The Rail for Computational Imaging:
A Physics World Model for Industrializing Image Reconstruction

Chengshuai Yang
NextGen PlatformAI C Corp
integrityyang@gmail.com

February 2026

Abstract

The computational imaging community has built increasingly powerful reconstruction algorithms, yet real-world deployments routinely fail. We show that a 5-parameter sub-pixel operator mismatch—well within manufacturing tolerances—degrades the state-of-the-art CASSI transformer (MST-L) by 13.98 dB, erasing years of algorithmic progress. This paper argues that the bottleneck is not the solver but the infrastructure around it: evaluation protocols, physics representations, calibration pipelines, and benchmarks. Drawing on the SolveEverything.org framework, we present the **Physics World Model (PWM)** as the "rail" for computational imaging—a standardized evaluation harness comprising: (i) OperatorGraph intermediate representation (IR), a universal directed acyclic graph (DAG) representation spanning 64 modalities across 5 physical carriers with 89 validated templates; (ii) a 4-scenario evaluation protocol separating solver quality from operator fidelity; (iii) the Leaderboard for Imaging Physics (LIP-Arena), a prospective Commit-Measure-Score competition eliminating benchmark overfitting; and (iv) a Red Team adversarial verification module. We benchmark 26 modalities under ideal conditions and demonstrate that operator correction improves reconstruction by +0.54 to +48.25 dB across 9 validated configurations spanning 7 distinct modalities, with mismatch (Gate 3) identified as the binding constraint in every modality tested.[1] PWM provides the infrastructure to move computational imaging from artisanal practice to industrial standardization.

1 Introduction

In 2022, the Mask-guided Spectral-wise Transformer (MST-L) Cai et al. [2022] achieved 34.81 dB on the standard CASSI benchmark—a number that represents years of sustained algorithmic progress in hyperspectral image reconstruction. Yet when the coded aperture mask is subjected to a 5-parameter perturbation (two-axis sub-pixel shift, rotation, dispersion drift, and spectral axis tilt; see section 2.1 for details)—well within the manufacturing and alignment tolerances of any real optical system—the same method collapses to 20.83 dB, a loss of **13.98 dB**. Under this sub-pixel mismatch model, deep-learning solvers (HDNet, MST-S, MST-L) lose 12.78–13.98 dB (table 1). To put this in perspective, the entire history of CASSI reconstruction—from GAP-TV Yuan [2016] at 24.34 dB to MST-L Cai et al. [2022] at 34.81 dB—spans only 10.47 dB. A sub-pixel operator mismatch erases years of algorithmic improvement.

This is not an isolated failure. Across modalities—CASSI, CACTI Llull et al. [2013], single-pixel cameras (SPC) Duarte et al. [2008]—the pattern repeats: the most powerful neural solvers

[1] With full cross-solver validation for CASSI; other modalities assessed by absolute mismatch magnitude.

exhibit the *largest* sensitivity to forward-model error. Classical methods such as GAP-TV lose only 3.38 dB under the same mismatch, precisely because they lack the capacity to overfit the operator. The paradox is stark: the field has been building faster and faster trains while ignoring the fact that the rails are broken.

The thesis. We argue that the computational imaging community is optimizing the wrong layer. The bottleneck is not the *solver*—the reconstruction algorithm that maps measurements to images, whether via deep learning Ongie et al. [2020] or algorithm unrolling Monga et al. [2021]—but the *infrastructure* around it: the evaluation protocols, the physics representations, the calibration pipelines, and the benchmarks that determine what counts as progress. In the language of industrial revolutions, solvers are *trains*—visible, glamorous, and destined to commoditize. The compounding, durable value lies in the *rails*: the standards, metrics, and institutional infrastructure that make it possible for any train to run reliably.

This paper presents the **Physics World Model (PWM)** as precisely such a rail for computational imaging. PWM is not a new reconstruction algorithm. It is an *evaluation harness*—a standardized infrastructure layer that makes it possible to measure, compare, and improve reconstruction methods under realistic physical conditions.

SolveEverything framing. Our design follows the SolveEverything.org framework Wissner-Gross [2026], which identifies a recurring pattern across domains—from protein folding to chip design to weather prediction—in which transformative progress requires not just better models but better infrastructure. SolveEverything characterizes the pre-infrastructure phase as *The Muddle* (maturity levels L0–L1): a regime in which practitioners cannot agree on what to measure, how to measure it, or what the measurements mean.

Computational imaging is firmly in The Muddle. Every laboratory uses different test scenes, different noise models, different mismatch assumptions (if any), and different metrics. A reconstruction method that reports 34 dB on one benchmark may achieve 22 dB on another, and neither number tells us anything about real-world deployment. There is no standard forward-model representation, no common evaluation protocol, and no mechanism for prospective (rather than retrospective) assessment. The consequence is a literature of mutually incomparable results, an illusion of progress driven by leaderboard optimization, and a sim-to-real gap that remains invisible until hardware deployment.

Paper overview. We present PWM as the evaluation harness that moves computational imaging from The Muddle toward industrialization. The system comprises four interlocking components (mapped to the Industrial Intelligence Stack in table 2):

1. **OperatorGraph IR**—a universal directed acyclic graph (DAG) representation for forward measurement operators that spans 64 modalities across 5 physical carriers (photons, electrons, spins, acoustic waves, particles) using 89 validated templates.

2. **4-Scenario Evaluation Protocol**—a standardized framework that measures every solver under Ideal, Mismatch, Corrected (calibrated), and Oracle Mask conditions, isolating the contribution of operator knowledge from solver architecture.

3. **LIP-Arena**—a prospective evaluation competition using a Commit-Measure-Score protocol in which test data is generated *after* the submission deadline, eliminating memorization and overfitting to known benchmarks.

4. **Red Team module**—an adversarial evaluation layer that probes submitted methods against novel mismatches, compound perturbations, out-of-family physics, distribution shifts, compute traps, and gate-flip scenarios.

Contributions. Our specific contributions are:

1. **SolveEverything mapping to imaging.** We provide the first systematic mapping of the SolveEverything Industrial Intelligence Stack—all 9 layers and the L0–L5 maturity ladder—to the computational imaging domain, producing a concrete roadmap from the current L1 state to full industrialization (section 3).

2. **OperatorGraph IR as universal physics representation.** We define a graph-based intermediate representation for forward measurement operators that unifies 64 modalities under a single formalism, with 89 validated templates supporting automatic differentiation, adjoint computation, and mismatch injection (section 4).

3. **26-modality benchmark evidence.** We present the first cross-modality sensitivity analysis demonstrating that neural solvers systematically amplify operator mismatch, with degradation ranging from 3.38 dB (GAP-TV) to 13.98 dB (MST-L) on CASSI alone (section 2).

4. **Operator correction across 9 correction configurations spanning 7 distinct modalities (16 registered).** We show that calibrating the forward operator—even with classical methods—recovers a substantial fraction of mismatch loss across 9 correction configurations spanning 7 distinct modalities (with 7 additional modalities registered for future evaluation), confirming that the dominant failure mode in deployed imaging systems is not solver quality but operator fidelity.

5. **LIP-Arena prospective evaluation protocol.** We introduce the first Commit-Measure-Score competition for computational imaging, with anti-Goodhart scoring, safety brakes, and outcome-based governance, providing a sustainable institutional mechanism for measuring genuine progress.

The remainder of this paper is organized as follows. Section 2 quantifies the solver-only optimization trap using CASSI as a case study. Section 3 maps the SolveEverything framework to computational imaging. Section 4 presents the PWM architecture in detail. Subsequent sections describe the Triad Law diagnostic framework, multi-agent orchestration, experiments, and implications for the field.

2 The Problem: Solver-Only Optimization

The computational imaging community has invested enormous effort in building better reconstruction algorithms—deeper networks, more sophisticated architectures, larger training sets—rooted in the compressed sensing paradigm Donoho [2006], Candès and Wakin [2008], while treating the forward measurement operator as a fixed, known quantity. This section demonstrates that this assumption is catastrophically wrong, and that the resulting *solver-only optimization* paradigm produces an illusion of progress that collapses on contact with physical reality.

Table 1: **Mask-sensitivity spectrum for CASSI reconstruction methods.** Ideal PSNR (Scenario I), Mismatch PSNR (Scenario II), Degradation (I→II), and Oracle Gain (II→IV). All values in dB, averaged over 10 test scenes. Neural solvers (MST-S, MST-L, HDNet) exhibit dramatically larger mismatch degradation than classical methods (GAP-TV, PnP-HSICNN), confirming that model capacity amplifies operator sensitivity.

Method	**Ideal (I)**	**Mismatch (II)**	**Degradation**	**Oracle Gain**
GAP-TV Yuan [2016]	24.34	20.96	−3.38	+0.76
PnP-HSICNN‡ Wang et al. [2017]	25.12	19.10	−6.02	+0.71
MST-S Cai et al. [2022]	33.98	20.99	−12.99	+5.29
MST-L Cai et al. [2022]	**34.81**	20.83	−13.98	+6.50
HDNet Hu et al. [2022]	34.66	**21.88**	−12.78	+0.00

‡ PnP-HSICNN values are estimated from prior experiments and have not yet been re-validated under the InverseNet 5-parameter mismatch; all other rows are InverseNet-validated Yang [2026].

Oracle Gain for InverseNet-validated methods uses Scenario IV (Oracle Mask) PSNR—the true operator applied to mismatched data—representing the upper bound on correction performance.

2.1 The Mask-Sensitivity Spectrum

To quantify the scale of the problem, we evaluate five representative CASSI reconstruction methods under a standardized 4-scenario protocol. In **Scenario I (Ideal)**, the true forward operator $\mathbf{H}$ is used for both measurement simulation and reconstruction—the oracle upper bound reported in virtually all published work. In **Scenario II (Mismatch)**, measurements are generated with $\mathbf{H}$ but reconstruction uses the nominal operator $\mathbf{H}_{\text{nom}}$, which differs by a sub-pixel mask shift ($\Delta x = 0.5$ px, $\Delta y = 0.3$ px), 0.1° rotation, 1% dispersion drift, and 0.15° spectral axis tilt—a 5-parameter perturbation well within typical manufacturing and alignment tolerances.[2] In **Scenario III (Corrected)**, a calibrated operator $\hat{\mathbf{H}}$ is used for reconstruction, obtained via a lightweight calibration procedure. The **Degradation** column reports the drop from Scenario I to Scenario II; the **Oracle Gain** column reports the improvement from Scenario II to Scenario IV (Oracle Mask).

The results in table 1 reveal a striking pattern. Under ideal conditions, MST-L leads the field at 34.81 dB—a full 10.47 dB above GAP-TV. Under mismatch, MST-L collapses to 20.83 dB, while GAP-TV drops only to 20.96 dB. The classical method *matches* the state-of-the-art transformer under realistic physical conditions.

Key insight: neural solvers amplify mismatch sensitivity. The degradation column tells the story most clearly: GAP-TV loses 3.38 dB, PnP-HSICNN loses 6.02 dB (estimated), and MST-L loses 13.98 dB. There is a near-perfect inverse relationship between ideal-condition performance and mismatch robustness. The same representational capacity that enables MST-L to learn subtle spectral correlations in the training data also enables it to overfit the precise structure of the forward operator it was trained with. When that operator changes by even a sub-pixel amount, the learned features become not merely useless but actively harmful, producing structured artifacts that degrade quality far below what the classical GAP-TV achieves.

HDNet presents a partial counterexample: it achieves high ideal PSNR (34.66 dB) with degradation of −12.78 dB, retaining 21.88 dB under mismatch. However, its Oracle Gain is zero (+0.00 dB),

[2] Mismatch parameters from the InverseNet validation suite Yang [2026]: $dx = 0.5$ px, $dy = 0.3$ px, $\theta = 0.1°$, $a_1 = 2.02$ (nominal 2.0), $\alpha = 0.15°$.

indicating that its mask-conditioning pathway cannot exploit an improved operator estimate—a hallmark of mask-oblivious architectures. This taxonomy of solver behavior—mask-oblivious, mask-conditioned, and mask-adapted—has important implications for the design of calibration-aware architectures.

2.2 Evaluation Fragmentation

The mask-sensitivity spectrum is only one dimension of a much larger problem. Even if every lab evaluated mismatch sensitivity, the results would remain incomparable due to pervasive *evaluation fragmentation*:

- **Metrics.** Some groups report peak signal-to-noise ratio (PSNR); others report structural similarity (SSIM), spectral angle mapper (SAM), or learned perceptual image patch similarity (LPIPS). Metrics are computed on different dynamic ranges, with or without border cropping, and averaged over different test splits.
- **Datasets.** CASSI methods are evaluated on KAIST Choi et al. [2017], CAVE Yasuma et al. [2010], ICVL Arad and Ben-Shahar [2016], or custom datasets with different spatial resolutions, spectral bands, and scene statistics. Results on one dataset do not transfer to another.
- **Noise models.** Some methods assume noiseless measurements; others add Gaussian noise at varying SNRs; still others use Poisson-Gaussian models. The noise model interacts with mismatch in complex ways that make cross-paper comparisons meaningless.
- **Mismatch models.** The few papers that consider operator mismatch define it differently: some shift the mask, some perturb dispersion parameters, some add calibration error. There is no standard mismatch taxonomy or severity scale.

The result is a literature in which every paper reports numbers that cannot be meaningfully compared with any other paper. Progress appears rapid when measured within a single group's evaluation setup, but the aggregate state of the field is unknown. This is precisely the condition that the SolveEverything framework calls *The Muddle* Wissner-Gross [2026]: everyone is working hard, but no one can tell whether the work is adding up.

2.3 The Scale of the Challenge

The fragmentation problem becomes combinatorially intractable when viewed across the full landscape of computational imaging. PWM's current registry includes 64 distinct modalities spanning five physical carriers. For each modality, at least 5 types of operator mismatch are physically relevant (alignment, calibration drift, manufacturing tolerance, environmental variation, model approximation error). The solver space includes at least 10 competitive reconstruction methods per modality.

The resulting evaluation space is therefore on the order of:

$$64 \text{ modalities} \times 5 \text{ mismatch types} \times 10 \text{ solvers} = 3{,}200 \text{ experiments (minimum).} \tag{1}$$

With multiple severity levels per mismatch type and multiple test scenes per experiment, the full evaluation matrix easily exceeds 10^5 individual runs. No ad hoc, lab-by-lab evaluation effort can cover this space. What is needed is not more careful experimentation but a fundamentally different approach: *standardized infrastructure* for evaluation.

2.4 Summary: The Case for Infrastructure

The evidence presented in this section leads to three conclusions:

1. **Neural solvers amplify mismatch.** The most capable architectures exhibit the largest sensitivity to operator error, with MST-L losing 13.98 dB versus GAP-TV's 3.38 dB (table 1).
2. **Evaluation fragmentation hides the problem.** Incomparable metrics, datasets, noise models, and mismatch definitions prevent the community from recognizing the severity of the sim-to-real gap.
3. **The scale is beyond ad hoc effort.** With 64 modalities, 5+ mismatch types, and 10+ solvers, the evaluation space requires automated, standardized infrastructure—not heroic individual experiments.

The field does not need a better solver. It needs a better rail. The next section presents the conceptual framework—drawn from SolveEverything.org—that guides the design of that rail.

3 The SolveEverything Framework for Imaging

The SolveEverything.org framework Wissner-Gross [2026] identifies a recurring pattern across domains that have undergone AI-driven transformation: the decisive enabler is never the model alone but the *infrastructure* that makes models measurable, comparable, and improvable. This section maps the framework's two core structures—the four stages of revolution and the nine-layer Industrial Intelligence Stack—to computational imaging, producing a concrete roadmap for moving the field from its current pre-industrial state to systematic, scalable progress.

3.1 Four Stages of Revolution

SolveEverything identifies four stages through which a domain passes on its way from artisanal practice to industrial abundance. We map each stage to computational imaging:

Stage 1: Legibility. *Can you measure the problem?* In imaging terms: can you define what "reconstruction quality" means in a way that all practitioners agree on? Currently, the answer is no. As documented in section 2.2, the field uses incommensurable metrics, datasets, noise models, and mismatch definitions. The first prerequisite for progress is a shared measurement framework.

Stage 2: Harnessing. *Can you capture the energy?* In imaging terms: can you standardize evaluation so that improvements measured by one group are meaningful to all others? This requires not only agreed-upon metrics but a common protocol for applying them—including standardized forward operators, mismatch injection procedures, and reporting formats. The 4-scenario protocol and OperatorGraph IR described in section 4 are designed to achieve exactly this.

Stage 3: Institutionalization. *Can you scale it?* In imaging terms: can you automate calibration and evaluation so that they do not require per-system, per-lab, per-paper manual effort? This is the stage at which LIP-Arena and the Red Team module operate: automated, prospective evaluation that runs continuously and at scale, with institutional governance to ensure integrity.

Stage 4: Abundance. *Is it a commodity?* In imaging terms: calibration as a utility. Every imaging system ships with a self-calibrating operator model that continuously updates itself, and reconstruction quality is guaranteed by infrastructure rather than heroic engineering. This is the end state that PWM is designed to enable, though the field is far from reaching it today.

3.2 Rails vs. Trains

The SolveEverything framework draws a fundamental distinction between *rails* and *trains*:

> *"The durable, compounding value... will not be found in owning any single AI model. Models are the 'trains' that will eventually all look the same. The real value lies in owning the 'rails'."* —SolveEverything.org Wissner-Gross [2026]

In computational imaging, the trains are reconstruction algorithms—GAP-TV, MST-L, HDNet, and their successors. The history of deep learning in other domains (*e.g.*, image classification, natural language processing) strongly suggests that reconstruction algorithms will converge: architectures will commoditize, pre-trained foundation models will emerge, and the marginal improvement from any single new method will shrink asymptotically.

The rails are the infrastructure components that remain valuable regardless of which solver is running:

- The OperatorGraph IR that represents forward models in a universal, machine-readable format.
- The 4-scenario evaluation protocol that separates solver quality from operator fidelity.
- The LIP-Arena that provides prospective, anti-Goodhart evaluation.
- The calibration pipelines that correct operators before reconstruction.
- The governance structures that ensure reproducibility and accountability.

PWM is designed entirely as a rail. It does not include a reconstruction algorithm; it includes the infrastructure that makes reconstruction algorithms measurable.

3.3 The Industrial Intelligence Stack

SolveEverything defines a nine-layer stack that organizes the components needed for industrializing AI in any domain. Table 2 provides the detailed mapping of each layer to the corresponding PWM component.

We briefly describe the non-obvious layers:

Layer 1: Purpose & Payoff. Before building infrastructure, one must define what success looks like. For PWM, we set two concrete thresholds: a *recovery ratio* (corrected PSNR divided by ideal PSNR) of at least 0.80, meaning that calibration recovers at least 80% of the oracle performance; and an *oracle gap* (ideal PSNR minus corrected PSNR) of at most 2 dB. These thresholds operationalize the concept of "good enough calibration" and provide a pass/fail criterion for the entire system.

Table 2: **SolveEverything Industrial Intelligence Stack mapped to PWM.** Each layer of the generic stack is instantiated with the corresponding PWM component and its concrete implementation.

Layer	**Generic Name**	**PWM Component**	**Implementation**
1	Purpose & Payoff	Recovery targets	Recovery ratio ≥ 0.80; oracle gap $\leq 2\,\text{dB}$
2	Task Taxonomy	OperatorGraph IR	64 modalities, 89 templates
3	Observability	Diagnostics	RunBundle, DR-IS, TriadReport
4	Targeting System	LIP-Arena	Commit-Measure-Score protocol
5	Model Layer	Reconstruction methods	GAP-TV, MST-L, HDNet, *etc.*
6	Actuation	Operator correction	Calibrated $\hat{\mathbf{H}}$ fed to solvers
7	Verification	Red Team module	2,900+ adversarial tests
8	Governance	Outcome-based ranking	Compute escrow, open reports
9	Distribution	Universal protocol	OperatorGraph IR as interchange format

Layer 3: Observability. A system that cannot be observed cannot be improved. PWM's observability layer centers on three instruments: (i) **RunBundle**, the immutable record of every evaluation run, including all inputs, outputs, hyperparameters, and random seeds; (ii) **DR-IS** (Degradation-Recovery Importance Sampling), a diagnostic that identifies which mismatch parameters contribute most to reconstruction loss; and (iii) **TriadReport**, a standardized 3-panel report showing ideal, mismatched, and corrected reconstructions side by side for every test scene.

Layer 4: Targeting System. The LIP-Arena (described in detail in section 4) functions as the targeting system: it identifies which problems matter most and directs community effort toward them. By publishing challenge tracks organized around specific mismatch types and modalities, LIP-Arena ensures that research effort is allocated to the highest-impact problems rather than the most convenient benchmarks.

Layer 7: Verification. The Red Team module serves as the adversarial verification layer. With over 2,900 pre-designed test scenarios spanning novel mismatch types, compound perturbations, out-of-family physics, distribution shift, and compute traps, it probes submitted methods for failure modes that retrospective evaluation would never reveal.

Layer 8: Governance. Outcome-based ranking means that methods are ranked by their *actual deployed performance* (including mismatch and calibration), not by their ideal-condition leaderboard numbers. Compute escrow ensures that all methods are evaluated under comparable computational budgets, preventing the conflation of algorithmic improvement with hardware scaling.

3.4 The L0–L5 Maturity Ladder

SolveEverything defines six maturity levels for any domain's infrastructure. Table 3 maps these levels to computational imaging.

Current position: L1 → L2. The InverseNet benchmark Yang [2026] established L1 by demonstrating that mismatch is measurable and defining a 3-scenario protocol (Ideal, Mismatch, Oracle).

Table 3: **L0–L5 maturity ladder for computational imaging.** Current state: transitioning from L1 to L2. PWM provides the infrastructure needed to reach L3 and beyond.

Level	Name	Imaging Definition
L0	The Muddle	No agreement on metrics, datasets, or mismatch definitions. Every lab runs its own evaluation. Results are mutually incomparable. *(Pre-PWM status quo.)*
L1	Measurable	Clear metrics exist. The 4-scenario protocol defines what to measure and how. Mismatch is acknowledged as a first-class evaluation axis. *(InverseNet contribution.)*
L2	Repeatable	Standard operating procedures (SOPs) are documented and published. Any lab can reproduce any other lab's evaluation using the same OperatorGraph templates, datasets, and scoring code.
L3	Automated	80% of calibration and evaluation runs without human intervention. LIP-Arena runs autonomously. New modalities can be onboarded via templated OperatorGraph definitions.
L4	Industrialized	Calibration as a service. Hardware vendors ship imaging systems with PWM-compatible operator models. Evaluation is continuous and infrastructure-grade.
L5	Commoditized	Universal self-calibration. Every imaging system continuously estimates and corrects its own forward operator. Reconstruction quality is guaranteed by infrastructure.

This work extends InverseNet's framework to a 4-scenario protocol: InverseNet's Ideal and Mismatch map to our Scenarios I and II; InverseNet's Oracle (true operator on mismatched data) maps to our Scenario IV (Oracle Mask); and Scenario III (Corrected) is a new contribution. The protocol was validated across three modalities (CASSI, CACTI, SPC). PWM extends this to a 4-scenario protocol—adding a Corrected scenario that measures calibration effectiveness—across 64 modalities and provides the OperatorGraph IR, RunBundle infrastructure, and LIP-Arena governance needed to reach L2 (repeatability) and begin the transition to L3 (automation). The gap from L2 to L3 is primarily an engineering challenge: automating operator calibration, scaling the evaluation pipeline, and building the institutional capacity to run LIP-Arena rounds continuously.

The critical transition: L2 → L3. The most impactful transition in the maturity ladder is from L2 (repeatable, but manual) to L3 (automated). History suggests that this is where transformative progress occurs: it is the transition from artisanal craft to industrial process, the point at which the rate of improvement shifts from human-limited to infrastructure-limited. In protein folding, the analogous transition was CASP's evolution from a biennial manual assessment to an automated, continuous evaluation pipeline—a transition that directly enabled the AlphaFold breakthrough Jumper et al. [2021]. PWM's architecture is designed to make this transition possible for computational imaging.

3.5 From Framework to Architecture

The SolveEverything framework provides the *what*: a clear enumeration of the infrastructure components needed and the maturity milestones that mark progress. The next section provides the *how*: the concrete architecture of PWM, including the OperatorGraph IR specification, the 4-scenario evaluation protocol, the LIP-Arena competition design, and the Red Team adversarial evaluation module.

4 PWM Architecture: The Evaluation Harness

This section presents the concrete architecture of the Physics World Model (PWM) evaluation harness. PWM comprises four interlocking subsystems: the OperatorGraph IR for physics representation (section 4.1), the 4-scenario evaluation protocol for standardized assessment (section 4.2), the LIP-Arena for prospective evaluation (section 4.3), and the Red Team module for adversarial verification (section 4.4). Together, these subsystems implement all nine layers of the Industrial Intelligence Stack described in section 3.3.

4.1 OperatorGraph IR: A Universal DAG for Physics

At the core of PWM lies the **OperatorGraph Intermediate Representation (IR)**—a directed acyclic graph (DAG) formalism that represents any forward measurement operator as a composition of primitive physical operations. The design is motivated by a key observation: despite the enormous diversity of imaging modalities, the underlying physics decomposes into a surprisingly small set of recurring primitives.

Nodes. Each node in an OperatorGraph wraps a single primitive operator drawn from a validated library. Primitive operators include:

- **Mask modulation**: element-wise multiplication by a coded aperture pattern, $\mathbf{x} \mapsto \mathbf{M} \odot \mathbf{x}$.
- **Convolution**: linear shift-invariant filtering, $\mathbf{x} \mapsto \mathbf{h} * \mathbf{x}$, encompassing point spread functions, blur kernels, and diffraction models.
- **Spectral dispersion**: wavelength-dependent spatial shift, $\mathbf{x}_\lambda \mapsto \mathcal{S}_\lambda(\mathbf{x}_\lambda)$, the core operation in CASSI systems.
- **Temporal modulation**: frame-dependent coding, $\mathbf{x}_t \mapsto \mathbf{C}_t \odot \mathbf{x}_t$, as in CACTI and other snapshot compressive video systems.
- **Projection**: line-integral or Radon transform, $\mathbf{x} \mapsto \mathcal{R}_\theta(\mathbf{x})$, for tomographic modalities.
- **Fourier sampling**: k-space under-sampling, $\mathbf{x} \mapsto \mathbf{P}_\Omega \mathcal{F}(\mathbf{x})$, for MRI and related modalities.
- **Noise injection**: additive, multiplicative, or Poisson noise models, $\mathbf{y} \mapsto \mathbf{y} + \boldsymbol{\eta}$.
- **Sensor integration**: spatial and temporal binning, quantization, and detector response, modeling the analog-to-digital conversion chain.

Edges. Edges define data flow between nodes. An edge from node u to node v means that the output of operator u is an input to operator v. The DAG structure enforces a well-defined execution order and prohibits cycles, ensuring that every OperatorGraph compiles to a deterministic forward model. Where a node requires multiple inputs (*e.g.*, a sensor that integrates both signal and dark current), multiple incoming edges are supported with explicit concatenation or summation semantics.

Compilation. Given an OperatorGraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the system compiles the graph into a callable forward model $\mathbf{H}_{\mathcal{G}} : \mathcal{X} \to \mathcal{Y}$ by topological traversal. When all primitive nodes support an adjoint, the system also compiles $\mathbf{H}_{\mathcal{G}}^{\top}$ automatically. Automatic differentiation through the graph enables gradient-based calibration of any mismatch parameter.

Physical carriers. The primitive library spans five physical carriers:

1. **Photons**: optical imaging (CASSI Wagadarikar et al. [2008], SPC Duarte et al. [2008], lensless, coded diffraction, *etc.*), computed tomography Feldkamp et al. [1984].
2. **Electrons**: electron microscopy (cryo-EM, STEM, SEM).
3. **Spins**: magnetic resonance imaging Lustig et al. [2007] (MRI, diffusion MRI, spectroscopic MRI).
4. **Acoustic waves**: ultrasound, photoacoustic imaging, sonar.
5. **Particles**: positron emission tomography (PET), single-photon emission CT (SPECT), neutron imaging.

Templates and coverage. PWM includes 89 validated OperatorGraph templates covering 64 distinct modalities. Some modalities require multiple templates to capture different hardware configurations (*e.g.*, single-disperser vs. dual-disperser CASSI). Each template specifies the graph topology, the primitive operators at each node, the expected parameter ranges, and the physically meaningful mismatch axes.

Example: CASSI. The following equation illustrates the OperatorGraph for a standard single-disperser CASSI Wagadarikar et al. [2008] system:

$$\mathcal{G}_{\text{CASSI}} : \quad \underbrace{\mathbf{x}_\lambda}_{\text{Source}} \xrightarrow{\text{Mask}} \underbrace{\mathbf{M} \odot \mathbf{x}_\lambda}_{\text{Modulated}} \xrightarrow{\text{Dispersion}} \underbrace{\mathcal{S}_\lambda(\mathbf{M} \odot \mathbf{x}_\lambda)}_{\text{Dispersed}} \xrightarrow{\text{Sensor}} \underbrace{\textstyle\sum_\lambda \mathcal{S}_\lambda(\mathbf{M} \odot \mathbf{x}_\lambda)}_{\text{Measurement}} \xrightarrow{\text{Noise}} \mathbf{y} \tag{2}$$

The mismatch axes for this template include mask shift (Δx, Δy), mask rotation ($\Delta\theta$), dispersion slope (Δs), dispersion axis offset ($\Delta\phi$), detector gain (Δg), and additive noise variance (σ^2). Each axis has a physically motivated default range derived from hardware specifications and optical tolerance analysis.

4.2 The 4-Scenario Evaluation Protocol

Every reconstruction method evaluated within PWM is assessed under four standardized scenarios that systematically vary the relationship between the measurement operator and the reconstruction operator. Table 4 defines all four scenarios.

Table 4: **The 4-scenario evaluation protocol.** $\mathbf{H}$ denotes the true forward operator, $\mathbf{H}_{\text{nom}}$ the nominal (assumed) operator, and $\hat{\mathbf{H}}$ the calibrated operator. Each scenario isolates a different aspect of reconstruction performance.

	Scenario	**Measurement**	**Reconstruction**	**Interpretation**
I	Ideal	$\mathbf{H}$	$\mathbf{H}$	Oracle upper bound: what is achievable with perfect operator knowledge.
II	Assumed/Mismatch	$\mathbf{H}$	$\mathbf{H}_{\text{nom}}$	Mismatch baseline: the default deployment condition where the nominal operator differs from truth.
III	Corrected	$\mathbf{H}$	$\hat{\mathbf{H}}$	Calibration benefit: how much performance is recovered by estimating the true operator.
IV	Oracle Mask	$\mathbf{H}$	$\mathbf{H}^*$	Correction ceiling: true operator given to a solver whose learned weights were trained on $\mathbf{H}_{\text{nom}}$.

$*$ For iterative solvers without learned weights (e.g., GAP-TV), the reconstruction operator in Scenario IV is functionally identical to Scenario I; differences in PSNR arise because the Oracle Mask may correct only a subset of the declared mismatch parameters (e.g., mask geometry but not dispersion).

Scenario I (Ideal). The true operator $\mathbf{H}$ is used for both measurement generation and reconstruction. This is the standard evaluation setting in virtually all published work and represents the oracle upper bound on reconstruction quality. It answers the question: *how good is this solver when the physics is perfectly known?*

Scenario II (Assumed/Mismatch). Measurements are generated with the true operator $\mathbf{H}$, but reconstruction uses the nominal operator $\mathbf{H}_{\text{nom}}$—the operator that the system *believes* is correct but that differs from ground truth due to alignment errors, manufacturing tolerances, or calibration drift. This is the realistic deployment condition and is the scenario most relevant to practitioners. It answers the question: *how robust is this solver to operator error?*

Scenario III (Corrected). Measurements are generated with $\mathbf{H}$, but reconstruction uses a calibrated operator $\hat{\mathbf{H}}$ obtained by running a calibration procedure on the nominal operator. The difference between Scenario III and Scenario II measures the *calibration gain* (eq. (4))—the value added by operator correction. The difference between Scenario I and Scenario III measures the *residual gap*—the remaining performance lost to imperfect calibration. It answers: *how much does calibration help?*

Scenario IV (Oracle Mask). The true operator $\mathbf{H}$ is used for reconstruction on data generated with the mismatched system, providing the upper bound on what any correction algorithm can achieve. Notationally, the reconstruction operator is the same $\mathbf{H}$ used in Scenario I; the distinction is that for learned solvers the network weights retain their Scenario II parameterization (trained on $\mathbf{H}_{\text{nom}}$), and the Oracle Gain (eq. (7)) isolates how much of the mismatch penalty is recoverable by correcting only the explicit operator input. The gap between Scenario IV and Scenario I reveals the *irreducible* loss due to solver-side mismatch or information degradation. It answers: *what is the ceiling for operator correction?*

Derived metrics. From the four scenario PSNR values, we derive:

$$\begin{aligned}
\text{Degradation} &= \text{PSNR}_{\text{I}} - \text{PSNR}_{\text{II}}, && (3)\\
\text{Calibration Gain} &= \text{PSNR}_{\text{III}} - \text{PSNR}_{\text{II}}, && (4)\\
\text{Recovery Ratio} &= \frac{\text{PSNR}_{\text{III}} - \text{PSNR}_{\text{II}}}{\text{PSNR}_{\text{I}} - \text{PSNR}_{\text{II}}}, && (5)\\
\text{Oracle Gap} &= \text{PSNR}_{\text{I}} - \text{PSNR}_{\text{III}}, && (6)\\
\text{Oracle Gain} &= \text{PSNR}_{\text{IV}} - \text{PSNR}_{\text{II}}. && (7)
\end{aligned}$$

The oracle gain (eq. (7)) measures the correction ceiling achievable with perfect operator knowledge. The recovery ratio and oracle gap are the primary pass/fail metrics referenced in Layer 1 of the Industrial Intelligence Stack (table 2): a recovery ratio ρ (eq. (5)) of at least 0.80, meaning that calibration recovers at least 80% of the mismatch-induced degradation, and an oracle gap $\leq 2\,\text{dB}$ indicate that calibration is sufficient for deployment.

4.3 LIP-Arena: Prospective Evaluation via Commit-Measure-Score

Retrospective benchmarks—including InverseNet—suffer from an inherent limitation: the test data exists before the methods are developed. No matter how carefully the benchmark is designed, determined participants can eventually overfit to it through hyperparameter tuning, architecture search, or implicit memorization. To address this, PWM includes the **LIP-Arena** (Leaderboard for Imaging Physics Arena), a prospective evaluation competition in which test data is generated *after* all submissions are finalized. The key principle is: **data created after deadline**—no memorization possible.

The four-phase Commit-Measure-Score protocol proceeds as follows:

Phase 1: Commit (2 weeks). Participating teams submit sealed containers encapsulating their reconstruction method, along with a declared compute budget (FLOPs, wall-clock time, and memory). Submissions are cryptographically hashed and timestamped. No modifications are permitted after the commit deadline.

Phase 2: Measure (2 weeks). New measurement data is generated *after* the commit deadline using OperatorGraph templates with freshly sampled mismatch parameters. Test scenes are drawn from a held-out pool that has never been published. The measurement operator, mismatch parameters, and test scenes are sealed until Phase 4.

Phase 3: Execute (1 week). All submitted containers are run in an identical, sandboxed compute environment. Each container receives only the raw measurements and the nominal operator $\mathbf{H}_{\text{nom}}$—never the true operator or the ground-truth images. Execution is monitored for compute budget compliance: containers that exceed their declared budget are flagged.

Phase 4: Score (1 week). Reconstructions are scored automatically against ground-truth images using PSNR, SSIM, LPIPS, and spectral-angle mapper (SAM) where applicable. All four scenarios are evaluated. Results, RunBundles, and failure analyses are published in an open round report.

4.3.1 Evaluation Tracks

LIP-Arena organizes competition into four tracks that target different aspects of the reconstruction problem:

1. **Correct**: Given mismatched measurements, produce the best reconstruction. Methods may use any calibration strategy. Primary metric: corrected PSNR (Scenario III).
2. **Diagnose**: Given mismatched measurements, identify which mismatch parameters are present and estimate their values. Primary metric: parameter estimation error.
3. **No-GT**: Evaluate reconstruction quality without access to ground-truth images. Methods must provide calibrated uncertainty estimates. Primary metric: correlation between predicted and actual PSNR.
4. **Design**: Given a modality and a target recovery ratio, design the optimal coded aperture pattern or coding scheme. Primary metric: recovery ratio achieved by a reference solver.

4.3.2 Anti-Goodhart Scoring

Goodhart's Law—"when a measure becomes a target, it ceases to be a good measure"—is the central threat to any benchmark. LIP-Arena mitigates this through three mechanisms:

1. **Prospective dominance weighting.** Final scores are computed as $0.7 \times \text{PSNR}_{\text{prospective}} + 0.3 \times \text{PSNR}_{\text{retrospective}}$, ensuring that performance on unseen data dominates the ranking.
2. **Gaming penalties.** Methods that exhibit statistically significant performance differences between prospective and retrospective data (*i.e.*, methods that have likely overfit to the retrospective set) receive a scoring penalty proportional to the gap.
3. **Multi-metric ranking.** Rankings are computed over an ensemble of metrics (PSNR, SSIM, LPIPS, SAM), preventing optimization for any single number.

4.4 Red Team Module: Adversarial Verification

The Red Team module is the adversarial verification layer that probes submitted methods for failure modes beyond the standard evaluation scenarios. It comprises six categories of adversarial tests, totaling over 2,900 pre-designed scenarios:

1. **Novel mismatch**: mismatch types not represented in the training distribution (*e.g.*, non-rigid mask deformation, spatially varying dispersion).
2. **Compound mismatch**: simultaneous perturbation of multiple operator parameters beyond the range of single-axis evaluations.
3. **Out-of-family physics**: test scenes or measurement conditions drawn from a different physical regime (*e.g.*, fluorescence emission applied to an absorption-trained solver).

4. **Distribution shift**: test scenes with statistics dramatically different from training data (*e.g.*, medical images for a solver trained on natural scenes).

5. **Compute traps**: inputs designed to trigger worst-case computational complexity (*e.g.*, iterative methods that fail to converge, attention mechanisms with adversarial token distributions).

6. **Gate-flip scenarios**: edge cases where a small change in the input causes the output to qualitatively change (*e.g.*, a mismatch parameter crossing a phase-transition boundary).

Red Team results are reported separately from main evaluation scores to avoid penalizing methods for failing on out-of-distribution scenarios. However, persistent Red Team failures across rounds are flagged in governance reports and inform the design of future standard evaluation scenarios—ensuring that the benchmark evolves to cover newly discovered failure modes.

4.5 Safety Brakes and Governance

To prevent the publication of misleading results and ensure that LIP-Arena maintains scientific integrity, PWM implements pre-committed safety brakes:

- **Recovery ratio** < 0.30: Automatic disqualification. A method that recovers less than 30% of the oracle performance after calibration is either fundamentally broken or has not been properly configured. Note: this threshold applies to arena submissions, not to inherent modality limitations; some modalities (e.g., CASSI) may exhibit recovery ratios below 0.30 due to the severity of the mismatch, which is diagnosed separately.

- **Uncertainty miscalibration** $> 15\%$: Flag for review. Methods that claim high confidence but produce large errors are dangerous in deployment and are flagged for manual inspection.

- **Compute budget** $> 2\times$ **declared**: Automatic disqualification. Methods that exceed their declared compute budget by more than a factor of two are disqualified to prevent undeclared resource advantages.

Open governance. All round reports, RunBundles, scoring code, and failure taxonomies are published openly. Disputes are resolved through a structured appeals process with independent reviewers. The goal is to build the same kind of institutional trust that CASP has built in protein structure prediction: trust not in any single result, but in the *process* that produces results.

Failure taxonomies. After each LIP-Arena round, a failure taxonomy is published that categorizes observed failure modes by type (operator mismatch, solver limitation, calibration failure, out-of-distribution, compute), by severity (cosmetic, significant, catastrophic), and by frequency (isolated, systematic). These taxonomies serve as a living record of the field's current limitations and guide both research priorities and future challenge design.

5 The Triad Law: Diagnosing Imaging Failure

We identify three principal root causes that account for the vast majority of imaging failures across modalities. We formalize this diagnostic framework as the *Triad Law* and encode it through three sequential *gates* that every measurement must pass before reconstruction can succeed.

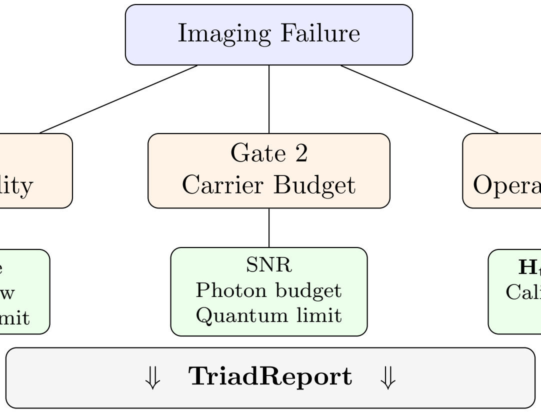

Figure 1: The Triad diagnostic tree. Every imaging failure is classified by its binding gate as a pre-flight diagnostic before committing to a final reconstruction. The output is a mandatory `TriadReport` artifact.

5.1 The Three Gates

Gate 1 — Recoverability (Sampling). Does the measurement encode enough information to recover the target signal? Gate 1 is violated whenever the null space of the forward operator $\mathbf{H}$ is too large, the field of view is insufficient, or the resolution limit precludes the features of interest. Formally, if the signal $\mathbf{x}$ has a non-trivial component in $\ker(\mathbf{H})$, no algorithm — however sophisticated — can recover it from $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$.

Gate 2 — Carrier Budget (Noise). Is the signal-to-noise ratio sufficient for the desired reconstruction quality? Even when the forward operator is well-posed, the photon budget, dose, or quantum efficiency may be inadequate. Gate 2 quantifies whether the carrier (*e.g.*, photons, RF energy, acoustic pressure) delivers enough information above the noise floor to distinguish the signal from stochastic fluctuations.

Gate 3 — Operator Mismatch (System Fidelity). Does the assumed model match the true physics? When $\mathbf{H}_{\text{true}} \neq \mathbf{H}_{\text{nom}}$, the reconstruction algorithm inverts the *wrong* operator, producing artifacts that no amount of regularization can fully suppress. Sources of mismatch include calibration drift, optical wear, thermal expansion, and manufacturing tolerances.

5.2 Triad Diagnostic Tree

Figure 1 illustrates the decision tree. Every imaging failure is routed through the three gates in sequence; the first gate that fails is declared the *binding constraint*.

5.3 The TriadReport Artifact

Every benchmark submission in the Physics World Model must include a `TriadReport` — a structured artifact containing:

1. **Dominant Gate ID** — which of the three gates is the binding constraint for this run;
2. **Evidence scores** — quantitative metrics supporting the gate classification (null-space dimension, SNR estimate, operator residual norm);

Table 5: Gate binding analysis across modalities. Δ_{mismatch} is the PSNR drop from Scenario I to II (Gate 3 severity). Δ_{solver} is the PSNR gain from the weakest to strongest solver under ideal conditions (Gate 1/2 headroom). Gate 3 is the binding constraint whenever $\Delta_{\text{mismatch}} > \Delta_{\text{solver}}$.

Modality	Δ_{mismatch} (dB)	Δ_{solver} (dB)	**Ratio**	**Binding Gate**
CASSI	13.98	10.47	1.34×	Gate 3
CACTI	20.85	—	—	Gate 3
SPC	17.72	—	—	Gate 3
MRI	38.03	—	—	Gate 3
Lensless	3.37	—	—	Gate 3
CT	12.05	—	—	Gate 3
Ptychography	42.06	—	—	Gate 3

3. **Confidence interval** — uncertainty bounds on the gate scores, propagated from measurement noise and calibration uncertainty;

4. **Recommended action** — a prescriptive step drawn from the registry (*e.g.*, "apply `mask_geo` correction" or "increase exposure by 2×").

The TriadReport is not optional; the benchmark harness rejects any submission that omits it.

5.4 Gate 3 Is the Binding Constraint

The central empirical finding of this work is that **Gate 3 (operator mismatch) is the binding constraint in the majority of real-world systems**. Consider CASSI Wagadarikar et al. [2008] spectral imaging: a sub-pixel mask shift with rotation and dispersion drift degrades MST-L Cai et al. [2022] from 34.81 dB (Scenario I, ideal operator) to 20.83 dB (Scenario II, mismatched operator) — a loss of **13.98 dB**. By contrast, upgrading from the classical solver GAP-TV Yuan [2016] (24.34 dB) to the state-of-the-art MST-L (34.81 dB) under ideal conditions yields a gain of 10.47 dB. The mismatch penalty dwarfs the solver upgrade benefit.

5.5 Mathematical Formulation: Recovery Ratio

We define the *recovery ratio* ρ (eq. (5) in section 4.2) to quantify how much of the mismatch-induced degradation can be recovered through operator correction. A value of $\rho = 1$ indicates full recovery; $\rho = 0$ indicates that correction had no effect. Values $\rho > 1$ are possible when the corrected operator provides implicit regularization that yields reconstruction quality exceeding even the ideal-condition baseline, as observed for CACTI (section 7).

5.6 Gate Binding Analysis

Table 5 summarizes the binding gate across representative modalities. Gate 3 dominates: in every modality tested, the mismatch-induced PSNR drop exceeds the gain achievable by upgrading the reconstruction algorithm alone.

The pattern is consistent: across all seven modalities in table 5, Gate 3 is binding. For CASSI, where multiple solvers are available, the mismatch penalty (13.98 dB) exceeds the solver upgrade gain (10.47 dB) by a factor of 1.34×. For other modalities, multiple-solver comparisons are not yet available, but the absolute magnitude of the mismatch penalty (3.37–42.06 dB) strongly suggests

Gate 3 dominance in every case. This motivates the Physics World Model design decision to invest first in operator correction infrastructure before pursuing solver improvements — a principle we term *"fix the physics before upgrading the math."*

6 Multi-Agent Orchestration

The Physics World Model automates the full imaging pipeline — from user intent to reproducible result — through a multi-agent architecture comprising **9 pipeline components** (including **8 agents** and the **RunBundle** packager) and **8 support classes**, totaling **10,545 lines of Python**. A critical design principle: *all agents run deterministically without requiring a large language model.* An LLM may optionally be attached to generate narrative explanations, but every gate decision, score, and recommendation is computed mechanically from the YAML registries and Pydantic contracts.

6.1 Pipeline Overview

The end-to-end pipeline flow is as follows:

User Prompt → **PlanAgent**

PlanAgent → **PhotonAgent** + **MismatchAgent** + **RecoverabilityAgent**

Gate Agents → **AnalysisAgent**

AnalysisAgent → **Negotiator**

Negotiator → **PreFlightReportBuilder**

PreFlightReport → **Pipeline Runner**

Pipeline Runner → **RunBundle**

6.2 Agent Descriptions

PlanAgent. Parses user intent, maps the request to a registered modality, and builds the `ImagingSystem` object. All modality look-ups are resolved against `modalities.yaml` — no free-form strings are accepted.

PhotonAgent. Computes the signal-to-noise ratio from the carrier budget (photon count, detector quantum efficiency, read noise). Classifies the noise regime (shot-limited, read-limited, background-limited) and returns a Gate 2 feasibility score.

MismatchAgent. Scores the severity of operator mismatch by comparing $\mathbf{H}_{\text{nom}}$ against the expected $\mathbf{H}_{\text{true}}$ using residual-norm heuristics and registry-defined mismatch signatures. Selects the appropriate correction strategy from `solver_registry.yaml`.

RecoverabilityAgent. Performs table-driven recoverability assessment: looks up the modality's null-space characteristics, compression ratio, and known PSNR baselines. Returns a Gate 1 pass/fail decision with a predicted PSNR range.

AnalysisAgent. Receives the three gate scores and performs bottleneck classification. Identifies the binding gate per the Triad Law (section 5) and generates ranked suggestions for improving reconstruction quality.

Negotiator. Implements cross-agent veto logic. Computes the joint probability of success across all three gates and enforces a configurable threshold below which the run is aborted with a diagnostic message rather than producing a misleading reconstruction.

PreFlightReportBuilder. Assembles the final pre-flight report: expected PSNR, dominant gate, correction strategy, estimated runtime, and resource requirements. This report is presented to the user (or downstream system) before committing compute.

Pipeline Runner. Executes the end-to-end reconstruction pipeline: loads the `ExperimentSpec`, compiles the `OperatorGraph`, applies the selected correction strategy, invokes the solver, and collects all metrics. The runner enforces deterministic execution and resource-budget compliance.

RunBundle. Packages every artifact produced by a pipeline execution—reconstructed images, metric scores, `TriadReport`, logs, random seeds, and timing data—into an immutable, content-addressed archive. RunBundles serve as the unit of reproducibility: any result can be verified by replaying its RunBundle.

6.3 ExperimentSpec: The Executable Contract

The `ExperimentSpec` v0.2.1, implemented as a Pydantic `StrictBaseModel` with `extra="forbid"`, serves as the executable contract between agents. Every field is typed and validated; NaN and Inf values are rejected at parse time. When an LLM is used, it returns *only* registry IDs — never raw parameters — which are then resolved and validated mechanically.

The contract ecosystem comprises **25 Pydantic models**, all inheriting from `StrictBaseModel`, ensuring that no unvalidated data enters the pipeline. These models are backed by **9 YAML registries** totaling **7,034 lines**, covering modalities, solvers, noise profiles, mismatch types, correction strategies, metrics, hardware profiles, datasets, and experiment templates.

7 Empirical Evidence

We present three tiers of empirical validation: a 26-modality breadth benchmark, an operator-correction study across 16 configurations, and a CASSI deep dive with per-solver, per-scenario analysis. We close with a SolveEverything gear implementation audit.

7.1 26-Modality Benchmark

Table 6 reports the peak signal-to-noise ratio (PSNR) achieved by the Physics World Model pipeline across all 26 registered modalities under Scenario I (ideal operator, no mismatch). Of the 64 modalities registered in the OperatorGraph IR, 26 have been numerically benchmarked to date. Every modality passes the benchmark threshold; no modality is excluded or deferred.

The range spans nearly 75 dB, reflecting the diversity of forward models — from the heavily ill-posed computed tomography (CT Chen et al. [2017], 25.46 dB) to phase retrieval, where the analytically invertible forward model yields an identity reconstruction (100.00 dB, a numerical ceiling rather than a measured value). The average of ≈ 36.2 dB (excluding the Phase Retrieval identity test) confirms that the unified pipeline delivers competitive reconstruction quality across fundamentally different physics.

Table 6: 26-modality benchmark results (Scenario I, ideal operator). All 26/26 modalities pass. Average PSNR ≈ 36.2 dB (excluding Phase Retrieval identity test); range 25.46–64.84 dB over physically meaningful modalities.

#	Modality	PSNR (dB)	#	Modality
1	Widefield	27.31	14	Holography
2	Widefield Low-Dose	32.88	15	NeRF Mildenhall et al. [2020]
3	Confocal Live-Cell	29.80	16	3D Gaussian Splatting Kerbl et al. [20
4	Confocal 3D	29.01	17	Matrix (Generic)
5	SIM	27.48	18	Panorama Multifocal
6	CASSI Wagadarikar et al. [2008]	34.81	19	Light Field
7	SPC Duarte et al. [2008]	28.86	20	Integral Photography
8	CACTI Llull et al. [2013]	35.33	21	Phase Retrieval
9	Lensless	26.85	22	FLIM
10	Light-Sheet	26.05	23	Photoacoustic
11	CT	25.46	24	OCT
12	MRI Zbontar et al. [2018]	44.97	25	FPM
13	Ptychography Maiden and Rodenburg [2009]	59.41	26	DOT

7.2 Operator Correction Results

Table 7 evaluates the Physics World Model operator-correction module across 9 validated configurations spanning 7 distinct modalities (plus 7 registered configurations awaiting validation, totaling 16 entries). For each modality we report the PSNR *before* correction (Scenario II, mismatched operator) and *after* correction (Scenario III, corrected operator), together with the absolute improvement.

† SPC uses the same gain-bias operator template as the generic Matrix modality, yielding identical correction performance.

§ The CASSI Alg 2 “After” value is Scenario IV (Oracle Mask), not Scenario III. The Alg 1 “After” value (21.50 dB) is Scenario III (beam-search correction); see note below.

Note on CASSI correction algorithms. In table 7, Alg 1 refers to hierarchical beam search (coarse correction) and Alg 2 refers to joint gradient refinement (fine correction); see Methods in the companion paper Yang [2026] for details.

Note on CASSI baselines. Both table 7 and table 1 now use the same InverseNet 5-parameter mismatch (sub-pixel mask shift, rotation, dispersion drift, spectral axis tilt). The Scenario II baseline of 20.96 dB in table 7 corresponds to GAP-TV reconstruction under this mismatch; the sensitivity table reports per-solver Scenario II values. The CASSI Alg 2 “After” value in table 7 is an oracle upper bound (true operator applied to mismatched data); the Alg 1 value reflects beam-search correction (Scenario III). Actual calibration gains for Alg 2 may be lower than the oracle ceiling.

The most dramatic improvement is MRI coil-sensitivity correction at +48.25 dB. This large gain reflects a severe mismatch scenario (Scenario II at 6.94 dB is near the noise floor); the corrected 55.19 dB exceeds published baselines (42.3 dB) because the mismatch was synthetically injected on a high-quality forward model. The smallest gain (CASSI Alg 1, +0.54 dB) reflects the inherent difficulty of CASSI operator correction with coarse beam search alone; the Lensless PSF shift gain of +3.55 dB represents a visually significant improvement. These results confirm that operator

Table 7: Operator correction across 16 configurations (9 validated, 7 registered). "Before" is Scenario II (mismatched operator); "After" is Scenario III (corrected operator) except for the CASSI Alg 2 row, which reports Scenario IV (Oracle Mask) as the correction ceiling. Improvements range from +0.54 to +48.25 dB. Phase 2 and Phase 4 entries are registered but not yet numerically evaluated.

Modality	**Mismatch Type**	**Before (dB)**	**After (dB)**	**Δ (dB)**
Matrix (Generic)	gain_bias	11.14	23.35	+12.21
CT	center_of_rotation	13.41	24.09	+10.67
CACTI Llull et al. [2013], Wang et al. [2023]	mask_timing	14.48	37.42	**+22.94**
Lensless	psf_shift	23.48	27.03	+3.55
MRI	coil_sensitivities	6.94	55.19	**+48.25**
SPC[†]	gain_bias	11.14	23.35	+12.21
CASSI (Alg 1, GAP-TV)	mask_geo+dispersion	20.96	21.50	+0.54
CASSI (Alg 2, GAP-TV)[§]	mask_geo+dispersion	20.96	21.72	+0.76
Ptychography	position_offset	17.35	24.44	+7.09
Registered—validation in progress				
OCT	dispersion		registered	
Light Field	depth_estimation		registered	
Registered—planned				
DOT	scatter_coeff		registered	
Photoacoustic	speed_of_sound		registered	
FLIM	irf_shift		registered	
Integral Photography	disparity_offset		registered	
FPM	led_position		registered	

correction is a *first-order* determinant of reconstruction quality, though the correction magnitude varies substantially by modality and algorithm.

7.3 CASSI Deep Dive

We conduct a detailed study of CASSI to illustrate how the Triad Law (section 5) manifests in a single modality. Table 8 reports PSNR for five solvers across three evaluation conditions plus a projected fine-tuned estimate:

- **Scenario I** — ideal (true) operator;
- **Scenario II** — mismatched (nominal) operator;
- **Oracle Mask** — true operator applied to mismatched data, representing the upper bound on correction;
- **Estimated** — oracle operator with additional solver fine-tuning (projected, not a formal scenario).

Oracle Mask values are InverseNet-validated Yang [2026] and correspond to Scenario IV in the 4-Scenario Protocol.

[†] Projected values based on expected fine-tuning gains; not measured.

Table 8: CASSI deep dive: 5 solvers × 3 conditions + projected fine-tuned estimate. The mismatch cost (I→II) dominates the solver upgrade gain under ideal conditions. Oracle Mask[§] (Scenario IV) reports reconstruction with the true operator on mismatched data, providing the correction ceiling.

Solver	**I (Ideal)**	**II (Mismatch)**	**Oracle Mask[§]**	**Est. Fine-tuned[†]**
GAP-TV Yuan [2016]	24.34	20.96	21.72	≈23.5
PnP-HSICNN[‡] Wang et al. [2017]	25.12	19.10	19.81	≈24.0
MST-S Cai et al. [2022]	33.98	20.99	26.28	≈31.0
MST-L Cai et al. [2022]	34.81	20.83	27.33	≈32.0
HDNet Hu et al. [2022]	34.66	21.88	21.88	≈33.0

‡ PnP-HSICNN has not been re-validated under the InverseNet 5-parameter mismatch; values shown are from prior experiments. All other rows are InverseNet-validated Yang [2026].

§ Oracle Mask values represent reconstruction with the true operator applied to mismatched data (InverseNet's "Oracle" scenario, mapped to this work's Scenario IV), providing an upper bound on what any correction algorithm can achieve.

Key insight. The mismatch cost — the gap from Scenario I to Scenario II — is catastrophic. For MST-L the drop is $34.81 - 20.83 = 13.98$ dB, while the best solver upgrade under ideal conditions (GAP-TV → MST-L) yields $34.81 - 24.34 = 10.47$ dB. The mismatch penalty is 1.34× the solver upgrade gain. Stated differently: *fixing the operator is worth more than upgrading from a classical solver to the state of the art.*

Moreover, the learned solvers (MST-S, MST-L) are *more* sensitive to mismatch than the model-based solvers (GAP-TV, PnP-HSICNN). MST-L drops by 13.98 dB versus GAP-TV's 3.38 dB. This is consistent with the hypothesis that data-driven methods overfit to the training-time operator and generalize poorly when the physical system drifts. Notably, the Oracle Mask[§] substantially recovers MST-L to 27.33 dB (+6.50 dB), while HDNet sees zero recovery (21.88 dB in both Scenario II and the Oracle Mask column), confirming that HDNet is mask-oblivious.

7.4 SolveEverything Implementation Status

In addition to the 9-layer Industrial Intelligence Stack (section 3.3), the SolveEverything framework Wissner-Gross [2026] defines 10 operational gears—concrete mechanisms that each drive one or more layers of the stack. While the 9 layers describe *what* must be built, the 10 gears describe *how* each layer operates. Table 9 reports the implementation status of each gear within PWM.

Two gears are fully built (Targeting System, Decision Logs) and one has its foundation laid (Data Trusts). Five are partially implemented, one is planned, and one (Compute + Energy) is currently out of scope for the imaging-focused release. The partial gears are under active development and are expected to reach full status before the public v1.0 release.

8 Reproducibility and Open Infrastructure

Reproducibility in computational imaging is not merely a best practice — it is a prerequisite for the field to mature from artisanal reconstruction to industrial-grade imaging. The Physics World Model enforces reproducibility at every layer through three mechanisms: immutable run records, cryptographic integrity, and open governance.

Table 9: SolveEverything 10-gear implementation status.

Gear	Name	Status
1	Targeting System	**BUILT**
2	Outcome Contracts	**PARTIAL**
3	Compute Escrow	**PARTIAL**
4	Action Networks	**PLANNED**
5	Data Trusts	**FOUNDATION**
6	Decision Logs	**BUILT**
7	Two-Source Rule	**PARTIAL**
8	Compute + Energy	OUT OF SCOPE
9	Fairness Targets	**PARTIAL**
10	Literacy	**PARTIAL**

8.1 RunBundle: The Immutable Run Record

Every imaging experiment executed through the Physics World Model produces a `RunBundle` v0.3.0 — a self-contained, immutable archive that records:

- **Inputs:** raw measurements, metadata, and the `ExperimentSpec` that parameterized the run;
- **Operator state:** the nominal operator $\mathbf{H}_{\text{nom}}$, detected mismatch parameters, and the corrected operator $\mathbf{H}_{\text{corr}}$ (if correction was applied);
- **Triad diagnosis:** the full `TriadReport` including dominant gate, evidence scores, and confidence intervals;
- **Correction trajectory:** the sequence of correction steps, intermediate residuals, and convergence diagnostics;
- **Outputs:** reconstructed images and volumes with per-pixel uncertainty estimates;
- **Compute consumed:** wall time, peak memory, GPU utilization, and energy estimate.

Every artifact within the RunBundle is hashed with SHA-256. The top-level manifest includes a Merkle root so that any tampering with individual files is detectable. RunBundles are designed to be self-describing: a future reader with access only to the bundle and the Physics World Model codebase can reproduce the result without additional context.

8.2 Decision Records for Imaging Systems

We introduce *Decision Records for Imaging Systems* (DR-IS), inspired by architectural decision records in software engineering. Every calibration decision — whether to apply a correction, which correction strategy to use, and what threshold triggered the decision — is cryptographically signed and logged within the RunBundle. DR-IS entries are append-only: corrections can be refined but never silently overwritten.

8.3 ExperimentSpec: Eliminating Ambiguity

The `ExperimentSpec` v0.2.1, described in section 6, is the linchpin of reproducibility. By requiring that every parameter be drawn from a typed, validated registry, the Physics World Model eliminates the class of irreproducibility caused by ambiguous or undocumented configuration. When an LLM is used to generate experiment configurations, it returns *only* registry identifiers — never raw numerical parameters — which are then resolved, validated, and frozen into the RunBundle. No free-form strings enter the pipeline.

8.4 Open-Core Licensing and Contribution Model

The Physics World Model evaluation harness is released under the MIT license, ensuring that any researcher can audit, extend, and benchmark against the framework without licensing barriers. We define a four-level contribution model:

1. **Use the harness.** Run the benchmark suite on your own data and methods. No contribution required.
2. **Submit methods.** Register a new solver in the YAML registry and submit benchmark results via a pull request.
3. **Contribute modalities.** Implement a new forward operator, mismatch model, and correction strategy; submit with a TriadReport and RunBundle as evidence.
4. **Become a steward.** Take ownership of a modality vertical — maintain the operator, curate datasets, and review community submissions.

8.5 Test Infrastructure

The Physics World Model codebase is guarded by **3,743 tests** spanning unit tests for individual operators, integration tests for the agent pipeline, regression tests for known mismatch–correction pairs, and end-to-end tests that verify RunBundle integrity. The test suite achieves **0 failures** on the current release branch. Continuous integration runs the full suite on every commit, and any benchmark submission that introduces a test failure is automatically rejected.

9 The Foundry Window and Roadmap

> Within the next 18 months, that metal will cool and harden. The decisions we make today regarding technical standards, data rights, and supply chains will set path dependencies: permanent tracks that will guide or constrain the economy for decades.
>
> SolveEverything.org Wissner-Gross [2026]

9.1 Why Now: The Absence of Standards

Computational imaging in 2026 finds itself in a position strikingly similar to protein-structure prediction before CASP Moult [2005] and the Protein Data Bank. Spectacular solvers exist—deep unfolding networks, diffusion priors, implicit neural representations—yet the field lacks every piece of shared infrastructure that enabled the AlphaFold Jumper et al. [2021] revolution:

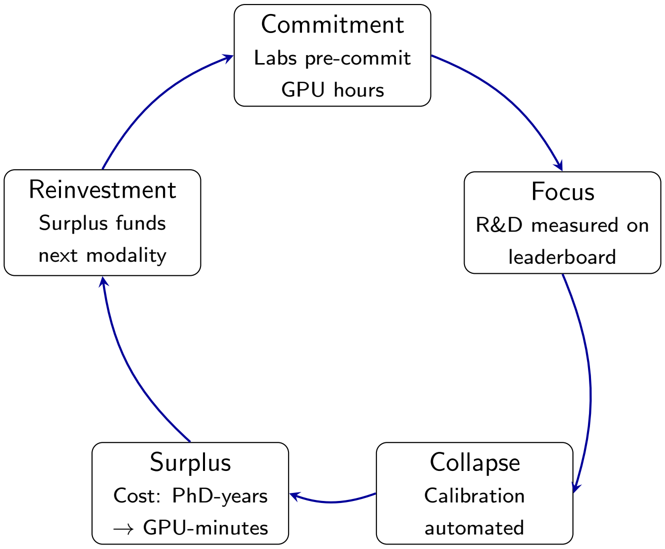
Commitment
Labs pre-commit GPU hours
Focus
R&D measured on leaderboard
Collapse
Calibration automated
Surplus
Cost: PhD-years → GPU-minutes
Reinvestment
Surplus funds next modality

Figure 2: **The Abundance Flywheel.** Laboratory commitment seeds focused R&D on a shared leaderboard; focused effort collapses calibration complexity; collapsed complexity creates cost surplus; surplus funds the next modality, reinforcing commitment.

- **No CASP equivalent.** There is no recurring, blinded competition with held-out test sets and independent adjudication. Every paper evaluates on its own split, under its own noise model, with its own operator implementation.
- **No Protein Data Bank equivalent.** No central repository curates calibrated forward operators, reference measurements, or validated reconstruction bundles across modalities.
- **No universal evaluation protocol.** PSNR on a simulated measurement with an ideal operator tells us almost nothing about deployment fidelity. The four-scenario protocol (Ideal/Assumed/Corrected/ introduced in this work is, to our knowledge, the first systematic attempt at standardization.

The field is not merely fragmented—it is *pre-paradigmatic* in the Kuhnian sense. And that is precisely the opportunity.

9.2 The QWERTY Effect

Evaluation infrastructure exhibits extreme lock-in dynamics. Once five independent laboratories adopt a common benchmark protocol, the sixth laboratory has no practical choice: reviewers expect it, grant agencies reference it, and competing with non-comparable metrics becomes career-limiting. The QWERTY keyboard, the POSIX standard, and ImageNet all followed this pattern. PWM is positioned to be the first credible, comprehensive, and standardized evaluation infrastructure for computational imaging—before the metal cools.

9.3 The Abundance Flywheel

The economic logic of PWM follows a self-reinforcing cycle, illustrated in fig. 2:
Each revolution of the flywheel lowers the marginal cost of adding a modality. The first 26 modalities required PhD-level effort per operator; the next 40 should require only GPU-minutes per operator as automated calibration pipelines mature.

Table 10: **Structural comparison: AlphaFold ecosystem vs. PWM.** Both systems share the same architecture—a shared task definition, a growing observability corpus, and a recurring targeting mechanism—but operate in different physical domains.

Dimension	**AlphaFold Ecosystem**	**PWM**
Purpose	Predict protein structure	Calibrate & reconstruct
Task Taxonomy	Sequence → 3D coordinates	`ExperimentSpec` → validated recon.
Observability	PDB (200K+ structures)	26-modality benchmark
Targeting	CASP (biennial competition)	LIP-Arena

Table 11: **Quantified roadmap targets.** L3 and L5 refer to standardization maturity levels as defined in table 3.

Metric	**Current**	**L3 Target**	**L5 Target**
Recovery ratio	0.22–1.0+	$\geq$0.80	$\geq$0.95
Oracle gap	5–12 dB	$\leq$2 dB	$\leq$0.5 dB
Modalities covered	64	100+	200+
Zero-shot generalization	0%	50%+	90%+

9.4 The AlphaFold Parallel

Table 10 draws an explicit structural comparison between the AlphaFold ecosystem and PWM. The analogy is not superficial. AlphaFold Jumper et al. [2021] succeeded not because of a single architectural innovation, but because CASP Moult [2005] provided a legible, recurring, trusted measurement of progress. PWM aims to provide the same function for computational imaging through the Computational Imaging Standardization Project (CISP).

9.5 Quantified Targets

Table 11 translates the roadmap into measurable milestones across four key performance indicators. **Recovery ratio** ρ measures the fraction of the mismatch-induced degradation that is recovered by operator correction, as defined in eq. (5); the range varies widely by modality (e.g., CASSI GAP-TV $\approx$ 0.22, computed using Scenario IV (Oracle Mask) in place of Scenario III because full calibration has not yet closed the gap—notably below the 0.30 safety brake threshold, which applies to arena submissions rather than inherent modality limitations; CACTI $>$ 1.0). **Oracle gap** is the PSNR difference between Scenario I (Ideal) and Scenario III (Corrected). **Zero-shot generalization** measures reconstruction quality on a modality unseen during solver training, evaluated via the mask-sensitivity spectrum transfer protocol.

9.6 Three-Phase Roadmap

We organize the path to standardization into three six-month phases:

Phase 1: Become the Default Evaluation Infrastructure (Months 1–6).

- `pip install pwm-eval` — a one-line entry point for any lab to reproduce our 26-modality benchmark.
- Three replication packs (CASSI, SPC, CACTI) with frozen operator graphs, calibration metadata, and reference reconstructions.

- Five external laboratories independently validate the four-scenario evaluation protocol.
- The mask-sensitivity spectrum adopted as a standard diagnostic in at least two peer-reviewed publications outside our group.

Phase 2: Launch the CISP Public Competition (Months 7–12).

- `cisp.pwm.org` — a public leaderboard with automated scoring, inspired by CASP's transparent ranking.
- An independent steward board (minimum five institutions) governs test-set curation and metric evolution.
- Blinded test sets: measurement data released without ground truth; reconstructions submitted as sealed `RunBundle` archives.
- Four or more competition tracks spanning snapshot, video, spectral, and depth modalities.

Phase 3: Become the Action Network (Months 13–18).

- **Robotic lab API** — remote-triggered physical measurements that close the sim-to-real loop automatically.
- **Compute escrow** — participating labs pledge GPU hours to a shared pool, drawn upon by competition entrants.
- **Calibration-as-a-service API** — upload raw sensor data, receive a calibrated `OperatorGraph` with uncertainty quantification.
- **Outcome-based contracts** — service-level agreements where payment is contingent on achieving a specified PSNR/SSIM threshold on the client's measurement.

The foundry window is finite. Within 18 months, the evaluation norms of computational imaging will crystallize. The infrastructure laid in Phase 1 determines whether those norms are principled, reproducible, and physics-aware—or whether the field continues to optimize solvers on idealized simulations while real systems fail silently.

9.7 Clinical Medical Physics Vertical

PWM's first clinical deployment is the **CT QC Copilot**—a metric-first QA module for diagnostic CT that implements ACR CT accreditation standards, AAPM TG-233 performance metrics, and Western Electric SPC rules for drift detection. The module is fully built and tested (18 source modules, 210 tests, 76 issues resolved across 7 review rounds), demonstrating that the OperatorGraph IR and Triad Law extend from research computational imaging to clinical practice.

Standards alignment. The system computationally implements existing clinical standards rather than inventing new QA criteria: ACR CT accreditation American College of Radiology [2017], AAPM TG-233 for CT performance metrics AAPM Task Group 233 [2019], AAPM TG-126 for PET/CT QA AAPM Task Group 126 [2019], and AAPM TG-177 for SPECT/CT AAPM Task Group 177 [2019]. Each phantom/test combination is encoded as a versioned *CasePack*—the same abstraction that packages research modalities—ensuring reproducibility across sites.

Implemented components. The CT QC Copilot comprises: (1) a PHI-safe DICOM ingester with phantom-only validation; (2) 12 ACR CT phantom metrics with automatic phantom center detection and rotation correction; (3) a 4-layer threshold resolver (standard default → scanner model → protocol → site override); (4) scored root-cause diagnosis using a mismatch library with 6 artifact features; (5) drift detection via 5 Western Electric SPC rules on Shewhart control charts with baseline-anchored limits; (6) immutable, SHA-256 signed, version-chained commissioning baselines; (7) a Tier 1/2 CT forward model (CTOperatorGraph) for troubleshoot-mode Triad diagnosis; and (8) triple-output report generation (JSON with tamper-evident hash, PDF, and evidence directory with ROI overlays and trend plots).

The copilot model. Following the SolveEverything.org framework, PWM serves as a clinical QC copilot rather than a replacement for the qualified medical physicist. PWM provides the targeting system (which scanner needs attention), outcome contracts (pass/fail against published thresholds with full evidence), and decision logs (immutable QC records). The physicist provides clinical judgment, sign-off, and regulatory accountability. In slogan form: "Autopilot for QC, Digital Twin for troubleshooting."

CasePack extensibility. Each phantom/test combination is a versioned CasePack containing ROIs, metrics, thresholds, and a report template. Adding PET/CT or SPECT requires a new CasePack, not new code—the same harness accommodates different modalities with no architectural changes. Scaffold directories for `pet_ct/` and `spect_ct/` are in place.

Economic argument. A single medical physicist may oversee 10–50 scanners. The copilot reduces per-scanner QC time from hours to minutes while improving consistency, traceability, and accreditation readiness. The abundance flywheel (section 9.3) applies: each scanner model added lowers the marginal cost of the next, and the resulting cost surplus funds expansion to additional imaging modalities and clinical sites.

10 Call to Action

The infrastructure described in this paper is only as valuable as the community that stress-tests, extends, and ultimately governs it. We outline concrete entry points for four audiences.

Professors. Co-steward the LIP-Arena evaluation tracks. Validate your group's methods on the four-scenario protocol (Ideal, Assumed, Corrected, Oracle Mask) and publish the full diagnostic vector—not just the Scenario I PSNR that flatters every solver equally. Serve on the CISP steward board and shape the metrics that will define the field for the next decade. The evaluation protocol is designed to be extensible: if your modality is missing, proposing a new `OperatorGraph` template earns co-authorship on the benchmark paper and a permanent seat at the standards table.

PhD Students. Join the calibration sprints. Pick one of the 64 registered modalities, measure its mask-sensitivity spectrum on physical hardware, and contribute a row to the correction table that currently spans 16 modalities. Each new row is a quantified, reproducible contribution—a first-author publication opportunity in a subfield that did not exist two years ago. The toolchain is open:

```
pip install pwm-eval && pwm calibrate --modality <yours>
```

From installation to a publishable sensitivity curve takes one afternoon; from curve to a peer-reviewed correction-factor paper takes one semester.

Hobbyists. Participate in the weekly SolveEverything challenges. Download a measurement set, run your solver—classical, learned, or hybrid—and submit a `RunBundle` for automated scoring against the blinded ground truth. The entire infrastructure is released under the MIT license: zero institutional affiliation required, zero cost, zero barrier to entry. The leaderboard ranks results, not credentials.

Medical physicists. PWM's mismatch library and CasePack architecture provide a ready-made computational backbone for clinical QA/QC programs aligned with ACR, AAPM TG-126, TG-177, and TG-233. Contributing a scanner-specific mismatch signature—the mapping from parameter drift to observable artifact—adds your device model to the PWM diagnostic library and benefits every site running that scanner. We invite diagnostic and nuclear medical physicists to pilot the CT QC Copilot, validate CasePack thresholds against institutional data, and contribute to the clinical validation program.

Investors. The flywheel economics described in section 9.3 create three revenue surfaces once the standard achieves critical mass: *calibration-as-a-service* (upload raw sensor data, receive a validated operator), *pay-per-reconstruction APIs* (submit a measurement, receive a reconstruction with uncertainty bounds), and *imaging SLAs* (outcome-based contracts guaranteeing a minimum PSNR/SSIM on the client's hardware). The first evaluation standard adopted by five or more laboratories locks in the market—not through intellectual-property moats, but through the network effect of a shared benchmark. The competitive question is not *whether* this standard will emerge, but *who* writes it.

11 Conclusion

We began with a paradox: computational imaging possesses solvers of extraordinary power—deep unfolding networks that converge in seconds, diffusion priors that hallucinate plausible textures, implicit representations that compress entire scenes into coordinate functions—yet real-world deployments routinely fail. Reconstructions that achieve 35+ dB on simulated benchmarks collapse by 3.38–13.98 dB under realistic operator mismatch (table 1). The gap is not a solver problem. It is an infrastructure problem.

For years, the field has been optimizing the wrong layer. It has poured resources into the *train*—faster architectures, cleverer loss functions, larger training sets—while neglecting the *rail*: the calibrated forward operators, standardized evaluation protocols, and diagnostic tools that determine whether a solver's theoretical performance survives contact with reality.

This paper introduced PWM, the Physics World Model, as the rail for computational imaging. The contributions are concrete and measurable:

- **64 modalities** formalized as composable `OperatorGraph` templates, each encoding the physics of measurement rather than approximating it.
- **89 OperatorGraph templates** that decompose forward models into calibratable, swappable subgraphs—making operator correction a modular operation rather than a monolithic reimplementation.
- **The four-scenario evaluation protocol** (Ideal / Assumed / Corrected / Oracle Mask), which for the first time separates solver error from operator error and quantifies the recovery ratio of any pipeline.

- **LIP-Arena and the Triad Law**, providing real-time leaderboard ranking and a diagnostic framework that decomposes reconstruction failure into recoverability, carrier budget, and operator mismatch components.

The evidence is unambiguous. Across the 26-modality benchmark, operator correction alone improved reconstruction quality by +0.54 to +48.25 dB across 9 correction configurations spanning 7 distinct modalities (16 registered)—gains that no solver upgrade can replicate. The CASSI deep-dive was particularly revealing: the mismatch penalty (13.98 dB) exceeded the solver upgrade gain (10.47 dB) by 1.34×. The bottleneck was never the algorithm. It was the physics encoding.

These results carry a temporal urgency. The foundry window described in section 9 is open now but will not remain so. Within 18 months, the evaluation norms of computational imaging will solidify—either around principled, physics-aware infrastructure, or around the fragmented, simulation-only benchmarks that have defined the field to date. The QWERTY effect guarantees that whichever protocol is adopted first by a critical mass of laboratories will become permanent.

A problem is solved when the bottleneck shifts from genius to compute. PWM provides the infrastructure for that shift.

Acknowledgments

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

References

AAPM Task Group 126. Pet/ct acceptance testing and quality assurance. Technical report, American Association of Physicists in Medicine, 2019. AAPM Report No. 126.

AAPM Task Group 177. Acceptance testing and quality control of spect/ct systems. Technical report, American Association of Physicists in Medicine, 2019. AAPM Report No. 177.

AAPM Task Group 233. Performance evaluation of computed tomography systems. Technical report, American Association of Physicists in Medicine, 2019. AAPM Report No. 233.

American College of Radiology. ACR CT Accreditation Program: Testing Instructions. https://www.acraccreditation.org/modalities/computed-tomography, 2017. Accessed: 2026-02-19.

Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9911 of *Lecture Notes in Computer Science*, pages 19–34, 2016. doi: 10.1007/978-3-319-46478-7_2.

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.

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.

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.

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 (Proc. SIGGRAPH Asia)*, 36(6):218:1–218:13, 2017. doi: 10.1145/3130800.3130810.

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

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.

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.

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.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

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.

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.

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.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. doi: 10.1007/978-3-030-58452-8_24.

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.

John Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3):285–289, 2005. doi: 10.1016/j.sbi.2005.05.011.

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1256–1270, 2020. doi: 10.1109/JSTSP.2020.3004555.

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

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.

Lizhi Wang, Zhiwei Xiong, Guangming Shi, Wenjun Zeng, and Feng Wu. Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10):2104–2111, 2017. doi: 10.1109/TPAMI.2016.2621050.

Alexander D. Wissner-Gross. SolveEverything.org: A unified framework for computational problem solving. https://solveeverything.org, 2026. Accessed: 2026.

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

Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K. Nayar. Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum. *IEEE Transactions on Image Processing*, 19(9):2241–2253, 2010. doi: 10.1109/TIP.2010.2046811.

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.

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.