DOES BETTER REPROJECTION MEAN BETTER 3D RECONSTRUCTION? CANDIDATE- AND EVIDENCE-STRATIFIED EVALUATION FOR FLFM

Anonymous authors
Paper under double-blind review

ABSTRACT

In Fourier light-field microscopy (FLFM), does better reproduction of the camera image imply more faithful recovery of three-dimensional fluorescence structure? Reprojection fit supports selection when paired volume references are unavailable. We introduce SCOPE-FLFM to examine what this criterion reveals. Comparable-Candidate Stratification (CCS) organizes comparisons across model families, related architectures, and states from the same training run. Contract-Separated Evidence Validation (CEV) distinguishes supplied-view fit, excluded-view prediction, and normalized agreement with known synthetic objects. Using a synthetic FLFM corpus and reference volumes spanning beads, clusters, filaments, and mixtures, we find that pooled rank agreement weakens among close candidates and prospective choices depend on the scored target. Under the tested additive background, max-normalized reprojection SSIM increases while object Pearson decreases across all core learned states. A matched HYPER-derived Calibra-LF case study improves clean-corpus held-out SSIM and object Pearson on average while worsening other endpoints. SCOPE-FLFM links each score to the imaging property that model selection aims to improve.

1 INTRODUCTION

When two FLFM models explain the same camera image, which recovers the three-dimensional fluorescence structure more faithfully? Light-field microscopy (LFM) encodes spatial and angular information for volumetric reconstruction (Levoy et al., 2006; Broxton et al., 2013), enabling whole-animal neural imaging with image-plane sampling (Prevedel et al., 2014). Fourier-plane recording and integral microscopy introduced pupil sampling (Llavador et al., 2016; Scrofani et al., 2018). Fourier light-field microscopy (FLFM) uses a Fourier-conjugate microlens array (Guo et al., 2019a); XLFM uses a distinct pupil-sampling configuration (Cong et al., 2017). Figure 1 shows two-color FLFM. HR-FLFM visualizes lipid nanoparticle-mediated delivery in lung and heart cells (Radmand et al., 2024) and rhesus CD34$^+$ and liver cells (Kim et al., 2025). FLFM tracks calcium transients and displacement in cardiac spheroids after spaceflight (Forghani et al., 2025); a medRxiv preprint assesses live tadpole heart function after Ets1 knockdown (Wang et al., 2026).

The HYPER-Net preprint uses point-spread-function (PSF) conditioning and reprojection training (Ling et al., 2026a), enabling selection without paired volumes. How well does reprojection fit rank close candidates, and does that ranking follow recovered structure? Different fluorescence distributions in depth can produce similar projections. A score can also separate model families while changing its ordering among nearby checkpoints. Model choice therefore depends on both the candidates compared and the property evaluated.

We introduce SCOPE-FLFM, Stratified Contract-Oriented Performance Evaluation for FLFM, an evaluation framework instantiated as a synthetic benchmark. Comparable-Candidate Stratification (CCS) defines comparisons by family, training run, and deduplicated run–epoch state, retaining checkpoint provenance. Contract-Separated Evidence Validation (CEV) evaluates supplied-view fit, excluded-view prediction, and normalized fidelity to known synthetic objects, linking each score to a reconstruction property. Our experiments show pooled rank agreement weakening among close HYPER-derived candidates and prospective choices depending on the scored target. Additive back-

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.

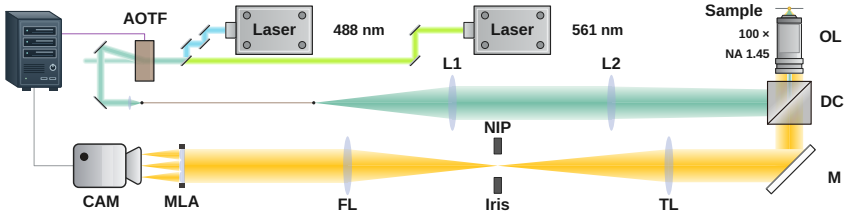


Figure 1: **Two-color FLFM imaging setup.** The system uses 488-nm and 561-nm lasers to excite fluorescence in the sample. The emitted light is collected by the objective lens (OL) and focused by the tube lens (TL) to form an intermediate image at the native image plane (NIP). The Fourier lens (FL) and microlens array (MLA) then form a set of sub-images, each showing the sample from a different viewing direction. A scientific CMOS (sCMOS) camera records these views together in a single image, which constitutes the light-field measurement. DC, dichroic cube; M, mirror; CAM, camera.

ground raises max-normalized reprojection SSIM while object Pearson falls. The matched Calibra-LF case study identifies which properties a small model change improves and which it worsens. Together, these findings connect FLFM model choice to the image prediction or spatial structure needed for its intended use.

2 Related Work

Conference reports developed Fourier-domain and integral models (Guo et al., 2019b; Liu et al., 2019). Journal and conference studies explored cellular resolution through HR-LFM/HR-FLFM (Li et al., 2019; Hua & Jia, 2020; Hua et al., 2021), wide-field fusion through wFLFM (Liu & Jia, 2020; 2021; 2022), and hybrid-PSF organoid imaging (Liu et al., 2022). Radiality-based rad-FLFM/SRRF (Han et al., 2022) and fluctuation-based SOFFLFM (Huang et al., 2023) exploit temporal information; Fourier DiffuserScope (Linda Liu et al., 2020) changes pupil encoding. FOLIC uses a concave multiaperture light-field architecture for live-cell and tissue imaging (Huang et al., 2026). These systems' spatial, angular, and complementary measurements and PSFs determine the information available to reconstruction.

Applications span autofluorescence (Ling et al., 2023b;a), LFC/LFC-plus cytometry (Hua et al., 2024; Ling et al., 2026b), and optofluidic, flow-imaging, and dual-color conference reports (Hua et al., 2022; 2023a; Liu et al., 2025). ACsN corrects sCMOS noise for HR-LFM microtubule reconstruction (Mandracchia et al., 2020) and HR-FLFM rhesus-cell imaging (Kim et al., 2025). MIRO supports FLFM cardiac restoration (Mandracchia et al., 2023); denoising-decomposition conference work (Ling et al., 2023c) and multiscale recursive unmixing (Ling et al., 2025) address signal separation. GRIN-array and multimodal endoscopy (Guo et al., 2020; Urner et al., 2022; Zheng & Jia, 2026), darkfield (Scrofani et al., 2020), and photothermal chemical FLFM (Jia et al., 2023) broaden configurations and contrast. A review surveys biomedical light-field imaging (Zhao et al., 2025).

Reconstruction includes Richardson–Lucy (Richardson, 1972; Lucy, 1974), PEARL acceleration (Wu et al., 2025), patch-wise Wiener–Butterworth deconvolution (Wu et al., 2026), and Gaussian adaptive reconstruction (Xu et al., 2025). Learned VCD, XLFMNet, HyLFM, and VsLFM (Wang et al., 2021; Vizcaino et al., 2021; Wagner et al., 2021; Lu et al., 2023), Fourier VCD (Yi et al., 2023), and the DFLFM conference report (Hua et al., 2023b) use differing supervision and operators. Single-molecule light-field and vortex methods target localization (Sims et al., 2020; Zhang et al., 2024); VLM trains V-Net and L-Net on system responses to reconstruct volumes and localize molecules from Fourier-aperture measurements (Han et al., 2025). The PAVR preprint uses simulated system responses (Hua et al., 2026); the VSM preprint addresses scattering through aperture-segmented FLFM (Gao et al., 2026). These methods differ in training data, reconstruction priors, and output representations, motivating separate comparisons across architectures and among related models. Our XLFMNet-inspired control uses a local FLFM adaptation.

SeReNet uses physics-driven self-supervised learning for LFM reconstruction (Lu et al., 2025). Measurement-withholding objectives restrict target access (Batson & Royer, 2019; Krull et al., 2019;

Laine et al., 2019; Yaman et al., 2020). Held-out views test camera prediction under shared acquisition assumptions; known synthetic reference volumes assess recovered structure within the declared simulator. SSIM measures image similarity (Wang et al., 2004); microscopy artifact mapping supports reference-based structural checks (Culley et al., 2018). Validation overfitting (Cawley & Talbot, 2010), underspecification (D'Amour et al., 2022), and inverse crime (Wirgin, 2004) motivate SCOPE-FLFM's separation of selection, readout, and reference targets.

3 SCOPE-FLFM: TESTING A SELECTION CLAIM

SCOPE-FLFM tests reprojection-based selection within a declared set (Figure 2): *who is compared* and *what evidence is scored*.

3.1 CANDIDATE SCOPE

Comparable-Candidate Stratification (CCS) defines the population for each selection question.

We distinguish family (architecture and output), run (optimization trajectory), state (weights at an epoch), and file (the rule-selected record); files may duplicate weights, and states within a run retain trajectory dependence.

Strata answer distinct questions: all learned states test broad family separation; HYPER-derived states test ordering within one FLFM reconstruction setting; matched signed-HYPER/Calibra-LF pairs test a small model change. The latter two are diagnostic populations; prospective deployment uses separately fixed cohorts.

Runs, checkpoint rules, family labels, state keys, splits, and prospective selection were fixed before final readout. The selected epochs determine the 23 states; files document selection, states define rankings, runs define paired comparisons and hierarchical resampling, and families provide broader strata.

3.2 EVIDENCE SCOPE

Contract-Separated Evidence Validation (CEV) assigns each score to its input access and reference target.

Let $Y = \{Y_i\}_{i=1}^{9}$ denote a normalized 3×3 FLFM measurement and M_K retain supplied views K. The PSF-slope vector $s_h \in \mathbb{R}^{18}$ conditions the reconstructor F_θ, while the PSF tensor h defines the projector $\mathcal{P}_h$:

$$V^{(K)} = F_\theta(M_K Y, s_h), \qquad \widehat{Y}^{(K)} = \mathcal{P}_h(V^{(K)}), \tag{1}$$

Here, $\mathcal{P}_h$ accumulates the magnitudes of depth-block convolution sums (Equation A.2); reprojection scores assess consistency with this reconstruction forward model. For excluded indices H disjoint from K,

$$Q_{\text{obs}}(K) = s(\widehat{Y}^{(K)}, Y; K), \tag{2}$$

$$Q_{\text{hold}}(K, H) = s(\widehat{Y}^{(K)}, Y; H). \tag{3}$$

Q_{obs} scores supplied-view consistency; Q_{hold} tests specified excluded-view prediction under the same synthetic acquisition contract. Volume fidelity and transfer across acquisitions, operators, instruments, wavelengths, or specimens require corresponding references.

When a known synthetic volume $V^\star$ is available, we additionally define

$$Q_{\text{obj}} = q\Big(T_{\text{eval}}(V^{(K)}), T_{\text{eval}}(V^\star)\Big). \tag{4}$$

T_{eval} is fixed: each prediction and target is clamped at zero, then independently divided by its own volume-wide maximum. Object metrics assess the spatial distribution and support of this nonnegative normalized representation on the synthetic grid; Appendix A.24 details the scope.

Input frames use global max normalization. Measurement NMSE is MSE after per-view-max normalization; SSIM uses these normalized views and a uniform 7×7 window (Appendix A.25). Pearson and support overlap complement object intensity agreement with spatial comparisons. We

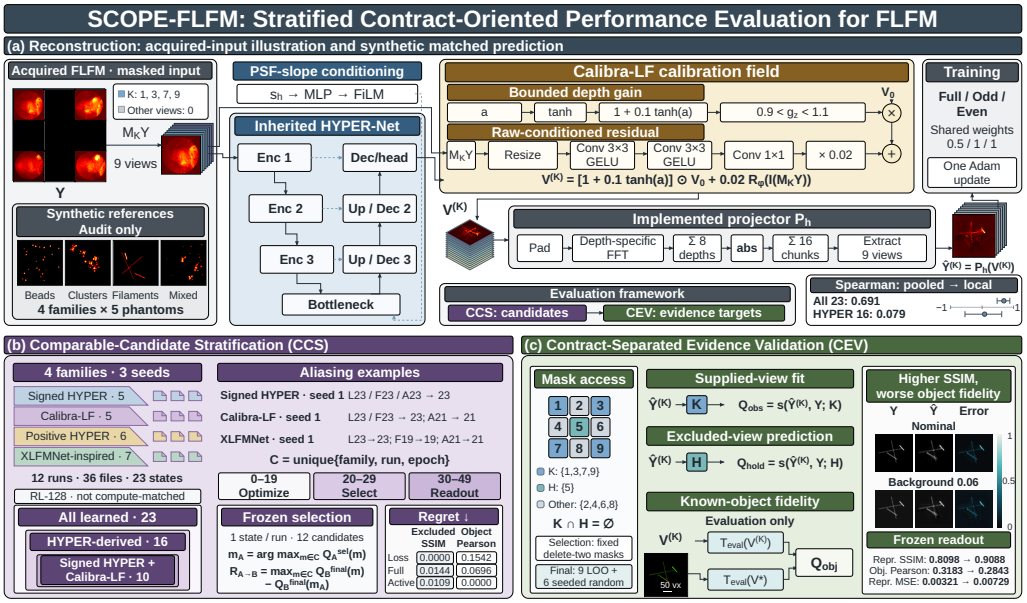


Figure 2: **SCOPE-FLFM: reconstruction, candidate scope, and evidence targets.** (a) The acquired nine-view input illustrates acquisition; the synthetic filament prediction and its reprojection use a separate synthetic input. Hot LUT encodes display intensity. Calibra-LF adds Bounded Depth Gain and Raw-Conditioned Residual after the inherited HYPER volume head, before the chunked projector; both paths respect the mask (Appendix A.25). (b) CCS maps 36 files from 12 runs to 23 states, retains rule aliases and run dependence, and freezes selection before final readout. Groups 0–19/20–29/30–49 serve optimization/selection/final evaluation. (c) CEV separates supplied-view fit (blue, $K = \{1, 3, 7, 9\}$), excluded-view prediction (teal, $H = \{5\}$), and normalized known-synthetic-object fidelity. The background example raises reprojection SSIM while object Pearson falls. Quantitative panels use synthetic evidence; the acquired input is not paired with the displayed volume.

preregister normalized full-volume Pearson and equal-budget top-k support overlap as primary object endpoints, retaining SSIM, NRMSE, IoU, F1, and axial diagnostics as non-interchangeable secondary views.

3.3 Prospective Selection and Diagnostic Ranking

For systems $m \in \mathcal{M}$, we report Spearman/Kendall correlation, top-k overlap, and discordant pairs. When both Q_A and Q_B are final endpoints,

$$\mathcal{R}^{\text{oracle}}_{A\to B} = \max_{m\in\mathcal{M}} Q_B(m) - Q_B\left(\arg\max_{m\in\mathcal{M}} Q_A(m)\right) \tag{5}$$

is an endpoint-oracle diagnostic, excluded from candidate selection. For each fixed rule-specific cohort $\mathcal{C}_r$, groups 20–29 select $m_A = \arg\max_{m\in\mathcal{C}_r} Q^{\text{sel}}_A(m)$ before final readout. We then compute

$$\mathcal{R}^{\text{pros}}_{A\to B} = \max_{m\in\mathcal{C}_r} Q^{\text{final}}_B(m) - Q^{\text{final}}_B(m_A). \tag{6}$$

For loss selection, Q^{sel}_A is negative selection loss. Zero regret matches the final endpoint optimum within that fixed cohort. Prospectivity concerns this rerun on historically accessible groups (Appendix A.2).

3.4 Data, Masks, Systems, and Controlled Tests

We use the released HYPER corpus of 500 synthetic FLFM measurements in 50 base groups with PSF/slope metadata (Ling et al., 2026a; Ling, 2026) for measurement evaluation and a separate

known-volume library for structural references. Groups 0–19 support optimization, 20–29 selection, and 30–49 final evaluation; scores aggregate by group. Appendix A.2 documents the released generation code (Shu Jia Lab, 2026), augmentations, and split diagnostics.

Training uses fixed odd/even partitions. Distinct deletion masks serve selection and final evaluation; final tests cover every omitted view and separately seeded random deletions (Appendix A.3). Corpus evaluation stays within odd/even software partitions, and frame-level wavelength provenance is absent. Monochrome known-volume phantoms generate all nine views from one object for same-channel angular prediction tests.

Candidates are signed HYPER, softplus-positive HYPER, Calibra-LF, and an XLFMNet-inspired local FLFM adaptation, with matched updates, three seeds, and loss/full/active checkpoint rules. Richardson–Lucy is an iterative reference selected by the full-view rule; budgets and the broader learned pool appear in Appendices A.4 and A.15.

HYPER-Net provides PSF conditioning, FiLM, self-supervised reprojection, and the FLFM formulation (Ling et al., 2026a). Calibra-LF adds identity-initialized Bounded Depth Gain and Raw-Conditioned Residual after the inherited volume head, retaining the projector, masks, targets, and selection rules. This bounded case study tests their joint effect within HYPER's model; Appendix A.25 specifies architecture, shared-weight training, and single-volume inference.

We evaluate measurement corruptions and 20 known monochrome synthetic objects generated for this study: sparse beads, clustered emitters, filaments, and mixtures. Controlled mismatch changes the data-generating PSF, background, noise, or throughput while reconstruction retains nominal conditioning. Appendices A.3, A.9, and A.11 specify stressors and distinguish structural from conditioning-only tests.

Figures 3–5 use nominal object-space sampling of $1.8\,\mu\mathrm{m}$ laterally and $3.2\,\mu\mathrm{m}$ axially from HYPER's released PSF code (Ling et al., 2026a; Shu Jia Lab, 2026); display conventions are in Appendix A.5.

4 Results

4.1 Rank Agreement Collapses Among Close Candidates

Pooled reprojection rankings show substantial agreement. Across 23 distinct learned states, supplied-view versus excluded-view Spearman is 0.691. To guide selection among close alternatives, this ordering should persist among candidates that share the same reconstruction setting.

CCS reveals a sharp loss of agreement among close candidates. The adapted XLFMNet control is separated from HYPER-derived states under both scores. Within the HYPER-derived stratum, Spearman falls to 0.079, with substantial pair reversals and limited top-ranked overlap; restricting to signed HYPER and Calibra-LF also leaves weak agreement (Appendix A.7). All 56 discordant pairs lie within the 120 HYPER-derived pairs; the remaining 133 pairs preserve their ordering. These counts explain how pooling raises agreement without resolving local ranking. The pattern persists with ten-augmentation group means: Spearman is 0.717 overall and 0.159 within the HYPER-derived stratum (Appendix A.13).

Checkpoint choice also changes final prediction within a training run. Full-view selection gives higher final excluded-view SSIM in eight of 12 runs, three tie, and one favors active-view selection; the mean full-minus-active difference is 0.0058 (Appendix A.7). These contrasts share an architecture and optimization trajectory. Deduplication prevents repeated files from giving one state extra weight; stratification tests ordering among the alternatives relevant to the decision.

Figure 3 gives the selection problem a spatial interpretation. Across dispersed, sparse, and dense punctate synthetic measurements, the early depth-calibrated variant displays more separated peaks and varied depth colors than HYPER. Common central crops expose changes in peak extent and diffuse signal. Depth hue identifies the reconstructed peak plane, making axial signal allocation visible. The paired SSIM/MSE plots quantify camera-image fit; separate known-volume comparisons in Figure 5 assess structural agreement with references. These historical models have unmatched weights and training histories, distinct from the matched Calibra-LF study below.

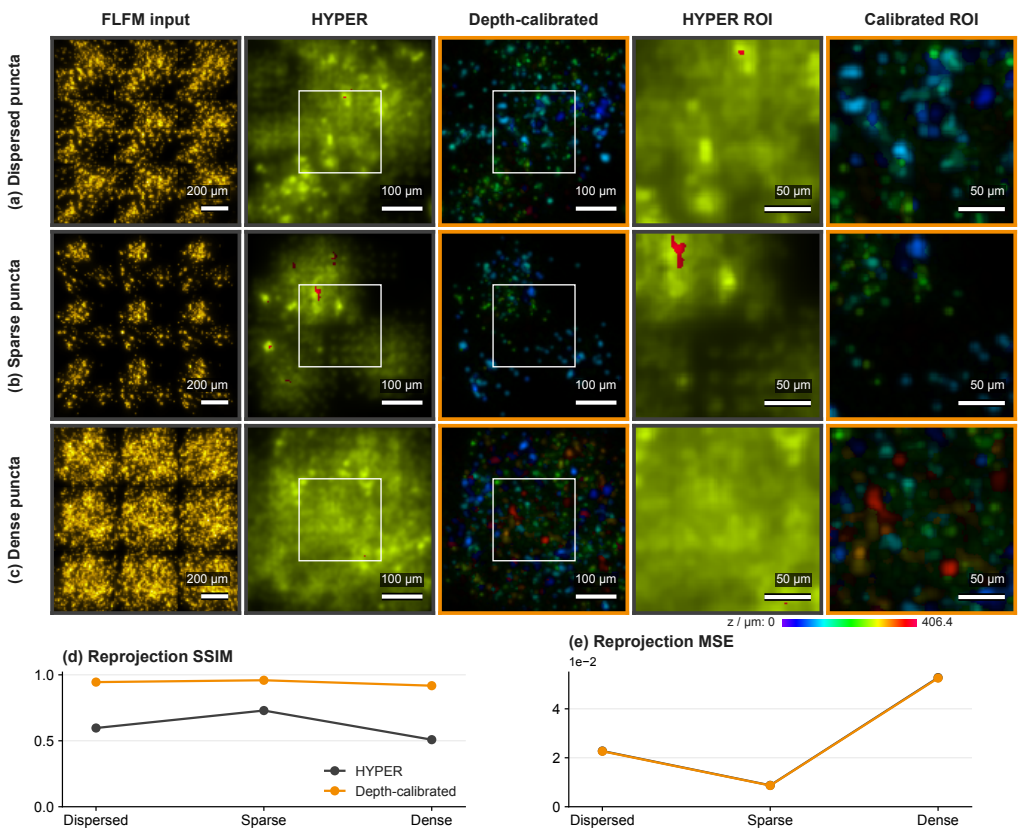


Figure 3: **Historical synthetic FLFM reconstructions and measurement fit.** Rows (a–c): dispersed, sparse, and dense puncta. Columns: nine-view input, HYPER reconstruction and depth-calibrated variant depth-coded XY maximum-intensity projections (MIPs), and central 112×112 ROIs. (d,e) Full-input reprojection SSIM and MSE; gray and orange denote HYPER and the depth-calibrated variant, respectively. Separated peaks and altered depth colors accompany score differences. Training is unmatched; the matched Calibra-LF comparison is separate. Scale bars: input 200 μm, full MIPs 100 μm, ROIs 50 μm.

4.2 Prospective Selection Still Depends on the Target

The selection target changes the model chosen before final readout. Rules fixed on groups 20–29 select candidates for evaluation on groups 30–49 and the known-volume phantoms. Loss selection chooses positive HYPER, with zero excluded-view regret and object-Pearson regret 0.1542. Active-view selection chooses Calibra-LF, with excluded-view regret 0.0109 and zero object-Pearson regret. Full-view selection chooses another Calibra-LF state, with excluded-view regret 0.0144 and object-Pearson regret 0.0696. Each value uses the optimum of its own fixed 12-candidate cohort (Appendix A.7), assessing a choice made before final evaluation.

CEV tests whether excluded-view scores better identify accurate objects. Both supplied-view and excluded-view SSIM correlate positively with object Pearson across all learned states, but this agreement weakens inside the HYPER-derived stratum and becomes negative within signed HYPER plus Calibra-LF (Appendix A.7). Withholding a view removes direct access to its image content and tests whether the remaining angular information predicts it. The target remains a projection of the same synthetic object through the same acquisition model. Depth distributions with similar projections can therefore receive similar measurement scores while differing in volume support. This distinction explains why stronger measurement testing still calls for separate structural evaluation.

Endpoint-specific optima reinforce this distinction. Within the reported synthetic table, selection-frozen RL-128 attains the highest normalized Pearson, equal-budget top-k support overlap, and

all-volume SSIM; positive HYPER has the lowest NRMSE and the highest SSIM among learned systems (Appendix A.8). Pearson measures centered volume agreement, whereas equal-budget support overlap asks whether the same number of high-intensity voxels occupies the reference support. NRMSE and volume SSIM add sensitivity to normalized amplitude and the large background region. Their differing preferences make the desired imaging property part of model choice.

Figure 5 shows where these distinctions arise in sparse beads and filaments. RL preserves separated bead locations and recognizable filament trajectories, while its brighter, broader support accompanies the less favorable NRMSE. The learned reconstructions recover portions of the lateral pattern but concentrate much of their signal into narrower depth bands in the XZ views. An XY projection can therefore retain recognizable geometry while hiding errors along the collapsed axis. Reference-selected slices expose whether signal is present at the same depth, and the full-volume readouts quantify agreement beyond the displayed plane. These different-seed exemplars support a spatial reading of the metrics; the population comparisons use the complete tables. The additional clustered and mixed phantoms extend this reading to emitter separation and combined structure (Appendices A.9 and A.17).

Calibra-LF shows how SCOPE-FLFM identifies the properties improved by a small model change. Matching signed HYPER by seed, split, update budget, and checkpoint rule yields six distinct state pairs nested within three seeds. Clean-corpus excluded-view SSIM and normalized object Pearson increase by 0.0043 and 0.0297 on average, each improving in five pairs; the clean-corpus SSIM interval crosses zero. Bounded Depth Gain adjusts depth-plane contributions; Raw-Conditioned Residual adds an input-dependent correction before the same projector. Known-volume excluded-view SSIM, object SSIM, NRMSE, and mean corruption performance worsen (Appendices A.19 and A.28, including intervals and full results). CCS supplies the matched comparison; CEV identifies the joint modification's endpoint-specific effects, without attributing them to either component alone.

4.3 A Measurement Score Can Rise While Known-Object Metrics Worsen

Controlled background provides a direct test of score improvement against object fidelity. With the synthetic objects and learned weights fixed, additive background at level 0.06 raises per-view-max-normalized full-reprojection SSIM and lowers normalized object Pearson in all 23 distinct learned states. This 23/23 reversal spans states nested within 12 training runs; hierarchical analysis retains that dependence. Appendix A.11 reports the mean changes, interval, and other mismatch conditions.

Figure 4 connects the reversal to a fixed filament target. The background condition adds diffuse signal to the measurement and reprojection, while the object-error map shows a broader low-level discrepancy around the filament structure. Reprojection MSE and independently normalized NMSE both worsen, as do object Pearson and NRMSE, despite the SSIM increase. These companion readouts locate the apparent gain in the particular image-similarity score; Appendix A.11 retains every nominal/background value.

Per-view maximum normalization changes what the comparison rewards. Dividing each image by its own peak suppresses absolute scale, and a shared broad background changes the local means and contrasts entering SSIM. Agreement in these image statistics can increase while the distribution of fluorescence in the reconstructed volume moves away from the reference. The counterexample thus tests the interpretation of an improved score under an explicit nuisance and preprocessing choice. Keeping the camera residual, normalized image similarity, and object error separate reveals which aspect improved and which deteriorated. Across the broader mismatch suite, their directions vary with the perturbation, reinforcing the need to state the scored target before treating a higher value as a reason to change reconstruction models.

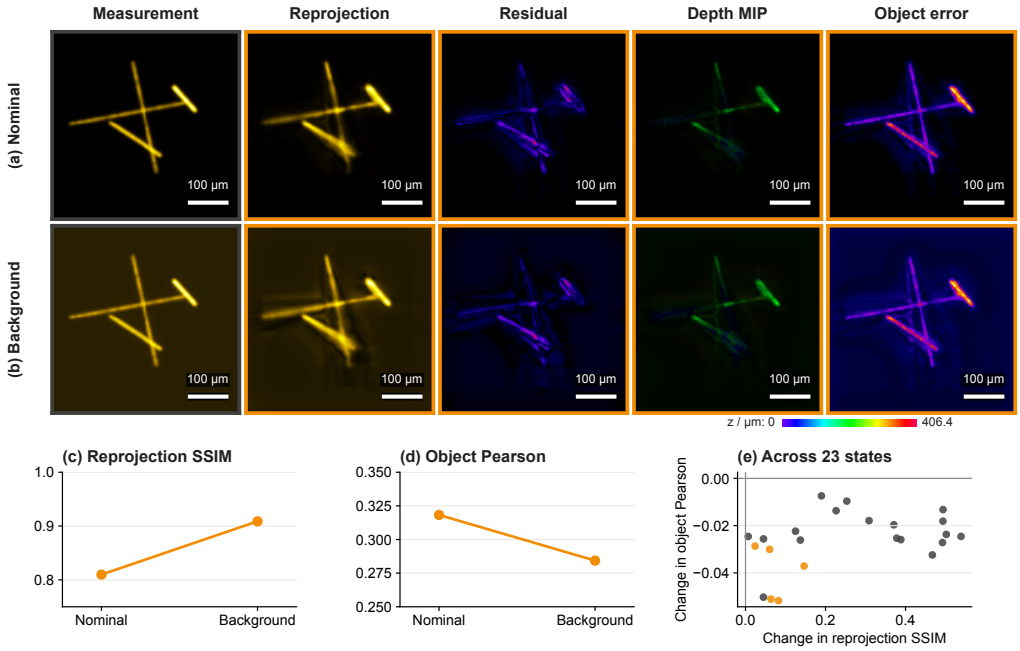


Figure 4: **A measurement score can move opposite to known-object fidelity.** Full-input Calibra-LF on a synthetic filament object under (a) nominal and (b) background-0.06 conditions. Columns: supplied view 4 (zero-based) measurement, reprojection, absolute normalized residual, XY depth MIP, and object error (maximum over z of absolute normalized-volume error). (c,d) Nine-view reprojection SSIM and volume Pearson for this example move oppositely. (e) Background-minus-nominal changes for 23 states nested within 12 runs, each endpoint averaged over the same 20 phantoms; orange denotes Calibra-LF, gray other learned families. All states gain SSIM and lose Pearson under the declared normalization. Bars: $100\,\mu$m.

5 Discussion: Matching Scores to Model Choice

SCOPE-FLFM ties selection to candidate scope and evidence target: CCS tests pooled versus close-candidate agreement, while CEV separates supplied-view fit, excluded-view prediction, and known-object fidelity. Withheld images test prediction; volume references assess recovered structure.

Background raises max-normalized SSIM while worsening measurement residuals and object estimates; selection therefore requires an imaging goal, normalization, and complementary evidence (Appendices A.1 and A.5; five-item reporting recommendation).

CCS distinguishes family choice, matched-modification tests, and within-run checkpoint selection, counting each state once while retaining run dependence. Selection data fix the rule before final readout; CEV evaluates the frozen choice with stated preprocessing and aggregation. The Calibra-LF case therefore requires reading prediction, normalized structure, and corruption response together (Appendices A.19 and A.28).

Known synthetic structures and controlled image-formation changes connect model choice to camera prediction, spatial agreement, and perturbation response. Appendix A.24 specifies the synthetic-data, historical-access, and comparison scope, as well as references needed for experimental anatomy, calibrated intensity, and instrument transfer.

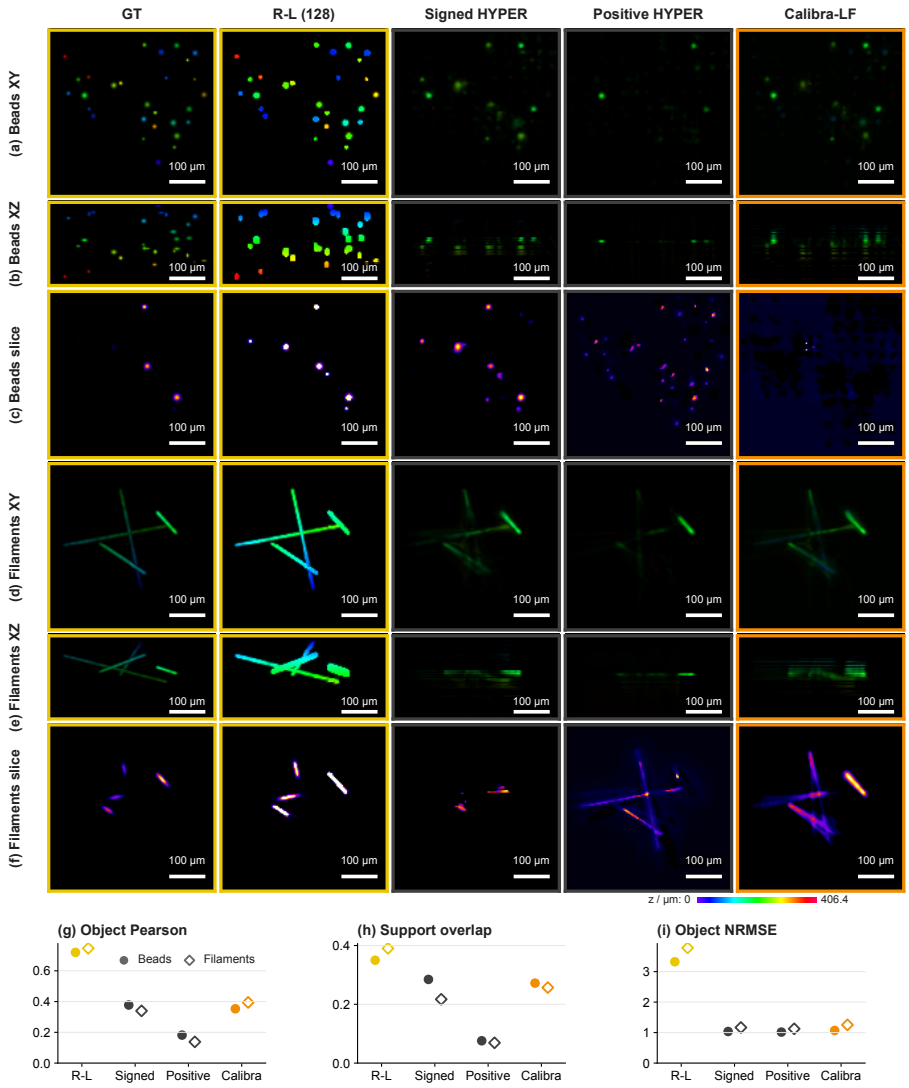


Figure 5: **Known synthetic structure and frozen reconstructions.** Sparse beads (a–c) and filaments (d–f): XY depth MIP, XZ maximum projection over Y, and shared-z slice. Columns: GT, RL-128, signed HYPER, positive HYPER, and Calibra-LF. Reference layer-intensity maxima fix $z = 73$ (beads) and $z = 54$ (filaments), zero-based, across methods. (g–i) Full-volume Pearson, equal-cardinality top-k support overlap (reference ≥ 0.2), and NRMSE from the displayed arrays. Filled circles: beads; open diamonds: filaments; no confidence intervals. Different-seed learned weights preclude matched-pair inference. Bars: $100\,\mu$m along X.

6 Conclusion

Camera-image agreement and three-dimensional recovery are distinct. In this synthetic FLFM benchmark, candidate stratification changes rank agreement, and additive background can raise max-normalized reprojection SSIM while lowering known-object agreement. SCOPE-FLFM separates supplied-view fit, excluded-view prediction, and recovered structure within defined candidate sets, linking each result to the imaging goal.

AI use statement

In this work, no generative AI tools were used.

REFERENCES

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 524–533, 2019.

Michael Broxton, Logan Grosenick, Samuel Yang, Noy Cohen, Aaron Andalman, Karl Deisseroth, and Marc Levoy. Wave optics theory and 3-D deconvolution for the light field microscope. *Optics Express*, 21(21):25418–25439, 2013. doi: 10.1364/oe.21.025418.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Lin Cong, Zeguan Wang, Yuming Chai, Wei Hang, Chunfeng Shang, Wenbin Yang, Lu Bai, Jiulin Du, Kai Wang, and Quan Wen. Rapid whole brain imaging of neural activity in freely behaving larval zebrafish (Danio rerio). *eLife*, 6:e28158, 2017. doi: 10.7554/elife.28158.

Siân Culley, David Albrecht, Caron Jacobs, Pedro Matos Pereira, Christophe Leterrier, Jason Mercer, and Ricardo Henriques. Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods*, 15(4):263–266, 2018. doi: 10.1038/nmeth.4605.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Parvin Forghani, Wenhao Liu, Zeyu Wang, Zhi Ling, Felipe Takaesu, Evan Yang, Gregory K. Tharp, Sheila Nielsen, Shankini Doraisingam, Stefanie Countryman, Michael E. Davis, Ronghu Wu, Shu Jia, and Chunhui Xu. Spaceflight alters protein levels and gene expression associated with stress response and metabolic characteristics in human cardiac spheroids. *Biomaterials*, 317:123080, 2025. doi: 10.1016/j.biomaterials.2024.123080.

Zijun Gao, Keyi Han, Zhi Ling, Hongmanlin Zhang, Edward A Botchwhey, Wen-hao Liu, Xuanwen Hua, Shuyi Nie, and Shu Jia. Volumetric Scattering Microscopy. *bioRxiv*, 2026. doi: 10.64898/2026.04.03.716429. Preprint.

Changliang Guo, Wenhao Liu, Xuanwen Hua, Haoyu Li, and Shu Jia. Fourier light-field microscopy. *Optics Express*, 27(18):25573–25594, 2019a. doi: 10.1364/oe.27.025573.

Changliang Guo, Wenhao Liu, and Shu Jia. Fourier-Domain Light-Field Microscopy. In *Biophotonics Congress: Optics in the Life Sciences Congress 2019 (BODA,BRAIN,NTM,OMA,OMP)*, pp. NS1B.3, 2019b. doi: 10.1364/ntm.2019.ns1b.3.

Changliang Guo, Tara Urner, and Shu Jia. 3D light-field endoscopic imaging using a GRIN lens array. *Applied Physics Letters*, 116(10):101105, 2020. doi: 10.1063/1.5143113.

Keyi Han, Xuanwen Hua, Vishwa Vasani, Ge-Ah R. Kim, Wenhao Liu, Shuichi Takayama, and Shu Jia. 3D super-resolution live-cell imaging with radial symmetry and Fourier light-field microscopy. *Biomedical Optics Express*, 13(11):5574–5584, 2022. doi: 10.1364/boe.471967.

Keyi Han, Xuanwen Hua, Tianrui Qi, Zijun Gao, Xiaopeng Wang, and Shu Jia. Volumetric localization microscopy with deep learning. *Nature Communications*, 16(1):10960, 2025. doi: 10.1038/s41467-025-65941-3.

Xuanwen Hua and Shu Jia. 3D Live Cell Imaging Using High-Resolution Fourier Light-Field Microscopy. In *Frontiers in Optics / Laser Science*, pp. FW7E.1, 2020. doi: 10.1364/fio.2020.fw7e.1.

Xuanwen Hua, Wenhao Liu, and Shu Jia. High-resolution Fourier light-field microscopy for volumetric multi-color live-cell imaging. *Optica*, 8(5):614–620, 2021. doi: 10.1364/optica.419236.

Xuanwen Hua, Biagio Mandracchia, Wenhao Liu, and Shu Jia. Volumetric and optofluidic live-cell imaging using high-resolution Fourier light-field microscopy (HR-FLFM). In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XX*, pp. 35, 2022. doi: 10.1117/12.2609264.

Xuanwen Hua, Keyi Han, Biagio Mandracchia, Wenhao Liu, and Shu Jia. High-resolution 3D Imaging Flow Cytometry with Fourier Light-Field Microscopy and Deep Learning. In *Frontiers in Optics + Laser Science 2023 (FiO, LS)*, pp. JW4A.29, 2023a. doi: 10.1364/fio.2023.jw4a.29.

Xuanwen Hua, Wenhao Liu, and Shu Jia. High-resolution Volumetric Cell Imaging with Deep Fourier Light-Field Microscopy (DFLFM). In *Optica Imaging Congress (3D, COSI, DH, FLatOptics, IS, pcAOP)*, pp. DM1A.3, 2023b. doi: 10.1364/3d.2023.dm1a.3.

Xuanwen Hua, Keyi Han, Biagio Mandracchia, Afsane Radmand, Wenhao Liu, Hyejin Kim, Zhou Yuan, Samuel M. Ehrlich, Kaitao Li, Corey Zheng, Jeonghwan Son, Aaron D. Silva Trenkle, Gabriel A. Kwong, Cheng Zhu, James E. Dahlman, and Shu Jia. Light-field flow cytometry for high-resolution, volumetric and multiparametric 3D single-cell analysis. *Nature Communications*, 15(1):1975, 2024. doi: 10.1038/s41467-024-46250-7.

Xuanwen Hua, Keyi Han, Zhi Ling, Olivia Reid, Zijun Gao, Hongmanlin Zhang, Edward A. Botchwey, Parvin Forghani, Wenhao Liu, Mithila Anil Sawant, Afsane Radmand, Hyejin Kim, James E Dahlman, Aparna Kesarwala, Chunhui Xu, and Shu Jia. PAVR: High-Resolution Cellular Imaging via a Physics-Aware Volumetric Reconstruction Framework. *bioRxiv*, 2026. doi: 10.64898/2026.03.04.709609. Preprint.

Haixin Huang, Haoyuan Qiu, Hanzhe Wu, Yihong Ji, Heng Li, Bin Yu, Danni Chen, and Junle Qu. SOFFLFM: Super-resolution optical fluctuation Fourier light-field microscopy. *Journal of Innovative Optical Health Sciences*, 16(03):2244007, 2023. doi: 10.1142/s1793545822440072.

Yan Huang, Corey Zheng, Zijun Gao, Wenhao Liu, and Shu Jia. Foveated light-field compound imager. *Science Advances*, 12(23):eaeb2738, 2026. doi: 10.1126/sciadv.aeb2738.

Danchen Jia, Yi Zhang, Qianwan Yang, Yujia Xue, Yuying Tan, Zhongyue Guo, Meng Zhang, Lei Tian, and Ji-Xin Cheng. 3D Chemical Imaging by Fluorescence-detected Mid-Infrared Photothermal Fourier Light Field Microscopy. *Chemical & Biomedical Imaging*, 1(3):260–267, 2023. doi: 10.1021/cbmi.3c00022.

Hyejin Kim, Ryan Zenhausern, Kara Gentry, Liming Lian, Sebastian G. Huayamares, Afsane Radmand, David Loughrey, Ananda R. Podilapu, Marine Z. C. Hatit, Huanzhen Ni, Andrea Li, Aram Shajii, Hannah E. Peck, Keyi Han, Xuanwen Hua, Shu Jia, Michele Martinez, Charles Lee, Philip J. Santangelo, Alice Tarantal, and James E. Dahlman. Lipid nanoparticle-mediated mRNA delivery to CD34+ cells in rhesus monkeys. *Nature Biotechnology*, 43(11):1813–1820, 2025. doi: 10.1038/s41587-024-02470-2.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void: Learning denoising from single noisy images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2129–2137, 2019. doi: 10.1109/cvpr.2019.00223.

Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. Light field microscopy. *ACM Transactions on Graphics*, 25(3):924–934, 2006. doi: 10.1145/1141911.1141976.

Haoyu Li, Changliang Guo, Deborah Kim-Holzapfel, Weiyi Li, Yelena Altshuller, Bryce Schroeder, Wenhao Liu, Yizhi Meng, Jarrod B. French, Ken-Ichi Takamaru, Michael A. Frohman, and Shu Jia. Fast, volumetric live-cell imaging using high-resolution light-field microscopy. *Biomedical Optics Express*, 10(1):29–49, 2019. doi: 10.1364/boe.10.000029.

Fanglin Linda Liu, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Fourier DiffuserScope: single-shot 3D Fourier light field microscopy with a diffuser. *Optics Express*, 28(20): 28969–28986, 2020. doi: 10.1364/oe.400876.

Zhi Ling. Training data sets containing 500 synthetic light-field data. figshare dataset, 2026. URL https://doi.org/10.6084/m9.figshare.32160783. CC BY 4.0.

Zhi Ling, Keyi Han, Wenhao Liu, Xuanwen Hua, and Shu Jia. Volumetric Live-Cell Autofluorescence Imaging Using Fourier Light-Field Microscopy. In *Frontiers in Optics + Laser Science 2023 (FiO, LS)*, pp. JM4A.79, 2023a. doi: 10.1364/fio.2023.jm4a.79.

Zhi Ling, Keyi Han, Wenhao Liu, Xuanwen Hua, and Shu Jia. Volumetric live-cell autofluorescence imaging using Fourier light-field microscopy. *Biomedical Optics Express*, 14(8):4237–4245, 2023b. doi: 10.1364/boe.495506.

Zhi Ling, Wenhao Liu, Biagio Mandracchia, Jessica Hou, Shuyi Nie, and Shu Jia. Enhanced Volumetric Reconstruction in Fourier Light-Field Microscopy through Denoising-Sparse Decomposition. In *Frontiers in Optics + Laser Science 2023 (FiO, LS)*, pp. FTh3D.4, 2023c. doi: 10.1364/fio.2023.fth3d.4.

Zhi Ling, Wenhao Liu, Kyungduck Yoon, Jessica Hou, Parvin Forghani, Xuanwen Hua, Hansol Yoon, Maryam Bagheri, Lakshmi P. Dasi, Biagio Mandracchia, Chunhui Xu, Shuyi Nie, and Shu Jia. Multiscale and recursive unmixing of spatiotemporal rhythms for live-cell and intravital cardiac microscopy. *Nature Cardiovascular Research*, 4(5):637–648, 2025. doi: 10.1038/s44161-025-00649-7.

Zhi Ling, Xuanwen Hua, Wenhao Liu, Hao Wu, Lucinda M Peng, Jessica Hou, Parvin Forghani, Christopher Pierce, Ge-Ah R Kim, Shuichi Takayama, Shuyi Nie, Chunhui Xu, Hang Lu, and Shu Jia. HYPER-Net: Physics-conditioned self-supervised reconstruction for Fourier light-field microscopy. *bioRxiv*, 2026a. doi: 10.64898/2026.04.14.718527. Preprint.

Zhi Ling, Wenhao Liu, Kyungduck Yoon, Zijun Gao, Keyi Han, Mithila Sawant, Aparna Kesarwala, and Shu Jia. LFC-plus: simultaneous multicolour volume cytometry for high-throughput single-cell analysis. *Lab on a Chip*, 26(3):564–575, 2026b. doi: 10.1039/d5lc00962f.

Wenhao Liu and Shu Jia. Enhancing the resolution of Fourier light-field microscopy using a hybrid wide-field microscopy image. In *Frontiers in Optics / Laser Science*, pp. FW7E.2, 2020. doi: 10.1364/fio.2020.fw7e.2.

Wenhao Liu and Shu Jia. wFLFM: enhancing the resolution of Fourier light-field microscopy using a hybrid wide-field image. *Applied Physics Express*, 14(1):012007, 2021. doi: 10.35848/1882-0786/abd3b7.

Wenhao Liu and Shu Jia. High-resolution volumetric imaging using hybrid image of Fourier light-field microscopy and wide-field microscopy. In *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XXIX*, pp. 16, 2022. doi: 10.1117/12.2610474.

Wenhao Liu, Changliang Guo, Xuanwen Hua, and Shu Jia. Fourier Light-Field Microscopy: An Integral Model and Experimental Verification. In *Biophotonics Congress: Optics in the Life Sciences Congress 2019 (BODA,BRAIN,NTM,OMA,OMP)*, pp. DT1B.4, 2019. doi: 10.1364/boda.2019.dt1b.4.

Wenhao Liu, Ge-Ah R Kim, Shuichi Takayama, and Shu Jia. Fourier light-field imaging of human organoids with a hybrid point-spread function. *Biosensors and Bioelectronics*, 208:114201, 2022. doi: 10.1016/j.bios.2022.114201.

Wenhao Liu, Zhi Ling, Shuyi Nie, and Shu Jia. Instant dual-color Fourier light-field microscopy. In *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XXXII*, pp. 5, 2025. doi: 10.1117/12.3043906.

A. Llavador, J. Sola-Pikabea, G. Saavedra, B. Javidi, and M. Martínez-Corral. Resolution improvements in integral microscopy with Fourier plane recording. *Optics Express*, 24(18):20792, 2016. doi: 10.1364/oe.24.020792.

Zhi Lu, Yu Liu, Manchang Jin, Xin Luo, Huanjing Yue, Zian Wang, Siqing Zuo, Yunmin Zeng, Jiaqi Fan, Yanwei Pang, Jiamin Wu, Jingyu Yang, and Qionghai Dai. Virtual-scanning light-field microscopy for robust snapshot high-resolution volumetric imaging. *Nature Methods*, 20(5): 735–746, 2023. doi: 10.1038/s41592-023-01839-6.

Zhi Lu, Manchang Jin, Shuai Chen, Xiaoge Wang, Feihao Sun, Qi Zhang, Zhifeng Zhao, Jiamin Wu, Jingyu Yang, and Qionghai Dai. Physics-driven self-supervised learning for fast high-resolution robust 3D reconstruction of light-field microscopy. *Nature Methods*, 22(7):1545–1555, 2025. doi: 10.1038/s41592-025-02698-z.

L. B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745, 1974. doi: 10.1086/111605.

Biagio Mandracchia, Xuanwen Hua, Changliang Guo, Jeonghwan Son, Tara Urner, and Shu Jia. Fast and accurate sCMOS noise correction for fluorescence microscopy. *Nature Communications*, 11 (1):94, 2020. doi: 10.1038/s41467-019-13841-8.

Biagio Mandracchia, Wenhao Liu, Xuanwen Hua, Parvin Forghani, Soojung Lee, Jessica Hou, Shuyi Nie, Chunhui Xu, and Shu Jia. Optimal sparsity allows reliable system-aware restoration of fluorescence microscopy images. *Science Advances*, 9(35):eadg9245, 2023. doi: 10.1126/sciadv.adg9245.

Robert Prevedel, Young-Gyu Yoon, Maximilian Hoffmann, Nikita Pak, Gordon Wetzstein, Saul Kato, Tina Schrödel, Ramesh Raskar, Manuel Zimmer, Edward S Boyden, and Alipasha Vaziri. Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nature Methods*, 11(7):727–730, 2014. doi: 10.1038/nmeth.2964.

Afsane Radmand, Hyejin Kim, Jared Beyersdorf, Curtis N. Dobrowolski, Ryan Zenhausern, Kalina Paunovska, Sebastian G. Huayamares, Xuanwen Hua, Keyi Han, David Loughrey, Marine Z. C. Hatit, Ada Del Cid, Huanzhen Ni, Aram Shajii, Andrea Li, Abinaya Muralidharan, Hannah E. Peck, Karen E. Tiegreen, Shu Jia, Philip J. Santangelo, and James E. Dahlman. Cationic cholesterol-dependent LNP delivery to lung stem cells, the liver, and heart. *Proceedings of the National Academy of Sciences*, 121(11):e2307801120, 2024. doi: 10.1073/pnas.2307801120.

William Hadley Richardson. Bayesian-Based Iterative Method of Image Restoration. *Journal of the Optical Society of America*, 62(1):55, 1972. doi: 10.1364/josa.62.000055.

G. Scrofani, J. Sola-Pikabea, A. Llavador, E. Sanchez-Ortiga, J. C. Barreiro, G. Saavedra, J. Garcia-Sucerquia, and M. Martínez-Corral. FIMic: design for ultimate 3D-integral microscopy of in-vivo biological samples. *Biomedical Optics Express*, 9(1):335, 2018. doi: 10.1364/boe.9.000335.

Gabriele Scrofani, Genaro Saavedra, Manuel Martínez-Corral, and Emilio Sánchez-Ortiga. Three-dimensional real-time darkfield imaging through Fourier lightfield microscopy. *Optics Express*, 28(21):30513, 2020. doi: 10.1364/oe.404961.

Shu Jia Lab. HYPER: A physics-conditioned, self-supervised reconstruction network for Fourier light-field microscopy. GitHub repository, 2026. URL `https://github.com/ShuJiaLab/HYPER/tree/2249fe5c23012b1c3b38faf9046920018932bb4d`. Revision 2249fe5c2301; full commit pinned in URL.

Ruth R. Sims, Sohaib Abdul Rehman, Martin O. Lenz, Sarah I. Benaissa, Ezra Bruggeman, Adam Clark, Edward W. Sanders, Aleks Ponjavic, Leila Muresan, Steven F. Lee, and Kevin O'Holleran. Single molecule light field microscopy. *Optica*, 7(9):1065, 2020. doi: 10.1364/optica.397172.

Tara M. Urner, Andrew Inman, Benjamin Lapid, and Shu Jia. Three-dimensional light-field microendoscopy with a GRIN lens array. *Biomedical Optics Express*, 13(2):590–607, 2022. doi: 10.1364/boe.447578.

Josue Page Vizcaino, Zeguan Wang, Panagiotis Symvoulidis, Paolo Favaro, Burcu Guner-Ataman, Edward S. Boyden, and Tobias Lasser. Real-time light field 3D microscopy via sparsity-driven learned deconvolution. In *IEEE International Conference on Computational Photography*, pp. 1–11, 2021. doi: 10.1109/iccp51581.2021.9466256.

Nils Wagner, Fynn Beuttenmueller, Nils Norlin, Jakob Gierten, Juan Carlos Boffi, Joachim Wittbrodt, Martin Weigert, Lars Hufnagel, Robert Prevedel, and Anna Kreshuk. Deep learning-enhanced light-field imaging with continuous validation. *Nature Methods*, 18(5):557–563, 2021. doi: 10.1038/s41592-021-01136-0.

Lu Wang, Analyne Schroeder, Katya Marchetti, Ingolf Reim, Matthias Blanc, K'leigh Guillotte, Amelie Mück, Elaina Bao, Wenhao Liu, Shu Jia, Denise Malicki, Rolf Bodmer, Shuyi Nie, Karen Ocorr, and Paul Grossfeld. Conserved roles of cardiac endothelial ETS factors in ventricular development and disease. *medRxiv*, 2026. doi: 10.64898/2026.09.08.26361286. Preprint.

Zhaoqiang Wang, Lanxin Zhu, Hao Zhang, Guo Li, Chengqiang Yi, Yi Li, Yicong Yang, Yichen Ding, Mei Zhen, Shangbang Gao, Tzung K. Hsiai, and Peng Fei. Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning. *Nature Methods*, 18(5):551–556, 2021. doi: 10.1038/s41592-021-01058-x.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/tip.2003.819861.

Armand Wirgin. The inverse crime. *arXiv preprint math-ph/0401050*, 2004.

Hao Wu, Zhi Ling, Zijun Gao, Keyi Han, Wenhao Zhang, Chunhui Xu, and Shu Jia. Fast volumetric light-field reconstruction via patch-wise Wiener–Butterworth deconvolution. *Optics Express*, 34 (12):21816–21825, 2026. doi: 10.1364/oe.596951.

Shunan Wu, Wenhao Liu, Xuanwen Hua, Jiajia Luo, Xunbin Wei, and Shu Jia. PEARL: projection-estimation accelerated Richardson-Lucy deconvolution for rapid volumetric imaging. *Optics Express*, 33(11):22685, 2025. doi: 10.1364/oe.559124.

Chenyu Xu, Zhouyu Jin, Chengkang Shen, Hao Zhu, Zhan Ma, Bo Xiong, You Zhou, Xun Cao, and Ning Gu. 3D Gaussian adaptive reconstruction for Fourier light-field microscopy. *Advanced Imaging*, 2(5):055001, 2025. doi: 10.3788/ai.2025.50000.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020. doi: 10.1002/mrm.28378.

Chengqiang Yi, Lanxin Zhu, Jiahao Sun, Zhaofei Wang, Meng Zhang, Fenghe Zhong, Luxin Yan, Jiang Tang, Liang Huang, Yu-Hui Zhang, Dongyu Li, and Peng Fei. Video-rate 3D imaging of living cells using Fourier view-channel-depth light field microscopy. *Communications Biology*, 6 (1):1259, 2023. doi: 10.1038/s42003-023-05636-x.

Boya Zhang, Sam Daly, Chengxi Zhu, Martin O. Lenz, Lucien E. Weiss, Lisa-Maria Needham, Ruby Peters, Steven F. Lee, and Kevin O'Holleran. Vortex light field microscopy: 3D spectral single-molecule imaging with a twist. *Optica*, 11(11):1519, 2024. doi: 10.1364/optica.534148.

Ruixuan Zhao, Xuanwen Hua, Woongjae Baek, Zhaoqiang Wang, Shu Jia, and Liang Gao. A review of light-field imaging in biomedical sciences. *Med-X*, 3(1):25, 2025. doi: 10.1007/s44258-025-00070-6.

Corey Zheng and Shu Jia. Multimodal 3D light-field and laser-speckle endoscopy. *Biomedical Optics Express*, 2026. doi: 10.1364/boe.610175.

A Appendix

A.1 Scope and Evidence Layers

CCS establishes the candidate hierarchy before CEV assigns meaning to a score. The appendix therefore follows the same order as the paper: first we account for 36 checkpoint files, 12 optimization runs, and 23 deduplicated states; then we separate evidence on supplied views, excluded views, and known synthetic objects.

1. **Supplied-view fit** scores measurements that entered the reconstruction branch and reports consistency with the observed camera image.
2. **Excluded-view prediction** scores views removed from that branch and tests prediction under the declared FLFM acquisition contract.
3. **Known-object fidelity** compares the reconstructed volume with a known synthetic 3D reference and therefore probes recovered structure under the simulator.

Candidate scope determines selection relevance. Rank correlation is conditional on the population over which it is computed. In a finite mixture of well-separated model families, two scores can agree on the coarse family ordering even when they disagree within each family. Such a coefficient describes separation across the pooled set; selecting a checkpoint among close alternatives requires agreement within that candidate set. SCOPE-FLFM therefore reports pooled, architecture-stratified, matched-pair, and within-run consequences, tying each result to the choice it can inform.

Each reference answers a distinct question. Excluded-view prediction is a stricter measurement test than scoring a view supplied to the same branch. It removes one direct route to copying or fitting the scored observation. Its target is still a camera image. The excluded view remains generated by the same synthetic object and acquisition contract, and the inverse problem can admit volumes with similar projections but different support or depth. Supplied-view, excluded-view, and known-object evidence therefore answer distinct questions about explanation, prediction, and recovered structure. Their relevance depends on the intended use.

Preprocessing is part of the evidence. The background counterexample must be interpreted with the score's normalization. Per-view maximum normalization suppresses absolute amplitude and can make broad low-frequency content appear structurally closer even as measurement residuals and object estimates worsen. The observed SSIM increase occurs under this preprocessing and nuisance. A useful selection report therefore states the normalization and companion errors, making clear which changes the score rewards and which it suppresses.

The benchmark contains 12 learned runs, 36 learned selection checkpoints, one fixed-iteration Richardson–Lucy reference, 20 prospective corpus test groups, 20 known-volume phantoms, nine leave-one-view-out masks, six acquisition corruptions plus clean, and three checkpoint-selection rules. Every learned configuration is repeated at three seeds. An expanded matched-protocol pool adds 30 HYPER/FLFM runs across 10 families; its role is to widen the candidate context, while the 23-state pool carries the principal CCS–CEV diagnosis.

A.2 Data Provenance and Nested Split

The frozen split makes the prospective decision traceable: groups 0–19 provide optimization records, groups 20–29 freeze checkpoint choices, and groups 30–49 provide the final corpus readout. We use the 500-frame synthetic FLFM training dataset released by Zhi Ling for HYPER-Net (Ling et al., 2026a; Ling, 2026). The released filenames group the measurements into 50 base groups with ten variants each; the generation driver uses tenfold augmentation. PSF and slope metadata are present, while paired 3D volumes are supplied only by our separate known-volume dataset.

Released data and generation method. The HYPER repository supplies the training dataset, PSF data, and generation implementation (Shu Jia Lab, 2026), inspected at commit `2249fe5c23012b1c3b38faf9046920018932bb4d`. The released generator samples

global, local, and smaller emitter clusters, assigns density-dependent intensities, smooths the emitters, and adds low-intensity background clouds. Its processing includes aperture masking, volume augmentation, wide-field PSF convolution, and Richardson–Lucy reconstruction. Light-field measurements are formed by FFT-based depthwise PSF convolution followed by depth summation. We cite HYPER-Net for the reconstruction framework and the fixed code revision for these generation operations. We evaluate the published TIFF dataset; exact regeneration of those TIFFs from the inspected implementation remains unverified. The 20 reference volumes in Appendix A.9 are generated separately for this study.

Table A1: Group-level split of the released 500-frame synthetic training corpus. Counts describe data access; augmentation 0 supplies the fixed selection and final evaluation records.

Role	Groups	Frames available	Use in this study
Optimization	0–19	200	Sampled for optimization updates
Checkpoint selection	20–29	100	Augmentation 0; 10 fixed records
Final corpus evaluation	30–49	200	Augmentation 0; 20 fixed records

The split was frozen before prospective training, and groups 30–49 never enter prospective gradients or checkpoint selection in this rerun. The groups were historically accessible during project development, so the final set is a frozen prospective readout within this rerun. The table therefore supports the stated checkpoint rule and its post-selection consequences.

We additionally audited all 500 frames after averaging the ten augmentations within each of 50 generator groups. Across the ten variants of each generator group, every split covers PSF indices 0–99 and has the same mean PSF-slope vector. These coverage statistics use all variants, while the fixed selection and final evaluations use augmentation 0. Table A2 summarizes the observable frame statistics.

The largest absolute standardized mean difference is 0.547 for the final-versus-optimization y center, only 0.00195 of the normalized field height. Balanced PSF-index coverage and similar first-order intensity statistics make the split suitable for the declared comparison, while the corpus itself supplies synthetic generator semantics rather than biological composition. The table therefore supports split traceability and candidate comparability, not a claim about experimental specimens.

Table A2: Observable split statistics after averaging ten augmentations within each generator group. P99 denotes the 99th intensity percentile; the y center uses normalized field coordinates. These distribution diagnostics carry no performance ranking.

Split	Signal mean	Signal std.	P99	Gradient energy	Normalized y center
Optimization	0.0913	0.0968	0.4477	0.0184	0.5003
Selection	0.0935	0.0975	0.4496	0.0180	0.5000
Final	0.0980	0.1012	0.4619	0.0181	0.4984

A.3 Exact Mask and Corruption Contracts

Masking defines what the reconstructor can access and therefore fixes the CEV contract for every score. We use zero-based indices: training uses the fixed odd/even software partitions, selection holds out $\{0, 8\}$ within odd positions and $\{3, 5\}$ within even positions, and final evaluation tests each of the nine positions once within its partition. Frozen-seed random deletions hold one to three views and use masks distinct from selection.

The released corpus lacks frame-level wavelength provenance, so corpus results use the term "within-partition." The monochrome known-volume phantoms generate all nine views from one volume and provide the verified same-channel angular tests. The corruption suite then stresses the same candidate under fixed changes in attenuation, clipping, noise, shifts, or conditioning metadata.

The corruption suite contains:

- attenuation by a factor of 0.5 on retained views;

- clipping at 60% of each retained view maximum;
- Poisson noise at 30 effective photons plus Gaussian noise with standard deviation 0.02 on normalized input;
- deterministic ± 2-pixel lenslet-local shifts;
- a global $(+2, -1)$ registration shift within retained tiles;
- a PSF-conditioning index offset of seven while preserving the true projector for evaluation.

All values were fixed before final results were read.

A.4 Models, Budgets, and Selection Rules

Model identity determines the CCS strata, so all learned candidates share the declared budget before their evidence is compared. Each predicts 128 depth planes with the same FLFM projector, train/selection records, 24 epochs, 80 updates per epoch, batch size one, optimizer family, learning rate 5×10^{-5}, and fixed-record validation frequency. The core set contains signed HYPER, softplus-positive HYPER, the XLFMNet-inspired local-FLFM control, and Calibra-LF; the expanded pool adds HYPER prior fusion and XLFM/CWFA variants under the same local projector. The rebuilt tables use the selection-frozen RL-128 reference, while Appendix A.12 retains the full iteration trajectory.

Calibra-LF is the matched HYPER-derived case study used later to show endpoint trade-offs. It retains HYPER PSF conditioning, FiLM, and self-supervised reprojection, then adds Bounded Depth Gain and Raw-Conditioned Residual after the single-volume head. The matched backbone uses base width 32, 128 depth planes, and a 256-dimensional embedding of the 18-dimensional PSF slope vector; it has 3,702,908 parameters versus 3,700,836 for signed HYPER, a 0.056% overhead. Appendix A.25 fixes the exact forward path and parameter accounting. Three full/odd/even forwards share all weights and each produce one volume; standard inference uses the full input and returns one volume.

Each run writes three rule-specific checkpoint files: loss-minimizing, full-view reprojection-SSIM-maximizing, and active-view SSIM-maximizing under the fixed selection mask. A shared run–epoch state can therefore have several file aliases. The tables retain aliases for provenance, while CCS collapses identical run–epoch pairs before ranking so that file multiplicity cannot manufacture evidence.

A.5 Visual Evidence and Display Conventions

The figures follow the evidence progression. Figure S1 establishes candidate identity and deduplication; Figure S2 fixes the three evidence paths; Figures S4–S11 then expose rank strata, mismatch behavior, object diversity, matched Calibra-LF effects, implementation budget, endpoint units, and the exact masks. Each panel uses the same anonymous records as the complete tables, connecting the numerical comparisons to their interpretation.

The figure metadata identify the source measurements, reference volumes, candidate states, and display transforms used in each panel. Main Figure 2 combines an acquisition illustration with separate synthetic quantitative panels. The known-object comparisons use fixed sparse-bead and filament examples, and the morphology atlas contains the 20 synthetic reference volumes. The matched Calibra-LF pairs below provide a separate state-level analysis.

Display convention. Main Figures 3–5 use HYPER's nominal object-space sampling, $1.8\,\mu\mathrm{m}$ laterally and $3.2\,\mu\mathrm{m}$ axially, declared in its released PSF code and consistent with the reported optics and volume extent of HYPER-Net (Ling et al., 2026a; Shu Jia Lab, 2026); this is not independent experimental calibration. Synthetic volumes retain $(Z, Y, X) = (128, 256, 256)$ sampling; crops preserve spacing, and the nine-view mosaic comprises separate views. Horizontal bars use length divided by lateral spacing and stay within one view. Rainbow hue encodes peak-intensity depth in XY MIPs and row depth in XZ, with intensity as brightness; depth is relative to the first stored plane, 0–$406.4\,\mu\mathrm{m}$. XZ retains array aspect ratio, not isotropic physical proportions. Volumes are clamped nonnegative and independently divided by their volume-wide maxima; ROIs inherit this normalization. Hot measurements/reprojections and Fire slices are individually max-normalized; residual/error maps use Fire on a fixed 0–1 range. No gamma or percentile clipping is applied. Borders identify the featured model (orange), GT/RL (yellow), and other inputs/models (dark gray).

Table conventions. Arrows indicate the favorable direction of a performance metric. Red and blue bold numbers mark the best and second distinct finite values within the comparison scope stated in each caption, using unrounded aggregates where available; exact ties share a color. Scores and normalized errors are dimensionless unless a unit is shown. Light red identifies the matched-protocol Calibra-LF (Ours); light blue identifies the non-Ours row with the most second-best performance cells, with table order breaking ties. In the historical frame audit, light blue identifies the Prior reference columns, which have the most second-best performance cells. Background identity is independent of numerical rank, and colors convey neither significance nor an overall winner. Metadata, sample counts, correlations between rankings, and signed-difference tables are diagnostic and remain unranked. Historical Prior/Selected checkpoints retain their original identities.

Figure S1 supplies the candidate map used by every later ranking. Its file-to-state collapse makes the pooled coefficient interpretable and keeps run-shared checkpoints in the correct dependence class.

Figure S2 supplies the evidence map used by every later endpoint. The same predicted volume can therefore be read as camera consistency, excluded-view prediction, or normalized object fidelity, with each path carrying its own claim.

Figure S3 makes the deduplication rule visible before the numerical tables. It explains why the appendix reports files for provenance, states for ranking, and runs for paired consequences.

Table A3 turns the CEV diagram into a reading rule: each exclusion pattern and aggregation unit supports a different conclusion. It is the reference point for the rank matrix, mismatch panels, and object tables that follow.

A minimum selection report. For an FLFM reprojection result to support model selection, we recommend reporting five linked items: (i) the candidate hierarchy and state-deduplication key; (ii) the data and rule used to choose a checkpoint; (iii) the information excluded from the scored branch; (iv) the scored target, preprocessing, and independent aggregation unit; and (v) the strongest conclusion that target permits. Measurement evidence remains valuable: supplied-view fit checks physical-loop consistency, and excluded-view prediction probes information withheld from that branch. A known object adds a structural target. Reporting these separately lets researchers choose for the imaging property they need.

Figure S4 is the bridge from candidate accounting to the local ranking result. Reading rows and columns together shows which apparent agreement comes from broad family separation and which survives within a close candidate scope or across evidence targets.

Figure S5 places measurement and object responses on separate axes under controlled mismatch. The background trajectories show how background and normalization affect camera and object scores together, providing context for the later 23/23 counterexample.

Figure S6 extends target dependence to an iterative reference. Its stopping-point markers show why an iteration selected for one evidence path cannot be silently reused as the optimum for another.

Figure S7 defines the object dataset used by CEV. The four morphology rows ensure that Pearson, support, and axial diagnostics are read over a stated range of synthetic structures.

Figure S8 is the local matched case study. Because split, budget, seed, and rule are paired, each stem shows how the same model change affects clean prediction and object structure; the accompanying tables extend this comparison to corruption response.

Figure S9 reports the implementation envelope behind the candidate rows. It lets the reader interpret endpoint differences with parameter and forward-cost context while keeping the scientific conclusion tied to the declared metrics.

Figure S10 closes the loop between morphology and aggregation. Its unit map prevents a mask, phantom, or run-shared state from being counted as an independent replicate and keeps each correlation attached to the evidence target it measures.

Figure S11 is the visual form of the access contract. It lets the reader verify which tiles entered a branch, which tile was scored, and why corpus masks and monochrome-object masks carry different interpretation.

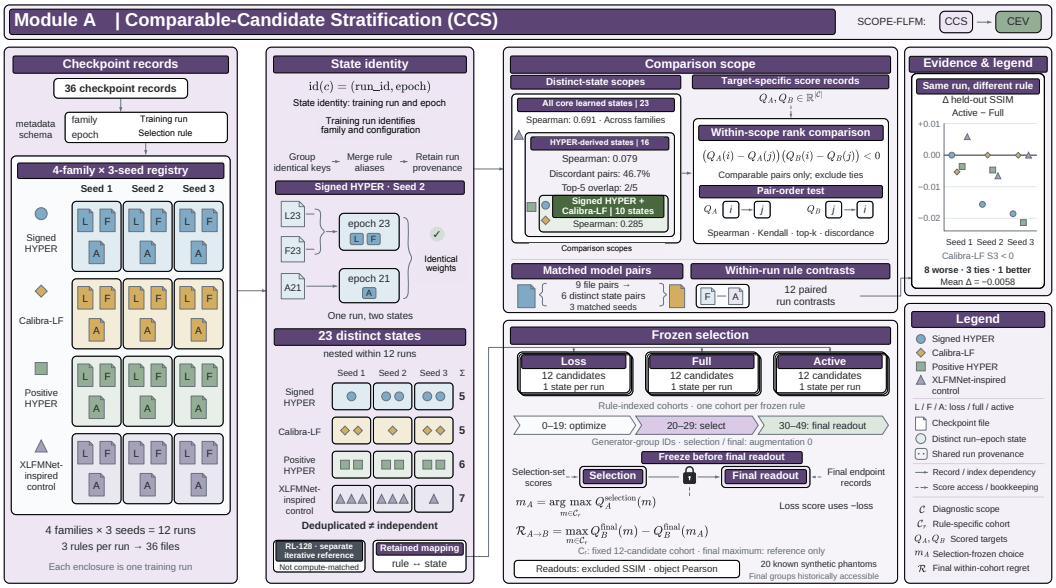


Figure S1: **Comparable-Candidate Stratification (CCS).** CCS maps 36 rule-specific files to 23 run–epoch states within 12 optimization runs, establishing pooled, local, matched-pair, and within-run comparison layers. The nested strata are All core learned states (23), HYPER-derived states (16), and Signed HYPER + Calibra-LF (10). Four families and three seeds supply the candidate coverage; six matched Calibra-LF/HYPER state pairs and 12 within-run rule contrasts retain the dependence needed for selection consequences. Each rule freezes one 12-candidate cohort on groups 20–29 before final corpus and known-object readout. Distinct-state accounting connects checkpoint files to the candidates being ranked. RL-128 remains a separate iterative reference.

Table A3: Minimum SCOPE-FLFM reporting contract. Each row pairs an exclusion rule and aggregation unit with the conclusion supported by that evidence layer. Masks within a group or phantom are repeated evaluations, not independent samples. The experimental-object row records the additional reference needed for measured-domain fidelity; the present benchmark supplies the first three rows.

Evidence	Required exclusion	Aggregation unit	Permitted conclusion
Observed measurement	None; scored views may enter the input	Group, averaged over masks	Supplied-measurement consistency
Held-out measurement	Scored views excluded from input and selection target	Group, averaged over masks	Specified measurement extrapolation
Known synthetic object	Target excluded from optimization and selection	Phantom, averaged over masks	Declared simulator/object fidelity
Experimental object	Independently acquired registered reference	Specimen or acquisition	Measured-domain fidelity

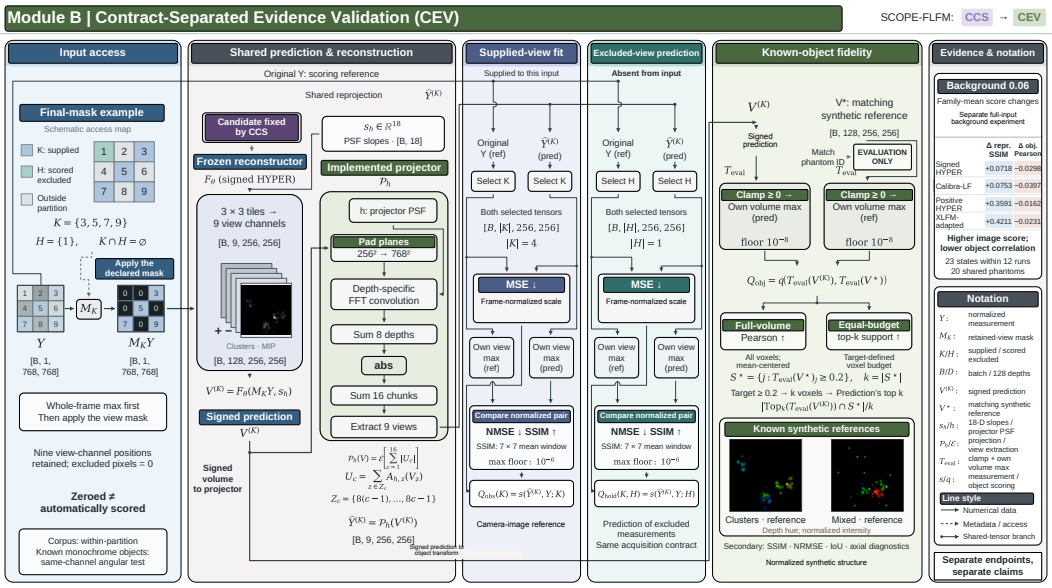


Figure S2: **Contract-Separated Evidence Validation (CEV).** One prediction feeds three declared paths: supplied-view fit, excluded-view prediction, and known-object fidelity. The illustrated mask retains one-based views $K = \{3, 5, 7, 9\}$ and scores $H = \{1\}$; whole-frame normalization precedes masking, measurement MSE uses frame-scale values, measurement NMSE/SSIM use independent view maxima, and object metrics use independently clamped volume maxima. The grayscale forward is a display-only signed-HYPER example on a clustered-emitter object (PSF index 69); the signed array remains the projection input. Clustered-emitter and mixed-structure thumbnails are reference-library examples from the frozen atlas; hue encodes the simulator depth plane. The background-0.06 result belongs to the separate full-input mismatch experiment. The three paths connect each score to its measurement or structural target.

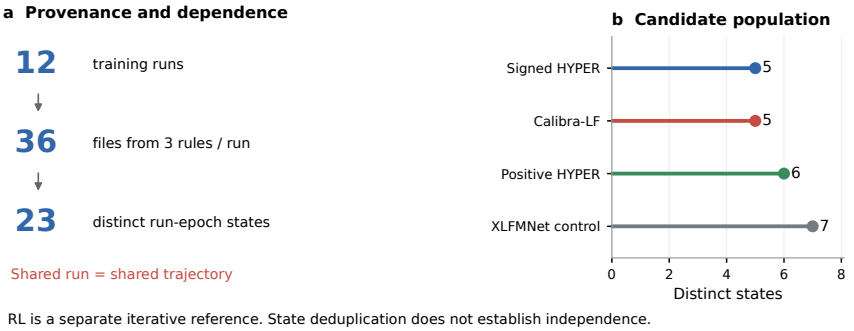


Figure S3: **Candidate provenance and deduplicated-state accounting.** The three rules in each of 12 runs produce 36 files, while identical run–epoch weights collapse to 23 learned states. Family counts describe coverage after this collapse; run identity preserves the shared optimization trajectory. This accounting is the prerequisite for interpreting rank correlations and bootstrap units, and the Richardson–Lucy sweep remains a separate iterative branch.

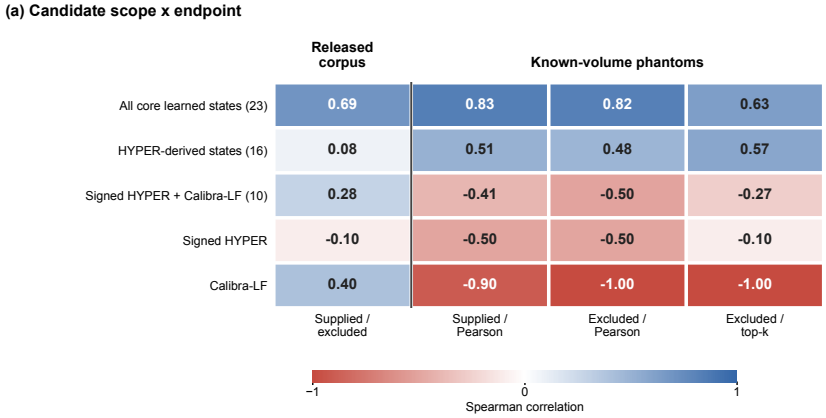


Figure S4: **Population- and endpoint-stratified rank agreement.** Rows change the candidate population and columns change the paired endpoints: supplied/excluded SSIM, supplied/Pearson, excluded/Pearson, and excluded/top-k support. The first column compares supplied-view and excluded-view rankings on the final corpus groups. The remaining columns compare measurement and object rankings on the 20 known-volume phantoms. Each row uses the same candidate stratum across these analyses: All core learned states (23), HYPER-derived states (16), Signed HYPER + Calibra-LF (10), signed HYPER (5), or Calibra-LF (5). Spearman values use a fixed $[-1, 1]$ scale. Pooled agreement can coexist with weak local agreement or endpoint disagreement, while run-shared states retain their dependence.

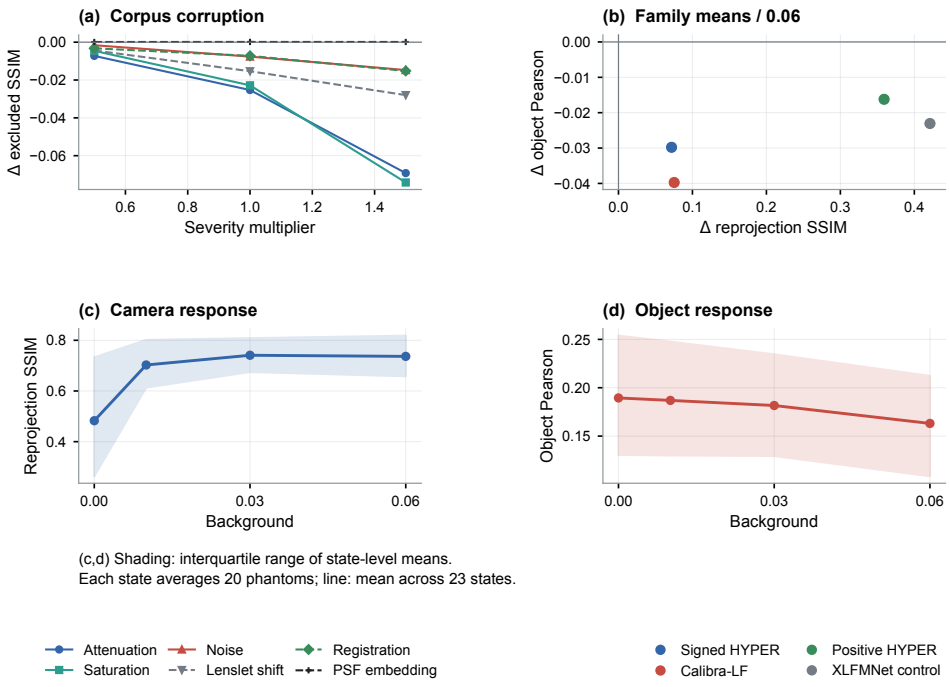


Figure S5: **Corruption and mismatch severity.** Panel (a) tracks excluded-SSIM changes for six corpus corruptions across frozen severity multipliers; the PSF row changes conditioning metadata while the data generator and final projector stay fixed. Panel (b) reports family-mean changes at background 0.06. Panels (c,d) trace camera and object scores across background levels: each state first averages the same 20 phantoms, lines average these state-level values, and shading spans their 25th–75th percentiles. These interquartile bands describe dispersion across 23 states nested within 12 runs, not confidence intervals. The figure shows how background and normalization affect camera and object scores together.

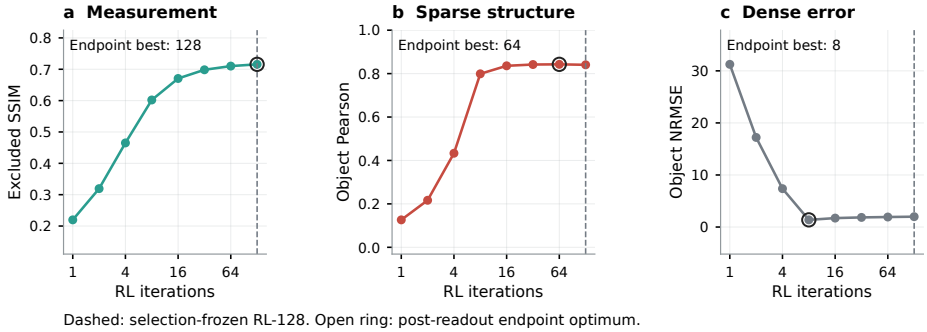


Figure S6: **Richardson–Lucy iteration trajectory.** The sweep follows excluded measurement SSIM, normalized object Pearson, and object NRMSE on their own axes. RL-128 is the iteration frozen by the selection-full rule on groups 20–29; open rings mark endpoint-specific optima read after freezing. The trajectory supplies a classical reference whose preferred iteration changes with the evidence target.

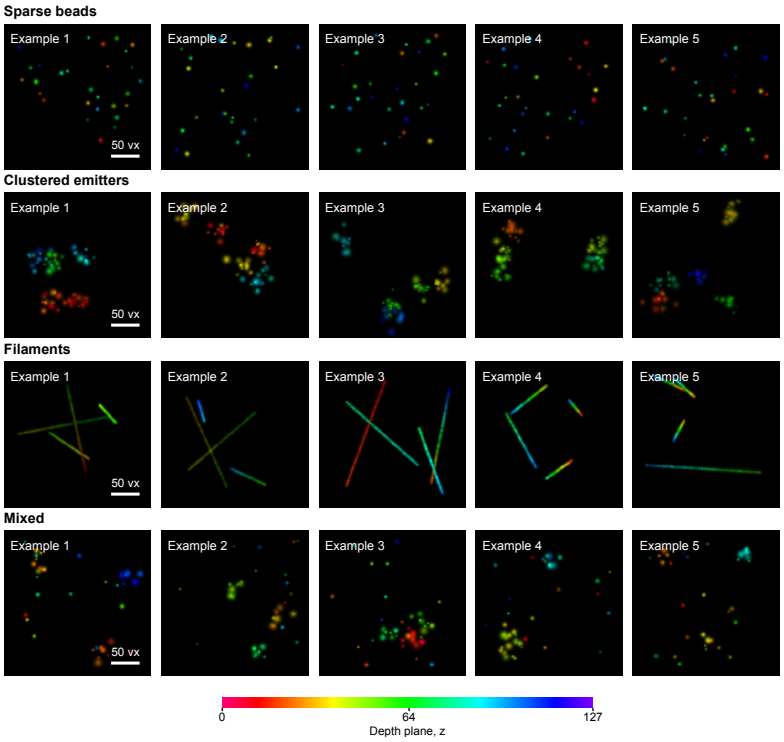


Figure S7: **Known-volume synthetic morphology atlas.** The 20 references span sparse beads, clustered emitters, filaments, and mixtures, with five phantoms per family. These strata probe support/localization, separation, continuity, and combined structure for the object branch of CEV. Projections use frozen global normalization; hue encodes the simulator depth plane, and each bar spans 50 of 256 simulator voxels.

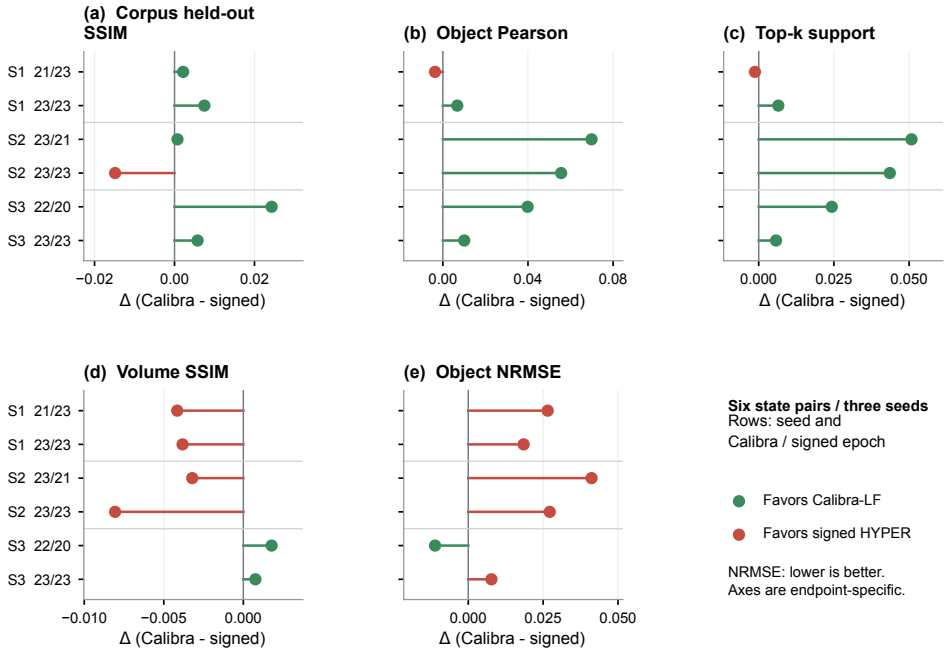


Figure S8: **Six distinct Calibra-LF versus signed-HYPER state pairs.** Each stem is a matched Calibra-LF-minus-signed-HYPER change in excluded SSIM, object Pearson, equal-budget top-k support, volume SSIM, or NRMSE. Seeds and paired epochs identify the six distinct state pairs after duplicate rule files are collapsed. Panel (a) uses the final corpus groups; panels (b–e) use the 20 known-volume phantoms. Green marks changes favoring Calibra-LF and vermilion marks changes favoring signed HYPER, with lower NRMSE favorable. The panels show how the same matched model change affects clean prediction, structural agreement, and normalized error.

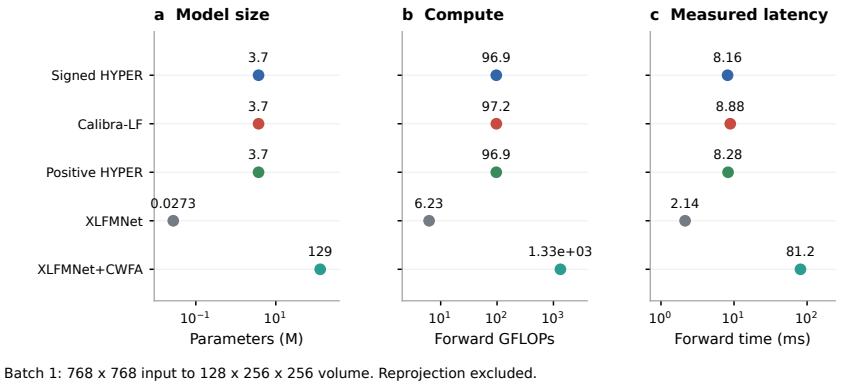
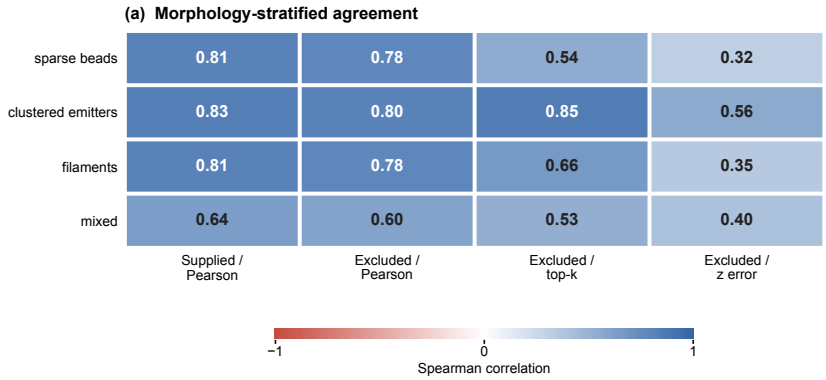


Figure S9: **Local complexity and runtime envelope.** Method rows report parameters, approximate network-forward GFLOPs, and local forward milliseconds for the common batch-one input/output contract. The FLFM projector is excluded, so the panel supplies implementation budget context for the candidate comparison while leaving optical latency and cross-paper speed outside this readout.



(b) Evidence inventory

Endpoint	Aggregation unit	Scored target
Supplied measurement	20 groups	Retained camera views
Excluded measurement	20 groups; multiple masks per group	Excluded camera views
Known synthetic object	20 phantoms	Normalized 3D reference
Controlled mismatch	20 shared phantoms	Synthetic camera + object

Figure S10: **Morphology strata and endpoint inventory.** Panel (a) reports rank agreement by morphology, with five phantoms per family on a fixed $[-1, 1]$ scale; the axial-centroid column uses the error itself, so its sign retains that endpoint's direction. Panel (b) maps each evidence target to its unit: groups for supplied/excluded measurements, phantoms for normalized objects, and shared phantoms for controlled mismatch. The inventory makes the CEV aggregation units explicit.

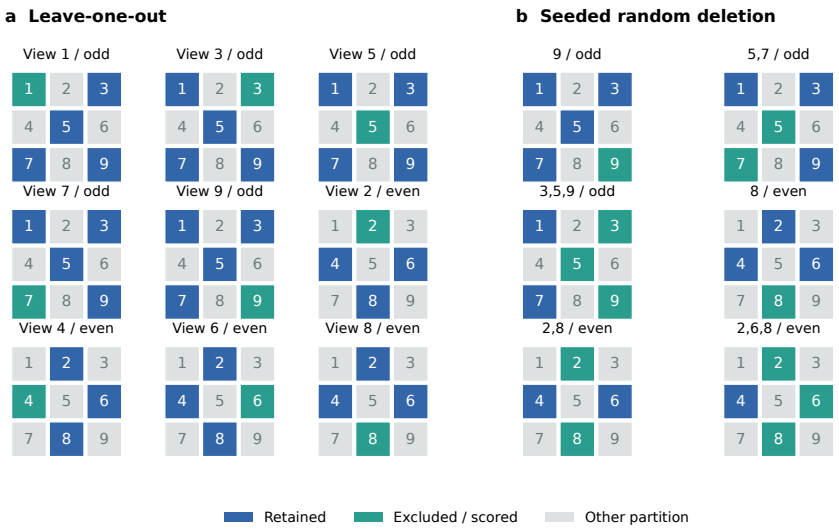


Figure S11: **Frozen final mask atlas.** Nine leave-one-out masks and six random-deletion masks define the final CEV access contracts. Blue tiles are retained, cyan tiles are scored, and gray tiles lie outside the branch partition; labels are one-based in the display and zero-based in Appendix A.3. Corpus masks are within-partition tests because frame-level wavelength provenance is absent, while monochrome phantoms provide the same-channel angular test.

A.6 All Prospective Measurement Checkpoints

This table is the provenance layer for the measurement contract: it retains all 36 rule-specific files and the selection-frozen RL-128 row. The active and held-out ranks are intentionally shown at file level, while the CCS conclusions use 23 unique learned states so repeated aliases do not become extra evidence. The numerical rank columns preserve the file ordering, including rule aliases.

Table A4: Complete prospective measurement ranking on groups 30–49 under nine leave-one-view-out masks, aggregated by group. SSIM is compared per column across the displayed checkpoints and RL-128; higher is favorable, and the iterative reference is not compute-matched. Color conventions follow Appendix A.5. Seed, epoch, and file-order ranks are uncolored.

System	Seed	Selection	Epoch	Supplied-view SSIM ↑	Excluded-view SSIM ↑	Supplied-view file rank	Excluded-view file rank
Positive HYPER	S3	Full	23	0.6326	**0.4933**	20	2
Positive HYPER	S3	Loss	23	0.6326	**0.4933**	21	3
Signed HYPER	S2	Full	23	0.7159	0.4829	17	4
Signed HYPER	S2	Loss	23	0.7159	0.4829	18	5
Calibra-LF	S3	Full	23	0.7272	0.4790	7	6
Calibra-LF	S3	Loss	23	0.7272	0.4790	8	7
Calibra-LF	S3	Active	22	0.7243	0.4789	9	8
Positive HYPER	S1	Full	23	0.6318	0.4761	22	9
Positive HYPER	S1	Loss	23	0.6318	0.4761	23	10
Signed HYPER	S3	Full	23	0.7289	0.4732	5	11
Signed HYPER	S3	Loss	23	0.7289	0.4732	6	12
Positive HYPER	S1	Active	18	0.6205	0.4725	25	13
Positive HYPER	S3	Active	16	0.6302	0.4720	24	14
Calibra-LF	S1	Full	23	0.7229	0.4711	10	15
Calibra-LF	S1	Loss	23	0.7229	0.4711	11	16
Positive HYPER	S2	Full	23	0.6169	0.4686	27	17
Positive HYPER	S2	Loss	23	0.6169	0.4686	28	18
Calibra-LF	S2	Full	23	**0.7322**	0.4681	2	19
Calibra-LF	S2	Loss	23	**0.7322**	0.4681	3	20
Calibra-LF	S2	Active	23	**0.7322**	0.4681	4	21
Signed HYPER	S2	Active	21	0.7121	0.4673	19	22
Calibra-LF	S1	Active	21	0.7212	0.4658	15	23
RL-128	–	Fixed iter.	–	**0.9970**	**0.7157**	1	1

System	Seed	Selection	Epoch	Supplied-view SSIM ↑	Excluded-view SSIM ↑	Supplied-view file rank	Excluded-view file rank
Positive HYPER	S2	Active	22	0.6183	0.4639	26	24
Signed HYPER	S1	Full	23	0.7221	0.4636	12	25
Signed HYPER	S1	Loss	23	0.7221	0.4636	13	26
Signed HYPER	S1	Active	23	0.7221	0.4636	14	27
Signed HYPER	S3	Active	20	0.7171	0.4547	16	28
XLFMNet	S2	Full	21	0.3861	0.3397	29	29
XLFMNet	S2	Active	23	0.3815	0.3332	30	30
XLFMNet	S2	Loss	22	0.3784	0.3323	31	31
XLFMNet	S1	Loss	23	0.3534	0.3249	32	32
XLFMNet	S3	Full	23	0.3469	0.3237	33	33
XLFMNet	S3	Loss	23	0.3469	0.3237	34	34
XLFMNet	S3	Active	23	0.3469	0.3237	35	35
XLFMNet	S1	Active	21	0.3414	0.3132	36	36
XLFMNet	S1	Full	19	0.3370	0.3074	37	37

Among the 36 learned checkpoint files, the supplied-view and excluded-view top-five sets have zero overlap. Table A4 also lists RL-128 as a separate iterative reference and records checkpoint aliases; the CCS analysis ranks the 23 distinct learned states. After deduplication, supplied-view versus excluded-view Spearman is 0.691 across all core learned states, falls to 0.079 among the 16 HYPER-derived states, and has 46.7% discordant pairs. The file-level table supplies the provenance, while the state-level values answer the local-selection question.

A.7 Statistical Units and Stratified Rankings

Table A5: Supplied-view-to-excluded-view ranking on final corpus groups by candidate population: All core learned states (23), HYPER-derived states (16), Signed HYPER + Calibra-LF (10), and individual families. These strata identify which candidates each score ranks. For a five-state stratum, top-five overlap covers the entire stratum; Spearman, Kendall, and discordant-pair rate describe the ordering within it. Population diagnostics remain unranked.

Population	States n	Spearman ρ	Kendall τ	Discordant pairs	Top-5 overlap
All learned unique states	23	0.691	0.557	22.1%	2/5
HYPER family	16	0.079	0.067	46.7%	2/5
Signed HYPER + Calibra-LF	10	0.285	0.200	40.0%	4/5
Signed HYPER	5	-0.100	0.000	50.0%	5/5
Calibra-LF	5	0.400	0.400	30.0%	5/5
Positive HYPER	6	0.886	0.733	13.3%	4/5
XLFMNet	7	1.000	1.000	0.0%	5/5

The first table in Appendix A.7 answers the CCS question directly: how the supplied-view and excluded-view ordering changes with candidate population. Within the 16-state HYPER-derived stratum, 46.7% of candidate pairs reverse and top-five overlap is 2/5; signed HYPER plus Calibra-LF has Spearman 0.285. It is read alongside Figure S4, which adds object endpoints to the same population logic.

The paired checkpoint-rule comparison gives the corresponding run-level consequence. Active-view selection lowers final excluded-view SSIM in eight runs, ties in three, and improves it in one, with a mean change of -0.0058. These rule contrasts remain paired within each optimization trajectory, preserving dependence among states from the same training run.

Table A6: Observed and excluded measurement rankings against normalized object Pearson on the same 20 known volumes. The interval is the paired phantom-bootstrap confidence interval (CI) for excluded minus supplied Spearman; endpoint-oracle regrets report the object-Pearson cost of selecting with each measurement score. Population diagnostics remain unranked.

Population	Supplied-view ρ	Excluded-view ρ	$\Delta\rho$ CI	Supplied-view regret ↓	Excluded-view regret ↓
All learned unique states	0.831	0.823	[-0.056, -0.001]	0.0632	0.0696
HYPER family	0.506	0.482	[-0.153, +0.000]	0.0632	0.0696
Signed HYPER + Calibra-LF	-0.406	-0.503	[-0.170, +0.000]	0.0632	0.0696

We next retain the same candidate strata and evaluate measurement–object agreement on the 20 known-volume phantoms. Supplied-view and excluded-view scores are computed from these phantom measurements and compared with normalized object Pearson. Across all core learned states, supplied-view and excluded-view correlations are 0.831 and 0.823, with endpoint-oracle regrets of 0.0632 and 0.0696. Within the HYPER-derived states, the correlations fall to 0.506 and 0.482; within signed HYPER plus Calibra-LF, both are negative. The regrets quantify how a camera score can alter a structural decision even when the candidate set stays fixed.

Table A7: Prospective selection consequences. Each rule freezes one state per run and a deployment choice on groups 20–29 before final readout. Regret is relative to the optimum of that rule's fixed cohort on final corpus SSIM or known-volume Pearson. Regrets evaluate the single-state choice; the last column applies each rule to three-seed family means. Cohort-dependent regrets diagnose choices and are unranked across rules.

Rule	Selected family	Held-out regret ↓	Object regret ↓	Family-mean choice
Full	Calibra-LF	0.0144	0.0696	Calibra-LF
Loss	Positive HYPER	0.0000	0.1542	Positive HYPER
Active	Calibra-LF	0.0109	0.0000	Calibra-LF

The prospective table is the decision consequence of CCS and CEV together: each rule freezes one state per run before final readout, and the resulting excluded-view and object regrets identify which imaging target that rule serves.

A.8 All Known-Volume Checkpoints

The known-volume table is the third CEV evidence path: 20 phantoms are reconstructed under nine leave-one-view-out masks, producing 180 records per system and 6,660 records over the 36 learned checkpoints plus RL-128. It reports checkpoint aggregates for normalized structure, with phantom-level aggregation retaining mask dependence.

Table A8: Complete known-volume ranking over 20 phantoms and nine leave-one-view-out masks. Each metric is compared across displayed checkpoints and the non-compute-matched RL-128 reference; higher SSIM, Pearson, and top-k overlap and lower NRMSE are favorable. Color conventions follow Appendix A.5. Object rank preserves the file-level Pearson order.

System	Seed	Selection	Excluded-view SSIM ↑	Object Pearson ↑	Top-k overlap ↑	Volume SSIM ↑	NRMSE ↓	Object file rank
RL-128	–	Fixed iter.	**0.9748**	**0.8405**	**0.6534**	**0.9900**	1.9892	1
Calibra-LF	S2	Full	0.3605	**0.2441**	**0.1635**	0.9671	1.1586	2
Calibra-LF	S2	Loss	0.3605	**0.2441**	**0.1635**	0.9671	1.1586	3
Calibra-LF	S2	Active	0.3605	**0.2441**	**0.1635**	0.9671	1.1586	4
Calibra-LF	S1	Full	0.5428	0.2140	0.1422	0.9789	1.1274	5
Calibra-LF	S1	Loss	0.5428	0.2140	0.1422	0.9789	1.1274	6
Signed HYPER	S1	Full	0.5767	0.2072	0.1356	0.9827	1.1089	7
Signed HYPER	S1	Loss	0.5767	0.2072	0.1356	0.9827	1.1089	8
Signed HYPER	S1	Active	0.5767	0.2072	0.1356	0.9827	1.1089	9
Calibra-LF	S1	Active	0.5480	0.2035	0.1344	0.9786	1.1354	10
Signed HYPER	S2	Full	0.5331	0.1885	0.1199	0.9751	1.1313	11
Signed HYPER	S2	Loss	0.5331	0.1885	0.1199	0.9751	1.1313	12
Calibra-LF	S3	Active	0.6686	0.1809	0.1251	0.9831	1.0700	13
Calibra-LF	S3	Full	**0.6729**	0.1744	0.1231	0.9825	1.0854	14
Calibra-LF	S3	Loss	**0.6729**	0.1744	0.1231	0.9825	1.0854	15
Signed HYPER	S2	Active	0.3714	0.1742	0.1127	0.9703	1.1174	16
Signed HYPER	S3	Full	0.6631	0.1644	0.1173	0.9818	1.0776	17
Signed HYPER	S3	Loss	0.6631	0.1644	0.1173	0.9818	1.0776	18
Signed HYPER	S3	Active	0.6224	0.1410	0.1008	0.9813	1.0811	19
Positive HYPER	S2	Full	0.2406	0.1290	0.0518	0.9809	1.0347	20
Positive HYPER	S2	Loss	0.2406	0.1290	0.0518	0.9809	1.0347	21
Positive HYPER	S2	Active	0.2003	0.1284	0.0522	0.9766	1.0501	22
Positive HYPER	S1	Full	0.2389	0.0975	0.0404	0.9797	1.0713	23

System	Seed	Selection	Excluded-view SSIM ↑	Object Pearson ↑	Top-k overlap ↑	Volume SSIM ↑	NRMSE ↓	Object file rank
Positive HYPER	S1	Loss	0.2389	0.0975	0.0404	0.9797	1.0713	24
Positive HYPER	S1	Active	0.2565	0.0926	0.0411	0.9812	1.0830	25
Positive HYPER	S3	Full	0.5249	0.0899	0.0422	**0.9885**	**1.0068**	26
Positive HYPER	S3	Loss	0.5249	0.0899	0.0422	**0.9885**	**1.0068**	27
Positive HYPER	S3	Active	0.3293	0.0827	0.0371	0.9860	**1.0263**	28
XLFMNet	S2	Active	0.0487	0.0793	0.0430	0.8179	1.6482	29
XLFMNet	S2	Full	0.0481	0.0783	0.0427	0.8148	1.6249	30
XLFMNet	S2	Loss	0.0468	0.0780	0.0424	0.8136	1.6625	31
XLFMNet	S1	Loss	0.0337	0.0701	0.0488	0.7547	2.0762	32
XLFMNet	S1	Active	0.0295	0.0652	0.0465	0.7392	2.3836	33
XLFMNet	S1	Full	0.0280	0.0630	0.0454	0.7338	2.4020	34
XLFMNet	S3	Full	0.0319	0.0614	0.0396	0.7397	2.1852	35
XLFMNet	S3	Loss	0.0319	0.0614	0.0396	0.7397	2.1852	36
XLFMNet	S3	Active	0.0319	0.0614	0.0396	0.7397	2.1852	37

The object table makes the target choice concrete: selection-frozen RL-128 has the highest normalized object Pearson, top-k support overlap, and volume SSIM while carrying poor NRMSE; positive HYPER has the best volume SSIM among learned systems and the lowest NRMSE in the learned/RL table, with lower sparse-structure correlation. Supplied and excluded measurement scores correlate with object Pearson at 0.831 and 0.823, with endpoint regrets 0.0632 and 0.0696. These rows teach the reader which object property each metric rewards, rather than producing one universal ranking.

A.9 Phantom Families and Object Metrics

We generate the object-evaluation library for this study: four synthetic morphology strata—sparse beads, clustered emitters, filaments, and mixtures—with five phantoms per family. Gaussian emitters vary in axial/lateral width and amplitude, and filaments interpolate emitter kernels along random three-dimensional line segments. Each target is globally normalized once before projection, fixing the structural reference used by CEV. This in-house phantom construction is separate from the released HYPER training corpus and its generation code.

CEV applies one fixed object transform to every prediction V and reference $V^\star$: clamp at zero, then divide by the volume-wide maximum. Primary Pearson uses all centered voxels; equal-budget top-k overlap sets k from target voxels above 0.2 and measures predicted support recovery. IoU, foreground MSE, normalized RMSE, average slice SSIM, cosine similarity, and normalized z-centroid error remain secondary views of the same normalized structure.

Pearson measures centered voxelwise intensity agreement, while equal-budget top-k overlap measures recovery of the reference support. Volume SSIM and NRMSE provide complementary summaries of normalized voxelwise agreement, including sensitivity to background and normalized amplitude. Pearson and top-k remain the primary object readouts, with the remaining metrics exposing complementary structure and scale behavior.

A.10 Corruption Results

Table A9: Condition-level held-out results on final corpus groups across 12 active-selected learned checkpoints. The corruption suite, including its clean condition, averages nine leave-one-view-out masks and six fixed random-deletion masks. Delta is condition minus clean, and rank ρ is clean-versus-condition Spearman across checkpoints. Conditions are stress tests, so these mean SSIM, difference, and correlation diagnostics remain unranked.

Condition	Excluded-view SSIM ↑	Δ vs clean	Rank ρ
attenuation	0.3819	-0.0253	0.944
clean	0.4072	+0.0000	1.000
lenslet shift	0.3918	-0.0154	0.965
poisson gaussian noise	0.3996	-0.0076	0.986
psf mismatch	0.4073	+0.0001	1.000
registration shift	0.4000	-0.0072	0.972
saturation	0.3844	-0.0228	0.979

Table A10: Model-level corruption results for 12 active-selected checkpoints on final corpus groups. The corruption suite, including its clean condition, averages nine leave-one-view-out masks and six fixed random-deletion masks. Mean and worst SSIM summarize seven condition-level aggregates, including clean; higher is favorable within each column across the 12 checkpoints. Color conventions follow Appendix A.5.

System	Seed	Mean excluded-view SSIM ↑	Worst excluded-view SSIM ↑
Positive HYPER	S3	**0.4623**	**0.4560**
Positive HYPER	S1	**0.4610**	**0.4553**
Positive HYPER	S2	0.4400	0.4319
Signed HYPER	S1	0.4366	0.4229
Calibra-LF	S1	0.4319	0.4166
Signed HYPER	S2	0.4226	0.4068
Calibra-LF	S3	0.4208	0.3955
Calibra-LF	S2	0.4182	0.4052
Signed HYPER	S3	0.4175	0.3813
XLFMNet	S2	0.3015	0.2703
XLFMNet	S3	0.2783	0.1902
XLFMNet	S1	0.2614	0.1955

The two Appendix A.10 tables separate condition-level severity from model-level aggregation. Together they show how excluded-view prediction changes under a stressor, while Appendix A.11 and Figure S5 ask whether the same score tracks normalized object structure.

The corruption tables are robustness diagnostics for the same excluded-view measurement contract; they keep the stressor within the declared candidate set. Attenuation and saturation produce the largest mean drops; the PSF-conditioning row is nearly neutral because it changes metadata while the true PSF remains fixed for projection and the XLFMNet-inspired control ignores that embedding. Appendix A.11 then changes the data-generating operator itself to separate metadata sensitivity from true operator mismatch.

A.11 True Data/Operator Mismatch on Known Synthetic Volumes

The mismatch panel extends CEV from a nominal simulator contract to controlled changes in the data-generating operator. Learned weights stay fixed while the data-generating PSF, additive background, shot/read noise, or per-view throughput changes and the model retains nominal PSF conditioning. The corpus test varies PSF conditioning while retaining the projection operator; the known-

volume test varies the data-generating PSF. These tests separate metadata sensitivity from operator mismatch.

Table A11: Mean change from the nominal full-input condition over 20 known volumes and 23 unique learned states. PSF offsets change the data-generating PSF while reconstruction retains nominal conditioning. Counts separately record positive reprojection-SSIM changes and negative object-Pearson changes. Positive differences increase each score; a reversal requires both directions in the same state. Stressor differences and counts remain unranked.

Condition	Δ reprojection SSIM	Δ object Pearson	States with SSIM +	States with Pearson −
PSF offset 3	-0.0029	-0.0007	1/23	21/23
PSF offset 7	-0.0031	-0.0005	4/23	20/23
PSF offset 15	-0.0045	-0.0017	7/23	21/23
Background 0.01	+0.2199	-0.0025	18/23	17/23
Background 0.03	+0.2582	-0.0078	22/23	17/23
Background 0.06	+0.2538	-0.0264	23/23	23/23
Noise mild	+0.0555	-0.0031	23/23	21/23
Noise moderate	+0.0614	-0.0065	17/23	21/23
Noise severe	+0.0083	-0.0150	14/23	21/23
Throughput mild	-0.0044	-0.0011	6/23	14/23
Throughput moderate	-0.0021	-0.0013	9/23	19/23
Throughput severe	-0.0113	-0.0021	5/23	18/23

Background 0.06 supplies the sharpest CEV counterexample: reprojection SSIM rises in all 23 states (mean $+0.2538$) while normalized object Pearson falls in all 23 (mean -0.0264, hierarchical mean-delta interval $[-0.0336, -0.0215]$). The 23 states are nested within 12 runs, with within-run dependence retained in the hierarchical analysis. Per-view spatial maximum normalization rewards broad low-frequency agreement under additive background, whereas the volume target records the structural change. Reading the two columns together identifies the score's preference and the evidence it leaves unresolved.

For the full-input Calibra-LF filament example in Figure 4, max-normalized reprojection SSIM rises from 0.810 to 0.909 under background 0.06. Reprojection MSE worsens from 0.00321 to 0.00729, independently max-normalized NMSE from 0.001825 to 0.001850, normalized object Pearson falls from 0.318 to 0.284, and object NRMSE worsens from 1.163 to 1.422. MSE uses the frame-normalized measurement scale before the extra per-view normalization; these values preserve the distinction between that residual and the normalized similarity score. Together, these readouts resolve three responses to the same background: frame-scale camera residual, independently normalized image similarity, and agreement with the known three-dimensional object.

Table A12: Nominal full-input endpoints on 20 known volumes, averaged equally over deduplicated states within each of four learned families. Family means are compared per metric, with higher SSIM, Pearson, overlap, F1, and axial correlation and lower peak error favorable. Color conventions follow Appendix A.5. F1@2 uses a two-voxel support tolerance; axial peak error is in simulator-grid planes.

Family	Reprojection SSIM ↑	Object Pearson ↑	Top-k overlap ↑	F1@2 ↑	Axial ρ ↑	Peak error (planes) ↓
Calibra-LF (Ours)	**0.7341**	**0.2734**	**0.1805**	**0.4627**	**0.1482**	**21.29**
Signed HYPER	**0.7044**	**0.2386**	**0.1569**	**0.4064**	**0.1295**	**21.23**
Positive HYPER	0.4309	0.1475	0.0538	0.1640	0.0326	23.33
Adapted XLFMNet	0.1893	0.1303	0.0696	0.2276	0.0116	33.14

The auxiliary endpoints make the object target concrete: distance-tolerant F1 compares dilated support at one, two, or four simulator-grid cells, axial-profile Pearson compares normalized z profiles,

and peak error measures plane displacement. These metrics complement volume Pearson and top-k overlap by measuring spatial tolerance and axial signal allocation. Distances in this table are expressed in simulator-grid units; Figures 3–5 use the nominal sampling intervals stated in Appendix A.5 for display.

A.12 Richardson–Lucy Iteration Selection

Different evaluation targets favor different points along the same Richardson–Lucy trajectory. Selection on groups 20–29 fixes the reference at 128 iterations. On the final known-object evaluation, Pearson reaches its maximum at 64 iterations and NRMSE reaches its minimum at 8 iterations. The trajectory therefore separates the stopping point chosen from measurement data from the optima identified by subsequent object evaluation.

Table A13: RL iteration trajectory. Selection SSIM uses groups 20–29; final SSIM uses groups 30–49 and object scores use 20 known volumes, read after selection is frozen. Values are compared across iterations within each column, with higher SSIM and Pearson and lower NRMSE favorable; final optima are diagnostic. Color conventions follow Appendix A.5.

	Selection SSIM ↑			Final SSIM ↑		Known object	
Iter.	Supplied	Full	Excluded	Supplied	Excluded	Pearson ↑	NRMSE ↓
1	0.4499	0.3982	0.2191	0.4610	0.2197	0.1264	31.2524
2	0.6762	0.5970	0.3224	0.6784	0.3193	0.2166	17.2118
4	0.8732	0.7824	0.4668	0.8740	0.4651	0.4332	7.3744
8	0.9653	0.8835	0.5990	0.9662	0.6021	0.7991	**1.3783**
16	0.9886	0.9166	0.6668	0.9890	0.6706	0.8359	**1.7181**
32	0.9941	0.9268	0.6935	0.9943	0.6983	**0.8419**	1.8493
64	**0.9959**	**0.9307**	**0.7044**	**0.9961**	**0.7101**	**0.8427**	1.9249
128	**0.9968**	**0.9325**	**0.7091**	**0.9970**	**0.7157**	0.8405	1.9892

Selection-active, selection-full, and selection-held-out freeze iterations 128, 128, and 128. Their final held-out/object-Pearson pairs are 0.7157/0.8405, 0.7157/0.8405, and 0.7157/0.8405. The table keeps NRMSE as an object diagnostic because the corpus selection contract contains measurements rather than an object target.

A.13 Augmentation and Corruption Sensitivity

Augmentation averaging preserves the group as the independent unit while checking whether the measurement ranking depends on augmentation 0. Replacing augmentation 0 with the ten-variant group mean gives active and held-out rank stability 0.996 and 0.977; active-to-held-out Spearman is 0.717 overall, 0.159 within HYPER, and 0.321 within signed HYPER plus Calibra-LF.

Severity curves separate access effects from operator effects. At severity 1.5, attenuation and saturation reach mean held-out deltas of -0.0692 and -0.0742, while lenslet and registration shifts mainly reduce active fit. Conditioning offsets remain near zero with a fixed data generator and projector; Appendix A.11 supplies the complementary true-mismatch evidence.

A.14 Cutoff and Additional Learned-Control Diagnostics

The continuation experiment starts from each full-selected checkpoint and adds 12 training epochs at the original learning rate with a fresh optimizer state. Full SSIM improves in 12/12 runs and loss improves in 12/12, with the new boundary selected in 10/12 and 9/12 runs. The control examines continuation along each training trajectory within the same model family.

Table A14: Mean corruption-minus-clean SSIM over one active-selected checkpoint per learned run. Conditions are stressors within the same candidate set, so the table reports changes rather than a model ranking. The PSF row changes conditioning metadata; Appendix A.11 changes the data-generating operator.

Severity	Condition	Δ active SSIM	Δ held-out SSIM
0.5	attenuation	-0.0212	-0.0072
0.5	lenslet shift	-0.0169	-0.0044
0.5	poisson gaussian noise	-0.0018	-0.0016
0.5	psf mismatch	+0.0001	+0.0001
0.5	registration shift	-0.0189	-0.0032
0.5	saturation	+0.0009	-0.0046
1.0	attenuation	-0.0579	-0.0253
1.0	lenslet shift	-0.0622	-0.0154
1.0	poisson gaussian noise	-0.0074	-0.0076
1.0	psf mismatch	+0.0002	+0.0001
1.0	registration shift	-0.0441	-0.0072
1.0	saturation	-0.0078	-0.0228
1.5	attenuation	-0.1326	-0.0692
1.5	lenslet shift	-0.1184	-0.0280
1.5	poisson gaussian noise	-0.0152	-0.0147
1.5	psf mismatch	+0.0002	+0.0001
1.5	registration shift	-0.0942	-0.0152
1.5	saturation	-0.0665	-0.0742

Table A15: Three-seed means for expanded XLFMNet+CWFA and original XLFMNet controls under the local FLFM projector. Measurement SSIM uses final corpus groups; object metrics use 20 known volumes. Means are compared between the two controls per metric, with higher SSIM and Pearson and lower NRMSE favorable. Color conventions follow Appendix A.5. Both controls are local adaptations rather than official XLFM reproductions.

System	Supplied-view SSIM ↑	Excluded-view SSIM ↑	Object Pearson ↑	Object SSIM ↑	Object NRMSE ↓
XLFMNet+CWFA	**0.6602**	**0.4215**	**0.4131**	**0.8728**	**1.3138**
XLFMNet	**0.3566**	**0.3233**	**0.0686**	**0.7656**	**2.0723**

Table A16: Representative network-forward cost under the common $1 \times 1 \times 768 \times 768$ input contract. Approximate PyTorch FLOPs and local latency exclude the common FLFM projector. Costs are compared per column among displayed systems, with lower values favorable. Color conventions follow Appendix A.5. Cost ranks provide local implementation context, independent of reconstruction quality.

System	Parameters ↓	Approx. GFLOPs ↓	Forward (ms) ↓	Output shape
Signed HYPER	**3,700,836**	**96.91**	**8.16**	1×128×256×256
Calibra-LF	3,702,908	97.15	8.88	1×128×256×256
Positive HYPER	**3,700,836**	**96.91**	8.28	1×128×256×256
XLFMNet	**27,264**	**6.23**	**2.14**	1×128×256×256
XLFMNet+CWFA	128,941,279	1329.31	81.17	1×128×256×256

A.15 Expanded Learned-Model Comparison

The expanded learned-model pool contains 30 training runs across ten model families, with three seeds per family. All runs use the same HYPER/FLFM protocol, groups 0–19 for optimization, groups 20–29 for checkpoint selection, and selection supplied-view mean SSIM as the ranking metric. Partition SSIM averages odd/even masked-input reprojection; it is a software-partition score, while the core 12-run population supplies the deduplicated CEV diagnosis.

Table A17: Expanded family-level matched-protocol training results on selection groups 20–29. Best and mean summarize runs within each family for the named metric; different best columns can come from different runs. Aggregates are compared across families per column, with higher SSIM and lower loss favorable. Color conventions follow Appendix A.5. These selection scores are separate from final-endpoint analysis.

Family	Best selection supplied-view SSIM ↑	Mean selection supplied-view SSIM ↑	Best partition SSIM ↑	Best full SSIM ↑	Best loss ↓
Calibra-LF	**0.790961**	**0.760641**	0.748534	0.709452	0.087685
HYPER-Net	**0.787721**	**0.762893**	0.746373	0.699020	0.087913
HYPER + prior fusion	0.773730	0.731509	0.741914	0.694976	0.054621
XLFM/CWFA + SLNet-prior	0.689924	0.663806	**0.773036**	**0.809266**	**0.044392**
Positive HYPER	0.672824	0.648923	0.658858	0.620676	0.052302
XLFM/CWFA-wide	0.662496	0.652863	0.727762	0.730560	0.052889
XLFM/CWFA + SLNet-prior-wide	0.649823	0.647126	**0.765209**	**0.810650**	**0.047298**
XLFM/CWFA	0.648048	0.634006	0.740275	0.765095	0.050363
XLFMNet-wide	0.492221	0.463509	0.524944	0.540741	0.093725
XLFMNet	0.392496	0.370941	0.435020	0.474528	0.111556

Table A18: All 30 matched-protocol training runs, scored on selection groups 20–29. Selection supplied-view SSIM is the pre-declared ranking metric; partition/full SSIM and loss remain secondary readouts. Values are compared across runs per metric column, with higher SSIM and lower loss favorable. Color conventions follow Appendix A.5. Rank, seed, and epoch are uncolored.

Rank	Family	Seed	Selection supplied-view SSIM ↑	Partition SSIM ↑	Full SSIM ↑	Loss ↓	Epoch
1	Calibra-LF	76203	**0.790961**	0.748534	0.709452	0.087685	23
2	HYPER-Net	76203	**0.787721**	0.746373	0.698925	0.087913	23
3	HYPER-Net	76201	0.778181	0.742596	0.699020	0.088057	23
4	Calibra-LF	76201	0.774856	0.747044	0.700746	0.087941	23
5	HYPER+prior	76201	0.773730	0.741914	0.694976	0.088117	23
6	HYPER+prior	76203	0.769522	0.726145	0.687890	0.089976	23
7	HYPER-Net	76202	0.722776	0.722222	0.685119	0.090901	16
8	Calibra-LF	76202	0.716106	0.724265	0.691390	0.090632	13
9	XLFM/CWFA+SLNet	76203	0.689924	0.751378	0.799196	0.047569	16
10	Positive HYPER	76202	0.672824	0.658858	0.620676	0.052302	23
11	XLFM/CWFA+SLNet	76202	0.666627	**0.768297**	**0.809266**	0.049665	6
12	XLFM/CWFA-wide	76202	0.662496	0.709058	0.700194	0.068684	20
13	XLFM/CWFA-wide	76201	0.657538	0.727762	0.730560	0.052889	22
14	HYPER+prior	76202	0.651275	0.650449	0.625793	0.054621	23
15	XLFM/CWFA+SLNet-wide	76203	0.649823	0.765209	0.792618	0.047396	13
16	Positive HYPER	76201	0.649789	0.612907	0.582184	0.061045	19
17	XLFM/CWFA+SLNet-wide	76202	0.648807	0.740495	**0.810650**	0.053101	7
18	XLFM/CWFA	76202	0.648048	0.688747	0.726181	0.064002	20
19	XLFM/CWFA+SLNet-wide	76201	0.642747	0.746009	0.800850	**0.047298**	8
20	XLFM/CWFA-wide	76203	0.638554	0.718238	0.715721	0.055056	21
21	XLFM/CWFA+SLNet	76201	0.634869	**0.773036**	0.797908	**0.044392**	9
22	XLFM/CWFA	76203	0.630648	0.707031	0.738075	0.057870	18
23	Positive HYPER	76203	0.624156	0.593925	0.561998	0.064008	21
24	XLFM/CWFA	76201	0.623323	0.740275	0.765095	0.050363	18
25	XLFMNet-wide	76201	0.492221	0.524944	0.540741	0.093725	22
26	XLFMNet-wide	76202	0.474876	0.448370	0.514119	0.103685	22
27	XLFMNet-wide	76203	0.423429	0.402672	0.516395	0.116120	20
28	XLFMNet	76201	0.392496	0.435020	0.474528	0.111556	22
29	XLFMNet	76203	0.366771	0.332630	0.469051	0.136507	22
30	XLFMNet	76202	0.353556	0.431355	0.471386	0.112564	22

A.16 Comparison Pools and Reference Methods

We organize methods by their role in the study. The core pool supports candidate-stratified selection analysis, and the expanded learned-model pool broadens the architectural comparison under the local HYPER/FLFM protocol. Historical configurations provide development-stage comparisons, while additional published methods provide literature context.

Table A19: Comparison pools and reference methods. The categorical rows distinguish local numerical comparisons, historical development results, and literature context; they are not metric-ranked.

Method group	Role in this study	Evidence source
Expanded learned-model pool	Architectural comparison	Results for 30/30 local HYPER/FLFM runs, including HYPER-Net; family summaries and individual runs appear in Appendix A.15 and Table A18.
Core 12-run/RL pool	Candidate-stratified selection and object evaluation	Complete checkpoint comparisons in Appendices A.6–A.10, using the declared candidate hierarchy and evidence targets.
Recent published methods	Literature context	34 methods screened by publication venue and reported performance; most have no local FLFM-protocol numerical comparison.
Historical exploratory methods	Development-stage comparisons	Historical coverage and controls under their original protocols, without a recent-top-venue screening restriction; separate from the current matched FLFM comparison.
VsLFM	Literature context	Published reconstruction method (Lu et al., 2023); no local HYPER/FLFM numerical result is included.
SeReNet	Literature context	Published physics-driven self-supervised method (Lu et al., 2025); no local HYPER/FLFM numerical result is included.

A.17 Known-Volume Qualitative Comparisons

The qualitative panels give the object branch a spatial reading alongside Pearson, top-k, and NRMSE. Every row shares the synthetic target and nominal simulator for its phantom, while display normalization remains separate from the quantitative aggregation. The displayed Calibra-LF state belongs to the matched-protocol family. These examples visualize individual reconstruction outputs from different seeds, complementing the six seed-paired numerical comparisons in Appendix A.19.

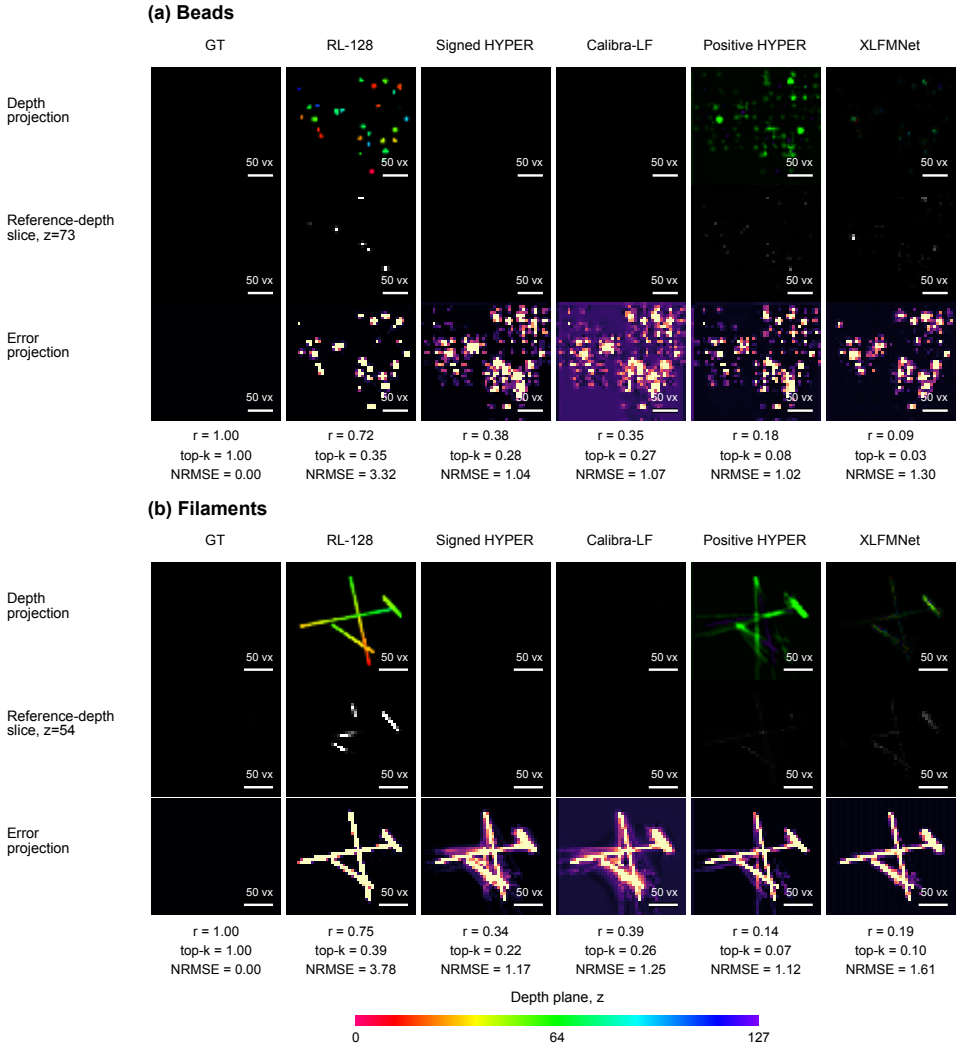


Figure S12: Additional synthetic-object display for sparse beads and filaments, with the reference, RL-128, signed HYPER, Calibra-LF, positive HYPER, and adapted XLFMNet control. Each phantom uses a common display scale and crop distinct from the quantitative volume normalization; fixed slices use the reference maximum-energy plane and errors use a shared within-phantom scale. Hue encodes the simulator depth plane, white bars span 50 lateral simulator voxels, and full-volume Pearson, top-k, and NRMSE remain the quantitative object readouts.

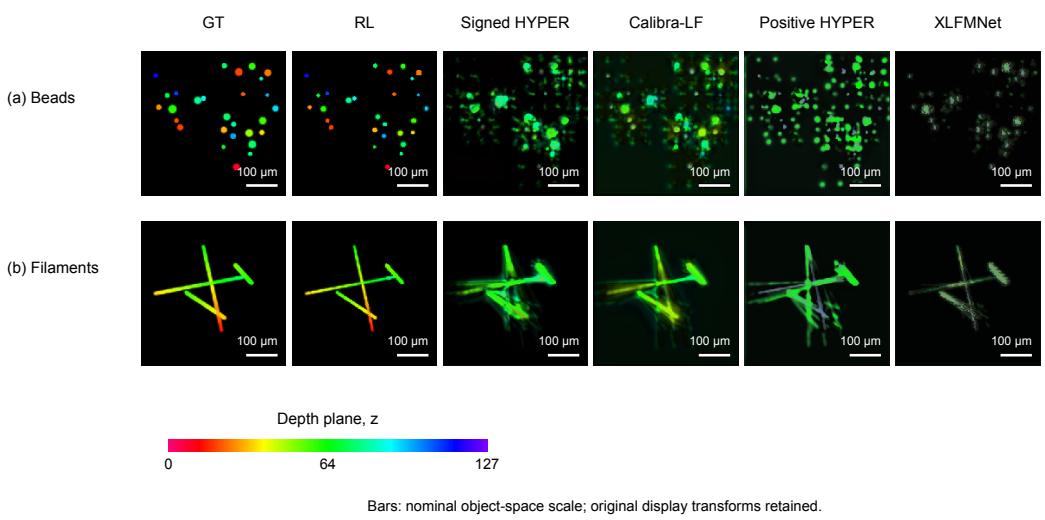


Figure S13: Additional depth-encoded XY maximum-intensity projections for representative bead and filament phantoms. Methods within each phantom share the synthetic target and nominal FLFM simulator; hue encodes the simulator depth plane and brightness encodes normalized intensity. The panel uses robust-normalized volumes, complementing the maximum-normalized main views.

The additional slices fix the target maximum-energy plane and normalize each displayed slice independently; error MIPs follow the projection manifest. These historical depth overlays combine intensity-weighted colors across planes; the main XY MIPs encode peak-plane depth. The axial profiles add a depth-distribution readout, while the tables retain the quantitative object metrics.

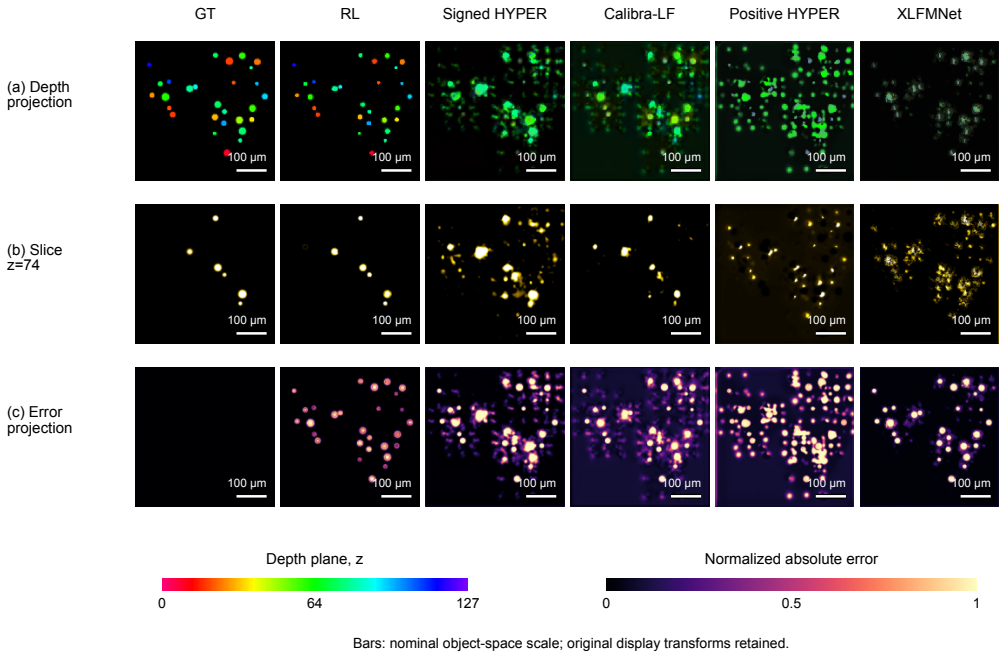


Figure S14: Historical synthetic sparse-bead comparison. Rows show depth-encoded XY MIPs, independently normalized slices at the reference-chosen $z = 74$ (zero-based), and absolute-error MIPs relative to GT. Errors use independently nonnegative robust-normalized volumes. Columns: GT, RL, signed HYPER, Calibra-LF, positive HYPER, and the adapted XLFMNet control.

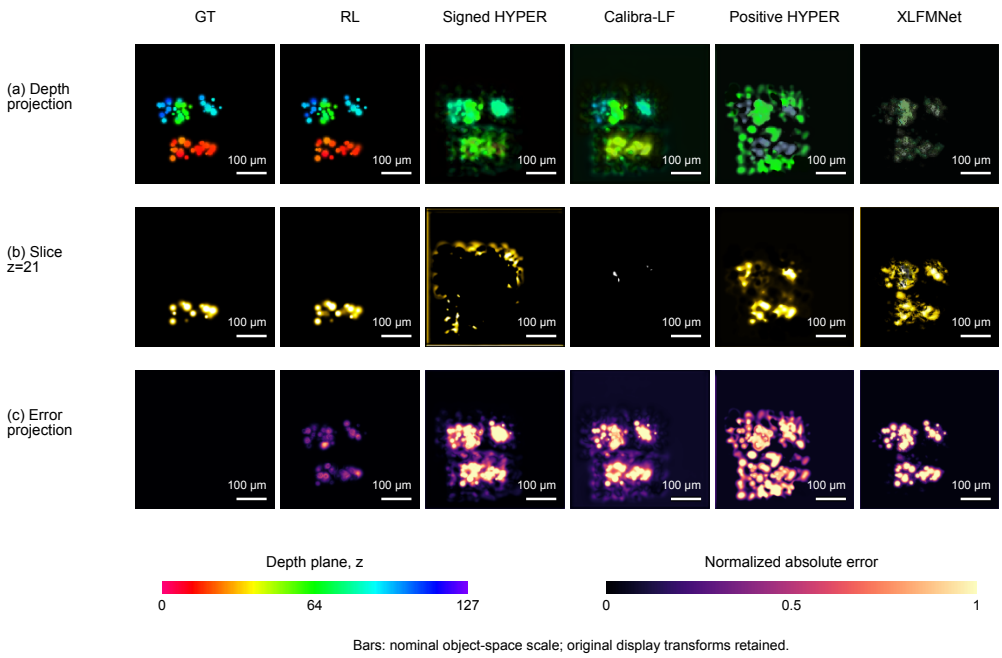


Figure S15: Historical synthetic clustered-emitter comparison. Rows show depth-encoded XY MIPs, independently normalized slices at the reference-chosen $z = 21$ (zero-based), and absolute-error MIPs relative to GT. Errors use independently nonnegative robust-normalized volumes. Columns: GT, RL, signed HYPER, Calibra-LF, positive HYPER, and the adapted XLFMNet control.

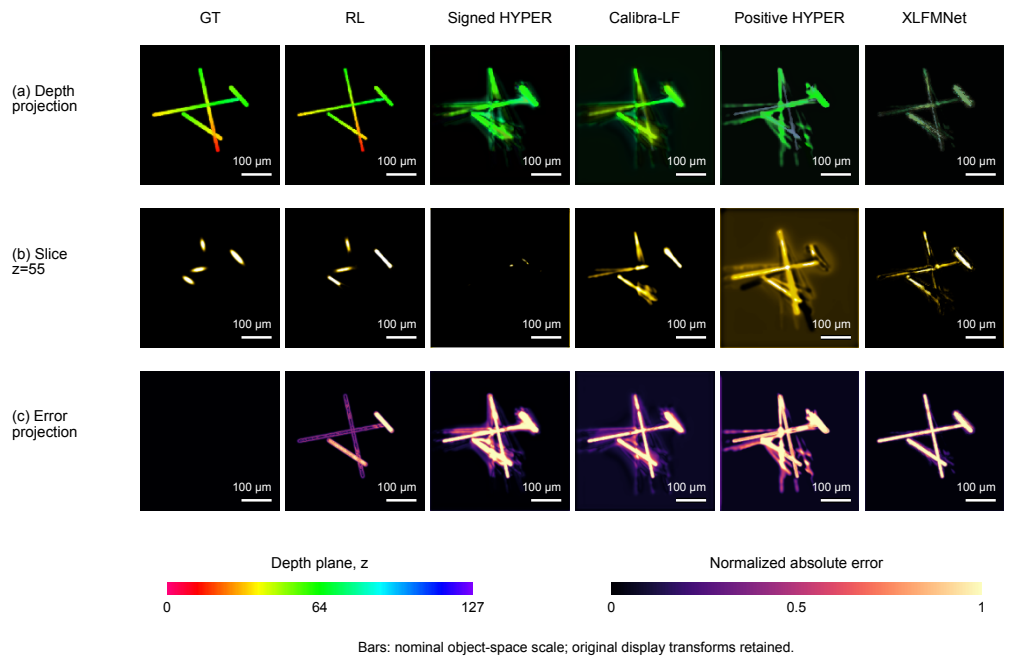


Figure S16: Historical synthetic filament comparison. Rows show depth-encoded XY MIPs, independently normalized slices at the reference-chosen $z = 55$ (zero-based), and absolute-error MIPs relative to GT. Errors use independently nonnegative robust-normalized volumes. Columns: GT, RL, signed HYPER, Calibra-LF, positive HYPER, and the adapted XLFMNet control.

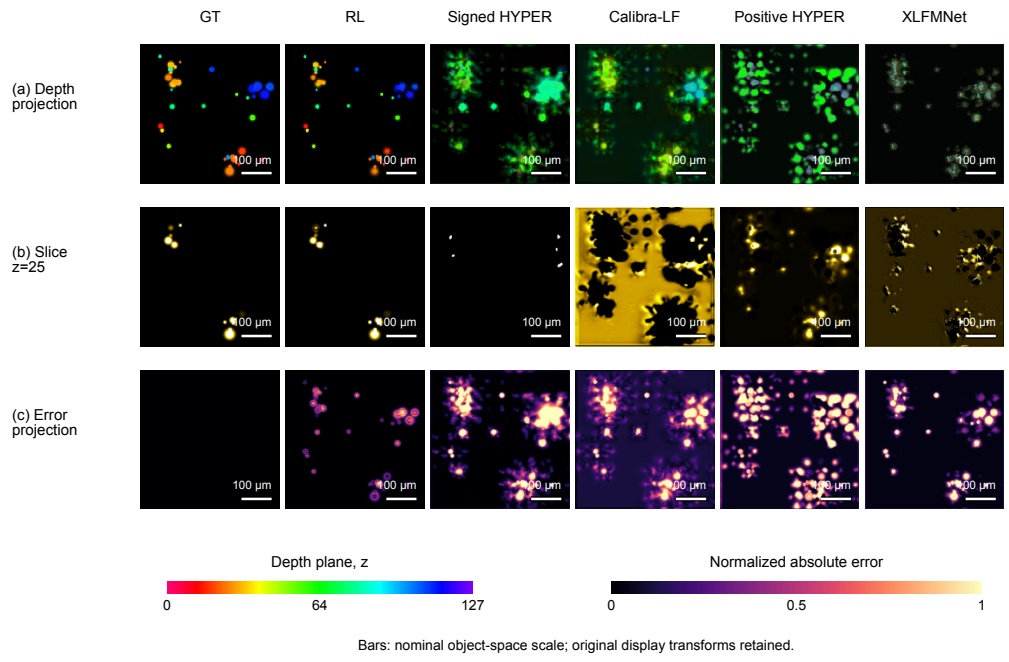


Figure S17: Historical synthetic mixed-structure comparison. Rows show depth-encoded XY MIPs, independently normalized slices at the reference-chosen $z = 25$ (zero-based), and absolute-error MIPs relative to GT. Errors use independently nonnegative robust-normalized volumes. Columns: GT, RL, signed HYPER, Calibra-LF, positive HYPER, and the adapted XLFMNet control.

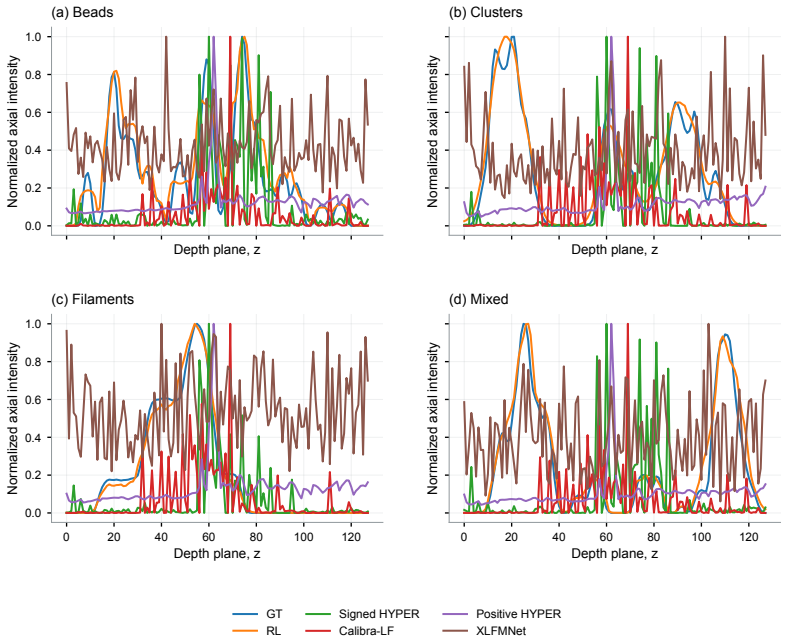


Figure S18: Historical synthetic sparse-bead, clustered-emitter, filament, and mixed-structure axial intensity profiles for GT, RL, signed HYPER, Calibra-LF, positive HYPER, and the adapted XLFMNet control. Each curve sums the nonnegative robust-normalized volume over X, Y at each simulator z plane and is independently scaled to unit peak.

A.18 HISTORICAL SYNTHETIC FLFM QUALITATIVE COMPARISON

This development-stage comparison shows how depth calibration changes reconstructed punctate structures and reprojection scores. The early depth-calibrated configuration is described below; the matched Calibra-LF comparison appears in Appendix A.19.

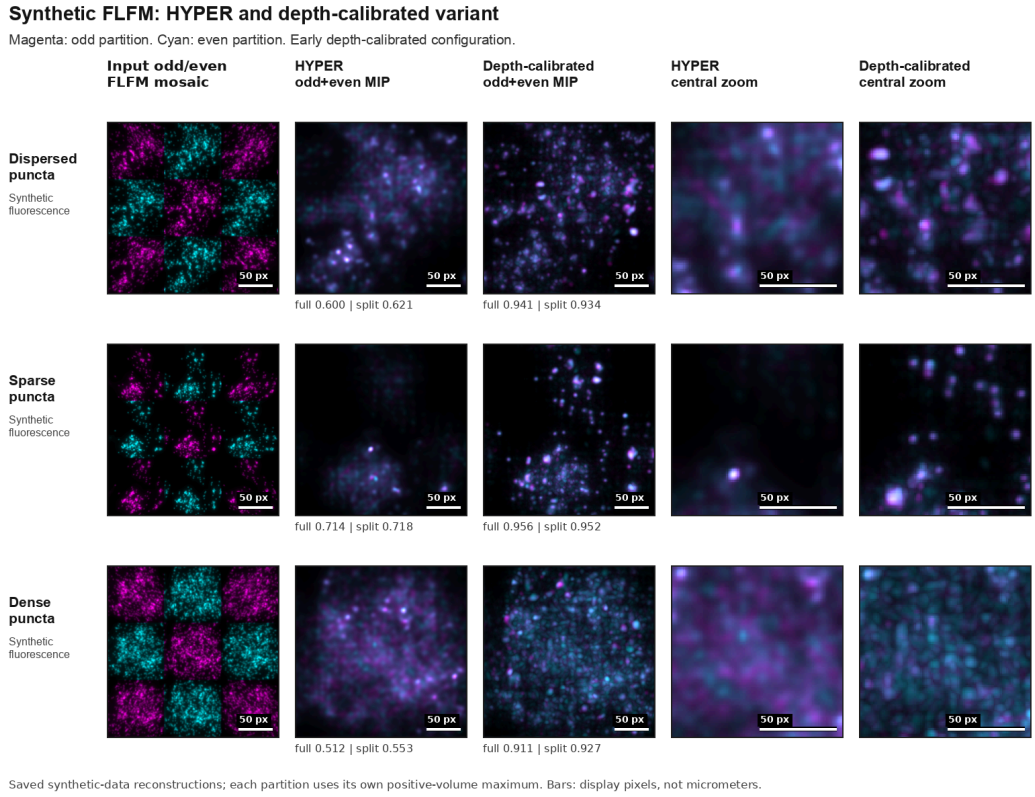


Figure S19: Historical comparison of synthetic FLFM measurements with dispersed, sparse, and dense puncta (top to bottom). The input and branch MIPs use the fixed percentile, crop, and per-volume normalization recorded in the display contract; odd-magenta and even-cyan colors identify software partitions. The warm-started base-width-48 depth-calibrated variant uses depth gain 0.08 and depth mixing 0.25, while the matched Calibra-LF case uses the base-width-32 protocol. Reported values are full-input and masked-input reprojection SSIM. The samples have PSF/slope metadata without paired object references or frame-level wavelength labels; their identities remain in the figure's provenance record.

The released dataset contains synthetic FLFM frames with PSF/slope metadata and no paired volume references. Its two reconstructions differ in configuration and training history; magenta and cyan mark the software-partition branches. The panel therefore contributes provenance for the historical record while the current CEV claims remain tied to the frozen synthetic contracts.

A.19 COMPLETE CALIBRA-LF–HYPER MATCHED PAIRS

Across six distinct state pairs nested within three matched seeds, Calibra-LF increases clean-corpus held-out SSIM by +0.0043 and normalized object Pearson by +0.0297 on average, with improvements in 5/6 of the pairs and 5/6 of the pairs, respectively. The object-Pearson interval is $[+0.0015, +0.0607]$, while the clean-corpus held-out SSIM interval crosses zero. Tables A20–A22 show how the matched candidates respond under the other measurement, object, and corruption metrics. Table A20 retains all nine seed–rule file pairs; repeated epoch pairs collapse to six distinct pairs in Table A21.

Table A20: Calibra-LF minus signed HYPER for every file pair matched by seed, budget, split, and selection rule, on 20 known volumes. Columns are signed differences: positive favors Calibra-LF except for NRMSE, where negative is favorable. Duplicate epochs remain for provenance; differences are unranked. Clean-corpus SSIM appears in Table A21.

Seed	Rule	Δ excluded-view SSIM	Δ object Pearson	Δ top-k overlap	Δ volume SSIM	Δ NRMSE
S1	F	-0.03387	+0.00678	+0.00652	-0.00382	+0.01849
S1	L	-0.03387	+0.00678	+0.00652	-0.00382	+0.01849
S1	A	-0.02865	-0.00367	-0.00124	-0.00416	+0.02652
S2	F	-0.17262	+0.05558	+0.04357	-0.00806	+0.02722
S2	L	-0.17262	+0.05558	+0.04357	-0.00806	+0.02722
S2	A	-0.01090	+0.06987	+0.05078	-0.00321	+0.04117
S3	F	+0.00978	+0.01006	+0.00576	+0.00076	+0.00774
S3	L	+0.00978	+0.01006	+0.00576	+0.00076	+0.00774
S3	A	+0.04623	+0.03984	+0.02429	+0.00178	-0.01110

The first matched table preserves every file alias so the provenance layer remains auditable. Its repeated epochs motivate the six-pair collapse used in Figure S8.

Table A21: Six unique Calibra-LF-minus-signed-HYPER run–epoch pairs after duplicate rule files are collapsed. Held-out SSIM uses clean final corpus groups; other columns use 20 known volumes. Clean-corpus scores use the nine leave-one-view-out masks. Positive differences favor Calibra-LF except for NRMSE, where negative is favorable. C/H denotes Calibra-LF/HYPER epochs. State-level differences remain unranked.

Seed	Epochs C/H	Source rules	Δ excluded-view SSIM	Δ object Pearson	Δ top-k overlap	Δ volume SSIM	Δ NRMSE
S1	21/23	Active	+0.00214	-0.00367	-0.00124	-0.00416	+0.02652
S1	23/23	Full/Loss	+0.00747	+0.00678	+0.00652	-0.00382	+0.01849
S2	23/21	Active	+0.00074	+0.06987	+0.05078	-0.00321	+0.04117
S2	23/23	Full/Loss	-0.01486	+0.05558	+0.04357	-0.00806	+0.02722
S3	22/20	Active	+0.02428	+0.03984	+0.02429	+0.00178	-0.01110
S3	23/23	Full/Loss	+0.00577	+0.01006	+0.00576	+0.00076	+0.00774

The unique-pair table aligns clean-corpus excluded-view reprojection SSIM with object Pearson, top-k, volume SSIM, and NRMSE on the same six state pairs. The clean condition in the corruption suite includes the additional random-deletion masks, giving it a different aggregation scope from Table A21.

Known-volume excluded-view SSIM, mean corruption SSIM, and volume SSIM decrease on average, while normalized RMSE increases. Together with the clean-corpus and Pearson gains summarized above, these results show a target-dependent profile within three seeds: model selection depends on which imaging property is sought.

Table A22: All 21 Calibra-LF-minus-signed-HYPER corruption pairs: three seeds and seven conditions on final corpus groups. The corruption suite, including its clean condition, averages nine leave-one-view-out masks and six fixed random-deletion masks. Positive SSIM differences favor Calibra-LF. Paired differences remain unranked.

Seed	Condition	Excluded-view reprojection SSIM delta
S1	attenuation	-0.02055
S1	clean	-0.00131
S1	lenslet shift	-0.00028
S1	poisson gaussian noise	-0.00035
S1	psf mismatch	-0.00131
S1	registration shift	-0.00024
S1	saturation	-0.00863
S2	attenuation	-0.00740
S2	clean	-0.00252
S2	lenslet shift	-0.00788
S2	poisson gaussian noise	-0.00631
S2	psf mismatch	-0.00250
S2	registration shift	-0.00272
S2	saturation	-0.00154
S3	attenuation	+0.00395
S3	clean	+0.00116
S3	lenslet shift	-0.00459
S3	poisson gaussian noise	+0.00844
S3	psf mismatch	+0.00113
S3	registration shift	-0.00062
S3	saturation	+0.01421

A.20 Selection Coverage and Evaluation Sensitivity

The development-stage comparisons examine how partition labels and validation coverage affect reconstruction scores. An earlier odd/even adaptation assigned software partitions to reported filter labels; the released synthetic frames lack frame-level wavelength provenance, so these labels identify software partitions and the cross-partition scores are partition diagnostics. The Selected checkpoint uses 16 draws over 12 distinct frames and Prior uses 40 draws over 26 distinct frames, with deterministic sampling with replacement within five reused development groups. The following comparisons quantify the consequences of these sampling choices and motivate the fixed masks, group units, and prospective split used in the current benchmark.

Historical group-delta audit

Table A23: Historical Selected-minus-Prior SSIM by base group. O→E and E→O denote odd-to-even and even-to-odd software partitions. Positive differences favor Selected for that score. These paired differences are unranked provenance diagnostics, separate from the matched Calibra-LF study.

Group	Δ full	Δ active mean	Δ O→E cross	Δ E→O cross	Δ cross mean
45	-0.00110	+0.00387	-0.00424	+0.01243	+0.00410
46	-0.00033	+0.00624	+0.00390	+0.00484	+0.00437
47	-0.00219	+0.00270	-0.00719	+0.01754	+0.00518
48	-0.00064	+0.00561	+0.00180	+0.01325	+0.00753
49	+0.00190	+0.01075	-0.00700	-0.00400	-0.00550

The group table records how a checkpoint change travels across evidence contracts. The active mean is positive in 5/5 groups with descriptive interval [0.00364, 0.00847], while the cross-partition mean

is positive in 4/5 groups with interval [-0.00145, 0.00621]. Five reused groups supply a transparent historical audit unit.

HISTORICAL VALIDATION-DRAW RECONSTRUCTION

Table A24: Deterministic validation-draw counts from each development group for the historical Prior and Selected checkpoints. Sampling is with replacement; counts include repeated frames and carry no performance ranking.

Group	Prior (40 draws)	Selected (16 draws)
45	5	5
46	9	1
47	8	6
48	12	3
49	6	1

Prior uses 26 unique frames in 40 draws; selected uses 12 unique frames in 16 draws. The deterministic sequence repeats at every validation epoch, and its coverage changes the readout when extended to all 50 frames: active SSIM changes by 0.00437 and full SSIM by 0.00848, with corresponding prior changes 0.00088 and 0.00164. These values motivate the fixed group-level split in the current contract.

HISTORICAL CONFIGURATION AND SELECTION DIFFERENCES

Table A25 summarizes the configurations used by the two historical checkpoints.

Table A25: Configuration and selection differences between the historical Prior and Selected checkpoints, separate from matched Calibra-LF. The rows record the joint protocol that produced each historical score profile; these metadata are unranked.

Field	Prior	Selected
Checkpoint selection	launch flag omitted; default unresolved	active-partition mean SSIM
Validation draws	40	16
Unique validation frames	26	12
Seed	151000	629015
Warm start	frozen source A	frozen source B
Learning rate	5×10^{-6}	10^{-5}
Scheduled epochs	72	48
Selected epoch	69	44
Per-view normalized MSE weight	0.2	0.4
Depth-gain scale	0.08	0.10
Depth-smoothness weight	0.00025	0.0005
Depth-collapse weight	0	0.01

HISTORICAL EVALUATOR SCALE AND POSITIVITY

The evaluator scale fixes how the historical scores should be read. The loader divides each frame by its maximum before network input, and per-view SSIM applies an independent maximum normalization to prediction and target. MSE, intensity ratios, and the Poisson-deviance-like diagnostic therefore live on the normalized, clamped evaluator scale; calibrated camera intensity and photon statistics require their own reference.

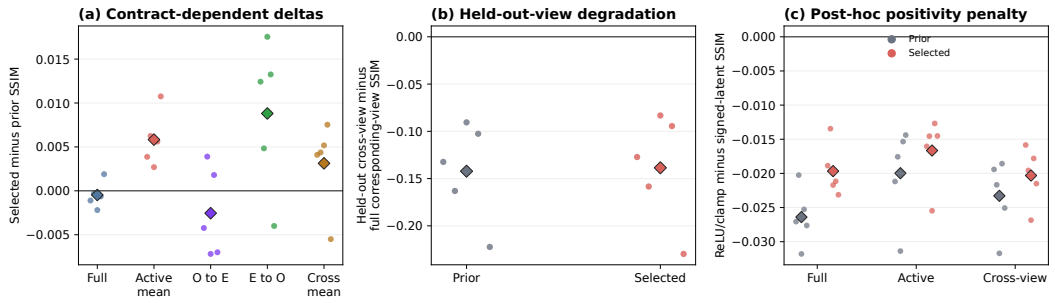


Figure S20: Group-level contract and post-hoc positivity sensitivity. Circles are the five reused development groups and diamonds are group means. The figure shows how evaluator scale and output sign alter the historical readout; trained positivity comparisons use the current candidate protocol.

The prior and selected negative-voxel fractions are 0.66993 and 0.62221. Post-hoc clamp decreases full SSIM by 0.02640 and 0.01966, and cross-partition mean by 0.02328 and 0.02031. Figure S20 shows group-level contract and positivity sensitivity.

COMPLETE 50-FRAME HISTORICAL AUDIT

The following table reports all 50 paired development frames. Historical Prior/Selected checkpoints are compared within each frame's SSIM or MSE pair, with higher SSIM and lower MSE favorable; negative-voxel fraction remains an unranked output diagnostic. Color conventions follow Appendix A.5. Group aggregation and selection reuse remain visible, keeping this audit separate from the matched Calibra-LF comparison and current CCS–CEV evidence.

Table A26: Complete audit of 50 historical frames. Prior and Sel. denote the earlier two checkpoints with unmatched configurations and training histories, separate from the matched Calibra-LF study. Performance values are compared within each frame and Prior/Selected metric pair, with higher SSIM and lower measurement MSE favorable. Color conventions follow Appendix A.5. MSE uses the frame-normalized measurement scale. Negative-voxel fractions are unranked diagnostics.

Group	Aug.	Full SSIM↑		Active mean↑		Cross mean↑		Measurement MSE↓		Negative fraction	
		Prior	Sel.	Prior	Sel.	Prior	Sel.	Prior	Sel.	Prior	Sel.
45	0	**0.94614**	**0.94493**	**0.94245**	**0.94642**	**0.79869**	**0.80484**	**0.022787**	**0.022659**	0.68200	0.63248
45	1	**0.94966**	**0.94836**	**0.94578**	**0.94951**	**0.82809**	**0.83102**	**0.025303**	**0.025163**	0.67603	0.62673
45	2	**0.95143**	**0.94912**	**0.94512**	**0.94899**	**0.80568**	**0.81121**	**0.023543**	**0.023411**	0.67161	0.62508
45	3	**0.95468**	**0.95500**	**0.94949**	**0.95274**	**0.82297**	**0.82748**	**0.022038**	**0.021914**	0.65330	0.61231
45	4	**0.95185**	**0.95164**	**0.94388**	**0.94864**	**0.82938**	**0.83692**	**0.021369**	**0.021247**	0.65838	0.61714
45	5	**0.95130**	**0.95007**	**0.94880**	**0.95180**	**0.83262**	**0.83230**	**0.023309**	**0.023178**	0.66324	0.61802
45	6	**0.94902**	**0.94839**	**0.94306**	**0.94690**	**0.80554**	**0.81107**	**0.022345**	**0.022220**	0.66770	0.62354
45	7	**0.95312**	**0.95127**	**0.94959**	**0.95375**	**0.83245**	**0.83206**	**0.026591**	**0.026444**	0.67053	0.62232
45	8	**0.95563**	**0.95512**	**0.94621**	**0.95052**	**0.82749**	**0.83149**	**0.024997**	**0.024857**	0.66595	0.61994
45	9	**0.94774**	**0.94564**	**0.94243**	**0.94625**	**0.80693**	**0.81240**	**0.021706**	**0.021584**	0.66553	0.62257
46	0	**0.92389**	**0.92450**	**0.92944**	**0.93428**	**0.76123**	**0.75844**	**0.027826**	**0.027674**	0.68026	0.62768
46	1	**0.93358**	**0.93494**	**0.93706**	**0.94384**	**0.76428**	**0.77413**	**0.033343**	**0.033166**	0.67593	0.62141
46	2	**0.92644**	**0.92460**	**0.92931**	**0.93439**	**0.75775**	**0.75887**	**0.026653**	**0.026506**	0.68493	0.63041
46	3	**0.92991**	**0.92951**	**0.92766**	**0.93715**	**0.75993**	**0.77072**	**0.029977**	**0.029816**	0.68665	0.63075
46	4	**0.93255**	**0.93425**	**0.93192**	**0.94020**	**0.78440**	**0.79102**	**0.018995**	**0.018887**	0.65953	0.61580
46	5	**0.92490**	**0.92602**	**0.92471**	**0.93357**	**0.76563**	**0.77540**	**0.023477**	**0.023347**	0.67343	0.62428
46	6	**0.92990**	**0.93041**	**0.93206**	**0.93825**	**0.77986**	**0.78167**	**0.021909**	**0.021785**	0.66082	0.61470
46	7	**0.93363**	**0.93148**	**0.93322**	**0.93613**	**0.77104**	**0.77491**	**0.028119**	**0.027964**	0.67212	0.62108
46	8	**0.92780**	**0.92539**	**0.92766**	**0.93358**	**0.75625**	**0.75850**	**0.030042**	**0.029879**	0.68029	0.62670
46	9	**0.92886**	**0.92712**	**0.93307**	**0.93711**	**0.76368**	**0.76407**	**0.031445**	**0.031277**	0.66974	0.61801
47	0	**0.95204**	**0.95042**	**0.94388**	**0.94785**	**0.84448**	**0.84854**	**0.011298**	**0.011229**	0.64713	0.62152
47	1	**0.96006**	**0.95671**	**0.95007**	**0.95457**	**0.87375**	**0.87584**	**0.007592**	**0.007541**	0.62384	0.60584
47	2	**0.95375**	**0.95187**	**0.95283**	**0.95479**	**0.86619**	**0.87794**	**0.010588**	**0.010520**	0.63278	0.60793
47	3	**0.95437**	**0.95123**	**0.94823**	**0.95077**	**0.86280**	**0.87255**	**0.010413**	**0.010347**	0.63403	0.60870
47	4	**0.96242**	**0.95892**	**0.95171**	**0.95684**	**0.87816**	**0.88232**	**0.008730**	**0.008672**	0.62595	0.60524

Group	Aug.	Full SSIM↑		Active mean↑		Cross mean↑		Measurement MSE↓		Negative fraction	
		Prior	Sel.	Prior	Sel.	Prior	Sel.	Prior	Sel.	Prior	Sel.
47	5	**0.96066**	**0.95822**	**0.95540**	**0.95450**	**0.87957**	**0.88169**	**0.010407**	**0.010341**	0.63376	0.61020
47	6	**0.95701**	**0.95477**	**0.94764**	**0.94914**	**0.86996**	**0.87364**	**0.009333**	**0.009273**	0.63118	0.61183
47	7	**0.95854**	**0.95633**	**0.94817**	**0.95008**	**0.87086**	**0.86622**	**0.010363**	**0.010296**	0.63287	0.61152
47	8	**0.95899**	**0.95743**	**0.95366**	**0.95525**	**0.86690**	**0.87345**	**0.010058**	**0.009994**	0.63127	0.60676
47	9	**0.95786**	**0.95785**	**0.95239**	**0.95713**	**0.86088**	**0.87312**	**0.010248**	**0.010181**	0.64738	0.62163
48	0	**0.94292**	**0.94130**	**0.93826**	**0.94451**	**0.82236**	**0.83244**	**0.021481**	**0.021363**	0.69706	0.64480
48	1	**0.94181**	**0.94211**	**0.93751**	**0.94514**	**0.84400**	**0.85179**	**0.015496**	**0.015405**	0.67262	0.62895
48	2	**0.94876**	**0.95022**	**0.94460**	**0.94985**	**0.85051**	**0.85989**	**0.020443**	**0.020328**	0.67808	0.62921
48	3	**0.94574**	**0.94576**	**0.94134**	**0.94719**	**0.84058**	**0.84497**	**0.021151**	**0.021033**	0.68913	0.63765
48	4	**0.94711**	**0.94281**	**0.93892**	**0.94287**	**0.84556**	**0.85006**	**0.016018**	**0.015924**	0.68835	0.64020
48	5	**0.94814**	**0.94670**	**0.94326**	**0.94888**	**0.83925**	**0.84672**	**0.021278**	**0.021160**	0.68413	0.63465
48	6	**0.94749**	**0.94743**	**0.94355**	**0.94897**	**0.85110**	**0.85688**	**0.014930**	**0.014842**	0.66638	0.62261
48	7	**0.94481**	**0.94514**	**0.93821**	**0.94689**	**0.84525**	**0.85371**	**0.020247**	**0.020132**	0.68608	0.63579
48	8	**0.95040**	**0.95004**	**0.94201**	**0.94541**	**0.85399**	**0.86152**	**0.016658**	**0.016564**	0.67485	0.62913
48	9	**0.95049**	**0.94978**	**0.94544**	**0.94951**	**0.85292**	**0.86280**	**0.017175**	**0.017076**	0.66935	0.62461
49	0	**0.90465**	**0.90784**	**0.92229**	**0.93036**	**0.66310**	**0.65838**	**0.057812**	**0.057519**	0.70266	0.62485
49	1	**0.90679**	**0.90717**	**0.91558**	**0.92654**	**0.66691**	**0.66749**	**0.047263**	**0.047023**	0.70319	0.62822
49	2	**0.91827**	**0.91703**	**0.92557**	**0.93799**	**0.69587**	**0.69109**	**0.046436**	**0.046196**	0.68559	0.61633
49	3	**0.91578**	**0.91997**	**0.91654**	**0.92954**	**0.70178**	**0.69715**	**0.039421**	**0.039217**	0.69804	0.62833
49	4	**0.91670**	**0.91980**	**0.92620**	**0.93830**	**0.69694**	**0.69449**	**0.046563**	**0.046321**	0.68772	0.61941
49	5	**0.91131**	**0.91434**	**0.92166**	**0.93284**	**0.68532**	**0.67907**	**0.052657**	**0.052389**	0.69787	0.62360
49	6	**0.91152**	**0.91349**	**0.92158**	**0.93252**	**0.70997**	**0.70160**	**0.040508**	**0.040294**	0.69604	0.62627
49	7	**0.91880**	**0.91821**	**0.92787**	**0.93681**	**0.69814**	**0.68700**	**0.048497**	**0.048246**	0.69011	0.61903
49	8	**0.91737**	**0.91966**	**0.93012**	**0.93826**	**0.70477**	**0.70087**	**0.042638**	**0.042414**	0.67451	0.61170
49	9	**0.91508**	**0.91775**	**0.92711**	**0.93892**	**0.69271**	**0.68335**	**0.052827**	**0.052557**	0.69648	0.62279

A.21 Exact Historical Validation Draws

This draw table makes the historical selection exposure inspectable. Repeated identities show the fixed with-replacement sequence directly, which is why those records are retained as provenance rather than pooled with the current group-level ranking.

Table A27: Exact deterministic validation draw sequence reconstructed from the frozen sampling code. Samples are identified by Group and Augmentation; repeated pairs denote repeated validation observations of the same frame.

Checkpoint	Draw	Group	Augmentation
Prior	0	46	6
Prior	1	48	0
Prior	2	45	9
Prior	3	46	4
Prior	4	47	1
Prior	5	46	6
Prior	6	48	7
Prior	7	49	4
Prior	8	47	1
Prior	9	46	1
Prior	10	49	6
Prior	11	48	7
Prior	12	46	6
Prior	13	46	9
Prior	14	49	3
Prior	15	46	6
Prior	16	48	0
Prior	17	48	2
Prior	18	47	0
Prior	19	49	8
Prior	20	48	9
Prior	21	48	5
Prior	22	48	5
Prior	23	45	8
Prior	24	49	8
Prior	25	47	0
Prior	26	49	7
Prior	27	45	1
Prior	28	47	1
Prior	29	47	5
Prior	30	45	3
Prior	31	47	2
Prior	32	46	3
Prior	33	47	0
Prior	34	48	8
Prior	35	48	2
Prior	36	48	0
Prior	37	48	9
Prior	38	46	7
Prior	39	45	7
Selected	0	45	3
Selected	1	47	2
Selected	2	45	4
Selected	3	45	8
Selected	4	49	3
Selected	5	47	3

Checkpoint	Draw	Group	Augmentation
Selected	6	47	2
Selected	7	45	4
Selected	8	47	2
Selected	9	47	1
Selected	10	45	0
Selected	11	48	6
Selected	12	48	8
Selected	13	46	5
Selected	14	48	7
Selected	15	47	2

A.22 Record Counts and Anonymized Data

The record counts show the scale of the declared contracts: $37 \times 20 \times 9 = 6{,}660$ clean measurement records, another 6,660 known-volume records, and $12 \times 20 \times 15 \times 7 = 25{,}200$ corruption records. The Calibra-LF analysis adds nine primary pairs, nine known-volume pairs, and 21 corruption pairs. Machine-readable JSON retains every raw record, while the PDF presents complete checkpoint aggregates and selected pair tables so the reader can audit both coverage and dependence. The historical archive contains 4,840 parseable records and 29 unreadable/incomplete files; it remains a provenance resource with no pooled ranking.

Table A28: Complete experiment inventory at freeze time. The table uses scientific group and configuration descriptions for the frozen aliases; every parseable record remains counted, while unreadable or incomplete source files are reported separately.

Study group	Configuration description	Records
Historical XLFM development	XLFMNet/CWFA-lite fusion and local prior variants for exploratory XLFM volume reconstruction.	642
Historical XLFM development	CWFA-lite encoder-based XLFM reconstruction sweeps; not a full normalizing-flow reproduction.	1591
Historical XLFM development	SLNet image-consistency sweeps and metric auditing, separate from volume-reconstruction comparisons.	275
Historical XLFM development	Local SLNet-inspired sparse/low-rank prior tuning on the XLFMNet/CWFA-lite fusion model.	96
Historical XLFM development	SSIM-prioritized loss weighting and fusion/gating variants for XLFM reconstruction.	25
Historical XLFM development	Warm-start refinement of SSIM loss weights and spatial gating, retaining PSNR as a secondary check.	16
Historical XLFM development	XLFMNet optimization over data subsets, crop/depth settings, and training hyperparameters.	1591
Historical XLFM development	Selected combinations of depth-dependent fusion, uncertainty gating, and high-frequency residual refinement.	8
Historical XLFM development	Gradient-consistency loss refinement on the selected XLFMNet/CWFA-lite fusion and local-prior model.	10
Historical XLFM development	Supplementary comparisons of training, selection, regularization, calibration, and data-subset choices.	8
Historical XLFM development	Component screen including fusion, depth/spatial gating, residual refinement, and conditioning on the XLFMNet/CWFA-lite model.	10
Historical HYPER/FLFM development	HYPER base-width 32/48 comparisons with SSIM-weight and learning-rate refinement.	10

Study group	Configuration description	Records
Historical HYPER/FLFM development	Selected depth-gain/raw-conditioned variants and HYPER controls, including PSF-index and objective-repair comparisons.	10
Historical HYPER/FLFM development	Warm-start HYPER-derived reconstruction with fixed odd/even view partitions and full/partition reprojection objectives.	10
Historical HYPER/FLFM development	Smoke tests of odd/even view-partition reprojection training for HYPER-derived models.	2
Historical HYPER/FLFM development	Exploratory HYPER FLFM comparisons of continued training, residual priors, and depth regularization.	19
Historical HYPER/FLFM development	Smoke test of a residual-prior HYPER variant and its FLFM reprojection workflow.	1
Historical HYPER-to-LFC adaptation	Supervised HYPER-to-LFC adaptation using widefield-stack targets and image-statistic or zero conditioning instead of measured PSF features.	11
Historical HYPER-to-LFC adaptation	Longer training and loss refinement of selected HYPER-to-LFC adaptations with image-statistic conditioning and widefield-stack targets.	13
Historical HYPER/FLFM development	Longer HYPER continuations of selected high-ranking configurations; exploratory reprojection results.	3
Historical HYPER/FLFM development	Early HYPER and locally adapted XLFM-family reprojection attempts on released FLFM data; superseded after output/projection defects.	21
Historical HYPER/FLFM development	Reprojection attempts after removing hard output rectification; superseded because PSF normalization overflow remained.	12
Historical HYPER/FLFM development	HYPER and local comparator reprojection tests after repairing PSF normalization overflow.	23
Historical HYPER/FLFM development	HYPER-derived comparisons of nonnegative outputs, negative-volume penalties, depth-collapse penalties, and odd/even balance.	10
Historical HYPER/FLFM development	Smoke test of nonnegative-output and volume-regularization settings in HYPER-derived reconstruction.	1
Historical HYPER/FLFM development	Workflow checks for repaired view/PSF indexing and hybrid-PSF handling, including minimal training smoke tests.	4
Historical HYPER/FLFM development	HYPER continuation and residual-prior comparisons after per-sample hybrid-PSF and index-mapping repairs.	21
Historical HYPER/FLFM development	Refinement of a visually selected HYPER checkpoint through learning rate, structural-loss weighting, and depth regularization.	10
Historical HYPER/FLFM development	HYPER component comparisons of depth gain, depth-local mixing, and raw-conditioned corrections against unmodified controls.	10
Historical HYPER/FLFM development	Smoke test of HYPER component loading and the FLFM projector.	1
Historical LFC/HR-FLFM development	Local XLFMNet, CWFA-fusion, and prior-model pilots using LFC widefield-stack supervision.	47

Study group	Configuration description	Records
Historical HYPER/FLFM development	No-training check of FLFM view extraction and PSF/slope indexing; not reconstruction-performance evidence.	1
Historical HYPER/FLFM development	Focused HYPER-derived component, objective, output-activation, and PSF-setting comparisons after an initial adaptive screen.	60
Historical HYPER/FLFM development	Additional adaptive refinement of selected HYPER-derived component and regularization settings.	50
Historical HYPER/FLFM development	Initial adaptive HYPER-derived screen of components, losses, output activation, and PSF settings under odd/even reprojection.	100
Historical HYPER/FLFM development	Smoke test of HYPER-derived component training with odd/even reprojection; a workflow check.	1
Historical FLFM and XLFM probes	HYPER-derived FLFM continuation checks and separate XLFM reconstruction probes; not cross-instrument validation.	10
Historical LFC/HR-FLFM development	Bead-only, biological-only, and pooled supervised training with within/cross-family widefield-stack validation.	10
Historical LFC/HR-FLFM development	Initial raw-light-field-to-widefield-stack supervised proof with depth, crop, loss, and residual-scale variants; validation only.	34
Historical LFC/HR-FLFM development	Smoke test of raw-light-field-to-widefield-stack supervised training and validation.	1
Historical LFC/HR-FLFM development	Warm-start supervised refinement from SSIM- and PSNR-selected checkpoints, varying loss balance and residual scale; validation only.	11
Historical LFC/HR-FLFM development	Supervised refinement of residual scale, crop context, and axial-gradient loss; validation only.	10
Historical LFC/HR-FLFM development	Selected-checkpoint continuations combining larger crops, loss balance, and residual scales; validation only.	10
Historical LFC/HR-FLFM development	Low-learning-rate continuations with nearby residual scales, SSIM/MSE weights, and intermediate crops; validation only.	10
Historical XLFM calibration transfer	Reduced-budget smoke test of measured-PSF synthetic-bead training, checkpoint selection, and biological-target evaluation.	1
Historical XLFM calibration transfer	Measured-PSF synthetic-bead supervision with varied calibration distributions, followed by biological native-target evaluation.	10
Historical XLFM calibration transfer	Repeated low-photon and clustered calibration training with matched controls; biological targets used only after synthetic-based selection.	10
All groups	Parsed experiment records	4840
All groups	Unreadable/incomplete result files	29
All groups	Source result files	4869

A.23 Evaluation Consistency

We maintain consistent candidate identities, input access, and aggregation throughout evaluation.

Data access and masks. Optimization and selection groups are disjoint, and prospective training and checkpoint selection exclude the final groups from their data discovery. Validation uses fixed records rather than sampling with replacement. The exact training, selection, and final masks are distinct. Known-volume targets are generated from a frozen seed before model comparison.

Checkpoint identity and selection. Each learned run saves loss-, full-, and active-selected checkpoints. Duplicate states are collapsed by run and epoch, and each fixed-rule cohort contains one frozen state per training run. Prospective regret uses selection-set scores to choose a candidate; final-endpoint maximization is reported separately as endpoint-oracle regret.

Paired aggregation and resampling. Group and phantom aggregation retains dependence among masks. Bootstrap sampling uses groups or phantoms rather than treating masks or checkpoint-rule files as independent samples, and within-run rule contrasts preserve their paired trajectories.

Result and figure generation. Evaluation records retain finite supplied-view and excluded-view scores together with group, mask, condition, and provenance information. Paper tables are generated from anonymous aggregates, and figure metadata identify their source records and display transforms. Best and second-best colors are assigned only within a comparable metric column.

A.24 Reproduction and Claim Boundary

The paper's reproducibility files contain anonymous aggregate JSON, generated TeX rows, figure builders, and the evaluation definitions needed to reproduce the tables and figures. Training and inference additionally use the released synthetic corpus, HYPER PSF data, and model modules. The recorded environment comprises two 64-core AMD EPYC 7B13 sockets, ten NVIDIA GeForce RTX 3080 GPUs with 10,240 MiB, Linux 5.15, Python 3.10.12, PyTorch 2.11.0 with CUDA 13.0, NumPy 2.2.6, SciPy 1.15.3, Matplotlib 3.10.8, and tifffile 2025.5.10; independent runs used one process per GPU with capped CPU math threads.

Data and comparison scope. We evaluate normalized synthetic FLFM structure using the released measurement corpus and 20 known reference volumes. Controlled mismatch varies the data-generating PSF, background, noise, or throughput within the same simulator family. Selection is fixed before the final readout of the rerun; the final groups had been accessible during earlier development. The HYPER corpus supplies neither paired experimental 3D references nor frame-level wavelength provenance. The XLFMNet control is adapted to the local FLFM operator, rather than an official XLFM reproduction, and RL uses the selection-frozen iteration specified in Appendix A.12, with an unmatched compute budget. Calibra-LF is near-parameter-matched to signed HYPER through six distinct state pairs nested within three seeds. Its paired results are reported separately for clean-corpus prediction, known-object structure, normalized error, and corruption response, including gains, losses, and the clean-corpus interval crossing zero. This joint-component comparison supports endpoint-specific model choice; separate component ablations and matched external comparisons would be needed to establish component-specific effects or broader superiority. Experimental reference volumes and instrument calibration provide corresponding targets for extending these comparisons to anatomy, fluorescence intensity, photon statistics, dual-color separation, and instrument transfer; other imaging models require their own evaluations.

A.25 Calibra-LF Definition and Optimization Details

Appendix A.25 fixes the Calibra-LF forward path so the matched case study can be reproduced under the same CCS candidate identity and CEV evidence contracts. Calibra-LF tests a model modification within SCOPE-FLFM. HYPER-Net contributes the PSF-conditioned inverse backbone, FiLM, and self-supervised reprojection. The matched configuration has base width 32, $D = 128$ output planes, and embedding dimension 256. The input is a globally max-normalized $1 \times 768 \times 768$ synthetic FLFM mosaic Y; M_K zeros every tile outside the supplied set K. The backbone stacks its nine 256×256 tiles as channels. The PSF slope vector $s_h \in \mathbb{R}^{18}$ conditions the backbone, while the PSF tensor h defines the projector.

The inherited PSF encoder maps s_h through an $18 \to 256 \to 512 \to 256 \to 256$ MLP with intermediate LeakyReLU activations. FiLM applies $\gamma(e_h) \odot U + \beta(e_h)$ to normalized feature maps U, where e_h is this embedding. The U-Net uses encoder widths 32/64/128, a 128-channel bottleneck, and bilinear upsampling with concatenated skip features. Its single-volume head has three 3×3 convolutions, $32 \to 128 \to 256 \to 128$, each followed by LeakyReLU. Signed HYPER and Calibra-LF have no additional positivity activation; the positive-HYPER control applies softplus during training and inference.

Let $V_0^{(K)} = F_{\mathrm{HYPER}}(M_K Y, s_h)$ denote the signed backbone volume, $a \in \mathbb{R}^D$ the learned depth-gain parameter, and R_ϕ the shallow residual network. The **Bounded Depth Gain** and **Raw-Conditioned Residual** combine as

$$V^{(K)} = [1 + 0.1 \tanh(a)] \odot V_0^{(K)} + 0.02 R_\phi(\mathcal{I}(M_K Y)), \tag{A.1}$$

where $\mathcal{I}$ bilinearly resizes the *masked* raw mosaic to 256×256 using pixel-center alignment. The residual is $\mathrm{Conv}_{1\times1}^{8\to128} \circ \mathrm{GELU} \circ \mathrm{Conv}_{3\times3}^{8\to8} \circ \mathrm{GELU} \circ \mathrm{Conv}_{3\times3}^{1\to8}$, with biases and unit padding in the 3×3 layers. We initialize $a = 0$ and both weights and bias of the final residual projection to zero, so the fused output initially equals the backbone output. The gain broadcasts over lateral positions and lies between 0.9 and 1.1, acting independently at each depth. The residual receives only the measurements retained by M_K.

Both components act after the inherited volume head and before the unchanged projector. Depth bias and depth mixing are disabled. The 2,072 additional registered parameters comprise 128 depth-gain parameters, 1,816 residual-network parameters, and 128 inactive depth-bias parameters; the resulting overhead is 0.056%. Both components preserve the PSF index, masks, targets, depth, and checkpoint rule. The image-formation model remains inherited from HYPER.

Implemented projector. The CEV measurement branch uses the implemented projector exactly, including its chunk-wise magnitude accumulation. The operator $\mathcal{P}_h$ in the main text includes chunk-wise magnitude accumulation, which makes its action on signed volumes nonlinear. With 16 depth blocks and zero-based depth indices $\mathcal{Z}_c = \{8(c-1), \ldots, 8c-1\}$,

$$\mathcal{P}_h(V) = \mathcal{P}_{\mathrm{impl},h}(V) = \mathcal{E}\left[\sum_{c=1}^{16} \left| \sum_{z \in \mathcal{Z}_c} \mathcal{A}_{h,z}(V_z) \right| \right]. \tag{A.2}$$

$\mathcal{A}_{h,z}$ includes zero-padding each 256×256 plane to 768×768, the upstream FFT padding and depth-specific PSF convolution, FFT shift, and central crop to the PSF grid. $\mathcal{E}$ extracts the nine 256×256 elemental views at the recorded lenslet coordinates. The implementation sums eight convolved planes *before* absolute value, then accumulates 16 chunks. Replacing this by a signed depth sum, a per-plane absolute sum, or an absolute value after summing all depths would change the implemented operator.

Shared-weight optimization and single-volume inference. The three forwards define one shared candidate state and one volume per branch. For $K \in \{\mathrm{full}, \mathrm{odd}, \mathrm{even}\}$, each forward predicts one $128 \times 256 \times 256$ volume and reprojects it with Equation A.2. All three forwards share the backbone, gain, and residual weights; there are no separate color heads or fusion of branch volumes. Let ℓ_K score only views in K against their original normalized measurements, using

$$\ell_{\mathrm{repr}}(\widehat{Y}, Y) = \mathrm{MSE}(\widehat{Y}, Y) + 0.4\,\mathrm{NMSE}(\widehat{Y}, Y) + 0.04\big(1 - \mathrm{SSIM}(\widehat{Y}, Y)\big), \tag{A.3}$$

$$\mathcal{L} = 0.5\ell_{\mathrm{full}} + \ell_{\mathrm{odd}} + \ell_{\mathrm{even}} + 0.0005 Z_2(V^{(\mathrm{full})}) + 0.01 C(V^{(\mathrm{full})}). \tag{A.4}$$

For one volume, the regularizers are defined as follows; training averages them over the batch:

$$Z_2(V) = \operatorname*{mean}_{1 \le z \le D-2,\, x,y} |2V_{z,x,y} - V_{z-1,x,y} - V_{z+1,x,y}|, \tag{A.5}$$

$$e_z(V) = \operatorname*{mean}_{x,y} \max(V_{z,x,y}, 0), \qquad p_z(V) = \frac{e_z(V)}{\max(\sum_j e_j(V), 10^{-8})}, \tag{A.6}$$

$$C(V) = \sum_{z=0}^{D-1} p_z(V)^2. \tag{A.7}$$

The axial regularizer penalizes the mean absolute second difference between neighboring depth planes. The concentration regularizer sums the squared normalized positive-depth mean energies. Both act on the full-input volume and are added to the weighted reprojection objective in Equation A.4, outside the 0.5 full-input reprojection weight. Training uses the camera measurements as supervision. Spatial TV, negative-volume, and odd/even energy-balance penalties are disabled in the matched run. The trainer accumulates gradients from the three forwards before one Adam update, with learning rate 5×10^{-5} and zero weight decay. Standard inference uses only $K = \mathrm{full}$ and returns one volume.

Exact measurement normalization and SSIM proxy. The normalization belongs to the evidence contract and is therefore fixed before comparing candidates. Define $N(X_i) = X_i / \max(\max_{x,y} X_i, 10^{-6})$ separately for each sample and view. MSE compares the projected and target views without this extra normalization; NMSE is

$$\mathrm{NMSE}(\widehat{Y}, Y) = \mathrm{mean}_{i,x,y} \left[N(\widehat{Y}_i) - N(Y_i) \right]^2, \tag{A.8}$$

not $\|\widehat{Y} - Y\|_2^2 / \|Y\|_2^2$. Measurement SSIM uses the same independently normalized views. For $X = N(\widehat{Y}_i)$ and $T = N(Y_i)$, its local score is

$$\frac{(2\mu_X\mu_T + c_1)(2\sigma_{XT} + c_2)}{(\mu_X^2 + \mu_T^2 + c_1)(\sigma_X^2 + \sigma_T^2 + c_2)}, \qquad c_1 = 0.01^2, \quad c_2 = 0.03^2. \tag{A.9}$$

Means and second moments use a uniform 7×7 average-pooling window with stride one and zero padding three, not a Gaussian-window implementation. Variance and covariance are pooled second moments minus products of pooled means. Nonfinite local scores map NaN/positive infinity/negative infinity to 0/1/0, respectively; scores are clipped to $[-1, 1]$ and averaged over pixels, views, and batch. Independent volume-wide nonnegative max normalization is a separate object-evaluation transform: $Q_{\mathrm{obj}} = q(T_{\mathrm{eval}}(V^{(K)}), T_{\mathrm{eval}}(V^\star))$, applied to *both* prediction and reference, not to the signed volume before reprojection.

At every epoch, all three checkpoint rules inspect the same ten fixed selection records. The loss rule minimizes the complete objective; the full rule maximizes full-input reprojection SSIM; and the active rule maximizes supplied-view SSIM under two frozen delete-two masks. After all checkpoint files are fixed, we evaluate the saved states using the final leave-one-out masks, random deletions, corruptions, and known-volume phantoms.

Table A29: Matched changes in clean-corpus prediction and structural agreement, expressed as Calibra-LF minus signed HYPER. Clean scores use final corpus groups and object scores use 20 known volumes. Positive differences favor Calibra-LF; "Improved" counts seed–selection-rule file pairs, including aliases. Differences and counts remain unranked; complementary metrics appear in Appendix A.28.

Endpoint	Mean Δ	Improved pairs
Clean active-view SSIM	+0.00634	6/9
Clean held-out SSIM	+0.00266	7/9
Known-volume Pearson	+0.02788	8/9
Known-volume top-k support overlap	+0.02061	8/9

A.26 Evaluation Records and Aggregation

We aggregate evaluation records according to the candidate hierarchy and evidence target. The clean measurement evaluation contains $37 \times 20 \times 9 = 6{,}660$ checkpoint–group–mask records: 36 learned checkpoints plus RL, 20 final groups, and nine leave-one-out masks. The known-volume evaluation contains another 6,660 records. Corruptions are evaluated for 12 active-selected learned checkpoints on nine leave-one-out and six frozen random-deletion masks, yielding $12 \times 20 \times 15 \times 7 = 25{,}200$ records including clean.

Four aggregation and interpretation choices connect each comparison to its evidence. Measurements aggregate by group and retain dependence among frames and masks; checkpoint tables expose all three selection rules while rankings use deduplicated run–epoch states; best/second colors stay within one metric column and population; and each conclusion follows the target directly tested. Supplied-view fit assesses camera consistency, excluded-view fit assesses specified measurement prediction, and known volumes assess normalized synthetic object fidelity.

A.27 Extended Selection Diagnostics and Controls

Appendix A.27 gives the local-ranking diagnosis an explicit pairwise form. SCOPE-FLFM treats model selection as a prediction problem. For candidates i and j, a selector using contract A makes

a discordant choice under contract B when

$$(Q_A(i) - Q_A(j))\,(Q_B(i) - Q_B(j)) < 0. \tag{A.10}$$

The pairwise wrong-selection rate is the fraction of all non-tied pairs satisfying Equation A.10. Spearman correlation weights all rank displacement, including easy family separations, whereas top-k overlap focuses on candidates likely to receive further compute. Selection regret measures the B-score sacrificed by deploying the A winner rather than the B winner.

Checkpoint multiplicity requires a separate accounting rule because the score is attached to a state, while files record selection provenance. Loss, full-SSIM, and active-SSIM selection can save the same epoch. Counting those files as distinct models would replicate one weight state. We expose all 36 checkpoint files for provenance, then collapse identical run–epoch pairs for primary selection analysis. Among 23 unique learned states, top-five overlap remains 2/5 and the final observed-view winner incurs 0.0253 held-out SSIM endpoint-oracle regret. This diagnostic uses final readouts; prospective deployment uses selection-set scores.

The controls probe the candidate differences that shape CCS and the endpoint profile. Signed HYPER tests the native signed output path; positive HYPER learns positive outputs during training. Richardson–Lucy provides an iterative reference across the 1–128 iteration trajectory in Table A13, including the initial four-iteration configuration. The selection-frozen comparison uses RL-128. XLFMNet supplies architectural diversity through a local FLFM adaptation, separate from official XLFM benchmarks. The identity-initialized Calibra-LF branch improves clean-corpus held-out SSIM and object Pearson on average, while known-volume held-out SSIM, volume SSIM, NRMSE, and corruption responses favor signed HYPER on average. This comparison measures the joint effect of Bounded Depth Gain and Raw-Conditioned Residual within the HYPER-derived model.

Table A30: Minimum reporting contract for masked FLFM reconstruction. Each row fixes an access rule, aggregation unit, question answered, and conclusion supported by that evidence level. Repeated masks or correlated patches do not create additional independent samples.

Evidence	Required exclusion	Aggregation unit	Question answered	Permitted conclusion
Observed measurement	None; scored views may enter input	Group, averaged over masks	How well does the reconstruction explain the supplied image?	Supplied-measurement consistency
Held-out measurement	Scored view excluded from input and selection target	Group, averaged over masks	How well does it predict the withheld view?	Specified measurement extrapolation
Known synthetic object	Target excluded from optimization and selection	Phantom, averaged over masks	How closely does it recover the known synthetic structure?	Declared simulator/object fidelity
Experimental object	Independently acquired registered reference	Specimen or acquisition	How closely does it match an independently acquired reference?	Measured-domain fidelity

A.28 Complementary Measurement and Object Metrics

We complete the matched comparison with known-volume prediction, normalized voxelwise error, and corruption responses. Together with Table A29, these measurements show how the same model change affects different imaging properties. The source-separated tables preserve the detailed Calibra-LF object and corruption results summarized in the main paper.

Table A31: Matched changes in known-volume prediction, normalized error, and corruption response, expressed as Calibra-LF minus signed HYPER. Known-volume scores use 20 phantoms and corruption scores use final corpus groups. The corruption suite, including its clean condition, averages nine leave-one-view-out masks and six fixed random-deletion masks. Positive SSIM differences and negative NRMSE differences favor Calibra-LF. "Improved" counts seed–selection-rule file pairs for object endpoints and seed–condition pairs for corruption. Differences and counts remain unranked.

Endpoint	Mean Δ	Improved pairs
Known-volume held-out SSIM	-0.04297	3/9
Known-volume SSIM	-0.00309	3/9
Known-volume normalized RMSE	+0.01817	1/9
Corruption held-out SSIM	-0.00190	5/21

Table A32: Corruption-condition summary reproduced from Table A9 for interpreting the matched corruption differences in Table A31. The final corpus groups are evaluated over 12 active-selected learned checkpoints; the corruption suite, including its clean condition, averages nine leave-one-view-out masks and six fixed random-deletion masks. SSIM is the mean excluded-view score; delta is condition minus clean and rank ρ compares clean and corrupted checkpoint ordering. Conditions, differences, and rank correlations are unranked stress diagnostics.

Condition	Held-out SSIM $\uparrow$	Δ vs clean	Rank ρ
attenuation	0.3819	-0.0253	0.944
clean	0.4072	+0.0000	1.000
lenslet shift	0.3918	-0.0154	0.965
poisson gaussian noise	0.3996	-0.0076	0.986
psf mismatch	0.4073	+0.0001	1.000
registration shift	0.4000	-0.0072	0.972
saturation	0.3844	-0.0228	0.979