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

Chengshuai Yang
NextGen PlatformAI C Corp
integrityyang@gmail.com

February 2026

Abstract

Computational imaging systems routinely fail in practice because the assumed forward model diverges from the true physics, yet no existing framework systematically diagnoses *why* reconstruction degrades. We introduce Physics World Models (PWM), a universal diagnostic and correction framework grounded in the TRIAD LAW: every imaging failure decomposes into exactly three root causes—recoverability loss (**Gate 1**), carrier-noise budget violation (**Gate 2**), and operator mismatch (**Gate 3**). PWM compiles 64 modalities spanning five physical carriers (photons, electrons, spins, acoustic waves, and particles) into a unified OPERATORGRAPH intermediate representation comprising 89 validated operator templates. Autonomous, deterministic agents diagnose the dominant failure gate and correct the forward model without retraining any reconstruction algorithm. Across 7 distinct modalities (9 correction configurations, including two CASSI algorithms and the Matrix baseline; 16 registered), correction yields improvements ranging from +0.54 dB to +48.25 dB. **Gate 3** is identified as the dominant bottleneck in every validated modality, demonstrating that a decade of solver-centric progress has overlooked the principal source of imaging failure. The TRIAD LAW provides the first universal, quantitative language for imaging diagnosis.

Introduction

Why do state-of-the-art reconstruction algorithms fail in practice? The answer is deceptively simple: the assumed forward model is wrong, and nobody measures this systematically. The computational imaging community has devoted extraordinary effort to designing ever more powerful solvers—from compressed sensing[1,2] and plug-and-play priors[3] to end-to-end deep unrolling networks[4]—while treating the forward model as a fixed, trusted input. This implicit assumption is rarely justified. Optical masks shift during assembly, MRI coil sensitivities drift with patient positioning, and CT geometries deviate from their nominal calibration. When these mismatches arise, even the most sophisticated reconstruction algorithms collapse, and the resulting artifacts are routinely misattributed to solver limitations rather than to their true cause: an incorrect physics model.

The scale of this crisis is striking. Consider coded aperture snapshot spectral imaging (CASSI), a representative photon-domain modality. Under ideal conditions—where the true coded mask is known exactly—the state-of-the-art transformer solver MST-L [5] achieves 34.81 dB on a standard benchmark [6]. Introduce a realistic 5-parameter perturbation—sub-pixel mask shift, rotation, and multi-parameter dispersion drift (see Methods for full specification)—and MST-L drops to 20.83 dB, a catastrophic loss of 13.98 dB. To put this in perspective, the cumulative improvement from a decade of solver development in CASSI—progressing from early iterative methods through deep unrolling to modern transformer architectures—amounts to roughly 7 dB (from iterative TwIST at $\sim$27.8 dB to transformer MST-L at 34.81 dB). A sub-pixel mask perturbation erases roughly twice the gains of an entire research generation. This is not a pathological edge case; analogous degradations appear across modalities, from lensless imaging to magnetic resonance imaging [7,8] to computed tomography [9].

The root problem is a missing diagnostic layer. When a reconstruction fails, the practitioner faces a differential diagnosis with at least three distinct failure modes. First, the measurement may be fundamentally information-deficient: the null space of the forward operator may preclude recovery regardless of the solver or signal-to-noise ratio. Second, the carrier budget may be insufficient: too few photons, too low a dose, or too short an acquisition may push the measurement below the quantum or thermal noise floor. Third, the assumed forward model may diverge from the true physics: the operator used for reconstruction may not match the operator that generated the data. These three failure modes interact, compound, and masquerade as one another, yet no existing framework disentangles them.

Previous work has addressed fragments of this problem. Calibration methods exist for specific instruments [10,11], but they are modality-specific and do not generalize. Uncertainty quantification techniques can flag unreliable reconstructions, but they do not diagnose the *cause* of the unreliability. Robustness studies perturb individual systems [12], but they lack a unifying formalism that connects perturbation types across the electromagnetic, acoustic, and particle-physics domains. The imaging community thus remains in a pre-diagnostic era: systems are built, they fail, and the failure is addressed *ad hoc* if it is addressed at all.

This paper introduces Physics World Models (pwm), a universal framework that elevates imaging diagnosis to a first-class computational task alongside reconstruction. The theoretical backbone of pwm is the Triad Law, which asserts that every imaging failure decomposes into exactly three root causes, termed gates: **Gate 1** (recoverability), **Gate 2** (carrier budget), and **Gate 3** (operator mismatch). The Triad Law is not a heuristic; it is a structured decomposition grounded in the information-theoretic and physical constraints that govern all linear inverse problems. For every modality and every reconstruction failure, PWM produces a TRIADREPORT: a mandatory diagnostic artifact that identifies the dominant gate, quantifies the evidence, and prescribes a corrective action.

To apply the TRIAD LAW across the full landscape of computational imaging, PWM introduces the OPERATORGRAPH intermediate representation (IR): a directed acyclic graph (DAG) formalism that compiles forward models from 64 modalities spanning five physical carriers—photons, electrons, spins, acoustic waves, and particles—into a common computational substrate. Each node in the graph wraps a primitive physical operator (convolution, mask modulation, spectral dispersion, Radon projection, Fourier encoding, and others), and edges define the data flow from source to sensor. The OPERATORGRAPH IR currently comprises 89 validated templates, enabling PWM to reason about imaging systems as diverse as coded aperture spectral imaging[13], ptychography[14], accelerated MRI[15], photoacoustic tomography, and neutron computed tomography within a single formalism.

Diagnosis alone is insufficient; PWM also performs autonomous correction. Three diagnostic agents (part of a 6-agent deterministic system, plus 1 optional hybrid agent, described in Methods)—`RecoverabilityAgent`, `PhotonAgent`, and `MismatchAgent`—evaluate each gate without requiring any large language model or learned component. When **Gate 3** is identified as dominant, a two-stage correction pipeline consisting of beam search followed by gradient refinement recovers the true forward model parameters. Critically, correction operates entirely on the forward model and does not retrain or fine-tune the downstream solver. Across 7 distinct modalities (9 correction configurations, including two CASSI algorithms and the Matrix baseline; with 7 additional configurations registered for future validation), autonomous correction yields improvements ranging from +0.54 dB to +48.25 dB. In every validated modality, **Gate 3** is identified as the dominant failure gate, confirming that operator mismatch—not solver weakness or noise—is the principal bottleneck in modern computational imaging.

The Triad Law

The TRIAD LAW 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_{nom}, but the data were generated by a true operator H_{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 PSNR_{I} denote reconstruction quality under ideal conditions (true operator, high SNR), PSNR_{II} under mismatch conditions (nominal operator applied to data generated by the true operator), and PSNR_{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. 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 degradation comparable to halving the acceleration factor. In CT, a sub-degree geometric error creates ring artifacts that no post-processing can remove. The TRIAD LAW 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 LAW 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[13] forward model is represented as Source $\rightarrow$ MaskModulation $\rightarrow$ SpectralDispersion $\rightarrow$ SensorIntegration $\rightarrow$ PoissonNoise. MRI[7] becomes Source $\rightarrow$ CoilSensitivity $\rightarrow$ FourierEncoding $\rightarrow$ Undersampling $\rightarrow$ GaussianNoise. CT[16] is compiled as Source $\rightarrow$ RadonProjection $\rightarrow$ DetectorResponse $\rightarrow$ 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 Law 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. For the 64 modalities compiled in this work, Tier 1 and Tier 2 models suffice for diagnostic purposes; Tier 3 and Tier 4 are reserved for high-fidelity correction refinement.

Scale and validation. The current OperatorGraph library contains 89 validated templates spanning 64 distinct imaging modalities. 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). All 89 templates (composed of linear primitives at Tier 1 and Tier 2) pass all three checks. 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 Law. 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 $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)$ associated with the offending operator node. The parameter family is declared in the OperatorGraph template (*e.g.*, lateral shift dx, dy and rotation θ 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 $\boldsymbol{\psi}$ via backpropagation through the OperatorGraph DAG. The loss function combines a data-fidelity term $\|\mathbf{y} - H(\boldsymbol{\psi})\hat{\mathbf{x}}\|^2$ 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 $H(\hat{\boldsymbol{\psi}})$ 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_{true} with high SNR, establishing the performance ceiling. **Scenario II** (Mismatch): the solver reconstructs using the nominal operator H_{nom} applied to data generated by H_{true}, quantifying the mismatch penalty. **Scenario III** (Corrected): the solver reconstructs using the PWM-corrected operator $H(\hat{\boldsymbol{\psi}})$, measuring correction effectiveness. **Scenario IV** (Oracle Mask): the true operator H_{true} is used for reconstruction on data generated by the mismatched system, providing the upper bound on what any correction algorithm can achieve (the correction ceiling).

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

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

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). 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 across 7 distinct modalities (9 correction configurations, including two CASSI algorithms and the Matrix pipeline-consistency test, which shares the SPC operator template; 16 registered configurations total) and a broader 26-modality benchmark suite. All experiments use the 4-Scenario Protocol described above. Reconstruction quality is primarily measured by peak signal-to-noise ratio (PSNR in dB); SSIM and spectral angle mapper (SAM) values are recorded in the RunBundle manifests.

16-modality 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 $\Delta_{\mathrm{corr}} = \mathrm{PSNR}_{\mathrm{III}} - \mathrm{PSNR}_{\mathrm{II}}$ ranges from +0.54 dB (CASSI Alg 1) to +48.25 dB (accelerated MRI, where a coil sensitivity mismatch is severe). The validated modalities span photon-domain systems—CASSI (+0.76 dB oracle upper bound with GAP-TV; up to +6.50 dB with MST-L), CACTI (+22.94 dB), SPC (+12.21 dB), 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 LAW 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, θ=0.1°, a_1=2.02, α=0.15°). Under Scenario I (Ideal), GAP-TV[17] achieves 24.34 ± 1.90 dB (mean across 10 KAIST scenes), MST-L[5] achieves 34.81 dB, and HDNet[18] 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). 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)[19] exhibits the same pattern. The state-of-the-art method EfficientSCI[20] achieves 35.33 dB under ideal conditions but drops to 14.48 dB under mask mismatch—a loss of 20.85 dB. PWM correction recovers 22.94 dB, reaching 37.42 dB (Scenario III), corresponding to a recovery ratio of $\rho > 1.0$ (i.e., the corrected reconstruction slightly exceeds the ideal-condition baseline due to regularization benefits). The CACTI corrected PSNR (37.42 dB) exceeds the Scenario I ideal (35.33 dB), yielding $\rho > 1$. This occurs because the corrected operator provides implicit regularization that is absent in the ideal case—a phenomenon analogous to beneficial model mismatch in robust estimation. This is the second-largest correction gain among validated modalities. 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)[21] imaging presents a qualitatively different mismatch type: gain bias rather than geometric shift. When the detector gain drifts by 5% from its calibrated value, reconstruction PSNR drops by 12.21 dB. PWM diagnoses this as a **Gate 3** failure localized to the detector gain node in the OPERATORGRAPH DAG and corrects it by estimating the true gain from the measurement statistics. Correction recovers the full 12.21 dB, achieving $\rho = 1.0$.

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.

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 particle-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.

26-modality benchmark. Beyond the 16 registered correction configurations (of which 9 are fully validated across 7 distinct modalities), we compile a broader benchmark of 26 modalities for which the OperatorGraph template and adjoint check have been established; 8 have full Scenario I baselines with validated PSNR, while the remainder are in Phase 2 or Phase 4 validation (see Supplementary Table S3). All 26 modalities pass the automated validation suite (adjoint consistency, gradient flow, dimensional consistency). Among the 8 fully validated modalities, Scenario I PSNR values range from 24.09 dB (CT) to 55.19 dB (MRI). This benchmark establishes the breadth of the OperatorGraph IR and provides a foundation for scaling pwm to the full 64-modality target.

Discussion

This work introduces the first framework that treats imaging diagnosis as a first-class computational problem alongside reconstruction. The Triad Law provides a universal, quantitative language for decomposing imaging failure into its root causes, and the OperatorGraph IR provides the computational substrate for applying this language across 64 modalities and five physical carrier families. The empirical finding that **Gate 3** dominates in all validated modalities carries a clear implication for the field: 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 practical implications are substantial. In clinical MRI, even small coil sensitivity mismatches can produce diagnostic artifacts; 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 fundamentally information-limited or correctable through model refinement. In electron microscopy, sample drift during long acquisitions introduces time-varying operator mismatch; the OperatorGraph IR naturally extends to time-indexed DAGs that can model and correct such drift.

Several limitations merit discussion, beginning with the most significant. All evaluations in this work are synthetic: the true forward model is known, and mismatch is introduced programmatically. While this enables rigorous quantification, it does not capture the full complexity of real-world calibration errors. Hardware-in-the-loop validation is the essential next step. Second, the forward models used for many non-photon modalities are simplified (Tier 1 or 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 mismatch types that are not anticipated in the OperatorGraph template. Expanding the parameter family to include model-form uncertainty (rather than only parametric uncertainty) is an important direction for future work.

Looking forward, we envision three extensions. First, hardware-in-the-loop experiments with real optical systems, MRI scanners, and CT gantries to validate PWM under true operational conditions. Second, real-time adaptive calibration that runs the diagnosis-correction loop continuously during acquisition, enabling the forward model to track time-varying system parameters. Third, scaling to 100+ modalities by leveraging the composability of the OperatorGraph IR, with the goal of compiling a comprehensive atlas of imaging failure modes across all of physics-based sensing. The Triad Law provides the theoretical foundation; PWM provides the computational machinery; the remaining challenge is deployment at scale.

Acknowledgements. We thank the open-source computational imaging community for making reconstruction code and datasets publicly available. This work was supported by NextGen PlatformAI C Corp.

Author Contributions. C.Y. conceived the project, designed the Triad Law framework, developed the OperatorGraph IR, implemented the agent system, performed all experiments, and wrote the manuscript.

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

Data Availability. All synthetic measurement data used in this study can be regenerated using the OperatorGraph templates and mismatch parameters specified in the Supplementary Information. The KAIST hyperspectral dataset[6] used for CASSI experiments is publicly available.

Code Availability. The PWM codebase, including all OPERATORGRAPH templates, agent implementations, and evaluation scripts, is available at `https://github.com/integritynoble/Physics_World_Model` under the MIT license.

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_{trials}, $\delta_{\max}$, and $\bar{\delta}$ for audit. All 89 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 $A^* = A^*_{|\mathcal{V}|} \circ \cdots \circ A^*_1$ 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 89 templates organized across 64 modalities, 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 Law Formalization

The TRIAD LAW 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_{nom} and the true physical operator H_{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 ℓ_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 comparing 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 PSNR_{I} is the reconstruction PSNR under Scenario I (ideal operator), PSNR_{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 ρ 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_{noise} and C_{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[18], 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 θ) and dispersion parameters (slope a_1 and axis angle α); an optional sixth parameter, PSF width σ_{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, θ) 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, α) is searched over a 5×7 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 ± 0.1–0.2 pixels for spatial parameters and $\pm 0.05°$ 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, θ) with $9\times9\times7 = 567$ 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 ± 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_{true}. 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_{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_{true} instead of H_{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[6], 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 256×256 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 256×256, with 28 spectral bands dispersed along the spatial dimension. The five-parameter mismatch model $\boldsymbol{\psi} = (dx, dy, \theta, a_1, \alpha)$ describes: mask spatial shift in x (dx, pixels), mask spatial shift in y (dy, pixels), mask rotation angle (θ, degrees), dispersion slope (a_1, pixels per band), and dispersion axis angle (α, degrees). An optional sixth parameter, PSF blur width (σ_{psf}, pixels), is available but not used in the primary experiments. For the primary mismatch experiment (validated by InverseNet), the true mismatch parameters are $\boldsymbol{\psi}_{\mathrm{true}} = (dx = 0.5\text{ px},\ dy = 0.3\text{ px},\ \theta = 0.1°,\ a_1 = 2.02,\ \alpha = 0.15°)$. Solvers evaluated include TwIST[22], GAP-TV[17], DGSMP[23], MST-L[5], and CST-L[24], all of which receive the same operator and differ only in their reconstruction algorithm. The supplementary per-scene analysis additionally includes DeSCI[25] and HDNet[18].

CACTI configuration. The coded aperture compressive temporal imaging system uses binary temporal masks of dimensions 256×256, 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 GAP-TV with total-variation regularization.

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 a multiplicative gain bias on the measurement matrix. The default solver is ADMM-TV with total-variation regularization and a wavelet sparsifying transform.

MRI configuration. Cartesian k-space sampling with $4\times$ 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[15] with ℓ_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)[16] with a Ram-Lak filter, supplemented by iterative SART for comparison.

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 ± 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 ρ (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 ρ. 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:

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

where k indexes the mismatch parameter, i indexes the scene, and N_{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 ΔPSNR per modality and verify that each component contributes ≥ 0.5 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 MIT license 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] Candès, E. J. & Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Processing Magazine* **25**, 21–30 (2008).

[2] Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory* **52**, 1289–1306 (2006).

[3] Venkatakrishnan, S. V., Bouman, C. A. & Wohlberg, B. Plug-and-play priors for model based reconstruction. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 945–948 (2013).

[4] Monga, V., Li, Y. & Eldar, Y. C. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine* **38**, 18–44 (2021).

[5] Cai, Y. *et al.* Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17502–17511 (2022).

[6] Choi, I., Jeon, D. S., Nam, G., Gutierrez, D. & Kim, M. H. High-quality hyperspectral reconstruction using a spectral prior. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* **36**, 218:1–218:13 (2017).

[7] Lustig, M., Donoho, D. & Pauly, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine* **58**, 1182–1195 (2007).

[8] Zbontar, J. *et al.* fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839* (2018).

[9] Chen, H. *et al.* Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging* **36**, 2524–2535 (2017).

[10] Uecker, M. *et al.* ESPIRiT — an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. *Magnetic Resonance in Medicine* **71**, 990–1001 (2014).

[11] Maiden, A. M. & Rodenburg, J. M. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy* **109**, 1256–1262 (2009).

[12] Antun, V., Renna, F., Poon, C., Adcock, B. & Hansen, A. C. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences* **117**, 30088–30098 (2020).

[13] Wagadarikar, A. A., John, R., Willett, R. & Brady, D. J. Single disperser design for coded aperture snapshot spectral imaging. *Applied Optics* **47**, B44–B51 (2008).

[14] Rodenburg, J. M. & Faulkner, H. M. L. A phase retrieval algorithm for shifting illumination. *Applied Physics Letters* **85**, 4795–4797 (2004).

[15] Pruessmann, K. P., Weiger, M., Scheidegger, M. B. & Boesiger, P. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine* **42**, 952–962 (1999).

[16] Feldkamp, L. A., Davis, L. C. & Kress, J. W. Practical cone-beam algorithm. *Journal of the Optical Society of America A* **1**, 612–619 (1984).

[17] Yuan, X. Generalized alternating projection based total variation minimization for compressive sensing. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2539–2543 (2016).

[18] Hu, X. *et al.* HDNet: High-resolution dual-domain learning for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17542–17551 (2022).

[19] Llull, P. *et al.* Coded aperture compressive temporal imaging. *Optics Express* **21**, 10526–10545 (2013).

[20] Wang, L., Cao, M. & Yuan, X. 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)*, 18477–18486 (2023).

[21] Duarte, M. F. *et al.* Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine* **25**, 83–91 (2008).

[22] Bioucas-Dias, J. M. & Figueiredo, M. A. T. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing* **16**, 2992–3004 (2007).

[23] Huang, T., Dong, W., Yuan, X., Wu, J. & Shi, G. Deep gaussian scale mixture prior for spectral compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16216–16225 (2021).

[24] Cai, Y. *et al.* CST: Compact spectral transformer for hyperspectral image reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2022).

[25] Liu, Y., Yuan, X., Suo, J., Brady, D. J. & Dai, Q. Rank minimization for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 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 Law 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 (particle). 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: 89 templates, 64 modalities, 5 carrier families.

Figure 3 | Triad Law structure and gate binding. **a**, Decision tree for the Triad Law: 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 ρ distribution across all 9 correction configurations.

Figure 4 | Correction results across 9 validated configurations. Bar chart showing correction gain Δ_{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: EfficientSCI across 4 scenarios, showing 20.85 dB mismatch degradation and $\rho > 1.0$ (full recovery with regularization benefit). **c**, Example reconstructed spectral datacubes: Ideal, Mismatched, and Corrected.

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.