PROJECTED ENERGY MATCHING FOR GENERATIVE 3D PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transport-based generative models, which learn a time-dependent vector field that moves noise to data, have become a dominant paradigm. However, these models typically do not explicitly encode the data distribution. Energy-based models (EBMs) instead represent the data distribution explicitly through a scalar energy landscape, which *Energy Matching* learns by combining transport learning with contrastive refinement. Its transport objective, however, fits energy gradients to stochastic targets, whose variance degrades the training signal at scale. We introduce *Projected Energy Matching*, which learns this landscape through a more stable route: we first train a time-independent transport teacher, then freeze it and fit the negative energy gradient to its predicted velocities. This projection replaces noisy transport targets with deterministic supervision, while contrastive refinement shapes the landscape near the data manifold. On CIFAR-10, gradient-noise analysis reveals a cleaner training signal, accompanied by faster convergence than Energy Matching at matched, teacher-free training budgets. In the latent space of CT volumes, our method enables 3D CT generation and achieves better FID scores than flow models. The learned scalar potential serves as a zero-shot prior for the ill-posed inverse problem of sparse-view cone-beam CT reconstruction. By making explicit energy landscapes practical at volumetric scale, this work opens a path to wider adoption of energy-based formulations, bringing their flexibility to high-dimensional generation and inverse problems.

1 INTRODUCTION

Generative modeling for high-dimensional 3D data faces a trade-off between *sampling efficiency* and *explicit energy representation*. While Transport and Diffusion Models (Ho et al., 2020; Albergo and Vanden-Eijnden, 2023; Liu et al., 2023; Lipman et al., 2023; Tong et al., 2024) scale remarkably well, they typically model the distribution implicitly through their time-dependent dynamics. They do not directly expose an objective landscape $\phi(\mathbf{x})$ for comparing candidate states, combining the learned prior with other objectives, and out-of-distribution scoring. Energy-Based Models (EBMs) (Hinton, 2002; LeCun et al., 2007; Du and Mordatch, 2019) inherently provide this scalar potential but are notoriously difficult to train Song and Kingma (2021). Standard Markov chain Monte Carlo (MCMC)-based training (Song and Kingma, 2021; Song and Ermon, 2019; Schröder et al., 2023) suffers from prohibitive computational costs, training instability, and mode collapse. Consequently, practitioners often bypass MCMC entirely by employing heavily parameterized time-dependent ensembles, or cooperative architectures which train an auxiliary generator alongside the energy model for rapid sample proposal (Gao et al., 2021; Cui and Han, 2024; Guo et al., 2023; Zhang et al., 2024; Yoon et al., 2024). However, relying on auxiliary networks sacrifices the elegance of a single, unified, time-independent energy model.

Recent methods, including Energy Matching (Balcerak et al., 2025; 2026), Equilibrium Matching (Wang and Du, 2025), and Beckmann Transport Models (Lee et al., 2026), use simulation-free supervision to learn time-independent generative dynamics. In particular, Energy Matching learns a scalar potential whose negative gradient approximates a transport velocity field ($-\nabla\phi(\mathbf{x}) \approx v(\mathbf{x})$) away from the data manifold. The resulting transport guides samples toward the data distribution, providing an effective burn-in for subsequent MCMC sampling. The transport regime, however, is trained from scratch with a double-backward pass ($\nabla_\theta \nabla_{\mathbf{x}} \phi_\theta$) guided by a stochastic interpolant target. We observe that this target need not be stochastic. A velocity model, trained cheaply with first-order

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.

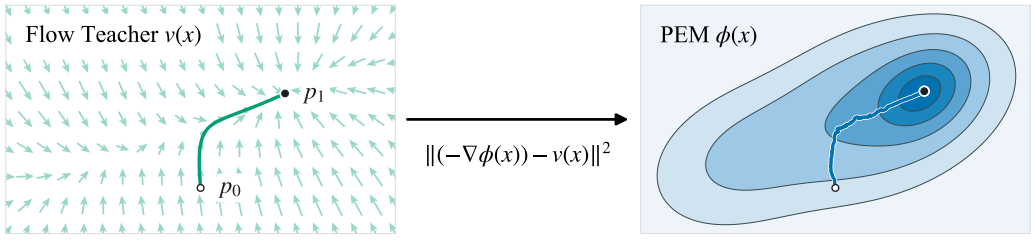


Figure 1: **Projected Energy Matching.** *Left:* a pre-trained, time-independent flow teacher $v(\mathbf{x})$ (arrows) transports a sample from a noise p_0 to a data distribution p_1 along a deterministic ODE path. *Right:* regressing a scalar potential onto that field under $\|v(\mathbf{x}) - (-\nabla\phi_\theta(\mathbf{x}))\|^2$ projects it onto the conservative subspace of vector fields, yielding an explicit energy landscape whose gradient reproduces the teacher dynamics (sampling path, same start and endpoint).

optimization, already carries the transport map. We then freeze this model and regress its predicted velocities onto the negative gradient of a parameterized scalar potential, $-\nabla\phi_\theta(\mathbf{x})$, replacing noisy regression targets with deterministic supervision.

Minimizing the ℓ_2 objective between $-\nabla\phi_\theta$ and the frozen teacher field ($v(x)$) projects the field onto the conservative subspace of vector fields as shown in Fig. 1. We term the resulting framework **Projected Energy Matching (PEM)**, which operates in three stages: (1) training a teacher model with an unconstrained flow objective to establish the transport map, (2) learning a single time-independent scalar potential by projecting the frozen teacher field onto the conservative subspace, and (3) refining the potential landscape using a contrastive loss. Both energy parameterizations of the score (Salimans and Ho, 2022; Du et al., 2023) and the conservativeness of learned fields (Chao et al., 2023) have been studied for time-conditioned diffusion. Teacher-supervised projection has also been used to stabilize the training of time-dependent diffusion energies (Thornton et al., 2025), while (Praneeth et al., 2026) use a learned conservative projection to amplify a field's rotational component at inference time. In PEM, an autonomous velocity teacher supervises a time-independent transport potential, with supervision retained during contrastive refinement. The student therefore inherits the integrable part of the teacher dynamics and discards the residual.

First, we investigate whether this projected target is a better optimization problem than learning the potential from scratch. To answer it, we use gradient noise scale estimation (McCandlish et al., 2018) and measure the signal-to-noise structure of the potential-model gradient under both regimes on CIFAR-10, holding training hyper parameters fixed across experiments. Projection against a frozen teacher yields a stronger gradient signal than training from scratch, indicating that the projected target is better conditioned. This cleaner signal is the mechanism behind the compute-friendly model at 3D.

Next, we train the potential on 25,692 computed tomography volumes (CT-RATE dataset) (Hamamci et al., 2026) in the latent space of a 3D variational autoencoder (MAISI) (Guo et al., 2025), which compresses full 3D CT volumes enough to make the double-backward pass tractable at useful spatial resolution, and evaluate the learned energy in two modes. The first is unconditional volumetric generation. The second evaluation uses the explicit scalar in regularizing an ill-posed inverse problem: sparse-view CBCT reconstruction from only 20 of 600 projections using Model-Based Iterative Reconstruction (MBIR) (Venkatakrishnan et al., 2013), in which the same unconditionally trained potential enters a measurement-consistency loop as a 3D prior.

Our main contributions are summarized as follows:

- **Projection improves the training signal:** We train a scalar potential by projecting the velocity field of a frozen, time-independent flow teacher onto the conservative subspace, replacing the stochastic transport targets of Energy Matching with deterministic supervision. A gradient-noise analysis on CIFAR-10 shows that the projected objective yields a stronger training signal than learning the potential from scratch, with improved convergence.
- **Energy-based priors at volumetric scale:** We train our PEM model on 25,692 CT-RATE volumes ($128\times128\times64$) in the latent space of the MAISI autoencoder, where a deterministic

training objective makes the training feasible with our computation budget, and the resulting potential serves as a zero-shot prior for out-of-distribution detection and sparse-view cone-beam CT reconstruction.

2 METHODOLOGY

Let $\mathbf{x} \in \mathbb{R}^D$ denote a state in the high-dimensional data space. We propose a framework for training an Energy Matching model on high-dimensional data by leveraging the transport efficiency of flow models. Our approach unfolds in three phases: (1) training a flow teacher to learn global transport; (2) projecting this teacher velocity field onto the conservative subspace of a scalar potential; and (3) refining the energy landscape using an Energy Matching contrastive loss.

2.1 PHASE 1: FLOW TEACHER

To provide a stable target for the projection, we first learn a transport map between a Gaussian prior p_0 and the data distribution p_1. We employ a transport objective, which learns a velocity field v_{teacher} that creates straight paths between distributions following (Balcerak et al., 2025).

To minimize the curvature of the transport trajectories and resolve path crossings, we utilize minibatch optimal transport (OT) couplings (Tong et al., 2024). Instead of random pairings, we compute the discrete OT map between batches of the Gaussian prior p_0 and the data distribution p_1, storing these distance-minimizing pairs $(\mathbf{x}_0, \mathbf{x}_1)$ in a memory bank $\mathcal{M}$. We then train our flow teacher to match these straight conditional paths:

$$\mathcal{L}_{\text{OT}} = \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{x}_1)\sim\mathcal{M}} \left[\| v_{\text{teacher}}(\mathbf{x}_t) - (\mathbf{x}_1 - \mathbf{x}_0) \|^2 \right], \tag{1}$$

where $\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$. By enforcing OT pairings, we drastically reduce trajectory intersections. This ensures the resulting marginal velocity field approaches a time-independent (autonomous) state ($\nabla_t v \approx 0$), providing a stable, low-variance target for the subsequent projection phase.

2.2 PHASE 2: PROJECTION

We now recover a scalar energy $\phi_\theta(\mathbf{x})$ whose gradient reproduces the teacher dynamics, $-\nabla_{\mathbf{x}}\phi_\theta \propto v_{\text{teacher}}$. The model is parametrized as a potential and differentiated with respect to its input, so the induced velocity field

$$v_\theta(\mathbf{x}) = -N\nabla_{\mathbf{x}}\phi_\theta(\mathbf{x}) \tag{2}$$

is conservative by construction and N represents the spatial dimension factor. The teacher, in contrast, was trained without this restriction. Minimizing the squared error between the two therefore does not fit the teacher pointwise but it returns the closest conservative field in L^2, weighted by the density of interpolants $\mathbf{x}_t$ along the transport paths. This is why we call the method *projected*, and why the OT couplings of Phase 1 are crucial, as the more closely the teacher approximates straight paths, the more of its transport structure ϕ_θ retains.

The projection objective is a mean squared error of the scalar potential against the frozen teacher:

$$\mathcal{L}_{\text{proj}} = \mathbb{E}_{t,\mathbf{x}_t} \left[\| -N\nabla_{\mathbf{x}}\phi_\theta(\mathbf{x}_t) - \text{sg}[v_{\text{teacher}}(\mathbf{x}_t)] \|^2 \right], \tag{3}$$

where $\text{sg}[\cdot]$ denotes the stop gradient operator.

This phase is efficient despite the double-backward pass because its target is a deterministic function of x_t, not a single-sample estimate of $\mathbb{E}\left[(x_1 - x_0) \mid x_t\right]$. Both losses share a minimizer, but the projected loss removes the dominant variance term of learning transport from scratch. We quantify the resulting difference in gradient noise structure in Sec. 3, and it is what makes the total step count affordable for us at 3D scale. The complete procedure is summarized in the Appendix 2 Algorithm 2.

2.3 PHASE 3: ENERGY MATCHING REFINEMENT

The projection supplies the global optimal transport structure, but the resulting potential must be fine-tuned near the data manifold. In this final phase we refine ϕ_θ using the Energy Matching framework (Balcerak et al., 2025), retaining the frozen teacher as the transport supervisor. This

Algorithm 1 Projected Energy Matching with Teacher Supervision.

Require: Projected Energy Matching potential ϕ_θ (from Phase 2), frozen teacher flow v_ψ (from Phase 1), contrastive weight $\lambda_{\text{contrastive}}$, dimension factor $N = D$
Require: Gradient accumulation steps K, Langevin steps M_{Langevin}, step size Δt
Require: Standard Gaussian noise vectors $\eta \sim \mathcal{N}(0, I)$

for iteration $n = 0, 1, \ldots$ **do**
  $g \leftarrow 0$ ▷ Initialize gradient accumulator
  $\mathcal{X}_{\text{cached}} \leftarrow \emptyset$ ▷ Clear negative sample cache
  for micro-batch $k = 1$ to K **do**
    Sample mini-batch $\{x^{(j)}_{\text{data}}\}_{j=1}^{B} \sim \mathcal{D}$ and $\{x^{(j)}_{\text{noise}}\}_{j=1}^{B} \sim \mathcal{N}(0, I)$
    Sample independent time steps $\{t^{(j)}\}_{j=1}^{B} \sim \mathcal{U}[0, 1]$
    $x_t^{(j)} \leftarrow t^{(j)} x^{(j)}_{\text{data}} + (1 - t^{(j)}) x^{(j)}_{\text{noise}}$ ▷ Interpolate along optimal transport paths
    $v^{(j)}_{\text{target}} \leftarrow \text{sg}[v_\psi(x_t^{(j)})]$ ▷ Frozen teacher velocity, single forward pass, no gradient
    $v^{(j)}_{\text{base}} \leftarrow -N \nabla_x \phi_\theta(x_t^{(j)})$ ▷ Conservative flow
    $\mathcal{L}_{\text{proj}} \leftarrow \frac{1}{B} \sum_{j=1}^{B} \|v^{(j)}_{\text{base}} - v^{(j)}_{\text{target}}\|_2^2$
    if $\mathcal{X}_{\text{cached}} = \emptyset$ **then** ▷ Perform MCMC only on first micro-batch
      Initialize negative samples $x^{(0)}_{\text{neg}}$ from noise distribution
      for $m = 0, 1, \ldots, M_{\text{Langevin}} - 1$ **do**
        Sample independent noise $\eta^{(m)} \sim \mathcal{N}(0, I)$
        $x^{(m+1)}_{\text{neg}} \leftarrow x^{(m)}_{\text{neg}} - \Delta t \times N \nabla_x \phi_\theta(x^{(m)}_{\text{neg}}) + \sqrt{2\Delta t}\, \eta^{(m)}$
      end for
      $\mathcal{X}_{\text{cached}} \leftarrow \text{sg}[x^{(M_{\text{Langevin}})}_{\text{neg}}]$ ▷ Detach and cache negative pool
    end if
    $x_{\text{neg}} \leftarrow \mathcal{X}_{\text{cached}}$ ▷ Retrieve cached negatives
    $\mathcal{L}_{\text{contrastive}} \leftarrow \frac{1}{B} \sum_{j=1}^{B} \left(\phi_\theta(x^{(j)}_{\text{data}}) - \phi_\theta(x^{(j)}_{\text{neg}}) \right)$ ▷ Raw potential values, no N
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{proj}} + \lambda_{\text{contrastive}} \mathcal{L}_{\text{contrastive}}$
    $g \leftarrow g + \frac{1}{K} \nabla_\theta \mathcal{L}_{\text{total}}$ ▷ Accumulate gradients across micro-batches
  end for
  Update $\theta \leftarrow \text{Optimizer}(\theta, g)$ ▷ Apply synchronized gradient update
end for
Return Refined potential ϕ_θ

joint objective combines the projection loss of Phase 2 ($\mathcal{L}_{\text{proj}}$), which maintains the transport, and a contrastive loss ($\mathcal{L}_{\text{contrastive}}$) driven by Langevin dynamics close to the data distribution. Keeping the teacher in the loop means the transport term remains a deterministic.

To adapt this framework for the massive memory footprint of high-dimensional 3D voxel grids, we introduce one efficiency modification to the standard Energy Matching pipeline. To manage VRAM limitations, 3D training requires gradient accumulation, but running deep Langevin chains for every micro-batch is computationally prohibitive. We solve this via *Negative Caching*: we execute the expensive MCMC sampling only during the first micro-batch, cache these generated negative samples, and reuse them as a frozen negative pool for the remaining accumulation steps.

The two phases share a single objective, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{proj}} + \lambda_{\text{contrastive}} \mathcal{L}_{\text{contrastive}}$.

Transport ($\mathcal{L}_{\text{proj}}$): We ensure the energy landscape retains the macro-level "funnel" shape required for efficient sampling from noise by continuing to regress the conservative field onto the frozen teacher as described in Eq. 3.

Local Refinement ($\mathcal{L}_{\text{contrastive}}$): Concurrently, near the data manifold, we transition to a diffusive regime, where we add scheduled noise, to capture high-frequency local modes. We carve out localized basins of attraction using a contrastive loss:

$$\mathcal{L}_{\text{contrastive}} = \mathbb{E}_{\mathbf{x}^+ \sim \mathcal{D}}[\phi_\theta(\mathbf{x}^+)] - \mathbb{E}_{\mathbf{x}^- \sim p_\theta}[\phi_\theta(\mathbf{x}^-)] \tag{4}$$

where $p_\theta(\mathbf{x}) \propto \exp(-\phi_\theta(\mathbf{x}))$ represents the model's currently learned distribution, and negatives $\mathbf{x}^-$ are drawn from it via Langevin dynamics on ϕ_θ.

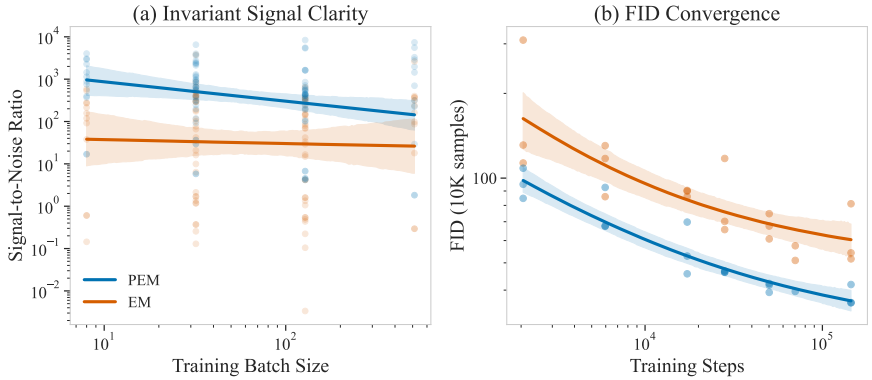


Figure 2: **Training signal and convergence on CIFAR-10. (a)** PEM exhibits higher gradient signal-to-noise ratios across the evaluated batch sizes. **(b)** FID during transport training shows that PEM converges faster than EM. Solid lines show posterior median fitted curves, and shaded regions indicate 95% credible intervals for the fitted mean.

3 Experiments

We first examine whether PEM improves the training signal and convergence of energy models on CIFAR-10. We then scale the method to 3D CT-RATE volumes, where the resulting prior is evaluated for generation, detecting distribution shifts, and sparse-view CBCT reconstruction. These applications test whether the same learned distribution can support different inference tasks without retraining.

3.1 Training Signal Quality & Dynamics in 2D

Each PEM run is paired with an Energy Matching (EM) run trained from scratch under identical architecture, initialisation, data ordering, optimiser, schedule, and seed. The only difference is the regression target: frozen teacher velocities for PEM and stochastic interpolant targets for EM. The experiment covers batch sizes $B \in \{8, 32, 128, 512\}$ and multiple learning rates, with three seeds at selected settings, giving 20 matched run pairs.

Gradient-noise measurements (McCandlish et al., 2018) show a higher signal-to-noise ratio for PEM across the evaluated batch sizes (Fig. 2a). The median squared norm of the mean parameter gradient is approximately $43\times$ larger than for EM. Gradient noise also increases, but by a smaller factor, so the average training signal becomes stronger relative to its variability.

This improvement is accompanied by better convergence behavior (Fig. 2b). After processing approximately 450,000 training examples per potential model, PEM achieves consistently lower FID as training progresses, suggesting faster convergence or better FID with the same compute budget. These results encourage us to learn a 3D generative prior in our next experiment under a constrained academic training budget, which otherwise seemed prohibitive. Statistical analysis details are provided in Appendix F.

3.2 Evaluating the 3D Energy Prior

To assess PEM at volumetric scale, we train an unconditional energy prior on CT-RATE volumes in the compressed latent space of the MAISI autoencoder. This representation makes training the 3D potential tractable. Dataset and preprocessing details are provided in Appendix B.2.

The resulting prior achieves generative quality competitive with time-dependent flow models (Table 1). Using the same U-Net backbone, PEM reaches an FID of 40.87, compared with 42.94 for OT Flow Matching and 42.39 for Rectified Flow. Because PEM requires no time conditioning, the potential is compatible with backbones that lack timestep embeddings. With MedNeXt (Roy et al., 2023), PEM reaches an FID of 35.87 with a comparable parameter count. FID is used to assess sample quality

Table 1: **Unconditional 3D CT generation.** FID evaluation uses 1,000 held-out volumes of size $128 \times 128 \times 64$, with a sampling budget of 1,000 function evaluations. The first three methods use the same U-Net backbone. The final row evaluates PEM with MedNeXt.

Method	Params	FID ↓
OT Flow Matching (Tong et al., 2024)	10.3M	42.94
Rectified Flow (Liu et al., 2023)	10.3M	42.39
PEM (Ours)	10.7M	40.87
PEM, MedNeXt backbone (Ours)	10.2M	**35.87**

(Woodland et al., 2024), and Figure 3 shows representative generated volumes. Ablations of the training phases are reported in Appendix E, Table 3.

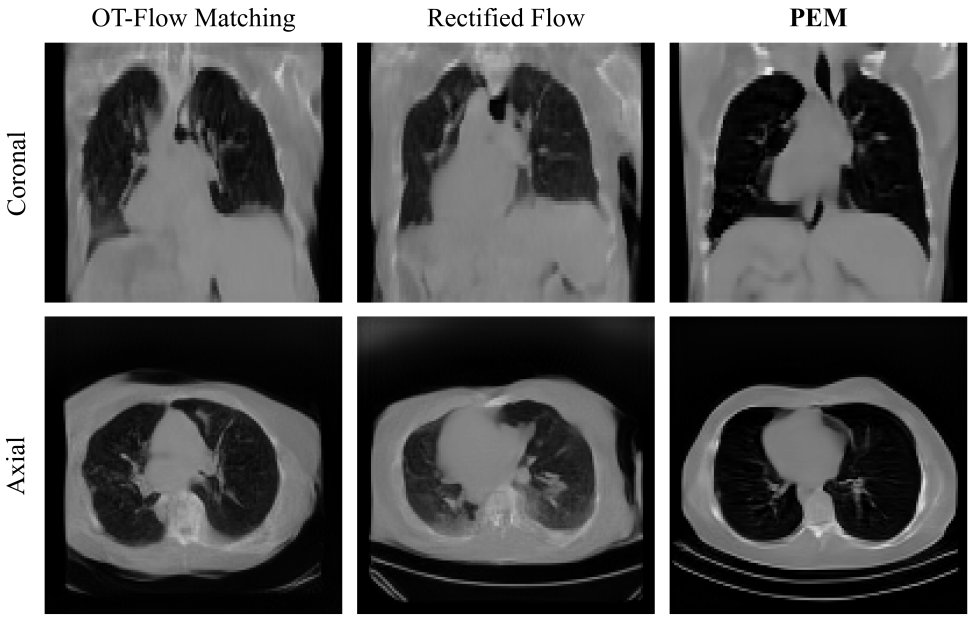


Figure 3: **Unconditionally generated CT volumes.** Coronal (top) and axial (bottom) slices from volumes generated by OT Flow Matching, Rectified Flow, and PEM using 1,000 function evaluations.

3.3 Detecting Distribution Shifts

A learned prior can also help assess how supplied volumes differ from its training distribution. Here, the energy is used to distinguish familiar CT volumes from inputs affected by different acquisition conditions or reconstruction artifacts, without training a separate detector.

Two settings test this capability across projection counts (Fig. 5). The first separates held-out CT-RATE volumes from real CBCT acquisitions. The second compares sparse-view reconstructions obtained with the Feldkamp-Davis-Kress (FDK) algorithm (Feldkamp et al., 1984) against their full-view, 600-projection FDK references. In this second setting, reducing the number of projections introduces artifacts while keeping the underlying anatomy fixed.

Each volume is scored as $s(\mathbf{x}) = \phi_\theta(\mathcal{E}(\mathbf{x}))$, where $\mathcal{E}$ is the MAISI encoder. After encoding, this requires a single evaluation of the potential. Comparisons use an OT Flow Matching trajectory score adapted from DiffPath (Heng et al., 2024), together with the L_2 norm and total variation (TV) of the volume.

At the lowest view counts, severe artifacts make separation straightforward for several scores. As additional projections improve reconstruction quality, the differences become less obvious. PEM

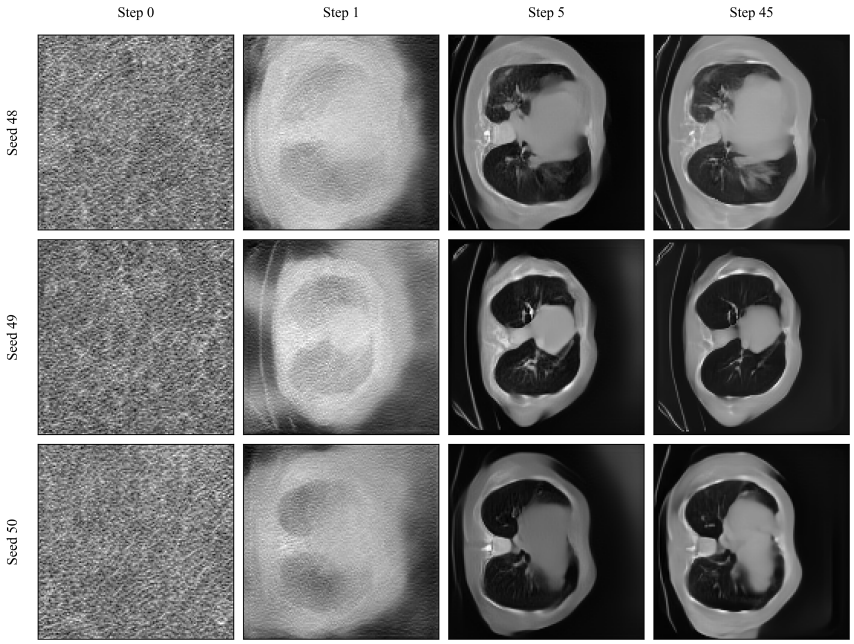


Figure 4: **Extended Manifold Exploration via Explicit Energy.** Unlike standard flow models that integrate forward from noise ($t = 0$) and terminate rigidly at the data manifold ($t = 1$), our PEM permits unbounded Langevin sampling. Starting from pure noise (Langevin Step 0), the global energy funnel rapidly projects the sample toward valid medical anatomy (Step 5). Continuing to sample extensively (Step 45) largely avoids the severe artifact accumulation typically seen when forcing time-dependent models past their endpoint.

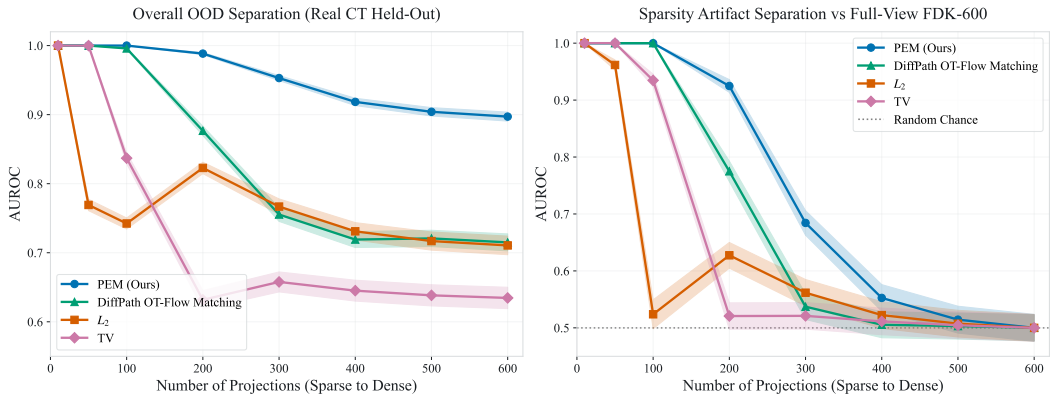


Figure 5: **Scoring acquisition differences and reconstruction artifacts.** We test whether the generative prior detects distribution shifts without additional training. AUROC across projection counts for separating held-out CT-RATE volumes from CBCT inputs (left), and sparse-view FDK reconstructions from their 600-projection FDK references (right). PEM scores each volume with a single evaluation of the learned potential after encoding. Higher values indicate better separation. At low projection counts, severe artifacts make separation straightforward for several scores. Increasing projection counts reduces artifacts while keeping anatomy fixed, making the second task harder. PEM retains AUROC above 0.9 at 200 projections in this task.

retains stronger separation in this regime: in the artifact-detection task, its AUROC remains above 0.9 at 200 of 600 projections, where the alternative scores are already less discriminative.

Table 2: **Sparse-View CBCT Reconstruction (20 of 600 projections).** Mean ± std over 10 held-out volumes. All iterative solvers share the forward operator, initialisation and step budget; only the regulariser differs. The energy prior admits a single conservative objective (Eq. 5) we show MAP estimation here.

Method	Prior	PSNR ↑	SSIM ↑	LPIPS ↓
FBP	–	20.91 ± 0.67	0.4455 ± 0.0122	1.0099 ± 0.0587
Iterative TV	–	26.75 ± 0.87	0.7410 ± 0.0380	0.3375 ± 0.0369
OT-Flow Matching MAP	Flow	31.92 ± 1.22	0.8881 ± 0.0303	0.0960 ± 0.0244
Ours, PEM variational MAP	Energy	**32.55 ± 1.30**	**0.9161 ± 0.0252**	**0.0621 ± 0.0154**

3.4 Manifold Exploration

The learned prior supports exploration both with and without measurements. In the unconstrained case, Langevin dynamics can continue to evolve a generated volume under the same time-independent potential. Fig. 4 illustrates the resulting anatomical variations and reports sample quality over extended sampling horizons (Appendix C Fig. 8).

Adding measurements turns this exploration into a reconstruction task. Given measurements $\mathbf{y}$ and a forward operator $\mathcal{A}$, we combine measurement consistency with the learned prior:

$$\mathbf{z}^\star = \arg\min_{\mathbf{z}} \ \|\mathcal{A}\mathcal{D}(\mathbf{z}) - \mathbf{y}\|^2 + \lambda\,\phi_\theta(\mathbf{z}), \tag{5}$$

where $\mathcal{D}$ is the MAISI decoder and λ weights the prior. For sparse-view CBCT, we retain only 20 of 600 projections and optimise Eq. 5 in latent space for 2500 steps, without retraining the prior or introducing measurement-specific conditioning. For the MAP estimate, we minimise Eq. 5 with Adam for 2500 steps, initialised from the encoded FBP reconstruction. For posterior sampling, we

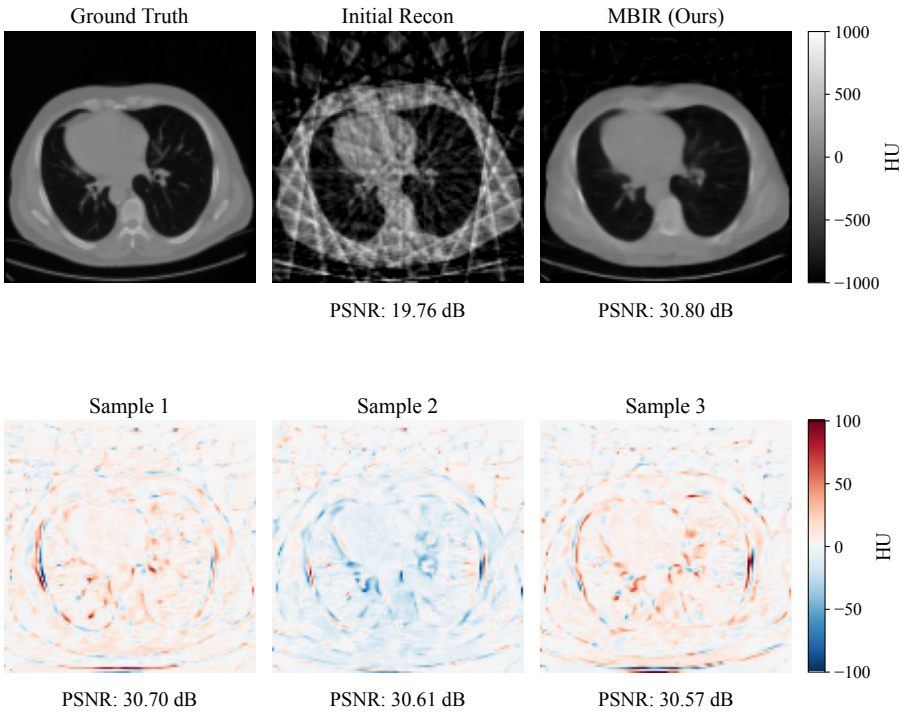


Figure 6: **Diversity under sparse measurements.** Top: ground truth, sparse-view initialisation, and ensemble-mean reconstruction. Bottom: three reconstructions obtained through Langevin sampling with approximate null-space repulsion, shown as differences from their ensemble mean. Bottom-row colours indicate intensity differences in Hounsfield units, and PSNR values refer to the underlying individual reconstructions.

integrate Langevin dynamics on the same objective (Appendix G 7). Reconstruction results are shown in Appendix C Fig. 7 and Table 2, with solver details in Appendix G.

Sparse measurements still leave room for multiple reconstructions. To explore this ambiguity, we draw three posterior samples jointly with Langevin dynamics (Appendix G Eq. 7 and add a repulsion term H. We apply this repulsion in the approximate null space, encouraging variation in directions weakly constrained by the measurements (see Appendix H for technical details). The bottom row of Figure 6 shows the resulting differences from the ensemble mean. Across 10 volumes and 3 seeds, repulsion improves the ensemble-mean PSNR from 30.02 to 30.22 dB compared with the same procedure without repulsion. Table 4 reports the corresponding individual and ensemble reconstruction metrics.

4 Discussion and Conclusion

We introduced Projected Energy Matching (PEM) as a more stable route to learning time-independent energy-based models through transport-field supervision, as in Energy Matching. A frozen velocity teacher replaces stochastic transport targets with deterministic supervision, while contrastive refinement shapes the potential near the data manifold. On CIFAR-10, this approach yields a stronger training signal, improved seed stability, and lower FID in a compute-matched comparison excluding teacher pretraining.

By making the energy landscape easier to learn, PEM enables Energy Matching at volumetric scale. Our pipeline learns a 3D latent energy prior that achieves lower FID than flow baselines using the same backbone, particularly with an architecture that natively omits time embeddings.

Limitations. PEM substantially amortises the transport regime, while the computational cost of contrastive Langevin sampling and the memory demands of double-backward differentiation remain. Both increase with dimensionality, but are manageable in the volumetric latent space studied here.

Broader Impacts. PEM opens opportunities for wider use of energy-based models in high-dimensional scientific and medical applications. Their scalar formulation allows learned priors to be combined directly with physical constraints and measurement objectives, creating opportunities for controlled generation and inference without task-specific retraining. In CT imaging, sparse-view reconstruction offers a route toward fewer projections and lower radiation dose. Translating this potential into clinical benefit requires validating anatomical fidelity, particularly for structures weakly constrained by the measurements, and robustness beyond the CT-RATE training distribution.

A single unconditionally trained potential generates volumes, scores out-of-distribution inputs, and supports sparse-view CBCT. Its energy can be evaluated directly and combined with measurement objectives and physical constraints without retraining. By combining this flexibility with lower FID than flow baselines, PEM advances a broader paradigm for volumetric generative modeling in which explicit energy landscapes serve as a common foundation for generation and scientific inference.

AI Use Statement

ChatGPT and Claude were used to assist with drafting and polishing the manuscript text, and Claude Code assisted with method implementation, coding and debugging. The research ideas and derivations were developed by the authors. The authors reviewed the code and evaluation, checked the final wording and citations, and take responsibility for the correctness and originality of the final submission.

References

O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, et al. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023. doi: 10.7717/peerj-cs.1516.

M. S. Albergo and E. Vanden-Eijnden. Building Normalizing Flows with Stochastic Interpolants. In *International Conference on Learning Representations*, 2023.

M. Balcerak, T. Amiranashvili, A. Terpin, S. Shit, L. Bogensperger, S. Kaltenbach, et al. Energy Matching: Unifying Flow Matching and Energy-Based Models for Generative Modeling. In *Advances in Neural Information Processing Systems*, volume 38, pages 8583–8609, 2025.

M. Balcerak, S. Shit, C. Prabhakar, S. Kaltenbach, M. S. Albergo, Y. Du, et al. Graph Energy Matching: Transport-Aligned Energy-Based Modeling for Graph Generation, 2026. arXiv:2603.23398.

M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, et al. MONAI: An open-source framework for deep learning in healthcare, 2022. arXiv:2211.02701.

C.-H. Chao, W.-F. Sun, B.-W. Cheng, and C.-Y. Lee. On Investigating the Conservative Property of Score-Based Generative Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4076–4095. PMLR, 2023.

J. Cui and T. Han. Learning Latent Space Hierarchical EBM Diffusion Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 9633–9645. PMLR, 2024.

Y. Du and I. Mordatch. Implicit Generation and Modeling with Energy Based Models. In *Advances in Neural Information Processing Systems*, volume 32, pages 3608–3618, 2019.

Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, et al. Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 8489–8510. PMLR, 2023.

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

R. Gao, Y. Song, B. Poole, Y. N. Wu, and D. P. Kingma. Learning Energy-Based Models by Diffusion Recovery Likelihood. In *International Conference on Learning Representations*, 2021.

P. Guo, C. Zhao, D. Yang, Z. Xu, V. Nath, Y. Tang, et al. MAISI: Medical AI for Synthetic Imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4430–4441. IEEE, 2025. doi: 10.1109/WACV61041.2025.00435.

Q. Guo, C. Ma, Y. Jiang, Z. Yuan, Y. Yu, and P. Luo. EGC: Image Generation and Classification via a Diffusion Energy-Based Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22952–22962, 2023.

I. E. Hamamci, S. Er, C. Wang, F. Almas, A. G. Simsek, S. N. Esirgun, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, 10(8):1610–1628, 2026. doi: 10.1038/s41551-025-01599-y.

A. Heng, A. H. Thiery, and H. Soh. Out-of-Distribution Detection with a Single Unconditional Diffusion Model. In *Advances in Neural Information Processing Systems*, volume 37, pages 43952–43974, 2024. doi: 10.52202/079017-1395.

G. E. Hinton. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

X. Jiang, G. J. Gang, and J. W. Stayman. CTorch: PyTorch-Compatible GPU-Accelerated Auto-Differentiable Projector Toolbox for Computed Tomography, 2025. arXiv:2503.16741.

A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang. Energy-Based Models. In *Predicting Structured Data*, pages 191–246. The MIT Press, 2007. doi: 10.7551/mitpress/7443.003.0014.

C.-K. Lee, F. Coeurdoux, Y. Chen, S. Tang, P. Potaptchik, Y. Du, et al. Beckmann Transport Models: From Autonomous Flows to One-Step Maps, 2026. arXiv:2608.01692.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

X. Liu, C. Gong, and Q. Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023.

S. McCandlish, J. Kaplan, D. Amodei, and OpenAI Dota Team. An Empirical Model of Large-Batch Training, 2018. arXiv:1812.06162.

S. Praneeth, G. Shanker, H. Dubey, and C. S. Seelamantula. Velocity Adaptation for Flow-Matching Models. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026.

S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, et al. MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*, pages 405–415. Springer, 2023. doi: 10.1007/978-3-031-43901-8_39.

T. Salimans and J. Ho. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*, 2022.

T. Schröder, Z. Ou, J. Lim, Y. Li, S. Vollmer, and A. Duncan. Energy Discrepancies: A Score-Independent Loss for Energy-Based Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 45300–45338. Curran Associates, Inc., 2023. doi: 10.52202/075280-1963.

Y. Song and S. Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11918–11930. Curran Associates, Inc., 2019.

Y. Song and D. P. Kingma. How to Train Your Energy-Based Models, 2021. arXiv:2101.03288.

J. Thornton, L. Béthune, R. Zhang, A. Bradley, P. Nakkiran, and S. Zhai. Composition and Control with Distilled Energy Diffusion Models and Sequential Monte Carlo. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 3259–3267. PMLR, 2025.

A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, et al. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. *Transactions on Machine Learning Research*, 2024.

S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg. Plug-and-Play Priors for Model Based Reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

R. Wang and Y. Du. Equilibrium Matching: Generative Modeling with Implicit Energy-Based Models, 2025. arXiv:2510.02300.

M. Woodland, A. Castelo, M. Al Taie, J. Albuquerque Marques Silva, M. Eltaher, F. Mohn, et al. Feature Extraction for Generative Medical Imaging Evaluation: New Evidence Against an Evolving Trend. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012 of *Lecture Notes in Computer Science*, pages 87–97. Springer, 2024. doi: 10.1007/978-3-031-72390-2_9.

S. Yoon, H. Hwang, D. Kwon, Y.-K. Noh, and F. C. Park. Maximum Entropy Inverse Reinforcement Learning of Diffusion Models with Energy-Based Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 24601–24624. Curran Associates, Inc., 2024. doi: 10.52202/079017-0776.

Y. Zhang, P. Yu, Y. Zhu, Y. Chang, F. Gao, Y. N. Wu, et al. Flow Priors for Linear Inverse Problems via Iterative Corrupted Trajectory Matching. In *Advances in Neural Information Processing Systems*, volume 37, pages 57389–57417. Curran Associates, Inc., 2024. doi: 10.52202/079017-1829.

A Appendix: Network Architectures and Implementation Details

All models operate on the compressed 4-channel latent representation of the 3D volumes. We instantiate two backbones. The first, MedNeXt, is used for our headline results and balances the long-range spatial dependencies typically handled by Transformers with the inductive biases of CNNs. The second, a 3D diffusion U-Net, compares against competing methods under a backbone standard in the flow- and diffusion-matching literature, so reported differences are attributable to the objective rather than the architecture. Both backbones are used for the Phase-1 flow teacher and for the Phase-2/3 potential model, and both share an identical potential head.

A.1 Backbone A: MedNeXt

MedNeXt is a fully convolutional 3D architecture inspired by the ConvNeXt design, which scales CNNs using Transformer-driven principles (Roy et al., 2023). The core computational unit is the residual inverted bottleneck block. Each block consists of three layers: (1) a depthwise convolution to aggregate spatial context, (2) an expansion layer (with an expansion ratio of $R = 2$) using $1 \times 1 \times 1$ convolutions and GELU activations to decouple width scaling from receptive field scaling, and (3) a compression layer projecting back to the original channel dimension (Roy et al., 2023). This backbone is time-independent: no timestep embedding is injected.

A.2 Backbone B: Diffusion U-Net (baseline-matched)

The second backbone is the standard 3D DiffusionModelUNet of MONAI Cardoso et al. (2022), with 4 input and output channels and 4 resolution levels of width $(16, 32, 64, 128)$. Each level contains 2 residual blocks with group normalisation over 16 groups. Self-attention is applied only at the lowest-resolution level, with 64 channels per attention head and attention logits upcast to `float32` for numerical stability; no cross-attention conditioning is used. Unlike Backbone A, this network receives a sinusoidal time step embedding which is only used for OT-Flow Matching Lipman et al. (2023) and Rectified Flow Liu et al. (2023). For PEM trained with this backbone, the models are time-independent, which is achieved by passing a dummy $t = 0.5$ on every pass.

A.3 Phase 1: Unconstrained Flow Teacher

In the first phase (learning the Optimal Transport flow), the network must output an unconstrained 3D marginal vector field v_{teacher}. For Backbone A we append a shallow *Velocity Head*: a single 3D MedNeXt block (expansion ratio 2, $3 \times 3 \times 3$ kernel, group normalisation) followed by a $1 \times 1 \times 1$ convolution mapping back to the 4-channel latent dimension. For Backbone B the U-Net already emits a 4-channel tensor at full latent resolution, so the velocity head is the identity. In both cases the teacher is a highly expressive, unconstrained functional mapping $\mathbf{x}_t \mapsto v(\mathbf{x}_t)$.

A.4 Phases 2 and 3: Projection and Potential Training

During the projection phase, and in the subsequent potential refinement phase, the architecture must shift from a dense vector field to a single global scalar potential $\phi_\theta(\mathbf{x}_t)$. We replace the velocity head with a *Potential Head* that is identical for both backbones, so that the two arms differ only in their feature extractor. The head is a small Transformer:

1. **Patch embedding:** a non-overlapping 3D patch embedding (patch size 2^3) maps the 4-channel backbone output to a sequence of tokens of dimension $d = 128$.
2. **Global mixing:** two pre-norm Transformer encoder layers ($d = 128$, 4 attention heads, feed-forward dimension 512, SiLU activation, no dropout), implemented with a `torch.compile`- and FlashAttention-compatible layer so that the double-backward pass through the head remains fused.
3. **Scalar projection:** mean pooling over the token dimension followed by a linear map $\mathbb{R}^{128} \to \mathbb{R}$, yielding one scalar per sample, $\phi_\theta(\mathbf{x}_t) \in \mathbb{R}^B$.

Self-attention is used here rather than further convolution because the potential is a single global functional of the volume: the head must aggregate evidence across the entire latent field, not refine

local structure. The forward pass of the functional velocity model computes the scalar $\phi_\theta(\mathbf{x}_t)$ via the potential model and then differentiates with respect to the input to output the conservative field $-\nabla_{\mathbf{x}}\phi_\theta(\mathbf{x}_t)$. Because the head is a scalar, the gradient is obtained in a single reverse-mode pass, and the training objective requires one further backward pass through that graph.

B Appendix: Datasets and Preprocessing Details

To rigorously evaluate the scalability and physical accuracy of our Projected Energy Matching Model, we use CIFAR-10 as a low-dimensional, fully controlled testbed for the gradient-noise analysis, alongside a 3D CT-dataset that exhibit complex, continuous topologies and strict domain constraints. This section details the preprocessing pipelines, autoencoder compression.

B.1 CIFAR-10: Low-Dimensional Variance Study

CIFAR-10 (Krizhevsky, 2009) consists of 60,000 RGB images at 32×32 resolution, split into 50,000 training and 10,000 test images over 10 classes. We use it exclusively for the gradient-noise and critical-batch-size analysis (Section 3.1), where the low dimensionality makes the paired scratch/projected sweep over batch sizes, learning rates, and seeds tractable. Class labels are discarded; all models are trained unconditionally. The 50,000 training images are further split 90/10 into 45,000 training and 5,000 held-out images, the latter providing the shared evaluation split used for the flow-matching loss closure reported across arms.

Preprocessing. Images are used at their native $32 \times 32 \times 3$ resolution, so both the teacher velocity field and the student potential operate directly in pixel space. The pipeline consists of random horizontal flipping, followed by an intensity rescaling from $[0, 255]$ to $[-1, 1]$ via per-channel normalization with mean and standard deviation 0.5. Samples are mapped back to $[0, 1]$ at generation time for evaluation.

B.2 CT-RATE: High-Dimensional 3D Medical Volumes

The CT-RATE dataset (Hamamci et al., 2026) is a large-scale, real-world cohort of volumetric chest CT scans, providing complex, continuous 3D anatomies. The dataset consists of 25,692 non-contrast chest CT volumes (expanded to 50,188 through various reconstructions) originating from 21,304 unique patients. Following Hamamci et al. (2026), we divided the cohort at the patient level: 20,000 patients were allocated to the training set, and the remaining 1,304 patients were reserved for the validation set.

Preprocessing and Normalization. To standardize the volumes for neural network processing, we employ a rigorous spatial and intensity normalization pipeline. First, all scans are spatially resampled to a voxel spacing of $1.5 \times 1.5 \times 2.367$ mm^3 using trilinear interpolation. To achieve a uniform high-resolution grid of $128 \times 128 \times 64$ voxels, we center-crop or pad the axial (XY) plane. If the z-axis (depth) exceeds 64 slices, it is center-cropped; if it is shorter, it is padded by replicating the first and last boundary slices.

Latent Space Embedding via MAISI. We compress the preprocessed volumes using a 3D Variational Auto Encoder (VAE) into a latent space Guo et al. (2025).

The model is trained on a combination of objectives to ensure the volume reconstructions adhere closely to the image manifold with high local realism. The overall objective integrates a voxel-space L_1 reconstruction loss (L_{recon}), an LPIPS perceptual loss (L_{lpips}), an adversarial loss (L_{adv}) utilizing a 3D discriminator to penalize unrealistic artifacts, and a Kullback-Leibler (KL) regularization term (L_{reg}) to enforce a standard normal distribution on the latent features, preventing high-variance latent spaces.

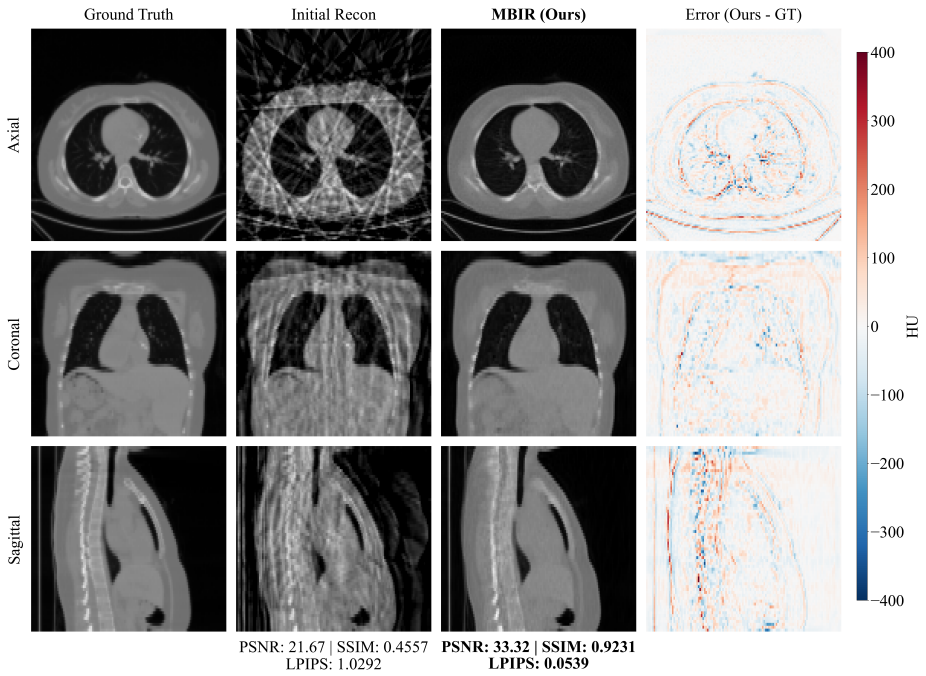


Figure 7: **Sparse-View CBCT Reconstruction via MBIR.** Qualitative evaluation on an extreme ill-posed inverse problem (20 of 600 projections).**Left:** full-view ground truth. **Middle:** the initial analytical reconstruction, heavily corrupted by sparse-view streak artifacts. **Right:** a *MAP sample* obtained by a joint objective of Eq. 5 (33.32 dB, SSIM 0.9231, LPIPS 0.0539)

C Appendix: Generative Quality

C.1 Ill-Posed Inverse Problem: Sparse-View CBCT Reconstruction

We evaluate the trained PEM as a zero-shot prior for sparse-view cone-beam CT (CBCT) reconstruction, a severely ill-posed inverse problem, using only 20 of 600 projections (Fig. 7).

C.2 Manifold Exploration

Since PEM is time-independent, Langevin sampling on the learned 3D scalar potential can continue far beyond the flow endpoint while staying on the data manifold (Fig. 8).

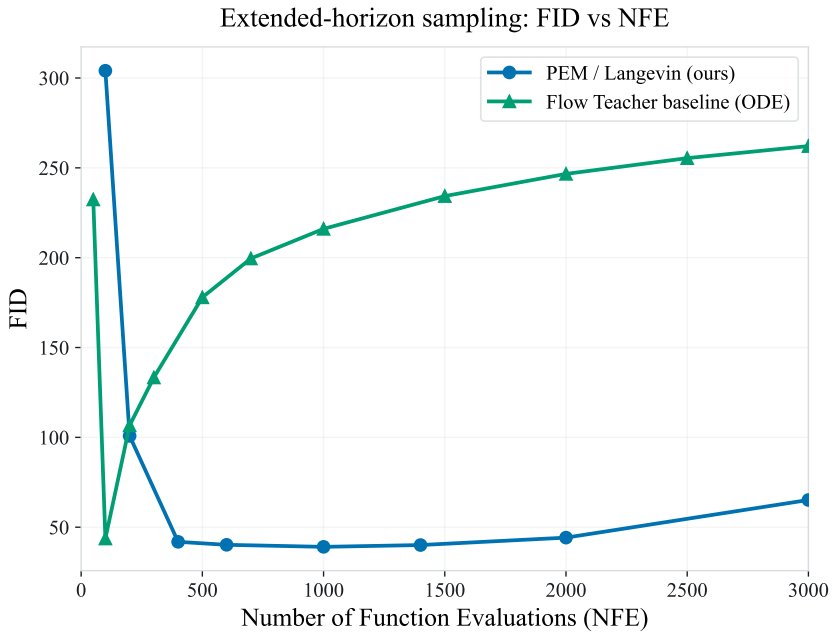


Figure 8: **Extended-Horizon Sample Quality.** Quantitative evaluation of Fréchet Inception Distance (FID) across cumulative evaluation costs (NFE). OT Flow Matching model (green) degrades sharply when the deterministic ODEs are integrated past its native $t = 1$ horizon. In contrast, our Projected Energy Matching (PEM, blue) safely permits unbounded Langevin sampling. By continuing to sample well beyond standard limits (extending past 2000 NFE), PEM maintains a stable, low FID, successfully avoiding the severe artifact accumulation typically seen when forcing time-dependent models past their endpoint.

D Appendix: Algorithms

Algorithm 2 Projection of the Flow Teacher (Phase 2).

Require: Pre-trained teacher flow v_ψ (frozen)
Require: Potential ϕ_θ, spatial dimension factor $N = D$
Initialize θ randomly
for iteration $0, 1, \ldots$ **do**
  Sample mini-batch $\{x^{(j)}_{\text{data}}\}_{j=1}^{B} \sim \mathcal{D}$ and $\{x^{(j)}_{\text{noise}}\}_{j=1}^{B} \sim \mathcal{N}(0, I)$
  Sample independent time steps $\{t^{(j)}\}_{j=1}^{B} \sim \mathcal{U}[0, 1]$
  $x_t^{(j)} \leftarrow t^{(j)} x^{(j)}_{\text{data}} + (1 - t^{(j)}) x^{(j)}_{\text{noise}}$ ▷ Interpolate along optimal transport paths
  $v_{\text{teacher}} \leftarrow v_\psi(x_t)$ ▷ Single forward pass, no gradient
  $v_{\text{base}} \leftarrow -N \nabla_x \phi_\theta(x_t)$ ▷ Conservative field by construction
  $\mathcal{L}_{\text{proj}} \leftarrow \frac{1}{B} \sum_{j=1}^{B} \| v^{(j)}_{\text{base}} - \text{sg}[v^{(j)}_{\text{teacher}}] \|_2^2$
  Update $\theta \leftarrow \text{Optimizer}(\nabla_\theta \mathcal{L}_{\text{proj}})$ ▷ Double-backward through $\nabla_x \phi_\theta$
end for
Return conservative potential ϕ_θ

E Appendix: Ablations

Table 3: Quantitative ablation of the progressive UNet training phases on the CT-Rate dataset. The table reports parameter counts and the resulting generative fidelity at each phase of the pipeline.

Method	Params	FID ↓
PEM Phase 1	10.3M	41.90
PEM Phases 1+2	10.7M	62.60
PEM Phases 1+2+3	10.7M	40.87

F Bayesian Analysis Details

All intervals are 95% highest-density intervals. Paired effects (Section 3.1 - 3.1) use conjugate or closed-form posteriors; the curves in Fig. 2 are fitted with NUTS in PyMC Abril-Pla et al. (2023).

Critical batch size. For each checkpoint j of run r admissible in both arms ($n = \mathbf{72}$ pairs, **8** runs), let δ_{rj} be the difference in $\log_{10} B_{\text{crit}}$ between PEM and EM. Checkpoints within a run are dependent, so we pool them through a run-level effect:

$$\delta_{rj} \sim \mathcal{N}(\theta_r, \sigma^2), \quad \theta_r \sim \mathcal{N}(\mu, \tau^2), \quad \mu \sim \mathcal{N}(0, 10^2), \quad \sigma^2, \tau^2 \sim \text{InvGamma}(2, 1), \tag{6}$$

and report 10^{μ}. We sample with Gibbs (40,000 iterations, 10,000 burn-in).

FID. Final-checkpoint log FID differences use a normal model with Jeffreys prior, whose posterior for the mean is Student-t. For the $B = 32$ long runs, the same hierarchical model as above is fitted with seed as the group. Seed dispersion at $B \in \{32, 128\}$ uses the scaled inverse-χ^2 posterior of the residual variance ($\nu = 4$); we report the ratio of standard deviations.

Admissibility. Beta-Binomial per arm with a $\text{Beta}(1, 1)$ prior.

Figure 2 (a) Per arm, $\log_{10}$ SNR is regressed on $\log B$ and $\log t$ with run intercepts and a Student-t likelihood. (b) Per arm, FID follows a saturating power law $F_r + A_r(t/t_0)^{-\alpha_r}$ with run-level parameters, fitted on $\log_{10}$ FID with a Student-t likelihood. Lines are posterior medians averaged over observed runs; bands are 95% HDIs. We use 4 chains of 2,000 draws after 3,000 tuning steps, with $\hat{R} \leq 1.01$ and no divergences.

G Appendix: Model-Based Iterative Reconstruction (MBIR) Frameworks

In this work, we deploy our learned 3D priors to solve ill-posed medical inverse problems, specifically Model-Based Iterative Reconstruction (MBIR) for sparse-view CBCT in latent space.

G.1 Forward Model and Data Consistency

The inference approach utilizes a physical forward modeling pipeline built on CTorch Jiang et al. (2025). The data consistency loss $\mathcal{L}_{\text{data}}(\mathbf{z})$ is formulated depending on the clinical scenario:

- **Sparse-View CT:** We utilize a standard Mean Squared Error (MSE) loss against the measured sparse sinogram $\mathbf{y}_{\text{meas}}$: $\mathcal{L}_{\text{data}} = \frac{1}{2}\|\mathbf{y}_{\text{meas}} - A(\mathcal{D}(\mathbf{z}))\|_2^2$.

G.2 Posterior Sampling via Explicit Langevin Dynamics

Our primary method leverages the explicit nature of our Projected Energy Matching to perform unconstrained posterior sampling. Instead of directly optimizing for a single point estimate, we sample from the posterior distribution $p(\mathbf{z}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{z})p(\mathbf{z})$ using Langevin dynamics.

As implemented in our solver, we initialize the latent state with random noise and integrate using a Heun-based Stochastic Differential Equation (SDE) solver Balcerak et al. (2025). The update step is

guided by the joint gradients of the explicit energy prior and the physical data loss:

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \gamma \nabla_{\mathbf{z}} \left(\lambda_{\text{data}} \mathcal{L}_{\text{data}}(\mathbf{z}_k) + \phi_\theta(\mathbf{z}_k) \right) + \sqrt{2\gamma}\boldsymbol{\epsilon} \tag{7}$$

where γ is the step size, λ_{data} controls the relative strength of the physical constraints, and $\phi_\theta(\mathbf{z})$ is the learned explicit scalar potential. The sampling procedure allows the state to settle into a low-energy basin on the data manifold while strictly adhering to the projection geometries.

H Appendix: Repulsion

We optimise $K = 3$ latents $\mathbf{z}_1, \ldots, \mathbf{z}_K$ jointly, initialised from the same FBP reconstruction plus Gaussian noise ($\sigma = 0.01$), with decoded volumes $\mathbf{x}_k = \mathcal{D}(\mathbf{z}_k)$. Each $\mathbf{z}_k$ descends Eq. 5 plus the repulsion term following Balcerak et al. (2025)

$$\mathcal{L}_{\text{rep}} = \sum_{i<j} \exp\left(-\frac{\|\mathcal{B}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2}{2h^2} \right), \qquad \mathcal{B} = \mathcal{I} - \mathcal{R}\mathcal{A}, \tag{8}$$

where $\mathcal{R}$ is filtered back-projection. $\mathcal{B}$ removes the part of a volume that is recoverable from its projections, so $\|\mathcal{B}(\mathbf{x}_i - \mathbf{x}_j)\|_2$ is the distance between two particles in the (approximate) null space of $\mathcal{A}$, and $\mathcal{L}_{\text{rep}}$ separates them only where the measurements are uninformative. The bandwidth h is the median of these pairwise distances, recomputed during optimisation and not differentiated. The gradient $\nabla_{\mathbf{z}_k} \mathcal{L}_{\text{rep}}$ is rescaled to norm $w_{\text{rep}} \|\nabla_{\mathbf{z}_k} \phi_\theta\|_2$ with $w_{\text{rep}} = 0.1$, so repulsion cannot override the prior. The baseline is identical with $w_{\text{rep}} = 0$, i.e. particles differ only through their initial noise. We report metrics of single particles, of the ensemble mean $\bar{\mathbf{x}} = \frac{1}{K} \sum_k \mathbf{x}_k$, and of the best particle selected with the ground truth (best-of-K). Calibration r is the Pearson correlation over voxels between the ensemble standard deviation and $|\bar{\mathbf{x}} - \mathbf{x}_{\text{gt}}|$.

Table 4: Sparse-view CBCT ($20/600$ views), $K = 3$ particles sampled with Langevin posterior sampling with and without null-space repulsion. Mean over 10 volumes $\times$ 3 seeds.

	PSNR ↑	SSIM ↑	LPIPS ↓
Without repulsion			
Single particle	29.88	0.808	0.202
Ensemble mean	30.02	0.815	0.198
With repulsion			
Single particle	29.92	0.811	0.196
Ensemble mean	**30.22**	**0.822**	**0.190**