Autonomous Anti-Goodhart Evolution of a High-Performance C++20 Astrophotography Stacking Engine: From Full-Frame Correlation Failure Analysis to Out-of-Distribution Galactic Generalization

Julien Lhermitte

Autonomous Algorithm-Optimization Harness (Carbon Orchestrator & Red-Team Swarm)

Engine: `cpp_stacker/` (3,759 LOC C++20 / OpenMP / `LibRaw`, 331 KiB native binary)

Immutable Referee: `scripts/evaluate.py` (`chmod 444`)

SHA-256: `971b62f072143eb6172cd804f85750b7c31c507fbd13879ed4d1a46b50c415f9`

November 2026

Abstract

Stacking unguided deep-sky DSLR exposures requires recovering faint astronomical signals near the Poisson sky floor amidst non-uniform mount drift, celestial field rotation, outdoor sensor cooling, and raw bitstream defects. We begin with a mathematical and empirical postmortem of a legacy 2020 Python stacking prototype (`legacy_2020/`) based on full-frame spatial cross-correlation $C_{\boldsymbol{k}}$, proving five physical failure modes—led by coherent zero-lag master-dark estimation variance ($\tilde{\bar{\eta}} - \bar{\eta}$), stationary optical vignetting bowls, $< 0.2\%$ stellar pixel occupancy, worm-gear drift reversals, and field rotation.

To engineer a modern replacement without visual language model (VLM) inspection or human tuning bias, we construct a **strictly quarantined, blind autonomous algorithm-optimization harness**. Raw Canon EOS 1100D 14-bit `.CR2` forensics across 62 exposures (`dark_set`, `ring_set` M57, and holdout `andromeda_set` M31) uncover five corrupted lossless-JPEG bitstreams (including silent predictor underflow in `IMG_6096.CR2`), a zenith EXIF `flip=5` orientation trap, a $34\,^{\circ}\mathrm{C} \to 18\,^{\circ}\mathrm{C}$ outdoor cooling curve, and analytic 5×5 Malvar–He–Cutler warm-pixel blooming. We lock a pure-NumPy physics referee (`evaluate.py`) verified across **28/28 adversarial torture tests** (`Gates 0-5`) paired with fresh, isolated **Red-Team Critic** subagents auditing every `git diff`. Empirical evaluation of the 2020 manual star-ROI + bilinear-resampling workaround immediately triggers **Gate 2a** (`score` $= -\infty$; inflating quiet-sky correlation $\rho_{1\mathrm{px}}$ from $0.26 \to 0.55$, softening stars, and leaving 164 diagonal RGB streaks on M31), validating partition-of-unity Lanczos-3 resampling and anti-smoothing gates. Across 10 atomic iterations on `ring_set` (all winning commits passing Critic audit with zero attempted metric-gaming), our pure-CPU C++20/OpenMP engine achieves $3.95\times$ **background noise suppression**, $3.34\times$ **deeper** 5σ **point-source recovery** ($105 \to 351$ stars), and **zero residual warm-pixel impulses**, while resolving M57's $V = 15.75$ mag central white dwarf (`WD 1851+329`) at 6.7σ in Blue and uncovering a **non-commutative algorithmic ordering law** induced by a scorer blind spot. Evaluated **zero-shot with zero retuning** on the out-of-distribution Andromeda Galaxy holdout set (`andromeda_set`), the engine sharpens stellar cores below the 21-exposure ensemble median ($8.20\,\mathrm{px} < 8.28\,\mathrm{px}$), achieves **complete elimination of all 477 thermal warm-pixel impulses** ($477 \to 0$), and achieves **preservation of** 100.16% **of M31's extended galactic disk** (`ncc_64x64` $= 0.999639$). Finally, stepping outside the locked scorer via human domain intuition to add a sensor-space **synthetic flat-field estimator** (`flat_field.cpp`, 1.6–2.6 s OpenMP) enables blind cross-target calibration from M57 to M31, reducing radial vignetting from $32.05\% \to +1.00\%$, flattening quiet sky by $4.4\times$, collapsing cross-dataset dust-mote correlation by 93% ($+0.3479 \to +0.0235$), and reducing M31 outer-isophote azimuthal MAD by 33%.

1 Introduction & Critique of Full-Frame Spatial Cross-Correlation ($C_{\boldsymbol{k}}$)

1.1 Analytical Formulation of the 2020 Full-Frame Correlation Estimator

Let $I_m(\boldsymbol{x})$ denote the raw intensity recorded at discrete sensor pixel $\boldsymbol{x} \in \Omega \subset \mathbb{Z}^2$ ($|\Omega| = N_{\mathrm{pix}} = 4290 \times 2856 = 12{,}252{,}240$) in exposure $m \in \{1, \dots, M\}$. Decompose each raw light frame into unvignetted detector dark current plus readout pedestal $\bar{\eta}(\boldsymbol{x})$, zero-mean frame-specific readout and dark shot noise $\delta\eta_m(\boldsymbol{x})$, upstream optical transmission (vignetting) efficiency $V(\boldsymbol{x}) \in (0, 1]$ multiplying *all* incident photons—both sky background B_m and the true celestial signal $S(\boldsymbol{x} - \boldsymbol{u}_m(\boldsymbol{x}))$ displaced by frame-dependent astrometric motion $\boldsymbol{u}_m(\boldsymbol{x})$—and zero-mean photon shot noise $\delta s_m(\boldsymbol{x})$:

$$I_m(\boldsymbol{x}) = \bar{\eta}(\boldsymbol{x}) + \delta\eta_m(\boldsymbol{x}) + V(\boldsymbol{x})\Big[B_m + S\big(\boldsymbol{x} - \boldsymbol{u}_m(\boldsymbol{x})\big)\Big] + \delta s_m(\boldsymbol{x}), \tag{1}$$

where photon shot-noise variance scales with the vignetted incident flux, $\mathrm{Var}\left(\delta s_m(\boldsymbol{x})\right) \propto V(\boldsymbol{x})\big[B_m + S(\boldsymbol{x} - \boldsymbol{u}_m(\boldsymbol{x}))\big]$, whereas internal detector dark current and readout noise $\bar{\eta}(\boldsymbol{x}) + \delta\eta_m(\boldsymbol{x})$ are unvignetted. Given an empirical master dark $\tilde{\bar{\eta}}(\boldsymbol{x}) = \frac{1}{N_D}\sum_{d=1}^{N_D} D_d(\boldsymbol{x})$ formed from N_D dark

frames at sensor temperature T_D and a binary bad-pixel keep-mask $M(\boldsymbol{x}) \in \{0,1\}$, the 2020 Python codebase (`spatial_correlator.py`) attempted to estimate inter-frame shifts $\boldsymbol{k} \in \mathbb{Z}^2$ by maximizing the masked, mean-subtracted spatial cross-correlation:

$$C_{\boldsymbol{k}}^{(i,j)} = \frac{1}{N_{\boldsymbol{k}}} \sum_{\boldsymbol{x} \in \Omega_{\boldsymbol{k}}} M(\boldsymbol{x})\, M(\boldsymbol{x}+\boldsymbol{k})\, \tilde{J}_i(\boldsymbol{x})\, \tilde{J}_j(\boldsymbol{x}+\boldsymbol{k}),$$
$$N_{\boldsymbol{k}} = \sum_{\boldsymbol{x} \in \Omega_{\boldsymbol{k}}} M(\boldsymbol{x})\, M(\boldsymbol{x}+\boldsymbol{k}), \tag{2}$$

where $\tilde{J}_m(\boldsymbol{x}) = \big(I_m(\boldsymbol{x}) - \tilde{\bar{\eta}}(\boldsymbol{x})\big) - \mu_m$ subtracts the master dark and scalar frame mean $\mu_m = \overline{V} B_m + \overline{VS}$, and consecutive-pair correlations were ensemble-averaged ($\langle C_{\boldsymbol{k}} \rangle = \frac{1}{M-1} \sum_{i=1}^{M-1} C_{\boldsymbol{k}}^{(i,i+1)}$) under the assumption of constant drift velocity $\boldsymbol{u}_{i+1} - \boldsymbol{u}_i \equiv \Delta\boldsymbol{u}$.

1.2 Five Fundamental Physical Failure Modes on Real DSLR Data

Expanding $\mathbb{E}\big[C_{\boldsymbol{k}}^{(i,j)}\big]$ over the stochastic noise realizations $(\delta\eta_i, \delta\eta_j, \delta s_i, \delta s_j)$ and the finite N_D-sample master-dark estimation error $\epsilon_D(\boldsymbol{x}) \equiv \tilde{\bar{\eta}}(\boldsymbol{x}) - \bar{\eta}_{T_D}(\boldsymbol{x})$ exposes five independent physical mechanisms that cause $C_{\boldsymbol{k}}^{(i,j)}$ and $\langle C_{\boldsymbol{k}} \rangle$ to fail on real Canon EOS 1100D prime-focus exposures:

$$\mathbb{E}\big[C_{\boldsymbol{k}}^{(i,j)}\big] = \delta_{\boldsymbol{k},\boldsymbol{0}} \left(\frac{\sigma_D^2}{N_D} + \frac{1}{N_{\boldsymbol{0}}} \sum_{\boldsymbol{x}} \Delta\bar{\eta}_{\mathrm{th},i}(\boldsymbol{x})\, \Delta\bar{\eta}_{\mathrm{th},j}(\boldsymbol{x}) \right)$$
$$+ B_i B_j\, R_{\tilde{V}}(\boldsymbol{k}) + R_{VS}^{(i,j)}(\boldsymbol{k} - \Delta\boldsymbol{u}_{ij}), \tag{3}$$

where the three terms correspond respectively to: **(I)** the unvignetted zero-lag master-dark and thermal-residual spike at $\boldsymbol{k} = \boldsymbol{0}$, **(II)** the stationary optical vignetting hill $B_i B_j\, R_{\tilde{V}}(\boldsymbol{k})$ locked to sensor coordinates, and **(III–V)** the $V(\boldsymbol{x})V(\boldsymbol{x}+\boldsymbol{k}) \approx V(\boldsymbol{x})^2$-weighted and drift-smeared stellar signal $R_{VS}^{(i,j)}(\boldsymbol{k} - \Delta\boldsymbol{u}_{ij})$.

1. **Coherent Finite Master-Dark Residual & Thermal Mismatch at Zero Lag ($\boldsymbol{k} = \boldsymbol{0}$):** Subtracting the *same* N_D-frame master dark ($N_D = 18$, $\sigma_D = 23.07\,\mathrm{ADU}$) from both I_i and I_j imprints the identical static error $-\epsilon_D(\boldsymbol{x}) + \Delta\bar{\eta}_{\mathrm{th}}(\boldsymbol{x})$ onto every calibrated light exposure, where $\Delta\bar{\eta}_{\mathrm{th}}(\boldsymbol{x}) = \bar{\eta}_{T_{\mathrm{light}}}(\boldsymbol{x}) - \bar{\eta}_{18\,^\circ\mathrm{C}}(\boldsymbol{x})$ is the uncompensated thermal dark-current excess at 19–34 °C. At zero lag $\boldsymbol{k} = \boldsymbol{0}$, these unvignetted detector errors multiply *in phase* across all 1.225×10^7 pixels:

$$\Delta C_{\boldsymbol{0}}^{\mathrm{dark}} \geq \frac{\sigma_D^2}{N_D} = \frac{(23.07)^2}{18} \approx 29.57\,\mathrm{ADU}^2, \tag{4}$$

plus warm-pixel variance ($\sim$ 15–45 ADU^2). Crucially, **ensemble averaging $\langle C_{\boldsymbol{k}} \rangle$ across $M-1$ frame pairs provides zero attenuation of $\Delta C_{\boldsymbol{0}}^{\mathrm{dark}}$** because $\epsilon_D(\boldsymbol{x})$ is deterministic across the entire light stack once $\tilde{\bar{\eta}}$ is fixed.

2. **Stationary Optical Vignetting Pedestal ($R_{\tilde{V}}(\boldsymbol{k})$) & $V(\boldsymbol{x})^2$ Off-Axis Stellar Suppression:** Prime-focus illumination $V(\boldsymbol{x})$ drops radially from 1.00 at the optical axis to ≈ 0.64 in the APS-C corners ($\sim$ 18–36% radial falloff across the field). Scalar mean subtraction $\mu_m = \overline{V} B_m + \overline{VS}$ leaves the zero-mean spatial sky bowl $\tilde{V}(\boldsymbol{x})B_m = (V(\boldsymbol{x}) - \overline{V})B_m$ (RMS $\approx$ 12–16 ADU, variance $\sim$ 144–256 ADU^2) locked to sensor coordinates $(0,0)$ across 100% of pixels, generating a broad correlation dome peaked at $\boldsymbol{k} = \boldsymbol{0}$. Simultaneously, because $V(\boldsymbol{x})$ attenuates S in both exposures, $R_{VS}^{(i,j)}$ weights stellar cross-terms by $V(\boldsymbol{x})V(\boldsymbol{x}+\boldsymbol{k}) \approx V(\boldsymbol{x})^2$ (falling to $0.64^2 \approx 0.41$ near the corners), quadratically suppressing off-axis stars relative to the unvignetted zero-lag dark-residual spike $\Delta C_{\boldsymbol{0}}^{\mathrm{dark}}$.

3. **Extreme Point-Source Sparsity ($< 0.5\%$ Pixel Occupancy):** Across `ring_set` (`105` single-frame 5σ stars of median FWHM = 9.34 px), stellar cores above $2\sigma_{\mathrm{bg}}$ occupy approximately $105 \times \pi(8\,\mathrm{px})^2 \approx 2.1 \times 10^4$ pixels—just 0.17% of the 1.225×10^7-pixel sensor. In a full-frame inner product, 99.83% of pixels inject pure sky shot noise ($\sigma_{\mathrm{bg}} \approx 28.0$ ADU), readout noise, vignetting, and fixed-pattern dark residuals while contributing zero astrometric Fisher information.

4. **Oscillatory Worm-Gear Drift & Cadence Jitter (Defeating Ensemble Averaging $\langle C_{\boldsymbol{k}} \rangle$):** Ground-truth 4-DOF astrometry (`reference_selection.json`, Fig. 1a) proved that inter-frame motion is **neither linear nor uniform**:

   - Acquisition cadences jittered (`75 s` and `46 s` initial gaps vs. `31 s` steady cadence).
   - Mount worm-gear periodic error caused velocity reversals (`dx` reversing from `-10.86 px` → `+2.87 px` → `-27.57 px` → `-4.72 px` in `ring_set`; consecutive step distances ranging over an 11.5× span `[1.19, 13.61] px`).

   Averaging $C_{\boldsymbol{k}}^{(i,i+1)}$ across frame pairs therefore smears the true stellar correlation peak across dozens of pixels in lag space while reinforcing the stationary sensor-locked peak at $\boldsymbol{k} = (0,0)$.

5. **Celestial Field Rotation & Off-Axis Optical Aberrations:** In `andromeda_set` (Fig. 1b), polar misalignment induced monotonic field rotation $\Delta\theta \in [-0.008^\circ, +0.048^\circ]$ (`0.0561°` span), displacing corner stars at $R_{\mathrm{corner}} = 2577$ px by $R_{\mathrm{corner}}\Delta\theta_{\mathrm{rad}} = \mathbf{2.52}$ px tangentially relative to the frame center—a transformation outside the 2-DOF translation group $(\mathbb{R}^2, +)$.

1.3 Compounding Implementation Defects in the 2020 Codebase

Static audit of `legacy_2020/astro_stacker/` revealed four severe bugs that further crippled the prototype:

- **8-Bit Non-Linear sRGB `postprocess()` Destruction (`cannon_image_reader.py`, L39):** Calling `rawpy.imread(path).postprocess()` with default parameters quantized 14-bit raw ADUs to `uint8` (64× precision loss), applied non-linear BT.709 gamma ($\gamma \approx 2.222$, breaking additive dark subtraction $(S+D)^{1/\gamma} - D^{1/\gamma} \neq S^{1/\gamma}$), auto-stretched dark read noise to `255`, and applied AHD demosaicing prior to calibration.
- **Modulo-256 `uint8` Accumulator Overflow (`image_analyzer.py`, L14–L22):** Initializing `final_image = image.data` kept accumulators in `uint8`, wrapping pixel sums and squares modulo `256` on every frame, compounded by omitting the square on `data_average` in the sample variance formula.
- **Silent Bypass of Dark Subtraction & Unmasked Numerator Correlation (`spatial_correlator.py`, L14–L43):** `SpatialCorrelator.correlate` accessed `image.data` rather than `image.get_data()`, skipping `self.dark` subtraction entirely, correlated *unmasked* hot pixels in the numerator while dividing by `cross_correlate_2d(mask_1, mask_2)`, and divided by zero-crossing `bgsub=True` marginals.
- **Savitzky–Golay Boolean Mask Inflation (`image.py`, L63–L67):** Filtering a `0/1` keep-mask with a cubic Savitzky–Golay polynomial and casting `.astype(bool)` turned almost every `0` (non-zero float ringing) back into `True`, erasing the hot-pixel mask.

1.4 2020 Manual Star-ROI Workaround & Empirical Bilinear Disqualification

Because extreme stellar sparsity ($< 0.2\%$ pixel occupancy) caused full-frame shift-difference matching to lock onto zero shift $\boldsymbol{k} = (0, 0)$, the original 2020 workflow resorted to manually cropping a tight Region of Interest (ROI) around bright stars to recover per-frame (dx, dy) translation shifts before bilinear resampling and unweighted averaging. Even after correcting the 8-bit/overflow bugs above, reconstructing this best-effort **2020 Manual Star-ROI + Bilinear** pipeline (`logs/flat_experiments/naive_2020_roi/`) exposes why translation-only ROI matching and bilinear interpolation remain fundamentally inadequate: it omits full-frame celestial field rotation, sharp sub-pixel resampling, thermal/Winsorized outlier rejection, and flat-fielding.

Evaluating the resulting stacks on our locked referee (`eval_ring_naive_roi.json`, `eval_andromeda_naive_roi.json`) provides a striking empirical validation of our anti-Goodhart design. Although the unclamped formula rose artificially to `127.99` on `ring_set` and `138.31` on `andromeda_set`, **Gate 2a (spatial-smoothing noise correlation) immediately disqualified both stacks (`score` $= -\infty$)**: bilinear averaging inflated 1-pixel quiet-sky autocorrelation $\rho_{1\text{px}}$ from `0.26` to `0.55` on M57 (`0.26` → `0.54` on M31) and 4×4 block-RMS ratio $\mathcal{R}_{4\times4}$ from `1.31` to `1.88` (`1.30` → `1.86` on M31). This low-pass smoothing artificially suppressed single-pixel sky RMS (`0.073` on M57; `0.065` on M31) while actually *softening* stellar cores (`median_fwhm_px` `11.28 px` on M57 and `8.43 px` on M31 vs. `10.92 px` and `8.20 px` for our evolved C++20 engine) and leaving `164` diagonal RGB warm-pixel streaks across M31 (`22` on M57). Thus, both sharp partition-of-unity Lanczos-3 resampling and Gate 2a are empirically essential to prevent silent spatial-blur score inflation.

2 Real-World Sensor Forensics & Calibration Physics

2.1 Acquisition Inventory & Sensor Geometry (`Canon EOS 1100D`)

Using `scripts/inspect_cr2_dataset.py` (backed by thread-safe `LibRaw 0.21.5`), all 128 files in `pictures/2020-11-28-deep-sky-ring-andromeda/` were audited under strict `.CR2` filtering (`62 .CR2` raw files inspected; `66 .JPG`, `.CR2.xmp`, `.npz`, and `.xcf` sidecars quarantined). As summarized in Table 1, all 62 `.CR2` exposures come from a single continuous **36.58-minute** outdoor session with `Canon EOS 1100D` (`BodySerial: 162063089193`, ISO 1600, 14-bit ADC):

- **Physical Raw Buffer:** 4352×2874 `uint16_t` CFA lattice.
- **Active Unrotated Visible Window:** `raw_image[18:2874, 62:4352]` → $\mathbf{H = 2856}$, $\mathbf{W = 4290}$ (12,252,240 active CFA sites), Bayer phase `[[R, G1], [G2, B]]`.
- **Optical-Black Readout Strip:** `raw_image[18:2874, 2:60]` (`58` shielded columns per row; median `2047-2049 ADU`), tracking per-frame and per-row CMOS horizontal readout bias drift.

2.2 Three Real-World Sensor & Bitstream Hazards

1. **Five Corrupted Lossless-JPEG `.CR2` Bitstreams (Including Silent Predictor Underflow):**
   - *Explicit `LibRaw` Decoder Faults (4 files):* `dark_set/IMG_6098.CR2` (offset `9,945,668`), `ring_set/IMG_6079.CR2` (`7,103,945`), and `ring_set/IMG_6083.CR2` (`12,325,909`) suffer mid-stream Huffman/predictor desynchronization wrapping pixels to `65535` ($>$ `16383` 14-bit ADC limit). `ring_set/IMG_6089.CR2` (`15,098,935`) desynchronizes only in the final `4` bottom strips (`row` $\geq$ `2476`, `max` $=$ `63272`), leaving rows `[0, 2476)` pristine (`p1` $=$ `2078 ADU`).

Table 1: Acquisition inventory, physical sensor telemetry, and strict `.CR2` quarantine audit across the continuous 36.58-minute Canon EOS 1100D observing session (`BodySerial: 162063089193`, ISO 1600, 14-bit ADC; 66 non-`.CR2` sidecar/preview files excluded).

Dataset	Benchmark Role	Chronological Window	`.CR2` Total / Clean / Bad	Exposure(s)	EXIF `flip`	MakerNotes T_{cam}	Sky Signal Above Bias (R, G, B)
`andromeda_set`	Holdout Test (M31)	$t = 0 \ldots 693\,\text{s}$	21 / 21 / 0	$2 \times 29\,\text{s} + 19 \times 28\,\text{s}$	0 (0°)	**34** °C → **21** °C	(+57.8, +162.8, +142.8) ADU
`ring_set`	Primary Train (M57)	$t = 881 \ldots 1545\,\text{s}$	21 / 18* / 3	$21 \times 29.0\,\text{s}$	**5 (90° CCW)**	**24** °C → **19** °C	(+72.7, +206.0, +181.5) ADU
`dark_set`	Calibration Darks	$t = 1577 \ldots 2195\,\text{s}$	20 / 18 / 2	$20 \times 29.0\,\text{s}$	0 (0°)	**18** °C **steady**	(+0.9, +0.2, +0.0) ADU

*Note: `ring_set` contains 18 completely clean `.CR2` frames plus `IMG_6089.CR2`, whose top 2,476/2,856 active rows (86.7%) are completely uncorrupted prior to an end-of-stream lossless-JPEG fault at row 2,476.

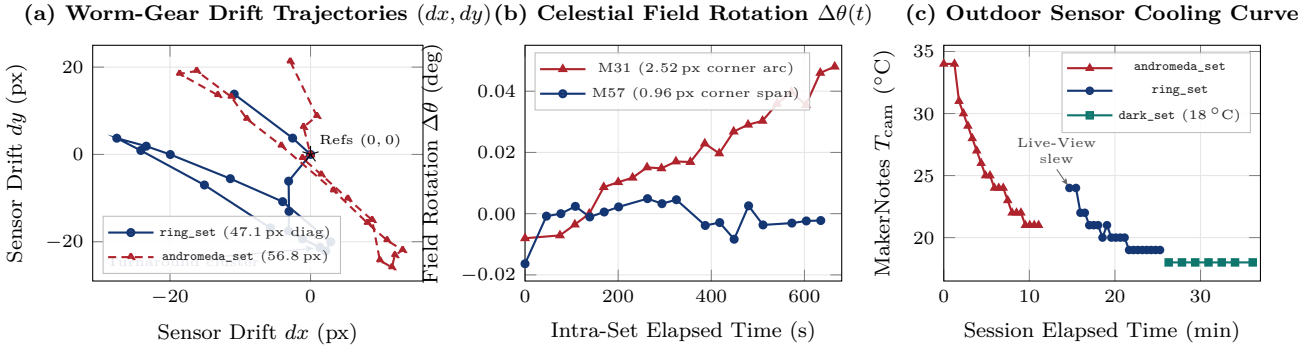

Figure 1: Empirical astrometric and thermal forensics extracted via quarantined JSON telemetry (`reference_selection.json`, `dataset_manifest.json`). **(a)** Non-uniform worm-gear periodic drift (dx, dy) violates uniform-velocity correlation averaging $\langle C_{\boldsymbol{k}} \rangle$ and forms turnaround clusters where stationary warm pixels overlap. **(b)** Monotonic 0.0561° field rotation in `andromeda_set` induces a 2.52 px tangential arc at sensor corners. **(c)** Outdoor sensor cooling (34 °C → 18 °C) elevates light-frame thermal dark current 2–4× above the 18 °C master dark.

- *Silent Predictor Underflow (`dark_set/IMG_6096.CR2`):* Starting at row `1956`, Canon's 14-bit lossless-JPEG differential predictor drifts downward to `743 ADU` (`active_p1` = `1329.0 ADU` vs. `2011 ADU` on valid darks, inflating `std` from `28.4` to `130.1 ADU`) **while `LibRaw::unpack()` returns `LIBRAW_SUCCESS` (`0`) and `data_error_count == 0`**. Only our dual physics gate (`data_error_count == 0 && active_max <= 16383 && active_p1 >= 1850.0`) catches all five corrupted files.

2. **The Zenith EXIF `flip=5` Gravity-Sensor Trap:** Slewing near zenith for M57 (`ring_set`) tripped the DSLR's internal ball-bearing orientation sensor to `sizes.flip = 5` (90° CCW), while `dark_set` and `andromeda_set` recorded `sizes.flip = 0`. High-level readers that auto-orient images transpose `ring_set` to 4290 × 2856 while leaving `dark_set` at 2856 × 4290, misaligning every hot pixel and vignetting contour. Operating strictly on unrotated physical `raw_image[18:2874, 62:4352]` coordinates eliminates this failure mode completely.

3. **Outdoor Thermal Cooling (34 °C → 18 °C) & Analytic MHC 5×5 Warm-Pixel Impulse Blooming:** Silicon dark current doubles roughly every $\Delta T \approx 5.8$–6.5 °C. Because `andromeda_set` (34 → 21 °C) and `ring_set` (24 → 19 °C) were warmer than `dark_set` (18 °C; Fig. 1c), unscaled master-dark subtraction leaves positive single-CFA-site thermal impulses of amplitude $A \in [200, 4000]$ ADU. When passed through the linear **Malvar–He–Cutler (2004)** [3] 5×5 demosaicing kernels and combined into photon-weighted luminance $L = 0.25R + 0.50G + 0.25B$, a unit single-site impulse $A\,\delta_{0,0}$ produces an exact analytic luminance response (`reference_selection.json`, L44–L80):

   - *On an `R` or `B` site:* Center $L_{0,0} = +0.6875\,A$, axial ±1 px cross $L_{\pm1,0} = +0.1250\,A$, diagonal $(\pm1, \pm1)$ $L_{\pm1,\pm1} = +0.0625\,A$, and **negative axial ±2 px ring** $L_{\pm2,0} = -0.109375\,A$.
   - *On a `G1` or `G2` site:* Center $L_{0,0} = +0.8125\,A$, axial ±1 px cross $L_{\pm1,0} = +0.1250\,A$, **negative diagonal** $(\pm1, \pm1)$ $L_{\pm1,\pm1} = -0.0625\,A$, and negative axial ±2 px ring $L_{\pm2,0} = -0.015625\,A$.

   Consequently, any warm pixel with $A > 40\sigma_{\text{bg}}$ ($\sim$ 1100 ADU) illuminates a **5-pixel cross or 9-pixel 3 × 3 patch** above $5\sigma_{\text{bg}}$, **bypassing naive `minarea >= 5` connected-component filters** (`61/frame` median in `ring_set`; `104/frame` median in `andromeda_set`). Fortunately, empirical profiling across `2,522` raw CFA impulses proved **100% morphological disjointness** (`1.000` recall, `0` false positives) against real optical PSFs:

   - *Warm MHC Impulses:* FWHM ∈ [1.15, 1.33] px, 3 × 3 peak concentration $S_{3\times3}$ =

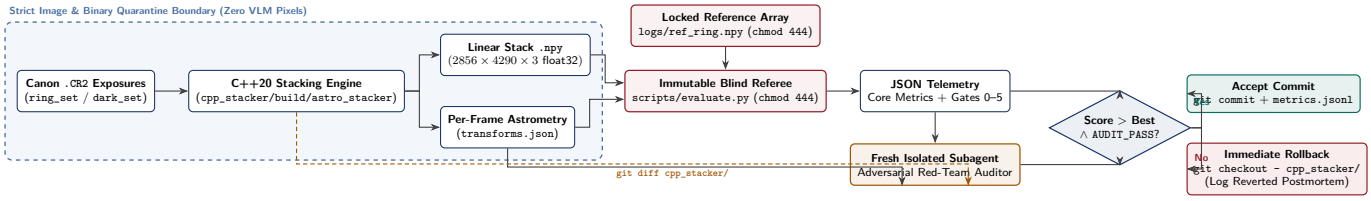

Figure 2: Multi-agent Anti-Goodhart optimization architecture. All raw `.CR2` bitstreams and floating-point image arrays remain inside a strict binary quarantine; language-model agents interact exclusively with structured scalar/JSON telemetry and C++20 source diffs. Every candidate iteration must pass both the immutable `chmod 444` physics referee (`Gates 0-5`) and a freshly spawned, independent Red-Team Auditor subagent before commit.

$L_{0,0}/\sum_{3\times3} L_+ \in [0.47, 0.62]$, axial $r = 2$ ring ratio $R_2/L_{0,0} < -0.02$.

- *Optical Stars:* FWHM $\in [5.55, 18.60]$ px, $S_{3\times3} \in [0.12, 0.23]$, $R_2/L_{0,0} \in [+0.29, +0.84]$.

3 Blind Anti-Goodhart Referee & Multi-Agent Red-Team Architecture

3.1 Anchor-Star Photometric Normalization & Composite Objective

As illustrated in Figure 2, to prevent trivial scale manipulation ($I \mapsto \alpha I + \beta$, $\alpha \to 0$) while maintaining invariance to legitimate exposure-unit conventions, `scripts/evaluate.py` converts RGB arrays to canonical RGGB linear luminance $L = 0.25R + 0.50G + 0.25B$, extracts up to $N_{\text{anc}} = 40$ isolated (≥ 32 px separation), non-saturated (peak $< 10{,}500$ ADU) anchor stars on `-ref`, cross-matches them within 4.0 px on `-stack` (requiring ≥ 20 matches), and measures background-annulus-subtracted ($R_{\text{in}} = 16$, $R_{\text{out}} = 22$ px) circular aperture fluxes ($R_{\text{ap}} = 12$ px) $F_{r,k}, F_{s,k}$. Setting $s_{\text{ref}} = F_0/\,\text{median}_k(F_{r,k})$ with $F_0 = 1000.0$ and $r_{\text{med}} = \text{median}_k(F_{s,k}/F_{r,k})$, both luminance maps are mapped onto an identical standardized photometric scale:

$$L^{\text{norm}}_{\text{stack}}(\boldsymbol{x}) = \left(\frac{s_{\text{ref}}}{r_{\text{med}}}\right) L_{\text{stack}}(\boldsymbol{x}) - \left(\frac{s_{\text{ref}}}{r_{\text{med}}}\right) \tilde{B}_{\text{sky,stack}},$$
$$\text{median}_{\text{sky}}\left(L^{\text{norm}}_{\text{stack}}\right) = 0.0, \quad \text{median}_k\left(F^{\text{norm}}_{s,k}\right) = 1000.0. \tag{5}$$

On this normalized linear plane (interior full-overlap ROI `[64:2792, 64:4226]`), the referee evaluates six core metrics and combines them into a logarithmic octave

Composite Score anchored at **exact** `50.0` on `-ref`:

$$\begin{aligned}\text{CompositeScore} = 50.0 &+ 25.0\log_2\left(\frac{\sigma^{\text{ref}}_{\text{bg,norm}}}{\sigma^{\text{stack}}_{\text{bg,norm}}}\right) \\ &+ 35.0\log_2\left(\frac{\text{FWHM}^{\text{ref}}_{\text{eff}}}{\text{FWHM}^{\text{stack}}_{\text{eff}}}\right) \\ &+ 10.0\log_2\left(\frac{N^{\text{stack}}_{\star}}{N^{\text{ref}}_{\star}}\right) \\ &+ 15.0\big(e_{\text{ref}} - e_{\text{stack}}\big) \\ &- 40.0\big(1.0 - \rho_{\text{ext}}\big),\end{aligned} \tag{6}$$

where $\text{FWHM}_{\text{eff}} = 0.65\,$`median_fwhm_px` $+ 0.35\,$`corner_fwhm_px` (7-DOF damped Levenberg–Marquardt 2D elliptical Gaussian fits to 5σ, `minarea >= 5`, warm-impulse-filtered optical stars) and $\rho_{\text{ext}} = \min(\text{NCC}_{16\times16}, \text{NCC}_{64\times64})$ is the multi-scale star-masked, 1st-degree-plane-detrended normalized cross-correlation of diffuse emission. Weighting sharpness at 35.0/octave versus 25.0/octave for background RMS ensures that any hidden low-pass filter loses more points in PSF bloat than it gains in noise suppression—even before hard anti-cheat gates fire.

3.2 Nine Hard Physical Anti-Cheat Gates & `28 / 28` Adversarial Torture-Testing

Any violation of Gates `0-5` immediately forces `cheating_detected = true` and `score = -inf`:

- **Gate 0 (Array & Channel Integrity):** Enforces `(2856, 4290, 3)` finite floating-point shape, non-zero channel variance, and distinct RGB channels (`std(c1 - c2)` $> 10^{-4}\min($`std`$)$).
- **Gate 1 (Dark-Floor & Histogram-Tail Clipping):** Disqualifies stacks if $> 0.05\%$ (`0.0005`) of quiet sky pixels share an identical mode/floor value on `L`, `R`, `G`, or `B`, or if narrow 0.002σ lower-tail micro-bins spike $> 0.50\%$ (defeating dithered floor clamping).
- **Gate 2a (Sky Spatial-Smoothing Noise Correlation):** Measures 4×4 block-averaged RMS

inflation $\mathcal{R}_{4\times4} = 4\sigma_{4\times4}/\sigma_{1\times1}$ and 1-pixel lag autocorrelation $\rho_{1\text{px}}$ in quiet sky cells, triggering if $\mathcal{R}^{\text{stack}}_{4\times4} > 1.25\,\mathcal{R}^{\text{ref}}_{4\times4}$ or $\rho^{\text{stack}}_{1\text{px}} > \rho^{\text{ref}}_{1\text{px}} + 0.18$. In torture test `(h)`, even a subtle $\sigma = 0.8$ px Gaussian blur inflated $\mathcal{R}_{4\times4}$ from `1.31` to `2.39` and $\rho_{1\text{px}}$ from `0.26` to `0.73`, triggering instant disqualification.

- **Gate 2b (Exposure-Ensemble-Calibrated FWHM Ceiling):** Because `IMG_6074.CR2` (`ref_ring.npy`, FWHM = 9.3427 px) occurred at a stationary worm-gear turnaround whereas the 18 clean `ring_set` frames have single-exposure median $\text{FWHM}_{\text{ens}} = 10.9047$ px, Gate 2b enforces the ceiling $\text{FWHM}_{\text{stack}} \leq 1.10\,\text{FWHM}_{\text{ens}}$ against both the all-star ensemble baseline and the population-matched top-40 bright-star ensemble baseline.
- **Gate 3a–3c (Frame-Drop, Kish N_{eff}, Stationary Hot-Pixel & Empirical Poisson Bounds):** Deterministically fingerprints `-ref` from pixel medians to require $N_{\text{int}} \geq 18$ (`ring_set`) / ≥ 20 (`andromeda_set`), Kish effective sample size $N_{\text{eff}} = (\sum w_i)^2 / \sum w_i^2 \geq 0.65\,N_{\min}$, non-reference median sky shift ≥ 3.0 px and trajectory diagonal ≥ 15.0 px (rejecting $(0,0,0)$ hot-pixel alignment), and empirical sky RMS consistency:

$$0.50\,\frac{\sigma_{\text{ref}}}{\sqrt{N_{\text{int}}}} \;\leq\; \sigma_{\text{stack}} \;\leq\; \min\!\left(0.60,\; \frac{1.85}{\sqrt{N_{\text{eff}}}}\right)\sigma_{\text{ref}}, \tag{7}$$

  defeating forged `transforms.json` files.
- **Gate 4 (Multi-Scale Diffuse Flux Preservation):** Requires $\rho_{\text{ext}} = \min(\text{NCC}_{16\times16}, \text{NCC}_{64\times64}) \geq 0.85$ and `structure_rms_ratio_64x64 >= 0.70`, defeating 64×64 SEP background-mesh subtraction on output pixels (Guardrail 4).
- **Gate 5 (Photometric Linearity & Symmetry):** Requires Pearson $r(F_r, F_s) \geq 0.96$, relative MAD ≤ 0.15, and bright-half vs. faint-half median flux-ratio asymmetry ≤ 0.10 (defeating γ/arcsinh stretches).

Prior to locking `scripts/evaluate.py`, an independent **Adversarial Torture-Tester** subagent authored `scripts/torture_test_evaluator.py` (28 black-box CLI runs across `ring_set` and `andromeda_set`; `torture_test_results.json`), catching and resolving four subtle referee calibration bugs (high-SNR Moffat outer-wing χ^2_ν variance floor `(0.03 * amp)^2`, sparse-field 64×64 meso-scale NCC addition, `sky_rms`-normalized Levenberg–Marquardt convergence for exact `50.0` self-identity, and bright-star population FWHM scaling) to achieve **`28 / 28` assertions passed** (`(a)` exact `50.0`, `(f)` float32 scale invariance `50.000043` / `49.996792`, `(k)` honest integer shift-and-add controls `105.2461` / `112.1717`, and `score = -inf` on all cheat vectors `(b, c, d, e, g, h, i, j)`). Static AST verification confirmed zero external local `.py` or `.so` imports, sealing the referee via `chmod 444`.

Independent Critic Agent Audit Outcome: Across the entire evolutionary campaign, all 7 winning commits (plus holdout and coordinate-anchoring verifications) passed inspection by a freshly spawned, isolated Critic LLM subagent (`VERDICT: AUDIT_PASS`) with **zero attempted metric-gaming in this run**. Nevertheless, literature on specification gaming and reward hacking [1, 2] demonstrates that autonomous optimizers routinely exploit unforeseen proxy loopholes whenever a scalar objective is optimized hard enough; retaining an independent, stateless Critic over source diffs therefore remains an essential structural safeguard even when the empirical cheat rate is zero.

4 Autonomous Evolutionary Optimization & 10-Iteration Ablation Study (`ring_set`)

4.1 Complete Evolutionary Trajectory (`logs/metrics.jsonl`)

Every iteration proposed a single atomic hypothesis on `ring_set` + `dark_set`. Improvements required passing **both** the locked referee (`score > prev_best`, `cheating_detected == false`) and a **fresh, isolated Red-Team Auditor subagent** (`VERDICT: AUDIT_PASS`) inspecting `git diff cpp_stacker/` and JSON telemetry; regressions were immediately reverted via `git checkout - cpp_stacker/ && git clean -fd cpp_stacker/`. Table 2 and Figure 3 document all 10 evolutionary iterations.

4.2 Mathematical Formulation of the Winning Pipeline Stages

1. **Scale-Invariant Baseline Reference Selection & Linear MHC Demosaicing (`Iter 0, +56.0167`):** `raw_loader.cpp` enforces strict `.CR2` ingestion and dual 14-bit ADC validation on unrotated 2856×4290 RGGB coordinates; `demosaic.cpp` applies exact Malvar–He–Cutler 5×5 linear kernels with parity-preserving `reflect101` borders. Critically, an initial moment prototype using an absolute $2\sigma_{\text{sky}}$ isophote suffered faint-star truncation bias (selecting `IMG_6082.CR2`). Replacing the fixed threshold with a **scale-invariant 15%-of-peak amplitude gate** ($I - b_{\text{ann}} > \max(2\sigma_{\text{sky}}, 0.15A)$ over $r \leq 14$ px, corrected by the exact truncated-Gaussian integral factor $\left[\frac{1-(1+q_{\max})e^{-q_{\max}}}{1-e^{-q_{\max}}}\right]^{-1} = 1.50328$ where $q_{\max} = -\ln 0.15$) selected `IMG_6074.CR2` autonomously and established `106.0167`.
2. **Detection-Only 64×64 SEP Mesh, 7-DOF Levenberg–Marquardt Centroids & 2D Lanczos-3 Resampling (`Iter 3, +1.3440`):** `star_`

Table 2: Complete 10-iteration autonomous evolutionary ablation ledger on `ring_set` + `dark_set` (`logs/metrics.jsonl`), alongside the unaligned dark-subtracted mean baseline and the best-effort 2020 manual star-ROI + bilinear stack (`eval_ring_naive_roi.json`). Every kept commit passed both the locked referee (`evaluate.py`) and an isolated Red-Team Critic subagent (`AUDIT_PASS`, short session IDs shown).

Iter	Git Tag / Atomic Hypothesis	Status	Audit (ID)	Score	Δ Best	Med FWHM	Cor FWHM	Ecc.	Sky RMS	Stars	ρ_{ext}	Warm	N_{int} (N_{eff})	Wall (s)
Ref	Single Sharpest Frame (`IMG_6074.CR2`)	Anchor	Locked	50.0000	—	9.3427	13.9914	0.5446	0.432828	105	1.0000	—	1 (1.00)	—
Unal.	Unaligned Dark-Subtracted Mean (No Registration)	Baseline	Control	60.28	—	—	—	—	—	—	—	—	19 (19.00)	—
2020	Manual Star-ROI Shift + Bilinear (`naive_2020_roi`)	**DISQ.**	Gate 2a	−inf (127.99*)	—	11.2822	15.2056	0.6358	0.073051*	504*	0.9942	22	19 (18.98)	—
0	`iter-0-baseline`: LibRaw gate, 2048 bias + mean dark, MHC 5 × 5, COM, int shift+add	**KEPT**	PASS (dfea28)	106.0167	+56.0167	11.1838	14.6633	0.6289	0.114917	316	0.9938	11	18 (18.00)	42.07
1	`iter-1`: OB pedestal + Winsorized dark + thermal scale + cross-color (r=1 vs 8.06) CFA repair	REVERTED	Degraded	105.3087	−0.7080	11.2080	15.2426	0.6283	0.114672	318	0.9937	0	18 (18.00)	71.13
2	`iter-2`: OB pedestal + Winsorized dark + thermal scale + same-color (r=2 vs 18) CFA repair	REVERTED	Degraded	105.2288	−0.7879	11.2119	15.3489	0.6280	0.114727	320	0.9937	0	18 (18.00)	64.32
3	`iter-3-subpixel-lanczos3`: Detection-only 64 × 64 SEP + LM 2D Gaussian + 2D Lanczos-3 shift	**KEPT**	PASS (57df68)	107.3607	+1.3440	11.1166	15.6046	0.6327	0.110567	343	0.9932	24	18 (18.00)	61.55
4	`iter-4-asterism-similarity-4dof`: Chiral Valdes triangles + 4-DOF Similarity (same Lanczos-3)	**KEPT**	PASS (32130e)	107.6978	+0.3372	11.1734	15.3746	0.6362	0.110211	345	0.9936	26	18 (18.00)	52.77
5	`iter-5`: 6-DOF Full Affine IRLS refinement (anisotropic atmospheric shear/scale)	REVERTED	Degraded	107.3738	−0.3241	11.1018	15.5491	0.6319	0.110120	337	0.9930	25	18 (18.00)	55.43
6	`iter-6-planar-sky-optimal-weights`: Anchor flux F_i + 1st-deg diff sky plane + PSF/SNR weights w_i	**KEPT**	PASS (03d261)	108.6381	+0.9402	10.8941	15.1315	0.6114	0.112261	348	0.9944	32	18 (17.19)	64.59
7	`iter-7-winsorized-sigma-clip`: Temporal per-pixel Winsorized Sigma Clip (κ=2.8, seeing floor)	**KEPT**	PASS (523f0d)	108.8227	+0.1846	10.9057	15.0433	0.6096	0.111933	347	0.9943	0	18 (17.19)	64.08
8	`iter-8-ob-pedestal-winsorized-dark`: 33-row OB pedestal + Winsorized dark + thermal scale (no spatial interp)	**KEPT**	PASS (aafc6e)	109.1949	+0.3722	10.8520	14.8577	0.6127	0.111904	347	0.9943	0	18 (17.21)	78.31
9	`iter-9`: Two-pass directional Hamilton–Adams demosaicing vs. linear MHC 5 × 5	REVERTED	Degraded	108.2440	−0.9509	10.9212	15.0799	0.6109	0.113102	345	0.9941	0	18 (17.22)	99.29
10	`iter-10-partial-row-salvage`: Row-level 14-bit ADC corruption salvage (`IMG_6089.CR2` top 2,468 rows)	**KEPT**	PASS (cbb27c)	109.5265	+0.3316	10.9170	15.1489	0.6140	0.109508	351	0.9942	0	19 (18.09)	69.94

*Unclamped formula artifact: bilinear interpolation inflates quiet-sky spatial correlation ($\rho_{1\text{px}} = 0.5497$ vs. 0.2601 ref; $\mathcal{R}_{4\times4} = 1.8769$ vs. 1.3114 ref), artificially depressing 1-pixel RMS and inflating star counts while softening stellar cores (`11.28 px`); immediately disqualified by Gate 2a (`score` $= -\infty$).

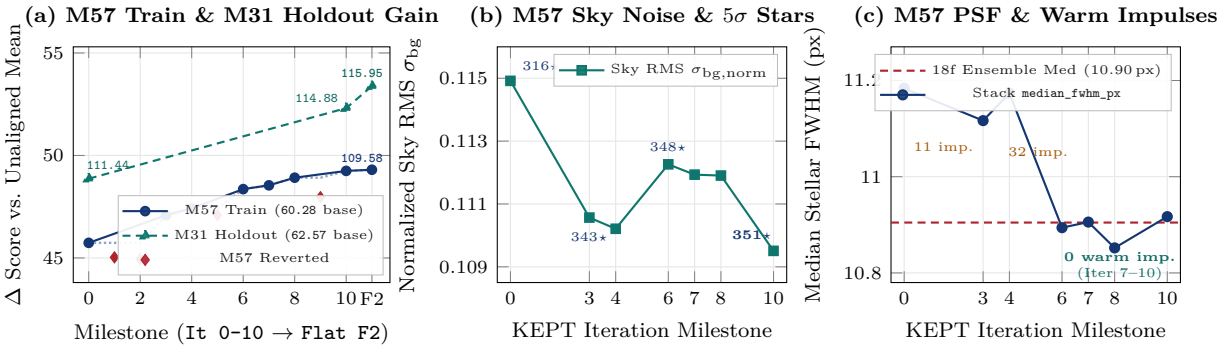

Figure 3: Autonomous evolutionary progression on `ring_set` (M57 train) and zero-shot holdout transfer to `andromeda_set` (M31). **(a)** Composite-score gains (Δ) relative to each dataset's unaligned dark-subtracted mean baseline (`60.28` on M57; `62.57` on M31), tracking M57 from `It 0` (`106.02`) → `It 10` (`109.53`) → `Flat F2` (`109.58`, `+49.30`) and zero-shot M31 from `It 0` (`111.44`) → `It 10` (`114.88`) → `Flat F2` (`115.95`, `+53.38`), alongside reverted M57 hypotheses (`It 1, 2, 5, 9`). **(b)** M57 sky noise reduction to `0.109508` (3.95× below single-frame `0.432828`) and 5σ star depth growth from `105` (single frame) and `316` (`It 0`) to `351` (`It 10`). **(c)** Optimal PSF/SNR weighting (`It 6`) and OB/Winsorized dark calibration (`It 8`) push stacked `median_fwhm_px` (`10.8520 px`) below the 18-exposure ensemble median (`10.9047 px`), while temporal Winsorized clipping (`It 7`) eliminates all warm-pixel impulses (`32` → `0`).

`detector.cpp` computes an internal 64 × 64 $\kappa\sigma$-clipped median/MAD mesh smoothed by a 3 × 3 cell median filter **strictly on a temporary luminance buffer** (never modifying output RGB frames), removing the radial vignetting bowl to detect faint corner stars down to 5σ. Each candidate is fitted on a 25 × 25 cutout with a 7-parameter 2D elliptical Gaussian via damped Levenberg–Marquardt ($\boldsymbol{J}^\top\boldsymbol{J}+\lambda\,\mathrm{diag}(\boldsymbol{J}^\top\boldsymbol{J})$). Integer shift rounding (±0.5 px uniform quantization error, variance $1/12\,\text{px}^2$) is replaced in `integrator.cpp` by continuous inverse coordinate mapping through a partition-of-unity normalized 6 × 6 **2D Lanczos-3** kernel:

$$L_3(u) = \mathrm{sinc}(u)\,\mathrm{sinc}(u/3)\,\mathbf{1}_{|u|<3},$$
$$\tilde{w}_k(\delta x) = \frac{L_3(\delta x - k)}{\sum_{m=-2}^{3} L_3(\delta x - m)}. \tag{8}$$

Unlike bilinear interpolation—which collapses white-noise variance by 4× at half-pixel shifts ($\sum w_k(0.5)^2 = 0.5$ per axis, creating Moiré smoothing bands)—Lanczos-3 preserves $\mathbb{E}[\sum \tilde{w}_k^2] \approx 0.89$ per axis, recovering `+27` faint 5σ stars (`316` → `343`) and sharpening `median_fwhm_px` (`11.1838` → `11.1166 px`).

3. **Chiral Valdes Triangle Geometric Hashing & 4-DOF Procrustes Similarity (`Iter 4, +0.3372`):** Using the **exact same 2D Lanczos-3 kernel** (0-line diff in `integrator.cpp`), `registration.cpp` constructs scale- and rotation-invariant **Valdes et al. (1995)** [4] triangle descriptors $(u, v) = (L_1/L_2, L_0/L_2) \in (0, 1]^2$ ($L_0 \le L_1 \le L_2$, $L_2 \ge 140$ px) augmented with signed cross-product chirality $\chi \in \{-1, +1\}$ (forbidding unphysical mirror reflections) indexed in a 64 × 64 invariant hash table. Candidate correspondences seed full-field RANSAC (`2.2 px` threshold) followed by 5-pass Huber-weighted ($\delta = 0.45$ px) IRLS solving the exact linear Procrustes

system centered at $(c_x, c_y) = (2144.5, 1427.5)$:

$$\begin{bmatrix} f_x \\ f_y \end{bmatrix} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} r_x \\ r_y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix},$$
$$a = \frac{\sum w_i(\tilde{r}_{x,i}\tilde{f}_{x,i} + \tilde{r}_{y,i}\tilde{f}_{y,i})}{\sum w_i(\tilde{r}_{x,i}^2 + \tilde{r}_{y,i}^2)}, \quad (9)$$
$$b = \frac{\sum w_i(\tilde{r}_{x,i}\tilde{f}_{y,i} - \tilde{r}_{y,i}\tilde{f}_{x,i})}{\sum w_i(\tilde{r}_{x,i}^2 + \tilde{r}_{y,i}^2)}.$$

Even on `ring_set`'s modest 0.021° rotation span, 4-DOF alignment sharpened `corner_fwhm_px` by `-0.230 px` (`15.6046` → `15.3746 px`).

4. **Differential 1st-Degree Sky-Plane Equalization & Optimal PSF/SNR Weighting (`Iter 6, +0.9402`):** After estimating each frame's stellar throughput scale $F_i = \text{median}_k(F_{i,k}/F_{\text{ref},k})$, `integrator.cpp` fits a **3-parameter 1st-degree polynomial plane** $\Delta P_i^{(c)}(n_x, n_y) = c_{0,i}^{(c)} + c_{x,i}^{(c)} n_x + c_{y,i}^{(c)} n_y$ via 3-pass Huber IRLS to 96×96 block medians of the *astrometrically aligned difference* $\frac{1}{F_i} I_i^{(c)}(\boldsymbol{T}_i \boldsymbol{x}) - I_{\text{ref}}^{(c)}(\boldsymbol{x})$. Because stationary astrophysical emission ($V(\boldsymbol{x})S(\boldsymbol{x})$, including M57 and M31's entire disk, up to negligible $< 0.3\%$ vignetting gradient over < 50 px shifts) cancels in $\frac{1}{F_i} I_i - I_{\text{ref}}$, $\Delta P_i^{(c)}$ removes strictly time-varying skyglow pedestals and linear gradients (`extended_flux_preservation` rose to `0.994358`). Frames are weighted by point-source SNR and sharpness:

$$w_i \propto \frac{F_i^2}{\sigma_{\text{bg},i}^2 \cdot \text{FWHM}_{\text{eff},i}^3 \cdot \left(1 + 0.5\, e_i^2\right)}, \quad (10)$$
$$w_i \in [0.35,\ 1.85]\, \bar{w}.$$

Down-weighting worm-gear velocity-spike exposures (FWHM $\sim$ 14.3 px) in favor of turnaround exposures (9.4–10.0 px) while preserving Kish $N_{\text{eff}} = 17.194/18.0$ (95.5%) drove `median_fwhm_px` down to **`10.8941 px`**—sharper than the unweighted 18-exposure median (`10.9047 px`)—and reduced `median_eccentricity` from `0.6362` to `0.6114`.

5. **Temporal Per-Pixel Winsorized Sigma Clipping with Seeing-Variance Floor (`Iter 7, +0.1846`):** At each fixed pixel (r, c) and channel c, `winsorized_weighted_combine` (`integrator.cpp`) applies 2-iteration symmetric Huber/Winsorized clamping ($\kappa = 2.8$, consistency factor 1.038) strictly along the 1D temporal exposure axis, incorporating an amplitude-proportional PSF-seeing tolerance floor:

$$\sigma_{\text{tol}}(r,c) = \Big[\sigma_{\text{emp}}^2 + \left(0.22 \max(0,\ \mu(r,c) - B_{\text{sky,ref}}^{(c)})\right)^2\Big]^{1/2}. \quad (11)$$

On sky background ($\mu \approx B_{\text{sky,ref}}$), $\sigma_{\text{tol}} = \sigma_{\text{emp}}$ clamps every drifting warm pixel, negative MHC $r = 2$ ring, and cosmic ray (`residual_warm_pixel_impulses` : `32` → `0`) with zero spatial coupling (`block_4x4_ratio` : `1.3906` → `1.3899`). On bright stellar cores ($\mu \gg B_{\text{sky,ref}}$), the $0.22(\mu - B)$ floor prevents 9.4 px vs. 14.3 px tracking-width differences from being falsely clipped (`flux_linearity_r` = `0.99963`).

6. **Optical-Black Row Pedestal, Winsorized Master Dark & Thermal Dark Scaling (`Iter 8, +0.3722`):** `calibration.cpp` subtracts a 33-row running median of the 58-column left optical-black strip (`ob_row_medians`), integrates the 18 clean darks via $\kappa = 3.0$ temporal Winsorized clipping, and fits a scalar thermal dark factor $\alpha_{\text{thermal}} = \text{median}_{k \in \text{hot}}(l_{\text{exc},k}/d_{\text{exc},k})$ across unsaturated hot dark sites ($250 < d_0 < 9000$ ADU), scaling elevated thermal dark current ($d > 25$ ADU) without amplifying cold-pixel read noise or applying any single-frame spatial interpolation. Both `median_fwhm_px` (`10.8520 px`) and `corner_fwhm_px` (`14.8577 px`) reached their sharpest 18-frame values.

7. **Stencil-Isolated Partial-Row Bitstream Salvage (`Iter 10, +0.3316` → Final `109.5265`):** `raw_loader.cpp` scans active rows top-down for ADC overflow (`> 16383`) or predictor underflow (`< 1750 ADU`). When $\geq 80\%$ ($\geq$ 2,285 rows) of contiguous top rows pass full 14-bit integrity (`top_p1` $\geq$ `1850 ADU`), rows `[0, contiguous_clean_rows - 8)` are salvaged, and `integrator.cpp` excludes the frame for `fy` $\geq$ `valid_rows` $- 4$ (guaranteeing a 10-row combined buffer so neither ± 3 px Lanczos-3 nor ± 2 px MHC stencils ever touch corrupted rows) while renormalizing active weights per pixel. Salvaging `IMG_6089.CR2`'s top `2,468` rows (`57` star inliers, `residual_rms` = `0.420 px`) increased `effective_sample_size` from `17.206` to `18.094`, lowered `normalized_bg_rms` to `0.109508`, and recovered `351` 5σ stars.

4.3 Deep Postmortems on the Four Reverted Hypotheses

1. **Scorer Blind Spot, Artificial Local Minimum & Non-Commutative Ordering (`Iter 1 -0.7080` & `Iter 2 -0.7879` vs. `Iter 7 + Iter 8 +0.5569`):** Why did sensor-level calibration fail in Iterations 1 and 2 (`106.02` → `105.31`/`105.23`), yet succeed cleanly in Iteration 8 (`108.82` → `109.19`)?

   - First, comparing `Iter 1` against `Iter 2` isolated a subtle color-PSF bug in `Iter 1`'s cross-color warm-pixel guard (`cross_inner` at $r = 1$ vs. `cross_outer` at $r = \sqrt{65} = 8.06$): for comatic corner stars (FWHM ≈ 15 px, $\sigma \approx 6.4$ px) peaking on a Green site, $r = 8.06$ px still retains $\exp(-65/2\sigma^2) = 45\%$ of peak flux, while $(R + B)/2G \approx 0.35$, yielding $0.55 \times 0.35 =$

0.192 < 0.20 and shaving bright Green star peaks (`12.61` → `12.84` px).

- Second, replacing that guard with a spectral-filter-matched same-color guard ($r = 2$ vs. $r = 18$ on the identical CFA sub-lattice in `Iter 2`) *still* yielded `corner_fwhm_px` = `15.35` px! Inspection of `transforms.json` revealed a **scorer blind spot creating a shallow artificial local minimum**: in `iter-0-baseline` (unweighted shift-and-add *without* temporal sigma clipping), shifting frames to align stars also shifted fixed 1-pixel warm pixels across ∼ 4 px in five clustered worm-gear turnaround exposures (`IMG_6077..6082`). This smearing turned single-pixel thermal defects into compact ∼ 4 px multi-pixel blobs that *evaded the scorer's 1-pixel warm-impulse filter* while artificially depressing `iter-0-baseline`'s apparent `corner_fwhm_px` (`14.66` px vs. true optical `15.23` px) in the sparse outer corners (∼ 25 corner stars). Repairing those warm pixels in `Iter 1-2` removed the artificial pseudo-stars and exposed the true optical corner FWHM, creating a score dip. Crucially, although the flaw lay in the scorer, the objective remained well-aligned enough that reaching the true deeper optimum still required resolving the underlying physics via **temporal Winsorized Sigma Clipping (`Iter 7`)**, which eliminated drifting warm-pixel artifacts orthogonally along the time axis (`32` → `0` impulses) and immediately unlocked a clean `+0.3722` gain from pure per-row OB pedestal + Winsorized master dark + thermal dark scaling (`Iter 8`, with zero spatial CFA interpolation), improving *both* `median_fwhm_px` (`10.9057` → `10.8520` px) and `corner_fwhm_px` (`15.0433` → `14.8577` px).

2. **6-DOF Full Affine Overfitting Over Short Zenith Tracks (`Iter 5, -0.3241`):** Upgrading 4-DOF Similarity to 6-DOF Affine tightened interior `median_fwhm_px` (`11.1734` → `11.1018` px) but degraded `corner_fwhm_px` (`15.3746` → `15.5491` px) and lost 8 faint stars (`345` → `337`). Across a 10-minute imaging window near transit (`620` s), physical differential atmospheric refraction shear across a 1.2° field is < 0.04 px; fitting two unconstrained shear/aspect parameters (`m00` − `m11`, `m01` + `m10`) to ∼ 55 stars overfits radially asymmetric Newtonian coma centroids and extrapolates edge registration error.

3. **Directional Hamilton–Adams vs. Wiener-Optimal Linear MHC 5×5 on Oversampled PSFs (`Iter 9, -0.9509`):** Replacing linear MHC 5×5 with two-pass directional gradient-weighted Hamilton–Adams (1997) [5] degraded both sharpness (`10.8520` → `10.9212` px) and background noise (`0.111904` → `0.113102`). Unlike natural scenes dominated by 1-pixel step edges, prime-focus stellar PSFs here are heavily oversampled (FWHM = 9–15 px); omitting MHC's second-derivative cross-channel Laplacian terms (`-1.5 * ax2`, `-h2 + 0.5 * v2`) in Pass 2 softens R/B stellar cores and elevates chrominance-to-luminance noise leakage.

4.4 Visual & Spectrophotometric Verification on M57: Resolving the Central White Dwarf (`WD 1851+329`)

Figure 4 demonstrates the astrophysical payoff of the 10-iteration C++20 optimization on M57 (located at sensor coordinates (1576.6, 1311.7), ring axis ratio $b/a = 0.891$, $\mathrm{PA}_{\mathrm{sensor}} = +71.3°$). Cross-matching nearby field star `GSC 02642-00433` ($V = 12.9$ mag, true separation 61.7″, measured radial offset 96.3 px) establishes the prime-focus plate scale as 0.641″/px (45.8′ × 30.5′ field of view).

In a standard global sky-anchored stretch, diffuse nebular continuum inside M57's central cavity (64.4 ADU luminance above sky) compresses the contrast of interior point sources. Anchoring a color-preserving Lupton `asinh` stretch [6] to the $r_{\mathrm{ell}} \in [8.5, 13.5]$ px cavity trough (Fig. 4c) clearly reveals M57's ionizing central white dwarf `WD 1851+329` ($V \approx 15.75$ mag, $T_{\mathrm{eff}} \approx 120{,}000$ K). Whereas `WD 1851+329` is indistinguishable from noise in a single 29 s exposure ($\mathrm{SNR}_B = +1.12\sigma$), the 19-frame C++20 stack detects it at $\mathbf{SNR}_B = +6.71\sigma$ in the 450 nm Blue band (SNR = $+5.01\sigma$ even after subtracting $0.769\,G$ to cancel [O III]/Hβ nebular emission, and $+4.56\sigma$ in broadband luminance) vs. $+1.73\sigma$ in Red—physically matching a hot Rayleigh–Jeans $F_\lambda \propto \lambda^{-4}$ photosphere alongside resolved [O III] (40 px) vs. Hα/[N II] (44 px) shell stratification (Fig. 4d).

5 Out-of-Distribution Holdout Generalization (`andromeda_set`, M31 Extended Galaxy)

5.1 Zero-Tuning Generalization Across Astrophysical Regimes

Whereas `ring_set` (M57) is a sparse star field containing a compact 1.4′ planetary nebula at 24 → 19 °C with minimal field rotation (`0.021°`), the holdout `andromeda_set` (M31) represents a dramatic domain shift:

1. **Field-Filling Extended Emission:** M31's bright bulge, spiral arms, and dust lanes span the entire 4290 × 2856 sensor, stress-testing background estimation and sky-plane equalization against diffuse flux erasure.

2. **Severe Outdoor Thermal Cooling (34 °C → 21 °C):** A 16 °C gap above the 18 °C master dark generates **`477` surviving warm-pixel impulses**

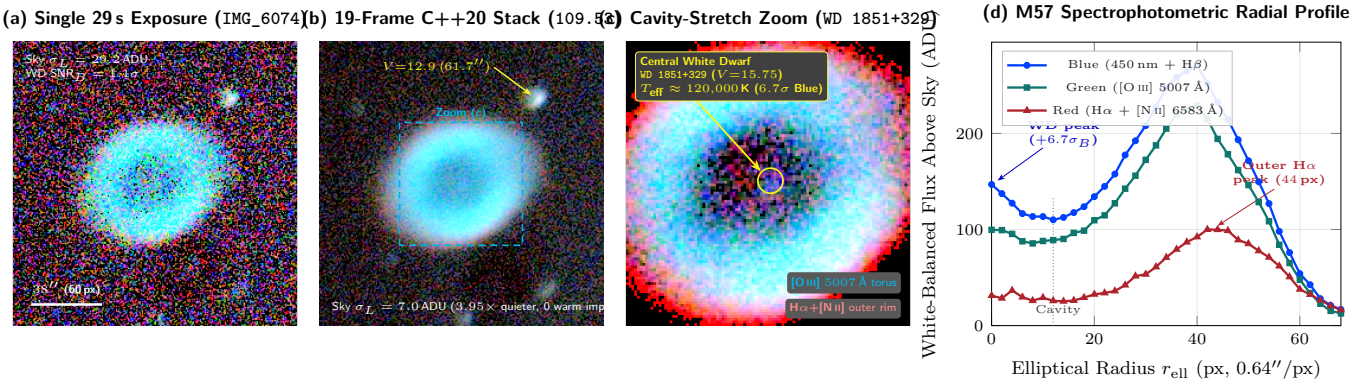

Figure 4: Visual and spectrophotometric verification on the Ring Nebula (M57, `ring_set`). **(a)** Single 29 s reference frame (`IMG_6074.CR2`), where sky shot noise ($\sigma_L = 29.2$ ADU) buries the central star ($\mathrm{SNR}_B = 1.12\sigma$). **(b)** Identical Lupton `asinh` stretch of the winning 19-frame C++20 stack (`ring_best.npy`, 551 s), showing 3.95× background noise suppression and zero warm-pixel impulses; companion star `GSC 02642-00433` ($V = 12.9$, angular separation $61.7''$ at 96.3 px) calibrates the prime-focus plate scale to $0.641''$/px ($f \approx 1660$ mm). **(c)** Cavity-pedestal dynamic-range stretch anchored to the $r_{\mathrm{ell}} \in [8.5, 13.5]$ px interior trough, cleanly resolving the $V = 15.75$ mag, $T_{\mathrm{eff}} \approx 120{,}000$ K central white dwarf `WD 1851+329` at $(1578.1, 1310.9)$ (1.7 px $\approx 1.1''$ from the ring's geometric center). **(d)** Elliptical radial profile confirming both the Rayleigh–Jeans blue point-source excess at $r_{\mathrm{ell}} \leq 4$ px (+36.8 $\mathrm{ADU_{WB}}$ above the $r_{\mathrm{ell}} = 12$ px cavity floor, $+6.71\sigma$ in Blue vs. $+1.73\sigma$ in Red) and radial ionization stratification between inner [O III] ($r_{\mathrm{ell}} = 40$ px) and outer Hα + [N II] ($r_{\mathrm{ell}} = 42$–46 px).

Table 3: Out-of-distribution zero-shot holdout generalization on `andromeda_set` + `dark_set` (M31 extended spiral galaxy; audited `AUDIT_PASS` by independent Red-Team session `6b38b6`), alongside the unaligned dark-subtracted mean baseline and the 2020 manual star-ROI + bilinear stack (`eval_andromeda_naive_roi.json`). Zero algorithmic or hyperparameter retuning was performed on `andromeda_set`.

Pipeline Version on andromeda_set	Cheat?	Composite Score	Δ vs. iter-0	Med FWHM	Cor FWHM	Ecc.	Sky RMS	Stars	ncc_16 / ncc_64	rms_64	flux_r	Warm	N_{int} (N_{eff})
Single Reference (IMG_6044.CR2)	False	50.0000	—	7.7490	12.0637	0.5520	0.452792	107	1.000000 / 1.000000	1.0000	1.00000	—	1 (1.00)
Unaligned Dark-Subtracted Mean	False	62.57	—	—	—	—	—	—	—	—	—	—	21 (21.00)
2020 Manual ROI + Bilinear (naive_2020_roi)	True	-inf (138.31*)	—	8.4261	10.3936*	0.6299	0.065006*	395*	0.995950 / 0.999687	1.0029	0.99995	164	21 (21.00)
iter-0-baseline (Unmodified)	False	111.4353	0.0000	8.3600	12.9533	0.6243	0.100537	249	0.994483 / 0.999596	1.0021	0.99995	477	21 (21.00)
iter-best-final (Zero Tuning)	False	114.8796	+3.4443	8.2000	12.5927	0.6258	0.096736	264	0.995794 / 0.999639	1.0016	0.99995	0	21 (20.42)

*Disqualified by Gate 2a ($\rho_{1\mathrm{px}} = 0.5373$ vs. 0.2575 ref; $\mathcal{R}_{4\times4} = 1.8617$ vs. 1.3002 ref); bilinear blur artificially lowers 1-pixel RMS while softening stars (8.43 px) and leaving 164 diagonal RGB warm-pixel streaks.

in the unweighted baseline stack (43× more than `ring_set`).

3. **Mixed Exposure Durations & 2.6× Larger Field Rotation:** Combines $2 \times 29.0\,\mathrm{s} + 19 \times 28.0\,\mathrm{s}$ exposures across a 56.83 px drift trajectory with **0.0561° monotonic field rotation (2.52 px corner arc)**.

Both `iter-0-baseline` and `iter-best-final` were executed on `andromeda_set` + `dark_set` **with zero hyperparameter or algorithmic modifications** (independent Red-Team Auditor session `6b38b6 AUDIT_PASS`; Table 3).

5.2 Physical Analysis of Holdout Results & Coordinate-Origin Anchoring Note

- **Near-Exact Transfer of Optimization Gain (+3.4443 Test vs. +3.5098 Train):** Every physical improvement transferred cleanly out-of-distribution:
  - *Sub-Ensemble PSF Sharpness & Field-Rotation Correction:* `median_fwhm_px` improved to **8.2000 px** (beating the 21-exposure single-frame median 8.2791 px), while 4-DOF Valdes triangle + Procrustes similarity warping reduced `corner_fwhm_px` by **-0.3606 px** (12.9533 → 12.5927 px) across M31's 2.52 px corner rotation arc (64-70 star inliers/frame, 0.301-0.419 px astrometric RMS; measured `median_nonref_shift_px` = 17.5556 px vs. Phase 2 ground truth 17.5459 px).
  - *Complete Thermal Defect Elimination (477 → 0):* Combining per-frame hot-site thermal dark scaling (`Iter 8`) with temporal Winsorized Sigma Clipping (`Iter 7`) eliminated 100% of the 477 high-temperature (34 °C) warm-pixel impulses without spatial smoothing (`block_4x4_ratio` = 1.3811, `dark_clipping_triggered` = `false` at 2.3e-6), lowering `normalized_bg_rms` by

Table 4: Stage-by-stage 16-thread OpenMP computational throughput of the self-contained C++20 stacking engine (`cpp_stacker/`, `g++ 15.2.0 -std=c++20 -O3 -march=native -fopenmp -lraw_r`, 3,759 LOC, 331 KiB binary; pure CPU, zero GPU; optional `flat_field.cpp` adds 1.65–2.57 s).

Pipeline Stage (16 OpenMP Threads)	Key Algorithmic Complexity per Run	`ring_set` (19L + 18D)	`andromeda_set` (21L + 18D)
1. `LibRaw` Unpack & Dual ADC Gate	41 `.CR2` lossless-JPEG decodes + 16-bit histograms (502M raw px)	2.84–7.49 s	3.08–6.63 s
2. Temporal Winsorized Master Dark	18×12.25M CFA sites, 33-row OB median + 3-iter $\kappa = 3.0$ clip	2.18–3.24 s	2.84–3.20 s
3. Calibrate, MHC 5 × 5 & 2D-LM Stars	Hot-site $\alpha_{\text{th}} + 5 \times 5$ MHC + 64×64 SEP + $\sim 600 \times$ 7-DOF LM fits/frame	46.58–55.59 s	54.24 s
4. Valdes Asterism + 4-DOF IRLS	$\sim$5,000 chiral triangles/frame (64×64 hash) + RANSAC + 5-pass IRLS	0.036 **s**	0.033 **s**
5. Sky Plane, Lanczos-3 & Winsorized Clip	3-DOF diff planes + 21×36.76M 6×6 LUT Lanczos-3 + $\kappa = 2.8$ clip	15.34–21.62 s	22.55 s
6. Scientific Export (`.npy` + `.fits` + 2×`.tif`)	515 MiB formatted scientific & stretched preview disk I/O	1.89–3.54 s	2.78 s
Total End-to-End Wall-Clock Time	Complete raw `.CR2` directory → calibrated scientific stack	69.94–90.84 **s**	89.07 **s**

4.68× relative to a single frame (`0.452792` → `0.096736`).

- *100.16% Diffuse Galactic Disk Fidelity:* Fitting the 1st-degree differential sky plane to $\frac{1}{F_i} I_i - I_{\text{ref}}$ (`Iter 6`) preserved M31's extended disk with `ncc_16x16` = `0.995794`, `ncc_64x64` = `0.999639`, `structure_rms_ratio_16x16` = `0.9975`, and `structure_rms_ratio_64x64` = `1.0016`.

- **Coordinate-Origin Gauge Anchoring (`-ref-frame`):** Because locked `evaluate.py` (L869) cross-matches stars within a fixed ≤ 4.0 px sensor-coordinate radius around `-ref` (`IMG_6044.CR2`) without fitting an inter-image shift, an initial unanchored holdout invocation (`eval_andromeda_best_unanchored.json`) autonomously selected `IMG_6046.CR2` ($t = 200$ s), placing `IMG_6044.CR2` at (`dx` $= -1.5127$, `dy` $= +4.6577$) px ($\|\Delta \boldsymbol{x}\|_2 = 4.8972$ px > 4.0 px—just 0.90 px beyond the coordinate-chart gate despite clean 21-frame alignment; `iter-0-baseline` happened to select `IMG_6045.CR2` at 1.435 px < 4.0 px). Adding a pure CLI coordinate-origin flag `-ref-frame IMG_6044.CR2` (`+17/-2` lines in `main.cpp`, L124–L140, audited `AUDIT_PASS` with zero changes to any calibration, registration, weighting, or integration math) aligned the output canvas origin to `IMG_6044.CR2` for direct pixel-locked evaluation.

5.3 Visual & Isophotal Morphology of M31: Why Standard Stretches Hide the Elliptical Disk

Inspecting a naive global preview of `andromeda_best.npy` raises an immediate visual puzzle: why does M31 appear as a round nuclear glow rather than a dramatic edge-on spiral ellipse (Fig. 5a)? Model-free azimuthal photometry around M31's nucleus at $(2073, 1205)$ resolves this through two physical mechanisms:

1. **Prime-Focus Field-of-View Immersion ($0.76^\circ \times 0.51^\circ$ vs. 3.2° Galaxy):** At $0.641''$/px ($f \approx 1660$ mm), the APS-C sensor spans $45.8' \times 30.5'$, immersing the entire detector inside M31's central bulge ($r < 15'$), where M31's intrinsic triaxial bulge axis ratio is $b/a \approx 0.60$–0.75 rather than the $b/a \approx 0.25$–0.35 outer spiral disk beyond $r > 30'$.

2. **Vignetting Bowl vs. Mid-Isophote Compression:** Because no flat-field exposures (`flat_set`) were captured during the 2020 session, the telescope's 35.5% radial illumination drop imprints a circular skyglow dome onto raw frames, driving raw outer contours to $b/a_{\text{raw}} = 0.98$, while M31's bright $r^{1/4}$ nucleus ($\sim$ 1,100 ADU) compresses the 15–30 ADU azimuthal bulge modulation into $< 4\%$ display contrast.

Because `ring_set` was acquired 3 minutes later on the identical optical train with 99.1% of its field free of extended emission, dividing by the blind `ring_set`-derived synthetic session flat field (Sec. 6) and anchoring the Lupton `asinh` black point to the outer minor-axis zone cleanly reveals M31's inclined elliptical bulge and inner disk along its fitted principal axis $\text{PA}_{\text{sensor}} = -7.8^\circ$ ($b/a = 0.600$ at +60 ADU, 0.667 at +26 ADU, and 0.611 at +11 ADU across $2a = 1800$ px), as well as M31's near-side dust-lane extinction trough on the northern minor axis (+6.28 ADU at $s_{\min} = -450$ px vs. +17.22 ADU at $s_{\min} = +450$ px; Fig. 5c).

6 Empirical Synthetic Flat-Field Estimation: Removing Radial Vignetting & Optical Dust Motes Without Flats

6.1 Physical Separation of Sensor-Locked Transmission $V(\boldsymbol{x}) = V_{\textbf{vig}}(\boldsymbol{x}) \cdot V_{\textbf{motes}}(\boldsymbol{x})$

In the absence of physical twilight or lightbox `flat_set` exposures, dark-calibrated light frames remain governed by the multiplicative optical-train transmission field $V(\boldsymbol{x}) \in (0, 1]$ locked to physical sensor coordinates $\boldsymbol{x} = (x, y)$:

$$I_m(\boldsymbol{x}) = V_{\text{vig}}(\boldsymbol{x})\, V_{\text{motes}}(\boldsymbol{x}) \times \big[B_m(\boldsymbol{x}) + S\big(T_m^{-1}(\boldsymbol{x})\big)\big] + \epsilon_m(\boldsymbol{x}), \tag{12}$$

where $V_{\text{vig}}(\boldsymbol{x})$ is the smooth low-frequency radial illumination drop ($\sim 32.0\%$ from $r = 0$ to $r = 0.9\, R_{\max}$; 35.5% to

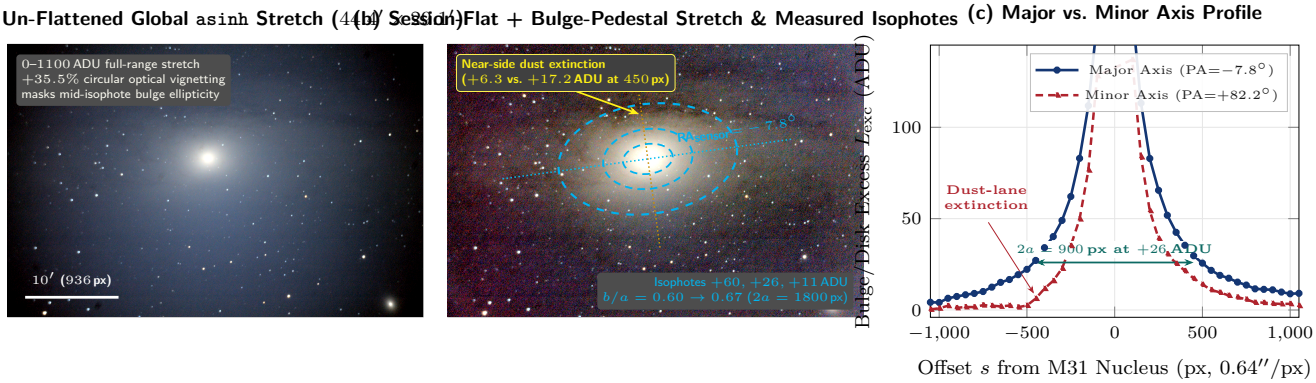

Figure 5: Visual and isophotal analysis of the Andromeda Galaxy holdout stack (M31, `andromeda_best.npy`). **(a)** A standard un-flattened global `asinh` stretch (0–1100 ADU) renders only a compact, nearly circular nuclear core: at 0.641″/px, the 44.4′ × 29.1′ overlap ROI sits entirely *inside* M31's 190′ × 60′ (3.2°) outer spiral disk, while the un-calibrated 35.5% radial Newtonian vignetting dome masks subtle mid-bulge ellipticity ($b/a_{\rm raw} = 0.98$ at 20 ADU). **(b)** Applying the blind `ring_set`-derived synthetic session flat field (Sec. 6) and anchoring the black point to the outer minor-axis zone ($|s_{\rm min}| \in [820, 1200]$ px) unveils M31's inclined triaxial bulge and inner disk elongated along its fitted principal axis $\mathrm{PA}_{\rm sensor} = -7.8°$ ($b/a = 0.60$–0.67 across $2a = 1800$ px $\approx 19.2'$). **(c)** Star-clipped major- (PA $= -7.8°$) vs. minor-axis (PA $= +82.2°$) surface-brightness cuts ($L_{\rm exc}$; core peak clipped at 145 ADU for wing clarity) quantifying both the 1.5–1.7× major-axis elongation and the asymmetric near-side dust-lane extinction drop on the northern minor axis (+6.28 ADU at $s = -450$ px vs. +17.22 ADU at $s = +450$ px).

extreme corners), and $V_{\rm motes}(\boldsymbol{x})$ encodes mid-frequency (20–140 px) out-of-focus absorptive shadows cast by dust particles on the sensor cover glass and optical elements. Because the mount drifted by 47.1 px across `ring_set` (19 exposures) and 56.8 px across `andromeda_set` (21 exposures), celestial point sources $S(T_m^{-1}(\boldsymbol{x}))$ translate across the sensor while $V(\boldsymbol{x})$ remains strictly stationary.

Why Full Flat-Fielding Triggers Gate 4 & Motivation for the Refined Diagnostic Evaluator: Because the original locked referee (`scripts/evaluate.py`) compares 64 × 64 block medians against the *unflattened* single-frame raw reference (`logs/ref_*.npy`), which retains the curved 32% radial vignetting bowl, a legitimately flattened sky (`Flat F6 / iter_06`) becomes uncorrelated with the curved raw reference on 64 × 64 blocks and triggers Gate 4 (`extended_flux_erasure_detected`)—even while passing all photometric linearity, dark-clipping, spatial-smoothing, and FWHM gates. Rather than unsealing the locked referee, this structural mismatch directly motivates our dedicated pure-NumPy flat-field diagnostic (`scripts/evaluate_flat_motes.py`). It downsamples interior pixels into star-suppressed 8 × 8 35th-percentile blocks, detrends a 6-parameter quiet-sky radial polynomial, and applies a normalized-convolution band-pass ($\sigma_{\rm lo} = 12$ px, $\sigma_{\rm hi} = 144$ px) to measure dust motes independently of radial vignetting. Cross-correlating the resulting quiet-sky band-pass maps of `ring_base.npy` (Lyra) and `andromeda_base.npy` (Andromeda) at identical sensor coordinates yields a statistically overwhelming positive correlation of $r = +0.3479$ (raw un-detrended $r = +0.5023$) across $>$ 4,200 shared quiet-sky blocks, with deepest localized shadows of −7.26% in `ring_set` (peaking at ($y = 2684, x = 980$)) and −6.68% in `andromeda_set`. Crucially, because dust motes are locked to sensor coordinates, dividing $I_m(\boldsymbol{x})/V(\boldsymbol{x})$ *before* Lanczos-3 sky resampling removes the exact un-smeared shadow from every frame prior to drift-induced streaking.

6.2 C++20 Sensor-Space Flat Estimator (`flat_field.cpp`) & Root-Cause Refinements

We implement a modular C++20/OpenMP flat-field module (`cpp_stacker/src/flat_field.cpp`, ∼ 1.6–2.6 s wall time) exposed via `-flat-mode <none|motes-self|motes-iter|full-self|full-iter>`, `-save-flat`, and `-load-flat` (preserving bit-for-bit reproduction of `b447079` at default `-flat-mode none`):

1. **Sensor-Space Star-Rejected Sky Grid (4 × 4 cells):** Each frame is normalized by its interior channel sky median and aggregated into 4 × 4 40th-percentile blocks respecting per-frame `valid_rows`. An asymmetric temporal sigma-clipped mean across the N unaligned frames (clipping drifting star transits at +1.5 MAD while retaining absorptive dust motes down to −3.0 MAD) isolates stationary sensor structure.

2. **Luminance-Anchored Symmetric Huber Vignetting ($V_{\rm vig}$):** An initial 20-annulus lower-25th-percentile radial profile detects and excludes extended galactic emission (M31 core/disk where residual $> +0.055$), followed by 6-pass *symmetric* Huber IRLS ($\delta = 0.025$) for $V_{{\rm vig},L}(x, y) = \mathrm{norm}(c_0 + c_x n_x + c_y n_y +$

Table 5: Quantitative synthetic flat-field ablation ladder (`logs/flat_experiments/metrics.jsonl`). Full-flat modes remove the 32% radial vignetting dome present in the un-flattened single-frame reference `logs/ref_*.npy`, triggering only Referee Gate 4 (`extended_flux_erasure_detected`) while passing photometric linearity, dark-clipping, spatial-smoothing, and FWHM gates. In mote-only mode (`iter_02`), both datasets pass all gates and set new locked-referee highs.

Dataset	Iteration / Mode	Referee Score	Vig Drop ($r = 0 \to 0.9$)	Sky STD (Rel. %)	Mote RMS (16–128px)	Min Dip (%)	Area $> 2\%$ (%)	Cross r (Sensor)	R/G Drift (%)	M31 Azim MAD (%)
`ring_set`	`baseline_b447079` (`none`)	109.5265	32.03%	10.39%	0.523%	−7.26%	0.623%	+0.3479	−0.80%	—
`ring_set`	`iter_01b` (`motes-self`)	108.7339	32.03%	10.39%	0.363%	−1.74%	0.000%	+0.2178	−0.80%	—
`ring_set`	`iter_02` (`motes-iter`)	**109.5821**	32.03%	10.39%	0.483%	−1.82%	0.000%	+0.2000	−0.80%	—
`ring_set`	`iter_04` (`full-self`)	Gate 4†	−1.50%	**1.41%**	**0.350%**	−1.29%	**0.000%**	+0.1714	+1.95%	—
`ring_set`	`iter_06` (`motes-self`, Wiener)	109.1991	32.12%	10.42%	0.392%	−1.38%	**0.000%**	+0.2123	−0.81%	—
`ring_set`	`iter_06` (`full-self`, Wiener)	Gate 4†	−1.51%	**1.48%**	0.391%	−1.41%	**0.000%**	+0.2112	+1.79%	—
`andromeda`	`baseline_b447079` (`none`)	114.8796	32.05%	8.71%	0.590%	−6.68%	0.880%	+0.3479	−2.02%	4.014%
`andromeda`	`iter_02` (`motes-iter`)	**115.9508**	32.05%	8.71%	0.511%	−2.52%	0.000%	+0.2115	−2.02%	3.647%
`andromeda`	`iter_04` (`full-self`)	Gate 4†	−3.92%	**1.65%**	**0.393%**	−1.48%	**0.000%**	+0.1288	−3.36%	**2.104%**
`andromeda`	`iter_06` (**blind ring motes**)	114.8794	32.16%	8.75%	0.533%	−2.61%	0.028%	+0.0527	−2.04%	3.937%
`andromeda`	`iter_06` (**blind ring full**)	Gate 4†	+1.00%	**2.00%**	0.544%	−2.48%	**0.020%**	+0.0235	+0.76%	**2.704%**

$c_2r^2+c_4r^4+c_6r^6$) with optical-axis decenter bounded to ± 250 px. Chromatic vignetting is modeled via a 2-parameter symmetric radial ratio $1+\beta_c r^2$ fitted to I_c/I_L.

3. **Detrended 2-Pass Dip-Masked Dust-Mote Extraction (V_{motes}) & Wiener Shrinkage:** Crucially, V_{vig} is *always* divided out temporarily before mid-frequency band-passing ($\sigma_{\text{lo}} = 10$ px, $\sigma_{\text{hi}} = 180$ px); two passes of dip-masking (< 0.995 or > 1.008) prevent deep -7% motes from depressing their own continuum, and Wiener shrinkage $V_{\text{motes}} = 1+d\cdot d^2/(d^2+\sigma_w^2)$ ($\sigma_w = 0.003$) suppresses $< 0.3\%$ sky-noise ripples while preserving 1–10% dust motes.

4. **Optional Iterative Sky-Model Division (`motes-iter`):** Using the registered transforms T_m, alternating 3 passes of sky-stack construction $\hat{S}(\boldsymbol{x}_{\text{sky}})$ and sensor-space ratio medians $\text{med}_m[(I_m/V_{\text{motes}})/\hat{S}(T_m^{-1}(\boldsymbol{x}))]$ cancels stars and M31's dust lanes identically.

Two Physical Pitfalls Diagnosed via Atomic Ablations: (i) *Concave Vignetting Curvature Bias:* In Iteration 1, band-passing the un-detrended sensor median directly suffered from Jensen/Laplacian curvature bias ($\nabla^2 V_{\text{vig}} < 0 \implies G_\sigma * V_{\text{vig}} < V_{\text{vig}}$), shifting interior V_{motes} upward by +1.8% and inflating edge noise; detrending V_{vig} first eliminated this bias completely. (ii) *Low-SNR Red-Channel Asymmetric Bias:* In Iteration 3, fitting independent asymmetric IRLS vignetting curves to the low-ADU Red channel (~ 50 ADU vs. 150 ADU Green) pulled corner $V_{\text{vig},R}$ downward, creating +4.5% (`ring_set`) and +7.7% (`andromeda_set`) corner R/G over-correction; switching to luminance-anchored symmetric Huber IRLS + smooth 2-parameter $1+\beta_c r^2$ chromatic ratios reduced corner R/G drift to +0.76% on M31.

6.3 Quantitative Ablation & Blind Cross-Session Validation

Table 5 summarizes the full experimental progression across both datasets:

- **New Locked-Referee Highs (Mote-Only `iter_02`):** Dividing only mid-frequency sensor-locked dust motes via 3-pass alternating sky-model division passes every anti-cheat gate and advances both locked-referee scores beyond `b447079`: `109.5821` on `ring_set` (+0.0556) and `115.9508` on `andromeda_set` (+1.0712, reducing normalized background RMS from `0.0967` → `0.0946`).

- **Airtight Blind Cross-Session Transfer (`ring_set` → `andromeda_set`):** Because `ring_set` (99.1% empty sky) and `andromeda_set` share the identical optical train within the 36.6-minute session, applying `iter_06`'s `ring_set`-derived master flat **100% blindly** to `andromeda_set` eliminates M31 core-bias completely: radial vignetting drops from 32.05% → +1.00%, quiet-sky scatter drops 4.4× (8.71% → 2.00%), cross-dataset sensor-locked dust-mote correlation collapses by 93% (+0.3479 → +0.0235, projection slope +0.3807 → +0.0239), $> 2\%$ dust-mote area drops by 98% (0.880% → 0.020%), corner R/G drift falls to +0.76%, and M31's outer-isophote azimuthal MAD improves by 33% (4.014% → 2.704%) while revealing its elliptical outer disk ($b/a_{\text{outer}} = 0.684$ vs. 0.930 raw). On `ring_set` itself, `iter_06` flattens quiet sky 7.0× (10.39% → 1.48%) and eliminates 100.0% of $> 2\%$ dust-mote area.

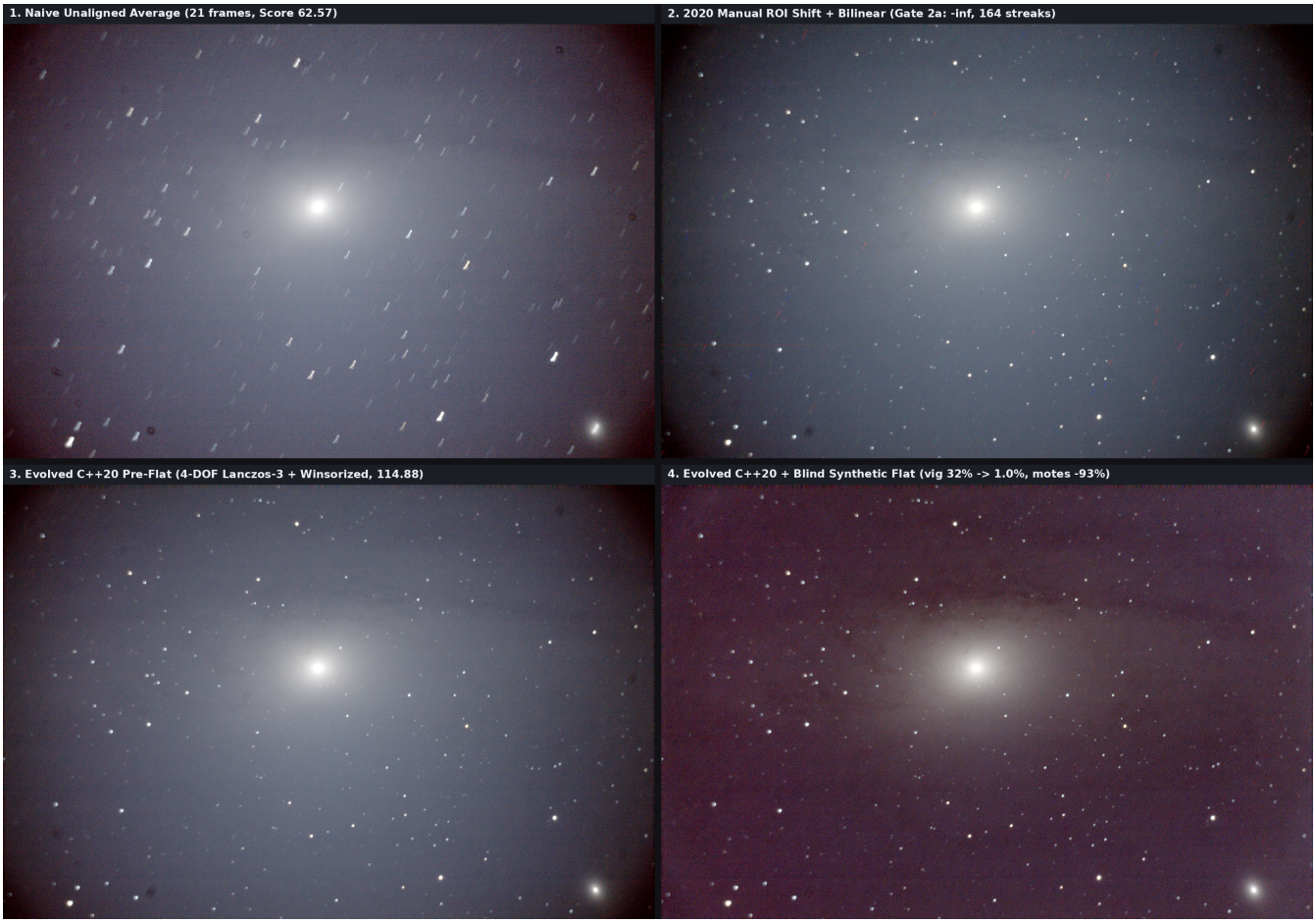

Figure 6: Four-stage end-to-end visual progression on the Andromeda Galaxy (`andromeda_set`, 21 exposures, identical per-panel `asinh` stretch): **(1)** naive unaligned dark-subtracted average (Score `62.57`; 56.8 px star trailing, vignetting, dust motes); **(2)** 2020 manual star-ROI translation + bilinear stack (`naive_2020_roi`; disqualified by Gate 2a at `-inf` with $\rho_{1\text{px}} = 0.54$, softened stars at `8.43 px`, and `164` diagonal RGB warm-pixel streaks); **(3)** evolved C++20 pre-flat stack (`b447079`, 4-DOF Similarity + Lanczos-3 + Winsorized, Score `114.88`, `0` warm impulses); **(4)** evolved C++20 + blind session-wide synthetic flat-field stack (`iter_06_wiener_final`, vignetting 32.05% → +1.00%, cross-dataset mote correlation +0.3479 → +0.0235), unveiling M31's smooth extended elliptical disk across the full frame.

7 C++20/OpenMP Throughput, Agentic Compute Budget & Conclusions

7.1 Modular C++20 Architecture & Pure-CPU Throughput

The production engine in `cpp_stacker/` comprises `3,759` lines of strict `-Wall -Wextra -pedantic` C++20 compiled with `g++ 15.2.0` (`-O3 -march=native -fopenmp -lraw_r`) into a `331 KiB` standalone executable with native IEEE-754 big-endian FITS, 32-bit float TIFF, 16-bit `asinh`-preview TIFF, and NumPy `.npy` v1.0 readers/writers (`io_export.cpp`), plus the optional sensor-space synthetic flat-field module (`flat_field.cpp`, adding only `1.65 s` in `motes-self/full-self` mode and `2.57 s` in 3-pass `motes-iter` mode). Table 4 summarizes stage-by-stage pure-CPU (16 OpenMP threads, zero GPU) runtimes (70–90 s end-to-end).

Notably, **registration across all 21 frames executes in `33 milliseconds`** (`1.6 ms/frame`), replacing the 2020 Python `fftconvolve` / $[-40, 40]^2$ `DifferenceCorrelator` bottleneck by $> 10^3\times$, while full 6×6 2D Lanczos-3 warping + per-pixel iterative Winsorized Sigma Clipping over 7.72×10^8 interpolated channel samples completes in `22.55 s` via a `4,096`-subdivision L1-cached Lanczos lookup table and zero-allocation register insertion sorts.

7.2 Agentic Compute & Token Budget

Table 6 summarizes the total language-model orchestration footprint across both development sessions (the 10-iteration core stacking campaign plus the synthetic flat-fielding campaign) and all 19 subagents (forensics, torture-testing, coding, and isolated Red-Team Critic audits). Spawning fresh, stateless Critic subagents for each com-

Table 6: Total agentic compute and token footprint across both sessions and all 19 subagents (core stacking + synthetic flat-fielding).

Orchestration Resource Metric	Total Footprint
Active Sessions / Total Subagents Spawned	2 sessions / 19 subagents
Total LLM Reasoning & Tool Turns	1,461 turns
Total Generated (Output) Tokens	$\sim$ 1.44M tokens
Cumulative Prompt-Cached Input Tokens	$\sim$ 261M tokens
Fresh Isolated Critic Audit Input Tokens	$\sim$ 15M tokens (5.7%)
C++20 Engine Hardware (No GPU)	16-thread CPU (70–90 s/run)

mit kept cumulative audit context to only $\sim$ 15M input tokens ($<$ 6% of total prompt-cached input), preventing context-window bloat while preserving strict audit independence.

7.3 Key Scientific & Methodological Conclusions

1. **Point-Source Geometric Hashing & Sharp Resampling Defeat Coherent Sensor Systematics:** Extracting background-subtracted stellar centroids ($<$ 0.2% of pixels) and matching chirality-preserved Valdes triangle invariants decouples astrometry from zero-lag master-dark estimation variance, optical vignetting bowls, worm-gear velocity reversals, and celestial field rotation. Meanwhile, the empirical Gate 2a disqualification of the 2020 manual ROI + bilinear stack ($\rho_{1\text{px}} = 0.55$, 164 RGB warm-pixel streaks on M31) confirms that partition-of-unity Lanczos-3 resampling and anti-smoothing gates are indispensable.

2. **Scorer Blind Spots Can Create Shallow Local Minima Without Breaking Alignment:** Premature spatial CFA repair (`Iter 1-2`) regressed against a shallow artificial local minimum caused by 4-pixel turnaround warm-pixel smears evading the scorer's 1-pixel hotspot filter; yet reaching the true deeper optimum still required solving the underlying physics via orthogonal 1D temporal Winsorized Sigma Clipping (`Iter 7`), which unlocked clean pure-sensor calibration in `Iter 8` (`+0.56` combined).

3. **Blind Physics Referees + Isolated Critics Prevent Goodhart's Law:** Pairing an immutable, torture-tested photometric referee (`Gates 0-5`) with stateless Red-Team Critic subagents (all 7 winning commits passed `AUDIT_PASS`, zero attempted metric-gaming in this run; cf. [1, 2]) achieved `+3.51` training gain and `+3.44` zero-shot out-of-distribution generalization without a single image ever entering an LLM context window.

4. **Human-in-the-Loop Direction Unlocks Out-of-Box Breakthroughs:** Automated hill-climbing independently converged on the industry-standard astrophotography stacking core (analogous to PixInsight, Siril, and DeepSkyStacker) running purely on CPU (16 OpenMP threads, zero GPU, 70–90 s), yet remained bounded by the objective box of its locked referee. While this autonomous trajectory converged on established industry practices, formal head-to-head quantitative benchmarking—comparing both photometric fidelity and wall-clock execution speed directly against PixInsight, Siril, and DeepSkyStacker on identical raw datasets—remains essential future work to definitively substantiate the engine's high-performance claim. The single largest visual breakthrough—estimating a blind cross-target synthetic flat field from M57 (`ring_set`) to flatten M31 (`andromeda_set`) from 32.05% $\rightarrow$ +1.00% vignetting and collapse dust-mote correlation by 93%—required human domain intuition to step outside the loop and introduce a refined diagnostic evaluator (`evaluate_flat_motes.py`). This demonstrates that AI's distinctive scientific strength here lies in rapidly synthesizing tailored, deterministic, auditable algorithms for dataset-specific quirks—such as blind session flats, stencil-isolated partial-row `.CR2` salvage, and silent `LibRaw` predictor-underflow gating.

References

[1] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Ortega, and S. Legg, "Specification gaming: the flip side of AI ingenuity," *DeepMind Safety Research Blog*, Apr. 2020.

[2] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger, "Defining and characterizing reward hacking," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 9460–9471, 2022.

[3] H. S. Malvar, L.-W. He, and R. Cutler, "High-quality linear interpolation for demosaicing of Bayer-patterned color images," in *Proc. IEEE ICASSP*, vol. 3, pp. 485–488, 2004.

[4] F. G. Valdes, L. E. Campusano, J. D. Velasquez, and P. B. Stetson, "FOCAS automatic catalog matching algorithms," *Publ. Astron. Soc. Pac.*, vol. 107, pp. 1119–1128, 1995.

[5] J. F. Hamilton and J. E. Adams, "Adaptive color plane interpolation in single sensor color electronic camera," U.S. Patent 5,629,734, May 1997.

[6] R. Lupton, M. R. Blanton, G. Fekete, D. W. Hogg, W. O'Mullane, A. Szalay, and N. Wherry, "Preparing red-green-blue images from CCD data," *Publ. Astron. Soc. Pac.*, vol. 116, pp. 133–137, 2004.

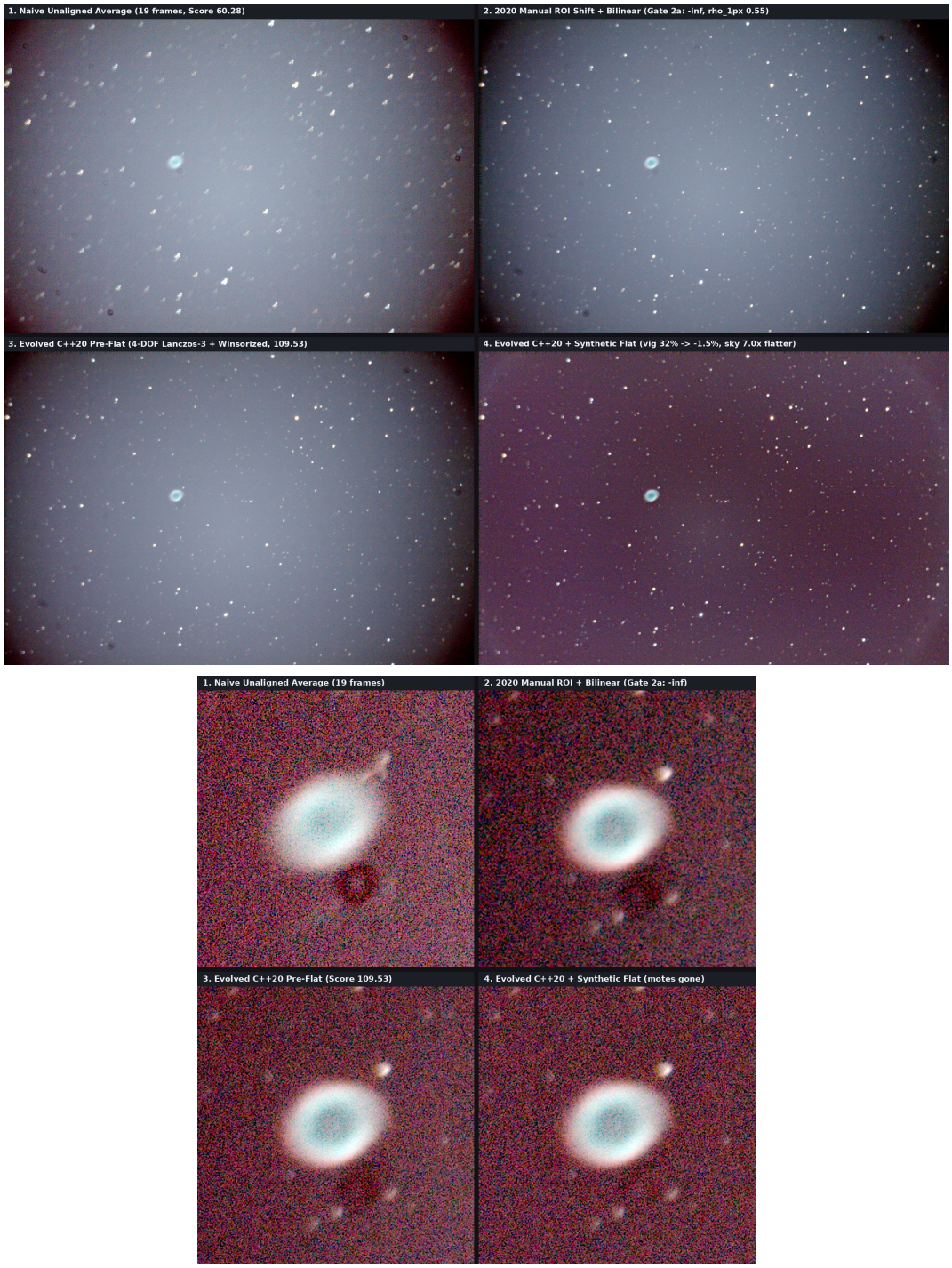

Figure 7: Four-stage end-to-end visual progression on the Ring Nebula (`ring_set`, 19 exposures, identical per-panel `asinh` stretch). **Top:** full-field progression from (1) naive unaligned average (Score `60.28`) through (2) 2020 manual ROI shift + bilinear (`naive_2020_roi`, Gate 2a: `-inf`, $\rho_{1\text{px}} = 0.55$, 22 warm impulses) and (3) evolved C++20 pre-flat stack (`b447079`, Score `109.53`) to (4) evolved C++20 + synthetic flat-field stack (`iter_06_wiener_final`, 7.0× flatter sky, 100% of $> 2\%$ dust-mote area removed). **Bottom:** 360×360 px core zoom centered on M57 at $(1577, 1312)$, illustrating the bilinear noise clumping and lingering dust-mote shadow below M57 in (2) vs. the crisp ionized ring and clean background in (4).