CRYOFIG: FEED-FORWARD INFERENCE OF 3D GAUSSIANS FOR AB INITIO CRYO-EM RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ab initio cryo-electron microscopy (cryo-EM) reconstruction recovers molecular density from noisy particle images with unknown poses. However, extending learned pose inference to direct density prediction requires combining weak image evidence across views. We introduce CryoFIG (Feed-forward Inference of Gaussians), a shared model for rapid reconstruction that predicts particle geometry and anisotropic 3D Gaussian densities. Its geometry-guided decoder aggregates spatial image features across particles using viewing geometry and known contrast transfer function metadata. Training combines simulated pretraining, shared adaptation on synthetic and experimental particles, and reprojection supervision from disjoint target images. After shared-model training, one checkpoint reconstructs density maps from 32 support particles in approximately seven seconds without an initial volume or inference-time optimization. Across two synthetic benchmarks, CryoFIG achieves finer resolution than the evaluated baselines, as measured by map-to-map Fourier shell correlation. It also achieves competitive reconstruction quality on two experimental benchmarks. These results support learned Gaussian density prediction as a route to rapid cryo-EM reconstruction. Code and pretrained model weights will be released.

1 INTRODUCTION

Ab initio reconstruction in single-particle cryo-electron microscopy (cryo-EM) couples molecular density recovery with particle pose estimation (Singer & Sigworth, 2020). Particle orientations and in-plane shifts are unknown, and each observed image is modulated by the microscope's contrast transfer function (CTF). Weak image signals complicate pose estimation, while inaccurate poses obscure the structural information shared across particles.

Cryo-EM reconstruction increasingly uses learned inference to estimate particle geometry. RELION and cryoSPARC estimate particle poses and molecular density iteratively (Scheres, 2012; Punjani et al., 2017). Neural approaches introduce flexible density representations and, in some cases, amortized pose inference, but still fit a density model to each reconstruction dataset (Zhong et al., 2021a; Levy et al., 2022a;b; 2025). CryoFastAR further predicts particle geometry and reconstructs a volume through Fourier backprojection (Zhang et al., 2025). This progress motivates extending shared inference from particle geometry to the density representation itself.

Anisotropic 3D Gaussians offer a continuous representation suited to direct density prediction (Kerbl et al., 2023). CryoSplat and GEM use this representation for molecular reconstruction through dataset-specific optimization (Chen & Ling, 2025; Qu et al., 2025). In computer vision, pixelSplat and NoPoSplat predict Gaussian scene representations directly from sparse input views (Charatan et al., 2024; Ye et al., 2025).

Transferring direct Gaussian prediction to cryo-EM requires aggregating volumetric evidence across particle images. Each image integrates density along viewing rays and is modulated by the CTF (Levy et al., 2022a). A local image feature therefore cannot be assigned directly to a visible surface point. Interpreting that feature in three dimensions requires complementary evidence from other projections. We ask whether a shared model can directly predict molecular density from a small unposed particle set.

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.

We introduce CryoFIG (Feed-forward Inference of Gaussians), a shared framework that predicts particle poses and an anisotropic Gaussian density from unposed particle images. A relative pose frontend establishes a shared coordinate frame, while an independent encoder extracts spatial image features. Its geometry-guided Gaussian query decoder samples spatial features at projected 3D locations and aggregates them using estimated poses and known CTF metadata. Simulated pretraining and shared adaptation prepare one model for reconstruction without an initial volume or optimization during inference.

Our contributions are threefold:

- **Direct Gaussian density prediction.** We jointly predict particle poses and anisotropic 3D Gaussian density parameters from unposed cryo-EM images using a shared model.
- **Geometry-guided cross-particle decoding.** We aggregate image evidence at projected query locations using viewing geometry and CTF metadata.
- **Shared-model reconstruction with low inference latency.** One checkpoint produces density maps from 32 support particles in approximately seven seconds. It achieves finer resolution than evaluated baselines on synthetic data and competitive quality on experimental data.

2 Related Work

Neural density models and amortized inference. Neural methods improve density modeling and pose inference while retaining dataset-specific density fitting. CryoDRGN and CryoFormer model structural heterogeneity through learned density representations (Zhong et al., 2021a; Liu et al., 2023). CryoDRGN2 and CryoDRGN-AI support ab initio reconstruction through pose search and refinement (Zhong et al., 2021b; Levy et al., 2025). CryoPoseNet, CryoAI and cryoFIRE amortize orientation or pose estimation with image encoders (Nashed et al., 2021; Levy et al., 2022a;b). CryoFIG predicts density from cross-particle evidence with decoder weights fixed during inference.

Pretrained representations and geometric inference. Pretrained representations support both particle analysis and geometric reconstruction. DUSt3R predicts dense point maps in a common coordinate frame from uncalibrated image pairs (Wang et al., 2024). Within cryo-EM, Cryo-IEF learns particle representations for classification, pose-based clustering and image quality assessment (Yan et al., 2026). CryoFastAR predicts particle geometry and reconstructs a molecular volume through Fourier backprojection (Zhang et al., 2025). CryoFIG extends geometric inference to Gaussian density prediction through geometry-guided feature sampling and aggregation.

Gaussian representations of molecular density. Gaussian representations model molecular density and structural variability with continuous primitives. The mixed-dimensional Gaussian mixture model of Chen & Ludtke (2021) encodes conformational and compositional changes in a learned latent space. DynaMight learns deformations of a Gaussian consensus model (Schwab et al., 2024). CryoSplat and GEM adapt Gaussian splatting to cryo-EM through optimization against particle observations (Chen & Ling, 2025; Qu et al., 2025). These methods fit individual specimens, while CryoFIG predicts Gaussian parameters through a shared decoder without optimization during inference.

Feed-forward Gaussian scene reconstruction. Building on 3D Gaussian Splatting (Kerbl et al., 2023), pixelSplat and MVSplat predict Gaussian scene representations from calibrated input views (Charatan et al., 2024; Chen et al., 2024). NoPoSplat and SPFSplat accept unposed images at inference and predict Gaussians in a common reference frame (Ye et al., 2025; Huang & Mikolajczyk, 2025). Their radiance and visibility models differ from cryo-EM's noisy, CTF-modulated density integrals (Levy et al., 2022a; Chen & Ling, 2025). CryoFIG addresses these measurements through geometry-guided aggregation, known CTF conditioning and physical reprojection supervision.

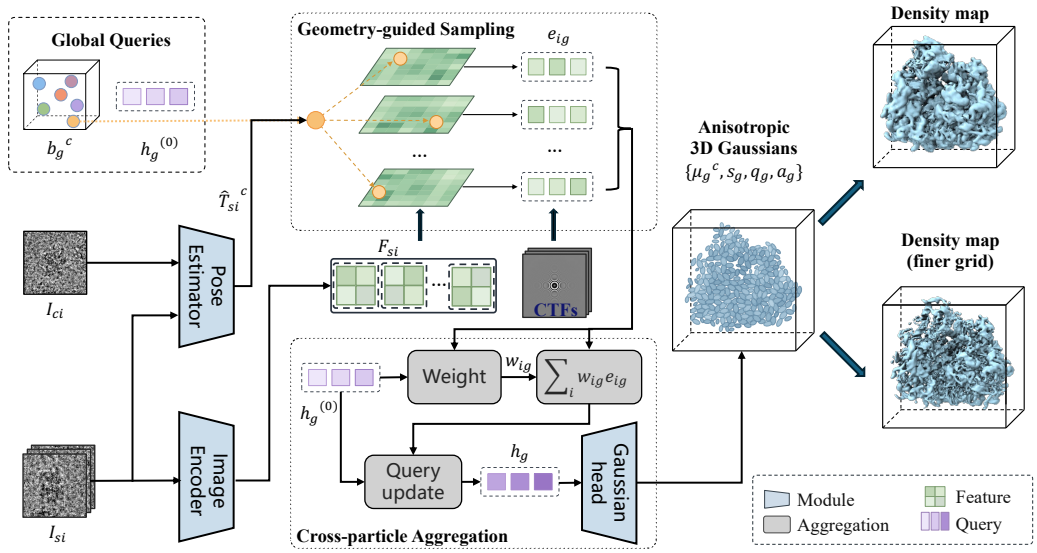


Figure 1: Feed-forward reconstruction with CryoFIG. Support geometry $\widehat{T}^{c}_{\mathrm{si}}$ guides feature sampling from F_{si} for learned 3D queries. Aggregated image, geometry and CTF evidence yields anisotropic Gaussians, whose density we evaluate at voxel centers.

3 METHOD

CryoFIG predicts anisotropic Gaussian density from unposed cryo-EM images through geometry-guided cross-particle aggregation, as shown in Figure 1. A pose frontend and independent image encoder provide geometry and spatial features, which learned 3D queries combine with known CTF metadata.

Notation. The labels s and q identify support and target images, while c marks the reference image and frame. The index $\tau \in \{\mathrm{t}, \mathrm{i}\}$ distinguishes training and inference, with support index set $\mathcal{S}_\tau$. We write support geometry as $T^c_{s\tau} = \{(R^c_i, \mathbf{t}_i)\}_{i\in\mathcal{S}_\tau}$, where R^c_i is a particle rotation and $\mathbf{t}_i$ its detector-plane shift. Indices i and g denote particles and Gaussian queries, and $\mathcal{Q}_{\mathrm{t}}$ is the training target index set. Hats mark predictions and tildes mark gate-selected geometry.

3.1 ANISOTROPIC GAUSSIAN DENSITY

We represent molecular density as a weighted sum of normalized anisotropic Gaussians (Chen & Ludtke, 2021; Chen & Ling, 2025). In reference frame c, the representation contains G components, $\mathcal{G}^c = \{(\boldsymbol{\mu}^c_g, \mathbf{s}_g, \mathbf{q}_g, a_g)\}_{g=1}^{G}$. Each component has a center $\boldsymbol{\mu}^c_g$, positive principal standard deviations $\mathbf{s}_g$, a unit quaternion $\mathbf{q}_g$ and a nonnegative integrated weight a_g. The scales and rotation determine the covariance (Kerbl et al., 2023), giving

$$\rho_{\mathcal{G}^c}(\mathbf{x}) = \sum_{g=1}^{G} a_g \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}^c_g, \Sigma^c_g),$$
$$\Sigma^c_g = U(\mathbf{q}_g)\,\mathrm{diag}(\mathbf{s}^2_g)U(\mathbf{q}_g)^\top, \tag{1}$$

where $\mathcal{N}$ is a normalized three-dimensional Gaussian and U converts a quaternion to a rotation matrix. This continuous representation supports differentiable projection and evaluation on a chosen voxel grid.

3.2 GEOMETRY-GUIDED GAUSSIAN DECODING

Image and pose encoding. The pose frontend uses an image encoder initialized with Cryo-IEF weights (Yan et al., 2026) to estimate rotations and detector-plane shifts relative to a fixed reference.

An independent spatial encoder produces a feature map F_i for each support particle. We denote the collection of support feature maps by $F_{s\tau}$ and detail pose estimation in Appendix A.1.

Separate structure embeddings condition the two branches during training, with the decoder embedding modulating initial query features as detailed in Appendix A.2.

Geometry-guided sampling. Each query consists of a learned base position $\mathbf{b}_g^c$ and an initial feature $\mathbf{h}_g^{(0)}$. We project query positions into support images (Wang et al., 2022) and combine sampled spatial features with geometry and CTF encodings,

$$\begin{aligned} \mathbf{u}_{ig} &= \frac{\Pi R_i^c \mathbf{b}_g^c}{e} - \frac{2\mathbf{t}_i}{H_i}, \\ \mathbf{e}_{ig} &= \mathrm{LN}\left[W_f\,\mathrm{Sample}(F_i, \mathbf{u}_{ig}) + f_{\mathrm{geom}}(\mathbf{g}_{ig}) + f_{\mathrm{CTF}}(\mathbf{c}_i)\right]. \end{aligned} \tag{2}$$

Here, Π selects image-plane coordinates, e is the spatial half-extent, and H_i is the native image size for pixel shifts $\mathbf{t}_i$. Sample performs bilinear interpolation (Jaderberg et al., 2015), and W_f projects sampled features into the evidence space. $\mathbf{g}_{ig}$ encodes viewing geometry and sampling validity, while $\mathbf{c}_i$ contains known CTF metadata. The functions f_{geom} and f_{CTF} embed these inputs, and LN denotes layer normalization (Ba et al., 2016).

Cross-particle aggregation. The decoder refines each query by repeatedly reweighting the same sampled evidence across support particles. At layer ℓ, normalized query and evidence features $\bar{\mathbf{h}}_g^{(\ell)}$ and $\bar{\mathbf{e}}_{ig}^{(\ell)}$ determine the aggregation weights. The update is

$$\begin{aligned} w_{ig}^{(\ell)} &= \mathrm{softmax}_i\left[f_{\mathrm{score}}^{(\ell)}\left(\bar{\mathbf{h}}_g^{(\ell)} + \bar{\mathbf{e}}_{ig}^{(\ell)}\right)\right], \\ \mathbf{h}_g^{(\ell+1)} &= \mathbf{h}_g^{(\ell)} + f_{\mathrm{update}}^{(\ell)}\left(\left[\bar{\mathbf{h}}_g^{(\ell)}; \sum_{i\in\mathcal{S}_\tau} w_{ig}^{(\ell)}\mathbf{e}_{ig}\right]\right). \end{aligned} \tag{3}$$

Softmax normalizes over support particles for each query, and square brackets with a semicolon denote concatenation.

Gaussian parameter prediction. A shared head maps final query features to center corrections around $\mathbf{b}_g^c$, positive scales, normalized quaternions and nonnegative integrated weights.

3.3 Training and Inference

Each training episode uses disjoint support and target sets from the same structure. The decoder predicts density from support images; target images provide reprojection supervision, as shown in Figure 2.

Accuracy-gated geometry. For structure d at training step k, an accuracy gate selects predicted geometry with probability $p_d(k)$,

$$\begin{aligned} z_k &\sim \mathrm{Bernoulli}(p_d(k)), \\ \widetilde{T}_{\mathrm{st}}^c &= \begin{cases} T_{\mathrm{st}}^{\mathrm{gt},c}, & z_k = 0, \\ \mathrm{sg}(\widehat{T}_{\mathrm{st}}^c), & z_k = 1, \end{cases} \end{aligned} \tag{4}$$

where sg stops gradients and the same selection applies to the entire support set.

The probability $p_d(k)$ starts at zero and follows validation performance. We increase it after consecutive validation checks satisfy rotation, pairwise-rotation and shift criteria while preserving reconstruction quality under labeled geometry. It decreases on regression and otherwise remains unchanged.

Reconstruction-condition dropout. We prepare the decoder for inference without structure conditioning by dropping its feature-wise linear modulation (FiLM) (Perez et al., 2018) with probability $p_{\mathrm{drop}} = 0.5$. All queries and decoder evaluations within an episode share one binary mask.

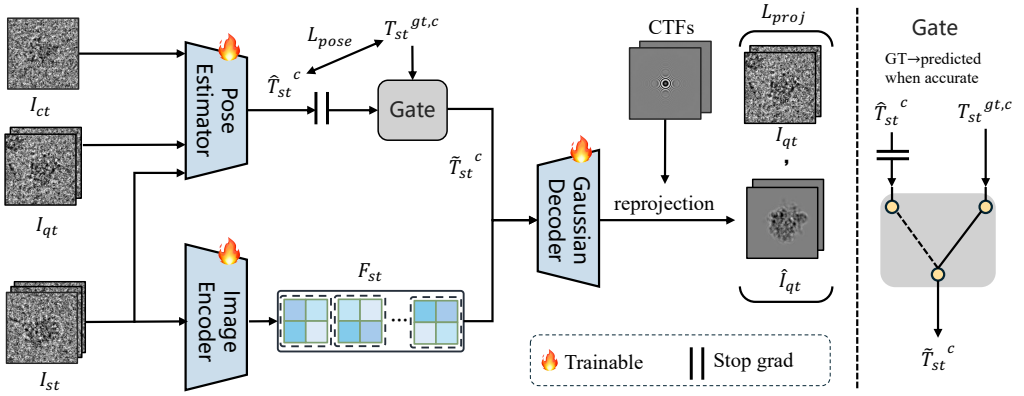


Figure 2: Reconstruction training with CryoFIG. An accuracy gate selects labeled or stopped-gradient support geometry. Labeled target poses and CTFs supervise Gaussian reprojection through $\mathcal{L}_{\text{proj}}$. Separate objectives train the pose frontend. Flames mark trainable modules and double bars mark stopped gradients.

Reprojection supervision. Reprojection supervision updates the reconstruction branch while treating support geometry as fixed. A differentiable renderer applies orthographic Gaussian rasterization (Chen & Ling, 2025), CTF modulation and detector-plane shifting under the cryo-EM image formation model (Levy et al., 2022a). Using labeled target poses, we render $\widehat{I}_{i,\text{qt}}$ and minimize

$$\mathcal{L}_{\text{proj}} = \mathbb{E}\left[\frac{1}{|\mathcal{Q}_{\text{t}}|}\sum_{i\in\mathcal{Q}_{\text{t}}} \ell(\widehat{I}_{i,\text{qt}}, I_{i,\text{qt}})\right], \tag{5}$$

where ℓ is the normalized Fourier-domain image discrepancy defined in Appendix A.3.

Pose supervision. We jointly adapt the pretrained pose frontend with supervised geometry loss $\mathcal{L}_{\text{pose}}$ and auxiliary physical loss $\mathcal{L}_{\text{pose-proj}}$. For the physical loss, the same decoder uses labeled support geometry to predict a density, which we detach before reprojection with predicted target poses. Reprojection gradients update the pose frontend through the target poses.

With loss weights λ_{pose} and λ_{phys}, the objective is

$$\mathcal{L} = \mathcal{L}_{\text{proj}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{phys}}\mathcal{L}_{\text{pose-proj}}. \tag{6}$$

Appendix A.4 details the pose objectives and their gradient routing.

Inference. CryoFIG predicts density in a single forward pass with fixed network parameters, requiring neither an initial volume nor optimization during inference. For all main evaluations, we disable structure conditioning in both the pose frontend and Gaussian decoder. We evaluate the predicted density at voxel centers to produce the output map.

4 Experiments

We evaluate one shared CryoFIG checkpoint on two synthetic and two experimental cryo-EM datasets. The experiments assess reconstruction quality, inference efficiency, and the effects of reconstruction-condition dropout and support-set size.

4.1 Experimental Setup

Pretraining. We pretrain the complete CryoFIG model on 2.5 million simulated projections from 1,000 Protein Data Bank structures (Berman et al., 2000). Each structure contributes 2,500 projections, and we sample fresh Gaussian noise online during synthetic training. Pretraining takes approximately 20 days on four NVIDIA GeForce RTX 4090 GPUs, totaling roughly 1,900 GPU-hours.

Shared-model adaptation. We then adapt one shared model on the training partitions of all four benchmark datasets. We follow the procedure in Section 3.3, weighting each structure equally in the training objective. This stage includes experimental particles to address the simulation-to-experiment domain gap, following the real-data adaptation strategy of CryoFastAR (Zhang et al., 2025). All main benchmark results use the resulting shared checkpoint for reconstruction.

Datasets. The synthetic benchmarks, Spike and HSP, follow the simulation protocol of CryoAI (Levy et al., 2022a). Each contains 10,000 clean training projections, which we corrupt with online noise at a signal-to-noise ratio (SNR) of 0.1. Each test set contains 50,000 independently generated projections at the same SNR. We compare each synthetic reconstruction with the corresponding ground-truth density map.

We evaluate experimental reconstruction on two datasets, EMPIAR-10028 (Wong et al., 2014) and EMPIAR-10180 (Plaschka et al., 2017). We downsample the images to 128×128 and use 10,000 particles per dataset for shared adaptation. We draw evaluation support sets from held-out particles and compare the reconstructions with reference density maps.

Baselines. We compare CryoFIG with CryoSPARC (Punjani et al., 2017) and CryoDRGN-AI (Levy et al., 2025) under the configurations described below. We evaluate CryoSPARC before and after refinement, initializing refinement with its own ab initio reconstruction. On experimental data, the baselines use the full particle collections, while CryoFIG reconstructs from 32 held-out support particles.

Evaluation metrics. We evaluate reconstruction quality using masked map-to-map Fourier shell correlation (FSC) between aligned density maps (Penczek, 2010). We compute FSC with EMAN2 and report resolution at FSC$= 0.5$ for all datasets (Tang et al., 2007). For experimental data, we additionally report the resolution at FSC$= 0.143$. We also assess synthetic reconstructions using ground-truth-based local resolution and visualize density maps with UCSF ChimeraX (Pettersen et al., 2021).

We report a resolution upper bound when a curve remains above the threshold at the highest evaluated frequency. We denote this bound by $\leq$ and report the resolution corresponding to that frequency. Appendix A.7 defines FSC and explains the interpretation of threshold crossings and Nyquist-limited bounds.

Reconstruction settings. CryoFIG predicts 32,768 anisotropic Gaussians per reconstruction and evaluates density maps directly from these Gaussians. We compare all methods on a 128^3 output grid and disable structure conditioning in both the pose frontend and Gaussian decoder for all main evaluations. Appendix A.6 provides training configurations, image preprocessing, alignment, masking and software details.

4.2 Reconstruction Quality Across Datasets

Synthetic data. CryoFIG achieves the finest FSC$= 0.5$ resolution among the evaluated methods on both synthetic benchmarks, as summarized in Table 1. Relative to refined CryoSPARC, the strongest baseline on these datasets, its resolution values in Å are approximately 22% lower on Spike and 37% lower on HSP. Figures 3 and 4 show the reconstructed densities, FSC curves and local resolution distributions. CryoFIG's local resolution estimates shift toward finer values, consistent with the greater structural detail visible in its maps.

Experimental data. On EMPIAR-10028, CryoFIG maintains FSC above 0.5 up to the Nyquist frequency. We therefore report a resolution bound of ≤ 7.54 Å, while both refined CryoSPARC and CryoDRGN-AI cross the threshold at lower frequencies. Figure 5 compares the reconstructed densities and their FSC curves against the experimental reference maps.

On EMPIAR-10180, CryoDRGN-AI achieves finer FSC$= 0.5$ resolution than CryoFIG, with values of 8.77 Å and 9.37 Å, respectively. Both methods recover the overall molecular architecture, while CryoSPARC produces an incomplete reconstruction even after refinement under the evaluated configuration. The CryoFIG and CryoDRGN-AI curves both remain above FSC$= 0.143$ throughout the evaluated frequency range.

Table 1: Masked map-to-map FSC resolution at the 0.5 threshold in Å. Lower values indicate finer resolution, and bold entries identify the best results. CryoFIG uses one shared checkpoint and 32 support particles per reconstruction. † FSC remains above 0.5 at the Nyquist frequency, so the reported value is an upper bound. Failed denotes an incomplete reconstruction under the evaluated configuration.

Method	Spike	HSP	EMPIAR-10028	EMPIAR-10180
CryoSPARC (Punjani et al., 2017)	10.91	14.77	11.77	Failed
CryoSPARC (refined) (Punjani et al., 2017)	7.79	6.19	7.91	Failed
CryoDRGN-AI (Levy et al., 2025)	8.02	6.62	8.04	**8.77**
CryoFIG (Ours)	**6.06**	**3.92**	$\leq$ **7.54**†	9.37

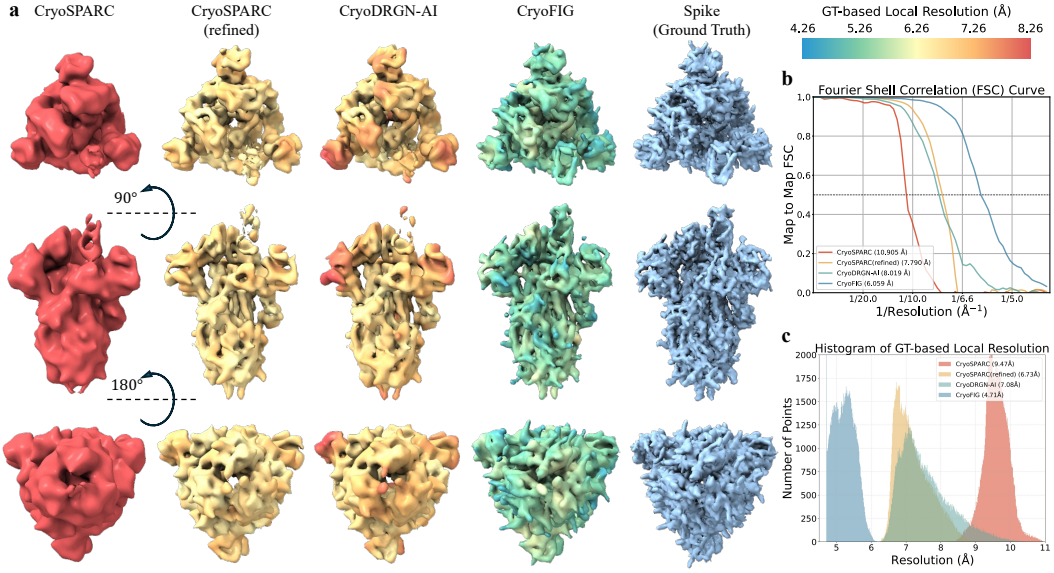


Figure 3: **Reconstruction quality on Spike. (a)** Reconstructed maps and ground-truth density in three orientations. Reconstructions are colored by ground-truth-based local resolution, with cooler colors indicating finer resolution. **(b)** Map-to-map FSC against the ground truth, with the dashed line marking 0.5. **(c)** Corresponding local resolution distributions for the reconstructed density maps.

4.3 Feed-Forward Reconstruction Efficiency

CryoFIG produces a density map from 32 support particles in 7.03–7.12 seconds across the four datasets, with fixed model parameters. We time pose estimation, image encoding, Gaussian decoding and density evaluation on a 128^3 grid. This latency excludes both offline pretraining and shared-model adaptation of the checkpoint.

Baseline runtimes span minutes for CryoSPARC and hours for CryoDRGN-AI under their respective reconstruction settings, as summarized in Table 2. These runtimes include pose and density estimation, with refinement also included for refined CryoSPARC. On experimental data, the baselines process full particle collections, while CryoFIG uses 32 held-out support particles per reconstruction.

4.4 Ablation Studies

We evaluate reconstruction-condition dropout and support-set size with reconstruction FiLM disabled at inference, as summarized in Table 3. The dropout comparison uses 32 support particles, while all support-size variants use the same checkpoint trained with $p_{\text{drop}} = 0.5$.

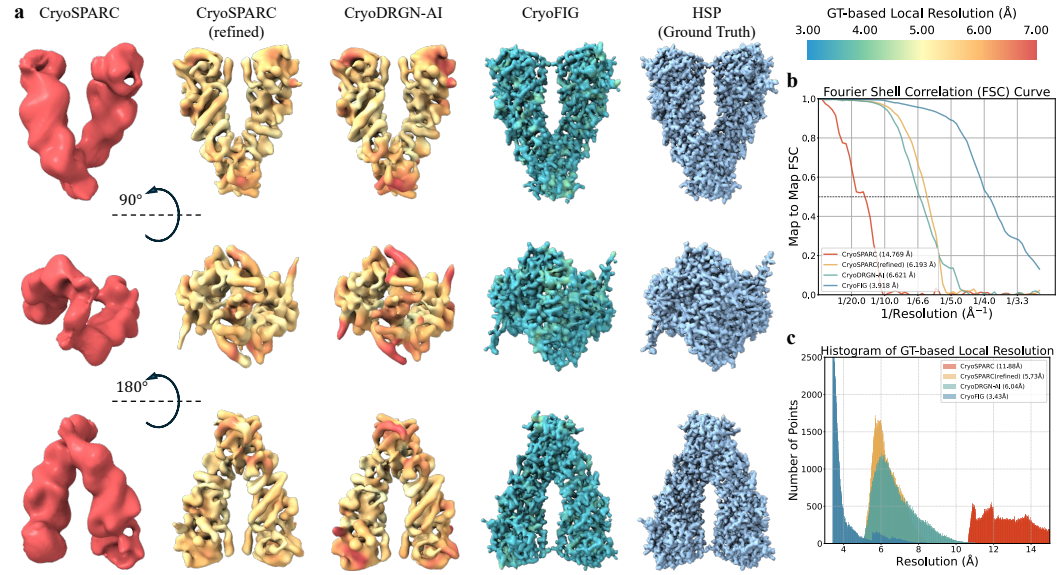


Figure 4: **Reconstruction quality on HSP. (a)** Reconstructed maps and ground-truth density in three orientations. Reconstructions are colored by ground-truth-based local resolution, with cooler colors indicating finer resolution. **(b)** Map-to-map FSC against the ground truth, with the dashed line marking 0.5. **(c)** Corresponding local resolution distributions for the reconstructed density maps.

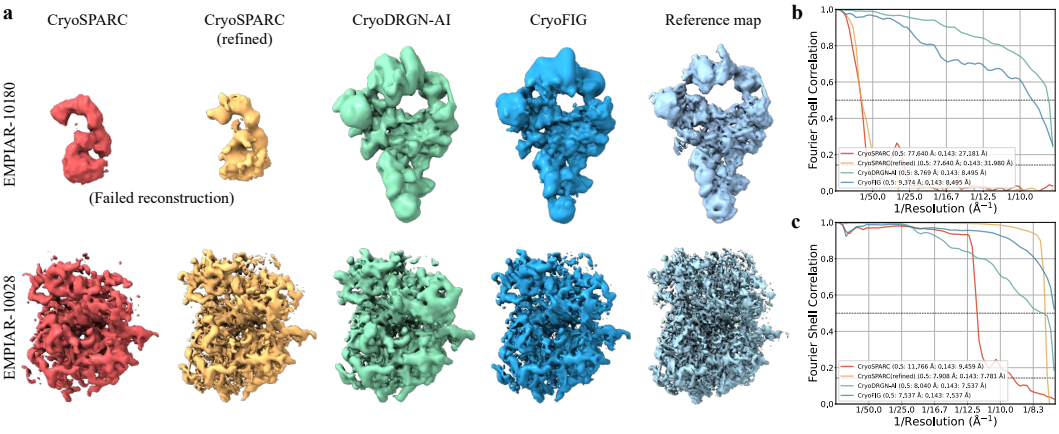


Figure 5: **Reconstruction quality on experimental data. (a)** Reconstructed densities and reference maps for EMPIAR-10180 at the top and EMPIAR-10028 at the bottom. Failure labels identify incomplete reconstructions under the evaluated configurations. **(b,c)** Map-to-map FSC against the corresponding references for EMPIAR-10180 and EMPIAR-10028, respectively. Dashed lines mark the two evaluated thresholds, FSC$=0.5$ and 0.143.

Reconstruction-condition dropout. Training with reconstruction-condition dropout improves FSC$=0.5$ resolution on all four datasets when reconstruction FiLM is disabled at inference. At 32 support particles, the largest absolute change occurs on EMPIAR-10180, from 135.870 Å with $p_{\text{drop}} = 0$ to 9.374 Å with $p_{\text{drop}} = 0.5$. On EMPIAR-10028, the FSC$=0.5$ result changes from 20.101 Å to a Nyquist-limited bound of ≤ 7.538 Å.

Support-set size. Increasing the support set from 8 to 32 particles improves FSC$=0.5$ resolution on Spike, HSP and EMPIAR-10180. Increasing the support set from 32 to 64 particles produces a smaller additional improvement only on Spike. Spike improves from 6.059 Å to 5.927 Å, while HSP and EMPIAR-10180 retain the same reported values.

Table 2: CryoFIG inference latency and baseline reconstruction runtimes in HH:MM:SS, with fractional seconds for CryoFIG. All outputs use a 128^3 grid, and bold entries identify the shortest times. CryoFIG timings include processing 32 support particles and evaluating the predicted density after shared-model training. Baseline timings include reconstruction optimization, with refinement also included for refined CryoSPARC. † The run produces an incomplete reconstruction under the evaluated configuration.

Method	Spike	HSP	EMPIAR-10028	EMPIAR-10180
CryoSPARC (Punjani et al., 2017)	00:04:25	00:03:05	00:09:33	00:11:51†
CryoSPARC (refined) (Punjani et al., 2017)	00:05:48	00:04:19	00:11:00	00:14:21†
CryoDRGN-AI (Levy et al., 2025)	04:39:18	04:37:58	09:55:39	30:50:30
CryoFIG (Ours)	**00:00:07.048**	**00:00:07.029**	**00:00:07.067**	**00:00:07.124**

Table 3: Ablation of reconstruction-condition dropout and support-set size using masked map-to-map FSC resolution in Å. Lower values indicate finer resolution, with FSC$= 0.5$ reported for synthetic data and FSC$= 0.5$ / 0.143 for experimental data. The symbol $\leq$ denotes a bound limited by the Nyquist frequency. Reconstruction FiLM remains disabled at inference for every setting in this table. Dropout is compared at 32 support particles, and all support-size variants use the same $p_{\mathrm{drop}} = 0.5$ checkpoint.

p_{drop}	Support	Spike	HSP	EMPIAR-10028	EMPIAR-10180
0	32	17.039	38.402	20.101 / ≤ 7.538	135.870 / 67.981
0.5	8	11.360	12.000	≤ 7.538 / ≤ 7.538	9.709 / ≤ 8.495
0.5	16	6.059	4.923	≤ 7.538 / ≤ 7.538	9.538 / ≤ 8.495
0.5	32	6.059	3.918	≤ 7.538 / ≤ 7.538	9.374 / ≤ 8.495
0.5	64	5.927	3.918	≤ 7.538 / ≤ 7.538	9.374 / ≤ 8.495

On EMPIAR-10028, all tested support sizes reach the Nyquist-limited bound at both FSC thresholds. The smallest support set of eight particles already yields a reported resolution of ≤ 7.538 Å at FSC$= 0.5$ and 0.143, and increasing the set to 64 particles leaves both bounds unchanged. These values indicate that the FSC does not cross either threshold within the evaluated frequency range. The reported resolutions therefore cannot distinguish reconstruction quality across support sizes on this dataset.

5 Conclusion

We present CryoFIG, a shared model that predicts particle poses and continuous molecular density from unposed cryo-EM images using anisotropic 3D Gaussians. Its geometry-guided decoder combines spatial image evidence across particles with viewing geometry and known CTF metadata to predict Gaussian density parameters. By integrating complementary projections in a common coordinate frame, the model connects local image observations to an explicit three-dimensional density. After simulated pretraining and shared adaptation, one checkpoint reconstructs density maps from 32 support particles in approximately seven seconds without optimization during inference. CryoFIG achieves finer map-to-map FSC resolution than the evaluated baselines on both synthetic benchmarks and competitive reconstruction quality on experimental data. These results support extending learned inference from particle geometry to continuous density prediction, providing a route to rapid reconstruction from small particle sets.

Limitations and future work. The current framework uses experimental adaptation and predicts a single density for each support set. Future work will investigate transfer with fewer experimental training particles to reduce the adaptation required across imaging conditions. We will also explore efficient aggregation over larger support sets to recover complementary structural information while controlling memory use and inference latency. Finally, we will study conformation-dependent Gaussian representations to model structural variability and extend the framework to heterogeneous reconstruction. A central goal is to separate conformational variation from particle pose while retaining shared inference across the particle set.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding. We have not used generative AI tools for research ideation, generating synthetic datasets, or proving mathematical claims. Additionally, we used generative AI tools to aid and polish writing, support literature retrieval and discovery, and draft sections of the paper. The AI-assisted code was manually reviewed and tested for correctness. The core method does not use LLMs as a methodological component. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available cryo-EM datasets and synthetic projections for computational structural reconstruction. It does not involve human subjects, private personal data, or deployed decision-making systems. We are not aware of ethical concerns beyond standard considerations of scientific data use and reproducibility.

REPRODUCIBILITY STATEMENT

The reconstruction method is described in Section 3, with additional methodological details in Appendix A. The pretraining corpus, benchmark datasets, shared-model training protocol, reconstruction settings, and evaluation metrics are described in Section 4.1. Details of synthetic projection generation, data preprocessing, optimization, and evaluation procedures are provided in Appendix A.6.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467. IEEE, 2024.

Muyuan Chen and Steven J Ludtke. Deep learning-based mixed-dimensional gaussian mixture model for characterizing variability in cryo-em. *Nature methods*, 18(8):930–936, 2021.

Suyi Chen and Haibin Ling. Cryosplat: Gaussian splatting for cryo-em homogeneous reconstruction. *arXiv preprint arXiv:2508.04929*, 2025.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European conference on computer vision*, pp. 370–386. Springer, 2024.

Joseph H Davis, Yong Zi Tan, Bridget Carragher, Clinton S Potter, Dmitry Lyumkis, and James R Williamson. Modular assembly of the bacterial large ribosomal subunit. *Cell*, 167(6):1610–1622, 2016.

Ranran Huang and Krystian Mikolajczyk. No pose at all: Self-supervised pose-free 3d gaussian splatting from sparse views. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27947–27957, 2025. doi: 10.1109/ICCV51701.2025.02594.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images. In *European Conference on Computer Vision*, pp. 540–557. Springer, 2022a.

Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022b.

Axel Levy, Rishwanth Raghu, J Ryan Feathers, Michal Grzadkowski, Frédéric Poitevin, Jake D Johnston, Francesca Vallese, Oliver Biggs Clarke, Gordon Wetzstein, and Ellen D Zhong. Cryodrgn-ai: neural ab initio reconstruction of challenging cryo-em and cryo-et datasets. *Nature Methods*, 22(7):1486–1494, 2025.

Xinhang Liu, Yan Zeng, Yifan Qin, Hao Li, Jiakai Zhang, Lan Xu, and Jingyi Yu. Cryoformer: Continuous heterogeneous cryo-em reconstruction using transformer-based neural representations. *arXiv preprint arXiv:2303.16254*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Youssef SG Nashed, Frédéric Poitevin, Harshit Gupta, Geoffrey Woollard, Michael Kagan, Chun Hong Yoon, and Daniel Ratner. Cryoposenet: End-to-end simultaneous learning of single-particle orientation and 3d map reconstruction from cryo-electron microscopy data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4049–4059. IEEE, 2021.

Pawel A Penczek. Resolution measures in molecular electron microscopy. In *Methods in enzymology*, volume 482, pp. 73–100. Elsevier, 2010.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Elaine C Meng, Gregory S Couch, Tristan I Croll, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein science*, 30(1):70–82, 2021.

Clemens Plaschka, Pei-Chun Lin, and Kiyoshi Nagai. Structure of a pre-catalytic spliceosome. *Nature*, 546(7660):617–621, 2017.

Ali Punjani, John L. Rubinstein, David J. Fleet, and Marcus A. Brubaker. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, 14(3):290–296, 2017. doi: 10.1038/nmeth.4169.

Huaizhi Qu, Xiao Wang, Gengwei Zhang, Jie Peng, and Tianlong Chen. Gem: 3d gaussian splatting for efficient and accurate cryo-em reconstruction. *arXiv preprint arXiv:2509.25075*, 2025.

Heng Ru, Melissa G Chambers, Tian-Min Fu, Alexander B Tong, Maofu Liao, and Hao Wu. Molecular mechanism of v (d) j recombination from synaptic rag1-rag2 complex structures. *Cell*, 163 (5):1138–1152, 2015.

Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530, 2012.

Johannes Schwab, Dari Kimanius, Alister Burt, Tom Dendooven, and Sjors HW Scheres. Dynamight: estimating molecular motions with improved reconstruction from cryo-em images. *Nature methods*, 21(10):1855–1862, 2024.

Amit Singer and Fred J Sigworth. Computational methods for single-particle electron cryomicroscopy. *Annual review of biomedical data science*, 3(1):163–190, 2020.

Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on robot learning*, pp. 180–191. PmLR, 2022.

Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *elife*, 3:e03080, 2014.

Yang Yan, Shiqi Fan, Fajie Yuan, and Huaizong Shen. A comprehensive foundation model for cryo-em image processing. *Nature Methods*, 23(1):88–95, 2026.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *International Conference on Learning Representations*, volume 2025, pp. 54009–54033, 2025.

Jiakai Zhang, Shouchen Zhou, Haizhao Dai, Xinhang Liu, Peihao Wang, Zhiwen Fan, Yuan Pei, and Jingyi Yu. Cryofastar: Fast cryo-em ab initio reconstruction made easy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8462–8471. IEEE, 2025.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021a.

Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4046–4055. IEEE, 2021b.

A Appendix

A.1 Pose Frontend and Reference Coordinates

The pose frontend uses a ViT-B/14 image encoder initialized with Cryo-IEF weights (Yan et al., 2026) and a reference–target interaction decoder. The decoder predicts dense point maps and confidence scores (Wang et al., 2024; Zhang et al., 2025). Confidence-weighted plane fitting followed by Gram–Schmidt orthogonalization recovers rotations, while a separate head predicts detector-plane shifts. A residual calibrator refines the rotations and shifts.

Let $\widehat{Q}_i$ and $\widehat{Q}_a$ denote calibrated rotations in a common intermediate frame, where a indexes the reference image. The reference-relative rotations are

$$\widehat{R}_i^c = \widehat{R}_{i|a} = \widehat{Q}_i \widehat{Q}_a^\top, \qquad \widehat{R}_a^c = I_3. \tag{7}$$

Calibration precedes this composition. Each particle retains its individual detector-plane shift; the reference shift is not subtracted. The fixed reference establishes coordinate frame c and participates in pose estimation, while Gaussian evidence is aggregated from the support images.

A.2 Structure Conditioning and Dropout

During training, the pose frontend and Gaussian decoder use separate learned embeddings indexed by structure d. The decoder embedding modulates its initial query features through feature-wise linear modulation (FiLM) (Perez et al., 2018).

Let $\mathbf{v}_g$ denote the learned unmodulated query feature and $(\boldsymbol{\gamma}_d, \boldsymbol{\beta}_d)$ the structure-conditioned FiLM outputs. Reconstruction-condition dropout is applied as

$$\begin{aligned} m_k &\sim \mathrm{Bernoulli}(1 - p_{\mathrm{drop}}), \\ \mathbf{h}_g^{(0)} &= (\mathbf{1} + m_k \boldsymbol{\gamma}_d) \odot \mathbf{v}_g + m_k \boldsymbol{\beta}_d. \end{aligned} \tag{8}$$

We use $p_{\mathrm{drop}} = 0.5$. The same mask is shared across all queries and decoder evaluations within an episode. When $m_k = 0$, the initial features reduce to $\mathbf{h}_g^{(0)} = \mathbf{v}_g$. This mask controls the decoder's FiLM modulation; image features and known CTF metadata enter the decoder in both settings.

A.3 Reprojection and Image Discrepancy

Given support features and gate-selected geometry, the Gaussian decoder D predicts

$$\mathcal{G}_{\mathrm{t}}^{c} = D\left(F_{\mathrm{st}}, \widetilde{T}_{\mathrm{st}}^{c}, C_{\mathrm{st}}; d, m_k\right). \tag{9}$$

The renderer $\mathcal{A}$ performs orthographic Gaussian rasterization (Chen & Ling, 2025), followed by CTF modulation and detector-plane shifting under the cryo-EM image formation model (Levy et al., 2022a). For query particle i, the synthesized image is

$$\widehat{I}_{i,q\mathrm{t}} = \mathcal{A}\left(\mathcal{G}_{\mathrm{t}}^{c}; T_{i,q\mathrm{t}}^{\mathrm{gt},c}, \mathbf{c}_{i,q\mathrm{t}}\right). \tag{10}$$

Rendered and observed images use a consistent dataset-level polarity and intensity convention.

With an orthonormal two-dimensional Fourier transform $\mathcal{F}$, the normalized discrepancy is

$$\ell(\widehat{I}, I) = \frac{\left\| M \odot [\mathcal{F}(\widehat{I}) - \mathcal{F}(I)] \right\|_2^2}{\max\left(\|M \odot \mathcal{F}(I)\|_2^2, \epsilon\right)}. \tag{11}$$

Here, M is a circular frequency mask excluding the zero-frequency component, and ϵ stabilizes normalization. The expectation in Eq. (5) is taken over the structure, support and query sampling, geometry gate, and reconstruction-conditioning mask. Structures contribute equally, with query losses averaged within each structure.

The reconstruction objective updates the spatial encoder, learned queries, Gaussian decoder, and reconstruction-conditioning parameters. Predicted support poses are detached before entering the decoder, so this objective does not update the pose frontend.

A.4 Pose Objectives and Gradient Routing

During joint adaptation, the pose heads, residual calibrator, and selected final encoder blocks are optimized alongside the reconstruction branch. The supervised geometry objective is

$$\mathcal{L}_{\mathrm{pose}} = \mathcal{L}_{\mathrm{dense}} + \lambda_R \mathcal{L}_R + \mathcal{L}_{\mathrm{pair}} + 0.2\mathcal{L}_{\mathrm{shift}} + 0.01\mathcal{L}_{\mathrm{conf}}, \tag{12}$$

where $\lambda_R = 1$. The outer weights in the joint objective are $\lambda_{\mathrm{pose}} = \lambda_{\mathrm{phys}} = 1$.

Let $\mathcal{I}$ denote the particles receiving geometry supervision in an episode, with $N = |\mathcal{I}|$. Stars denote labeled geometry. For each particle, the point-map target is $X_{ip}^{*} = u_p R_i^{*}$, where u_p is a grid point in box-length-normalized coordinates and $P = 256$ is the number of grid points. The dense and individual rotation losses are

$$\begin{aligned} \mathcal{L}_{\mathrm{dense}} &= \frac{1}{3NP} \sum_{i \in \mathcal{I}} \sum_{p=1}^{P} \left\| \widehat{X}_{ip} - X_{ip}^{*} \right\|_2^2, \\ \mathcal{L}_R &= \frac{1}{9N} \sum_{i \in \mathcal{I}} \left\| \widehat{R}_i - R_i^{*} \right\|_F^2. \end{aligned} \tag{13}$$

Thus, the dense loss averages over particles, grid points, and coordinate components, while the rotation loss averages over particles and all nine matrix entries.

Let a index the reference particle and define $\mathcal{P} = \{(i,j) : i,j \in \mathcal{I} \setminus \{a\},\ i < j\}$. The pairwise rotation loss includes all non-reference pairs, including support–query pairs:

$$\mathcal{L}_{\mathrm{pair}} = \frac{1}{9|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \left\| \widehat{R}_i \widehat{R}_j^{\top} - R_i^{*} (R_j^{*})^{\top} \right\|_F^2. \tag{14}$$

Shifts are expressed in pixels on the 128×128 grid. We divide both predicted and labeled shifts by four and apply SmoothL1 with $\beta = 1$, averaging over particles and the two shift components:

$$\mathcal{L}_{\text{shift}} = \frac{1}{2N} \sum_{i \in \mathcal{I}} \sum_{j=1}^{2} \text{SmoothL1}_{\beta=1} \left(\frac{\widehat{t}_{ij}}{4}, \frac{t^*_{ij}}{4} \right). \tag{15}$$

Confidence targets are derived from detached point-map errors:

$$\begin{aligned} e_{ip} &= \frac{1}{3} \left\| \widehat{X}_{ip} - X^*_{ip} \right\|_2^2, \\ y_{ip} &= \text{clip} \left(\exp \left[-\frac{\text{sg}(e_{ip})}{0.05} \right], 0.001, 0.999 \right). \end{aligned} \tag{16}$$

With predicted confidence clipped to the same interval, $\bar{c}_{ip} = \text{clip}(\widehat{c}_{ip}, 0.001, 0.999)$, the confidence loss is the mean binary cross-entropy:

$$\mathcal{L}_{\text{conf}} = -\frac{1}{NP} \sum_{i \in \mathcal{I}} \sum_{p=1}^{P} \left[y_{ip} \log \bar{c}_{ip} + (1 - y_{ip}) \log(1 - \bar{c}_{ip}) \right]. \tag{17}$$

The auxiliary physical objective uses the same Gaussian decoder with labeled support geometry. The resulting density is detached before rendering with predicted query poses:

$$\begin{aligned} \mathcal{G}^c_{\text{t,gt}} &= D \left(F_{\text{st}}, T^{\text{gt},c}_{\text{st}}, C_{\text{st}}; d, m_k \right), \\ \widehat{I}^{\text{pose}}_{i,q\text{t}} &= \mathcal{A} \left(\text{sg}(\mathcal{G}^c_{\text{t,gt}}); \widehat{T}^c_{i,q\text{t}}, \mathbf{c}_{i,q\text{t}} \right), \\ \mathcal{L}_{\text{pose-proj}} &= \mathbb{E} \left[\frac{1}{|\mathcal{Q}_{\text{t}}|} \sum_{i \in \mathcal{Q}_{\text{t}}} \ell \left(\widehat{I}^{\text{pose}}_{i,q\text{t}}, I_{i,q\text{t}} \right) \right]. \end{aligned} \tag{18}$$

The two physical objectives use the same renderer and image discrepancy. For $\mathcal{L}_{\text{proj}}$, query geometry is labeled and gradients update the reconstructed density. For $\mathcal{L}_{\text{pose-proj}}$, query geometry is predicted and gradients update the pose frontend through a fixed density. The supervised and physical pose terms update the pose frontend and calibrator, while the reconstruction term updates the Gaussian reconstruction branch.

A.5 Inference and Density Evaluation

At inference, structure conditioning is disabled in both the pose frontend and Gaussian decoder. The pose frontend estimates support geometry $\widehat{T}^c_{si}$ from the input particles and fixed reference. With reconstruction FiLM disabled, the decoder prediction is

$$\mathcal{G}^c_i = \mathcal{D} \Big(F_{si}, \widehat{T}^c_{si}, C_{si}; m_{\text{inf}} \Big), \qquad m_{\text{inf}} = 0. \tag{19}$$

No structure-indexed embedding is used by either branch during inference. The predicted Gaussian field is evaluated at voxel centers to produce a density map. Different output grids sample the same continuous field at different spatial intervals. Pose estimation and Gaussian decoding use fixed network parameters and require no optimization for the input particle set.

A.6 Additional Experimental Details

Pretraining data and online noise. We construct the pretraining corpus from 1,000 PDB structures (Berman et al., 2000). Each structure is converted into a three-dimensional density, from which 2,500 clean particle projections are simulated with known viewing geometry and CTF parameters, yielding 2.5 million projections in total. Both the pose frontend and Gaussian reconstruction network are pretrained on this corpus. Fresh Gaussian noise is sampled whenever a clean projection is used for training, exposing repeated presentations to different noise realizations.

For the synthetic benchmark training sets, the online noise level is SNR 0.1. Their noisy test projections are generated independently of the clean training projections.

Table 4: Accuracy thresholds for promotion to each predicted-pose probability. Rotation entries are angular errors in degrees; shift errors are measured in pixels on the 128×128 grid. Each entry is an upper bound. The criteria must hold for the fixed reference and both new-reference evaluations.

Target p_d	Rotation mean	Rotation 90th percentile	Pairwise rotation mean	Shift mean
0.10	45°	100°	60°	2.0 px
0.25	30°	75°	45°	1.5 px
0.50	20°	45°	30°	1.0 px

Experimental image preprocessing. EMPIAR-10028 (Wong et al., 2014) and EMPIAR-10180 (Plaschka et al., 2017) images are downsampled to 128×128 pixels, with effective pixel sizes of 3.76875 Å/pixel and 4.2475 Å/pixel, respectively. The corresponding Nyquist resolutions are 7.5375 Å and 8.495 Å. When an FSC curve remains above the evaluated threshold at this boundary, its resolution is reported as an upper bound. The EMPIAR-10028 boundary is rounded to 7.54 Å in Table 1 and 7.538 Å in Table 3.

Experimental geometry annotations. For EMPIAR-10180, training rotations and detector-plane shifts are obtained from the metadata distributed with the official EMPIAR archive.[1] For EMPIAR-10028, these labels are obtained from cryoSPARC particle exports (Punjani et al., 2017). The annotations provide geometric supervision during training. The superscript gt denotes the geometry labels used for supervision, including estimated poses for experimental data. During evaluation, support poses are predicted by CryoFIG from the particle images.

Optimization and implementation. Shared-model adaptation is performed on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. Training episodes contain 32 support particles and 16 disjoint query particles, with query projections rendered in micro-batches of four. The decoder predicts 32,768 anisotropic Gaussians.

The pose encoder is initialized with Cryo-IEF ViT-B/14 weights (Yan et al., 2026). During joint adaptation, the pose heads, residual calibrator, and up to the final six encoder blocks are optimized alongside the reconstruction branch. We use AdamW (Loshchilov & Hutter, 2019) with weight decay 0.01 in the pose stages and zero in the calibration and reconstruction stages. The reconstruction learning-rate schedule uses 10^{-4}, 3×10^{-5}, and 10^{-5} for three successive stages of 3,000 steps each. Reconstruction-condition dropout is applied with probability 0.5, and main evaluations use $m_{\mathrm{inf}} = 0$.

Accuracy-gate schedule. Each structure maintains its own predicted-pose probability $p_d \in \{0, 0.10, 0.25, 0.50\}$, initialized to zero. Training uses labeled support geometry for at least the first 600 optimization steps. Table 4 lists the validation thresholds required to enter each nonzero probability level. All thresholds must be satisfied under the fixed-reference protocol and under each of two new-reference evaluations. A promotion requires two consecutive passing validation checks separated by at least 300 optimization steps, and successive promotions must be at least 600 steps apart.

We additionally monitor diagnostic reconstruction NMSE using labeled geometry for each structure. Promotion requires this NMSE to be no more than 0.002 above its historical minimum. An increase greater than 0.002 triggers demotion to the immediately lower probability level, with a lower bound of zero. Otherwise, the probability is held constant unless all promotion conditions are met. The Bernoulli gate selects labeled or predicted geometry for the entire support set. Predicted support poses are detached before entering the reconstruction branch.

Map alignment and FSC evaluation. Reconstructed volumes are aligned to the corresponding ground-truth or experimental reference maps using CryoSPARC (Punjani et al., 2017). Masked map-to-map FSC (Penczek, 2010) is computed using EMAN2 (Tang et al., 2007). Soft masks use a threshold of 50% of the maximum density, dilation by 25 Å, and a 15 Å cosine edge, following the mask parameters reported for CryoDRGN-AI (Levy et al., 2025).

[1] `https://www.ebi.ac.uk/empiar/EMPIAR-10180/`

Local resolution and visualization. For the synthetic benchmarks, local resolution is estimated against the ground-truth density using CryoSPARC (Punjani et al., 2017). The estimates are mapped onto reconstructed density surfaces using UCSF ChimeraX (Pettersen et al., 2021), with the corresponding distributions shown in Figures 3 and 4.

A.7 Fourier Shell Correlation

Fourier shell correlation (FSC) quantifies the agreement between two density maps as a function of spatial frequency (Penczek, 2010). Let $A(\mathbf{k})$ and $B(\mathbf{k})$ denote the three-dimensional Fourier transforms of the aligned, soft-masked reconstruction and reference maps, respectively. For a Fourier shell $\mathcal{S}(f)$ centered at spatial frequency f, FSC is defined as

$$\mathrm{FSC}(f) = \frac{\mathrm{Re}\left[\sum_{\mathbf{k}\in\mathcal{S}(f)} A(\mathbf{k})B(\mathbf{k})^*\right]}{\sqrt{\left(\sum_{\mathbf{k}\in\mathcal{S}(f)} |A(\mathbf{k})|^2\right)\left(\sum_{\mathbf{k}\in\mathcal{S}(f)} |B(\mathbf{k})|^2\right)}}. \tag{20}$$

Here, * denotes complex conjugation. Values close to one indicate strong agreement within a frequency shell, whereas values near zero indicate weak correlation. Higher spatial frequencies correspond to finer structural detail.

Reference maps and threshold-based resolution. We compute map-to-map FSC against the ground-truth density for synthetic data and the experimental reference map for real data. Alignment and masking procedures are described in Appendix A.6. For a threshold τ, a crossing at spatial frequency f_τ corresponds to the resolution

$$d_\tau = \frac{1}{f_\tau}, \tag{21}$$

where f_τ is measured in Å^{-1} and d_τ in Å. A smaller d_τ indicates agreement with the reference extending to finer spatial scales. We use $\tau = 0.5$ for the main comparisons and additionally report $\tau = 0.143$ for experimental data. Both thresholds summarize reconstruction–reference agreement.

Nyquist-limited values. For voxel spacing Δ, the Nyquist frequency is $f_{\mathrm{Nyq}} = 1/(2\Delta)$, corresponding to $d_{\mathrm{Nyq}} = 2\Delta$. The experimental datasets have Nyquist resolutions of 7.5375 Å for EMPIAR-10028 and 8.495 Å for EMPIAR-10180. When an FSC curve remains above the specified threshold throughout the evaluated frequency range, we report $d_\tau \leq d_{\mathrm{Nyq}}$. This notation identifies a sampling-limited bound: the threshold crossing is unresolved within the available frequency range. Consequently, identical Nyquist-limited values do not establish identical reconstruction quality.

A.8 Feature Visualization Across Datasets

We examine particle representations extracted by the fine-tuned Cryo-IEF ViT-B/14 encoder (Yan et al., 2026) in the pose frontend. For each particle image, we extract the tokens after the final Transformer block and final layer normalization. We exclude the CLS token and average the remaining 256 patch tokens to obtain a 768-dimensional feature vector.

The visualization includes EMPIAR-10028 (Wong et al., 2014), EMPIAR-10049 (Ru et al., 2015), EMPIAR-10180 (Plaschka et al., 2017), EMPIAR-10076 (Davis et al., 2016), and the clean and SNR 0.1 versions of Spike and HSP. We use 2,000 particle images from each of these eight groups, yielding 16,000 feature vectors. All vectors are jointly embedded into a common two-dimensional space using UMAP (McInnes et al., 2018). The noise-free synthetic input condition and the EMPIAR-10049 and EMPIAR-10076 datasets are excluded from fine-tuning.

Figure 6 shows distinct groupings for Spike and HSP under both clean and noisy conditions, with visible displacement between the clean and noisy group centroids. Among the experimental datasets, EMPIAR-10049 forms a largely separate group, while EMPIAR-10028, EMPIAR-10180, and EMPIAR-10076 occupy neighboring, partially overlapping regions. The visualization provides a qualitative view of feature organization across datasets and noise conditions, including inputs beyond those used for fine-tuning.

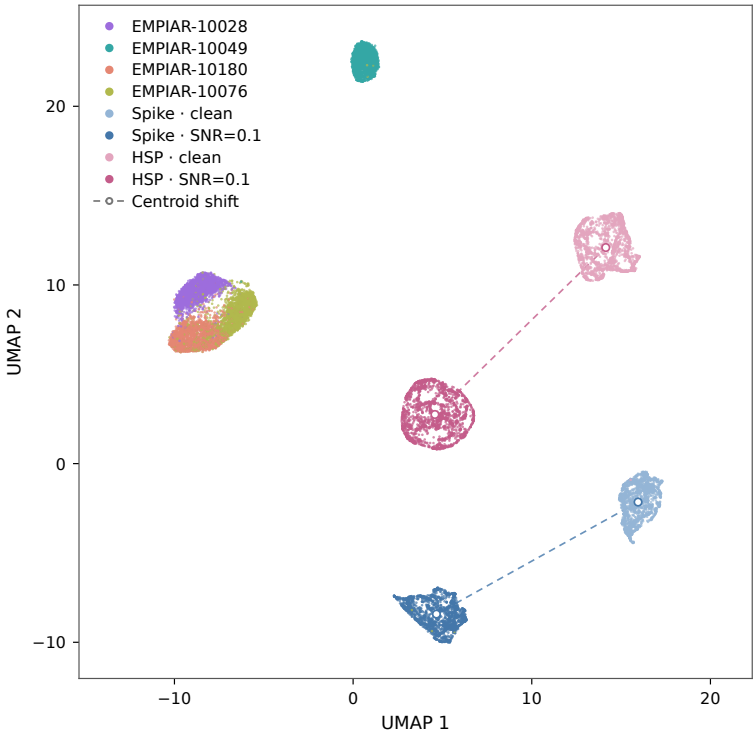


Figure 6: **UMAP visualization of pose-encoder features after fine-tuning.** Each solid point represents one particle image, encoded by averaging the 256 patch tokens after the final layer normalization. All 16,000 feature vectors are jointly embedded, with 2,000 images per group. Colors identify datasets and synthetic noise conditions. Hollow circles mark group centroids, and dashed lines connect the clean and noisy centroids of each synthetic dataset. The clean input condition, EMPIAR-10049, and EMPIAR-10076 are excluded from fine-tuning.