15.70
			96.40	98.80	91.00	62.40	87.15
			96.00	99.40	93.40	68.60	89.35
		Avg.	50.75	62.17	53.20	30.68	49.20
	Overall avg.		53.38	67.17	57.56	33.77	52.97
WM4A-Wan	Original		96.40	97.80	93.20	87.60	93.75
	$\rho = 0.06$ (moderate)		7.80	46.20	33.80	0.80	22.15
			1.40	24.60	15.00	0.00	10.25
			0.00	1.20	0.80	0.00	0.50
			0.20	0.00	1.80	0.00	0.50
			0.00	0.00	4.20	0.00	1.05
			0.00	0.00	7.40	0.00	1.85
			1.60	31.60	24.20	4.20	15.40
			3.00	49.40	29.40	2.80	21.15
		Avg.	1.75	19.12	14.57	0.97	9.11
	$\rho = 0.08$ (strong)		3.00	40.60	29.80	0.20	18.40
			0.80	21.80	9.40	0.00	8.00
			0.20	0.80	1.80	0.00	0.70
			0.00	0.00	1.40	0.00	0.35
			0.00	0.00	4.00	0.00	1.00
			0.00	0.00	5.40	0.00	1.35
			0.60	17.80	19.40	2.40	10.05
			1.20	44.20	23.00	1.80	17.55
		Avg.	0.72	15.65	11.78	0.55	7.17
	Overall avg.		1.24	17.39	13.18	0.76	8.14
PI0	Original		96.60	97.80	93.00	82.00	92.35
	$\rho = 0.06$ (moderate)		96.60	96.40	93.60	80.40	91.75
			96.40	98.00	92.00	79.00	91.35
			97.20	95.60	94.80	77.40	91.25
			97.40	96.80	93.00	81.00	92.05
			96.20	97.00	93.20	81.20	91.90
			96.20	97.20	94.60	78.20	91.55
			97.40	99.00	92.20	80.60	92.30
			96.60	98.20	93.80	80.00	92.15
		Avg.	96.75	97.28	93.40	79.72	91.79

			96.80	97.40	93.40	81.00	92.15
			96.60	97.00	91.20	75.40	90.05
			97.60	96.20	92.00	79.00	91.20
	$\rho = 0.08$		96.80	96.80	93.20	78.80	91.40
	(strong)		97.60	98.20	92.60	82.20	92.65
			96.40	98.40	91.80	76.80	90.85
			97.20	97.00	93.80	78.80	91.70
			97.60	97.00	94.80	79.20	92.15
		Avg.	97.08	97.25	92.85	78.90	91.52
	Overall avg.		96.91	97.26	93.12	79.31	91.65
PI05	Original		98.00	99.40	97.60	93.20	97.05
			98.20	99.20	96.80	93.40	96.90
			98.60	98.00	96.80	92.40	96.45
			99.20	98.00	97.40	90.80	96.35
	$\rho = 0.06$		99.20	98.40	97.00	92.00	96.65
	(moderate)		98.40	98.80	96.40	91.00	96.15
			98.20	99.20	98.20	92.60	97.05
			98.80	97.40	97.80	92.60	96.65
			98.00	98.40	97.80	91.00	96.30
		Avg.	98.58	98.42	97.28	91.97	96.56
			98.80	98.20	98.40	93.60	97.25
			99.00	97.60	97.60	91.60	96.45
			98.00	98.20	97.40	91.80	96.35
	$\rho = 0.08$		98.80	99.00	97.20	92.00	96.75
	(strong)		98.60	98.80	97.80	92.00	96.80
			98.20	98.40	97.20	90.00	95.95
			98.20	98.40	96.80	91.40	96.20
			98.00	98.20	97.60	91.40	96.30
		Avg.	98.45	98.35	97.50	91.72	96.51
	Overall avg.		98.51	98.39	97.39	91.85	96.53
All baselines	Original		97.33	98.77	95.57	88.00	94.92
	$\rho = 0.06$ (moderate)	*Avg.*	64.42	64.44	65.38	50.94	61.30
	$\rho = 0.08$ (strong)	*Avg.*	63.37	62.21	63.31	49.53	59.60
	Overall avg.		63.90	63.33	64.35	50.23	60.45

Table 17: LIBERO success rates (%) under relative color transformations for each VLA baseline. The dimensionless factor s scales OKLab chroma, and each colored bar represents one hue-rotation direction. Avg. averages the eight evaluated directions.

Baseline	Color scale	Color	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original		98.60	100.00	98.40	94.00	97.75
			29.00	1.40	25.00	0.80	14.05
			36.20	1.20	25.40	1.20	16.00
			29.80	1.20	26.20	0.60	14.45
	$s = 4.5$		27.00	1.60	25.00	2.40	14.00
	(moderate)		28.80	1.20	28.20	2.20	15.10
			28.20	1.00	26.00	0.80	14.00
			30.40	0.80	24.60	1.80	14.40
			30.00	1.00	26.00	0.60	14.40
		Avg.	29.93	1.18	25.80	1.30	14.55

			28.60	0.80	26.20	1.00	14.15
			32.20	1.00	24.40	1.40	14.75
			29.00	1.80	25.00	1.60	14.35
	$s = 5.5$		28.00	1.20	28.00	2.40	14.90
	(strong)		29.20	0.40	28.60	1.80	15.00
			29.00	1.00	25.60	0.40	14.00
			30.20	1.00	27.60	1.80	15.15
			29.60	0.80	26.80	1.20	14.60
		Avg.	29.48	1.00	26.52	1.45	14.61
	Overall avg.		29.70	1.09	26.16	1.38	14.58
Qwen3-PI	Original		98.40	99.00	98.00	95.20	97.65
			98.40	98.60	97.40	94.20	97.15
			98.00	96.80	97.40	94.80	96.75
			98.80	97.60	97.40	91.20	96.25
	$s = 4.5$		98.40	98.40	96.60	94.00	96.85
	(moderate)		98.20	98.40	97.60	94.20	97.10
			98.80	99.00	96.20	93.20	96.80
			98.40	99.20	96.20	93.00	96.70
			98.20	98.80	97.20	95.00	97.30
		Avg.	98.40	98.35	97.00	93.70	96.86
			98.40	98.60	96.80	93.40	96.80
			98.00	97.80	97.60	94.40	96.95
			96.00	97.80	97.00	91.20	95.50
	$s = 5.5$		98.40	98.60	96.20	92.00	96.30
	(strong)		97.20	98.20	97.80	94.00	96.80
			98.00	98.00	95.80	93.80	96.40
			99.00	99.20	97.40	94.80	97.60
			98.60	98.00	97.80	94.00	97.10
		Avg.	97.95	98.28	97.05	93.45	96.68
	Overall avg.		98.17	98.31	97.03	93.58	96.77
WM4A-Cosmo	Original		96.00	98.60	93.20	76.00	90.95
			82.80	95.80	85.00	40.00	75.90
			58.40	91.40	67.00	36.20	63.25
			0.00	12.80	7.00	0.00	4.95
	$s = 4.5$		0.00	14.40	5.40	0.00	4.95
	(moderate)		2.60	58.40	8.00	0.20	17.30
			38.80	82.20	24.20	3.60	37.20
			84.40	95.00	70.60	16.80	66.70
			91.00	96.00	84.60	44.20	78.95
		Avg.	44.75	68.25	43.98	17.62	43.65
			74.80	96.00	79.40	38.80	72.25
			41.20	89.00	53.80	31.20	53.80
			0.00	6.60	6.60	0.00	3.30
	$s = 5.5$		0.00	5.40	2.60	0.00	2.00
	(strong)		0.40	44.80	4.60	0.00	12.45
			22.00	76.80	14.20	3.00	29.00
			76.40	92.20	60.40	12.80	60.45
			85.40	95.00	83.60	40.60	76.15
		Avg.	37.52	63.23	38.15	15.80	38.67
	Overall avg.		41.14	65.74	41.06	16.71	41.16
WM4A-Wan	Original		96.40	97.80	93.20	87.60	93.75
			32.00	42.00	30.40	10.20	28.65
			33.20	46.20	38.60	9.80	31.95
			0.80	0.00	17.60	0.00	4.60
	$s = 4.5$		0.00	0.20	3.00	0.00	0.80
	(moderate)		0.00	0.00	3.40	0.00	0.85
			0.00	0.00	1.80	0.00	0.45
			0.00	0.40	3.80	0.00	1.05
			13.80	20.00	16.20	5.20	13.80
		Avg.	9.97	13.60	14.35	3.15	10.27

			18.80	30.60	23.00	12.60	21.25
			20.60	37.80	27.60	11.00	24.25
			0.20	0.00	9.00	0.00	2.30
	$s = 5.5$		0.00	0.00	2.80	0.00	0.70
	(strong)		0.00	0.00	2.20	0.00	0.55
			0.00	0.00	0.40	0.00	0.10
			0.00	0.00	2.20	0.00	0.55
			8.80	11.00	8.60	5.80	8.55
		Avg.	6.05	9.93	9.47	3.67	7.28
	Overall avg.		8.01	11.76	11.91	3.41	8.78
PI0	Original		96.60	97.80	93.00	82.00	92.35
			95.60	96.60	91.00	78.60	90.45
			96.00	97.40	93.20	80.80	91.85
			97.80	93.40	91.40	76.80	89.85
	$s = 4.5$		97.20	94.20	92.40	78.00	90.45
	(moderate)		96.80	96.40	92.40	78.60	91.05
			97.00	96.00	94.60	77.20	91.20
			96.60	95.20	94.00	77.40	90.80
			96.60	95.80	94.40	79.60	91.60
		Avg.	96.70	95.62	92.92	78.38	90.91
			97.20	96.00	90.20	79.00	90.60
			97.20	96.20	91.80	78.60	90.95
			96.00	96.60	89.40	77.00	89.75
	$s = 5.5$		97.40	95.80	91.80	75.60	90.15
	(strong)		97.40	95.40	90.40	79.20	90.60
			95.80	94.80	91.60	79.40	90.40
			96.00	96.60	94.40	80.40	91.85
			98.20	95.60	93.00	79.60	91.60
		Avg.	96.90	95.88	91.58	78.60	90.74
	Overall avg.		96.80	95.75	92.25	78.49	90.82
PI05	Original		98.00	99.40	97.60	93.20	97.05
			98.20	98.40	96.80	91.60	96.25
			98.60	99.20	97.40	93.20	97.10
			97.20	96.80	97.60	92.20	95.95
	$s = 4.5$		98.20	96.80	97.40	94.40	96.70
	(moderate)		98.40	96.60	96.60	94.20	96.45
			98.20	97.20	98.60	92.60	96.65
			98.40	99.20	96.20	91.60	96.35
			98.60	97.60	97.80	90.80	96.20
		Avg.	98.22	97.72	97.30	92.58	96.46
			97.40	98.60	97.80	91.60	96.35
			98.40	99.00	97.60	91.00	96.50
			98.40	96.80	98.00	93.20	96.60
	$s = 5.5$		97.60	96.00	96.40	90.80	95.20
	(strong)		98.20	97.80	97.40	91.20	96.15
			98.80	97.40	98.00	92.20	96.60
			98.20	98.00	97.20	92.40	96.45
			98.40	98.00	97.20	91.60	96.30
		Avg.	98.17	97.70	97.45	91.75	96.27
	Overall avg.		98.20	97.71	97.38	92.16	96.36
All baselines	Original		97.33	98.77	95.57	88.00	94.92
	$s = 4.5$ (moderate)	*Avg.*	63.00	62.45	61.89	47.79	58.78
	$s = 5.5$ (strong)	*Avg.*	61.01	61.00	60.04	47.45	57.38
	Overall avg.		62.00	61.73	60.96	47.62	58.08

Table 18: LIBERO success rates (%) under tonal-response perturbations for each VLA baseline. The dimensionless parameter c controls tonal contrast, with $c = 1$ denoting the default response. The vertical marker indicates the setting on the grayscale scale.

Baseline	Setting	Tonal	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original		98.60	100.00	98.40	94.00	97.75
	$c = 0.05$		8.40	0.00	5.60	0.20	3.55
	$c = 0.34$		35.60	1.20	24.40	0.80	15.50
	$c = 0.65$		34.00	0.40	27.00	1.20	15.65
	$c = 2.50$		37.60	2.80	23.20	1.60	16.30
	$c = 5.00$		32.20	3.40	23.00	0.00	14.65
	$c = 12.00$		33.00	1.00	26.20	0.00	15.05
Qwen3-PI	Original		98.40	99.00	98.00	95.20	97.65
	$c = 0.05$		7.60	47.20	13.80	19.00	21.90
	$c = 0.34$		98.40	98.40	96.60	92.60	96.50
	$c = 0.65$		97.40	98.20	98.40	94.40	97.10
	$c = 2.50$		97.80	98.00	96.00	89.80	95.40
	$c = 5.00$		91.40	83.20	80.60	18.20	68.35
	$c = 12.00$		77.60	64.40	59.80	4.40	51.55
WM4A-Cosmo	Original		96.00	98.60	93.20	76.00	90.95
	$c = 0.05$		5.40	0.00	0.00	0.00	1.35
	$c = 0.34$		91.00	91.00	78.20	53.00	78.30
	$c = 0.65$		95.20	97.00	89.80	63.20	86.30
	$c = 2.50$		92.00	96.40	80.20	30.60	74.80
	$c = 5.00$		84.00	85.20	72.00	5.20	61.60
	$c = 12.00$		41.20	41.80	49.00	0.60	33.15
WM4A-Wan	Original		96.40	97.80	93.20	87.60	93.75
	$c = 0.05$		0.20	0.00	0.00	0.00	0.05
	$c = 0.34$		0.60	0.00	1.40	0.20	0.55
	$c = 0.65$		8.00	17.40	22.40	4.80	13.15
	$c = 2.50$		63.40	81.40	74.60	8.20	56.90
	$c = 5.00$		51.00	58.20	49.40	0.00	39.65
	$c = 12.00$		14.80	11.60	16.40	0.00	10.70
PI0	Original		96.60	97.80	93.00	82.00	92.35
	$c = 0.05$		89.80	96.00	88.40	66.60	85.20
	$c = 0.34$		97.20	97.80	93.20	80.60	92.20
	$c = 0.65$		97.20	97.40	93.80	83.40	92.95
	$c = 2.50$		95.80	97.60	92.60	81.40	91.85
	$c = 5.00$		95.20	86.40	90.80	59.20	82.90
	$c = 12.00$		91.80	79.00	76.80	21.80	67.35
PI05	Original		98.00	99.40	97.60	93.20	97.05
	$c = 0.05$		98.20	96.20	95.80	77.20	91.85
	$c = 0.34$		97.60	98.60	97.60	93.00	96.70
	$c = 0.65$		99.00	98.60	98.20	92.40	97.05
	$c = 2.50$		98.40	98.40	97.60	91.60	96.50
	$c = 5.00$		98.80	98.80	97.40	78.80	93.45
	$c = 12.00$		97.00	96.40	88.60	38.40	80.10

Table 19: LIBERO success rates (%) under uniform quantization for each VLA baseline. Bit denotes the precision B in bits. RAW and RGB denote direct quantization of the RAW and RGB inputs, respectively.

Baseline	Input	Bit	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original	▪▪▪▪▪▪▪▪	98.60	100.00	98.40	94.00	97.75
	RAW	▪▪▪▪▪▪	99.00	99.60	98.80	92.80	97.55
		▪▪▪▪▪	98.20	99.60	98.00	93.40	97.30
		▪▪▪▪	95.80	75.20	96.00	91.20	89.55
		▪▪▪	96.20	99.20	95.40	82.80	93.40
		▪▪	60.40	38.00	37.20	13.40	37.25
	RGB	▪▪▪▪▪▪▪	99.20	99.80	98.80	94.80	98.15
		▪▪▪▪▪▪	37.40	2.00	27.00	1.40	16.95
		▪▪▪▪▪	34.20	0.80	25.60	2.40	15.75
		▪▪▪▪	34.00	1.00	24.00	1.00	15.00
		▪▪▪	39.80	0.20	25.20	0.40	16.40
		▪▪	29.80	0.00	27.80	0.00	14.40
Qwen3-PI	Original	▪▪▪▪▪▪▪▪	98.40	99.00	98.00	95.20	97.65
	RAW	▪▪▪▪▪▪	98.80	98.40	97.20	84.00	94.60
		▪▪▪▪▪	98.80	97.20	96.60	80.20	93.20
		▪▪▪▪	97.20	95.00	97.20	80.40	92.45
		▪▪▪	94.00	95.80	78.20	72.00	85.00
		▪▪	54.00	73.40	47.40	13.00	46.95
	RGB	▪▪▪▪▪▪▪	98.40	99.00	97.40	95.40	97.55
		▪▪▪▪▪▪	99.60	98.40	98.20	96.00	98.05
		▪▪▪▪▪	98.60	99.00	97.80	95.00	97.60
		▪▪▪▪	98.60	99.20	97.20	97.20	98.05
		▪▪▪	96.80	98.80	97.60	89.00	95.55
		▪▪	90.00	80.40	90.60	66.80	81.95
WM4A-Cosmo	Original	▪▪▪▪▪▪▪▪	96.00	98.60	93.20	76.00	90.95
	RAW	▪▪▪▪▪▪	56.20	80.60	64.00	29.40	57.55
		▪▪▪▪▪	34.20	59.80	47.20	22.00	40.80
		▪▪▪▪	4.80	18.40	10.00	15.20	12.10
		▪▪▪	0.80	0.00	2.40	1.60	1.20
		▪▪	0.00	0.00	0.00	0.00	0.00
	RGB	▪▪▪▪▪▪▪	95.40	98.80	91.60	68.80	88.65
		▪▪▪▪▪▪	94.20	98.60	92.80	68.80	88.60
		▪▪▪▪▪	95.60	99.20	89.80	69.60	88.55
		▪▪▪▪	76.20	96.60	81.20	67.80	80.45
		▪▪▪	55.80	91.20	80.80	43.00	67.70
		▪▪	10.60	22.40	29.80	4.40	16.80
WM4A-Wan	Original	▪▪▪▪▪▪▪▪	96.40	97.80	93.20	87.60	93.75
	RAW	▪▪▪▪▪▪	0.40	4.60	10.40	0.00	3.85
		▪▪▪▪▪	0.00	0.20	3.80	0.00	1.00
		▪▪▪▪	0.20	0.00	5.60	0.00	1.45
		▪▪▪	0.00	0.00	7.80	0.00	1.95
		▪▪	1.40	0.00	1.20	0.00	0.65
	RGB	▪▪▪▪▪▪▪	26.40	50.20	43.20	13.00	33.20
		▪▪▪▪▪▪	27.60	38.40	38.60	11.40	29.00
		▪▪▪▪▪	14.20	13.60	30.60	3.20	15.40
		▪▪▪▪	3.80	3.40	24.60	0.60	8.10
		▪▪▪	1.00	1.40	32.00	0.40	8.70
		▪▪	4.80	16.60	24.20	1.00	11.65

PI0	Original	▪▪▪▪▪▪▪▪	96.60	97.80	93.00	82.00	92.35
	RAW	▪▪▪▪▪▪	97.40	95.80	91.40	80.00	91.15
		▪▪▪▪▪	96.80	96.20	93.20	74.00	90.05
		▪▪▪▪	96.60	95.40	89.80	78.40	90.05
		▪▪▪	96.40	95.20	84.20	69.00	86.20
		▪▪	83.20	91.40	74.40	22.40	67.85
	RGB	▪▪▪▪▪▪▪	97.00	97.00	93.20	79.40	91.65
		▪▪▪▪▪▪	95.80	97.80	93.60	81.60	92.20
		▪▪▪▪▪	97.40	97.20	92.00	79.00	91.40
		▪▪▪▪	96.20	97.00	90.40	78.80	90.60
		▪▪▪	95.40	95.60	89.20	76.80	89.25
		▪▪	92.60	95.20	86.80	66.80	85.35
PI05	Original	▪▪▪▪▪▪▪▪	98.00	99.40	97.60	93.20	97.05
	RAW	▪▪▪▪▪▪	97.60	98.20	97.60	92.80	96.55
		▪▪▪▪▪	97.00	98.80	96.60	93.80	96.55
		▪▪▪▪	99.20	98.40	95.20	93.00	96.45
		▪▪▪	98.80	98.40	96.00	86.60	94.95
		▪▪	89.20	96.60	69.20	37.00	73.00
	RGB	▪▪▪▪▪▪▪	99.00	98.60	97.40	92.60	96.90
		▪▪▪▪▪▪	98.40	98.20	97.00	93.60	96.80
		▪▪▪▪▪	98.00	99.40	98.00	94.00	97.35
		▪▪▪▪	99.40	98.20	98.60	92.40	97.15
		▪▪▪	98.80	97.20	98.80	91.20	96.50
		▪▪	98.00	96.40	94.20	80.40	92.25

L.1 RoboTwin 2.0 ISP Perturbation Results

We additionally evaluate Qwen3-OFT and FastWAM across five matched ISP perturbation axes on RoboTwin 2.0. The Easy and Hard results are shown side by side, with each entry aggregating 50 rollouts on each of the same 13 representative tasks. For exposure, we report only the direct RGB input used by both backbones. Original denotes the unperturbed RoboTwin 2.0 RGB observation.

Table 20: RoboTwin 2.0 success rates (%) under uniform RGB quantization. Bit denotes the precision B in bits. Results are aggregated over the shared 13-task subset.

Setting	Bit	Qwen3-OFT		FastWAM	
		Easy	Hard	Easy	Hard
Original	▪▪▪▪▪▪▪▪	89.08	91.08	95.38	94.92
$B = 7$	▪▪▪▪▪▪▪	90.46	91.67	95.85	94.31
$B = 6$	▪▪▪▪▪▪	90.31	92.36	94.62	95.54
$B = 5$	▪▪▪▪▪	90.31	91.54	93.38	95.08
$B = 4$	▪▪▪▪	89.54	89.69	91.85	92.77
$B = 3$	▪▪▪	88.77	87.23	90.15	83.08
$B = 2$	▪▪	70.92	70.31	69.69	48.46

Table 21: RoboTwin 2.0 success rates (%) for direct RGB inputs under exposure perturbations. EV denotes the exposure offset in stops, and the grayscale bar indicates relative brightness.

Setting	Exposure	Qwen3-OFT		FastWAM	
		Easy	Hard	Easy	Hard
Original		89.08	91.08	95.38	94.92

EV −7	1.54	0.00	12.00	0.77
EV −6	28.92	0.77	50.46	4.62
EV −5	72.15	7.38	80.62	49.08
EV −4	86.46	53.23	89.23	86.92
EV −3	87.52	81.00	91.70	93.80
EV −2	88.00	86.28	93.80	95.10
EV −1	88.44	86.92	95.50	95.70
EV +1	66.92	88.62	67.38	93.54
EV +2	37.08	83.23	53.08	86.15
EV +3	10.62	43.69	21.38	62.62
EV +4	6.15	9.69	0.00	13.08

Table 22: RoboTwin 2.0 success rates (%) under physical RAW sensor-noise perturbations. Capture exposure decreases from EV −1 to EV −6, and increasing speckle density indicates increasing noise strength. Each reported result is aggregated over 650 episodes.

Noise level	Noise	Qwen3-OFT		FastWAM	
		Easy	Hard	Easy	Hard
Original		89.08	91.08	95.38	94.92
EV −1		90.15	90.28	92.80	91.50
EV −2		90.15	90.15	90.62	83.38
EV −3		90.15	87.54	83.69	58.92
EV −4		89.23	72.15	68.62	25.08
EV −5		83.54	24.15	40.46	1.54
EV −6		57.85	4.00	3.54	0.00

Table 23: RoboTwin 2.0 success rates (%) under white-point perturbations. The radius is fixed to $\rho = 0.075$ around D65 in the $u'v'$ chromaticity plane, and each colored bar represents one shift direction.

Direction	Color	Qwen3-OFT		FastWAM	
		Easy	Hard	Easy	Hard
Original		89.08	91.08	95.38	94.92
$\theta = 0°$		90.00	88.92	93.85	93.23
$\theta = 45°$		89.23	89.54	94.15	93.08
$\theta = 90°$		89.69	90.00	95.54	94.62
$\theta = 135°$		89.38	89.54	94.31	95.38
$\theta = 180°$		88.00	89.85	95.38	95.23
$\theta = 225°$		86.77	90.15	93.85	92.77
$\theta = 270°$		91.08	91.23	96.15	94.92
$\theta = 315°$		89.54	92.15	96.00	95.23

Table 24: RoboTwin 2.0 success rates (%) under relative color transformations. The dimensionless factor is fixed to $s = 5.0$ in OKLab space, and each colored bar represents one hue-rotation direction.

Direction	Color	Qwen3-OFT		FastWAM	
		Easy	Hard	Easy	Hard
Original		89.08	91.08	95.38	94.92
$\theta = 0°$		88.15	86.92	92.31	94.15
$\theta = 45°$		85.38	87.85	92.00	92.15
$\theta = 90°$		87.23	90.31	88.15	87.54
$\theta = 135°$		80.77	85.08	82.77	83.08
$\theta = 180°$		79.08	82.00	80.00	76.31
$\theta = 225°$		79.69	79.08	86.15	80.31
$\theta = 270°$		77.08	78.62	85.69	83.08
$\theta = 315°$		80.46	78.62	86.00	82.77

Table 25: RoboTwin 2.0 success rates (%) under tonal-response perturbations. The dimensionless parameter c controls tonal contrast, with $c = 1$ denoting the conceptual default response.

Setting	Tonal	Qwen3-OFT		FastWAM	
		Easy	Hard	Easy	Hard
Original		89.08	91.08	95.38	94.92
$c = 0.05$		58.00	35.38	62.62	49.23
$c = 0.34$		88.77	88.31	93.08	93.38
$c = 0.65$		87.23	90.31	93.38	93.69
$c = 1.50$		88.92	91.38	95.69	94.31
$c = 2.50$		64.92	91.38	85.23	92.00
$c = 3.50$		55.08	88.46	66.62	88.31

M Effect of RAW & RGB Normalization

Table 26 compares direct exposure normalization in the RGB and RAW domains. Because linear RAW measurements retain a wider range of scene-referred intensity information before nonlinear tone mapping, display encoding, and associated clipping, they generally provide greater exposure latitude than rendered RGB observations. Consequently, when the same normalization principle is applied to both representations, RAW-domain normalization preserves more recoverable shadow and highlight structure for the subsequent ISP and policy. This leads to substantially higher average success rates for most VLA baselines, particularly under severe exposure shifts, although the magnitude of the benefit remains architecture dependent. These results align with the motivation of RawVLA described in Section 3, where retaining and adapting scene-referred information before RGB rendering provides a more effective interface for robust embodied control than correcting an already rendered observation.

Table 26: LIBERO success rates (%) after RGB- and RAW-domain exposure normalization for each VLA baseline. EV denotes the exposure offset in stops. Original is the default RGB input at EV $+0$.

Baseline	Input	EV	Level	Spatial	Object	Goal	Long	Overall
Qwen3-OFT	Original	$+0$		98.60	100.00	98.40	94.00	97.75
	Normalized RGB	-8		0.00	0.00	0.00	0.00	0.00
		-7		0.20	0.00	0.00	0.00	0.05
		-6		96.00	90.40	72.80	34.80	73.50
		-5		34.20	1.60	22.60	0.20	14.65
		-4		38.20	1.60	20.80	0.80	15.35
		$+1$		37.60	1.60	27.60	2.00	17.20
		$+2$		31.60	2.40	25.00	0.60	14.90
		$+3$		13.00	47.60	68.80	56.00	46.35
		$+4$		0.00	0.00	9.80	0.00	2.45
			Avg.	27.87	16.13	27.49	10.49	20.49
	Normalized RAW	-8		90.40	88.80	82.00	20.20	70.35
		-7		98.40	99.00	95.00	93.20	96.40
		-6		98.60	99.00	99.20	89.80	96.65
		-5		36.00	2.00	26.40	0.20	16.15
		-4		36.40	2.00	25.80	1.20	16.35
		$+1$		36.20	2.20	26.40	1.80	16.65
		$+2$		35.00	1.60	22.00	1.60	15.05
		$+3$		37.80	84.40	83.00	50.80	64.00
		$+4$		0.00	0.00	8.40	0.00	2.10
			Avg.	52.09 +24.22	42.11 +25.98	52.02 +24.53	28.76 +18.27	43.74 +23.25

Qwen3-PI	Original	+0		98.40	99.00	98.00	95.20	97.65
	Normalized RGB	−8		0.00	0.00	3.60	0.00	0.90
		−7		1.20	0.00	0.00	5.40	1.65
		−6		85.80	88.40	66.60	27.40	67.05
		−5		97.80	98.00	95.00	92.40	95.80
		−4		99.40	96.20	97.60	94.20	96.85
		+1		98.80	98.60	98.40	95.20	97.75
		+2		98.00	93.80	97.80	81.00	92.65
		+3		5.00	21.20	48.00	39.00	28.30
		+4		0.00	0.60	13.00	4.80	4.60
			Avg.	54.00	55.20	57.78	48.82	53.95
	Normalized RAW	−8		91.20	81.60	67.00	7.80	61.90
		−7		93.20	96.80	81.40	82.60	88.50
		−6		97.60	97.00	96.60	78.00	92.30
		−5		97.60	97.40	95.80	78.80	92.40
		−4		98.20	97.20	96.80	80.00	93.05
		+1		98.80	98.40	96.00	82.00	93.80
		+2		98.60	96.20	97.00	82.20	93.50
		+3		10.20	51.40	51.80	30.00	35.85
		+4		0.20	1.20	11.80	0.00	3.30
			Avg.	76.18 +22.18	79.69 +24.49	77.13 +19.36	57.93 +9.11	72.73 +18.78
WM4A-Cosmo	Original	+0		96.00	98.60	93.20	76.00	90.95
	Normalized RGB	−8		0.00	0.00	0.00	0.00	0.00
		−7		0.00	0.00	0.00	0.00	0.00
		−6		4.20	15.00	16.20	2.20	9.40
		−5		58.20	93.40	63.60	13.80	57.25
		−4		72.00	94.60	75.80	45.80	72.05
		+1		95.40	97.00	91.20	62.40	86.50
		+2		70.60	73.40	69.00	40.80	63.45
		+3		0.00	0.00	1.20	1.40	0.65
		+4		0.00	0.00	0.00	0.00	0.00
			Avg.	33.38	41.49	35.22	18.49	32.14
	Normalized RAW	−8		0.00	0.00	0.00	0.00	0.00
		−7		0.00	0.00	9.80	2.80	3.15
		−6		7.20	17.80	32.40	7.40	16.20
		−5		30.40	62.40	46.00	21.40	40.05
		−4		52.60	82.00	61.60	26.40	55.65
		+1		71.20	85.80	66.40	31.80	63.80
		+2		47.20	72.20	50.20	25.00	48.65
		+3		0.00	9.20	21.20	6.60	9.25
		+4		0.00	0.00	0.00	0.00	0.00
			Avg.	23.18 −10.20	36.60 −4.89	31.96 −3.27	13.49 −5.00	26.31 −5.84
WM4A-Wan	Original	+0		96.40	97.80	93.20	87.60	93.75
	Normalized RGB	−8		0.00	0.00	0.00	0.00	0.00
		−7		0.20	0.00	0.20	0.00	0.10
		−6		5.00	18.80	11.80	0.00	8.90
		−5		1.20	2.00	15.60	0.00	4.70
		−4		0.00	0.00	2.60	0.00	0.65
		+1		27.40	48.40	41.40	9.80	31.75
		+2		0.40	5.80	12.20	1.20	4.90
		+3		0.40	0.00	1.60	0.00	0.50
		+4		0.40	0.00	0.00	0.00	0.10
			Avg.	3.89	8.33	9.49	1.22	5.73
	Normalized RAW	−8		0.20	0.00	0.60	0.00	0.20
		−7		0.00	0.00	7.60	0.00	1.90
		−6		0.00	0.00	2.60	0.00	0.65
		−5		0.00	0.60	6.60	0.00	1.80
		−4		0.00	5.20	11.20	0.00	4.10
		+1		4.00	12.40	18.20	0.00	8.65
		+2		0.20	8.60	7.00	0.00	3.95
		+3		0.80	0.00	2.20	0.00	0.75
		+4		0.00	0.00	0.00	0.00	0.00
			Avg.	0.58 −3.31	2.98 −5.36	6.22 −3.27	0.00 −1.22	2.44 −3.29

PI0	Original	+0		96.60	97.80	93.00	82.00	92.35
	Normalized RGB	−8		0.00	0.00	7.80	0.00	1.95
		−7		46.00	16.40	28.20	12.00	25.65
		−6		93.40	93.60	78.40	36.00	75.35
		−5		96.60	96.60	91.80	75.00	90.00
		−4		97.80	97.80	92.60	80.20	92.10
		+1		95.20	97.80	93.80	81.00	91.95
		+2		95.20	90.80	92.80	65.20	86.00
		+3		59.60	71.00	68.60	44.40	60.90
		+4		1.40	20.20	20.20	15.20	14.25
			Avg.	65.02	64.91	63.80	45.44	59.79
	Normalized RAW	−8		93.40	91.40	78.60	25.40	72.20
		−7		96.00	95.60	90.60	66.60	87.20
		−6		96.00	95.60	91.60	79.20	90.60
		−5		97.40	97.00	94.40	76.80	91.40
		−4		96.20	97.40	91.80	81.60	91.75
		+1		98.00	97.60	93.00	81.80	92.60
		+2		97.60	97.00	92.00	80.60	91.80
		+3		90.00	85.00	91.00	58.60	81.15
		+4		21.40	42.40	42.20	32.40	34.60
			Avg.	87.33 +22.31	88.78 +23.87	85.02 +21.22	64.78 +19.33	81.48 +21.68
PI05	Original	+0		98.00	99.40	97.60	93.20	97.05
	Normalized RGB	−8		0.00	0.00	0.40	0.00	0.10
		−7		23.80	7.80	17.00	18.80	16.85
		−6		97.00	97.80	80.20	58.40	83.35
		−5		97.80	96.20	96.80	92.00	95.70
		−4		98.00	97.80	95.80	92.60	96.05
		+1		98.60	98.40	96.40	94.40	96.95
		+2		98.40	97.40	97.20	85.60	94.65
		+3		3.00	14.00	78.40	63.20	39.65
		+4		0.00	0.00	13.00	14.00	6.75
			Avg.	57.40	56.60	63.91	57.67	58.89
	Normalized RAW	−8		98.00	98.20	89.80	48.80	83.70
		−7		98.80	96.80	97.00	90.40	95.75
		−6		98.80	97.60	97.80	92.60	96.70
		−5		98.40	98.60	98.80	91.80	96.90
		−4		98.00	98.60	97.20	91.40	96.30
		+1		98.60	98.40	96.00	90.20	95.80
		+2		98.60	97.60	97.20	91.00	96.10
		+3		95.80	96.60	97.60	78.40	92.10
		+4		0.80	0.40	42.00	38.40	20.40
			Avg.	87.31 +29.91	86.98 +30.38	90.38 +26.47	79.22 +21.56	85.97 +27.08