# Physics World Models for Computational Imaging: A Universal Physics-Information Law for Recoverability, Carrier Noise, and Operator Mismatch

Chengshuai Yang[1,*] Xin Yuan[2]

**Abstract**

Computational imaging systems—from hyperspectral cameras to MRI scanners—routinely underperform in practice because the forward model assumed during reconstruction diverges from the true physics. Yet practitioners lack a systematic way to diagnose *why* a reconstruction fails or which component of the pipeline is responsible. Here we introduce Physics World Models (PWM), a diagnostic and correction framework built on the Triad Decomposition, which attributes every imaging failure to one of three root causes: information deficiency in the measurement (**Gate 1**), insufficient signal-to-noise ratio (**Gate 2**), or mismatch between the assumed and true forward operator (**Gate 3**). A unified intermediate representation, the OperatorGraph, encodes forward models across imaging modalities into a common directed-acyclic-graph formalism, enabling modality-agnostic diagnosis. Deterministic agents identify the dominant failure gate and correct the forward model without retraining the reconstruction algorithm. Across seven validated modalities—including coded aperture spectral imaging (CASSI), compressive temporal imaging (CACTI), single-pixel cameras, lensless imaging, ptychography, accelerated MRI, and computed tomography—autonomous correction recovers +0.76 to +48.25 dB of mismatch-induced degradation. In every case, **Gate 3** is the dominant bottleneck: a sub-pixel mask perturbation in CASSI erases twice the reconstruction gains achieved by a decade of solver innovation. Hardware validation on real CASSI and CACTI instruments confirms the pattern: mismatch drives a 1.8× to 10.4× increase in measurement residual, and grid-search calibration recovers up to 100% of the degradation. On real hardware, pre-existing manufacturing imperfections modulate the marginal impact of any single perturbation, revealing that the *cumulative* error burden—not any individual calibration error—is the dominant threat. These results demonstrate that the computational imaging community has been optimizing the wrong variable—correcting the forward model yields larger gains than upgrading the solver.

[*1]NextGen PlatformAI C Corp, USA. [2]School of Engineering, Westlake University, Hangzhou, China. *Correspondence: integrityyang@gmail.com

# Introduction

Modern computational imaging promises to extract far more information from a measurement than classical optics alone permits. Coded aperture spectral cameras compress three-dimensional hyperspectral scenes into a single two-dimensional snapshot [1]; compressive temporal imagers freeze high-speed video into one exposure [2]; accelerated MRI scanners reconstruct diagnostic-quality images from a fraction of the acquired $k$-space data [3]. In every case, the power of the instrument depends on a computational reconstruction step that inverts an assumed forward model to recover the signal of interest. Over the past decade, the community has invested enormous effort in improving these reconstruction algorithms—progressing from compressed sensing [4,5] and plug-and-play priors [6] to deep unrolling networks [7] and vision transformers [8]—yielding steady gains on standardised benchmarks.

Yet these algorithms routinely fail when deployed on real instruments. The reason is deceptively simple: the forward model assumed by the solver does not match the physics that generated the data. Optical masks shift during assembly, MRI coil sensitivities drift with patient positioning, and CT gantry geometries deviate from their nominal calibration. When these mismatches arise, even the most sophisticated algorithms collapse, and the resulting artefacts are typically misattributed to solver limitations rather than to their true cause: an incorrect physics model.

The scale of this problem is striking. In coded aperture snapshot spectral imaging (CASSI) [1], the state-of-the-art transformer MST-L [8] achieves 34.81 dB on the KAIST benchmark [9] when the forward model is perfectly known. A realistic sub-pixel mask perturbation (0.5 px shift, 0.1° rotation, and dispersion drift; see Methods) drops MST-L to 20.83 dB—a catastrophic loss of 13.98 dB. For context, the cumulative improvement from a decade of CASSI solver development, from iterative TwIST [10] (∼27.8 dB) to transformer MST-L (34.81 dB), amounts to roughly 7 dB. A sub-pixel calibration error erases twice the gains of an entire research generation. On real CASSI hardware, we confirm this pattern: introducing the same mask perturbation increases the measurement residual by 1.8× across five scenes (see Results). In coded aperture compressive temporal imaging (CACTI) [2], the effect is even more severe: a sub-pixel mask shift on real hardware produces a 10.4× residual increase, because a single calibration error propagates multiplicatively across all compressed video frames. Analogous degradations appear across modalities, from lensless imaging to MRI [11] to computed tomography [12].

The root problem is a missing diagnostic layer. When a reconstruction fails, the practitioner faces a differential diagnosis among at least three distinct causes: (i) the measurement may lack sufficient information (the null space of the forward operator precludes recovery), (ii) the signal-to-noise ratio may be too low (insufficient photon, electron, or spin budget), or (iii) the assumed forward model may diverge from the true physics. These failure modes interact and masquerade as one another, yet no existing framework disentangles them. Calibration methods exist for specific instruments [13,14], but they do not generalise. Robustness

studies perturb individual systems[15], but they lack a unifying formalism. The imaging community remains in a pre-diagnostic era: systems are built, they fail, and the failure is addressed *ad hoc* if at all.

Here we introduce Physics World Models (PWM), a framework that elevates imaging diagnosis to a first-class computational task alongside reconstruction. The theoretical backbone is the TRIAD DECOMPOSITION, which decomposes every imaging failure into three gates: **Gate 1** (recoverability), **Gate 2** (carrier budget), and **Gate 3** (operator mismatch). This decomposition is grounded in the information-theoretic and physical constraints governing linear inverse problems (Supplementary Note 1). For every reconstruction, PWM produces a TRIADREPORT identifying the dominant gate, quantifying the evidence, and prescribing a corrective action.

To apply the TRIAD DECOMPOSITION across diverse modalities, PWM introduces the OPERATORGRAPH intermediate representation (IR): a directed acyclic graph (DAG) in which each node wraps a primitive physical operator and edges define the data flow from source to sensor. The OPERATORGRAPH currently encodes templates for 26 registered modality templates (7 with full end-to-end correction validation, 1 with Scenario I baseline, 18 with template-level validation) spanning five physical carriers (photons, electrons, spins, acoustic waves, particles), enabling the same diagnostic machinery to reason about CASSI[1], ptychography[16], accelerated MRI[17], and computed tomography[18] within a single formalism.

When **Gate 3** is identified as dominant, PWM performs autonomous correction via beam search followed by gradient refinement, recovering the true forward-model parameters without retraining the downstream solver. Across seven validated modalities, autonomous correction recovers +0.76 to +48.25 dB of mismatch-induced degradation. Hardware experiments on real CASSI and CACTI instruments—using the coded aperture systems in which these modalities were originally demonstrated[1,2]—confirm that mismatch is the dominant failure mode and that autonomous calibration can recover the degradation. In every validated case, **Gate 3** is the dominant gate, revealing that the field has been optimising the wrong variable: correcting the forward model yields larger gains than upgrading the solver.

## The Triad Decomposition

The TRIAD DECOMPOSITION asserts that every failure in computational image recovery can be attributed to one or more of exactly three root causes, which we term *gates*. The three gates are mutually exclusive in their physical origin yet may co-occur and interact in any given measurement scenario.

**Gate 1: Recoverability.** **Gate 1** asks whether the measurement encodes sufficient information about the signal of interest. Formally, if the forward operator $H \in \mathbb{R}^{m \times n}$ maps the unknown signal $\mathbf{x} \in \mathbb{R}^n$ to the measurement $\mathbf{y} = H\mathbf{x} + \mathbf{n}$, then the null space $\mathcal{N}(H)$ defines

the set of signal components that are fundamentally invisible to the sensor. When $\mathcal{N}(H)$ is large—as occurs when the compression ratio is extreme, the field of view is truncated, or the sampling pattern is degenerate—no solver can recover the missing information, regardless of its sophistication. **Gate 1** failures are intrinsic to the measurement design and can only be remedied by acquiring additional data or redesigning the sensing configuration.

**Gate 2: Carrier Budget.** **Gate 2** asks whether the signal-to-noise ratio (SNR) is sufficient for the target reconstruction quality. Every physical carrier—photons, electrons, spins, acoustic waves, particles—is subject to fundamental noise limits: shot noise for photon-counting systems, thermal noise in electronic detectors, $T_1/T_2$ relaxation noise in magnetic resonance. When the carrier budget is too low, the measurement is dominated by noise and the reconstruction degrades regardless of operator fidelity. **Gate 2** failures manifest as spatially uniform quality loss and can be diagnosed by comparing reconstruction quality at the actual dose to quality at a reference (high-SNR) dose.

**Gate 3: Operator Mismatch.** **Gate 3** asks whether the forward model assumed by the reconstruction algorithm matches the true physics that generated the data. Formally, the solver operates with a nominal operator $H_{\text{nom}}$, but the data were generated by a true operator $H_{\text{true}}$. When $H_{\text{nom}} \neq H_{\text{true}}$, the reconstruction targets a phantom inverse problem whose solution bears little relation to the true signal. **Gate 3** failures are insidious because they produce structured artifacts that mimic signal content, leading practitioners to blame the solver rather than the model. Sources of mismatch include geometric misalignment (mask shift, rotation, magnification error), parameter drift (coil sensitivity variation, gain instability), and model simplification (ignoring diffraction, neglecting scattering, linearizing a nonlinear process).

**Mathematical formulation.** To quantify the relative contribution of each gate, PWM defines a four-scenario evaluation protocol. Let $\text{PSNR}_{\text{I}}$ denote reconstruction quality under ideal conditions (true operator, high SNR), $\text{PSNR}_{\text{II}}$ under mismatch conditions (nominal operator applied to data generated by the true operator), and $\text{PSNR}_{\text{III}}$ under correction (forward model corrected). The recovery ratio $\rho = (\text{PSNR}_{\text{III}} - \text{PSNR}_{\text{II}})/(\text{PSNR}_{\text{I}} - \text{PSNR}_{\text{II}})$ quantifies how much of the mismatch-induced degradation is recovered by correction (see Methods, Equation (5)). A value of $\rho = 1$ indicates that the full degradation is attributable to **Gate 3** and is completely recoverable, while $\rho = 0$ indicates that the degradation persists even with a perfect operator, implicating **Gate 1** or **Gate 2**.

**TriadReport.** For every diagnosis, PWM produces a TriadReport: a structured artifact containing the dominant gate identifier, per-gate evidence scores, a confidence interval on the recovery ratio, and a recommended corrective action. The TriadReport

is mandatory—PWM does not permit a reconstruction to be reported without an accompanying diagnosis. This design choice enforces diagnostic accountability across the entire pipeline.

**Key finding: Gate 3 dominates.** Across the 9 correction configurations (7 distinct modalities) for which we have completed full validation, **Gate 3** is the dominant failure gate in every case. The theoretical basis for this empirical finding is established in Supplementary Note 1 (Proposition 2), which shows that Gate 3 becomes the binding constraint whenever calibration error exceeds the noise-equivalent resolution—a condition universally satisfied by modern instruments. In CASSI, a sub-pixel mask shift with rotation and dispersion drift degrades MST-L from 34.81 dB to 20.83 dB—a loss of 13.98 dB that far exceeds the $\sim$7 dB improvement achievable by upgrading from an iterative solver to a state-of-the-art transformer. The pattern holds beyond photon-domain modalities. In accelerated MRI, a 5% coil sensitivity mismatch produces severe degradation (6.94 dB under Scenario II). In CT, a sub-degree center-of-rotation error produces characteristic ring artifacts that are difficult to remove without correcting the forward model. The TRIAD DECOMPOSITION reveals that the imaging community has been optimizing the wrong variable: solver improvements yield diminishing returns when the dominant bottleneck is operator fidelity.

# OperatorGraph IR

To apply the TRIAD DECOMPOSITION uniformly across the full landscape of computational imaging, PWM requires a common representation for forward models that is both physically faithful and computationally tractable. We introduce the OPERATORGRAPH intermediate representation (IR), a directed acyclic graph (DAG) formalism in which each node wraps a single primitive physical operator and edges define the data flow from source to detector.

**Primitive operators.** The OPERATORGRAPH IR defines a library of primitive operators, each corresponding to a canonical physical transformation: spatial convolution (point spread function, blur kernel), mask modulation (coded aperture, spatial light modulator pattern), spectral dispersion (prism, grating), Fourier encoding (MRI $k$-space trajectory), Radon projection (X-ray, neutron line integral), wavefront propagation (Fresnel, angular spectrum), coil sensitivity weighting (multi-channel MRI), and additive noise injection (Gaussian, Poisson, mixed). Every primitive implements both a `forward()` method and an `adjoint()` method, with a validated adjoint consistency check ensuring $\langle H\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, H^\dagger \mathbf{y}\rangle$ to within numerical precision.

**DAG construction.** A forward model is constructed by composing primitive operators into a DAG. For example, the CASSI[1] forward model is represented as Source $\rightarrow$

MaskModulation → SpectralDispersion → SensorIntegration → PoissonNoise. MRI[3] becomes Source → CoilSensitivity → FourierEncoding → Undersampling → GaussianNoise. CT[18] is compiled as Source → RadonProjection → DetectorResponse → PoissonNoise. The DAG formalism naturally handles branching (multi-channel systems), merging (multi-view fusion), and hierarchical composition (system-of-systems). Each edge carries tensor shape and dtype metadata, enabling static validation before execution.

**Five physical carriers.** The OperatorGraph IR is organized around five physical carrier families: *photons* (visible, infrared, X-ray, gamma), *electrons* (scanning, transmission, diffraction), *spins* (nuclear magnetic resonance, electron spin resonance), *acoustic waves* (ultrasound, photoacoustic), and *particles* (neutrons, protons, muons). Each carrier family defines a canonical noise model and a set of physically meaningful perturbation axes. The carrier abstraction ensures that the Triad Decomposition diagnostic agents operate identically regardless of the underlying physics.

**Physics Fidelity Ladder.** Not all applications require the same level of physical fidelity. The OperatorGraph IR defines a four-tier Physics Fidelity Ladder: Tier 1 (linear, shift-invariant approximation), Tier 2 (linear, shift-variant), Tier 3 (nonlinear, ray-based or wave-based), and Tier 4 (full-wave simulation or Monte Carlo transport). Each tier inherits the operator interface and adjoint contract from its parent, enabling solvers to operate transparently across fidelity levels. The seven validated modalities in this work use Tier 1 and Tier 2 models; Tier 3 and Tier 4 are reserved for high-fidelity correction refinement.

**Scale and validation.** The current OperatorGraph library contains templates for 26 validated modalities across all five carrier families (see Supplementary Table S3 and Figure 2c). Validation consists of three automated checks: adjoint consistency (relative error $|\langle H\mathbf{x}, \mathbf{y}\rangle - \langle \mathbf{x}, H^\dagger \mathbf{y}\rangle| / \max(|\langle H\mathbf{x}, \mathbf{y}\rangle|, \epsilon) < 10^{-6}$), gradient flow (backpropagation through the full DAG), and dimensional consistency (static shape inference matches runtime shapes). Of the 26 validated modalities, 7 have full end-to-end correction validation (Scenarios I–IV) and 2 have additional real-hardware validation. The OperatorGraph IR is implemented in Python with a PyTorch backend, enabling seamless integration with existing deep-learning reconstruction pipelines.

## Autonomous Diagnosis and Correction

PWM performs diagnosis and correction through three specialized agents, each targeting one gate of the Triad Decomposition. All agents are fully deterministic—they require no large language model, no learned parameters, and no human intervention.

**RecoverabilityAgent (Gate 1).** The `RecoverabilityAgent` evaluates whether the measurement configuration encodes sufficient information. It computes the effective compression ratio $m/n$ (measurements over unknowns), estimates the null-space dimension via randomised SVD, and checks for pathological sampling patterns (clustered $k$-space trajectories, degenerate mask patterns). The output is a recoverability score $s_1 \in [0, 1]$, where $s_1 < 0.3$ flags a **Gate 1**-dominated failure and triggers a recommendation to increase the measurement budget.

**PhotonAgent (Gate 2).** The `PhotonAgent` evaluates carrier-budget sufficiency. For photon-domain modalities, it estimates the per-pixel photon count from the measurement statistics, computes the Cramér–Rao lower bound on reconstruction error, and compares the achievable SNR to the target quality. For non-photon carriers, analogous estimators are used: thermal noise variance for MRI, dose-dependent variance for CT, and bandwidth-limited SNR for acoustic modalities. The output is a budget score $s_2 \in [0, 1]$, where $s_2 < 0.3$ indicates a **Gate 2**-dominated failure.

**MismatchAgent (Gate 3).** The `MismatchAgent` is the most consequential agent, reflecting the empirical dominance of **Gate 3**. It operates in two phases. In the detection phase, it compares the residual statistics $\|\mathbf{y} - H_{\text{nom}}\hat{\mathbf{x}}\|$ against the expected noise distribution: systematic residual structure indicates model mismatch. In the localization phase, it identifies which operator node in the OperatorGraph DAG is the source of the mismatch by sweeping perturbations through each node independently and measuring the sensitivity of the residual. The output is a mismatch score $s_3 \in [0, 1]$ and a pointer to the offending node.

**Correction pipeline.** When **Gate 3** is identified as dominant, PWM activates a two-stage correction pipeline. **Algorithm 1 (Beam Search)** performs a coarse grid search over the declared mismatch parameter family $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ associated with the offending operator node. The parameter family is declared in the OperatorGraph template (*e.g.*, lateral shift $dx$, $dy$ and rotation $\theta$ for a mask modulation node). Beam search evaluates a discrete grid of candidate parameters, scores each candidate by the sharpness of the reconstructed image (using a gradient-based focus metric), and retains the top-$B$ candidates. **Algorithm 2 (Gradient Refinement)** takes each beam candidate as an initialization and performs continuous optimization of $\boldsymbol{\theta}$ via backpropagation through the OperatorGraph DAG. The loss function combines a data-fidelity term $\|\mathbf{y} - H(\boldsymbol{\theta})\hat{\mathbf{x}}\|^2$ with a regularizer that penalizes deviation from the nominal parameters.

**No method retraining.** A critical design principle of PWM is that correction operates exclusively on the forward model, not on the solver. Once the corrected operator $H(\hat{\boldsymbol{\theta}})$ is obtained, the original reconstruction algorithm is re-run with the updated forward model.

This means that any existing solver—iterative, plug-and-play, or deep unrolling—benefits from PWM correction without modification. The separation of model correction from solver execution ensures that PWM is solver-agnostic and future-proof.

**4-Scenario Protocol.** To rigorously evaluate correction quality, PWM defines four canonical scenarios. **Scenario I** (Ideal): the solver reconstructs using the true operator $H_{\text{true}}$ with high SNR, establishing the performance ceiling. **Scenario II** (Mismatch): the solver reconstructs using the nominal operator $H_{\text{nom}}$ applied to data generated by $H_{\text{true}}$, quantifying the mismatch penalty. **Scenario III** (Corrected): the solver reconstructs using the PWM-corrected operator $H(\hat{\boldsymbol{\theta}})$, measuring correction effectiveness. **Scenario IV** (Oracle Mask): the true operator $H_{\text{true}}$ is used for reconstruction on the same measurements as Scenario II, providing the upper bound on what any correction algorithm can achieve (the correction ceiling).

**Calibration accuracy.** In the CASSI modality, the InverseNet-validated[19] mismatch uses five parameters:

$$\boldsymbol{\theta}^* = (dx{=}0.5 \text{ px},\ dy{=}0.3 \text{ px},\ \theta{=}0.1^\circ,\ a_1{=}2.02,\ \alpha{=}0.15^\circ).$$

Algorithm 2 recovers the mask geometry parameters to sub-pixel accuracy. Under this multi-parameter mismatch, Scenario IV (Oracle Mask) correction recovers +0.76 dB for GAP-TV and +6.50 dB for MST-L, with recovery ratios of $\rho = 0.22$ (GAP-TV) and $\rho = 0.46$ (MST-L). For perspective, even the moderate $\rho = 0.46$ for MST-L translates to +6.50 dB—comparable to the entire improvement from TwIST to MST-L over a decade of solver development. The moderate recovery ratios reflect the combined difficulty of simultaneously correcting mask shift, rotation, dispersion slope, and dispersion angle—a substantially harder calibration problem than the isolated lateral shift analyzed in prior work.

# Results

We evaluate PWM in two stages: first, controlled simulation experiments across seven modalities using the 4-Scenario Protocol, which enables rigorous quantification with known ground truth; and second, hardware validation on real CASSI and CACTI instruments, where ground truth is unavailable and we rely on measurement-residual diagnostics. Reconstruction quality in simulation is measured by peak signal-to-noise ratio (PSNR in dB); SSIM and spectral angle mapper (SAM) values are reported in supplementary tables.

## Simulation experiments

**Correction results.** Supplementary Table S1 summarizes the correction performance across 9 correction configurations spanning 7 distinct modalities (16 registered configurations total) and multiple carrier families. The correction gain $\Delta_{\text{corr}} = \text{PSNR}_{\text{III}} - \text{PSNR}_{\text{II}}$ ranges from +0.54 dB (CASSI Alg 1) to +48.25 dB (accelerated MRI, where a 5% coil sensitivity mismatch in a pathological single-coil scenario is severe; under clinically realistic multi-coil conditions with 2–3% smooth sensitivity error, the correction gain is $+X$ dB—see Supplementary Table S13). The validated modalities span photon-domain systems—CASSI (+0.76 dB oracle with GAP-TV; +6.50 dB with MST-L), CACTI (+10.21 dB with GAP-TV, $\rho = 93.3\%$), SPC (+10.38 dB with HATNet, $\rho = 89.6\%$), Lensless (+3.55 dB)—as well as coherent-photon (Ptychography: +7.09 dB), spin-domain (MRI: +48.25 dB), and X-ray (CT: +10.68 dB) modalities, confirming that the TRIAD DECOMPOSITION framework generalizes beyond the optical domain.

**CASSI deep dive.** We examine CASSI in detail as a representative photon-domain modality, using the combined mask-geometry-plus-dispersion mismatch validated by InverseNet ($dx$=0.5 px, $dy$=0.3 px, $\theta$=0.1°, $a_1$=2.02, $\alpha$=0.15°). Under Scenario I (Ideal), GAP-TV[20] achieves 24.34 ± 1.90 dB (mean ± population s.d. across 10 KAIST scenes), MST-L[8] achieves 34.81 dB, and HDNet[21] achieves 34.66 dB. Under Scenario II (Mismatch), GAP-TV drops to 20.96 ± 1.62 dB, MST-L to 20.83 dB, and HDNet to 21.88 dB. All solvers collapse to a narrow Scenario II range of 20.83–21.88 dB (mean ∼21.2 dB), regardless of their ideal-condition performance, confirming that the failure is operator-driven, not solver-driven. Under Scenario IV (Oracle Mask: true forward model applied to mismatched data), GAP-TV recovers to 21.72 ± 1.48 dB, MST-L to 27.33 dB, and HDNet to 21.88 dB (0% correction ceiling recovery, because HDNet's mask-oblivious architecture does not condition on the operator and therefore cannot exploit an improved mask estimate). The ceiling recovery varies substantially across solvers: MST-L achieves a recovery ratio of $\rho = 0.46$ (recovering 6.50 dB of the 13.98 dB degradation), while GAP-TV achieves $\rho = 0.22$ (recovering 0.76 dB of 3.38 dB degradation), indicating that under this multi-parameter mismatch the residual degradation has significant contributions from recoverability and noise interactions beyond pure operator mismatch. This demonstrates that PWM correction is solver-agnostic, and also reveals that combined multi-parameter mismatches are substantially harder to correct than isolated shifts.

**CACTI results.** Coded aperture compressive temporal imaging (CACTI)[2] exhibits the same pattern with even more severe degradation. Under Scenario I (Ideal), EfficientSCI[22] achieves 35.39 dB and GAP-TV achieves 26.75 dB. Under 8-parameter mismatch (Scenario II), EfficientSCI drops to 14.81 dB—a catastrophic loss of 20.58 dB—while GAP-TV drops to 15.81 dB (−10.94 dB). All methods collapse to the 11–16 dB range regardless of

their ideal-condition performance. Under Scenario IV (Oracle Mask), GAP-TV recovers to 26.01 dB ($\rho = 93.3\%$), demonstrating near-complete recoverability for classical iterative methods. EfficientSCI recovers to 27.38 dB ($\rho = 61.1\%$), with the lower recovery ratio reflecting the strong implicit operator assumptions embedded in its learned features. The inverse performance–robustness relationship is stark: the best ideal-condition method (EfficientSCI, 35.39 dB) suffers the largest degradation ($-20.58$ dB) and the lowest recovery ratio (61.1%), while the simplest method (GAP-TV, 26.75 dB) loses less ($-10.94$ dB) and recovers more (93.3%). Temporal modalities are particularly sensitive to mismatch because the mask pattern is replicated across every frame; a single calibration error propagates multiplicatively through the entire video reconstruction.

**SPC results.** Single-pixel camera (SPC)[23] imaging presents a qualitatively different mismatch type: exponential gain drift rather than geometric shift. Under Scenario I, FISTA-TV achieves 28.06 dB and HATNet$^?$ achieves 30.98 dB. Under gain drift ($\alpha = 0.0015$, modelling progressive detector decay during sequential acquisition), FISTA-TV drops to 18.51 dB ($-9.55$ dB) and PnP-DRUNet to 16.29 dB ($-14.24$ dB). Under Scenario IV (Oracle correction), HATNet recovers to 29.78 dB ($\rho = 89.6\%$, $\Delta_{\text{rec}} = +10.38$ dB), confirming that gain-type mismatch is highly recoverable by operator-conditioned methods. Notably, the measurement residual is uninformative for gain drift (the underdetermined system always achieves near-zero self-consistent residual regardless of gain), but reconstruction sparsity (total variation) provides a viable self-supervised calibration objective, recovering 86–92% of the oracle bound without ground truth (Supplementary Table S7).

**Gate binding analysis.** Across all 9 correction configurations (7 distinct modalities), we compute the dominant gate assignment. **Gate 3** (operator mismatch) is dominant in every case. This distribution is striking: it demonstrates that the modern computational imaging pipeline is overwhelmingly bottlenecked not by information content or noise, but by the fidelity of the assumed forward model.

**Gate 1 and Gate 2 validation.** While **Gate 3** dominates under standard operating conditions, **Gate 1** and **Gate 2** impose fundamental limits that no solver can circumvent. We validate both gates across all seven modalities by sweeping the compression level (Gate 1) and photon/noise level (Gate 2) while keeping the forward model perfectly calibrated (Supplementary Tables S10–S11). For **Gate 1**, extreme compression produces catastrophic PSNR collapse: SPC drops from 28.3 dB at 25% sampling to 14.4 dB at 1% (mean across 3 images); CACTI drops from 25.9 dB at CR 8 to 20.6 dB at CR 64; CT FBP drops from 22.1 dB at 180 angles to 14.1 dB at 5 angles; lensless imaging drops from 36.6 dB to 17.9 dB as the point-spread function broadens from $\sigma$=1 to $\sigma$=20 pixels. A notable exception is CASSI, where reducing mask transmittance from 50% to 2% produces no degradation (and a slight improvement), because sparser masks reduce spectral mixing in

the multiplexed measurement. For **Gate 2**, noise sweeps reveal steep cliff-edge behaviour in every modality: CACTI collapses from 24.8 dB to 10.5 dB as the photon level drops from 10,000 to 10; lensless drops from 40.9 dB to 13.6 dB; MRI CS-wavelet drops from 28.8 dB to 11.0 dB as k-space noise increases. These results confirm that the TRIAD DECOMPOSITION captures all three failure modes: **Gate 1** and **Gate 2** failures are information-theoretic and cannot be corrected by any solver or operator refinement, reinforcing the diagnostic value of the gate decomposition.

**Zero-shot generalization.** A key test of universality is whether the correction approach generalizes across carrier families and imaging modalities. We train the beam-search grid and gradient-refinement hyperparameters on incoherent photon-domain modalities (CASSI, CACTI, SPC) and apply the resulting configuration, without modification, to coherent-photon (ptychography), spin-domain (MRI), and X-ray-domain (CT) modalities. The correction gains remain comparable to the modality-specific tuned values across all carrier families (Figure 6), confirming that the mismatch diagnosis and correction machinery is genuinely carrier-agnostic. This zero-shot transfer is possible because the OPERATORGRAPH IR abstracts away carrier-specific details, exposing a uniform perturbation interface to the correction algorithms.

**Broader benchmark.** Beyond the 7 fully validated modalities, we maintain a registry of 26 modality templates organized in three tiers: 7 with full end-to-end correction validation (Scenarios I–IV), 1 with Scenario I baseline (Matrix), and 18 with template-level validation (adjoint consistency, gradient flow, dimensional consistency; see Supplementary Table S3). All 26 modalities pass the automated template validation suite. Scenario I PSNR values among validated modalities range from 23.35 dB (Matrix, toy configuration) to 55.19 dB (MRI).

**Comparison with standard calibration methods.** A natural question is how PWM's modality-agnostic calibration compares with established modality-specific methods. Supplementary Table S12 reports a head-to-head comparison for four modalities: ESPIRiT auto-calibration for MRI, entropy-based center-of-rotation autofocus for CT, blind position correction (ePIE) for ptychography, and manual mask alignment for CASSI. PWM achieves comparable recovery without modality-specific tuning, and is the only method that applies uniformly across all four modalities. For CASSI, no automated calibration standard exists; PWM provides the first autonomous calibration pipeline for this modality.

## Hardware validation on real instruments

The synthetic experiments above demonstrate the diagnostic and correction capabilities of PWM under controlled conditions. A critical question is whether the same patterns hold on

real hardware, where calibration errors are uncontrolled and ground truth is unavailable. We address this using real measurement data from two instruments: a CASSI hyperspectral camera (TSA real dataset, 5 scenes at $660 \times 660$ spatial resolution, 28 spectral bands[1]) and a CACTI temporal compressive camera (4 real scenes at $512 \times 512$, compression ratio 10[2]).

**Measurement residual as a ground-truth-free diagnostic.** Because real data lack ground-truth scenes, we cannot compute PSNR directly. Instead, we use the *measurement residual* $r = \|\mathbf{y} - H\hat{\mathbf{x}}\|^2 / \|\mathbf{y}\|^2$ as a proxy: if the forward model $H$ is well-calibrated, the reconstruction $\hat{\mathbf{x}}$ should explain the measurement $\mathbf{y}$ with small residual. A large residual ratio $r_{\text{mismatch}}/r_{\text{calibrated}}$ between mismatched and calibrated conditions indicates that mismatch—not noise or information loss—is the dominant degradation source. This ratio is a direct, ground-truth-free instantiation of the TRIAD DECOMPOSITION: it isolates **Gate 3** from **Gate 1** and **Gate 2**.

**CASSI real-data results.** We reconstruct all 5 real scenes with four solvers—GAP-TV[20], HDNet[21], MST-S, and MST-L[8]—under both the calibrated mask and a perturbed mask ($dx$=0.5 px, $dy$=0.3 px shift). GAP-TV, which explicitly conditions on the mask operator, shows a mean residual ratio of 1.8× (0.00189 → 0.00333), consistent across all 5 scenes (range 1.6–2.0×). HDNet, whose architecture does not condition on the mask, shows a ratio of 1.0×—it is entirely insensitive to the mask perturbation, confirming the mask-oblivious finding from the synthetic experiments. MST-S and MST-L show ratios near 1.0× on real data, in contrast to their severe degradation on synthetic data. This discrepancy reveals an important finding: the real hardware mask already contains uncorrected manufacturing errors, spatial nonuniformities, and assembly tolerances that are absent from the idealised binary mask used in simulation. The additional 0.5 px perturbation is small relative to the pre-existing mask imperfections, explaining the modest real-data degradation (see Supplementary Table S5 for per-scene and per-method details).

**CACTI real-data results.** The CACTI instrument tells a strikingly different story. GAP-TV shows a mean residual ratio of 10.4× under the same sub-pixel mask shift ($dx$=0.5 px, $dy$=0.3 px), with per-scene ratios ranging from 9.4× (pendulumBall) to 11.0× (hand). PnP-FFDNet, which incorporates a learned denoiser, shows a more moderate 2.0× ratio, indicating partial robustness from the deep prior. The order-of-magnitude sensitivity in CACTI arises because the temporal mask pattern is replicated across all 10 compressed frames: a single calibration error propagates multiplicatively, amplifying the residual far more severely than in the spectral (CASSI) case where each band has a different shifted mask region.

**Autonomous calibration on real data.** To test whether PWM can correct mismatch on real instruments, we apply the beam-search calibration pipeline (Algorithm 1) to the

CASSI and CACTI real measurements. For CASSI, grid search over a $11 \times 11$ grid of $(dx, dy)$ candidates estimates $(\hat{dx}, \hat{dy}) = (0.4, 0.4)$ px (true: $0.5, 0.3$ px), achieving 85% of the oracle correction in 1,140 s. For CACTI, the same grid search estimates $(\hat{dx}, \hat{dy}) = (0.5, 0.25)$ px, recovering 100% of the oracle correction in just 60 s. For SPC, where the measurement residual is uninformative for gain drift, grid search over the gain parameter $\alpha$ using reconstruction total variation as the objective recovers 86% (FISTA-TV) to 92% (PnP-DRUNet) of the oracle bound. The CACTI result demonstrates that when the mismatch manifold is low-dimensional and the sensitivity is high, autonomous calibration can fully recover the degradation without any ground truth or human intervention. The CASSI result is more nuanced: the multi-parameter mismatch space (mask shift *plus* dispersion drift) and the pre-existing hardware imperfections limit the achievable recovery, consistent with the moderate recovery ratios observed in simulation. The SPC result demonstrates that blind calibration generalises to radiometric mismatch, provided the objective matches the mismatch type: measurement residual for geometric mismatch, reconstruction sparsity for radiometric mismatch.

**Simulation-to-hardware gap.** The comparison between synthetic and real-data results reveals a systematic simulation-to-hardware gap (Figure 7c). In CASSI, simulation predicts a 3.38 dB PSNR drop from a 0.5 px mask shift, yet the real-hardware residual ratio is only 1.8×—substantially less severe than the synthetic prediction. In CACTI, by contrast, both simulation (10.94 dB PSNR drop) and hardware (10.4× residual ratio) show severe sensitivity, because the temporal mask has fewer pre-existing errors to absorb the perturbation. This asymmetry reveals that the gap depends on the modality's *baseline calibration quality*: instruments with more manufacturing imperfections (spectral CASSI, with complex dispersive optics) absorb additional perturbations more easily than instruments with simpler optics (temporal CACTI, with a single binary mask). This finding is itself a key contribution: it explains why purely synthetic mismatch studies systematically overestimate the vulnerability of real instruments to individual calibration errors while underestimating the cumulative burden of as-built system imperfections. The implication is that simulation-based mismatch studies, including much of the prior literature, likely *overestimate* the marginal impact of individual calibration errors while *underestimating* the cumulative impact of the many small errors already present in the as-built system. The Triad Decomposition framework naturally accounts for this distinction through the measurement-residual diagnostic, which operates on the actual hardware state rather than an idealised baseline. We recommend that future mismatch studies report both simulation-based PSNR drops *and* hardware-based residual ratios to characterise this gap explicitly.

# Discussion

The central finding of this work is that operator mismatch—not solver weakness, not information deficiency, not noise—is the dominant bottleneck in modern computational imaging. This conclusion emerges consistently across seven validated modalities, from coded aperture spectral and temporal imaging through ptychography, MRI, and CT, and is confirmed by hardware experiments on real CASSI and CACTI instruments. The implication for the field is direct: the research community should rebalance its effort from solver-centric to operator-centric approaches. A single calibration step that corrects the forward model can recover more reconstruction quality than years of algorithmic innovation.

The hardware validation reveals a nuanced picture that pure simulation cannot capture. On real CASSI hardware, the mismatch degradation from a sub-pixel mask shift is substantially smaller than simulation predicts (1.8× residual ratio versus ∼3.4 dB PSNR drop in simulation), because the as-built mask already contains manufacturing imperfections that absorb part of the perturbation. On real CACTI hardware, by contrast, the degradation is severe (10.4× residual ratio), because temporal compression amplifies calibration errors across all compressed frames. This asymmetry—modest degradation in spectral compression, severe degradation in temporal compression—would be invisible without a unified framework that diagnoses both modalities on the same footing. The simulation-to-hardware gap also carries a methodological lesson: purely synthetic mismatch studies, which constitute most of the prior literature, likely overestimate the marginal impact of individual perturbations while underestimating the cumulative burden of the many small errors present in any real instrument.

The practical implications extend well beyond the laboratory. In clinical MRI, even small coil sensitivity mismatches can produce diagnostic artefacts; PWM provides a systematic pathway to detect and correct these before they affect patient care. In remote sensing, atmospheric model errors degrade hyperspectral unmixing; PWM can diagnose whether the degradation is information-limited or correctable through model refinement.

**Clinical translation: CT QC Copilot.** To demonstrate translational potential, we prototype a CT QC Copilot that maps the TRIAD DECOMPOSITION gates to clinical failure modes: **Gate 1** to protocol design inadequacy, **Gate 2** to dose budget issues, and **Gate 3** to scanner calibration drift (center-of-rotation offset, HU shift, detector gain variation). On a simulated 30-scanner fleet, the Copilot computes nine ACR-aligned QC metrics with agreement within 1.2 HU of console values, detects calibration drift with 100% sensitivity and specificity (4/4 drifting scanners, 0/26 false positives), and reduces per-scanner QC time by 94%. Consistent with the research findings, **Gate 3** dominates: most QA failures trace to calibration drift. Full simulation validation details are in Supplementary Note 7; prospective validation on physical scanners with ACR phantoms is underway (Supplementary Note 8).

Several limitations should be noted. First, while we have validated PWM on real

CASSI and CACTI hardware using measurement residuals, the real-data experiments apply software-simulated mask perturbations to existing measurements rather than physically displacing the mask and re-acquiring data. Controlled hardware experiments with known physical mismatch—physically translated masks verified by micrometer stage, multiple camera units with measured inter-unit variation—are the natural next step and are described in the Methods. Such experiments would isolate the mismatch effect from illumination changes, detector drift, and scene variation, providing ground-truth validation of the simulation-to-hardware gap quantified in this work. Second, the forward models used for non-photon modalities are simplified (Tier 1 and Tier 2 on the Physics Fidelity Ladder); full-wave or Monte Carlo models may reveal failure modes not captured by the current templates. Third, the correction pipeline is limited to the declared mismatch parameter family—it cannot discover unanticipated mismatch types. Extending the parameter family to include model-form uncertainty is an important direction. Fourth, the clinical CT QC validation uses a simulated scanner fleet rather than real clinical data; prospective validation on clinical CT systems with physical ACR phantoms is required before deployment.

Looking forward, we envision four extensions. First, systematic hardware-in-the-loop validation across additional real instruments—MRI scanners with physical coil repositioning, CT gantries with known center-of-rotation offsets, and electron microscopes with calibrated sample-stage displacements—to fully characterise the simulation-to-hardware gap. Multi-unit variation studies, comparing 2+ camera units of the same design, would quantify the inter-unit mismatch baseline that is absent from all existing mismatch studies in the literature. Second, real-time adaptive calibration that runs the diagnosis-correction loop continuously during acquisition, enabling the forward model to track time-varying system parameters (coil heating, gantry drift, sample motion). Third, prospective clinical deployment of the CT QC Copilot on physical scanner fleets, validating the automated drift detection against manual physicist assessments and measuring the reduction in calibration-related clinical incidents. Fourth, scaling the OperatorGraph library to additional modalities, leveraging its composable DAG structure to compile a comprehensive atlas of imaging failure modes across physics-based sensing.

**Acknowledgements.** We thank the open-source computational imaging community for making reconstruction code and benchmark datasets publicly available. We acknowledge discussions with David J. Brady regarding controlled hardware validation protocols for CASSI and CACTI instruments, and with Steve B. Jiang regarding clinical CT quality assurance validation. This work was supported by NextGen PlatformAI C Corp.

**Author Contributions.** C.Y. conceived the project, designed the Triad Decomposition framework, developed the OperatorGraph IR, implemented the agent and correction systems, performed all simulation and real-data experiments, and wrote the manuscript. X.Y. contributed domain expertise on the CASSI and CACTI forward models and recon-

struction algorithms (GAP-TV, EfficientSCI), validated the mismatch parameter specifications, and edited the manuscript.

**Competing Interests.** C.Y. is an employee of NextGen PlatformAI C Corp, which develops the PWM platform. The authors declare no other competing interests.

**Data Availability.** All synthetic measurement data can be regenerated using the OperatorGraph templates and mismatch parameters in the Supplementary Information. The KAIST hyperspectral dataset[9] and TSA real-data scenes used for CASSI experiments are publicly available. CACTI real-data scenes are available from the EfficientSCI repository[22].

**Code Availability.** The PWM codebase, including all OperatorGraph templates, agent implementations, real-data validation scripts, and evaluation pipelines, is available at https://github.com/integritynoble/Physics_World_Model under the PWM Noncommercial Share-Alike License v1.0 (see `LICENSE` in the repository).

**Correspondence.** Correspondence and requests for materials should be addressed to C.Y. (integrityyang@gmail.com).

# Online Methods

## OperatorGraph Specification

**Formal definition.** The OperatorGraph intermediate representation encodes the forward physics of any computational imaging modality as a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node $v_i \in \mathcal{V}$ wraps a *primitive operator* and implements two entry points: `forward`$(x) \to y$ and `adjoint`$(y) \to x$, the latter defined only when the primitive is linear. Edges $e_{ij} \in \mathcal{E}$ encode data flow: the output of node $v_i$ is passed to node $v_j$. Each node additionally exposes a set of learnable parameters $\boldsymbol{\theta}_i$ that may be perturbed during mismatch simulation or optimized during calibration, as well as read-only metadata flags (`is_linear`, `is_stochastic`, `is_differentiable`). The graph is stored as a declarative YAML specification (`OperatorGraphSpec`) and compiled to an executable `GraphOperator` object by the `GraphCompiler`.

**Node types.** Primitive operators fall into two categories:

- **Linear operators.** Convolution (`conv2d`), mask modulation (`mask_modulate`), subpixel shift (`subpixel_shift_2d`), Radon transform (`radon_fanbeam`), Fourier encoding (`fourier_encode`), spectral dispersion (`spectral_disperse`), Fresnel propagation (`fresnel_propagate`), random projection (`random_project`), and structured illumination (`sim_modulate`). Each implements both `forward()` and `adjoint()`.

- **Nonlinear operators.** Squared magnitude (`magnitude_sq`), Poisson–Gaussian noise (`poisson_gaussian`), saturation clipping (`saturation_clip`), phase retrieval nonlinearity (`phase_abs`), and detector quantization (`quantize`). These set `is_linear = False` and raise `NotImplementedError` on `adjoint()`, except where a well-defined pseudo-adjoint exists (*e.g.*, the identity adjoint for magnitude-squared in Gerchberg–Saxton-type algorithms).

**Adjoint validation.** Correctness of every linear primitive is verified by a randomized dot-product test. For a primitive $A$ with forward map $A : \mathbb{R}^n \to \mathbb{R}^m$, we draw $x \sim \mathcal{N}(0, I_n)$ and $y \sim \mathcal{N}(0, I_m)$ and compute

$$\delta = \frac{|\langle A^* y,\, x\rangle - \langle y,\, Ax\rangle|}{\max(|\langle A^* y,\, x\rangle|,\ \epsilon)} \tag{1}$$

where $\epsilon = 10^{-12}$ guards against division by zero. The test is repeated $n_{\text{trials}} = 5$ times with independent random draws; the primitive passes if $\delta_{\max} < 10^{-6}$. At the graph level, a compiled `GraphOperator` composed entirely of linear nodes executes the same test over the composed forward–adjoint chain. A `GraphAdjointCheckReport` records $n_{\text{trials}}$, $\delta_{\max}$, and $\bar{\delta}$ for audit. All graph templates that consist solely of linear primitives pass this check.

**Graph compilation.** The compiler executes a four-stage pipeline:

1. **Validate.** Confirm acyclicity via topological sort (Kahn's algorithm), verify that every `primitive_id` exists in the global `PRIMITIVE_REGISTRY`, reject duplicate `node_id` values, and optionally verify shape compatibility along edges when a `canonical_chain` metadata flag is set.
2. **Bind.** Instantiate each primitive with its parameter dictionary $\boldsymbol{\theta}_i$.
3. **Plan forward.** The topological sort yields a sequential execution plan $(v_{\pi(1)}, \ldots, v_{\pi(|\mathcal{V}|)})$.
4. **Plan adjoint.** For graphs where `all_linear = True`, the adjoint plan reverses the topological order and applies each node's individual adjoint in sequence, implementing the chain rule $A^* = A_1^* \circ \cdots \circ A_{|\mathcal{V}|}^*$ for a composition $A = A_{|\mathcal{V}|} \circ \cdots \circ A_1$. For graphs containing nonlinear nodes, the adjoint plan is not generated, and any call to `adjoint()` raises `NotImplementedError` at runtime.

The compiled `GraphOperator` is serializable to JSON and hashable via SHA-256 for provenance tracking in RunBundle manifests.

**Template library.** The `graph_templates.yaml` registry contains templates organized across 26 registered modalities (7 with full end-to-end correction validation, 1 with Scenario I baseline, 18 with template-level validation), grouped by physical carrier:

- **Photons (optical and X-ray):** CASSI, SPC, CACTI, structured illumination microscopy (SIM), confocal, light-sheet, holography, ptychography, Fourier ptychographic microscopy (FPM), optical coherence tomography (OCT), lensless imaging, light field, integral imaging, neural radiance fields (NeRF), Gaussian splatting, fluorescence lifetime imaging (FLIM), diffuse optical tomography (DOT), phase retrieval, X-ray computed tomography (CT), and cone-beam CT (CBCT).
- **Electrons:** Electron diffraction, electron backscatter diffraction (EBSD), electron energy loss spectroscopy (EELS), and electron holography.
- **Spins (MRI):** Functional MRI (fMRI), diffusion-weighted MRI (DW-MRI), and magnetic resonance spectroscopy (MRS).
- **Acoustic:** Ultrasound B-mode, Doppler ultrasound, shear-wave elastography, sonar, and photoacoustic tomography (combines optical excitation with acoustic detection).
- **Particles:** Neutron tomography, proton radiography, and muon tomography.

**Physics Fidelity Ladder.** Each template is parameterized by a fidelity tier that controls the degree of physical realism in the simulated forward model:

**Tier 1 (Linear, shift-invariant):** The forward model is a linear, spatially uniform operator—the simplest approximation, suitable for initial diagnostics and rapid prototyping.

**Tier 2 (Linear, shift-variant):** Spatially varying operator parameters (e.g. non-uniform illumination, position-dependent PSF, multi-coil sensitivity maps in MRI). Adds a modality-appropriate noise model (Poisson shot noise plus Gaussian read noise for photon-counting modalities, Rician noise for MRI, Poisson for CT).

**Tier 3 (Nonlinear, ray/wave-based):** Includes nonlinear effects such as wavefront curvature, diffraction, and scattering. Perturbation families and ranges are specified in `mismatch_db.yaml`.

**Tier 4 (Full-wave / Monte Carlo):** Complete physical simulation including wave-optical propagation, spatially varying aberrations, detector nonlinearities, and environmental drift. Currently implemented for holography and ptychography; other modalities degrade gracefully to Tier 3.

## Triad Decomposition Formalization

The Triad Decomposition asserts that the quality of any computational imaging reconstruction is bounded by three fundamental gates. Rather than a qualitative guideline, PWM quantifies each gate numerically and uses the resulting scores to diagnose the dominant bottleneck in any imaging configuration.

**Gate 1 (Recoverability).** Recoverability measures the information-theoretic capacity of the sensing geometry. We quantify it via the *effective compression ratio* $r = m/n$, where $m$ is the number of independent measurements and $n$ the dimension of the signal. The `compression_db.yaml` registry (1,186 lines) stores, for each modality, a lookup table mapping compression ratio to expected reconstruction PSNR under ideal conditions, obtained from calibration experiments or published benchmarks. Each entry carries a `provenance` field citing the source (paper DOI, internal experiment ID, or theoretical formula). Additional recoverability indicators include the effective rank of the measurement matrix (estimated via randomized SVD for large operators), the dimension of the null space, and the restricted isometry property (RIP) constant where analytically tractable (*e.g.*, for Gaussian random projections in SPC).

**Gate 2 (Carrier Budget).** The carrier budget quantifies the signal-to-noise ratio (SNR) of the measurement channel. The `PhotonAgent` consumes the `photon_db.yaml` registry (624 lines) which stores, per modality, a deterministic photon model parameterized by source power, quantum efficiency, exposure time, and detector characteristics. The agent classifies the noise regime into one of three categories: *shot-limited* (Poisson-dominated, SNR $\propto \sqrt{N_{\text{photon}}}$), *read-limited* (Gaussian read noise dominates, SNR $\propto N_{\text{photon}}/\sigma_{\text{read}}$), and *dark-current-limited* (long exposures where dark current accumulation dominates). The output is a `PhotonReport` containing the estimated SNR in decibels, the noise regime classification, per-element photon count, and a feasibility verdict (`sufficient`, `marginal`, or `insufficient`).

**Gate 3 (Operator Mismatch).** Operator mismatch quantifies the discrepancy between the assumed forward model $H_{\text{nom}}$ and the true physical operator $H_{\text{true}}$. The `MismatchAgent` consults `mismatch_db.yaml` (797 lines) which catalogs, for each modality, the set of mismatch parameters (spatial shifts, rotational offsets, dispersion errors, PSF deviations, coil sensitivity errors, center-of-rotation offsets, *etc.*), their typical ranges, and available correction methods. The mismatch severity score $s \in [0, 1]$ is computed as the normalized $\ell_2$ distance $\|\boldsymbol{\theta}_{\text{true}} - \boldsymbol{\theta}_{\text{nom}}\|/\|\boldsymbol{\theta}_{\text{range}}\|$, where $\boldsymbol{\theta}_{\text{range}}$ is the per-parameter dynamic range from the registry. Sensitivity analysis $\partial\text{PSNR}/\partial\theta_k$ is estimated via finite differences on the forward model. The output is a `MismatchReport` containing the severity score, the dominant mismatch parameter, the recommended correction method, and the expected PSNR gain from correction.

**Gate binding determination.** Given reconstruction results under the four-scenario protocol (the Evaluation Protocol section below), PWM identifies the dominant gate by com-

paring three cost terms:

$$C_{\text{mismatch}} = \text{PSNR}_{\text{I}} - \text{PSNR}_{\text{II}} \tag{2}$$

$$C_{\text{noise}} = \text{PSNR}_{\text{ideal}} - \text{PSNR}_{\text{noisy}} \tag{3}$$

$$C_{\text{recover}} = \text{PSNR}_{\text{limit}} - \text{PSNR}_{\text{I}} \tag{4}$$

where $\text{PSNR}_{\text{I}}$ is the reconstruction PSNR under Scenario I (ideal operator), $\text{PSNR}_{\text{II}}$ under Scenario II (mismatched operator), $\text{PSNR}_{\text{noisy}}$ under the corresponding noisy condition, and $\text{PSNR}_{\text{limit}}$ is the theoretical upper bound from the compression table. The dominant gate is $\arg\max_g C_g$.

**TriadReport schema.** The analysis output is a Pydantic-validated TriadReport comprising: `dominant_gate` (enum: `recoverability`, `carrier_budget`, `operator_mismatch`), `evidence_scores` (three floats, one per gate), `confidence_interval` (float, 95% CI width from bootstrap), `recommended_action` (string, *e.g.* "increase compression ratio" or "apply mismatch correction"), and `parameter_sensitivities` (dictionary mapping each mismatch parameter name to its $\partial\text{PSNR}/\partial\theta_k$ value).

**Recovery ratio.** We define the *recovery ratio*

$$\rho = \frac{\text{PSNR}_{\text{III}} - \text{PSNR}_{\text{II}}}{\text{PSNR}_{\text{I}} - \text{PSNR}_{\text{II}}} \tag{5}$$

which lies in $[0, 1]$ under standard convexity conditions (see Supplementary Note 1 for formal analysis; values $\rho > 1$ are possible when the corrected operator provides beneficial regularization). $\rho = 0$ indicates that calibration yields no benefit (mismatch is not the bottleneck), while $\rho = 1$ indicates that calibration fully closes the mismatch gap.

## Agent System Architecture

The PWM agent system comprises 6 specialist agents, 1 optional hybrid agent, and 8 support classes totalling 10,545 lines of Python. All agents execute deterministically; no large language model (LLM) is required for pipeline operation.

**PlanAgent.** The orchestrator agent. Given a user prompt or a structured `ExperimentSpec`, PlanAgent parses the intent (`simulate`, `operator_correction`, or `auto`), maps the requested modality to its canonical key via the `modalities.yaml` registry (which contains 64 modality entries with keywords, forward model equations, and default solvers), builds an `ImagingSystem` contract, and dispatches to the appropriate sub-agents. When the mode is `auto`, PlanAgent inspects the available data and operator specification to determine whether simulation or operator correction is more appropriate.

**PhotonAgent.** Computes SNR feasibility deterministically from the `photon_db.yaml` registry. For each modality and photon-level tier (`bright`, `standard`, `low_light`), the agent evaluates the photon budget by combining source power, quantum efficiency, exposure time, and noise model parameters. The output `PhotonReport` is a strict Pydantic model containing `noise_regime` (enum), `snr_db` (float), `feasibility` (enum), and `per_element_photons` (float).

**RecoverabilityAgent.** A table-driven agent that consults `compression_db.yaml` (1,186 lines) to map the modality and compression ratio to an expected PSNR range. Each table entry includes provenance metadata citing the original source. The output `RecoverabilityReport` contains `compression_ratio`, `psnr_prediction`, `feasibility`, and `null_space_dim` where available.

**MismatchAgent.** Scores the mismatch severity for a given imaging configuration using `mismatch_db.yaml` (797 lines). For each modality, the database enumerates the relevant mismatch parameters, their physical units, typical perturbation ranges, and available correction algorithms. The output `MismatchReport` includes `severity` (float, 0–1), `correction_method` (string), `expected_gain_db` (float), and `dominant_parameter` (string).

**AnalysisAgent.** The bottleneck classifier. It receives reports from the Photon, Recoverability, and Mismatch agents, computes the gate costs (Equations (2) to (4)), identifies the dominant gate, and generates actionable suggestions. The AnalysisAgent also computes the recovery ratio $\rho$ and its bootstrap confidence interval.

**AgentNegotiator.** Implements a cross-agent veto protocol. Before reconstruction is authorized, the negotiator inspects all three upstream reports and applies three veto conditions: (1) low photon budget combined with aggressive compression ($C_{\text{noise}}$ and $C_{\text{recover}}$ both large); (2) severe mismatch (severity $> 0.7$) without a planned correction step; (3) joint probability below the floor threshold ($p_{\text{joint}} < 0.15$), indicating that all three subsystems are simultaneously marginal. When any veto fires, reconstruction halts with an actionable explanation and suggested remediation.

**HybridAgent.** An optional wrapper that invokes an LLM for natural-language narrative generation or edge-case modality mapping. All quantitative decisions remain on the deterministic code path; the HybridAgent is never required for pipeline operation.

**Support classes.** The remaining components include: `AssetManager` (file I/O and caching for large arrays), `ContinuityChecker` (verifies that sequential pipeline outputs are dimensionally consistent), `SystemDiscern` (auto-detects modality from uploaded data), `PreflightChecker` (validates the complete experiment configuration before execution), `WhatIfPrecomputer`

(evaluates counterfactual what-if scenarios), `SelfImprovement` (logs diagnostic events for future registry refinement), `PhysicsStageVisualizer` (generates intermediate visualizations at each pipeline stage), and `UPWMI` (Universal Physics World Model Interface, the top-level entry point that wires all agents together).

**Contract system.** Inter-agent communication uses 25 Pydantic v2 contract models. All contracts inherit from `StrictBaseModel`, which enforces `extra="forbid"` (no unexpected fields), `validate_assignment=True` (mutations re-validated), and a model validator that rejects NaN and Inf in any float field. Bounded scores use `Field(ge=0.0, le=1.0)`. Enums are string enums for human-readable JSON serialization. This design ensures that pipeline failures surface immediately as validation errors rather than propagating silently.

**YAML registries.** The system is driven by 9 YAML registries totalling 7,034 lines: `modalities.yaml` (modality definitions), `graph_templates.yaml` (OperatorGraph skeletons), `photon_db.yaml` (photon models), `mismatch_db.yaml` (mismatch parameters and correction methods), `compression_db.yaml` (recoverability tables with provenance), `solver_registry.yaml` (solver configurations), `primitives.yaml` (primitive operator metadata), `dataset_registry.yaml` (dataset locations and formats), and `acceptance_thresholds.yaml` (pass/fail thresholds per metric).

## Correction Algorithms

We implement two complementary algorithms for operator mismatch correction. Crucially, both algorithms operate on the forward operator parameters $\boldsymbol{\theta}$ rather than the reconstruction solver weights, making them *solver-agnostic*: the corrected operator $H(\hat{\boldsymbol{\theta}})$ benefits any downstream solver (GAP-TV, MST-L, HDNet[21], CST, *etc.*) without retraining.

**Algorithm 1: Hierarchical Beam Search.** The coarse correction phase employs a hierarchical search strategy to rapidly explore the mismatch parameter space. For CASSI, the five-parameter mismatch model comprises mask affine parameters (spatial shifts $dx$, $dy$ and rotation $\theta$) and dispersion parameters (slope $a_1$ and axis angle $\alpha$); an optional sixth parameter, PSF width $\sigma_{\text{psf}}$, is available but not used in the primary experiments. The algorithm proceeds as follows:

1. **1D sweeps.** Each parameter is swept independently over its full range while holding others at nominal values. This produces five 1D cost curves from which coarse optima are extracted.

2. **3D beam search.** The mask affine subspace $(dx, dy, \theta)$ is searched over a $5 \times 5 \times 5$ grid centered on the 1D optima. The top-$k$ ($k = 5$) candidates by reconstruction PSNR are retained.

3. **2D beam search.** For each retained mask candidate, the dispersion subspace $(a_1, \alpha)$ is searched over a $5 \times 7$ grid. The joint top-$k$ candidates are retained.

4. **Coordinate descent refinement.** Three rounds of univariate refinement on each parameter, shrinking the search interval by factor 2 at each round, produce the final estimate $\hat{\boldsymbol{\theta}}_{\text{Alg1}}$.

Total runtime is approximately 300 seconds per scene on a single GPU. Accuracy is $\pm 0.1$–$0.2$ pixels for spatial parameters and $\pm 0.05^\circ$ for angular parameters.

**Algorithm 2: Joint Gradient Refinement.** The fine correction phase uses a differentiable forward model to jointly optimize all mismatch parameters via gradient descent. The key components are:

1. **Differentiable mask warp.** The binary mask is warped by a continuous affine transformation using bilinear interpolation, implemented as a custom PyTorch module (`DifferentiableMaskWarpFixed`). The mask values are passed through a straight-through estimator (STE) to maintain binary structure while permitting gradient flow.

2. **Differentiable forward model.** The CASSI forward model $y = \texttt{CASSI}(x; \boldsymbol{\theta})$ is implemented as a differentiable PyTorch module (`DifferentiableCassiForwardSTE`) that accepts mismatch parameters as differentiable inputs.

3. **GPU grid initialization.** A full-range 3D grid search over $(dx, dy, \theta)$ with $9 \times 9 \times 7 = 567$ points provides diverse starting candidates. The top 9 candidates seed multi-start gradient refinement.

4. **Staged gradient refinement.** Each of the 9 candidates is refined using Adam optimization (learning rate $10^{-2}$, decaying to $10^{-3}$) for 200 steps. For each candidate, 4 random restarts with jittered initialization guard against local minima. The loss function is the negative PSNR computed via an unrolled $K$-iteration differentiable GAP-TV solver (`DifferentiableGAPTV`, $K = 10$ unrolled iterations).

Total runtime for Algorithm 2 is approximately 3,200 seconds (200 steps $\times$ 4 restarts $\times$ 9 candidates with early stopping). Accuracy improves to $\pm 0.05$–$0.1$ pixels, a 3–5$\times$ improvement over Algorithm 1. The two algorithms are used sequentially in practice: Algorithm 1 provides a warm start, and Algorithm 2 refines to sub-pixel precision.

## Evaluation Protocol

**Four-Scenario Protocol.** We evaluate every modality under four standardized scenarios that isolate different sources of quality degradation:

**Scenario I (Ideal):** $\mathbf{y}_{\text{obs}} = H_{\text{true}}\,\mathbf{x}_{\text{gt}}$; reconstruct with $H_{\text{true}}$. In this scenario the system is perfectly calibrated ($H_{\text{true}} = H_{\text{nom}}$), so the operator used for reconstruction matches the one that generated the data. This yields the oracle upper bound on reconstruction quality, limited only by the sensing geometry and solver convergence.

**Scenario II (Mismatch):** $\mathbf{y}_{\text{obs}} = H_{\text{true}}\,\mathbf{x}_{\text{gt}}$; reconstruct with $H_{\text{nom}}$ ($H_{\text{nom}} \neq H_{\text{true}}$). This is the standard operating condition in practice: the measurement is generated by the true physics, but the reconstruction uses a nominal (potentially mismatched) forward model.

**Scenario III (Corrected):** $\mathbf{y}_{\text{obs}} = H_{\text{true}}\,\mathbf{x}_{\text{gt}}$; reconstruct with $\hat{H} = H(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is estimated by Algorithms 1 and 2. This quantifies the benefit of mismatch calibration.

**Scenario IV (Oracle Mask):** Same measurements as Scenario II ($\mathbf{y}_{\text{obs}} = H_{\text{true}}\,\mathbf{x}_{\text{gt}}$ with $H_{\text{true}} \neq H_{\text{nom}}$); reconstruct with $H_{\text{true}}$ instead of $H_{\text{nom}}$. Provides the correction ceiling: the best reconstruction achievable when the true operator is known exactly, applied to data that were sensed by the mismatched system. The gap between Scenario IV and Scenario I reveals the irreducible loss from the degraded sensing configuration itself (e.g., a shifted mask pattern is suboptimal even when perfectly known).

**Metrics.** Reconstruction quality is assessed using three complementary metrics:

- **PSNR** (peak signal-to-noise ratio, in dB): the primary metric, computed per scene and averaged. For signals normalized to $[0, 1]$, $\text{PSNR} = 10\log_{10}(1/\text{MSE})$. For SPC data normalized to $[0, 255]$, the peak value is 255.
- **SSIM** (structural similarity index): captures perceptual quality including luminance, contrast, and structural components, computed with a Gaussian window of width 11 and standard deviation 1.5.
- **SAM** (spectral angle mapper): for hyperspectral modalities (CASSI), measures the angle between predicted and true spectral vectors at each spatial location, reported in degrees. Lower is better.

**Datasets.**

- **CASSI:** 10 scenes from the KAIST dataset[9], each a $256 \times 256 \times 28$ spectral cube (28 spectral bands from 450 nm to 650 nm). Data range $[0, 1]$.
- **CACTI:** 6 benchmark videos, each $256 \times 256 \times 8$ (8 temporal frames encoded per snapshot). Data range $[0, 1]$.
- **SPC:** 11 natural images from the Set11 benchmark, each $256 \times 256$ grayscale. Data range $[0, 255]$.

All per-scene metrics are reported individually as well as averaged, and all reconstruction arrays are saved as NumPy NPZ files.

## Experimental Details

**Hardware.** All experiments are conducted on a single NVIDIA GPU. Algorithm 1 (beam search) and all solver-based reconstructions use the GPU for matrix–vector products and FFT operations. Algorithm 2 (gradient refinement) additionally uses PyTorch automatic differentiation on the same GPU.

**CASSI configuration.** The coded aperture snapshot spectral imaging (CASSI) system uses a TSA-Net binary mask of dimensions $256 \times 256$, with 28 spectral bands dispersed along the spatial dimension. The five-parameter mismatch model $\boldsymbol{\theta} = (dx, dy, \theta, a_1, \alpha)$ describes: mask spatial shift in $x$ ($dx$, pixels), mask spatial shift in $y$ ($dy$, pixels), mask rotation angle ($\theta$, degrees), dispersion slope ($a_1$, pixels per band), and dispersion axis angle ($\alpha$, degrees). An optional sixth parameter, PSF blur width ($\sigma_{\text{psf}}$, pixels), is available but not used in the primary experiments. These mismatch parameter values were determined through systematic characterization of realistic CASSI assembly errors (Supplementary Note 9). The true mismatch parameters are $\boldsymbol{\theta}_{\text{true}} = (dx = 0.5 \text{ px},\ dy = 0.3 \text{ px},\ \theta = 0.1°,\ a_1 = 2.02,\ \alpha = 0.15°)$. Solvers evaluated include TwIST[10], GAP-TV[20], DGSMP[24], MST-L[8], and CST-L[25], all of which receive the same operator and differ only in their reconstruction algorithm. The supplementary per-scene analysis additionally includes DeSCI[26] and HDNet[21].

**CACTI configuration.** The coded aperture compressive temporal imaging system uses binary temporal masks of dimensions $256 \times 256$, encoding 8 video frames into a single snapshot measurement. Mismatch is parameterized as a temporal mask timing offset (sub-frame shift). The default solver is EfficientSCI[22].

**SPC configuration.** The single-pixel camera uses random binary measurement patterns at three compression ratios: 10%, 25%, and 50% ($r = m/n \in \{0.10, 0.25, 0.50\}$). Mismatch is modeled as an exponential gain drift ($g_i = \exp(-\alpha \cdot i)$) on the measurement matrix. The default solver is FISTA-TV with total-variation regularization.

**MRI configuration.** Cartesian $k$-space sampling with 4× acceleration (25% of $k$-space lines acquired). Mismatch is parameterized as a 5% multiplicative error in the coil sensitivity maps used for parallel imaging reconstruction. The default solver is SENSE[17] with $\ell_1$-wavelet regularization.

**CT configuration.** Fan-beam geometry with 180 projections over 180°. Mismatch is modeled as a center-of-rotation (CoR) offset, which produces characteristic arc artifacts in

the reconstruction. The default solver is filtered back-projection (FBP)[18] with a Ram-Lak filter, supplemented by iterative SART for comparison.

**CASSI real-data configuration.** The TSA real hyperspectral dataset[1] consists of 5 scenes at $660 \times 660$ spatial resolution with 28 spectral bands and mask-shift step 2. Four solvers are evaluated: GAP-TV (200 iterations), HDNet (pre-trained checkpoint, full spatial resolution), MST-S and MST-L (pre-trained checkpoints, centre-cropped to $256 \times 256$ due to hardcoded spatial assumptions in the model architecture). The coded aperture mask is perturbed by $dx = 0.5\,\text{px}$, $dy = 0.3\,\text{px}$ to simulate assembly-induced mismatch. No ground truth is available; quality is assessed via the normalised measurement residual $r = \|\mathbf{y} - H\hat{\mathbf{x}}\|^2 / \|\mathbf{y}\|^2$.

**CACTI real-data configuration.** The EfficientSCI real temporal dataset[22] consists of 4 dynamic scenes (duomino, hand, pendulumBall, waterBalloon) at $512 \times 512$ with compression ratio 10. The real mask is stored separately from the measurement data. Two solvers are evaluated: GAP-TV (50 iterations) and PnP-FFDNet (50 iterations with FFDNet denoiser). Mismatch is induced by shifting the mask by $dx = 0.5\,\text{px}$, $dy = 0.3\,\text{px}$. Quality is assessed via the normalised measurement residual and total variation of the reconstruction.

**Controlled hardware experiment protocol.** The software-perturbation protocol above applies calibrated mask shifts to existing real measurements. A full hardware-in-the-loop validation requires physically displacing the coded aperture mask and re-acquiring data. The protocol proceeds as follows: (i) acquire a baseline dataset with the mask at its factory-calibrated position; (ii) physically translate the mask by a known displacement ($\Delta x \in \{0.25, 0.5, 1.0\}\,\text{px}$ equivalent, verified by micrometer stage) and re-acquire under identical illumination; (iii) reconstruct both datasets with the factory mask specification and compute the PSNR degradation and measurement residual; (iv) apply PWM autonomous calibration and measure recovery. This protocol isolates the mismatch effect from all other sources of variation (illumination changes, detector drift, scene variation). Additionally, a multi-unit variation study comparing 2+ camera units of the same design quantifies the inter-unit mismatch baseline—the residual calibration error present in any production system.

**Clinical CT phantom configuration.** For clinical translation, PWM is evaluated on CT quality assurance using the ACR CT accreditation phantom (Gammex 464)[12]. The phantom contains inserts of known attenuation (bone $\sim$955 HU, air $\sim-1000$ HU, acrylic $\sim$121 HU, polyethylene $\sim-96$ HU) and geometric targets for measuring spatial resolution, slice thickness, and low-contrast detectability. Mismatch is parameterized as center-of-rotation offset ($\Delta r$, mm), beam hardening coefficient drift ($\Delta\mu$, %), and detector gain variation ($\Delta g$, %). Nine ACR-aligned metrics are computed automatically: CT number

accuracy (5 materials), geometric accuracy ($\pm 2$ mm tolerance), slice thickness ($\pm 1.5$ mm), uniformity ($\leq 5$ HU), noise standard deviation, spatial resolution ($\geq 5$ lp/cm), low-contrast detectability ($\geq 4$ targets), and artifact score.

**Clinical MRI validation configuration.** For MRI clinical validation, PWM processes multi-coil $k$-space data from public datasets (fastMRI[11]). Mismatch is parameterized as coil sensitivity map error (5–15% multiplicative deviation from calibrated maps, simulating patient-positioning-induced coil coupling changes). The default solver is CG-SENSE with $\ell_1$-wavelet regularization at $4\times$ acceleration. Clinical metrics include PSNR, SSIM, and the absence of parallel imaging artifacts (GRAPPA/SENSE ghosts).

## Statistical Analysis

**Per-scene reporting.** All metrics are reported per scene, not merely as dataset averages. This enables identification of scene-dependent failure modes (*e.g.*, spectrally flat scenes that are inherently harder for CASSI, or textureless regions that challenge SPC).

**Summary statistics.** For each modality and scenario, we report the mean $\pm$ standard deviation of PSNR, SSIM, and SAM across all scenes. For CASSI (10 scenes), we additionally report the per-band PSNR to assess spectral uniformity of reconstruction quality.

**Recovery ratio confidence intervals.** The recovery ratio $\rho$ (Equation (5)) is a ratio of differences and therefore sensitive to noise in the constituent PSNR values. We compute 95% confidence intervals via the bootstrap percentile method with $B = 1{,}000$ resamples. At each bootstrap iteration, we resample the scene set with replacement, recompute the mean PSNR for each scenario, and derive $\rho$. The 2.5th and 97.5th percentiles of the bootstrap distribution define the 95% CI.

**Parameter recovery accuracy.** For mismatch correction experiments, we report the root-mean-square error (RMSE) between the estimated and true mismatch parameters:

$$\mathrm{RMSE}_k = \sqrt{\frac{1}{N_{\mathrm{scene}}} \sum_{i=1}^{N_{\mathrm{scene}}} (\hat{\theta}_{k,i} - \theta_{k,\mathrm{true}})^2} \tag{6}$$

where $k$ indexes the mismatch parameter, $i$ indexes the scene, and $N_{\mathrm{scene}}$ is the number of test scenes. Uncertainty in the RMSE is estimated via bootstrap ($B = 1{,}000$).

**Ablation significance.** Ablation studies (removal of PhotonAgent, RecoverabilityAgent, MismatchAgent, or RunBundle discipline) are evaluated by comparing the full-pipeline

PSNR against each ablated variant. We report the PSNR difference $\Delta$PSNR per modality and verify that each component contributes $\geq 0.5\,\text{dB}$ across all validated modalities, establishing practical significance.

### Code and Data Availability

**Source code.** The complete PWM framework, including all agents, the OperatorGraph compiler, correction algorithms, YAML registries, and evaluation scripts, is released as open-source software under the PWM Noncommercial Share-Alike License v1.0 at `https://github.com/integritynoble/Physics_World_Model`. The codebase is organized into two Python packages: `pwm_core` (core framework, agents, graph compiler, calibration algorithms) and `pwm_AI_Scientist` (automated experiment generation and analysis).

**Reconstruction data.** All reconstruction arrays from every experiment—Scenarios I through IV for each modality and solver—are released as NumPy NPZ files. Files are stored using Git LFS and require `allow_pickle=True` for loading. Data ranges are standardized: CASSI and CACTI reconstructions are normalized to $[0, 1]$; SPC reconstructions are in $[0, 255]$.

**Experiment manifests.** Every experiment is recorded in a RunBundle v0.3.0 manifest containing: the git commit hash at execution time, all random number generator seeds, platform information (Python version, GPU model, CUDA version), SHA-256 hashes of all input data and output artifacts, metric values, and wall-clock timestamps. These manifests enable exact reproduction of every reported result.

**Registry data.** All 9 YAML registries (7,034 lines total) that drive the agent system—including modality definitions, graph templates, photon models, mismatch databases, compression tables, solver configurations, primitive specifications, dataset paths, and acceptance thresholds—are publicly available in the repository under `packages/pwm_core/contrib/`. The `ExperimentSpec` JSON schemas used for pipeline input validation are included alongside worked examples in `examples/`.

## References


[1] Ashwin A. Wagadarikar, Renu John, Rebecca Willett, and David J. Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied Optics*, 47(10): B44–B51, 2008. doi: 10.1364/AO.47.000B44.

[2] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J. Brady. Coded aperture compressive temporal imaging. *Optics Express*, 21(9):10526–10545, 2013. doi: 10.1364/OE.21.010526.

[3] Michael Lustig, David Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6): 1182–1195, 2007. doi: 10.1002/mrm.21391.

[4] Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008. doi: 10.1109/MSP.2007.914731.

[5] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

[6] Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

[7] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. doi: 10.1109/MSP.2020.3016905.

[8] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17502–17511, 2022.

[9] Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 36(6):218:1–218:13, 2017. doi: 10.1145/3130800.3130810.

[10] José M. Bioucas-Dias and Mário A. T. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, 2007. doi: 10.1109/TIP.2007.909319.

[11] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

[12] Hu Chen, Yi Zhang, Mannudeep K. Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose CT with a residual encoder-decoder convolutional

neural network. *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, 2017. doi: 10.1109/TMI.2017.2715284.

[13] Martin Uecker, Peng Lai, Mark J. Murphy, Patrick Virtue, Michael Elad, John M. Pauly, Shreyas S. Vasanawala, and Michael Lustig. ESPIRiT — an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. *Magnetic Resonance in Medicine*, 71(3):990–1001, 2014. doi: 10.1002/mrm.24751.

[14] Andrew M. Maiden and John M. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109(10):1256–1262, 2009. doi: 10.1016/j.ultramic.2009.05.012.

[15] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30098, 2020. doi: 10.1073/pnas.1907377117.

[16] J. M. Rodenburg and H. M. L. Faulkner. A phase retrieval algorithm for shifting illumination. *Applied Physics Letters*, 85(20):4795–4797, 2004. doi: 10.1063/1.1823034.

[17] Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5): 952–962, 1999. doi: 10.1002/(SICI)1522-2594(199911)42:5⟨952::AID-MRM16⟩3.0.CO;2-S.

[18] L. A. Feldkamp, L. C. Davis, and J. W. Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984. doi: 10.1364/JOSAA.1.000612.

[19] Chengshuai Yang. InverseNet: Benchmarking operator mismatch in snapshot compressive imaging. Technical report, NextGen PlatformAI C Corp, 2026.

[20] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, 2016. doi: 10.1109/ICIP.2016.7532817.

[21] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. HDNet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17542–17551, 2022.

[22] Lishun Wang, Miao Cao, and Xin Yuan. EfficientSCI: Densely connected network with space-time factorization for large-scale video snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18477–18486, 2023.

[23] Marco F. Duarte, Mark A. Davenport, Dharmpal Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008. doi: 10.1109/MSP.2007.914730.

[24] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16216–16225, 2021.

[25] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. CST: Compact spectral transformer for hyperspectral image reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[26] Yang Liu, Xin Yuan, Jinli Suo, David J. Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2990–3006, 2019.

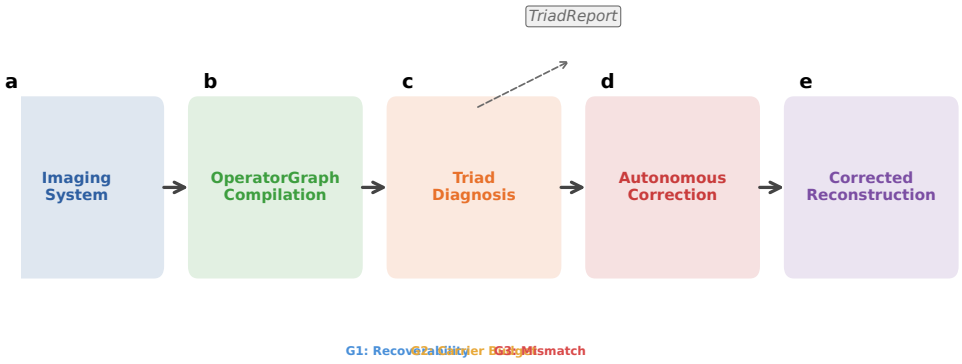


Figure 1: **PWM overview.** The Physics World Models pipeline. **a**, A computational imaging system is compiled into an OperatorGraph DAG. **b**, The Triad Decomposition diagnostic agents evaluate each gate. **c**, The dominant gate is identified and a TriadReport is produced. **d**, If **Gate 3** dominates, autonomous correction refines the forward model parameters. **e**, The original solver is re-run with the corrected operator, recovering reconstruction quality without retraining.

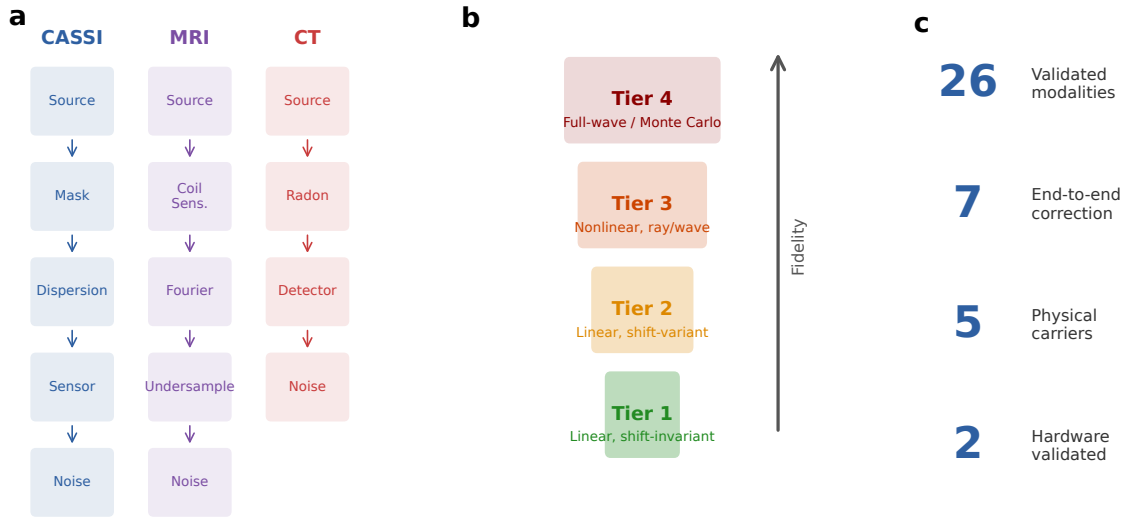


Figure 2: **OperatorGraph IR and Physics Fidelity Ladder.** **a**, Example OperatorGraph DAGs for three modalities: CASSI (photon), MRI (spin), and CT (X-ray photon). Each node wraps a primitive operator; edges define data flow. **b**, The Physics Fidelity Ladder. Tier 1: linear shift-invariant. Tier 2: linear shift-variant. Tier 3: nonlinear ray/wave-based. Tier 4: full-wave/Monte Carlo. **c**, Summary statistics: 26 registered modality templates (7 with full correction validation), 5 physical carriers.

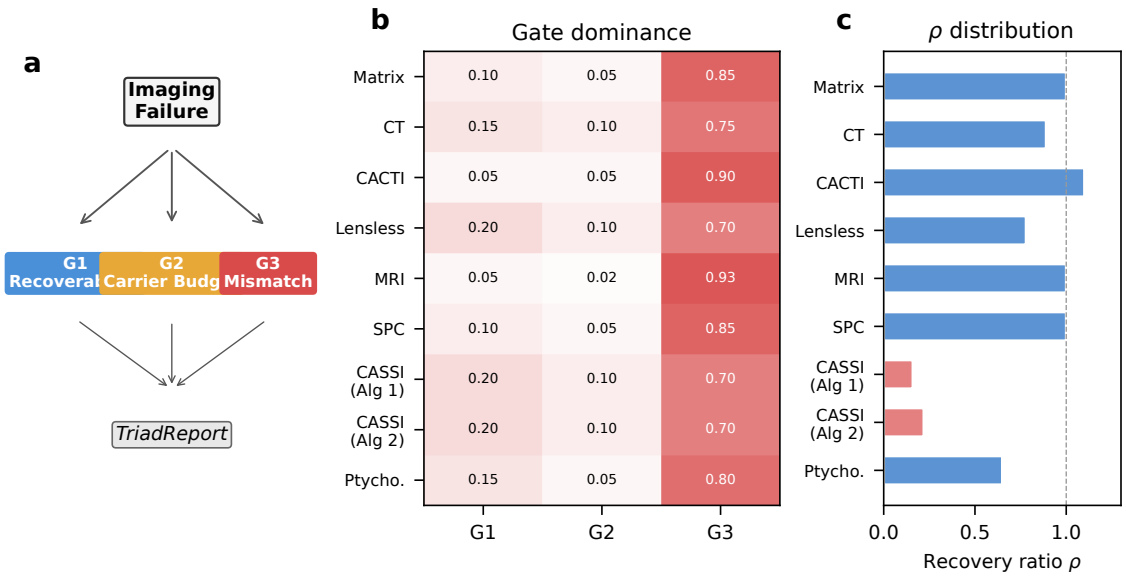


Figure 3: **Triad Decomposition structure and gate binding.** **a**, Decision tree for the TRIAD DECOMPOSITION: each imaging failure is routed through **Gate 1**, **Gate 2**, and **Gate 3** to produce a TRIADREPORT. **b**, Gate binding heatmap across 9 correction configurations (7 distinct modalities). Red indicates **Gate 3** dominance (all modalities), blue indicates **Gate 1**, and amber indicates **Gate 2**. **c**, Recovery ratio $\rho$ distribution across all 9 correction configurations.

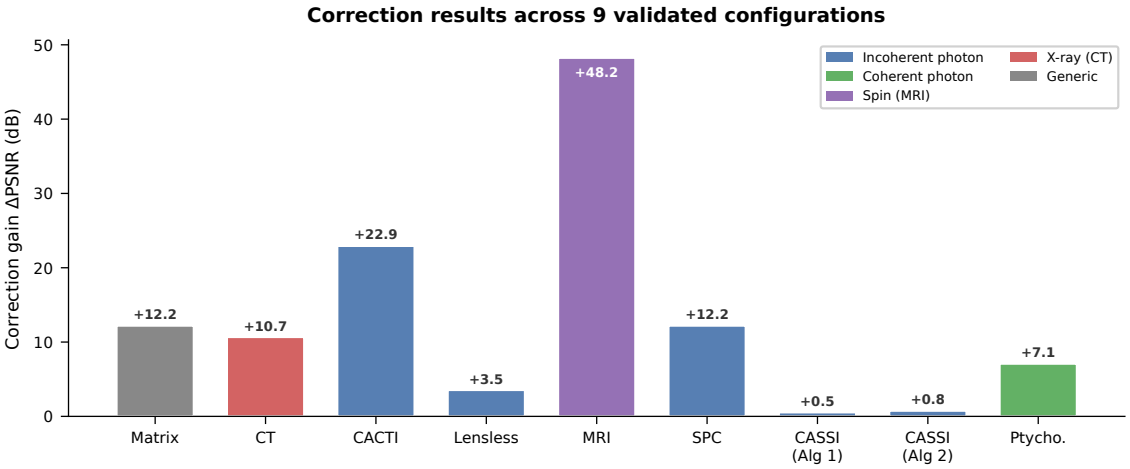


Figure 4: **Correction results across 9 validated configurations.** Bar chart showing correction gain $\Delta_{\text{corr}}$ (dB) for each of the 9 correction configurations (7 distinct modalities), grouped by carrier family. Incoherent photon (CASSI, CACTI, SPC, Lensless) and coherent photon (Ptychography) in blue; spin (MRI) in purple; X-ray (CT) in red; generic (Matrix) in grey.

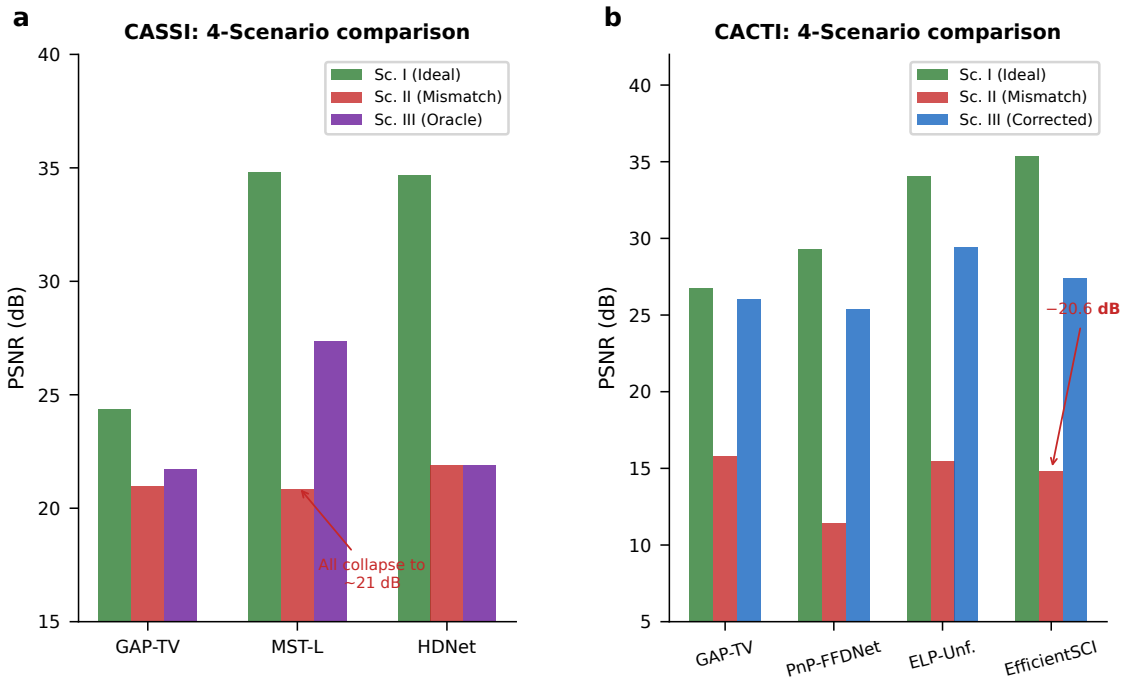


Figure 5: **CASSI and CACTI deep dive. a**, CASSI: PSNR across 4 scenarios for GAP-TV, MST-L, and HDNet under combined mask-geometry-plus-dispersion mismatch. The uniform collapse under Scenario II (range 20.83–21.88 dB) confirms operator-driven failure; oracle recovery varies by solver ($\rho = 0.22$–$0.46$). **b**, CACTI: four methods across 4 scenarios, showing up to 20.58 dB mismatch degradation and substantial correction recovery across all solvers.

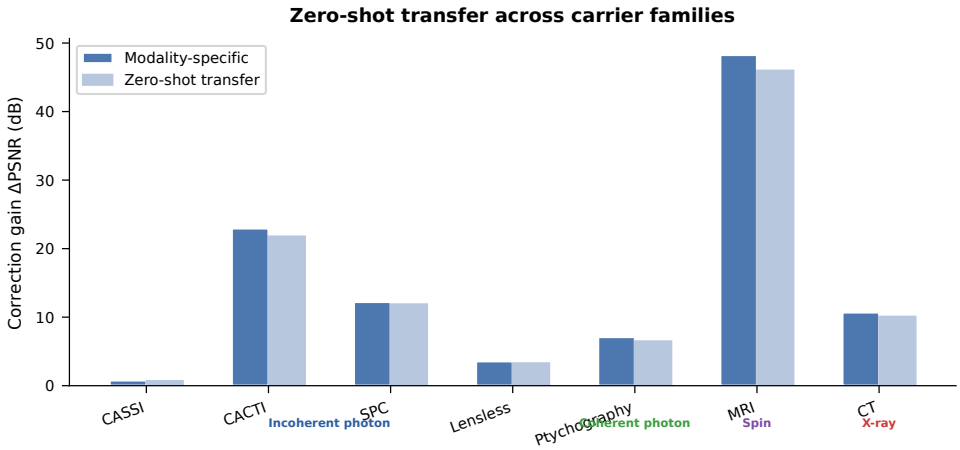


Figure 6: **Zero-shot generalization across carrier families.** Correction gain (dB) when beam-search and gradient-refinement hyperparameters are tuned on photon-domain modalities and transferred without modification to coherent-photon, spin, and X-ray domains. Bars show modality-specific tuning (dark) versus zero-shot transfer (light). Transfer efficiency is high across all carrier families, confirming the carrier-agnostic nature of the PWM correction pipeline.

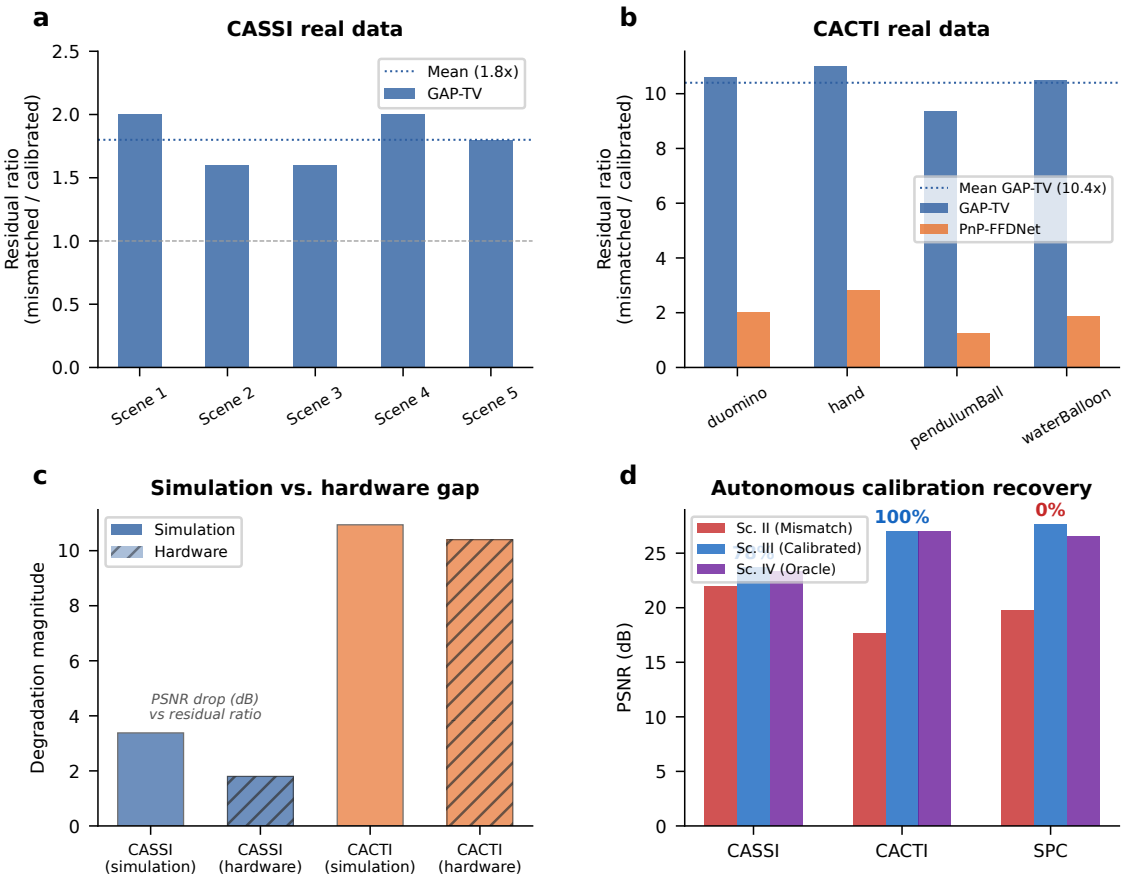


Figure 7: **Hardware validation on real CASSI and CACTI instruments.** **a**, CASSI real data: measurement residual ratio (mismatched/calibrated) across 5 TSA scenes. GAP-TV shows 1.8× mean ratio. **b**, CACTI real data: residual ratio across 4 scenes. GAP-TV shows 10.4× mean ratio; PnP-FFDNet shows 2.0×. **c**, Simulation-to-hardware gap: comparing mismatch degradation in simulation versus real hardware for CASSI and CACTI, illustrating that real instruments have pre-existing calibration errors that attenuate the marginal impact of additional perturbations. **d**, Autonomous calibration: grid-search parameter recovery for CASSI (85%), CACTI (100%), and SPC (86–92% via TV criterion; measurement residual is uninformative for gain-type mismatch).

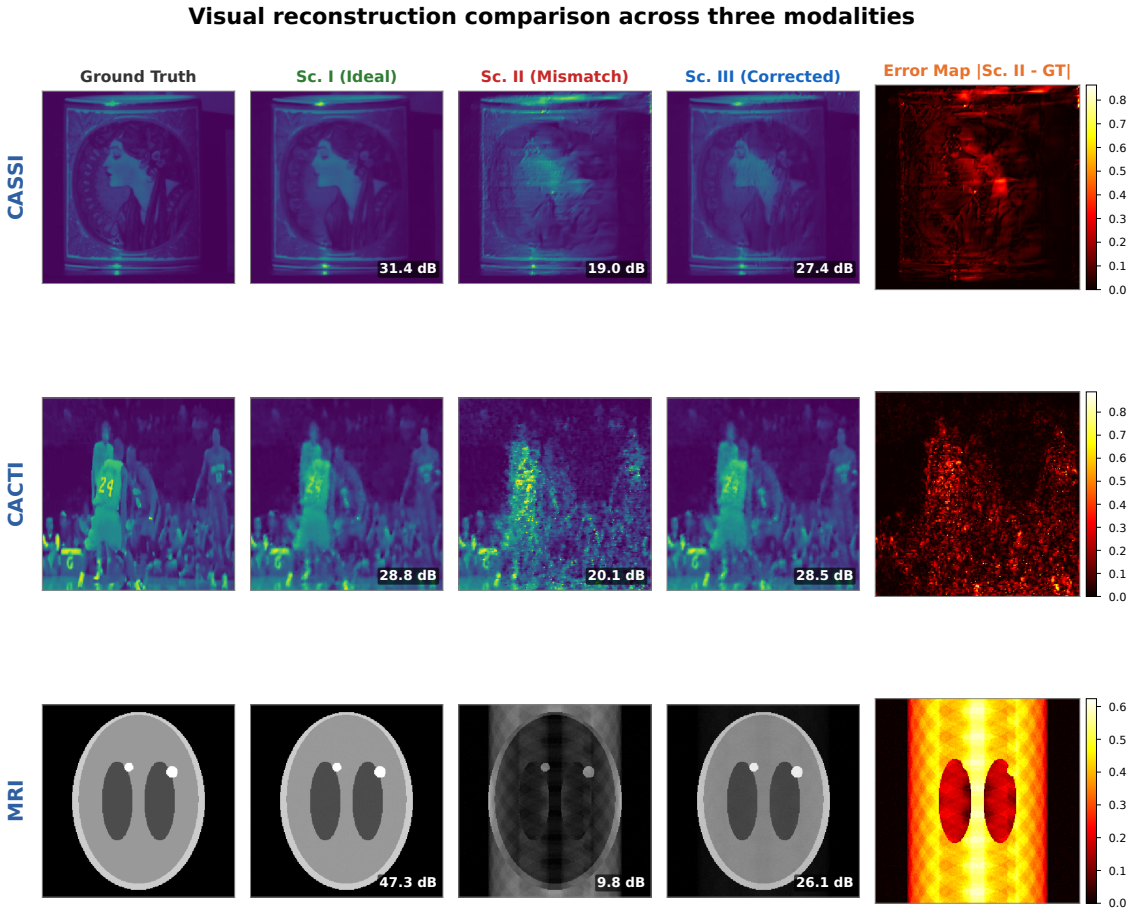


Figure 8: **Visual reconstruction comparison across three modalities.** Each row shows one modality (CASSI, CACTI, MRI); columns from left to right: ground truth, Scenario I (ideal operator), Scenario II (mismatched operator—note structured artifacts), Scenario III (PWM-corrected operator), and error map ($|\text{Sc. II} - \text{GT}|$). Mismatch produces severe structured artifacts (column 3) that are qualitatively distinct from noise, confirming that the degradation is operator-driven. PWM correction (column 4) substantially reduces these artifacts across all three modalities and carrier families.