LEARNING HIGH-ORDER EQUILIBRIUM DYNAMICS WITH SINKHORN-CONSTRAINED HETEROGENEOUS PRIORS

Anonymous authors

ABSTRACT

Deep equilibrium models (DEMs) provide an effective framework for solving inverse problems by modeling the reconstruction process as finding the fixed point of a learned operator. However, they are fundamentally limited by the dilemma between expressivity and stability, *i.e.*, incorporating more expressive or heterogeneous priors violating the contractivity required for convergence. The implicit equilibrium solver could suffer from unstable solving dynamics or degenerated solutions by simply integrating multiple priors that represent complementary image statistics, even well captured by varying neural architectures. To address this challenge, we propose a manifold-constrained hyper-connected deep equilibrium model (MHC-DEM) that achieves multi-prior fusion with a structured transport process. Specifically, we construct a learnable interaction matrix via Sinkhorn normalization that lies on the Birkhoff polytope to project a query distribution into convex fusion weights, and consequently, enforce doubly-stochastic constraints on prior interactions to simultaneously ensure Lipschitz constraints with bounded Lipschitz constants and prevent collapse to a single dominant prior. Furthermore, we design a high-order equilibrium solver based on the stable multi-prior fusion mechanism. We introduce an inertial second-order update in an augmented state space to accelerate convergence, and embed a short inner optimization trajectory within each equilibrium step to perform multi-stage spatial refinement with non-shared parameters. We demonstrate in theory that the proposed multi-prior fusion preserves the contractivity property of neural network based operator represented by MHC-DEM. Extensive experiments show that MHC-DEM achieves consistent gains over existing methods across diverse inverse problems, including image restoration and interferometric imaging, with stable convergence guarantees.

1 INTRODUCTION

Imaging inverse problems such as deblurring Nah et al. (2017), inpainting Pathak et al. (2016), and interferometric reconstruction Akiyama et al. (2019) aim to recover a high-fidelity image from incomplete or degraded observations y. These problems are typically formulated as Maximum A Posteriori (MAP) estimation:

$$\hat{z} = \arg\min_{z} \mathcal{E}(z; y) + \lambda R(z), \tag{1}$$

where $\mathcal{E}$ enforces data fidelity and R encodes prior knowledge. Due to the ill-posed nature of inverse problems, the quality of reconstruction critically depends on the expressiveness of the prior.

Recent approaches such as Plug-and-Play (PnP) Venkatakrishnan et al. (2013), Regularization by Denoising (RED) Romano et al. (2017), and Deep Equilibrium Models (DEMs) Gilton et al. (2021) replace handcrafted regularizers with learned neural priors and yield substantial performance gains. Various neural network architectures have been established to capture complementary image statistics. Convolutional neural networks are particularly effective at modeling local textures and fine details Zhang et al. (2017a; 2021), while Transformer-based models better capture long-range dependencies and global structures Liang et al. (2021); Zamir et al. (2022). The complementary advantages of varying architectures suggest that combining multiple priors could potentially enhance reconstruction quality. For better understanding, we perform an empirical analysis in Fig. 1 and observe that different architectures exhibit distinct spatial preferences and produce nearly orthogonal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

residual gradient fields. Each prior is shown to provide unique descent directions in the reconstruction space, and relying on a single monolithic prior inevitably limits achieving high-quality solutionv (see Appendix A for details). Several recent works have explored combining multiple priors or denoisers to improve reconstruction quality Mataev et al. (2019); Zhao et al. (2026), and demonstrate that multi-prior strategies can outperform approaches using single prior.

However, integrating multiple priors into iterative solvers introduces significant challenges on stability guarantees. In RED, convergence guarantees rely on strong assumptions such as non-expansiveness or bounded Lipschitz constants of the denoiser Reehorst & Schniter (2018). Combining multiple priors could easily violate these assumptions and lead to unstable or divergent convergence behavior. This issue becomes more pronounced in DEMs, where the same operator is applied recursively until convergence. Small deviations in operator contractivity could be iteratively accumulated, making naive multi-prior integration particularly fragile in guaranteeing stability.

Existing DEM-based approaches usually circumvent this issue by resorting to a single prior operator Gilton et al. (2021); Zou et al. (2023). This design limits the expressive power of the equilibrium solution and prevents the model from leveraging complementary information across architectures. Recent attempts to improve DEM expressivity focus on modifying network structure or training strategies Zheng et al. (2025), but do not explicitly address the challenge of stable multi-prior integration. There remains a fundamental problem of *incorporating multiple complementary priors into a deep equilibrium model with stability guarantees*.

To address this problem, we propose a **Manifold-constrained Hyper-Connected Deep Equilibrium Model (MHC-DEM)** that integrates stable multi-prior fusion with efficient equilibrium optimization. Our key idea is to formulate prior fusion as a transport process on the Birkhoff polytope, which enforces structured and balanced interactions among priors. This design ensures that the fused operator remains well-conditioned to avoid the instability and degeneration commonly observed in naive multi-prior formulations.

Building on this formulation, we further enhance the equilibrium solver with two complementary components. First, we introduce a second-order update with inertial connections to improve convergence efficiency. Second, we incorporate a short inner refinement trajectory within each equilibrium step to enable more expressive spatial reconstruction.

Our contributions are summarized as follows:

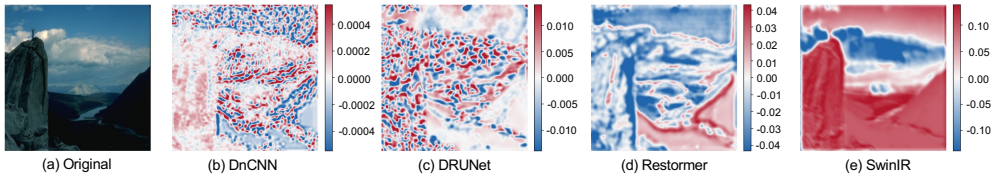
- **Manifold-Constrained Multi-Prior Fusion:** We propose a Sinkhorn-constrained fusion mechanism that enforces structured, doubly-stochastic interactions among priors. This formulation provably bounds the Lipschitz constant of the fused operator while preventing degenerated weight distributions to enable stable integration of heterogeneous priors within implicit models.

- **High-Order Equilibrium Dynamics with Decoupled Refinement:** We introduce a second-order equilibrium formulation with inertial memory, coupled with a nested inner refinement trajectory. This design decouples global convergence dynamics from local spatial reconstruction to simultaneously allow expressive multi-step updates within each equilibrium iteration and preserve the fixed-point structure.

- **Theoretic-Empirical Synergy:** We provide a rigorous mathematical stability analysis for the augmented non-linear system. Empirically, MHC-DEM achieves state-of-the-art performance across highly diverse and challenging domains, including standard natural image deblurring and highly ill-posed black hole interferometric imaging.

2 Proposed Method

In this section, we propose a novel DEM framework that explicitly decouples the inverse problem optimization into second-order temporal dynamics and nested spatial refinement, and introduces a Sinkhorn-constrained interaction mechanism to dynamically fuse heterogeneous learned priors.



(a) Original (b) DnCNN (c) DRUNet (d) Restormer (e) SwinIR

Figure 1: Spatial complementarity of heterogeneous residual fields. From left to right: a real image, the distinct residual gradient fields $r_i(\mathbf{z})$ generated by representative CNN and Transformer architectures. This intrinsic architectural complementarity necessitates our multi-prior fusion mechanism.

2.1 Deep Equilibrium Inverse Solver

Solving inverse problems typically involves balancing data fidelity and prior regularization. Classical approaches such as proximal gradient descent alternate between a gradient step on the data term and a projection onto the prior:

$$\tilde{z}_k = z_{k-1} - \gamma \nabla_z \mathcal{E}(z_{k-1}; y), \quad z_k = \text{prox}_{\lambda R}(\tilde{z}_k), \tag{2}$$

where γ is the step size. RED replace the proximal operator with a learned denoiser $\mathcal{D}_\theta$, yielding a residual correction form:

$$z_k = \tilde{z}_k + \alpha\big(\mathcal{D}_\theta(\tilde{z}_k) - \tilde{z}_k\big), \tag{3}$$

where the residual term $r(\tilde{z}_k) = \mathcal{D}_\theta(\tilde{z}_k) - \tilde{z}_k$ acts as an explicit descent direction, and α controls the update magnitude. DEMs extend this paradigm by treating the entire iterative process as an implicit network. The solution is defined as the fixed point of a nonlinear operator:

$$z^\star = \mathcal{T}_\theta(z^\star), \tag{4}$$

where $z^\star$ denotes equilibrium point of the operator. However, standard DEMs rely on first-order updates $z_{k+1} = \mathcal{T}_\theta(z_k)$, which may limit the expressivity of the iterative dynamics. To enrich the trajectory while preserving the equilibrium formulation, we introduce a second-order update with inertial memory:

$$z_{k+1} = \mathcal{F}_{\text{inner}}(z_k) + \alpha\beta_k(z_k - z_{k-1}), \tag{5}$$

where $\mathcal{F}_{\text{inner}}$ denotes a spatial refinement operator (Section 2.2), and β_k controls the momentum strength. This formulation can be rewritten as a first-order system in an augmented state space $\mathbf{u}_k = [z_k^\top, z_{k-1}^\top]^\top$, enabling standard fixed-point solvers. Notably, at equilibrium where $z_k^\star = z_{k-1}^\star$, the inertial term vanishes, ensuring consistency with the original optimization objective.

In practice, the inertial connection improves the conditioning of the iterative dynamics and accelerates convergence, while the spatial operator $\mathcal{F}_{\text{inner}}$ provides expressive reconstruction within each step. Together, they define a high-order equilibrium solver that integrates temporal acceleration and structured spatial refinement. As illustrated in Fig. 2, we reformulate the reconstruction process as a high-order equilibrium system. Compared with the standard DEM iteration, our framework augments the latent state with temporal hyper-connections and integrates nested spatial refinement within each equilibrium step. Algorithm 1 summarizes the overall pipeline of the proposed method.

2.2 Nested Spatial Refinement

While the augmented state $\mathbf{u}_k$ governs the temporal macro-trajectory, the operator $\mathcal{F}_{\text{inner}}$ is responsible for spatial reconstruction. Instead of using a single-step update, we model $\mathcal{F}_{\text{inner}}$ as a nested T-step optimization trajectory. This design serves two complementary purposes. First, since the outer DEM iterations are weight-tied, directly relying on them leads to a time-invariant and limited operator. By introducing a short inner trajectory with non-shared parameters (step sizes η_t and Sinkhorn logits $\mathbf{L}_t$), we enable stage-wise adaptive refinement within each equilibrium step, allowing the model to capture more complex reconstruction behaviors. Second, the inner loop provides a clean separation between spatial refinement and temporal acceleration. In particular, the inner updates operate without momentum and focus purely on projecting the intermediate state toward the image manifold. This prevents the outer inertial dynamics from amplifying residual artifacts when applied to under-refined states, thereby improving overall stability.

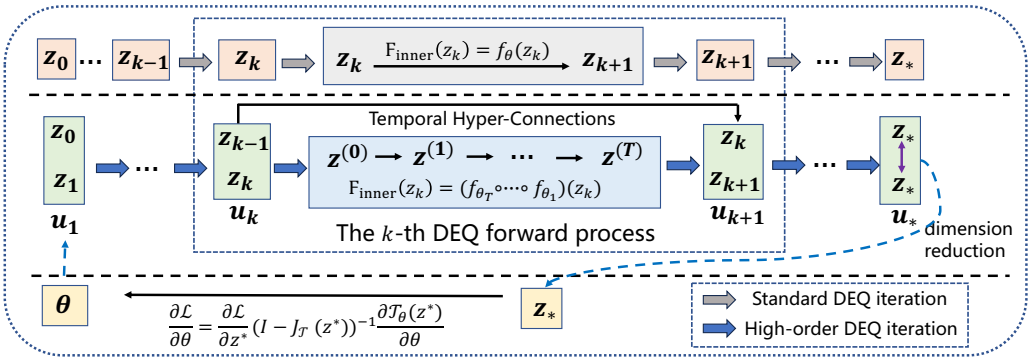


Figure 2: Overall architecture of the proposed MHC-DEM framework. The model unifies temporal hyper-connections and nested spatial refinement within an augmented-state equilibrium formulation. We explicitly illustrate both the forward and backward passes of DEM training, and provide a side-by-side comparison between standard DEM iteration and the proposed high-order dynamical iteration.

Concretely, each macro equilibrium step ($k \to k+1$) executes T alternating updates of data fidelity and prior correction. Let $z^{(0)} = z_k$. The inner updates are defined as:

$$z^{(t+1)} = z^{(t)} - \eta_t \nabla \mathcal{E}(z^{(t)}) + \alpha v_{\text{fused}}(z^{(t)}; \mathbf{L}_t), \tag{6}$$

where $\nabla \mathcal{E}(z^{(t)}) = A^\top(Az^{(t)} - y)$ for linear inverse problems. The parameters $\{\eta_t, \mathbf{L}_t\}$ are non-shared across steps, enabling progressive refinement. $v_{\text{fused}}(z^{(t)}; \mathbf{L}_t)$ is sinkhorn-constrained multi-prior interaction (Section 2.3). The final output defines the spatial operator as $\mathcal{F}_{\text{inner}}(z_k) = z^{(T)}$. The k-th MHC-DEM iteration is illustrated in Fig. 3.

2.3 Sinkhorn-Constrained Multi-Prior Interaction

A central challenge in integrating heterogeneous priors within DEMs is to control the expansivity of the fused operator. A naive summation of residuals leads to an accumulation of Lipschitz constants, resulting in an overly expansive mapping that breaks the contractivity required for equilibrium convergence ($L_{\text{fused}} \leq \sum_{i=1}^{N} L_{r_i} \leq 2N$).

A natural remedy is to enforce convex combination weights (e.g., via Softmax), which bounds the Lipschitz constant of the fused operator. However, in the context of deep equilibrium models, such unconstrained simplex parameterizations often lead to degenerate solutions in practice. Because the same operator is repeatedly applied during fixed-point iterations, small biases in the optimization can accumulate and drive the weight distribution toward sparse, near one-hot configurations, effectively collapsing the multi-prior system into a single dominant prior.

To address this issue, we introduce a Sinkhorn-constrained fusion mechanism that enforces a structured interaction among priors. Specifically, we construct a transport matrix $W \in \mathcal{B}_N$ on the Birkhoff polytope Sinkhorn (1964), i.e., the set of doubly stochastic matrices:

$$\mathcal{B}_N = \left\{ W \in \mathbb{R}^{N \times N} \mid W_{ij} \geq 0,\ W\mathbf{1} = \mathbf{1},\ W^\top \mathbf{1} = \mathbf{1} \right\}. \tag{7}$$

where $\mathbf{1} \in \mathbb{R}^N$ is the vector of all ones. The matrix W is efficiently computed via the differentiable Sinkhorn-Knopp algorithm Cuturi (2013). For notational simplicity, we omit the inner-step subscript t in this subsection. Specifically, given a learnable logit matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$ from Eq. equation 6 and a temperature parameter $\tau > 0$, we initialize a strictly positive matrix $M^{(0)} = \exp(\mathbf{L}/\tau)$. The Sinkhorn algorithm iteratively projects this matrix onto the row and column marginals of the Birkhoff polytope via alternating scaling operations:

$$\begin{aligned} M^{(2k+1)} &= \text{diag}\big(M^{(2k)}\mathbf{1}\big)^{-1} M^{(2k)}, && \text{(Row normalization)} \\ M^{(2k+2)} &= M^{(2k+1)} \text{diag}\big((M^{(2k+1)})^\top \mathbf{1}\big)^{-1}, && \text{(Column normalization)} \end{aligned} \tag{8}$$

for iteration steps $k = 0, 1, \ldots, K-1$, where K is a predefined number of iterations. The final transport matrix is obtained as $W = M^{(2K)}$.

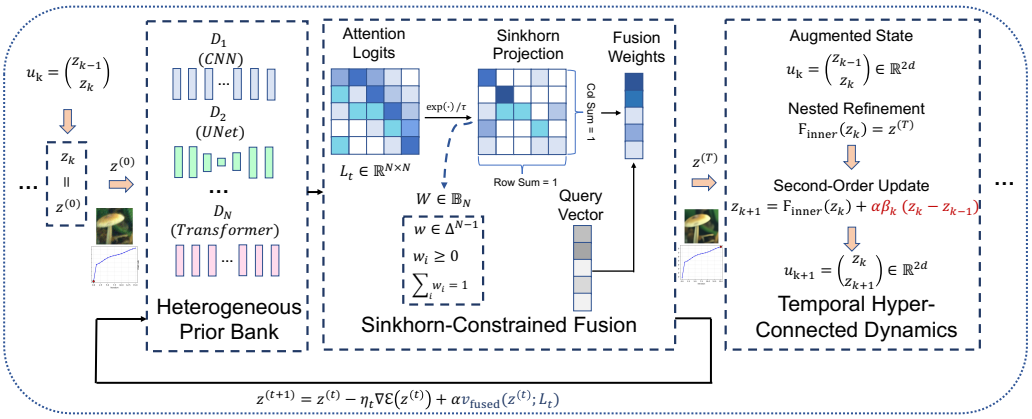


Figure 3: Detailed illustration of the k-th MHC-DEM iteration. The process integrates three key components: (1) A Heterogeneous Prior Bank that extracts diverse features; (2) A Sinkhorn-Constrained Fusion module that projects residuals onto the Birkhoff polytope; and (3) Temporal Hyper-Connections that inject temporal momentum.

To ensure strictly convex aggregation, we introduce a learnable query vector q, constrained to the $(N-1)$-dimensional probability simplex Δ^{N-1} via the Softmax function: $q = \mathrm{Softmax}(v) \in \Delta^{N-1}$, where $v \in \mathbb{R}^N$ is a learnable, unconstrained parameter vector, and the probability simplex is formally defined as $\Delta^{N-1} = \{x \in \mathbb{R}^N \mid x_i \geq 0, \sum_{i=1}^N x_i = 1\}$.

The final fusion weights are derived by projecting the query vector through the transport matrix: $w = Wq$. Since W is column-stochastic ($W^\top \mathbf{1} = \mathbf{1}$) and q lies on the simplex ($\mathbf{1}^\top q = 1$), it holds that $\mathbf{1}^\top w = \mathbf{1}^\top W q = \mathbf{1}^\top q = 1$. Combined with the non-negativity of both W and q, this mathematically guarantees that the resultant weight vector w strictly resides in Δ^{N-1}. The fused residual field is therefore computed as:

$$v_{\text{fused}}(z) = R(z)w = \sum_{i=1}^{N} w_i r_i(z). \tag{9}$$

Geometric Interpretation and Stability Implications. Our Sinkhorn-constrained formulation provides two distinct stability properties whose origins should be carefully separated. We use *simplex constraint* to refer to any parameterization that ensures $w \in \Delta^{N-1}$, and reserve *Sinkhorn / Birkhoff constraint* for the additional requirement that W be doubly stochastic.

Bounded expansivity from the simplex constraint alone. Because $w \in \Delta^{N-1}$, the fused field v_{fused} is a strict convex combination of residual operators. Under Assumption 1, $L_{\text{fused}} \leq \sum_{i=1}^N w_i \|r_i\|_{\text{Lip}} \leq 2\sum_{i=1}^N w_i = 2$, independent of N. We emphasize that this bound is not unique to Sinkhorn: any simplex constraint (e.g., a plain softmax) realizes the same bound. The relevant point of contrast is therefore the naive aggregation $\sum_i r_i$, whose Lipschitz constant grows linearly in N, rather than other simplex parameterizations.

Anti-collapse from the doubly-stochastic constraint. What separates Birkhoff from a generic simplex constraint is that $W\mathbf{1} = \mathbf{1}$ and $W^\top \mathbf{1} = \mathbf{1}$ enforce balanced probability transport across both rows and columns. In recursive equilibrium iteration, unconstrained simplex maps such as softmax tend to amplify minor optimization biases until the routing concentrates onto a single dominant denoiser, effectively reducing the multi-prior system to a single-prior one (see Table 8, softmax places 97.2% of the mass on one prior). The doubly-stochastic constraint resists this concentration and keeps heterogeneous priors continuously active throughout the dynamics.

Together, these two properties decouple operator expansivity from prior-diversity preservation: the first defines a stable working regime for inertial acceleration (Theorem 1); the second ensures that the diversity exploited inside that regime does not silently dissolve as iterations progress. As before, the derived Lipschitz bound should be read as a sufficient stability condition rather than a tight characterization of practical convergence.

3 THEORETICAL ANALYSIS

In this section, we provide a convergence analysis of the proposed MHC-DEM framework. Unlike standard DEMs or RED methods that rely on memoryless fixed-point iterations ($z_{k+1} = f(z_k)$), our introduction of **Temporal Hyper-Connections** creates a **second-order dynamical system** dependent on historical states.

The primary theoretical challenge is to maintain stability when incorporating both heterogeneous priors and inertial dynamics into an implicit equilibrium system. In general, combining multiple nonlinear priors increases operator expansivity, while momentum further amplifies sensitivity to perturbations across recursive iterations. Without explicit control, these effects can jointly destroy the contractivity required for equilibrium convergence.

Our key theoretical insight is two-fold. First, the simplex constraint $w \in \Delta^{N-1}$ alone preserves a stability budget whose magnitude is independent of the number of priors, creating a controllable margin that safely accommodates inertial acceleration. Second, the Birkhoff (doubly-stochastic) refinement ensures that this budget is actually exploited by all priors rather than monopolized by a single one across recursive iterations.

3.1 DYNAMICAL SYSTEM FORMULATION

Let $\mathcal{H} = \mathbb{R}^d$ denote the image space. The fixed-point update rule at iteration k, incorporating Data Consistency, Sinkhorn-constrained Multi-Prior Fusion, and Inertial Memory, is:

$$z_{k+1} = \Phi(z_k, z_{k-1}) := \underbrace{z_k - \eta\nabla\mathcal{E}(z_k)}_{\text{Data Consistency}} + \alpha \underbrace{v_{\text{fused}}(z_k)}_{\text{Sinkhorn Prior}} + \underbrace{\alpha\beta_k(z_k - z_{k-1})}_{\text{Inertial Hyper-Connection}}, \tag{10}$$

where η is the step size, α is the regularization strength, β_k is the momentum coefficient, and $v_{\text{fused}}(z)$ is the fused residual field from the prior bank.

To analyze this history-dependent sequence, we employ **State-Space Augmentation**. We define the augmented state $\mathbf{u}_k = [z_k^\top, z_{k-1}^\top]^\top \in \mathbb{R}^{2d}$ and reformulate the system as a first-order mapping in the augmented space: $\mathbf{u}_{k+1} = \mathbf{T}(\mathbf{u}_k) = [\Phi(z_k, z_{k-1})^\top, z_k^\top]^\top$. Asymptotic stability requires the spectral radius of the augmented Jacobian $\rho(J_{\mathbf{T}})$ to be strictly less than 1. For rigorous analysis, we make standard assumptions: Pre-trained denoisers are 1-Lipschitz.

3.2 MAIN CONVERGENCE THEOREMS

By bounding the Lipschitz constant of the Sinkhorn projection (Lemma 1 in Appendix D.2) and analyzing the block companion matrix of the augmented system (Lemma 2 in Appendix D.3), we establish the sufficient conditions for global stability.

Theorem 1 (Asymptotic Stability of MHC-DEM)**.** *Under the stated assumptions, the augmented dynamical system* $\mathbf{u}_{k+1} = \mathbf{T}(\mathbf{u}_k)$ *converges asymptotically to a unique fixed point if the following* ***Stability Condition*** *holds:*

$$\underbrace{\|I - \eta\nabla^2\mathcal{E}(z^\star)\|_2}_{\textit{Data Fidelity}} + \underbrace{2\alpha}_{\textit{Spatial Prior}} + \underbrace{2\alpha\beta_k}_{\textit{Temporal Inertia}} < 1. \tag{11}$$

Proof Sketch. The Jacobian of $\mathbf{T}$ forms a 2×2 block matrix. The spatial sensitivity is bounded by the contraction of the data term and the Sinkhorn-constrained prior (bounded to 2). The temporal sensitivity is purely defined by the inertia $\alpha\beta_k$. Applying spectral bounds on block matrices yields Eq. equation 11. The full proof is provided in Appendix D.4.

Theorem 1 highlights two complementary roles of our fusion in equilibrium stabilization. First, the simplex constraint $w \in \Delta^{N-1}$ alone preserves a bounded stability region independent of prior cardinality—this is what enables the inertial budget $2\alpha\beta_k$ to coexist with the spatial budget 2α in Eq. equation 11. Without simplex normalization, the multi-prior Lipschitz constant grows with the number of priors and forces momentum to be chosen extremely conservatively to avoid divergence. Second, in recursive equilibrium iteration, unconstrained simplex parameterizations such as softmax tend to collapse onto a single prior (Table 8), effectively re-introducing the single-prior regime that

motivated our design. The Birkhoff constraint enforced by Sinkhorn iterations resists this collapse, so that the budget granted by Theorem 1 is exploited by a genuinely heterogeneous operator rather than silently spent on one prior. Although the derived bound is conservative, it provides a useful theoretical blueprint for balancing reconstruction expressivity and equilibrium stability in practice.

When extending the single-step update to a nested unrolled block (depth T), the stability condition must account for the composite operator expansivity.

Theorem 2 (Conditional Stability of Nested DEM and Depth Bound)**.** *Let L_t be the Lipschitz constant of the t-th inner update. The nested dynamical system converges to a unique fixed point if the inner depth T satisfies the* ***Bounded Expansion Condition****:*

$$\underbrace{\prod_{t=0}^{T-1} L_t}_{\textit{Composite Expansion}} + \underbrace{2\alpha\beta_k}_{\textit{Inertial Cost}} < 1. \tag{12}$$

Proof provided in Appendix D.5.

Eq. equation 12 formalizes a critical design trade-off. Increasing T improves *Expressivity*, allowing the model to capture highly non-linear projections onto the image manifold. However, since individual steps may be slightly expansive ($L_t \geq 1$), the product $\prod L_t$ accumulates, degrading *Stability*. Thus, T cannot be arbitrarily large.

4 Experiments

4.1 Experimental Setup

MHC-DEM is implemented within the deep equilibrium framework with a Jacobian-free solver. Anderson acceleration is used for implicit differentiation, with a history size of 5 and a maximum of 30 iterations. Both forward and backward solvers use a tolerance of 10^{-3}. The model is trained for 100 epochs using Adam with a learning rate of 5×10^{-5} and weight decay of 10^{-6}. All experiments are conducted on an NVIDIA H200 GPU. Additional implementation are provided in Appendix C.

4.2 Optimized Trajectory

To investigate the effect of the proposed temporal hyper-connections, we visualize the optimization trajectories on the image deblurring task using PCA and t-SNE projections of the latent representations across iterations. As shown in Fig. 4, compared with standard memoryless equilibrium dynamics, MHC-DEM exhibits more direct and faster trajectories toward the target solution. Without inertial dynamics, the iterative process tends to evolve slowly along locally stagnant paths. In contrast, the proposed second-order dynamics introduce characteristic heavy-ball behavior, enabling the solver to traverse flat regions more efficiently and reach the equilibrium state with fewer iterations. These observations empirically support our theoretical analysis that the proposed Sinkhorn-constrained formulation provides a stable optimization region for inertial acceleration.

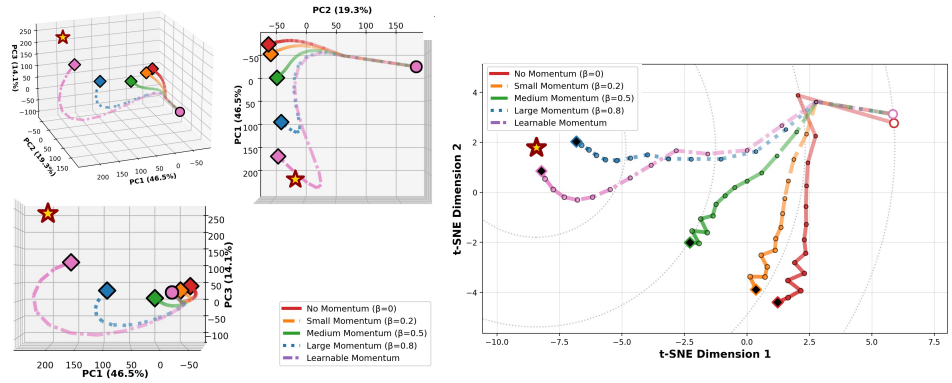
4.3 Black Hole Interferometric Imaging

Black hole imaging from Event Horizon Telescope (EHT) observations is a severely ill-posed and highly non-linear inverse problem. EHT reconstruction relies on closure phases and log-closure amplitudes to suppress atmospheric corruption, resulting in a challenging non-convex optimization landscape. Appendix E.1 provides detailed additional analyses for both synthetic simulations and real M87* 2017 observations, together with the complete EHT reconstruction formulation within the proposed framework.

Table 1: Quantitative results on synthetic M87 simulation data.

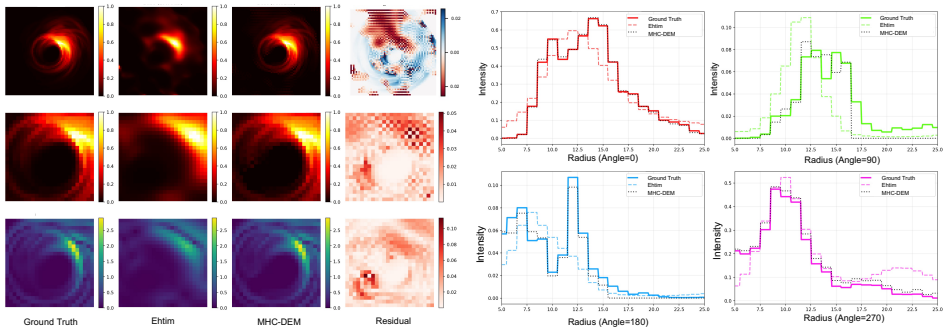
Method	MSE ($\times 10^{-4}$)	Corr. (ρ)
Ehtim	16.64	0.944
DPS	7.45	0.913
PnP-DM	3.52	0.936
MHC-DEM	**1.32**	**0.996**

We compare MHC-DEM with the Regularized Maximum Likelihood solver implemented in the official Ehtim library Akiyama et al. (2019); Chael et al. (2018), as well as diffusion-based recon-



(a) Trajectories projected onto PCA's three principal directions. (b) Trajectories in t-SNE embeddings.

Figure 4: Iteration trajectories under different momentum coefficients visualized via PCA and t-SNE. For PCA, three projections are presented: (i) the trajectory in the space of the first three principal components, (ii) the projection onto the first and second principal components (PC1–PC2), and (iii) the projection onto the first and third principal components (PC1–PC3). t-SNE provides a complementary nonlinear embedding to further illustrate the trajectory evolution under different momentum settings.



(a) Reconstruction Results (b) Radial Profiles

Figure 5: Visual evaluation and morphological fidelity on synthetic black hole imaging. (a) Multi-scale comparative visualization. Columns from left to right: Ground truth, Ehtim Akiyama et al. (2019), proposed MHC-DEM, and the residual error between MHC-DEM and GT. Rows from top to bottom display the global structural comparison, local zoomed-in views of the emission region, and edge-level enhancement details. (b) Radial intensity distribution profiles sampled at four distinct azimuthal angles $(0°, 90°, 180°, 270°)$, demonstrating MHC-DEM's precise alignment with the GT flux variations.

struction methods Chung et al. (2023); Wu et al. (2024). As showed in Table 1, MHC-DEM achieves the best quantitative performance, reaching 1.32×10^{-4} MSE and a correlation coefficient (Corr.) of 0.996, substantially outperforming existing baselines. Qualitative comparisons in Fig. 5 further demonstrate the advantages of the proposed framework. MHC-DEM accurately reconstructs the bright Doppler emission region and the central black hole shadow while producing significantly fewer structural artifacts. The zoomed-in and edge-level visualizations reveal substantially sharper photon-ring boundaries than those reconstructed by Ehtim. In addition, the radial intensity profiles in Fig. 5(b) closely follow the ground-truth distribution across multiple azimuthal directions, indicating more accurate recovery of ring thickness and local flux variations.

4.4 Image Deblurring

To evaluate the generalization and robustness of the MHC-DEM on standard linear inverse problems, we conduct image deblurring experiments on the *CBSD68* dataset. Following standard protocols, images are degraded using the eight classical *Levin09* blur kernels with additive Gaussian noise ($\sigma = 0.01$). Due to space limitations, Table 2 reports results for four representative kernels

Table 2: Comparison of PSNR (dB) and SSIM for image deblurring on *CBSD68*. We report results on representative kernels and the Average across all 8 *Levin09* kernels. The best and second-best results are **bolded** and underlined.

Models	#0	#1	#4	#5	Average
DE-Prox	32.35 (0.9166)	32.27 (0.9156)	33.28 (0.9308)	33.40 (0.9319)	32.64 (0.9196)
ELDER	32.11 (0.9166)	31.93 (0.9064)	33.17 (0.9293)	33.25 (0.9296)	32.43 (0.9152)
DiffIR	29.54 (0.8521)	29.43 (0.8497)	30.73 (0.8781)	30.38 (0.8705)	29.85 (0.8589)
DAPS	30.24 (0.8674)	30.59 (0.8751)	31.77 (0.9009)	30.95 (0.8830)	30.88 (0.8815)
ProxPnP-DRS	32.27 (0.9109)	32.00 (0.9078)	32.73 (0.9230)	32.90 (0.9227)	32.10 (0.9108)
DenseDiff	32.40 (0.9170)	32.42 (0.9168)	33.37 (0.9362)	33.49 (0.9381)	32.69 (0.9214)
MHC-DEM	**32.73** (**0.9173**)	**32.91** (**0.9186**)	**33.81** (**0.9397**)	**33.67** (**0.9385**)	**32.91** (**0.9224**)

Table 3: Deblurring performance under different architectures and hyper-connection settings.

Hyper-Connection	DnCNN	DRUNet	Restormer	SwinIR(w/o SN)	PSNR (dB)	Time (ms)
None	✓				30.89	730
Fixed ($\beta = 0.5$)	✓				31.09	558
Learnable	✓				31.35	526
Learnable		✓			31.66	517
Learnable	✓	✓			32.65	548
Learnable	✓	✓	✓		**32.91**	591
Learnable	✓	✓	✓	✓	32.32	672

together with the overall average, while the complete evaluation over all kernels is provided in Appendix E.2.As shown in Table 2, MHC-DEM achieves the best average performance of 32.91 dB PSNR across all eight kernels, establishing a new state of the art. Compared with recent generative restoration approaches (e.g., DiffIR and DAPS), our method produces more faithful reconstructions by explicitly enforcing data consistency within the equilibrium optimization process, avoiding the structural hallucinations introduced by stochastic sampling. Compared with recent equilibrium-based methods (e.g., ELDER, DenseDiff, and DE-Prox), the proposed framework benefits from stable heterogeneous prior interaction and high-order equilibrium dynamics, enabling more accurate reconstruction under challenging blur degradations.

4.5 Ablation Study

We conduct ablation studies on the image deblurring task (Levin kernel #1, $\sigma = 0.01$). Table 3 reports the PSNR and average inference time per image. Compared with the vanilla equilibrium solver, temporal hyper-connections improve both reconstruction quality and inference efficiency, demonstrating the benefit of the proposed inertial dynamics. Heterogeneous prior fusion further improves reconstruction performance, with the combination of DnCNN, DRUNet, and Restormer achieving the best performance among the evaluated configurations with only marginal additional cost. Increasing the number of priors does not necessarily yield further improvements.

Adding SwinIR without spectral normalization degrades performance, suggesting that the stability of the equilibrium dynamics depends not only on the fusion mechanism but also on the operator properties of individual priors. Further analyses of convergence, spectral properties, fusion behavior, and inference efficiency are provided in Appendix E.4.

5 Conclusion

We presented the manifold-constrained hyper-connected deep equilibrium model. By integrating Sinkhorn-constrained heterogeneous prior fusion with temporal hyper-connections, the proposed framework enhances both representation capacity and convergence stability of DEM-based methods. Theoretical analysis guarantees operator stability, while experiments on image deblurring and black hole imaging demonstrate consistent improvements over state-of-the-art approaches.

AI USE STATEMENT

We used generative AI tools to aid in polishing and improving the clarity, grammar, and readability of the manuscript. These tools were used for language refinement and writing assistance, rather than for generating experimental results, conducting experiments, or establishing the core scientific contributions of this work.

We have reviewed all AI-assisted content and take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

Gregery T Buzzard, Stanley H Chan, Suhas Sreehari, and Charles A Bouman. Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium. *SIAM Journal on Imaging Sciences*, 11(3):2001–2020, 2018.

Andrew A Chael, Michael D Johnson, Katherine L Bouman, Lindy L Blackburn, Kazunori Akiyama, and Ramesh Narayan. Interferometric imaging directly with closure phases and closure amplitudes. *The Astrophysical Journal*, 857(1):23, 2018.

Qi Chen, Yifei Wang, Zhengyang Geng, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Equilibrium image denoising with implicit differentiation. *IEEE Transactions on Image Processing*, 32: 1868–1881, 2023.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. The International Conference on Learning Representations, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. Jfb: Jacobian-free backpropagation for implicit networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6648–6656, 2022.

Davis Gilton, Gregory Ongie, and Rebecca Willett. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:1123–1133, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Mike Heddes, Adel Javanmard, Kyriakos Axiotis, Gang Fu, Mohammadhossein Bateni, and Vahab Mirrokni. Deepcrossattention: Supercharging transformer residual connections. In *International Conference on Machine Learning*, pp. 22881–22903. PMLR, 2025.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 4700–4708, 2017.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *the Tenth International Conference on Learning Representations*, 2022.

Samuel Hurault, Antonin Chambolle, Arthur Leclaire, and Nicolas Papadakis. Convergent plug-and-play with proximal denoiser and unconstrained regularization parameter. *Journal of Mathematical Imaging and Vision*, 66(4):616–638, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jingyun Liang, Jlezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021.

Brian Mak and Jeffrey Flanigan. Residual matrix transformers: Scaling the size of the residual stream. In *Forty-second International Conference on Machine Learning*, 2025.

Gary Mataev, Peyman Milanfar, and Michael Elad. DeepRED: Deep image prior powered by red. In *Proceedings of the IEEE/CVF International Conference on Computer Vision workshops*, pp. 0–0, 2019.

Alexander Mehta, Ruangrawee Kitichotkul, Vivek K Goyal, and Julián Tachella. Equivariant deep equilibrium models for imaging inverse problems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11437–11441. IEEE, 2026.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891, 2017.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

Aniket Pramanik, M Bridget Zimmerman, and Mathews Jacob. Memory-efficient model-based deep learning with convergence and robustness guarantees. *IEEE Transactions on Computational Imaging*, 9:260–275, 2023.

Edward T Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Transactions on Computational Imaging*, 5(1):52–67, 2018.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pp. 5546–5557. PMLR, 2019.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948. IEEE, 2013.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine L Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37:118389–118427, 2024.

Da Xiao, Qingye Meng, Shengping Li, and Xingyuan Yuan. MUDDFormer: Breaking residual bottlenecks in transformers via multiway dynamic dense connections. In *International Conference on Machine Learning*, pp. 68440–68458. PMLR, 2025.

Duoduo Xue, Ziyang Zheng, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. On the convergence of non-convex phase retrieval with denoising priors. *IEEE Transactions on Signal Processing*, 70:4424–4439, 2022.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20895–20905, 2025.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26 (7):3142–3155, 2017a.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Rrecognition*, pp. 3929–3938, 2017b.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

Yifan Zhao, Ziyang Zheng, Duoduo Xue, Yong Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Unfolding convolutional sparse coding with low-rank-guided hybrid priors for image denoising. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(3):1–29, 2026.

Ziyang Zheng, Wenrui Dai, Jingwei Liang, Xinyu Peng, Duoduo Xue, and Hongkai Xiong. Deep equilibrium model with weighted densely-connected iteration differences for convergent image restoration. *IEEE Journal of Selected Topics in Signal Processing*, 2025.

Guoru Zhou, Yixin Zuo, Zhe Zhang, Bingchen Zhang, and Yirong Wu. CR-DEQ-SAR: A deep equilibrium sparse SAR imaging method for compound regularization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:4680–4695, 2025.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. In *The Thirteenth International Conference on Learning Representations*, 2024.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1219–1229, 2023.

Zi Zou, Jialin Liu, Brendt Wohlberg, and Ulugbek S Kamilov. Deep equilibrium learning of explicit regularization functionals for imaging inverse problems. *IEEE Open Journal of Signal Processing*, 4:390–398, 2023.

A Empirical Analysis of Heterogeneous Prior Complementarity

A core motivation of the proposed MHC-DEM framework is the integration of multiple heterogeneous priors. In standard RED frameworks, pre-trained denoisers are treated as monolithic proximal operators. In this section, we empirically demonstrate that denoisers with different architectures inherently learn distinct structural sub-manifolds, thereby providing orthogonal and complementary descent directions during the inverse optimization process.

A.1 Formulation of Implicit Gradient Fields

In residual-based evolutionary dynamics, the output of the i-th denoiser $D_i(\cdot)$ given a degraded intermediate state $\mathbf{z}$ can be reformulated as a state update driven by a residual field:

$$r_i(\mathbf{z}) = D_i(\mathbf{z}) - \mathbf{z}. \tag{13}$$

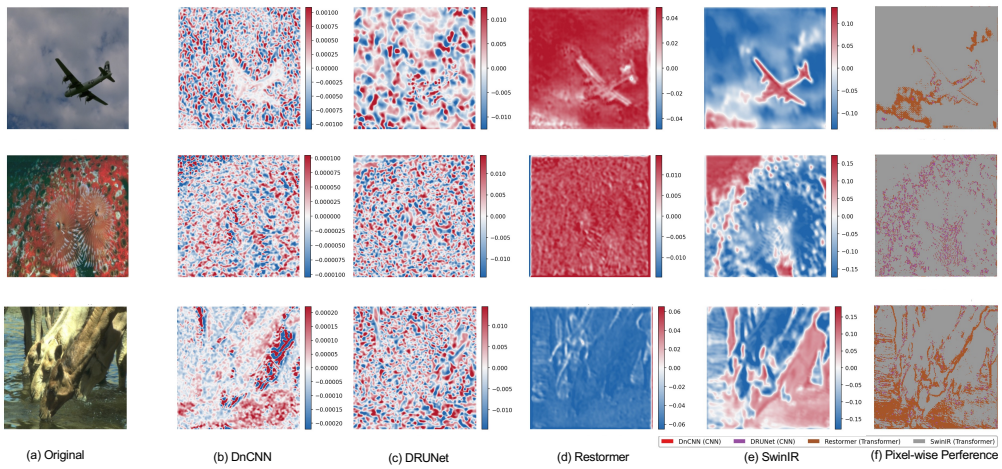


Figure 6: Visualizing the spatial complementarity of heterogeneous priors. For a given degraded image (left), different network architectures generate distinct residual fields $r_i(\mathbf{z})$ (middle columns), highlighting completely different structural features. The "Pixel-wise Denoiser Performance" map (right) illustrates the spatial regions where each denoiser achieves the lowest reconstruction error. The legend in the bottom-right corner indicates the correspondence between colors and different denoisers in the performance map. CNNs Zhang et al. (2017b; 2021) often excel at high-frequency local textures, while Transformers Zamir et al. (2022); Liang et al. (2021) reconstruct smooth backgrounds and sharp edges more effectively.

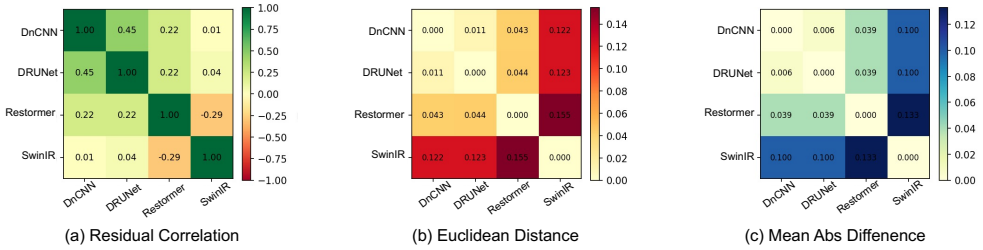


Figure 7: Statistical and geometric diversity among residual fields. Three similarity matrices computed across different denoiser architectures on natural images: (a) Pearson Correlation (ρ), (b) Euclidean Distance, and (c) Mean Absolute Difference (MAD). The matrices quantitatively confirm that while intra-class architectures (e.g., CNN vs. CNN) exhibit moderate correlation, inter-class architectures (e.g., CNN vs. Transformer) are nearly orthogonal ($\rho \to 0$) and separated by significant geometric distances.

From an optimization perspective, $r_i(\mathbf{z})$ acts as an explicit descent direction (the negative gradient) of an implicit regularization functional governed by the specific inductive bias of the i-th network. Consequently, investigating the spatial behavior and statistical diversity among different residual fields $\{r_i(\mathbf{z})\}_{i=1}^{N}$ provides a principled foundation to justify the necessity of multi-prior fusion. To systematically characterize these fields, we evaluate representative network architectures on natural images from the *CBSD68* dataset. All four denoisers were retrained in our experiments. Since applying spectral normalization to SwinIR Liang et al. (2021) led to noticeably degraded denoising performance, spectral normalization was only imposed on the other three models.

A.2 Spatial Preference and Region-Wise Dominance

We first analyze the spatial behaviors of different priors on complex natural images. As shown in Fig. 6, the residual fields $r_i(\mathbf{z})$ produced by different architectures exhibit clear spatial discrepancies. To quantify these differences, we construct a *pixel-wise denoiser preference map*, where each pixel is assigned to the denoiser achieving the lowest local reconstruction error.

The resulting maps reveal highly heterogeneous spatial distributions, indicating that the optimal denoiser varies across semantic regions. CNN-based priors tend to dominate highly textured and local-detail regions, while Transformer-based priors are more effective at recovering smooth backgrounds and long-range structural boundaries. These observations suggest that different architectures capture complementary reconstruction behaviors, and relying on a single prior inevitably restricts reconstruction quality in diverse scenes.

A.3 Statistical Diversity of Residual Fields

Beyond spatial visualization, we further analyze the statistical relationships between different residual fields. Fig. 7 reports the Pearson Correlation Coefficient, Euclidean Distance, and Mean Absolute Difference between priors. The correlation analysis shows that CNN-based and Transformer-based priors exhibit very low correlation, indicating nearly orthogonal descent directions. In contrast, priors within the same architectural family exhibit higher correlation. The Euclidean Distance and MAD statistics further confirm that heterogeneous priors differ not only in direction but also in magnitude and spatial response patterns.

These results demonstrate that heterogeneous denoisers capture complementary image manifolds rather than redundant representations. Consequently, properly combining these diverse residual fields can significantly improve the expressivity of equilibrium-based reconstruction.

A.4 Implications for Deep Equilibrium Solvers

The above observations provide important insights into equilibrium-based inverse solvers.

1. **Learned denoisers behave as biased descent operators** Rather than acting as exact proximal mappings onto a universal image manifold, different denoisers induce distinct reconstruction trajectories shaped by their architectural biases.
2. **Single-prior DEMs limit equilibrium expressivity** Using only one denoiser constrains the equilibrium dynamics to a restricted optimization path, which is suboptimal for images containing mixed textures and structures.

These findings directly motivate the design of MHC-DEM. Since heterogeneous priors provide complementary yet statistically diverse gradient directions, naively combining them may destabilize the equilibrium dynamics. Our proposed Sinkhorn-constrained fusion addresses this issue by enabling structured and bounded multi-prior interaction, allowing the equilibrium solver to benefit from complementary priors while preserving stable optimization behavior.

B Related Work

B.1 Deep Equilibrium Models in Inverse Problem

DEMs Bai et al. (2019) formulate the restoration trajectory as the search for the fixed point $\mathbf{x}^\star = \mathcal{T}_\theta(\mathbf{x}^\star)$ of an implicit, weight-tied layer. By leveraging the implicit function theorem to compute gradients analytically at the equilibrium state Bai et al. (2019); Fung et al. (2022), DEMs entirely decouple the forward iteration depth from the backward memory footprint. This paradigm enables theoretically infinite-depth feature processing with low memory cost Pramanik et al. (2023); Mehta et al. (2026). In the specific context of solving computational inverse problems, DEMs generally construct their implicit operator $\mathcal{T}_\theta$ by intertwining a data-consistency step with a learned regularizer. For instance, DE-Prox Gilton et al. (2021) directly wraps a standard Proximal Gradient Descent step and a pre-trained denoiser within the equilibrium loop, treating the entire process as a differentiable fixed-point solver. To enhance flexibility, ELDER Zou et al. (2023) moves beyond generic denoisers by explicitly parameterizing a learnable regularization functional whose gradient drives the equilibrium updates. More recently, multi-stage or nested DEMs have been proposed, where short unrolled blocks are iterated to equilibrium, aiming to capture more complex degradation patterns Chen et al. (2023); Zhou et al. (2025).

Despite these architectural advances, existing DEMs for image restoration predominantly rely on a monolithic prior network. Whether utilizing a standard U-Net or a basic ResNet backbone, they

define the implicit prior space using a single structural inductive bias. This rigid design fails to exploit the complementary modeling capacities of heterogeneous architectures, such as the high-frequency local texture recovery inherent to Convolutional Neural Networks Zhang et al. (2017b) and the global, long-range dependency modeling characteristic of Vision Transformers Liang et al. (2021); Zamir et al. (2022). Consequently, the equilibrium trajectory in standard DEMs is strictly confined to a suboptimal, homogeneous feature manifold, severely restricting their performance on highly ill-posed or complex non-linear inverse problems.

B.2 Multi-Prior Fusion in Inverse Problems

Incorporating diverse regularizers has a rich history in inverse problems, initially starting with hybrid handcrafted priors (e.g., combining Total Variation with wavelet sparsity). In the deep learning era, this paradigm evolved into explicit multi-prior optimization frameworks. A prominent example is Multi-Agent Consensus Equilibrium Buzzard et al. (2018), which generalizes Plug-and-Play Venkatakrishnan et al. (2013) by formulating the reconstruction as finding a consensus among multiple independent expert denoisers using ADMM. Similarly, within deep unrolled networks, recent methods have explored parallel feature branches or cascaded blocks of heterogeneous architectures Mataev et al. (2019); Zhao et al. (2026) to dynamically aggregate diverse spatial representations.

To artificially balance these multi-prior systems, an intuitive approach is to enforce a convex combination using dynamic gating networks, structurally similar to Mixture of Experts Jacobs et al. (1991). However, within recursive equilibrium solvers, unconstrained simplex mappings (e.g., Softmax) frequently suffer from severe routing degeneracy. Because the spatial operator is repeatedly evaluated across hundreds of steps to reach a fixed point, minor initial optimization biases cause the routing mechanism to prematurely lock onto a single dominant denoiser. This phenomenon essentially collapses the multi-prior system into a single-prior model. Our work explicitly departs from these heuristic gating mechanisms by rigorously formulating the prior interaction as an optimal transport problem on the Birkhoff polytope, which guarantees uniform and balanced prior participation without premature collapse.

B.3 Temporal Dynamics and Feature Aggregation

Beyond spatial feature extraction, facilitating gradient flow and mitigating information loss across sequential stages is critical He et al. (2016); Huang et al. (2017); Heddes et al. (2025). In finite-depth iterative networks, diverse mechanisms have been explored to reuse historical structural representations. Recurrent states are often embedded to pass memory across unrolled stages Zhu et al. (2024), while cross-layer concatenation explicitly fuses multi-scale features from previous layers to adaptively modulate current updates Xiao et al. (2025); Mak & Flanigan (2025). Translating these concepts to infinite-depth DEMs presents a unique challenge. Recent works like DenseDiff Zheng et al. (2025) have attempted to reintroduce history by aggregating successive iteration differences via learned dense connections, functioning as a quasi-inertial scheme to accelerate the implicit solver.

However, applying such heuristic temporal augmentations to implicit models introduces severe theoretical vulnerabilities. Unconstrained linear combinations of historical states lack geometric bounding, risking the violation of strict Lipschitz continuity required by the Banach Fixed-Point Theorem for forward convergence. Without precise calibration akin to formal momentum methods (e.g., Polyak's Heavy-Ball Nesterov et al. (2018)), combining memory reuse with multi-prior fusion easily leads to numerical divergence in an infinite-depth loop. This is where our framework provides a holistic solution. The simplex-projected component of our fusion bounds the spatial operator's Lipschitz constant and therefore creates a mathematical stability margin; the Sinkhorn / Birkhoff component Cuturi (2013) additionally guarantees that this margin is realized by a non-degenerate, multi-prior convex combination rather than a near one-hot one. Together, this bounded *and* multi-prior spatial formulation is what safely permits the integration of temporal hyper-connections, enabling a provably convergent, second-order dynamical system.

C Additional Implementation Details

C.1 Symbol Table for MHC-DEM

Table 4 provides a summary of the symbols used throughout this paper and their corresponding definitions.

Table 4: Symbol Table for MHC-DEM

Symbol	Meaning	Notes / Units
z	Reconstructed image state	$\mathbb{R}^d$
$z^\star$	Equilibrium point of the operator	$\mathbb{R}^d$
$D_i(z)$	i-th denoiser in the heterogeneous prior bank	$\mathbb{R}^d \to \mathbb{R}^d$
$r_i(z)$	Residual of the i-th denoiser	$r_i(z) = D_i(z) - z$
$R(z)$	Residual matrix	$R(z) = [r_1(z), \dots, r_N(z)]$
w	Fused convex weights from Sinkhorn projection	$w \in \Delta^{N-1}$
$v_{\text{fused}}(z)$	Fused residual vector	$v_{\text{fused}}(z) = R(z)w$
α	Regularization strength for spatial prior	scalar
β	Momentum / inertial coefficient	scalar
η	Step size for data fidelity gradient descent	scalar
$\mathbf{u}_k$	Augmented state	$\mathbf{u}_k = [z_k^\top, z_{k-1}^\top]^\top \in \mathbb{R}^{2d}$
$\mathcal{T}_\theta$	Nested DEM fixed-point operator	$\mathbb{R}^d \to \mathbb{R}^d$
T	Number of inner nested refinement steps	integer
$J_\mathcal{T}$	Jacobian of the operator $\mathcal{T}_\theta$	$\partial \mathcal{T}_\theta / \partial z \in \mathbb{R}^{d \times d}$
y	Observed degraded measurement	domain-specific
$\mathcal{E}(z; y)$	General data-fidelity energy	scalar
$\nabla \mathcal{E}(z)$	Gradient of the data-fidelity term	$\mathbb{R}^d$
$\nabla^2 \mathcal{E}(z)$	Hessian matrix of the data-fidelity term	$\mathbb{R}^{d \times d}$

C.2 Implicit Backpropagation

The proposed MHC-DEM framework operates as an implicit network. To make this computationally feasible and memory-efficient, we adopt distinct strategies for the training and inference phases, leveraging implicit differentiation and fixed-point solvers.

End-to-End Training via Implicit Differentiation Standard deep unrolling networks require storing intermediate activations for all K layers to compute gradients via backpropagation through time, resulting in $O(K)$ memory complexity. In contrast, our DEM formulation leverages the *implicit function theorem* to decouple the forward pass depth from the memory cost. The training is formulated as a bilevel optimization problem:

$$\min_\theta \mathcal{L}(z^\star, z_{\text{gt}}) \quad \text{s.t.} \quad z^\star = \mathcal{T}_\theta(z^\star; y). \tag{14}$$

During the **forward pass**, we solve for the fixed point $z^\star$ using a black-box solver. Since we only need the final equilibrium state $z^\star$ for gradient computation, intermediate states are discarded, achieving $O(1)$ memory complexity regarding depth.

During the **backward pass**, the gradient of the loss $\mathcal{L}$ with respect to parameters θ is computed via the fixed-point condition:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial z^\star}(I - J_\mathcal{T}(z^\star))^{-1} \frac{\partial \mathcal{T}_\theta(z^\star)}{\partial \theta}, \tag{15}$$

where $J_\mathcal{T}$ is the effective Jacobian of the spatial operator. Crucially, at equilibrium ($z_k = z_{k-1}$), the dependency on the history simplifies. While the forward pass evolves in the augmented space $\mathbb{R}^{2d}$, the Jacobian inversion in Eq. equation 15 can be efficiently reduced to the effective spatial dimension $\mathbb{R}^d$ due to the vanishing inertial gradients. To invert the Jacobian term $(I - J_\mathcal{T})^{-1}$ without computing the full $O(d^3)$ Hessian inverse, we solve the vector-Jacobian product (VJP) system $v^\top(I - J_\mathcal{T}) = \partial \mathcal{L} / \partial z^\star$ using a Jacobian-free fixed-point iteration scheme. Specifically, Anderson acceleration

Algorithm 1 MHC-DEM Algorithm

Require: Observed y, data-fidelity gradient function $\nabla\mathcal{E}(\cdot; y)$, Denoiser bank $\mathcal{D} = \{D_i\}_{i=1}^{N}$
Require: Learnable parameters θ (attention logits $\mathbf{L}$, $\{\eta_t\}_{t=0}^{T-1}$ and β_k), Hyperparameters α, ϵ

— Augmented Step Dynamics —

function StepDynamics($\mathbf{u}; y, \theta$)
  1. Parse state:
    $(z_k, z_{k-1}) \leftarrow \mathbf{u}$; $z^{(0)} \leftarrow z_k$
  2. Nested Spatial Refinement:
  for $t = 0$ to $T - 1$ **do**
    Heterogeneous residual fields: $R^{(t)} = [r_1(z^{(t)}), \ldots, r_N(z^{(t)})]$
    Sinkhorn-constrained fusion weights: $w^{(t)} = W_t q_t, W_t \in \mathcal{B}_N, q_t = \mathrm{Softmax}(v_t)$
    Fuse residual directions: $v_{\text{fused}}^{(t)} = R^{(t)} w^{(t)}$
    Perform spatial refinement: $z^{(t+1)} = z^{(t)} - \eta_t \nabla\mathcal{E}(z^{(t)}; y) + \alpha v_{\text{fused}}^{(t)}$
  end for
  3. Temporal Hyper-Connection (Eq. equation 5):
    $z_{k+1} = z^{(T)} + \alpha\beta_k(z_k - z_{k-1})$
  return $\mathbf{u}_{\text{next}} = [z_{k+1}, z_k]^\top$
end function

— Stage 1: Forward Pass —

Determine domain-specific initialization z_{init}
Initialize $\mathbf{u}_0 = [z_{\text{init}}, z_{\text{init}}]^\top$
Find $\mathbf{u}^\star$ such that $\|\text{StepDynamics}(\mathbf{u}^\star) - \mathbf{u}^\star\| \le \epsilon$
Extract spatial component $z^\star$ from $\mathbf{u}^\star$
Compute loss $\mathcal{L}(z^\star, z_{\text{gt}})$

— Stage 2: Backward Pass —

At convergence, $z_{k+1} \approx z_k \approx z_{k-1}$.
Solve $v^\top(I - J_{\mathcal{T}}(z^\star)) = \partial\mathcal{L}/\partial z^\star$
$\nabla_\theta\mathcal{L} = v^\top \frac{\partial \text{StepDynamics}(z^\star;\theta)}{\partial\theta}$
$\theta \leftarrow \theta - \text{Optimizer}(\nabla_\theta\mathcal{L})$

— Stage 3: Inference Process —

Input: Test measurement y_{test}.
Initialize $\mathbf{u}_0 = [z_{\text{init}}, z_{\text{init}}]^\top$.
Fixed-Point Iteration:
  Repeat $k = 0, 1, \ldots$ $\mathbf{u}_{k+1} = \mathcal{T}_\theta(\mathbf{u}_k; y_{\text{test}})$
    Apply Anderson Acceleration to $\{\mathbf{u}_k\}$ sequence.
  Until $\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_2 / \|\mathbf{u}_k\|_2 < \epsilon_{\text{test}}$
return Restored image $z^\star$

is applied during backpropagation to improve convergence speed and stability, utilizing a history size of 5 and a maximum of 30 inner iterations. Both forward and backward solvers adopt a strict convergence threshold of 10^{-3}.

C.3 Experimental Settings

General Training & Optimization Details To ensure a fair comparison, we consider multiple representative denoiser backbones, including DnCNN Zhang et al. (2017b), DRUNet Zhang et al. (2021), Restormer Zamir et al. (2022), and SwinIR Liang et al. (2021). All of the denoisers are thoroughly retrained on our curated datasets, including *DIV2K*, *Flickr2K*, *WED*, and *CBSD500*. All experiments are implemented using PyTorch with a fixed random seed of 42 for reproducibility. The batch size is fixed to 2. All models are optimized using the Adam optimizer for 100 epochs. The initial learning rate is set to 5×10^{-5} with a weight decay of 10^{-6}. To stabilize the Deep Equilibrium training, gradient clipping with a norm threshold of 1.0 is enforced. We adopt a MultiStep learning rate schedule, decaying the learning rate by a factor of 0.5 at predefined milestones. Furthermore, early stopping is implemented with a patience of 30 epochs and a minimum improvement threshold of 10^{-6} to prevent overfitting.

Image Deblurring Task For the image deblurring task, we evaluate the performance on the standard *CBSD68* dataset using the classical eight *Levin09* blur kernels. The inverse problem is solved utilizing the RED framework with 15 unrolled iterations. During training, images are sampled from the *DIV2K*, *Flickr2K*, *WED*, *CBSD500*, and *ImageNet* datasets. The training data is randomly cropped into 256×256 RGB patches, with horizontal flipping applied at a probability of 0.5.

Black Hole Interferometric Imaging Task To validate our framework on highly non-linear, under-sampled inverse problems, we apply MHC-DEM to both synthetic and real-world Event Horizon Telescope (EHT) imaging tasks:

- **Synthetic M87 Simulation:** We benchmark algorithmic limits following the setup in PnP-DM Wu et al. (2024). Synthetic VLBI observations (`obs.uvfits`) are generated containing visibility amplitudes, closure phases, and log closure amplitudes. The reconstruction grid is set to 64×64 (FOV $= 158\,\mu$as). During the gradient computation of data-fidelity, the Regularized Maximum Likelihood (RML) weights are uniformly set to $\alpha_{d_1} = \alpha_{d_2} = \alpha_{d_3} = 100$. The maximum iteration limit is 2000, and the total flux is strictly constrained to $F_{\text{target}} = 222.02$.
- **Real M87* 2017 Observations:** We evaluate on the official EHT 2017 LO-band observations of M87* (MJD 57854) Akiyama et al. (2019). The raw interferometric data (`StokesI.uvfits`) was collected by the global EHT network. The EHT Collaboration SR1 official reconstruction (generated via CLEAN) serves as the reference Ground Truth. Unlike simulations, real EHT data is severely corrupted by atmospheric turbulence and station-based phase/gain errors. For this inversion, the total flux target is explicitly enforced at $F_{\text{target}} \approx 0.5964$ Jy. To prevent overfitting to real-world noise while maintaining physical geometry, the structural GradLoss weight is optimally set to $\zeta = 0.5$.

C.4 MHC-DEM Algorithm

We present the proposed MHC-DEM algorithm in Algorithm 1, which includes the augmented step dynamics in the network architecture, as well as the forward pass, backward pass during training, and the inference procedure.

D Detailed Theoretical Analysis and Proofs

This section provides the rigorous definitions, formal assumptions, and auxiliary lemmas supporting the theoretical analysis presented in Section 3 of the main paper. To establish a coherent analytical framework, we categorize these prerequisites into three parts: the geometric spaces constraining the fusion module, the mathematical formulation of our temporal dynamical system, and the regularity assumptions required for convergence.

D.1 Definition and Assumption

Definition 1 (Birkhoff Polytope)**.** The Birkhoff polytope $\mathcal{B}_N$ is the set of all $N \times N$ doubly stochastic matrices:

$$\mathcal{B}_N = \left\{ W \in \mathbb{R}^{N \times N} \;\middle|\; W_{ij} \geq 0,\ \sum_{i=1}^{N} W_{ij} = 1,\ \sum_{j=1}^{N} W_{ij} = 1 \right\}. \tag{16}$$

According to the Birkhoff–von Neumann theorem, $\mathcal{B}_N$ is the convex hull of all $N \times N$ permutation matrices. In our Sinkhorn-constrained attention, the projection operator maps the raw attention matrix onto $\mathcal{B}_N$. The critical property here is that multiplying any doubly stochastic matrix $W \in \mathcal{B}_N$ with a prior probability vector $q \in \Delta^{N-1}$ strictly yields a new vector $w = Wq$ that remains in Δ^{N-1}. This mathematically bridges the optimal transport theory with the convex combination requirement in our proofs.

Definition 2 (Heavy-Ball Method and Second-Order System)**.** The discrete update rule with an inertial term, historically known as the Polyak heavy-ball method, is given by $z_{k+1} = z_k - \eta \nabla \mathcal{E}(z_k) + \beta_k (z_k - z_{k-1})$.

Because the update of z_{k+1} depends on both current (z_k) and historical (z_{k-1}) states, the sequence forms a second-order discrete Markov process. Traditional fixed-point theorems (e.g., Banach Fixed-Point Theorem), which only apply to single-step mappings $z_{k+1} = \Phi(z_k)$, cannot be directly used to prove convergence. This necessitates the introduction of state space augmentation.

To ensure the theoretical tractability of deep neural networks within iterative frameworks, we adopt the following standard assumptions, which are widely accepted in the analysis of RED and DEMs.

Assumption 1 (Normalized Denoisers)**.** *Each pre-trained base denoiser D_i in the network bank is strictly 1-Lipschitz continuous, satisfying $\|D_i(x) - D_i(y)\|_2 \leq \|x - y\|_2$ for any inputs x, y.*

This is a standard requisite for convergent RED algorithms Ryu et al. (2019); Xue et al. (2022). It bounds the Lipschitz constant of the residual mapping $r_i(z) = D_i(z) - z$ to at most 2. This is the cornerstone of Lemma 1, guaranteeing that the deep denoisers do not cause numerical explosions.

D.2 Proof of Lemma 1

Before analyzing the stability of the overall system, we first establish the Lipschitz upper bound for the fusion operator under the Sinkhorn constraint.

Lemma 1 (Lipschitz Stability of Sinkhorn-Constrained Fusion)**.** *Let the fusion weights w be generated via the Sinkhorn projection, i.e., $w = Wq$, where $W \in \mathcal{B}_N$ (the Birkhoff polytope of doubly stochastic matrices) and $q \in \Delta^{N-1}$ is the prior probability distribution. The fused residual operator*

$$v_{\textit{fused}}(z) = \sum_{i=1}^{N} w_i \big(D_i(z) - z \big)$$

satisfies the following Lipschitz continuity condition:

$$\|v_{\textit{fused}}(x) - v_{\textit{fused}}(y)\|_2 \leq 2\|x - y\|_2. \tag{17}$$

Proof. By the properties of the Sinkhorn projection, the output matrix W is strictly doubly stochastic (i.e., all entries are non-negative, and both row and column sums equal 1). Since q lies on the probability simplex Δ^{N-1}, their product $w = Wq$ is guaranteed to reside on the probability simplex as well. Thus, the weights satisfy the strict non-negativity and normalization conditions:

$$w_i \geq 0 \quad \forall i, \quad \text{and} \quad \sum_{i=1}^{N} w_i = 1.$$

We define the residual operator for the i-th base denoiser as $r_i(z) = D_i(z) - z$. Under Assumption 1, each denoiser satisfies $\|D_i(x) - D_i(y)\|_2 \leq \|x - y\|_2$. Utilizing the triangle inequality, we can bound

the Lipschitz constant of an individual residual operator:

$$\begin{aligned}\|r_i(x) - r_i(y)\|_2 &= \|\big(D_i(x) - x\big) - \big(D_i(y) - y\big)\|_2 \\ &= \|\big(D_i(x) - D_i(y)\big) - (x - y)\|_2 \\ &\le \|D_i(x) - D_i(y)\|_2 + \|x - y\|_2 \\ &\le \|x - y\|_2 + \|x - y\|_2 = 2\|x - y\|_2. \end{aligned} \tag{18}$$

Thus, $\|r_i\|_{\text{Lip}} \le 2$ holds for all i.

Consequently, the fused residual operator $v_{\text{fused}}(z)$ acts as a convex combination of the individual residual operators $r_i(z)$:

$$v_{\text{fused}}(z) = \sum_{i=1}^{N} w_i r_i(z).$$

By the properties of Lipschitz constants under convex combinations, the Lipschitz constant of the composite operator is bounded by the weighted sum of the components' Lipschitz constants:

$$\begin{aligned}\|v_{\text{fused}}(x) - v_{\text{fused}}(y)\|_2 &= \left\| \sum_{i=1}^{N} w_i \big(r_i(x) - r_i(y)\big) \right\|_2 \\ &\le \sum_{i=1}^{N} w_i \|r_i(x) - r_i(y)\|_2 \\ &\le \sum_{i=1}^{N} w_i \big(2\|x - y\|_2\big) \\ &= 2\|x - y\|_2 \left(\sum_{i=1}^{N} w_i \right) = 2\|x - y\|_2. \end{aligned} \tag{19}$$

Therefore, we conclude that $\|v_{\text{fused}}\|_{\text{Lip}} \le 2$.

Remark:

The Lipschitz bound proven above relies only on $w \in \Delta^{N-1}$ and is therefore shared by any simplex constraint, including softmax. The additional doubly-stochastic structure (Sinkhorn / Birkhoff) is *not* required for this bound; its role, established empirically in Table 8 and discussed in Section 2.3, is to prevent the simplex point from collapsing to a near-vertex during recursive equilibrium iteration—i.e., to ensure the bound is realized by a genuinely mixed convex combination rather than a degenerate one. Without any simplex constraint, the freely learned weights w_i may violate normalization or become unbounded, in which case the Lipschitz constant amplifies to $\sum_i |w_i| \cdot \|r_i\|_{\text{Lip}}$ and forward/backward numerical stability is lost.

□

D.3 Proof of Lemma 2

Lemma 2 (Spectral Radius Bound for Block Companion Matrices)**.** *Consider a block companion matrix of the form* $M = \begin{bmatrix} A & B \\ I & 0 \end{bmatrix}$*. If the sub-matrices satisfy the norm condition* $\|A\|_2 + \|B\|_2 < 1$*, then the spectral radius of the block matrix is strictly less than 1, i.e.,* $\rho(M) < 1$.

Proof. Let $\lambda \in \mathbb{C}$ be an arbitrary eigenvalue of M, with the corresponding non-zero eigenvector $v = [x^\top, y^\top]^\top \neq \mathbf{0}$. From the characteristic equation $Mv = \lambda v$, we obtain the following system:

$$Ax + By = \lambda x, \tag{20}$$

$$x = \lambda y. \tag{21}$$

Substituting Eq. equation 21 into Eq. equation 20 to eliminate x yields:

$$A(\lambda y) + By = \lambda(\lambda y) \implies \lambda^2 y = \lambda A y + By.$$

Taking the ℓ_2-norm on both sides and applying the sub-multiplicativity and the triangle inequality of matrix norms, we have:

$$\begin{aligned} \|\lambda^2 y\|_2 &= \|\lambda A y + B y\|_2 \\ |\lambda|^2 \|y\|_2 &\leq |\lambda| \cdot \|A\|_2 \|y\|_2 + \|B\|_2 \|y\|_2. \end{aligned} \tag{22}$$

We claim that $y \neq \mathbf{0}$. If $y = \mathbf{0}$, Eq. equation 21 dictates that $x = \mathbf{0}$, which implies $v = \mathbf{0}$, contradicting the definition of an eigenvector. Hence, $\|y\|_2 > 0$. Dividing both sides by $\|y\|_2$ gives:

$$|\lambda|^2 \leq \|A\|_2 |\lambda| + \|B\|_2 \implies |\lambda|^2 - \|A\|_2 |\lambda| - \|B\|_2 \leq 0.$$

To find the upper bound for $|\lambda|$, we define the real quadratic function:

$$f(t) = t^2 - \|A\|_2 t - \|B\|_2.$$

The condition $f(|\lambda|) \leq 0$ indicates that $|\lambda|$ must be bounded by the largest real root of the equation $f(t) = 0$, which is given by:

$$t^* = \frac{\|A\|_2 + \sqrt{\|A\|_2^2 + 4\|B\|_2}}{2}.$$

Since the parabola opens upwards, to prove $t^* < 1$, it is sufficient to show that the function value at $t = 1$ is strictly positive ($f(1) > 0$) and the axis of symmetry is less than 1 ($t = \|A\|_2/2 < 1$).

Evaluating the function at $t = 1$:

$$f(1) = 1^2 - \|A\|_2(1) - \|B\|_2 = 1 - (\|A\|_2 + \|B\|_2).$$

Given the hypothesis $\|A\|_2 + \|B\|_2 < 1$, we have $f(1) > 0$. This guarantees that the largest root t^* is strictly less than 1. Consequently, for any eigenvalue λ, its magnitude satisfies:

$$|\lambda| \leq t^* < 1.$$

Since the spectral radius $\rho(M)$ is the supremum among the absolute values of all eigenvalues, it follows that $\rho(M) < 1$. □

D.4 Proof of Theorem 1

Proof. Because our system incorporates an inertia term from the previous time step, standard single-step Markovian dynamics no longer apply. We augment the state space and define:

$$\mathbf{u}_k = \begin{bmatrix} z_k \\ z_{k-1} \end{bmatrix}, \quad \mathbf{T}(\mathbf{u}_k) = \begin{bmatrix} \Phi(z_k, z_{k-1}) \\ z_k \end{bmatrix},$$

where the update function $\Phi(z_k, z_{k-1})$ integrates the general data-fidelity gradient descent, spatial priors, and temporal inertia:

$$\Phi(z_k, z_{k-1}) = z_k - \eta \nabla \mathcal{E}(z_k) + \alpha v_{\text{fused}}(z_k) + \alpha \beta_k (z_k - z_{k-1}).$$

To analyze the stability within the local neighborhood of the fixed point, we differentiate the augmented operator $\mathbf{T}$ with respect to $\mathbf{u}_k$ to obtain the block Jacobian matrix:

$$J_{\mathbf{T}} = \frac{\partial \mathbf{T}}{\partial \mathbf{u}_k} = \begin{bmatrix} \frac{\partial \Phi}{\partial z_k} & \frac{\partial \Phi}{\partial z_{k-1}} \\ I & 0 \end{bmatrix} = \begin{bmatrix} J_{z_k} & J_{z_{k-1}} \\ I & 0 \end{bmatrix}.$$

We evaluate the spectral norm bounds for the two sub-blocks respectively. Note that the derivative of the data-fidelity gradient $\nabla \mathcal{E}(z_k)$ is precisely its Hessian matrix $\nabla^2 \mathcal{E}(z_k)$:

$$J_{z_k} = (I - \eta \nabla^2 \mathcal{E}(z_k)) + \alpha J_{\mathcal{P}} + \alpha \beta_k I, \tag{23}$$

$$J_{z_{k-1}} = -\alpha \beta_k I. \tag{24}$$

Here, $J_{\mathcal{P}}$ is the Jacobian of the fusion operator v_{fused}.

(1) Temporal Sensitivity (Inertia Overhead, corresponding to block B in Lemma 2):

$$\|J_{z_{k-1}}\|_2 = \| - \alpha \beta_k I\|_2 = \alpha \beta_k.$$

(2) Spatial Sensitivity (Corresponding to block A in Lemma 2): By the sub-additivity of matrix norms and applying Lemma 1 ($\|J_{\mathcal{P}}\|_2 \leq 2$), we obtain:

$$\begin{aligned}\|J_{z_k}\|_2 &\leq \|I - \eta\nabla^2\mathcal{E}(z_k)\|_2 + \alpha\|J_{\mathcal{P}}\|_2 + \alpha\beta_k\|I\|_2 \\ &\leq \|I - \eta\nabla^2\mathcal{E}(z_k)\|_2 + 2\alpha + \alpha\beta_k. \end{aligned} \tag{25}$$

(3) Total Stability Budget: For the system to be asymptotically stable, we require the spectral radius of the augmented Jacobian to be strictly less than 1 ($\rho(J_{\mathbf{T}}) < 1$). According to Lemma 2, a sufficient condition for this is $\|J_{z_k}\|_2 + \|J_{z_{k-1}}\|_2 < 1$. Summing the derived bounds yields:

$$\|J_{z_k}\|_2 + \|J_{z_{k-1}}\|_2 \leq \big(\|I - \eta\nabla^2\mathcal{E}(z_k)\|_2 + 2\alpha + \alpha\beta_k\big) + \alpha\beta_k = \|I - \eta\nabla^2\mathcal{E}(z_k)\|_2 + 2\alpha + 2\alpha\beta_k.$$

When this sum is strictly less than 1 in the neighborhood of the equilibrium (i.e., Eq. equation 11 holds), Lemma 2 ensures $\rho(J_{\mathbf{T}}) < 1$. Consequently, the augmented operator $\mathbf{T}$ operates as a local contraction mapping. By the Banach Fixed-Point Theorem, the iterative system is guaranteed to stably converge to a unique fixed point.

Interpretation of the Stability Budget: The inequality decomposes the system's stability into three overhead components: the data-fidelity curvature $\|I - \eta\nabla^2\mathcal{E}(z_k)\|_2$, the spatial prior term 2α, and the temporal inertia term $2\alpha\beta_k$. The simplex constraint $w \in \Delta^{N-1}$ is what bounds the spatial Jacobian $\|J_{\mathcal{P}}\|_2$ by 2α and thus preserves a stability margin for the inertia momentum β_k, regardless of whether the physical forward model is linear or non-linear. The Sinkhorn / Birkhoff refinement makes a complementary contribution that is not visible at the level of this scalar bound: as shown in Section 2.3 and Table 8, it keeps the realized fusion weights non-degenerate so that the margin is actually used by all priors rather than monopolized by one across recursive iterations.

□

D.5 Proof of Theorem 2

Proof. In a nested DEM, the basic single-step mapping is replaced by a deep composite mapping network $\mathcal{T}_\theta$. According to the multivariate chain rule and the composition properties of Lipschitz continuity, the Lipschitz constant of the composite operator $L_{\mathcal{T}}$ is bounded by the product of the individual layer constants:

$$L_{\mathcal{T}} \leq \prod_{t=0}^{T-1} L_t. \tag{26}$$

When analyzing the nested framework as an augmented dynamical system over $\mathbf{u}_k$, the norm bound of the spatial sensitivity block (Block A) is dictated by $\|J_{\mathcal{T}}\|_2$. Analogous to the derivation in Theorem 1, if the system incorporates an internal inertia feedback scaled by $\alpha\beta_k$, the upper bound for the spatial Jacobian norm becomes:

$$\|A\|_2 \leq L_{\mathcal{T}} + \alpha\beta_k.$$

The temporal sensitivity block (Block B, external inertia connection) remains bounded by:

$$\|B\|_2 \leq \alpha\beta_k.$$

Substituting these into the block matrix stability criterion $\|A\|_2 + \|B\|_2 < 1$ from Lemma 2, the global convergence requirement for the nested system is:

$$(L_{\mathcal{T}} + \alpha\beta_k) + \alpha\beta_k < 1 \implies L_{\mathcal{T}} + 2\alpha\beta_k < 1.$$

Finally, replacing $L_{\mathcal{T}}$ with the composite bound $\prod_{t=0}^{T-1} L_t$ yields the ultimate stability condition:

$$\prod_{t=0}^{T-1} L_t + 2\alpha\beta_k < 1.$$

This implies that as long as the Lipschitz constant L_t of each internal layer is strictly regulated (e.g., via Spectral Normalization or Weight Decay) such that the inequality holds, the augmented system preserves strict contractivity, thereby ensuring stable convergence during the forward root-finding process. □

In deep equilibrium models, the gradient of the loss function with respect to parameters relies on implicit differentiation, which involves computing the inverse $(I - J_{\mathbf{T}}^{\star})^{-1}$ at the fixed point. When the stability condition $\rho(J_{\mathbf{T}}^{\star}) < 1$ holds, the inverse can be safely computed via the Neumann series expansion:

$$(I - J_{\mathbf{T}}^{\star})^{-1} = \sum_{k=0}^{\infty} (J_{\mathbf{T}}^{\star})^{k}. \tag{27}$$

This guarantees that gradients neither explode nor vanish during backpropagation. To handle transient instability during the early training phases of implicit networks, we additionally employ gradient clipping on the backward pass to ensure the Jacobian inversion remains numerically well-conditioned.

E Additional Experiments

E.1 Black Hole Interferometric Imaging

Unlike standard linear image restoration, black hole interferometric imaging Chael et al. (2018) represents a highly ill-posed, strictly non-linear inverse problem. We deploy our MHC-DEM framework on this challenging astronomical task to demonstrate its extraordinary stability and generalization capability when confronted with severe information loss and complex physical forward models.

E.1.1 Problem Formulation

The Event Horizon Telescope (EHT) is a global network of radio telescopes. It does not capture images directly; instead, it measures the complex visibility $V(\mathbf{u})$, which corresponds to the 2D Fourier transform of the target black hole's sky brightness distribution $I(\boldsymbol{\theta})$:

$$V(\mathbf{u}) = \iint I(\boldsymbol{\theta})\, e^{-2\pi i\, \mathbf{u}\cdot\boldsymbol{\theta}}\, d\boldsymbol{\theta}. \tag{28}$$

Here, $\boldsymbol{\theta} = (\theta_x, \theta_y)$ denotes the 2D angular coordinates on the sky, and $\mathbf{u} = (u, v)$ represents the spatial frequency coordinates sampled by a pair of telescopes.

In our computational framework, the continuous emission $I(\boldsymbol{\theta})$ is discretized into a pixelized image state $z \in \mathbb{R}_{+}^{H\times W}$ (where $H = W = 64$, pixel size $\Delta\theta = 2.4687\,\mu$as, representing a Field of View of $158\,\mu\text{as}^2$). If the measurements were complete and perfect, recovering z would be a straightforward linear inverse Fourier transform. However, EHT data is extremely sparse and the raw complex visibilities are critically corrupted by unpredictable phase delays caused by the Earth's atmospheric turbulence above each telescope.

To achieve robust reconstruction, domain experts rely on *closure quantities*—cleverly derived combinations of measurements that are mathematically invariant to these station-specific atmospheric errors. The actual observed data y fed into our inverse solver consists of:

$$y = \left\{ |V_m|^{\text{obs}},\ \phi_{\text{cp},n}^{\text{obs}},\ \log \mathcal{C}_p^{\text{obs}} \right\}, \tag{29}$$

where the components are defined as follows:

- $|V_m|^{\text{obs}}$: The **visibility amplitudes** measured at baseline m.
- $\phi_{\text{cp},n}^{\text{obs}}$: The **closure phase** for a triangle of three telescopes $n = (i, j, k)$. It is defined as the sum of their visibility phases ($\phi_{ij} + \phi_{jk} + \phi_{ki}$). This cyclic summation exactly cancels out the local atmospheric phase delays at the three stations.
- $\log \mathcal{C}_p^{\text{obs}}$: The **log closure amplitude** for a quadrangle of four telescopes $p = (i, j, k, l)$. It is defined as the log-ratio of their visibility amplitudes $\log(|V_{ij}||V_{kl}|/|V_{ik}||V_{jl}|)$, which elegantly cancels out station-specific telescope gain errors.

While these closure quantities successfully eliminate atmospheric noise, they fundamentally alter the inverse problem. The forward mapping from the image pixels z to these robust observables ($z \mapsto \phi_{\text{cp}}, \log \mathcal{C}$) requires complex multiplications, divisions, and trigonometric operations over the Fourier coefficients. This transforms a simple linear projection into a **highly non-linear, severely non-convex optimization problem**, serving as an extreme stress test for the stability of our implicit DEM framework.

E.1.2 SOLVING EHT WITHIN THE MHC-DEM FRAMEWORK

Traditional Regularized Maximum Likelihood (RML) methods tackle this by relying on heavy "human-in-the-loop" tuning of hand-crafted priors (e.g., Total Variation or Maximum Entropy), which inherently biases the astrophysical interpretation. Recent advanced probabilistic methods replace these with deep priors, but often require computationally expensive Langevin dynamics. We seamlessly unify the EHT reconstruction pipeline into our highly efficient MHC-DEM framework by instantiating three physically-motivated components:

1. Data Fidelity as a Non-Linear Gradient Flow We replace the linear $A^\top(Az - y)$ projection with the analytical gradient of the RML objective (computed via the *Ehtim* library). The data-fidelity energy is formulated as:

$$\mathcal{E}_{\text{EHT}}(z) = \alpha_{d_1}\chi^2_{\text{amp}}(z) + \alpha_{d_2}\chi^2_{\text{cp}}(z) + \alpha_{d_3}\chi^2_{\text{logcamp}}(z), \tag{30}$$

where $\chi^2_{\text{amp}} = \sum_m \left(\frac{|V_m|^{\text{obs}} - |V_m|^{\text{pred}}(z)}{\sigma_m}\right)^2$, and χ^2_{cp} and χ^2_{logcamp} strictly enforce closure consistency. The non-linear gradient $\nabla\mathcal{E}_{\text{EHT}}(z)$ directly drives the data-consistency step (Eq. 6). To enforce astrophysical validity, the latent state is implicitly clamped to the non-negative orthant and normalized to a target flux F_{target} during the nested loop. Concurrently, the bias of hand-crafted regularizers is entirely circumvented by our Sinkhorn-Constrained Heterogeneous Prior Bank, which intrinsically projects the intermediate states onto the valid, high-resolution manifold of black hole emissions without causing numerical explosions.

2. Physically-Motivated Inertial Initialization (z_0) Standard DEM solvers initialize the state z_0 with a pseudo-inverse (e.g., $A^\top y$). However, the extreme sparsity of the EHT (u, v) coverage makes inverse-FFT initializations pathologically noisy. Instead, we initiate the DEM trajectory using an untrained, static parametric model representing coarse physical assumptions. Let $\tilde{z}_0$ be defined as:

$$\tilde{z}_0 = A_{\text{ring}}\mathbf{R}(r_0, \sigma_r) + \sum_{i=1}^{8}\mathbf{S}_i + \mathbf{B}, \tag{31}$$

where $\mathbf{R}$ is an elliptical ring (radius $r_0 = 12.0\,\mu$as, width $\sigma_r = 3.0$), $\mathbf{S}_i$ are Gaussian spots capturing relativistic Doppler beaming, and $\mathbf{B}$ is the background flow. We enforce exact flux conservation by initializing $z_0 = \tilde{z}_0 \frac{F_{\text{target}}}{\|\tilde{z}_0\|_1}$. Crucially, to align with our theoretical stability bounds, we set $z_{-1} = z_0$. This guarantees the initial temporal momentum (Eq. 5) is precisely zero, ensuring the second-order dynamical system accelerates smoothly away from the physical prior without erratic initial shocks.

3. Implicit Supervision via Structural GradLoss To supervise the implicit operator $\mathcal{T}_\theta$ via the fixed-point gradient (Eq. 15), standard Mean Squared Error (MSE) is insufficient, as it blurs the sharp, sub-pixel photon ring boundaries critical for assessing black hole mass and shadow depth. We define the equilibrium training loss as a composite of MSE and a Structural Sobel Gradient Loss:

$$\mathcal{L}(z^\star, z_{\text{GT}}) = \frac{1}{HW}\|z_c^\star - z_{\text{GT}}\|_2^2 + \zeta\frac{1}{HW}\|\nabla_S z_c^\star - \nabla_S z_{\text{GT}}\|_2^2, \tag{32}$$

where $z_c^\star = \max(z^\star, 0)$ is the strictly non-negative clamped equilibrium state, $\nabla_S = (G_x*, G_y*)$ represents the 2D spatial gradient via Sobel operators, and ζ balances the terms. This guides the heterogeneous prior bank to stabilize not just intensities, but the high-frequency structural topology.

The assessment of black hole image reconstruction requires strict adherence to astrophysical properties, moving beyond standard visual metrics. We employ the following quantitative and domain-specific criteria:

Quantitative Metrics.

1. **Mean Squared Error (MSE):** Evaluates the pixel-wise reconstruction fidelity between the flux-normalized reconstruction and the reference image:

$$\text{MSE}(z_c^\star, z_{\text{GT}}) = \frac{1}{HW}\|z_c^\star - z_{\text{GT}}\|_2^2. \tag{33}$$

2. **Pearson Correlation (ρ, or NXCOR):** In astronomical interferometry, Normalized Cross-Correlation (NXCOR) evaluates pure structural and morphological similarity, entirely in-

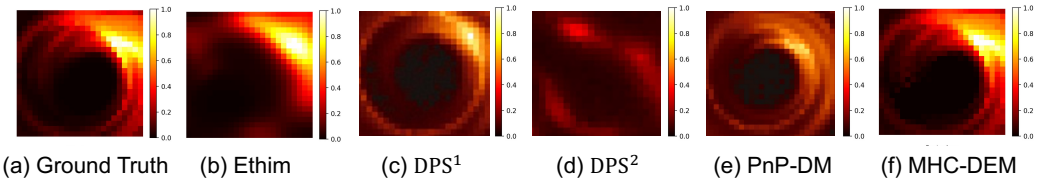


Figure 8: Extended local zoomed-in evaluation comparing deterministic and generative reconstruction methods. Generative methods (DPS Chung et al. (2023) and PnP-DM Wu et al. (2024)) exhibit a noticeable gap in reconstruction quality compared to the Ground Truth. Notably, the diffusion-based DPS method demonstrates severe instability across different sampling trajectories; while DPS[1] yields a heavily blurred ring, DPS[2] suffers from a catastrophic reconstruction failure. In contrast, our deterministic MHC-DEM guarantees stable, artifact-free, and highly exact structural recovery.

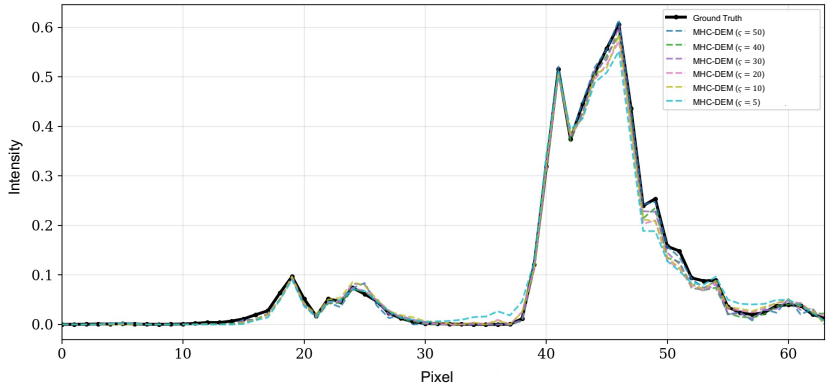


Figure 9: 1D horizontal cross-section intensity profiles comparing the Ground Truth against MHC-DEM under varying structural constraint weights (ζ). While all configurations closely track the true structure, $\zeta = 50.0$ yields the optimal fit, perfectly capturing the deep contrast of the central shadow and the multi-path physical characteristics (sharp multiple peaks representing primary and lensed photon rings) without introducing artifacts.

variant to global flux scaling or DC offsets:

$$\rho = \frac{\mathrm{Cov}(\mathrm{vec}(z_c^\star), \mathrm{vec}(z_{\mathrm{GT}}))}{\sigma_{z_c^\star}\, \sigma_{\mathrm{GT}}}. \tag{34}$$

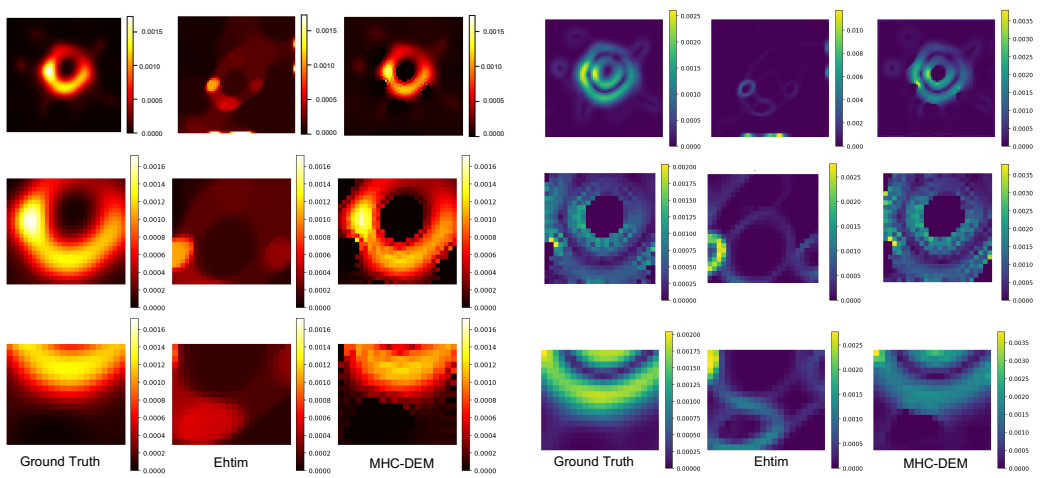
Achieving $\rho \approx 1.0$ is critical to confirm that the reconstructed asymmetry accurately reflects relativistic beaming rather than algorithmic artifacts.

Qualitative Criteria.

- **Morphological Fidelity (Photon Ring & Shadow):** The algorithm must precisely recover the central shadow depression and the asymmetric brightness ring caused by the Doppler boosting of the accretion flow.
- **Artifact Suppression:** Due to the severe under-sampling of the (u, v) plane, naive inversions generate pervasive, scattered high-frequency speckles. Superior reconstructions must exhibit a strictly clean background devoid of false sources.

E.1.3 Simulation Data Experiment

Precision vs. Generative Diversity. Recent diffusion-based methods such as DPS Chung et al. (2023) and PnP-DM Wu et al. (2024) have shown strong generative capability in inverse problems



(a) Reconstruction Results (b) Edge Detection

Figure 10: Main reconstruction results on real M87* 2017 observations. (a) Visual comparisons among Ground Truth (SR1), Ehtim, and MHC-DEM. Rows from top to bottom display: overall morphology, isolated main photon ring, and highly zoomed-in ring edges. (b) Corresponding structural edge maps extracted for the three rows, explicitly demonstrating MHC-DEM's superior capacity to recover sharp, continuous astrophysical boundaries compared to the fragmented Ehtim results.

by modeling posterior distributions through stochastic sampling. However, black hole interferometric imaging is an extremely ill-posed reconstruction problem under severely sparse measurements, where strict data fidelity is critical to avoid hallucinated structures.

As shown in Fig. 8, diffusion-based approaches exhibit noticeable instability in this low-data regime. Different sampling runs of DPS produce inconsistent reconstructions, ranging from overly smooth structures to severe artifact collapse, while PnP-DM struggles to preserve real ring boundaries. In contrast, MHC-DEM performs deterministic equilibrium optimization under explicit data-consistency constraints. By directly solving for a stable MAP equilibrium state rather than relying on stochastic sampling, our method reconstructs clearer photon-ring structures and more reliable shadow boundaries, which are essential for scientific measurements such as black hole diameter estimation.

Parameter Sensitivity and Multi-Path Physical Characteristics. Fig. 9 further presents 1D horizontal intensity profiles under different gradient penalty parameters ζ. MHC-DEM consistently follows the ground-truth distribution across different settings, accurately recovering both the peak emission regions and the deep central shadow. Importantly, the reconstructed profiles preserve the multi-path characteristics of black hole imaging, including multiple sharp intensity peaks produced by gravitational lensing and higher-order photon rings. Smaller ζ values slightly underestimate peak responses, while larger values improve edge sharpness and better match the ground-truth gradient distribution. In particular, $\zeta = 50.0$ achieves the best balance, accurately recovering fine-grained optical structures without introducing noticeable high-frequency artifacts.

E.1.4 Real M87* 2017 Observations

Table 5: Quantitative results on real EHT 2017 M87* observations.

Method	MSE ($\times 10^{-7}$) $\downarrow$	Correlation (ρ) $\uparrow$
Ehtim Akiyama et al. (2019) + GS prior	2.43	-0.007
Ehtim Akiyama et al. (2019) + TV prior	2.36	-0.007
Ehtim Akiyama et al. (2019) + TVLog prior	1.05	0.113
MHC-DEM ($\zeta = 0.5$)	**0.08**	**0.954**

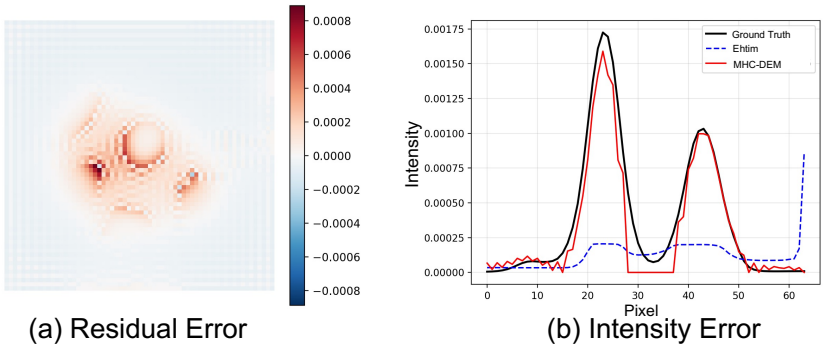


Figure 11: Detailed error and intensity profiling on real M87* data. (a) Residual map computing the pixel-wise difference between MHC-DEM and Ground Truth, indicating near-zero structural error. (b) 1D horizontal cross-section intensity profiles. MHC-DEM accurately matches the GT peak emission locations and shadow contrast, whereas Ehtim exhibits massive structural deviations and flux blurring.

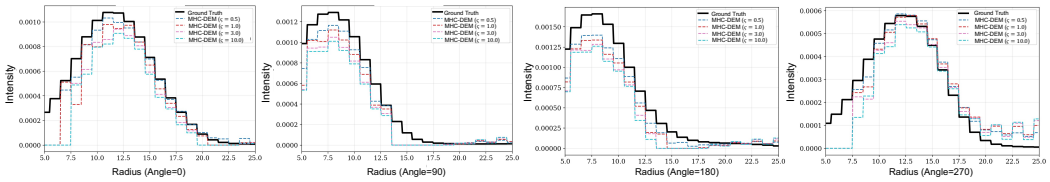


Figure 12: Radial intensity distributions sampled at four azimuthal angles ($0°, 90°, 180°, 270°$) under varying ζ values. The profiles demonstrate physically consistent ring radii and capture the Doppler beaming asymmetry. For real-world data corrupted by observation noise, a gentle structural constraint ($\zeta = 0.5$) yields the optimal alignment with the GT, avoiding the noise-amplification caused by larger ζ values.

Morphological Fidelity and Edge Precision. The real-world results in Table 5 reveal a key limitation of traditional inverse solvers. Although Ehtim achieves numerically plausible MSE values, its Pearson correlation remains extremely low, indicating poor structural consistency with the ground truth. In contrast, MHC-DEM substantially improves both MSE and correlation, demonstrating stronger preservation of global morphology and spatial energy distribution.

These improvements are further confirmed in Fig. 10. The Ehtim produces overly diffuse structures and fails to clearly recover the central shadow region, whereas MHC-DEM reconstructs a well-defined crescent morphology with the characteristic brightness asymmetry of M87*. The zoomed-in views and edge maps further show that our method produces sharper photon-ring boundaries and more accurate structural contours, closely matching the GT.

Error Analysis and Cross-Sectional Accuracy. Fig. 11 provides residual and cross-sectional analyses of the reconstructed images. The residual maps demonstrate that MHC-DEM introduces significantly fewer structural artifacts than Ehtim. In addition, the horizontal intensity profiles show that our method more accurately captures both the sharp emission peaks and the deep flux depression corresponding to the black hole shadow, while Ehtim exhibits noticeable spatial displacement and oversmoothing effects.

Radial Physics and the ζ Discrepancy. To further validate the recovered astrophysical geometry, Fig. 12 presents radial intensity profiles sampled at four azimuthal angles. MHC-DEM consistently preserves the physically correct ring radius and brightness asymmetry across different directions.

The optimal structural penalty differs between synthetic and real observations. While synthetic experiments favor a larger ζ value for sharper reconstruction, real EHT data achieves the best performance at a much smaller $\zeta = 0.5$. This difference arises because real observations contain substantial atmospheric noise and calibration uncertainty. A large structural penalty may over-enhance

noisy high-frequency components, whereas a smaller ζ provides gentler structural guidance and better balances morphological regularization with measurement fidelity.

E.2 Image Deblurring

While black hole imaging serves as a strictly non-linear stress test, image deblurring represents the fundamental benchmark for linear inverse problems. We apply the MHC-DEM framework to this classic task to demonstrate its capacity to recover heavily attenuated high-frequency textures using standard physical degradation models.

E.2.1 Problem Formulation

In image deblurring, the degradation process is modeled as the spatial convolution of a clean image $\mathbf{z}^\star$ with a known blur kernel k, followed by the addition of Additive White Gaussian Noise (AWGN) ϵ. Mathematically, following standard PnP restoration literature Zhang et al. (2021), this is formulated as a linear system:

$$y = A\mathbf{z}^\star + \epsilon, \tag{35}$$

where $A \in \mathbb{R}^{d\times d}$ denotes the degradation matrix. Under circular boundary conditions, A is a block circulant matrix with circulant blocks (BCCB) representing 2D convolution ($A\mathbf{z} \Leftrightarrow k \otimes \mathbf{z}$). Because the blur kernel k acts as a low-pass filter, the matrix A is highly ill-conditioned. Consequently, direct inversion (e.g., $z = A^{-1}y$) inevitably amplifies the high-frequency noise, necessitating robust, manifold-constrained regularization.

E.2.2 Solving Deblurring within the MHC-DEM Framework

Traditional RED methods typically solve this via half-quadratic splitting or proximal gradient descent, alternating between a closed-form data-fidelity step and a single deep denoiser step. MHC-DEM elevates this process by integrating it into a continuous dynamical system:

1. Data Fidelity as a Linear Gradient Flow For the deblurring task, the data-fidelity energy is the standard quadratic Gaussian log-likelihood:

$$\mathcal{E}_{\text{deblur}}(z) = \frac{1}{2}\|Az - y\|_2^2. \tag{36}$$

The analytical gradient of this energy is strictly linear: $\nabla\mathcal{E}_{\text{deblur}}(z) = A^\top(Az - y)$. Here, the transposed operator $A^\top$ corresponds to the 2D convolution with the flipped blur kernel ($k^\top \otimes \cdot$). We inject this explicit gradient into our nested micro-dynamic loop (Eq. 6). Unlike traditional RED approaches that rely on a single CNN denoiser to invert the blur Zhang et al. (2021), our Sinkhorn-Constrained Heterogeneous Prior Bank dynamically fuses CNNs (for sharp local edges) and Transformers (for non-local self-similarity) to holistically reconstruct the missing high-frequency spectrum without violating Lipschitz stability bounds.

2. Observation-Driven Inertial Initialization (z_0) Unlike the highly sparse interferometric imaging task requiring physical priors, the dense observations in deblurring provide a natural starting point. We directly initialize the deep equilibrium trajectory using the degraded observation:

$$z_0 = y. \tag{37}$$

Consistent with our theoretical stability prerequisites for the second-order system, we initialize the historical state as $z_{-1} = z_0 = y$. This ensures that the inertial momentum term $\alpha\beta_k(z_0 - z_{-1})$ evaluates to zero at the first iteration, allowing the MHC-DEM solver to initiate the gradient flow smoothly without erratic momentum shocks.

3. Implicit Supervision via Robust Charbonnier Loss To supervise the implicit operator $\mathcal{T}_\theta$ through implicit differentiation, we require a loss function that tightly penalizes structural deviations without over-smoothing. For deblurring, we supervise the equilibrium state $z^\star$ using the Charbonnier penalty (a differentiable approximation of the ℓ_1 norm), which is highly effective in restoring sharp image gradients:

$$\mathcal{L}(z^\star, z_{\text{GT}}) = \frac{1}{HW}\sum_{i=1}^{HW}\sqrt{(z_i^\star - z_{\text{GT},i})^2 + \epsilon^2}, \tag{38}$$

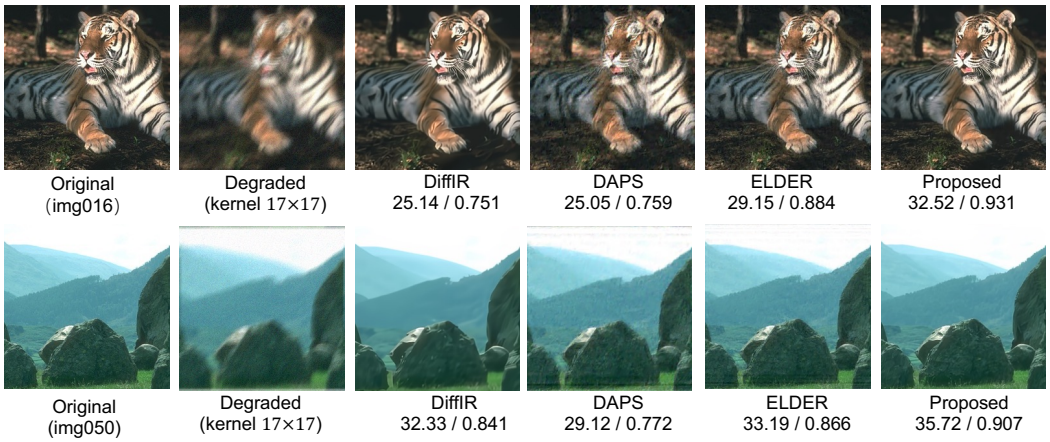


Figure 13: Visual comparison on *CBSD68* (images 'img016' and 'img050') under Gaussian noise ($\sigma = 0.03$) and a 17×17 blur kernel.

Table 6: Comparison of average PSNR (dB) and SSIM for image deblurring on *CBSD68*. The best results are in **bold** and the second best are underlined. All test images are center-cropped to 256×256.

Models	#0	#1	#2	#3	#4	#5	#6	#7
GS-PnP	31.70 (0.8936)	31.92 (0.9037)	32.19 (0.9163)	30.66 (0.8708)	33.19 (0.9314)	33.01 (0.9237)	32.15 (0.9146)	31.91 (0.9101)
DE-Prox	32.35 (0.9166)	32.27 (0.9156)	32.85 (0.9221)	31.69 (0.9019)	33.28 (0.9308)	33.40 (0.9319)	32.88 (0.9172)	32.43 (**0.9203**)
ELDER	32.11 (0.9166)	31.93 (0.9064)	32.67 (0.9190)	31.52 (0.8990)	33.17 (0.9293)	33.25 (0.9296)	32.59 (0.9165)	32.18 (0.9051)
DPS	28.22 (0.8232)	28.07 (0.8200)	28.89 (0.8379)	27.52 (0.8079)	29.08 (0.8420)	28.32 (0.8254)	27.83 (0.8147)	27.56 (0.8088)
DiffIR	29.54 (0.8521)	29.43 (0.8497)	29.98 (0.8617)	29.13 (0.8431)	30.73 (0.8781)	30.38 (0.8705)	30.02 (0.8626)	29.59 (0.8532)
DAPS	30.24 (0.8674)	30.59 (0.8751)	31.30 (0.8906)	30.26 (0.8679)	31.77 (0.9009)	30.95 (0.8830)	31.20 (0.8884)	30.76 (0.8788)
ProxPnP-DRS	32.27 (0.9109)	32.00 (0.9078)	31.79 (0.9085)	31.83 (0.9038)	32.73 (0.9230)	32.90 (0.9227)	31.78 (0.9074)	31.51 (0.9020)
DenseDiff	32.40 (0.9170)	32.42 (0.9168)	32.78 (0.9228)	31.79 (0.9027)	33.37 (0.9362)	33.49 (0.9381)	32.81 (**0.9188**)	**32.47** (0.9190)
MHC-DEM	**32.73** (**0.9173**)	**32.91** (**0.9186**)	**32.87** (**0.9239**)	**31.96** (**0.9040**)	**33.81** (**0.9397**)	**33.67** (**0.9385**)	**32.89** (0.9181)	**32.47** (0.9192)

where $\epsilon = 10^{-3}$ is a small stabilization constant. Driven by this robust gradient through the Jacobian-free implicit backward pass (Eq. 15), the temporal hyper-connections learn to aggressively accelerate the trajectory toward the true sharp image manifold.

E.2.3 Experimental Results

To evaluate the generalization ability of MHC-DEM on classical linear inverse problems, we conduct image deblurring experiments on the *CBSD68* dataset (center-cropped to 256×256). Following standard protocols, images are degraded using the eight Levin blur kernels with additive Gaussian noise ($\sigma = 0.01$). Table 6 compares our method with recent deterministic restoration frameworks (GS-PnP Hurault et al. (2022), DE-Prox Gilton et al. (2021), ELDER Zou et al. (2023), ProxPnP-DRS Hurault et al. (2024), DenseDiff Zheng et al. (2025)) and diffusion-based approaches (DPS Chung et al. (2023), DiffIR Zhu et al. (2023), DAPS Zhang et al. (2025)).

As shown in Table 6, MHC-DEM achieves the best PSNR on all 8 kernels (tying on Kernel #7) and the best SSIM on 6 out of 8 kernels, demonstrating strong robustness across diverse degradations.

Table 7: Comparison of PSNR (dB) and SSIM for image inpainting on *CBSD68*.

Models	30%	50%	70%
DE-Prox	35.66 (0.9702)	32.11 (0.9354)	28.90 (0.8736)
ELDER	35.53 (0.9696)	31.93 (0.9340)	28.57 (0.8694)
DiffIR	34.84 (0.9593)	30.51 (0.8971)	26.29 (0.8269)
DAPS	34.48 (0.9533)	30.56 (0.8980)	26.52 (0.8307)
ProxPnP-DRS	35.05 (0.9630)	31.42 (0.9125)	27.84 (0.8200)
DenseDiff	35.87 (0.9710)	32.16 (0.9349)	28.99 (0.8738)
MHC-DEM	**36.12** (**0.9726**)	**32.49** (**0.9398**)	**29.10** (**0.8811**)

In particularly challenging cases such as Kernel #1 and Kernel #4, our method improves PSNR by +0.49 dB and +0.44 dB over the strongest baseline (DenseDiff), respectively. These gains indicate that the proposed Sinkhorn-constrained heterogeneous prior interaction effectively preserves both local textures and global structures under severe blur.

The results also reveal a clear difference between deterministic equilibrium solvers and diffusion-based restoration methods. Although diffusion models can generate visually plausible textures, their stochastic sampling process often struggles to strictly enforce data consistency, resulting in noticeably lower PSNR/SSIM values. In contrast, MHC-DEM performs restoration through a deterministic equilibrium optimization process, iteratively refining the solution under explicit data-fidelity constraints. This allows the model to achieve more accurate and stable reconstructions without introducing hallucinated structures. Visual comparisons further supporting these observations are provided in Fig. 13.

E.3 Image Inpainting

To further assess the generalization of MHC-DEM beyond deblurring and black hole imaging, we conduct image inpainting experiments on *CBSD68* with missing ratios of 30%, 50%, and 70%. As reported in Table 7, MHC-DEM consistently achieves the highest PSNR and SSIM under all missing ratios, reaching 29.10 dB and 0.8811 SSIM even with 70% missing observations. These results further demonstrate the effectiveness and generalization of MHC-DEM for diverse image inverse problems.

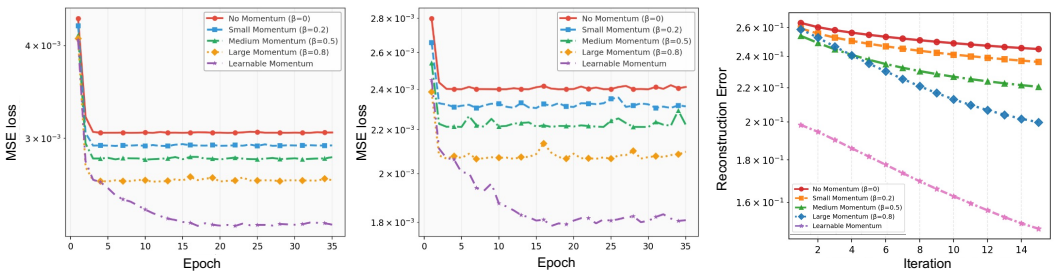
E.4 Model Analysis and Ablation Study

E.4.1 Convergence Property

As illustrated in Fig. 14, we analyze the convergence behavior under different momentum settings from multiple perspectives in the deblurring task. Fig. 14(a) and (b) present the evolution of the MSE loss with respect to DEM training epochs under different momentum coefficients. Both figures show that incorporating momentum consistently accelerates convergence and leads to lower final loss values compared to the non-momentum baseline. Fig. 14(c) further compares the reconstruction error as a function of DEM iterations. For clarity, only the first 15 iterations are shown, highlighting the early-stage convergence behavior. It can be observed that momentum significantly improves convergence speed, allowing the model to reach a lower error within fewer iterations. Among all configurations, the learned momentum parameter achieves the best performance, exhibiting both the fastest convergence rate and the lowest final reconstruction error. These results demonstrate the effectiveness of momentum, especially when adaptively learned, in improving both optimization efficiency and solution quality.

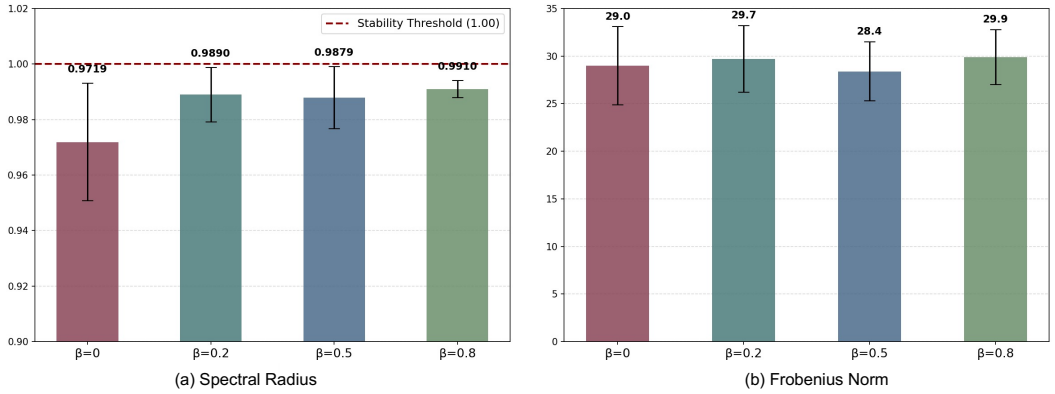
E.4.2 Spectral Radius and Frobenius Norm

We further analyze the spectral properties of the Jacobian in the deblurring task. Since directly computing the full Jacobian matrix J is computationally intractable due to its large scale, we estimate the dominant singular value at the fixed point using the power method. In particular, the spectral



(a) Training loss vs. DEQ epoch (b) Test loss vs. DEQ epoch (c) Error vs. DEQ iteration

Figure 14: Convergence analysis of MHC-DEM under different momentum coefficients: (a) Training loss, (b) Test loss, (c) Reconstruction error.



(a) Spectral Radius (b) Frobenius Norm

Figure 15: Spectral analysis under different momentum coefficients: (a) spectral radius, (b) Frobenius norm.

radius is defined as $\rho(J) = \max_i |\lambda_i|$, i.e., the maximum absolute value among the eigenvalues of the Jacobian matrix.

According to Lemma 2, the condition $\rho(J) < 1$ guarantees local convergence stability. Fig. 15 presents the mean values and variation ranges of the spectral radius and the Frobenius norm under different momentum coefficients. It can be observed that the spectral radius of our method remains close to 1 across different momentum settings. Notably, when using the learned momentum (which is close to $\beta = 0.8$), the spectral radius is strictly smaller than 1, satisfying Lemma 2 and thus ensuring convergence.

We further evaluate the Frobenius norm of the Jacobian, defined as $\|J\|_F = \sqrt{\sum_i \sum_j |J_{ij}|^2} = \sqrt{\text{trace}(J^\top J)} = \sqrt{\sum_i \sigma_i^2}$. Based on the Hutchinson estimator, we have

$$\mathbb{E}[\|Jv\|^2] = \mathbb{E}[v^\top J^\top J v] = \frac{1}{n}\text{trace}(J^\top J) = \frac{1}{n}\|J\|_F^2,$$

for a random probe vector v. Therefore, $\|J\|_F^2 = n\,\mathbb{E}[\|Jv\|^2]$, and $\|J\|_F = \sqrt{n\,\mathbb{E}[\|Jv\|^2]}$.

Empirically, under different momentum configurations, the Frobenius norm remains within a similar range (approximately 28–30), indicating that the overall structural stability of the iterative mapping is preserved. Moreover, no clear correlation is observed between the Frobenius norm and the magnitude of the momentum coefficient, suggesting that momentum primarily influences the convergence direction and rate rather than the overall sensitivity of the mapping.

E.4.3 Effect of Residual Scaling Coefficient

To evaluate the impact of the residual scaling coefficient α on restoration performance, we conduct an analysis under the degradation setting of a 15×15 blur kernel with a noise level of $\sigma = 0.01$.

Table 8: Evolution of fusion weights (%) during training.

Epoch	Softmax Fusion			Sinkhorn Fusion		
	DnCNN	DRUNet	Restormer	DnCNN	DRUNet	Restormer
1	33.5	33.2	33.3	33.4	33.3	33.3
5	58.2	28.4	13.4	41.5	33.2	25.3
20	84.1	12.3	3.6	48.6	31.8	19.6
50	94.5	4.2	1.3	50.8	31.0	18.2
100	97.2	2.4	0.4	51.2	30.9	17.9

By varying this hyperparameter, we observe PSNR values of 32.76 dB, 32.87 dB, and 32.63 dB for $\alpha \in \{0.05, 0.1, 0.3\}$, respectively. These results demonstrate that our framework maintains a strong degree of robustness to the choice of α within a certain range. However, performance visibly degrades when the coefficient becomes excessively large. This empirical observation strictly aligns with our theoretical guarantees, which dictate that the update magnitude must be properly bounded to preserve the contractivity required for equilibrium convergence. Based on these findings, we set $\alpha = 0.1$ as the default optimal configuration for all practical implementations.

E.4.4 Analysis of Multi-prior Fusion and Prior Collapse

Fusion Weight Evolution and Robustness We first investigate the behavior of the proposed multi-prior interaction mechanism by analyzing the evolution of fusion weights during training. Experiments are conducted on the image deblurring task with the Levin kernel #1 and noise level $\sigma = 0.01$. All fusion weights are initialized uniformly, and their evolution is tracked throughout the optimization process. As reported in Table 8, Softmax fusion exhibits a rapid concentration of weights on a single prior. For the three-prior configuration (DnCNN, DRUNet, and Restormer), the maximum weight increases from approximately 33.5% at initialization to 97.2% after 100 epochs, while the contributions of the other two priors decrease to 2.4% and 0.4%, respectively. In contrast, Sinkhorn-constrained fusion maintains contributions from all three priors throughout training. Its final weight distribution is 51.2%, 30.9%, and 17.9% for DnCNN, DRUNet, and Restormer, respectively.

The observed behavior is further examined under different prior configurations. With two priors (DnCNN and DRUNet), the Sinkhorn-based fusion produces weights of 62.1% and 37.9%, respectively. When Restormer is introduced, the distribution becomes more balanced, with the three priors contributing 51.2%, 30.9%, and 17.9%. These results suggest that the proposed interaction mechanism can preserve non-degenerate contributions from heterogeneous priors as the prior bank expands. To evaluate the robustness of this behavior, we further investigate the maximum fusion weight across different datasets, random seeds, and prior configurations. As summarized in Table 9, Softmax fusion consistently exhibits strong prior dominance, with maximum weights reaching 96.5%–99.1% in the evaluated settings. In contrast, Sinkhorn fusion maintains lower maximum weights, ranging from 49.1% to 62.1%. These observations indicate that the reduced prior concentration is not limited to a single training configuration.

The above results support the role of the doubly stochastic interaction constraint in maintaining multi-prior participation. Unlike unconstrained Softmax fusion, which may amplify an initial preference toward a particular prior during recursive equilibrium updates, the proposed formulation imposes a structured constraint on prior interactions. This provides a mechanism for mitigating prior collapse while retaining the ability to assign different importance to heterogeneous restoration operators.

Comparison with Alternative Anti-collapse Strategies We next compare the proposed Sinkhorn-constrained fusion with alternative strategies for mitigating prior collapse. Specifically, we consider four fusion schemes: (1) standard Softmax fusion, (2) Softmax fusion with entropy regularization, (3) Softmax fusion with load-balancing regularization, and (4) the proposed Sinkhorn-constrained fusion.

Table 9: Maximum fusion weight across datasets and random seeds.

Dataset	Seed	Softmax	Sinkhorn	Priors
CBSD68	42	97.2%	51.2%	3
CBSD68	40	96.9%	50.8%	3
ImageNet	42	98.1%	53.7%	3
CelebA	42	96.5%	49.1%	3
CBSD68	42	99.1%	62.1%	2

The results are reported in Table 10. Entropy and load-balancing regularization reduce prior concentration compared with standard Softmax fusion, improving the reconstruction performance from 30.96 dB to 31.79 dB and 31.84 dB, respectively. The proposed Sinkhorn fusion achieves the highest PSNR of 32.91 dB in this comparison, together with a maximum fusion weight of 0.51.

These results highlight the distinction between auxiliary regularization and structural constraints on prior interactions. Entropy and load-balancing regularization encourage diversified weights through additional optimization objectives. In contrast, Sinkhorn fusion incorporates a doubly stochastic interaction structure directly into the fusion mechanism. This design is particularly relevant to MHC-DEM, where the fusion operation is repeatedly embedded within the equilibrium solver.

Although the results demonstrate improved reconstruction quality and reduced prior concentration, the observed gains should not be attributed exclusively to the anti-collapse effect. The Sinkhorn formulation also changes the structure of prior interactions and their propagation through the equilibrium operator. Therefore, the comparison supports the combined effect of structured fusion and prior diversification, rather than establishing that the performance improvement is solely caused by weight balancing.

Table 10: Ablation study of different prior fusion strategies.

Method	PSNR	Entropy	Max weight
Softmax	30.96±0.16	0.15	0.97
Entropy	31.79±0.19	0.63	0.62
Load balance	31.84±0.21	0.90	0.40
Sinkhorn	32.91±0.32	0.82	0.51

E.4.5 Ablation on Nested Inner Refinement

The nested inner refinement is designed to decouple local spatial reconstruction from the global equilibrium dynamics. Specifically, the outer equilibrium iteration governs the evolution of the global latent state, while the inner refinement performs additional spatial reconstruction updates using non-shared parameters. This design allows the model to progressively refine the recovered image within each equilibrium step.

We evaluate the effect of the nested refinement depth T by varying the number of inner layers while keeping all other components unchanged. The results are summarized in Table 11. As T increases from 1 to 15, the PSNR improves from 31.88 dB to 32.91 dB, accompanied by consistent improvements in SSIM and MSE. This result suggests that additional inner refinement steps contribute to improved reconstruction quality rather than merely increasing the computational cost.

When T is further increased from 15 to 20, the PSNR slightly decreases from 32.91 dB to 32.86 dB, while the inference time increases from 591 to 685. This indicates that deeper inner refinement does not necessarily yield continuous performance improvements. In the evaluated setting, $T = 15$ provides a favorable balance between reconstruction quality and computational cost. The observed performance saturation is also consistent with the practical trade-off suggested by the stability analysis in Theorem 2, although the theorem does not by itself establish that $T = 15$ is the optimal refinement depth.

Table 11: Ablation study of nested inner refinement depth T.

Nested Layers	PSNR	SSIM	MSE($\times 10^{-4}$)	Time
1	31.88	0.9105	6.486	263
5	32.28	0.9124	5.916	340
10	32.56	0.9138	5.546	458
15	32.91	0.9186	5.117	591
20	32.86	0.9157	5.176	685

E.4.6 Momentum Parameter Analysis

We further investigate the behavior of the learned momentum parameter β_k during training. The learned values stabilize within the range of 0.8–0.9, indicating that the solver retains information from previous equilibrium states during iterative updates. This inertial mechanism provides a means of incorporating historical state information into the current refinement process.

To examine the relationship between the learned momentum and the stability behavior, we compare the learned parameters with the spectral radius and Frobenius norm measurements shown in Fig. 15. The observed parameters remain within the empirically stable region identified in our analysis. This observation is consistent with the theoretical stability conditions derived for the proposed solver, while the formal convergence guarantee follows from the assumptions and conditions established in the corresponding theoretical analysis.

E.4.7 Inference Efficiency

We evaluate the inference time under different momentum settings and nested refinement depths, as reported in Table 12. Since the denoisers are frozen during inference, the memory consumption remains approximately constant across the evaluated configurations at around 42 GB. Moreover, the results show that incorporating momentum reduces the measured inference time compared with the corresponding no-momentum and fixed-momentum settings.

For a fixed momentum configuration, the inference time increases with the number of nested refinement layers. This trend is expected because deeper refinement introduces additional iterative computations within each equilibrium step. In particular, the learned-momentum configuration requires 340, 458, 591, and 685 ms for $T = 5$, 10, 15, and 20, respectively. The measured runtime therefore reflects a trade-off between refinement depth and computational cost.

Compared with the no-momentum setting, the learned-momentum configuration reduces the measured inference time for all evaluated nested depths. These results suggest that the inertial update can improve computational efficiency in the evaluated implementation. However, the measured runtime should be interpreted together with the hardware, batch size, and implementation details to ensure a fair comparison across configurations.

Table 12: Inference time (ms) under different momentum settings and nested refinement depths.

Nested Layers	No Momentum	Medium ($\beta = 0.5$)	Learned Momentum
5	542	381	340
10	663	490	458
15	800	625	591
20	891	715	685