Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Simulation-based inverse imaging of dynamic material-interface states from radiographic surrogates and sparse projections

Nadav Schneider
Israel Atomic Energy Commission
Shai Stamler-Grossman
Israel Atomic Energy Commission
Re'em Harel
Israel Atomic Energy Commission
Yehonatan Fridman
Israel Atomic Energy Commission
Gershon Hanoch
Israel Atomic Energy Commission
Assaf Shimony
Nuclear Research Center Negev
Gal Oren
galoren@stanford.edu
Stanford University

Article

Keywords:

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10982889/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Simulation-based inverse imaging of dynamic material-interface states from radiographic surrogates and sparse projections

Nadav Schneider[1], Shai Stamler-Grossman[1], Re'em Harel[1], Yehonatan Fridman[1], Gershon Hanoch[1], Assaf Shimony[2], and Gal Oren[3,4,*]

[1]Israel Atomic Energy Commission, Israel
[2]Nuclear Research Center Negev, Beer Sheva, Israel
[3]Technion - Israel Institute of Technology, Haifa, Israel
[4]Stanford University, Stanford, California, USA
[*]Correspondence: galoren@stanford.edu

ABSTRACT

Fast radiography of dynamic material interfaces is blurred, sparse, and difficult to interpret in three dimensions. Restoration, state inference, and sparse-view reconstruction are often studied separately, which makes their outputs difficult to compare against known simulation labels. We address this problem with a curated Rayleigh–Taylor instability (RTI) corpus containing 101,250 interface images from 1,350 simulations, each labeled by Atwood number, acceleration, initial perturbation amplitude, and time. The simulations are sampled on a common uniform time grid, although acceleration and time are only partly independent under the normalized-time scaling of the underlying RTI problem. Controlled temporal averaging and spatial blur generate radiograph-like surrogates with known target states. On these surrogates, a fine-tuned restoration model improves mean SSIM from 0.905 to 0.962 and mean PSNR from 20.67 to 22.05 dB relative to Restormer. The same corpus supports retrieval, regression of Atwood number and time, local ambiguity diagnostics, short-horizon temporal densification, and a lifted 4D benchmark with 54 dynamic 3D sequences and 72 synthetic projections per volume. Sparse-view neural fields reach about 35 dB on held-out views in that synthetic benchmark. At sufficiently late times, some frames are best interpreted as controlled image-processing cases because finite-domain effects can influence the morphology. Within that scope, the RTI corpus provides a shared benchmark for restoration, label-space inference, ambiguity analysis, and sparse-view reconstruction.

Introduction

Dynamic material interfaces are central to hydrodynamic mixing in liquids, plasmas, and other accelerated multi-material systems. For sufficiently small initial perturbations, the Rayleigh–Taylor instability evolves from linear perturbation growth to nonlinear bubble–spike development and secondary roll-up; the resulting morphology depends on the density contrast, acceleration, initial perturbation amplitude, wavelength, and time.[1–4] Single-mode liquid experiments and high-resolution radiography on high-energy-density platforms both exploit this sensitivity, which makes RTI a useful benchmark for inverse imaging of evolving interfaces in a well-defined class of flows.[5,6]

The imaging problem is intrinsically difficult. Fast radiographs in high-energy-density physics and dynamic materials experiments integrate over motion, blur through finite source and detector response, and often provide only a few views of an evolving 3D state.[7–9] Similar tradeoffs appear in time-resolved tomography and fast micro-CT, where angular sparsity and acquisition speed must be balanced against image quality and temporal resolution.[10–12] In such settings the image is not a complete observation of the interface; it is a sparse, instrument-filtered projection whose interpretation becomes increasingly dependent on three-dimensional modeling.[13]

This makes dynamic interface analysis an inverse-imaging problem. Data-driven inverse methods are most useful when they preserve the connection between observations and the latent physical variables that generated them, not just image appearance.[14,15] In fluid mechanics and scientific machine learning, curated simulations are increasingly used to learn observables, surrogates, and inverse mappings when experiments are scarce or expensive.[16,17] For the single-mode RTI family used here, acceleration and physical time are also not fully independent morphology coordinates: much of the evolution can be organized by normalized time $\tilde{t} = \sqrt{\mathscr{A} g/\lambda}\, t$, so different (g,t) pairs can produce similar interface shapes.[17] A simulation library with known labels can therefore serve as a reference space for interpreting sparse or blurred observations, provided that this degeneracy is kept in view.

Existing methods usually address only one part of this problem. Established restoration backbones improve image fidelity,

compact learned embeddings and regression models can map morphology back to simulation labels, and sparse-view CT or neural-field methods provide strong priors for volumetric recovery under angular scarcity.[17–27] Yet these tools are usually evaluated with task-specific metrics and only limited connection to the physical variables that generated the data. A shared evaluation protocol linking restoration, state inference, ambiguity analysis, and 4D reconstruction is still missing.

This article uses the released *RayleAI*, CLARIPHY, and *SDT* resources to turn that idea into a common benchmark. We build a labeled RTI corpus, convert it into radiograph-like surrogates through controlled degradation, test restoration on morphologically complex cases, infer simulation labels through retrieval and regression, examine underdetermination with local matching and temporal densification, and then lift the same family into a sparse-view 4D benchmark based on synthetic projections.

The evidence is separated deliberately. A low-Atwood liquid RTI experiment provides a qualitative morphology anchor. Controlled 2D degradations test restoration and state inference on label-preserving radiograph-like surrogates. A synthetic 2D-to-3D lift defines the volumetric benchmark, and sparse-view neural fields are evaluated only in that simulation-derived setting rather than against real 3D experimental ground truth.

The contribution is therefore a reproducible RTI-labeled evaluation protocol linking restoration, state inference, ambiguity analysis, temporal densification, and sparse-view reconstruction. It measures agreement with the assigned simulation labels or with held-out synthetic radiographs, rather than direct validation of the governing RTI physics. Newer models can be inserted into the same setting without changing that question.

Results

We begin with the labeled RTI corpus because every later test depends on whether morphology varies systematically with the assigned simulation labels. We then examine how much structure survives controlled degradation, how well restored or retrieved images stay near the correct labels, where local morphology becomes ambiguous, and how far a synthetic sparse-view 4D benchmark can be pushed before angular ambiguity dominates.

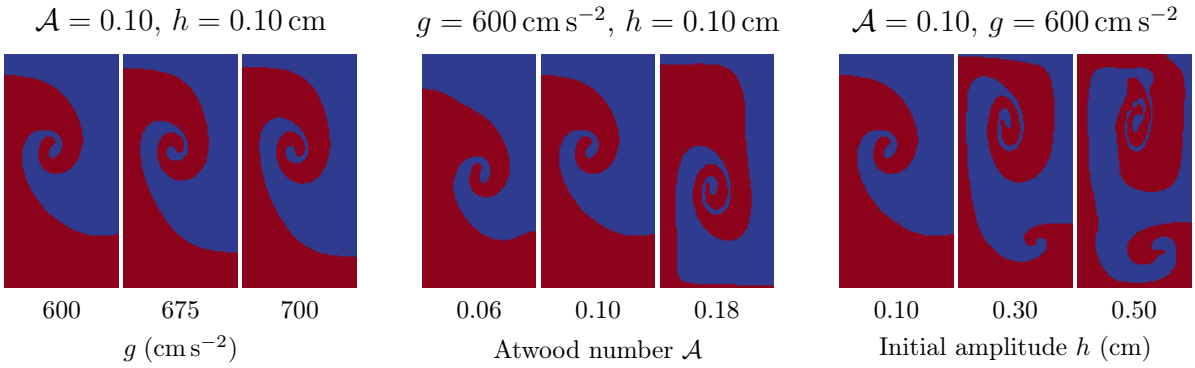
A labeled RTI simulation space for dynamic interface morphology

The study begins with an RTI benchmark containing 101,250 thresholded interface images from 1,350 simulations, with 75 saved times per run on a common grid from $t = 0.00$ to 0.74 s in 0.01 s increments.[17] Every frame inherits four assigned state variables: Atwood number $\mathscr{A}$, gravitational acceleration g, initial perturbation amplitude h, and time t. The simulations span $\mathscr{A} \in [0.02, 0.50]$, $g \in [600, 800]$ cm s^{-2}, and $h \in [0.1, 0.6]$ cm on a 64×128 grid covering a 2.7 cm horizontal by 5.4 cm vertical domain. In the single-mode setup, the 2.7 cm horizontal extent corresponds to half of the imposed wavelength, so $\lambda = 5.4$ cm in the normalized-time definition $\tilde{t} = \sqrt{\mathscr{A} g/\lambda}\, t$ used in Ref.[17]. Across the full corpus, $\tilde{t}$ therefore spans approximately 0 to 6.37.

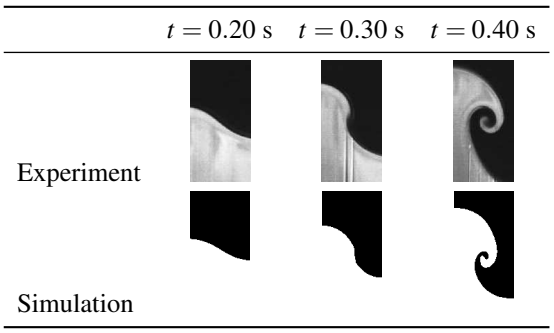
These labels are useful but not fully independent morphology coordinates. Because g and t partly collapse under normalized time, raw proximity in $(\mathscr{A}, g, h, t)$ can overcount states that remain visually similar after acceleration-time rescaling.[17] In addition, the finite rectangular domain and reflecting boundaries can influence the latest frames. At sufficiently late times, the interface may interact with the computational boundaries or with reflected disturbances. Those frames remain useful as controlled image-processing benchmarks, but they should not be interpreted as the unconstrained late-time evolution of the original RTI problem.

Figure 1 illustrates how the corpus is used. At fixed raw time $t = 0.55$ s, changing g, $\mathscr{A}$, or h changes spike penetration, bubble width, and roll-up. The Atwood-number sweep is ordered by increasing $\mathscr{A}$, making the corresponding increase in interface excursion visible directly. The low-Atwood single-mode experiment of Waddell *et al.* occupies the same qualitative morphology family and serves as an experimental anchor for the benchmark.[5] The corpus provides a shared reference for later error analysis while preserving genuine ambiguity in the inverse problem.

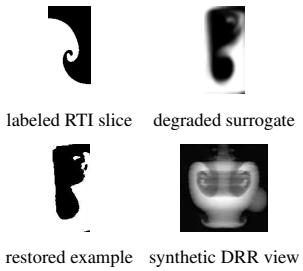
Table 1. Public datasets and labels used across the study.



(a) Examples at fixed raw time $t = 0.55$ s. The Atwood-number sweep shows $\mathscr{A} = 0.06$, 0.10, and 0.18 from left to right at fixed $g = 600$ cm s^{-2} and $h = 0.10$ cm, illustrating the increase in interface excursion with increasing $\mathscr{A}$.



(b) Qualitative comparison with the low-Atwood single-mode experiment of Waddell *et al.*



(c) Conceptual map of the benchmark layers. The images are illustrative examples from the released resources rather than successive transformations of one sample.

Figure 1. The RTI benchmark defines a labeled simulation space. The verified parameter sweep shows how morphology changes with the RTI state variables at fixed raw time, the experiment–simulation strip places the benchmark in the canonical low-Atwood liquid regime, and the conceptual map identifies the benchmark layers with illustrative examples.

Layer	Data representation	Scale and labels	Scientific role
RayleAI	Labeled 2D interface images	101,250 images from 1,350 simulations, 75 saved times each on a common 0.00–0.74 s grid; labels $(\mathscr{A}, g, h, t)$; 64×128 cells over a 2.7 cm $\times$ 5.4 cm domain; $\tilde{t} = \sqrt{\mathscr{A} g/\lambda}\, t$ with $\lambda = 5.4$ cm	Reference RTI label space for morphology, retrieval, regression, temporal completion, degradation, and 2D-to-3D lifting
CLARIPHY pairs	Sharp/degraded 2D image pairs	Temporal averaging over 3, 5, 7, 9, or 13 frames (0.02–0.12 s first-to-last span) and spatial blur radii of 3, 5, 7, or 9 pixels (about 0.10–0.30 cm horizontally and 0.09–0.28 cm vertically after preprocessing); labels inherited from *RayleAI*	Experiment-like observation model for restoration under limited temporal and spatial resolution
SDT	Dynamic 3D volumes and sparse radiographs	54 dynamic 3D sequences, 72 projections per volume, symmetric and symmetry-broken variants	Sparse-view 4D reconstruction benchmark with held-out-view evaluation

Controlled degradation generates label-preserving radiographic surrogates

Starting from the labeled RTI frames, the study constructs radiograph-like surrogates by combining temporal averaging over 3, 5, 7, 9, or 13 consecutive saved frames with spatial blur from spherical kernels of radius 3, 5, 7, or 9 pixels. On the 0.01 s saved-time grid, these windows span 0.02 to 0.12 s between the first and last averaged frames. Under the normalized-time scaling above, the same windows span about 0.03 to 0.17 for a 3-frame window and 0.18 to 1.03 for a 13-frame window across the full $(\mathscr{A}, g)$ range. The target remains the sharp central frame, so every degraded image inherits the central-state labels while emulating motion integration and finite spatial resolution.

The procedure isolates two dominant sources of measurement loss while preserving the central-frame labels needed for controlled evaluation. The degraded inputs therefore remain tied to the same state variables used later for retrieval, regression, and 3D lifting. Figure 2a shows how local structure is softened first by temporal averaging and then by additional spatial blur. The lower example is centered at a later raw time and is included as a difficult image-processing case; at such stages, depending on $\mathscr{A}$ and g, finite-domain effects can already influence the morphology.

The most severe information loss appears when temporal averaging and spatial blur are combined in morphologically complex cases, where rolled-up boundaries merge and narrow spikes broaden.

A restoration module preserves late-time nonlinear morphology

Restormer provides the baseline restoration model, and CLARIPHY is the RTI-specific fine-tuned version.[19,28] This experiment tests whether restoration recovers contours that are closer to the sharp RTI states. Averaged over 57 released time-ordered validation cases, which the public benchmark presents as progressively harder evaluation examples rather than as a sweep over unseen parameter tuples, the blurred inputs have mean SSIM 0.837 and mean PSNR 19.48 dB, the Restormer baseline reaches 0.905 and 20.67 dB, and CLARIPHY reaches 0.962 and 22.05 dB (Table 2). The public release does not provide a simulation-level manifest for these 57 cases, so they should not be interpreted as independent-by-simulation test averages.

Restoration error generally increases for morphologically complex states at larger normalized times, particularly when roll-up produces fine boundaries. In the released 57-case sequence, however, the later examples also mix parameter-dependent acceleration-time scaling, stronger blur, and, in the most extreme cases, finite-domain effects. Over the final 12 time-ordered validation cases, mean SSIM increases from 0.782 with Restormer to 0.916 with CLARIPHY, while mean PSNR increases from 16.30 to 18.07 dB. Figure 2b illustrates the main difference: CLARIPHY preserves sharper bubble caps, narrower spikes, and clearer roll-up than the pretrained baseline in the cases that are most informative for state inference. A reduced-data run trained on only 10% of the paired corpus follows similar validation trends, indicating that the restoration prior is learnable without exhaustive exposure to the full dataset. Because the public 10% setup downsamples images rather than documented simulation groups and does not expose a recoverable split manifest, this reduced-data result is best read as a benchmark sensitivity check rather than as evidence of generalization to unseen trajectories. Residual error remains concentrated near late-time roll-up boundaries, where small contour shifts change local morphology substantially.

Retrieval and coordinate inference align morphology with the simulation labels

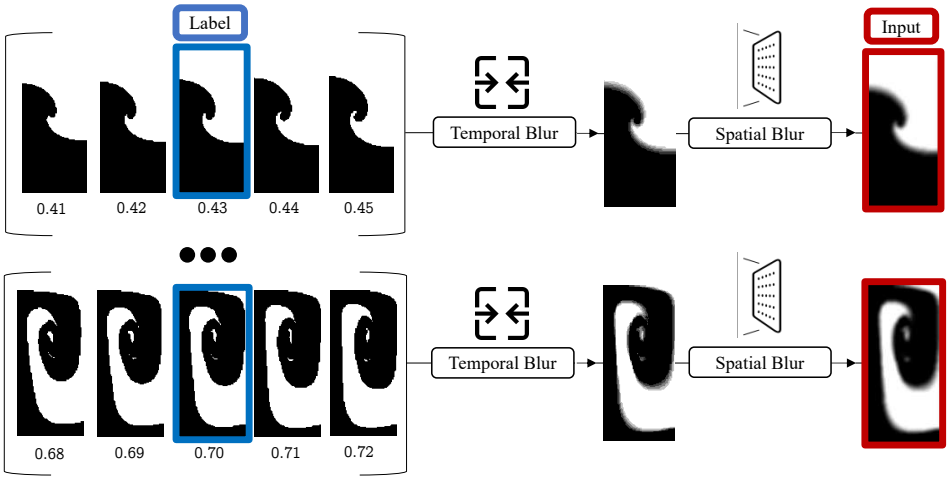
The RTI corpus also supports a direct test of whether image similarity reflects proximity in the assigned label space. LIRE provides a hand-crafted control, and InfoGAN provides a compact learned embedding; both are judged by label-space proximity rather than pixel error.[17,29,30] On 13,000 random queries and a second set of about 10,000 complex images with $t > 0.25$ s, the learned embedding places closer label-space neighbors earlier in the ranking, especially in the nonlinear regime. For the complex queries, the first returned indices cluster around an average Atwood-number normalized label-space error of about 0.25 for InfoGAN, compared with about 0.4 for LIRE.

Because g and t partly collapse under normalized time, raw four-parameter distances are only a proxy for morphological similarity. The released retrieval plots therefore emphasize Atwood-number ranking, and the reduced pReg probe is restricted to $\mathscr{A}$ and t. Figure 3d,e shows that this ordering tracks the normalized label-space error closely and produces visually coherent late-time neighbors. In a representative example, the model predicts $\mathscr{A} = 0.1482$ and $t = 0.4050$ s for a reference state with $\mathscr{A} = 0.155$ and $t = 0.400$ s. This is not full state recovery: gravity and initial perturbation amplitude remain unresolved in the reduced pReg setting.

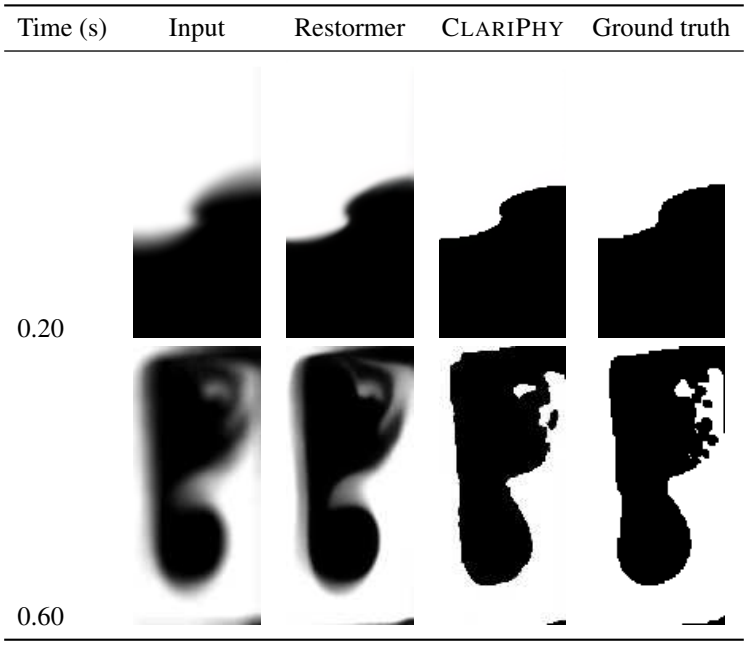
Ambiguity-aware local matching and temporal densification reveal underdetermination

Not every observation determines a unique global state, and the benchmark makes that ambiguity measurable. The local-matching stage is intentionally simple and interpretable: sixteen experimental templates were searched against the RTI corpus with Quality-Aware Template Matching (QATM), and the returned patches were analyzed with PCA and k-means clustering.[17,31] Figure 4a–c distinguishes three regimes: distinctive templates produce a compact high-score neighborhood, partially distinctive templates remain informative over only a limited rank range, and non-distinctive templates distribute across several clusters. Non-unique matches therefore identify cases in which local morphology is insufficient to determine a unique global state.

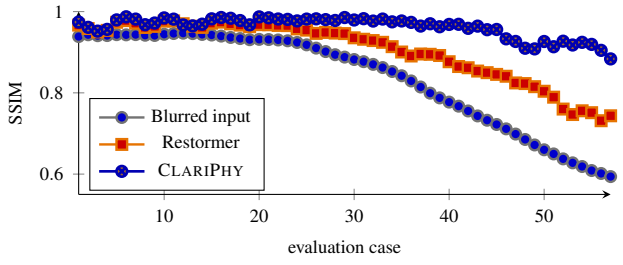
Temporal completion provides a complementary short-horizon constraint. PredRNN is used here as a lightweight temporal-densification module.[17,32] Figure 4d,e shows that SSIM and PSNR remain high over short prediction horizons and then decay as nonlinear roll-up accumulates, with experimental trajectories remaining more stable than long-horizon simulation extrapolations. The useful regime is therefore short-gap completion; the error grows steadily once the forecast extends into strongly nonlinear evolution.



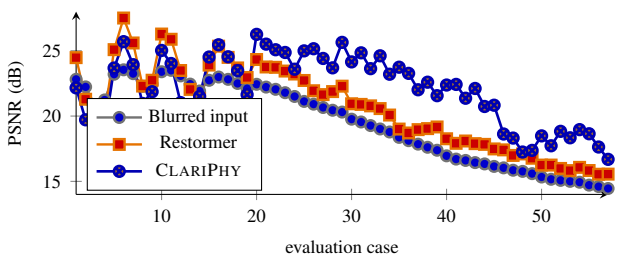
(a) Label-preserving temporal and spatial blur generation.



(b) Representative early and late restoration examples.

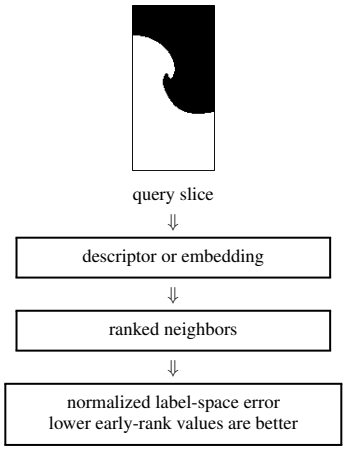


(c) SSIM across the 57 released evaluation cases.

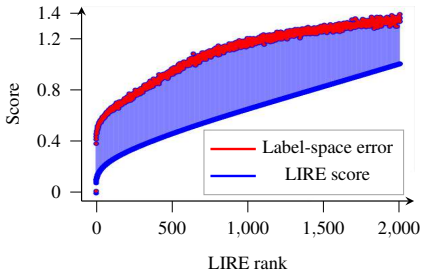


(d) PSNR across the 57 released evaluation cases.

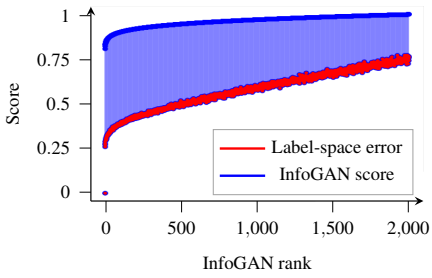
Figure 2. Radiograph-like degradation and restoration. Temporal averaging and spatial convolution convert sharp RTI states into degraded images while preserving the central-frame labels. The later examples are included as challenging benchmark cases; depending on $(\mathscr{A}, g, h, t)$, comparable raw times can already be influenced by the finite domain. Fine-tuned restoration modestly improves average PSNR and preserves spike, bubble, and roll-up boundaries that are important for later label-space analysis.



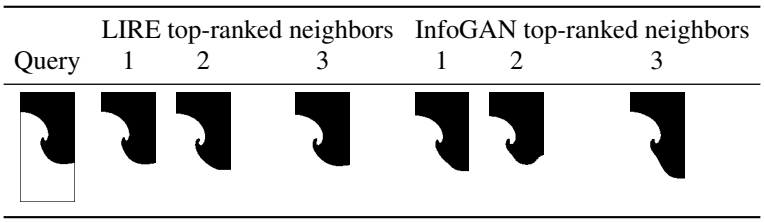
(a) Query-to-ranking evaluation logic.



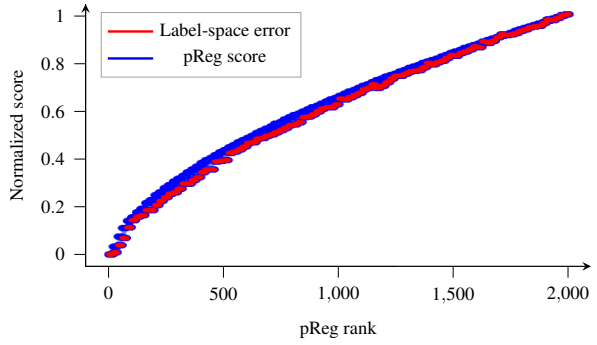
(b) Appearance-level retrieval with LIRE.



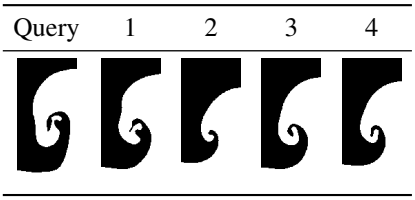
(c) Learned retrieval with InfoGAN.



(d) Representative late-time query and first retrieved neighbors.

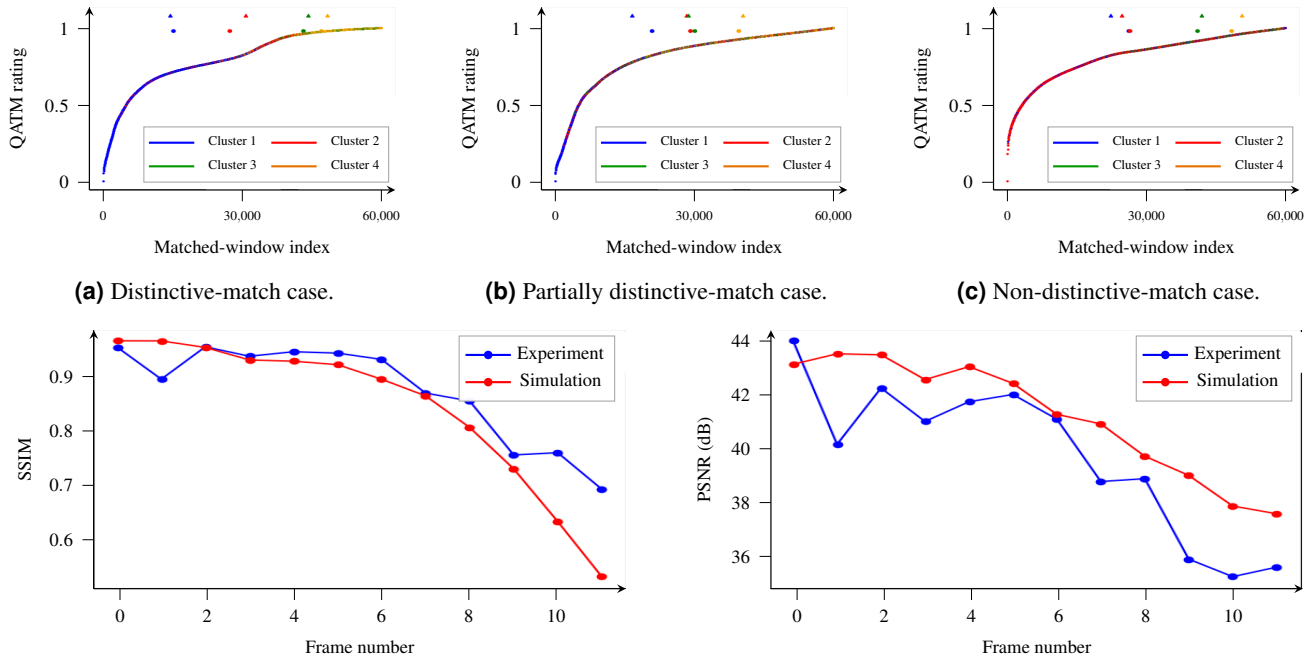


(e) pReg ordering over the regressed $(\mathscr{A}, t)$ coordinates.



(f) Representative pReg neighbors for a late-time query.

Figure 3. Retrieval and coordinate inference in label space. LIRE is a classical content-based retrieval baseline with hand-crafted descriptors, whereas InfoGAN provides a compact learned embedding. For the retrieval curves, the red trace is normalized label-space error, so lower values at early ranks indicate closer agreement with the assigned RTI parameters. Because acceleration and raw time are partly degenerate under normalized time, the released comparison emphasizes Atwood-number error when contrasting LIRE and InfoGAN. The example strip shows that the late-time query is matched to a tighter neighborhood by InfoGAN than by LIRE. The pReg panels summarize direct regression of $\mathscr{A}$ and t: the ranking curve follows label-space error closely, and the neighbor strip remains coherent in the nonlinear regime.



(a) Distinctive-match case. **(b)** Partially distinctive-match case. **(c)** Non-distinctive-match case.

(d) SSIM over densification horizon. **(e)** PSNR over densification horizon.

Figure 4. Ambiguity diagnostics and short-horizon densification. QATM is applied to 16 experimental templates and the returned matches are grouped with PCA and k-means. Distinctive templates create a compact early-rank cluster, partially distinctive templates remain informative only over a limited range, and non-distinctive templates mix clusters quickly, indicating local underdetermination. PredRNN then provides short-horizon temporal densification: SSIM and PSNR remain high initially and decay as nonlinear roll-up accumulates.

Table 2. Representative quantitative results across the main study stages.

Stage and metric	Reported outcome
Restoration mean SSIM	Degraded input 0.837, Restormer 0.905, and CLARIPHY 0.962 across 57 released evaluation cases.
Restoration mean PSNR	Degraded input 19.48 dB, Restormer 20.67 dB, and CLARIPHY 22.05 dB across the same 57 released cases.
Later released restoration cases	Over the final 12 released cases, SSIM values of 0.647, 0.782, and 0.916 and PSNR values of 15.21, 16.30, and 18.07 dB for degraded input, Restormer, and CLARIPHY, respectively.
Retrieval on complex RTI queries	For later-time queries, the first returned neighbors cluster around an average Atwood-number normalized label-space error of about 0.25 for InfoGAN versus about 0.4 for LIRE.
Representative state regression	pReg predicts $\mathscr{A} = 0.1482$ and $t = 0.4050$ s for a reference state with $\mathscr{A} = 0.155$ and $t = 0.400$ s.
Sparse-view 4D reconstruction	Symmetric regime: about 35 dB held-out-view PSNR after 20,000 iterations. Symmetry-broken regime: about 35 dB after 100,500 iterations.
Single-view latent adaptation	Representative progression from approximately PSNR/SSIM $20/0.6$ to $30/0.9$ during test-time latent refinement.

A sparse-view neural field produces candidate 4D reconstructions from sparse projections

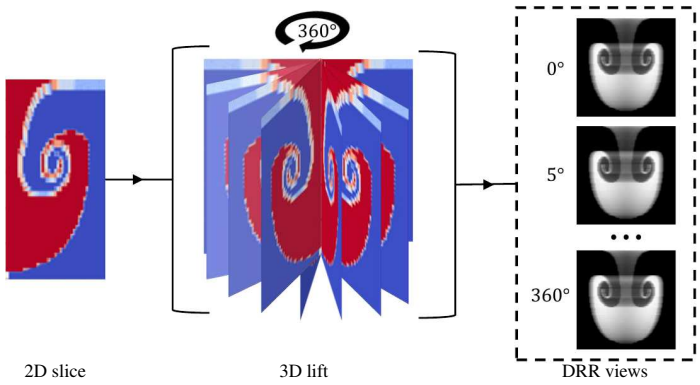
The same simulation family also supports a reproducible sparse-view 4D reconstruction setting for prior-guided volumetric inference. Each 2D RTI slice is revolved into a dynamic 3D attenuation volume, optionally perturbed to break perfect symmetry, and rendered into 72 synthetic radiographs over 360°. This produces 54 dynamic 3D sequences that retain their original RTI labels while exposing the inverse problem faced in dynamic sparse-view tomography.

A MedNeRF-derived time-conditioned model is used for this benchmark.[23,33] In the symmetric regime, it reaches about 35 dB PSNR on held-out views after 20,000 training iterations, corresponding to roughly 5 h on a Tesla V100 GPU. In the symmetry-broken regime, convergence requires 100,500 iterations, or about 30 h, to reach a similar held-out-view PSNR. Figure 5b shows that held-out agreement is strongest near 0° and 360°, while fidelity declines around 180°–270°, where angular ambiguity is largest. Figure 5c shows representative single-view latent-adaptation stages, in which PSNR/SSIM improve from about $20/0.6$ to $30/0.9$ as boundaries sharpen and morphology becomes more coherent. The main failure mode is angular ambiguity once symmetry is broken. The reconstructions are therefore projection-consistent candidates, not unique recovered volumes from a single projection.

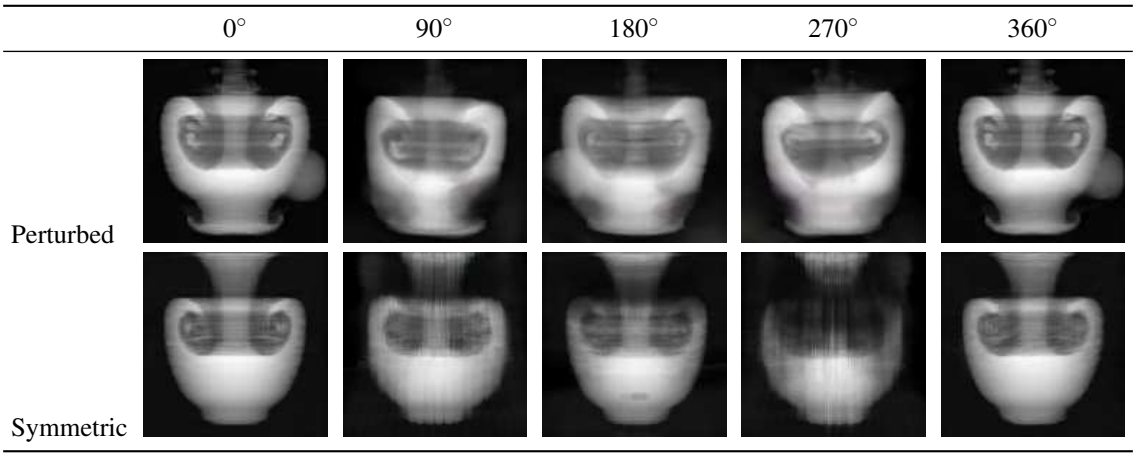
What the shared labels reveal

The common label set exposes different levels of constraint across the inverse problem. Restoration can recover contours that remain close to the known central frame, especially in morphologically complex cases where degradation merges spikes and roll-up boundaries. Retrieval and regression show that morphology carries usable information about $\mathscr{A}$ and t, but the local-matching results also show that some image regions are not globally diagnostic.

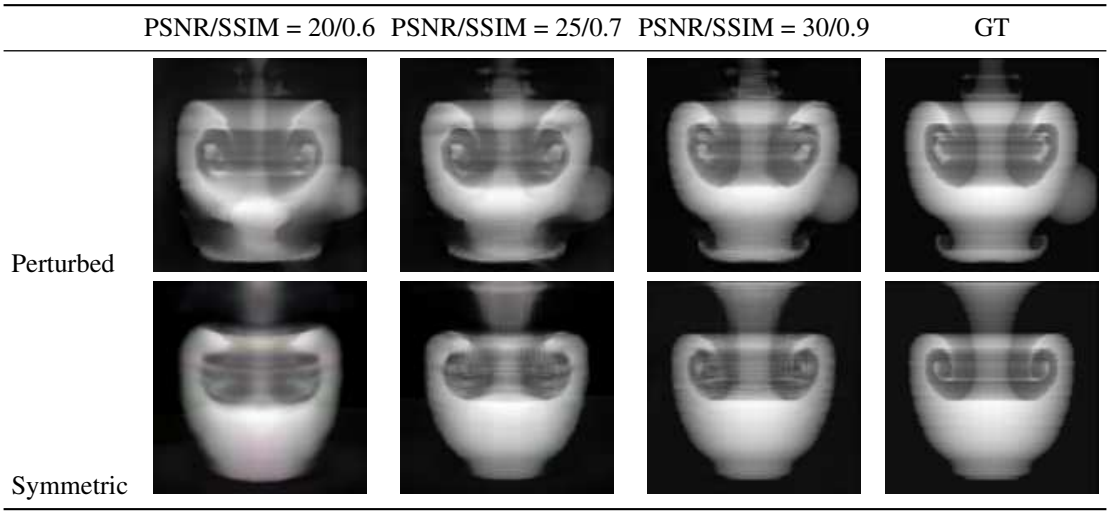
The same distinction carries into the 4D benchmark. Held-out-view PSNR tests whether a reconstructed field remains consistent with unseen synthetic radiographs, but it does not resolve volumetric uniqueness once angular ambiguity grows. These comparisons do not validate the governing RTI equations directly. They test agreement with assigned simulation parameters or with held-out synthetic radiographs. That distinction matters most where acceleration-time degeneracy, finite-domain effects, or angular scarcity reduce uniqueness. Table 2 summarizes the strongest quantitative outcomes across the study.



(a) 2D slice, axisymmetric lift, and DRR generation.



(b) Angular self-consistency of representative single-view reconstructions.



(c) Representative latent-adaptation stages for single-view inference.

Figure 5. Sparse-view neural fields produce candidate 4D reconstructions consistent with synthetic radiographs. Labeled 2D RTI slices are lifted into dynamic 3D attenuation volumes, rendered into synthetic radiographs, and reconstructed with a time-conditioned neural field. Held-out-view fidelity remains strongest for the symmetric family, while controlled symmetry breaking exposes the cost of underdetermination. The single-view panels show representative latent-adaptation stages, not averaged performance, and they should be interpreted as prior-guided refinements of candidate 3D structure.

Discussion

The same RTI labels are used in the degradation model, restoration targets, label-space tests, ambiguity analysis, and synthetic 4D benchmark. That makes it possible to compare image fidelity with agreement in the assigned simulation labels. A restored contour can have high SSIM and still move away from the correct region of label space, and a sparse-view volume can match held-out projections without being uniquely determined in three dimensions.

The strongest evidence in the present benchmark comes from the restoration and label-space retrieval results. RTI-specific fine-tuning preserves bubble caps, spikes, and roll-up more faithfully than the generic restoration baseline in morphologically complex cases. The learned embedding also improves early-rank Atwood-number agreement relative to the hand-crafted retrieval control. The ambiguity analysis is informative for a different reason: distinctive templates identify local regions that are diagnostically useful, whereas non-distinctive templates show where local appearance alone cannot support global state inference. The sparse-view benchmark is the weakest part of the chain in terms of uniqueness. Once symmetry is broken, the main difficulty is angular ambiguity, so held-out-view agreement is best interpreted as projection consistency rather than direct recovery of a single 3D state.[14,21,27,34]

Several limits on interpretation are important. The degradation model is controlled and not detector calibrated. The raw label tuple $(\mathscr{A}, g, h, t)$ is not a set of four independent morphology axes because acceleration and time partly collapse under normalized time. Some of the latest benchmark frames are influenced by the finite domain and reflecting boundaries, so they remain most useful as controlled image-processing cases rather than as unconstrained late-time RTI states. The 57 released restoration cases are time ordered benchmark examples rather than documented simulation-disjoint test averages, which means that the reduced-data result does not establish generalization to unseen trajectories. pReg resolves only the $\mathscr{A}$ and t axes of the full label set, PredRNN is limited to short-horizon densification, the 2D-to-3D lift is synthetic and does not replace direct 3D hydrodynamics, and the sparse-view neural-field outputs are prior-guided candidate reconstructions rather than unique volumetric solutions.

The evidence is therefore strongest for RTI as a benchmark for dynamic two-material interfaces, not as a fully validated route to real 3D experimental inversion. Extending the approach to calibrated dynamic radiographs, direct 3D simulations, uncertainty-aware inversion, domain adaptation, and additional interface families such as Richtmyer–Meshkov or shock-driven mixing will be essential for broader use.[4,11–13] The value of the present study is that those future models can be tested against the same labeled RTI benchmark rather than against disconnected task-specific criteria.

Methods

The benchmark combines three released resources. *RayleAI* provides labeled 2D RTI interface images, CLARIPHY provides controlled sharp/degraded pairs for restoration, and *SDT* provides lifted dynamic 3D volumes with synthetic projections. The following sections describe how each resource was used and how the reported metrics were computed.

RTI simulation corpus and assigned state labels

The study uses single-mode RTI simulations inspired by the low-Atwood single-mode experiment of Waddell *et al.*[5,17] The initial interface is sinusoidal, $y = h\cos(2\pi x/\lambda)$, and is evolved with the DAFNA interface-tracking hydrodynamic code on a finite rectangular domain with reflecting boundary conditions in all directions.[17,35] The grid contains 64×128 cells over a 2.7 cm horizontal by 5.4 cm vertical domain.[17] In the single-mode setup, the 2.7 cm horizontal extent corresponds to half of the imposed wavelength, so $\lambda = 5.4$ cm for the normalized-time definition used below. The archived setup uses an ideal-gas equation of state with adiabatic index $\gamma = 5/3$ and hydrostatic equilibrium with 1 bar at the interface. The swept control parameters are Atwood number $\mathscr{A}$, gravitational acceleration g, initial perturbation amplitude h, and time t.

The public *RayleAI* release spans $\mathscr{A} = 0.02$ to 0.50 in steps of 0.02, $g = 600$ to 800 cm s^{-2} in steps of 25 cm s^{-2}, and $h = 0.1$ to 0.6 cm in steps of 0.1 cm.[17] This yields 1,350 simulations, each sampled at 75 saved times, for a total of 101,250 thresholded interface images. The public file-conversion and sequence-processing scripts use the same saved-time grid for all simulations, $t = 0.00, 0.01, \ldots, 0.74$ s.[17] Following Ref.[17], we define normalized time as

$$\tilde{t} = \sqrt{\frac{\mathscr{A} g}{\lambda}}\, t.$$

Over the full corpus, $\tilde{t}$ ranges from 0 to about 6.37. Because acceleration and raw time partly collapse under this rescaling, distances in $(\mathscr{A}, g, h, t)$ are not equivalent to distances between morphologically distinct states. Each image inherits the labels of its parent simulation and time index, so the corpus can be navigated either in image space or in the assigned coordinates $(\mathscr{A}, g, h, t)$. At late times, reflected disturbances or direct interaction with the finite boundaries can alter the interface morphology. These frames remain part of the released benchmark and are useful for controlled image-processing tests, but they are interpreted cautiously when discussing late-time RTI evolution. Experimental images used for qualitative

alignment were derived from grayscale radiographs by expert-selected preprocessing, including erosion–dilation or histogram equalization, cropping, resizing, and binarization.[17]

Controlled degradation and restoration module

Let I_t denote the sharp thresholded interface image at time t. A temporally blurred surrogate is formed as

$$\bar{I}_t = \frac{1}{|W_t|} \sum_{\tau \in W_t} I_\tau,$$

where W_t is a centered window of size 3, 5, 7, 9, or 13 frames. Spatial blur is then applied by convolving $\bar{I}_t$ with a normalized spherical kernel K_r of radius $r \in \{3,5,7,9\}$ pixels,

$$B_t = K_r * \bar{I}_t.$$

Because the saved-frame spacing is 0.01 s, the first-to-last support of the temporal windows is 0.02, 0.04, 0.06, 0.08, and 0.12 s for the 3, 5, 7, 9, and 13 frame averages, respectively. Under the normalized-time scaling above, the corresponding spans range from about 0.03 to 0.17 for the 3-frame window and from about 0.18 to 1.03 for the 13-frame window across the full $(\mathscr{A}, g)$ range. Public CLARIPHY preprocessing applies the blur after cropping the RTI images to 176×80 pixels. Relative to the 2.7 cm horizontal and 5.4 cm vertical domain extents, this corresponds to about 0.0338 cm per pixel horizontally and 0.0307 cm per pixel vertically. The 3, 5, 7, and 9 pixel radii therefore correspond to approximately (0.10, 0.09), (0.17, 0.15), (0.24, 0.21), and (0.30, 0.28) cm along the horizontal and vertical image axes. Because the restoration target remains the sharp central frame I_t, every degraded surrogate B_t retains the central-state labels $(\mathscr{A}, g, h, t)$ while emulating motion integration and finite spatial resolution.[28] This controlled surrogate observation model emulates two dominant degradation mechanisms, but it is not a detector-calibrated radiographic forward model.

Restoration uses Restormer as the baseline backbone and fine-tunes it on the RTI blur pairs to obtain CLARIPHY.[19,28] Inputs and targets remain at the native RTI image resolution. Training uses 30 epochs, batch size 8, the Adam optimizer, and an L1 loss in PyTorch on a 40 GB NVIDIA A100 GPU. In the public database-construction script, complete simulation folders are shuffled and assigned to an 80/20 train/validation split, which keeps all temporal and spatial degradation variants of a trajectory in the same split. The public release does not expose the exact shuffle seed or the manifest used for the 57 reported evaluation traces. A reduced-data variant is trained on 10% of the paired corpus to test sensitivity to dataset size, but the public materials describe two different reduction paths: a stride-based reduced database generator and an additional random 10% subsampling step inside the training script. In the released `train.py`, the validation subset is also sized from the already reduced training-set length rather than from the full validation-set length. The released 10% curves therefore do not come with a recoverable grouped manifest. The main-text summary statistics are recomputed directly from the 57 released SSIM and PSNR traces, which the public manuscript and metric files present as time-ordered validation examples of increasing complexity. The results therefore characterize performance across increasingly difficult released cases and should not be interpreted as independent-by-simulation averages.

Retrieval controls, coordinate regression, and normalized label-space error

Retrieval compares an InfoGAN-based embedding with LIRE, a classical content-based retrieval library using hand-crafted texture descriptors.[17,29,30] InfoGAN serves as a compact learned morphology embedding, and LIRE provides a simple appearance-level control. The public evaluation uses 13,000 random queries and a second set of about 10,000 complex images with $t > 0.25$ s. Returned candidates are ranked by model score, then evaluated with a normalized simulation-label error, called physical loss in Ref.[17],

$$L^{(S)}_{\mathrm{label}}(i) = \frac{1}{|S|} \sum_{k \in S} \left(\frac{y_k - y_{i,k}}{r_k} \right)^2,$$

where y and y_i are the label vectors of the query and the ith candidate, r_k is the sampled range of coordinate k, and S is the subset of coordinates under evaluation.[17] For the long-time retrieval plots, the principal reported coordinate is Atwood number, which is the most discriminative label in the released comparison figures. Because $\mathscr{A}$, g, h, and t do not all affect morphology independently, this metric is always interpreted with respect to the chosen subset of labels S. In particular, the long-time retrieval plots emphasize Atwood-number error because acceleration and time are partly degenerate under $\tilde{t}$, and differences in initial perturbation amplitude can become time-shift-like in late evolution.[17]

Direct regression uses pReg, with three 5×5 convolutional layers of 64 filters each, max-pooling, L2 regularization, dropout at rate 0.1, and batch normalization, followed by fully connected layers of 250 and 200 units and a two-output sigmoid head for $\mathscr{A}$ and t.[17] Training uses a reduced RTI subset with $\mathscr{A} \in [0.08, 0.50]$ in steps of 0.02, $g \in \{625, 700, 750, 800\}$ cm s^{-2},

$h \in [0.1, 0.5]$ cm in steps of 0.1 cm, and $t \in [0.1, 0.6]$ s in steps of 0.01 s, giving about 16,000 training images and 4,000 validation images. Augmentation is restricted to zooming with range 0.1 so that orientation and topology remain physically meaningful. Candidate states are then ordered by distance in the regressed $(\mathscr{A}, t)$ space, so the method probes which coordinates can be recovered from morphology without attempting full-state inference.

Ambiguity diagnostics and temporal densification

Local morphology matching uses QATM in its 1-to-1 configuration on 16 templates cropped from experimental RTI images.[17,31] For each template, the best-matched subfigure in every *RayleAI* image is scored and ranked. Because local templates do not map uniquely to global labels, evaluation is performed by projecting the returned set with principal-component analysis and clustering it with k-means. This makes the module an ambiguity diagnostic: it separates templates with distinctive, partially distinctive, and non-distinctive physical neighborhoods without forcing a single global interpretation.

Temporal completion uses PredRNN as a short-horizon sequence-densification module.[32] The model is trained on *RayleAI* sequences sampled every 0.01 s. The experiment-like demonstration uses a 12-frame binarized RTI sequence with approximately 0.033 s spacing between frames; missing intermediate states are generated iteratively, one step at a time. A representative simulation test uses $g = 725$ cm s^{-2}, $\mathscr{A} = 0.14$, and $h = 0.3$ cm, with the first 10 frames given as input and 49 future steps predicted. The intended use is short-gap densification; long-range forecasting of nonlinear RTI lies outside the scope of this benchmark.

2D-to-3D lifting and sparse-radiograph generation

Each 2D RTI frame $S_t(x,z)$ is converted to a 3D attenuation volume $V_t(x,y,z)$ by solid of revolution about the symmetry axis.[33] Straight revolutions define the symmetric object family. A harder symmetry-broken family is created by perturbing the lifted volume with localized inhomogeneities, such as embedded high-density spheres or local density distortions, while preserving the parent RTI trajectory.

This 2D-to-3D lift is a controlled benchmark construction and not a direct 3D hydrodynamic simulation. Its purpose is to create a reproducible sparse-view inverse problem whose hidden state remains tied to the RTI labels used in the 2D tasks.

Synthetic radiographs are rendered by line integration through the attenuation field,

$$p_\theta(u,v) = \int \mu(x,y,z)\, dl,$$

for 72 uniformly spaced angles covering 360°. The released benchmark contains 54 dynamic 3D sequences and 72 radiographs per volume. These views can be subsampled to emulate few-view settings such as 8, 4, or 2 inputs, corresponding to angular separations of about 45°, 90°, and 180° for uniformly spaced selections, as well as single-view inference, while retaining held-out angles for evaluation.[33] The dataset description also defines broader split families, including held-out sequences, held-out late times, and view-sparsity tests; the results summarized here use the released held-out-view and single-view evaluations for the symmetric and symmetry-broken regimes.

Time-conditioned neural-field reconstruction

Sparse-view reconstruction extends MedNeRF by introducing explicit time conditioning, so the model represents a continuous field $F_\Theta(x,y,z,t)$ over space and time.[23,33] Separate latent codes encode sequence-specific shape and appearance, following the MedNeRF/GRAF design. Training uses projection-only supervision and optimizes a combination of mean squared error and structural similarity losses on rendered versus target synthetic radiographs.[23]

The public SDT benchmark reports training on a single NVIDIA Tesla V100 GPU. The symmetric regime converges after 20,000 iterations, corresponding to about 5 h, and the symmetry-broken regime converges after 100,500 iterations, or about 30 h. Held-out-view PSNR is reported for both regimes. For single-view inference, the network weights are frozen and the latent codes are adapted to the observed radiograph for a few hundred iterations, yielding the progressive PSNR/SSIM improvements shown in Figure 5. The adapted outputs are interpreted as prior-guided dynamic reconstructions constrained by the observation and by the simulation-trained field.

Evaluation and reporting

Restoration and temporal completion are evaluated with structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) against the sharp target or held-out frame.[36] Retrieval and regression are evaluated with L_{label} on selected subsets of the assigned simulation labels. Sparse-view 4D reconstruction is evaluated by PSNR on held-out radiographs together with qualitative inspection of recovered morphology across time and angle. These metrics are stage specific and are not compared directly: SSIM and PSNR test image fidelity, L_{label} tests agreement with the assigned labels, and held-out-view PSNR tests consistency with unseen radiographs.

The quantitative values reported in the main text were taken from the publicly released benchmark figures and metric arrays associated with each task. Restoration means were recomputed from the 57 time-ordered traces distributed with CLARIPHY, while retrieval, regression, PredRNN, and SDT values follow the public evaluation figures and accompanying manuscript descriptions.[17,28,33]

Data availability

The public resources used in this study are available from the following repositories and archives. *RayleAI* RTI corpus assets and experiment-alignment images: https://github.com/Scientific-Computing-Lab/SimulAI. CLARIPHY blurred/sharp RTI pairs, restoration metric traces, and representative restored outputs: https://github.com/Scientific-Computing-Lab/ClariPhy; archive DOI: https://doi.org/10.5281/zenodo.13385382. *SDT* lifted volumes, representative synthetic radiographs, reconstruction assets, and benchmark figures: https://github.com/Scientific-Computing-Lab/SDT; archive DOI: https://doi.org/10.5281/zenodo.16996679. Derived manuscript figures use only these public materials and their released metric arrays.

Code availability

Code for RTI simulation benchmarking, restoration training, sparse-view reconstruction, and the metric/plotting scripts used for the reported curves is available in the same public repositories. The SimulAI repository contains the RayleAI benchmarking and experiment-alignment workflow, the ClariPhy repository contains restoration training and evaluation code, and the SDT repository contains sparse-view reconstruction and rendering utilities.

References

1. Taylor, G. I. The instability of liquid surfaces when accelerated in a direction perpendicular to their planes. i. *Proc. Royal Soc. Lond. Ser. A. Math. Phys. Sci.* **201**, 192–196 (1950). DOI 10.1098/rspa.1950.0052.
2. Sharp, D. H. An overview of rayleigh–taylor instability. *Phys. D: Nonlinear Phenom.* **12**, 3–18 (1984). DOI 10.1016/0167-2789(84)90510-4.
3. Kull, H. J. Theory of the rayleigh–taylor instability. *Phys. Reports* **206**, 197–325 (1991). DOI 10.1016/0370-1573(91)90153-D.
4. Zhou, Y. *et al.* Rayleigh–taylor and richtmyer–meshkov instabilities: A journey through scales. *Phys. D: Nonlinear Phenom.* **423**, 132838 (2021). DOI 10.1016/j.physd.2020.132838.
5. Waddell, J. T., Niederhaus, C. E. & Jacobs, J. W. Experimental study of rayleigh–taylor instability: Low atwood number liquid systems with single-mode initial perturbations. *Phys. Fluids* **13**, 1263–1273 (2001). DOI 10.1063/1.1359762.
6. Angulo, A. M. *et al.* Design of a high-resolution rayleigh–taylor experiment with the crystal backlighter imager on the national ignition facility. *J. Instrumentation* **17**, P02025 (2022). DOI 10.1088/1748-0221/17/02/P02025.
7. Drake, R. P. High-energy-density physics. *Phys. Today* **63**, 28–33 (2010). DOI 10.1063/1.3455249.
8. Montgomery, D. S. Invited article: X-ray phase contrast imaging in inertial confinement fusion and high energy density research. *Rev. Sci. Instruments* **94**, 021103 (2023). DOI 10.1063/5.0127497.
9. Huntington, C. M. *et al.* Split radiographic tracer technique to measure the full width of a high energy density mixing layer. *High Energy Density Phys.* **35**, 100733 (2020). DOI 10.1016/j.hedp.2019.100733.
10. Heyndrickx, M., Bultreys, T., Goethals, W., Van Hoorebeke, L. & Boone, M. N. Improving image quality in fast, time-resolved micro-ct by weighted back projection. *Sci. Reports* **10**, 18029 (2020). DOI 10.1038/s41598-020-74827-x.
11. García-Moreno, F. *et al.* Tomoscopy: Time-resolved tomography for dynamic processes in materials. *Adv. Mater.* **33**, 2104659 (2021). DOI 10.1002/adma.202104659.
12. Zheng, Y. & Hatzell, K. B. Ultrasparse view x-ray computed tomography for 4d imaging. *ACS Appl. Mater. & Interfaces* **15**, 35024–35033 (2023). DOI 10.1021/acsami.3c06291.
13. Sauppe, J. P. *et al.* On the importance of three-dimensional modeling for high-energy-density physics experiments. *Phys. Plasmas* **30**, 062707 (2023). DOI 10.1063/5.0144517.
14. Arridge, S., Maass, P., Öktem, O. & Schönlieb, C.-B. Solving inverse problems using data-driven models. *Acta Numer.* **28**, 1–174 (2019). DOI 10.1017/S0962492919000059.

15. Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021). DOI 10.1038/s42254-021-00314-5.

16. Brunton, S. L., Noack, B. R. & Koumoutsakos, P. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52**, 477–508 (2020). DOI 10.1146/annurev-fluid-010719-060214.

17. Harel, R., Rusanovsky, M., Fridman, Y., Shimony, A. & Oren, G. Complete deep computer-vision methodology for investigating hydrodynamic instabilities. In *High Performance Computing*, Lecture Notes in Computer Science, 61–80 (Springer, 2020). DOI 10.1007/978-3-030-59851-8_5.

18. Nah, S., Kim, T. H. & Lee, K. M. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3883–3891 (2017). DOI 10.1109/CVPR.2017.35.

19. Zamir, S. W. *et al.* Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5728–5739 (2022). DOI 10.1109/CVPR52688.2022.00564.

20. Levesque, J. M., Merritt, E. C., Flippo, K. A., Rasmus, A. M. & Doss, F. W. Neural network denoising of x-ray images from high-energy-density experiments. *Rev. Sci. Instruments* **95**, 063508 (2024). DOI 10.1063/5.0207005.

21. Wang, G., Ye, J. C. & De Man, B. Deep learning for tomographic image reconstruction. *Nat. Mach. Intell.* **2**, 737–748 (2020). DOI 10.1038/s42256-020-00273-z.

22. Wolfe, B. T. *et al.* Neural network for 3d inertial confinement fusion shell reconstruction from single radiographs. *Rev. Sci. Instruments* **92**, 033506 (2021). DOI 10.1063/5.0043653.

23. Corona-Figueroa, A. *et al.* MedNeRF: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3843–3848 (2022). DOI 10.1109/EMBC48229.2022.9871757.

24. Zha, R., Zhang, Y. & Li, H. NAF: Neural attenuation fields for sparse-view CBCT reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, 442–452 (Springer, 2022). DOI 10.1007/978-3-031-16446-0_42.

25. Rückert, D., Wang, Y., Li, R., Idoughi, R. & Heidrich, W. NeAT: Neural adaptive tomography. *ACM Transactions on Graph.* **41**, 1–13 (2022). DOI 10.1145/3528223.3530121.

26. Cai, Y., Wang, J., Yuille, A., Zhou, Z. & Wang, A. Structure-aware sparse-view x-ray 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11174–11183 (2024). DOI 10.1109/CVPR52733.2024.01062.

27. Grega, I., Whitney, W. F. & Deshpande, V. S. Neural rendering enables dynamic tomography. *arXiv* (2024). URL https://arxiv.org/abs/2410.20558. 2410.20558.

28. Stamler-Grossman, S., Schneider, N., Hanoch, G. & Oren, G. CLARIPHY: Physics-informed image deblurring with transformers for hydrodynamic instability analysis. Machine Learning and the Physical Sciences Workshop, NeurIPS 2024 (2024). URL https://ml4physicalsciences.github.io/2024/files/NeurIPS_ML4PS_2024_19.pdf.

29. Chen, X. *et al.* InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, vol. 29, 2172–2180 (2016).

30. Lux, M. & Chatzichristofis, S. A. LIRE: Lucene image retrieval. In *Proceedings of the 16th ACM International Conference on Multimedia*, 1085–1088 (2008). DOI 10.1145/1459359.1459577.

31. Cheng, J., Wu, Y., AbdAlmageed, W. & Natarajan, P. QATM: Quality-aware template matching for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11553–11562 (2019). DOI 10.1109/CVPR.2019.01182.

32. Wang, Y., Long, M., Wang, J., Gao, Z. & Yu, P. S. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*, vol. 30, 879–888 (2017).

33. Schneider, N., Hanoch, G. & Oren, G. Learning 4d material-interface dynamics from few x-ray projections. AI for Accelerated Materials Discovery (AI4Mat) Workshop at ICLR 2026 manuscript (2026). URL https://openreview.net/pdf?id=MJXl8k46qr.

34. Barbano, R., Zhang, C., Arridge, S. & Jin, B. Quantifying model uncertainty in inverse problems via bayesian deep gradient descent. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 1392–1399 (2021). DOI 10.1109/ICPR48806.2021.9412521.

35. Klein, Y. Construction of a multidimensional parallel adaptive mesh refinement special relativistic hydrodynamics code for astrophysical applications. *Master's Thesis* (2010).

36. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Process.* **13**, 600–612 (2004). DOI 10.1109/TIP.2003.819861.

Acknowledgements

This work was supported by the Pazy Foundation and by the Lynn and William Frankel Center for Computer Science. Computational resources were provided by NegevHPC.

Author contributions statement

G.O. conceived and supervised the study. R.H., M.R., Y.F., A.S., and G.O. developed the RTI simulation benchmark and physical-state inference analyses. S.S.-G., N.S., G.H., and G.O. developed the restoration and sparse-view reconstruction components. All authors analyzed the results, revised the manuscript, and approved the final version.

Additional information

Parts of the benchmark resources analyzed here were previously described in Refs.[17,28,33]. The present article contributes the integrated RTI-labeled analysis reported here.

Competing interests: The authors declare no competing interests.