CryoGEM: Physics-Informed Generative Cryo-Electron Microscopy

Jiakai Zhang[1,2*] **Qihe Chen**[1,2*] **Yan Zeng**[1,2] **Wenyuan Gao**[1]

Xuming He[1] **Zhijie Liu**[1,3] **Jingyi Yu**[1]

[1]ShanghaiTech University
[2]Cellverse
[3]iHuman Institute

{zhangjk,chenqh2024, zengyan2024,gaowy,hexm,liuzhj,yujingyi}@shanghaitech.edu.cn

Abstract

In the past decade, deep conditional generative models have revolutionized the generation of realistic images, extending their application from entertainment to scientific domains. Single-particle cryo-electron microscopy (cryo-EM) is crucial in resolving near-atomic resolution 3D structures of proteins, such as the SARS-COV-2 spike protein. To achieve high-resolution reconstruction, a comprehensive data processing pipeline has been adopted. However, its performance is still limited as it lacks high-quality annotated datasets for training. To address this, we introduce physics-informed generative cryo-electron microscopy (CryoGEM), which for the first time integrates physics-based cryo-EM simulation with a generative unpaired noise translation to generate physically correct synthetic cryo-EM datasets with realistic noises. Initially, CryoGEM simulates the cryo-EM imaging process based on a virtual specimen. To generate realistic noises, we leverage an unpaired noise translation via contrastive learning with a novel mask-guided sampling scheme. Extensive experiments show that CryoGEM is capable of generating authentic cryo-EM images. The generated dataset can be used as training data for particle picking and pose estimation models, eventually improving the reconstruction resolution.

1 Introduction

In the past decade, deep generative models like VAEs [24], GANs [13], and diffusion models [16] have achieved significant success in conditional image generation. Recently, Stable Diffusion [43] can produce high-quality images given simple textual descriptions. Additional controls [62] can be imposed to produce tailored visual effects, e.g., theatrical lighting [39] and specific perspectives [31]. In fact, the successes of image generation have gone way beyond visual pleasantness, stimulating significant advances in scientific explorations. Examples include brain magnetic resonance imaging to computational tomography (MRI-to-CT) translation [58], X-ray image generation [49], etc. Different from entertainment applications, generation techniques for scientific imaging should faithfully follow the physical process: the generated results would eventually be applied to real-world downstream tasks such as medical image diagnosis. In this work, we extend image generation to a specific biomolecular imaging technique, single-particle cryo-electron microscopy (cryo-EM) to improve the performance of its downstream tasks. Cryo-EM aims to recover the near-atomic resolution 3D structure of proteins, with its latest application in recovering the SARS-COV-2 spike protein [59] structure for drug development.

*The authors contributed equally to this work.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2019

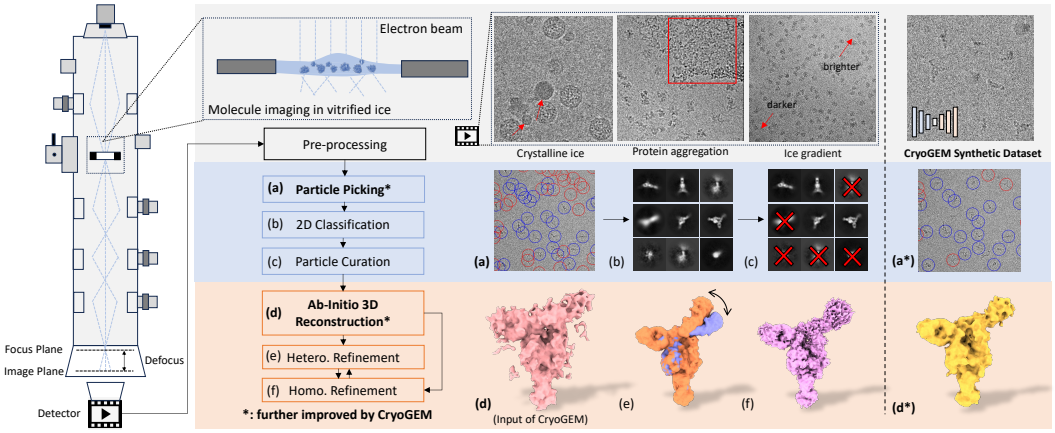

Figure 1: **CryoGEM improves cryo-EM data analysis.** Cryo-EM captures images of molecules in vitrified ice via electron beams. Data is processed for a high-resolution 3D reconstruction by a comprehensive pipeline. However, some modules like (a) particle picking and (d) ab-initio 3D reconstruction still lack high-quality training datasets. Given a coarse result as an input, CryoGEM can synthesize authentic single-particle micrographs as training dataset augmentation.

As shown in Figure 1, a comprehensive data processing pipeline of single-particle cryo-EM starts with capturing transmission images of flash-frozen purified specimens, termed as *micrographs*, using high-energy electron beams. The complete dataset contains hundreds of thousands of images of target particles with unknown locations, poses, and shapes. The main challenge in cryo-EM is to accurately estimate the locations and orientations of particles in the extremely low signal-to-noise ratio (SNR), mainly caused by detector shot noise in limited dosage conditions and other structural noises such as global ice gradients. Thus, the performance of existing methods for particle picking [4] and pose estimation [28] are limited due to the lack of high-quality annotated training datasets, which are labor-intensive for human experts.

In this paper, we present a novel physics-informed generative cryo-electron microscopy (CryoGEM) technique, which is capable of generating authentic annotated synthetic datasets using just 100 unannotated micrographs from a real dataset. To achieve highly controllable generation results, we introduced a simple yet highly controllable physical simulation process. Based on the coarse density volume, we achieve control at both the particle level and the micrograph level. However, this physical simulation still lacks critical authentic noise modeling. We thus adopt unpaired image-to-image translation to generate authentic noises. Existing methods like CycleGAN [68] assume that the source and target domains are bijective; however, real cryo-EM micrographs have random noises, existing contrastive learning methods, such as contrastive unpaired translation (CUT) [37], relax these assumptions by using a random sampling strategy for positive and negative samples. However, in cryo-EM, the randomly sampled positives and negatives can be semantically similar since particles are densely located in a micrograph. To address these challenges, we employ a novel contrast noise translation to transform the simulated micrographs into authentic cryo-EM micrographs. Furthermore, we use a particle-background segmentation paired with the simulated result as a guide for positive and negative sample selection. We show that precise guidance on sample selection can significantly improve the quality of image generation.

We validate CryoGEM on five diverse and challenging real cryo-EM datasets. Extensive experiments show that CryoGEM achieves significantly better visual quality compared with state-of-the-art methods. Also, we demonstrate that the performance of existing deep models in downstream tasks, including particle picking and particle pose estimation, can be significantly improved by training on our synthetic dataset. Notably, we achieve 44% picking performance improvements, leading to 22% better resolution of final reconstruction on average.

https://doi.org/10.52202/079017-2019

2 Related Works

Our work aims to extend generation methods to the field of cryo-EM. We therefore only discuss the most relevant works in respective fields.

Cryo-EM Pipeline. In recent decades, the CryoEM pipeline has rapidly evolved. Popular software such as CryoSPARC [41], Relion [70], and Warp [48] has been widely used for high-resolution 3D reconstruction of macromolecules. Two critical steps are particle picking and pose estimation. To achieve accurate particle picking, recent neural methods [4, 10, 51, 11, 14, 56] have built upon several manually annotated real datasets or incorporated with few-shot learning techniques. However, they still lack generalization capability as the training dataset only covers a small portion of real scenarios. In pose estimation, traditional methods [70, 41] propose a Maximum-A-Posteriori (MAP) optimization by Expectation–Maximization (EM). Recent self-supervised models [28, 27, 67] have been adopted for ab-initio reconstruction, i.e., 3D reconstruction from particle images with unknown poses. However, their performance is still limited without further fine-grained refinement processes. We improve the performance of particle picking and pose estimation by generating high-quality annotated synthetic datasets as training datasets.

Cryo-EM Simulations. Theoretical simulation techniques, based on physical priors, combine atomic-level simulations with global projection to accurately compute electron scattering during the imaging process [63, 50, 45, 30, 32, 15]. Traditional simulation methods such as InsilicoTEM [63] model the interaction between electrons by taking specimens as multi-slices to improve the algorithm's performance. However, they typically require expensive computational resources and may require complex adjustments of parameters to achieve ideal results. Our work also includes an efficient physics-based simulation module to present the structural information of particles. Additionally, we generate realistic noises via a novel unpaired noise translation technique.

Unpaired Image-to-image Translation. Deep generative models including GAN-based methods [68, 61, 23, 29, 17] and diffusion models [46, 55] are widely used in unpaired image-to-image translation tasks. CycleGAN [68] introduces generative adversarial networks to calculate cycle consistency losses, allowing for training on unpaired data [20, 26, 38, 22]. TraVeLGAN [2], DistanceGAN [3], and GcGAN [12] achieve one-way translation while avoiding traditional cycle consistency. Diffusion models [7, 46, 34, 55, 64, 57, 40] have been recently introduced to unpaired image-to-image translation. But they often demonstrate results in low resolution and they often rely on a large-scale training dataset. Recently, Contrastive Unpaired Translation (CUT) [37] introduces a new generative framework via contrastive learning [21, 52, 65] to propose a more efficient training framework. However, existing methods do not combine the physical process as additional constraints. In contrast, our method includes a physical simulation during authentic cryo-EM micrograph generation.

3 Physics-informed Generative Cryo-EM

We propose CryoGEM, the first method to combine a physics-based simulation process with a novel contrastive noise generation technique (Figure 2). Our approach begins with preparing a virtual specimen containing numerous uniformly distributed target protein structures (Section 3.1). We then emulate the cryo-EM imaging process to introduce physical constraints (Section 3.2). To generate authentic cryo-EM noise, we use a novel unpaired noise translation technique via contrastive learning guided by a particle-background mask and detail the final objective during training. (Section 4)

3.1 Virtual Specimen Preparation

To simulate the cryo-EM imaging process, we first prepare the virtual specimen $S(x, y, z)$, a large 3D density volume that contains multiple copies of the target molecule's coarse result with randomly generated locations, orientations, and conformations. To obtain the coarse result, we employ cryoSPARC [41] for ab-initio reconstruction, followed by cryoDRGN [66] for a continuous heterogeneous reconstruction to obtain a neural volume

$$V(\mathbf{p}, \mathbf{w}) : \mathbb{R}^3 \times \mathbb{R}^8 \mapsto \mathbb{R}, \tag{1}$$

where $\mathbf{p} = (x, y, z)^{\mathrm{T}} \in \mathbb{R}^3$ represents the spatial coordinates, $\mathbf{w} \in \mathbb{R}^8$ denotes a high-dimensional conformational embedding. This neural volume can generate a density volume given a learned conformational embedding following CryoDRGN [66]. Notably, such a coarse result can be easily

 https://doi.org/10.52202/079017-2019

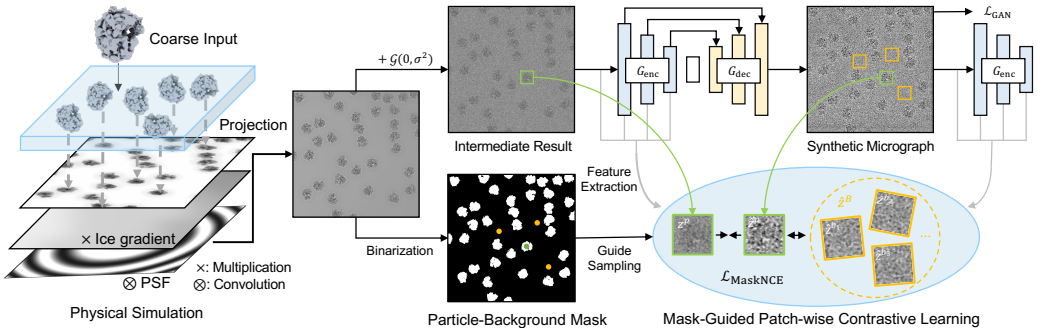

Figure 2: **Pipeline of CryoGEM.** We begin by creating a virtual specimen containing various initial reconstruction results. We then simulate the imaging process of cryo-EM, incorporating physical priors such as ice gradient and point spread function (PSF) to generate a physical simulation. By adding simple Gaussian noise to the physically simulated results, we introduce randomness within a contrastive learning framework. To enhance training efficiency and performance, we use the particle background mask as a guide for patch sampling. The sampled positive and negative instances are then encoded into multi-scale features for contrastive learning. Additionally, we introduce an adversarial loss to ensure realistic cryo-EM image synthesis.

resolved, but the final result requires further iterative optimizations of selected particles, estimated particle poses, 3D templates, etc.

During the specimen preparation, we sequentially add N particles $V(\mathbf{p}, \mathbf{w})$ into an empty virtual specimen. During every addition, we randomly sample orientation matrices $R_i \in \mathbb{R}^{3\times3}$ from SO(3) space, random conformations $\mathbf{w}_i$ from the learned conformational space, as well as translations $\mathbf{t}_i \in \mathbb{R}^3$ from unoccupied areas of the virtual specimen. Both sampled orientations and locations can be further used as ground-truth annotations. Thus, the virtual specimen can be expressed as:

$$S(x, y, z) = \sum_{i=1}^{N} V(R_i\mathbf{p} + \mathbf{t}_i; \mathbf{w}_i), \tag{2}$$

where the number of particles $N \sim \mathcal{N}(\mu_N, \sigma_N^2)$, with the mean μ_N and standard deviation σ_N derived from the actual distribution of particle count in real micrographs. We ensure that the minimum $N_{\min}$ and maximum $N_{\max}$ values are within two standard deviations from the mean.

3.2 Emulating the Imaging Process

We emulate the physical imaging process of cryo-EM including electron-specimen interaction, ice gradient, and cryo-EM optical aberration based on virtual specimen. For simplicity, we consider the complex electron-specimen interaction as an orthogonal projection of the specimen by applying the weak phase object approximation [53] (WPOA). To obtain the true ice gradients, we estimate them on real micrographs by IceBreaker [36], which can generate a weight map of ice, $W(x, y)$, whose every pixel represents the attenuation ratio compared to the maximum value of intensity. Also, we model the optical aberrations as a point spread function (PSF) g, which is the Fourier transform of the contrast transfer function (CTF), by off-the-shelf software CTFFIND4 [42]. To sum up, the 2D physics-based result $I_{\text{phy}}(x, y)$ can be expressed as:

$$I_{\text{phy}}(x, y) = g * \left[W(x, y) \cdot \int_{\mathbb{R}} S(x, y, z) \mathrm{d}z \right]. \tag{3}$$

During the simulation process, we randomly select a pair of weight maps and PSF estimated from real micrographs.

4 Contrastive Noise Generation

Real cryo-EM micrographs contain high-level noises related to specimen-electron interaction and detector [50]. Thus, accurate noise modeling is essential for authentic synthetic cryo-EM micrograph

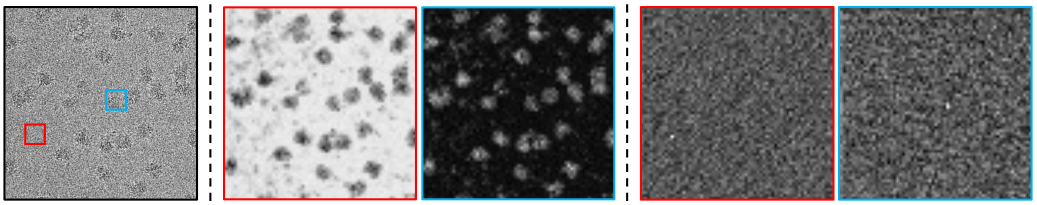

Figure 3: **Visualization of the learned similarity.** Given query patches (red for particle and blue for background) on the input real micrograph, we visualize the learned similarity maps of our and CUT's encoders G_{enc} by calculating $\exp(G_{enc}(v) \cdot G_{enc}(v^-)/\tau)$, where v denotes the query and v^- denotes the patches of real micrograph. The results imply that our encoder can recognize particles and backgrounds in real cryo-EM micrographs. However, CUT fails to learn that without the guidance of particle-background maps during training.

generation. We model the noise generation as an unpaired image-to-image translation task. The input, $I_{\text{phy}} \sim X$, learns the actual noise distribution from real cryo-EM images $I_{\text{real}} \sim Y$, to generate synthetic cryo-EM images $I_{\text{syn}} \sim \hat{Y}$. The generative model includes a discriminator D, and a generator $G = G_{\text{dec}} \circ G_{\text{enc}}$, where G_{enc} is a encoder and G_{dec} is a decoder.

Non-deterministic noise translation. Existing contrastive learning methods such as CUT [37] have shown significant promise for efficient training and generating high-quality natural images. Nonetheless, they do not account for generating random noise in cryo-EM, i.e., they can only generate synthetic micrographs in a deterministic way, against the nature of the random noise generation process, leading to an unstable training process and degraded performance. To address this, we introduce a random process by adding a zero-mean Gaussian random noise $\mathcal{G} \sim \mathcal{N}(0, \sigma^2)$ into the physical simulations, where $\sigma^2 = \text{var}(I_{phy})/\text{SNR}$ (SNR equals to 0.1 in our experiments). We introduce an intermediate result $I_{inter} = I_{phy} + \mathcal{G} \sim \hat{X}$, where $\hat{X} = X + \mathcal{N}(0, \sigma^2)$. Thus, the synthetic image can be defined as:

$$I_{\text{syn}} = G(I_{inter}) \tag{4}$$

This simple strategy significantly improves our noise generation quality by constructing a mapping of a single physical simulation result to infinite synthetic noisy results with varied noise patterns.

Mask-Guided sampling scheme. A fundamental assumption of patch-wise contrastive learning is that the randomly chosen negatively sampled patches are semantically different from the positive samples [8]. But in a single cryo-EM micrograph, hundreds of particles are densely distributed against an almost pure yet noisy background, which means that randomly chosen negative and positive patches are likely semantically similar. Taking advantage of our highly controllable physical simulation process, we introduce a mask-guided sampling scheme by generating a binary particle-background mask paired with I_{phy}, denoted as $M(x, y)$, where $M(x, y) = 1$ indicates a particle. Guided by this mask, our selection process can choose positive and negative samples from locations with contrasting mask labels. This significantly improves the encoder's performance, as shown in Figure 3, our encoder can generate accurate similarity maps given particles or backgrounds as references on real images, but CUT's encoder fails to recognize them.

Mutual information extraction. Guided by particle-background mask, we select Q paired patches in I_{phy} and I_{syn} as the positive samples and queries on particle region, respectively. We then choose K patches in I_{phy} as negative samples on the background region. The encoder G_{enc} maps these to normalized vectors $\boldsymbol{v} = \{\boldsymbol{v}_q\}_{q=1}^{Q}$, $\boldsymbol{v}^+ = \{\boldsymbol{v}_q^+\}_{q=1}^{Q}$, and $\boldsymbol{v}^- = \{\boldsymbol{v}_k^-\}_{k=1}^{K}$, respectively. Our objective is to maximize the likelihood of selecting a positive sample by minimizing the cross-entropy loss:

$$\ell(\boldsymbol{v}, \boldsymbol{v}^+, \boldsymbol{v}^-) = -\sum_{q=1}^{Q} \log \left[\frac{\exp(\boldsymbol{v}_q \cdot \boldsymbol{v}_q^+/\tau)}{\exp(\boldsymbol{v}_q \cdot \boldsymbol{v}_q^+/\tau) + \sum_{k=1}^{K} \exp(\boldsymbol{v}_q \cdot \boldsymbol{v}_k^-/\tau)} \right], \tag{5}$$

where τ is a temperature factor that scales the distance between the query and samples.

Mask-guided patch-wise contrastive learning. We leverages multiple intermediate layers of G_{enc} to extract multi-scale features from input patches. Notably, we align the receptive field with particle size by selecting the G_{enc}'s first L layers. Feature map from each layer is then passed through

https://doi.org/10.52202/079017-2019

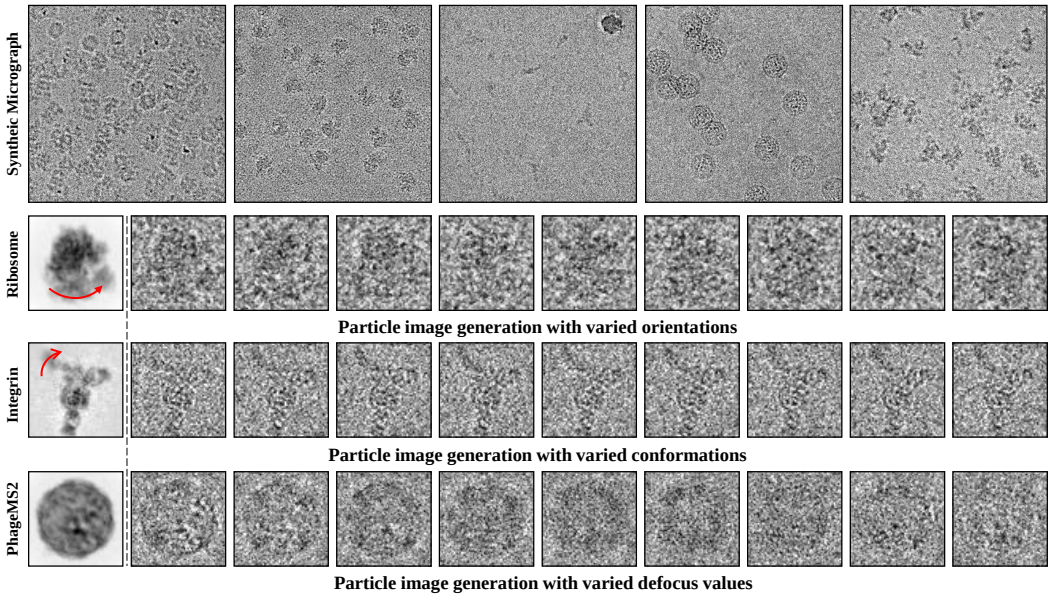

Figure 4: **Result gallery.** The first row showcases the diverse synthetic contents generated by CryoGEM (from left to right for Proteasome, Ribosome, Integrin, PhageMS2, and HumanBAF). The second, third, and fourth rows demonstrate CryoGEM's ability to control the particle's pose, conformation, and defocus during the image generation. In every row, the leftmost column is the clean particle image, and the controlled variable is smoothly adjusted from left to right.

compact two-layer MLP networks H_l ($l \in \{1, 2, ..., L\}$) to generate a feature $\boldsymbol{z}_l = H_l(G^l_{\text{enc}}(I_{\text{inter}}))$, where G^l_{enc} is the l-th layer. We define those on particles as $p \in P$, and those on the background as $b \in B$. The feature of l-th layer at a particle position p is denoted as $\boldsymbol{z}^p_l$, and the set represented as $\boldsymbol{z}^P_l = \{\boldsymbol{z}^p_l\}_{p \in P}$. Similarly, for background positions, we denote the l-th layer feature as $\boldsymbol{z}^b_l$, with the set represented as $\boldsymbol{z}^B_l = \{\boldsymbol{z}^b_l\}_{b \in B}$. Features encoded from the output I_{syn} are represented as $\hat{\boldsymbol{z}}$, where $\hat{\boldsymbol{z}}_l = H_l(G^l_{\text{enc}}(I_{\text{syn}}))$. We propose a novel mask-guided Noise Contrastive Estimation (NCE) loss for efficient contrastive learning in cryo-EM:

$$\mathcal{L}_{\text{MaskNCE}}(G, H, \hat{X}) = \mathbb{E}_{I_{\text{inter}} \sim \hat{X}} \sum_{l=1}^{L} \left[\sum_{p \in P} \ell\left(\boldsymbol{z}^p_l, \hat{\boldsymbol{z}}^p_l, \hat{\boldsymbol{z}}^B_l\right) + \sum_{b \in B} \ell\left(\boldsymbol{z}^b_l, \hat{\boldsymbol{z}}^b_l, \hat{\boldsymbol{z}}^P_l\right) \right]. \tag{6}$$

By minimizing $\mathcal{L}_{\text{MaskNCE}}$, we effectively differentiate particle and background features (Figure 3).

Final objective. We introduce an adversarial loss function, $\mathcal{L}_{\text{GAN}}$, to encourage the network to produce more realistic simulated cryo-EM images, I_{syn}, as follows:

$$\mathcal{L}_{\text{GAN}}(G, D, \hat{X}, Y) = \mathbb{E}_{I_{\text{real}} \sim Y}[\log D(I_{\text{real}})] + \mathbb{E}_{I_{\text{inter}} \sim \hat{X}}[\log(1 - D(G(I_{\text{inter}})))].$$

Combining this with the contrastive learning loss $\mathcal{L}_{\text{MaskNCE}}$, the overall training loss function is:

$$\mathcal{L}_{\text{GAN}}(G, D, \hat{X}, Y) + \lambda \mathcal{L}_{\text{MaskNCE}}(G, H, \hat{X}), \tag{7}$$

we set the hyper-parameter λ to 10.0 in all our experiments.

5 Experiments

Datasets. We evaluate CryoGEM across five challenging cryo-EM datasets. Each includes expert-curated particle-picking annotations for high-resolution reconstruction of target molecules. Note that the particle annotations in the dataset are solely used for validation of downstream tasks and

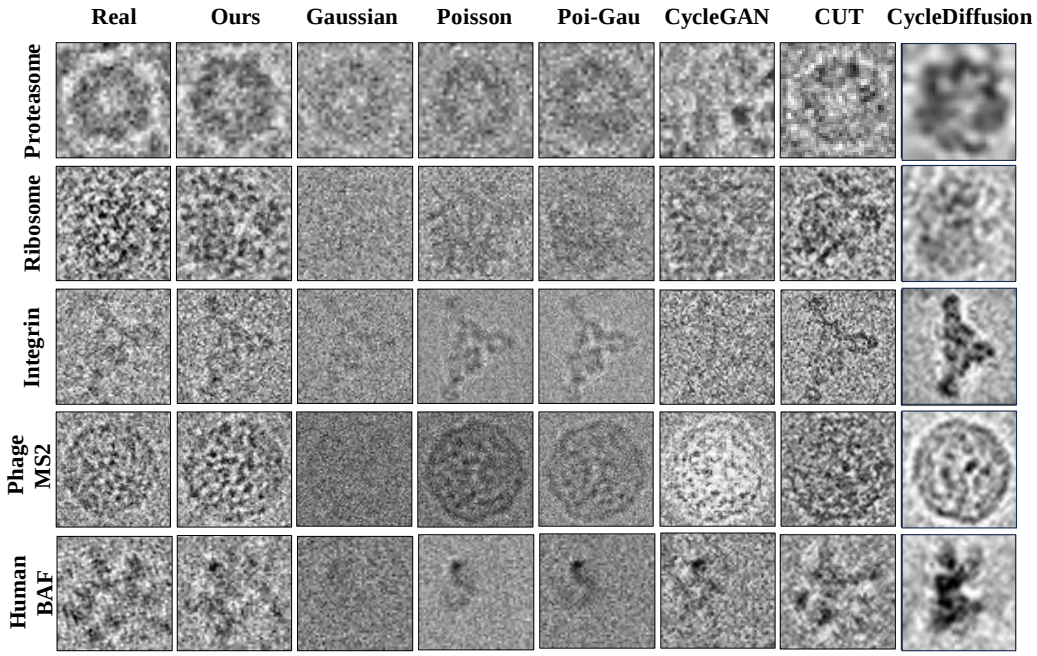

Figure 5: **Qualitative comparison results.** Our approach achieves the most authentic noise generation across all datasets. The traditional noise models succeed in preserving the structural information while lacking realistic noise patterns. CycleGAN, CUT, and CycleDiffusion introduce severe artifacts on generated results.

Table 1: **Quantitative comparison of visual quality.** Our approach consistently achieves the best performance in FID metric.

Metric	FID↓					
Dataset	Proteasome	Ribosome	Integrin	PhageMS2	HumanBAF	Avg.
Gaussian	89.87	177.01	174.94	162.81	46.48	130.22
Poisson	112.40	86.88	173.94	343.16	44.17	152.11
Poi-Gau	93.41	73.89	173.64	311.50	45.55	139.60
CycleGAN	89.33	27.73	54.83	422.80	137.41	146.42
CUT	46.61	44.23	49.96	88.65	74.76	60.84
CycleDiffusion	470.37	173.62	386.83	468.83	577.46	415.42
Ours	**42.96**	**6.54**	**42.46**	**63.11**	**34.50**	**37.91**

are not utilized during the training phase of our CryoGEM model. 1) T20S **Proteasome** (EMPIAR-10025) [6]' micrographs exhibit a high density of particles, leading to occlusions. posing significant challenges in particle picking. Also, the 3D structure is D7 symmetric, which brings ambiguities in the pose estimation task. 2) 80S **Ribosome** (EMPIAR-10028) [54] has a complex 3D structure, making it difficult to estimate the poses of particle images. 3) Asymmetric $\alpha \mathrm{V} \beta 8$ **Integrin** (EMPIAR-10345) [5] contains molecules with varied conformations, requiring a heterogeneous reconstruction. 4) **PhageMS2** (EMPIAR-10075) [25, 11]'s micrographs contain enormous spherically-shaped virus particles, suitable for testing the generalizability of the particle picking model. 5) Endogenous **HumanBAF** complex (EMPIAR-10590) [33, 11] has significant ice gradient artifacts. We kindly refer to Appendix B for a more detailed dataset description and illustration.

Baselines. We evaluate the performance of CryoGEM compared to several traditional noise baselines and deep generative models. We use **Poisson** noise to simulate the high-level detector shot noises in cryo-EM images, **Gaussian** noise for a general simplified noise modeling in cryo-EM, and Poisson-Gaussian mixed noise (**Pos-Gau**) as traditional baselines. We choose **CycleGAN**, **CUT**, and recent **CycleDiffusion** as deep generative baselines. In Appendix C, we detail their specific settings.

Implementation details. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. As a lightweight model, CryoGEM trains 100 epochs on a single dataset within an hour using less than 10 GB of memory. The training dataset (all held out for the evaluation) is comprised of 100 real images alongside 100 physics-based simulated results. We enhance the dataset through

https://doi.org/10.52202/079017-2019

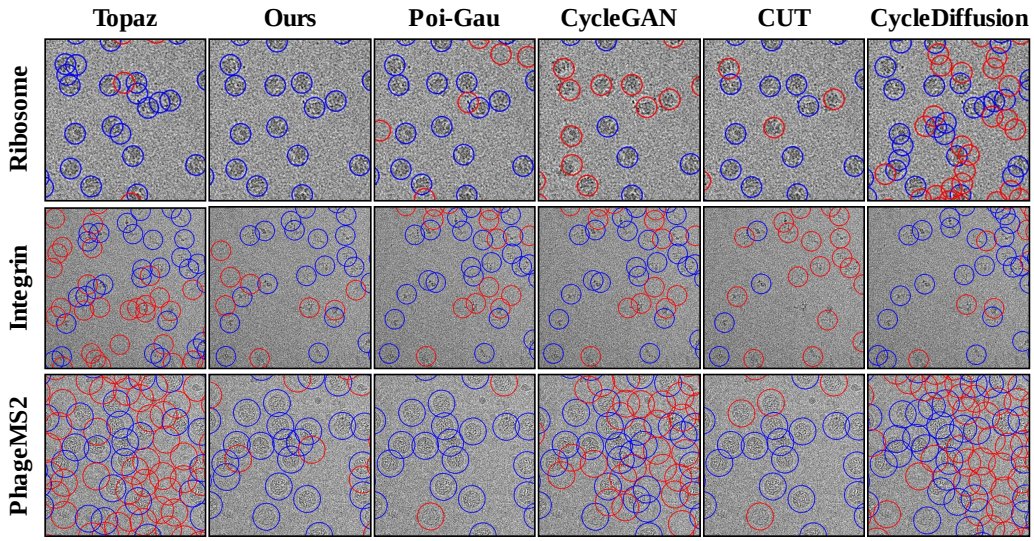

Figure 6: **Qualitative comparison results of particle picking.** The blue circles indicate matches with manual picking results, while the red circles represent misses or excess picks by the model.

Table 2: **Quantitative comparison of particle picking.** Our approach consistently achieves the best in AUPRC and Res(Å) metrics.

Metric	AUPRC↑						Res(Å) ↓					
Dataset	Proteasome	Ribosome	Integrin	PhageMS2	HumanBAF	Avg.	Proteasome	Ribosome	Integrin	PhageMS2	HumanBAF	Avg.
Gaussian	0.471	0.485	0.259	0.886	0.470	0.514	2.76	3.84	7.62	7.44	10.29	6.39
Poisson	0.469	0.375	0.213	0.490	0.483	0.406	2.77	4.16	7.52	10.50	10.67	7.12
Poi-Gau	0.455	0.618	0.210	0.706	0.381	0.474	2.80	3.87	8.13	10.90	13.18	7.77
CycleGAN	0.200	0.308	0.414	0.895	0.228	0.409	5.10	3.93	8.03	7.51	11.83	7.28
CUT	0.442	0.224	0.335	0.592	0.513	0.421	2.77	4.78	7.16	8.91	12.92	7.30
CycleDiffusion	0.346	0.205	0.233	0.469	0.392	0.329	2.92	4.56	6.35	9.51	6.68	6.66
Topaz	0.302	0.679	0.526	0.329	0.493	0.466	3.13	3.84	6.34	10.67	9.59	6.71
Ours	**0.490**	**0.797**	**0.606**	**0.915**	**0.562**	**0.674**	**2.68**	**3.25**	**5.54**	**7.16**	**7.74**	**5.27**

data augmentation, specifically by rotating the images by 90, 180, and 270 degrees. Please see Appendix B for more details. We choose PatchGAN [18] as our discriminator and UNet implemented by [37] as our generator. We fix the image resolution to 1024×1024 during training. Guided by the particle-background mask, we evenly sample 256 queries on particles and 256 on the background, ensuring that their corresponding negative samples are located where the labels are opposite.

5.1 Visual Quality

In Figure 4 , we demonstrate CryoGEM's ability to generate diverse content through multi-level controls, including adjustments to the ice gradient and defocus values of micrographs, as well as the particles' positions, orientations, and conformations. We then evaluate CryoGEM's visual quality compared to baselines. As illustrated in Figure 5, our approach consistently achieves the best qualitative results across all datasets in terms of overall noise patterns and the preservation of particle structural information. Traditional methods fail to mimic authentic noise distributions. CycleGAN struggles with convergence due to the breakdown of its bijection assumption. CUT can achieve better visual quality but its performance is still limited due to its random sampling scheme. CycleDiffusion cannot learn the the authentic noise patterns. As shown in Table 1, we also employed Frechet Inception Distance (**FID**) which is widely used in generative tasks to measure the similarity between generated and real data. Our results significantly outperform existing baseline methods in all five datasets. For additional conditional generated and zero-shot results, please refer to Figure 10 in the appendix.

5.2 Particle Picking

To evaluate how CryoGEM can enhance downstream particle picking task, we finetune the popular particle picking model **Topaz** [4] using synthetic annotated datasets from different baselines. Our

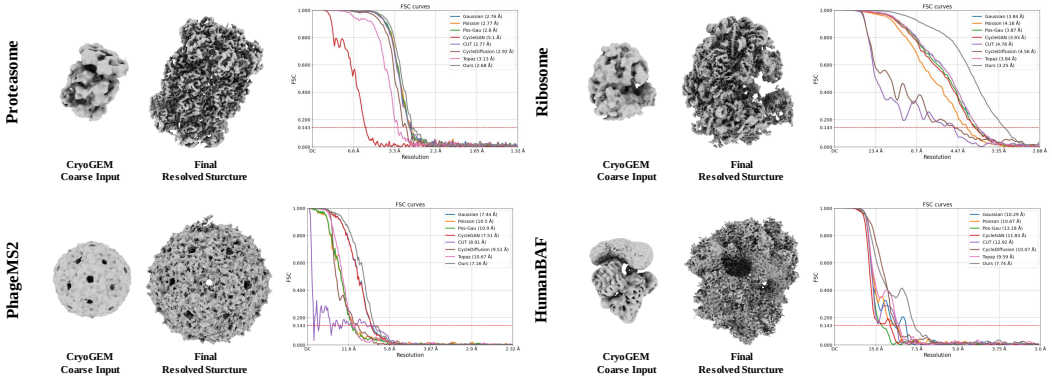

Figure 7: **Evaluation of CryoGEM's picked results on 3D Reconstruction.** We present both the initial coarse inputs from CryoGEM and the final refined structures, with particles selected using CryoGEM's refined particle picking model. Additionally, we provide a quantitative comparison of the reconstruction results across various baselines. The resolution of each reconstruction is evaluated using the Fourier Shell Correlation (FSC) criterion (threshold=0.143 for real datasets).

Table 3: **Quantitative comparison of pose estimation.** Our approach achieves the best performance in Res(px) and Rot.(rad) metrics.

Metric	Res(px)↓						Rot.(rad)↓					
Dataset	Proteasome	Ribosome	Integrin	PhageMS2	HumanBAF	Avg.	Proteasome	Ribosome	Integrin	PhageMS2	HumanBAF	Avg.
Gaussian	2.97	4.59	7.01	5.85	6.82	5.44	0.48	0.50	1.20	0.64	1.49	0.88
Poisson	3.02	4.912	7.01	5.98	8.03	5.79	1.20	0.90	1.19	0.69	1.47	1.10
Poi-Gau	3.02	4.39	9.06	5.90	8.12	6.09	1.05	0.39	1.40	0.61	1.43	0.98
CycleGAN	3.02	4.74	6.24	5.71	6.91	5.32	0.46	0.61	1.55	0.74	1.48	0.97
CUT	2.66	5.40	6.13	6.03	9.58	5.96	0.44	1.15	1.53	0.66	1.45	1.05
CycleDiffusion	3.61	5.79	8.5	5.91	9.33	6.62	0.44	1.42	1.55	0.58	1.53	1.10
CryoFIRE	5.94	16.92	13.87	17.23	6.98	12.18	1.55	0.64	0.93	0.75	1.53	1.08
Ours	**2.59**	**4.27**	**4.88**	**5.54**	**6.56**	**4.29**	**0.41**	**0.32**	**0.88**	**0.43**	**1.42**	**0.69**

approach outperforms the original Topaz and other baseline methods in terms of the quality of visual picking results. As illustrated in Figure 6, the fine-tuned Topaz model is employed to select particles across five datasets. The original Topaz often mistakenly picks false particles from contaminants, particle aggregations, and ice patches. Our fine-tuned Topaz significantly improves accuracy by reducing the incidence of false positives and redundant particle picks. Table 2 presents the quantitative particle picking results, where we utilize the Area Under the Precision-Recall Curve (**AUPRC**) metric that is widely used as particle picking metric [4, 51]. For the calculation of the AUPRC score, we retain the original particle-picking results for comparison against the ground truth labels, without any thresholding. The quantitative comparison shows that the finetuned Topaz using our data achieves the best performance across all datasets. We also show that our final reconstruction resolution (**Res(Å)**) is significantly improved using cryoSPARC [41]'s default reconstruction pipeline. Finally, we present the coarse volumes used for CryoGEM training, the final resolved structures from CryoGEM picked particles, and the FSC curves from different baselines in Figure 7.

5.3 Pose Estimation

In the cryo-EM reconstruction pipeline, the accuracy of pose estimation is crucial for the resolution of the final reconstruction. Existing cryo-EM ab-initio methods, such as CryoFIRE [28], leverage a self-supervised learning framework to predict particle poses and reconstruct 3D volume from images simultaneously. Given synthetic datasets generated from different baselines, we directly supervise its pose estimation module by minimizing the quadratic error between the matrix entries in the prediction and ground truth. The details of direct supervision and the evaluation metrics as well as the visualization of improved structures (Figure 14) can be found in Appendix C.3. We evaluate the pose estimation module on real datasets to obtain estimated poses as ground truths. We use a traditional reconstruction algorithm, filter back-projection (FBP), to obtain final reconstruction results. As demonstrated in Table 3, the pre-trained pose estimation module using our data consistently achieves the best performance in terms of pose accuracy (rotation error in radian) and reconstruction resolution (in pixel).

https://doi.org/10.52202/079017-2019

Dataset	Ribosome				
Task	Photo-realism	Particle Picking		Pose Estimation	
Metric	FID↓	AUPRC↑	Res(Å)↓	Res(px)↓	Rot(rad)↓
w/o Physical Priors	16.63	<u>0.789</u>	3.29	4.88	0.79
Third Layer Noise	15.83	0.764	3.31	4.45	0.47
Fifth Layer Noise	21.30	0.777	<u>3.26</u>	4.46	0.51
Second Channel Noise	<u>6.63</u>	0.012	9.76	4.29	0.34
SNR=0.01	22.16	0.313	3.80	7.11	1.50
SNR=1.0	17.95	0.749	3.40	**4.11**	**0.25**
w/o Gaussian Noise	50.74	0.735	3.31	4.27	0.35
w/o Particle	6.85	0.788	3.41	4.65	0.50
w/o Background	18.46	0.779	3.39	4.33	0.34
w/o Mask Guide	12.61	0.721	3.29	4.44	0.41
Ours	**6.54**	**0.797**	**3.25**	<u>4.27</u>	<u>0.32</u>

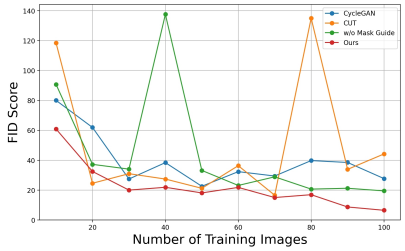

Figure 8: **Quantitative evaluation of ablation study.** Our complete model consistently achieves the **best** or <u>second</u> performance across all metrics, as shown in the left table. It demonstrates stable and efficient training compared to CycleGAN, CUT, and the ablation without Mask Guide (w/o Mask Guide), as illustrated in the right figure.

5.4 Ablation Study

We validate the effectiveness of our several key designs on the Ribosome dataset.

Physical priors. We denote the variation of our model without physical priors as **w/o Physical Priors**. Results imply that the physical priors narrow the domain gap between simulated results and real micrographs.

Introducing randomness. We introduce randomness by different kinds of noises, such as introducing feature-level noise in the encoder's third layer (**Third Layer Noise**) and fifth layer (**Fifth Layer Noise**). Additionally, we treat random noise as a second input channel, which may potentially preserve the content of physical simulation better (**Second Channel Noise**). Also, we demonstrate the impact of noise scaling relative to the signal with **SNR=0.01** and **SNR=1.0**. Additionally, we remove the Gaussian noise (**w/o Gaussian Noise**) to validate the effectiveness of this randomness.

Sampling scheme. We use a particle background mask to guide sampling in the contrastive learning process. For accurate particle position and shape control, the generator must preserve the input image's content. If we sample positives only in the background, the mutual information between particles can only be optimized indirectly, leading to inaccurate particle positions and shapes (**w/o Particle**). Conversely, sampling positives only in particles results in less realistic noise patterns (**w/o Background**). We also present the results without the mask-guided sampling scheme (**w/o Mask Guide**), which includes only the physics-based module and input domain randomness.

Training efficiency ablation. In the right of Figure 8, we validate that our method can generate realistic cryo-EM images with fewer samples compared to baselines in terms of FID. Notably, some high FID outliers indicate that removing the mask-guided sampling scheme or ignoring the physics-based module can decrease training stability.

6 Conclusion

Limitations. As the first trial to achieve authentic cryo-EM image generation with a novel combination of physics-based simulation and unpaired noise translation via contrastive learning, CryoGEM does have room for further improvements. First, our physics-based simulation relies on a coarse result as an input, which may be hard to obtain in some challenging cases, e.g., when the target molecule is very small or dynamic. The latest AlphaFold 3 [1] can help predict the structure directly. Furthermore, we sacrifice the generalization capability for a lightweight training framework. In the future, we will train a generalized version of CryoGEM to unlock more real applications.

Conclusion. We have presented the first generative approach in cryo-EM image synthesis, CryoGEM, that marries physics-based simulation with a contrastive noise generation, to enhance downstream deep models, finally improving cryo-EM reconstruction results. Extensive experiments have shown that CryoGEM's generated dataset with ground-truth annotations can effectively improve particle picking and pose estimation models, eventually improving reconstruction results. We believe that CryoGEM serves as a critical step for high-fidelity cryo-EM data generation using deep generative models, with diverse applications of structure discovery in in cryo-EM.

https://doi.org/10.52202/079017-2019

7 Acknowledgement

This work was supported by ShanghaiTech University's HPC Platform. We would like to thank the Cellverse team for their valuable discussions. We also extend our gratitude to Yutong Liu for contributing to the design of Figure 1.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[2] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8983–8992, 2019.

[3] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017.

[4] Tristan Bepler, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nature methods*, 16:1153 – 1160, 2018.

[5] Melody G. Campbell, Anthony Cormier, Saburo Ito, Robert I Seed, and Stephen L. Nishimura. Cryo-em reveals integrin-mediated tgf-β activation without release from latent tgf-β. *Cell*, 180:490–501.e16, 2020.

[6] Melody G. Campbell, David Veesler, Anchi Cheng, Clinton S. Potter, and Bridget Carragher. 2.8 å resolution reconstruction of the thermoplasma acidophilum 20s proteasome using cryo-electron microscopy. *eLife*, 4, 2015.

[7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[8] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[9] U Demir and G Unal. Patch-based image inpainting with generative adversarial networks. arxiv 2018. *arXiv preprint arXiv:1803.07422*.

[10] Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. Cryotransformer: A transformer model for picking protein particles from cryo-em micrographs. *bioRxiv*, 2023.

[11] Ashwin Dhakal, Rajan Gyawali, Liguo Wang, and Jianlin Cheng. A large expert-curated cryo-em image dataset for machine learning protein particle picking. *Scientific Data*, 10(1):392, 2023.

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[14] Rajan Gyawali, Ashwin Dhakal, Liguo Wang, and Jianlin Cheng. Accurate cryo-em protein particle picking by integrating the foundational ai image segmentation model and specialized u-net. *bioRxiv*, 2023.

[15] Benjamin Himes and Nikolaus Grigorieff. Cryo-TEM simulations of amorphous radiation-sensitive samples using multislice wave propagation. *IUCrJ*, 8(6):943–953, Nov 2021.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

 https://doi.org/10.52202/079017-2019

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Andrii Iudin, Paul K Korir, Sriram Somasundharam, Simone Weyand, Cesare Cattavitello, Neli Fonseca, Osman Salih, Gerard J Kleywegt, and Ardan Patwardhan. Empiar: the electron microscopy public image archive. *Nucleic Acids Research*, 51(D1):D1503–D1511, 2023.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ArXiv*, abs/1603.08155, 2016.

[21] Chanyong Jung, Gihyun Kwon, and Jong-Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18239–18248, 2022.

[22] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *ArXiv*, abs/1907.10830, 2019.

[23] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Roman I Koning, Josue Gomez-Blanco, Inara Akopjana, Javier Vargas, Andris Kazaks, Kaspars Tars, José María Carazo, and Abraham J Koster. Asymmetric cryo-em reconstruction of phage ms2 reveals genome structure in situ. *Nature communications*, 7(1):12524, 2016.

[26] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming Yang. Diverse image-to-image translation via disentangled representations. *ArXiv*, abs/1808.00948, 2018.

[27] Axel Levy, Frédéric Poitevin, Julien N. P. Martel, Youssef S. G. 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. *Computer vision - ECCV ... : ... European Conference on Computer Vision : proceedings. European Conference on Computer Vision*, 13681:540–557, 2022.

[28] Axel Levy, Gordon Wetzstein, Julien N. P. Martel, Frédéric Poitevin, and Ellen D. Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022.

[29] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[30] I Lobato and D Van Dyck. Multem: A new multislice program to perform accurate and fast electron diffraction and imaging simulations using graphics processing units with cuda. *Ultramicroscopy*, 156:9–17, 2015.

[31] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion, 2023.

[32] Jacob Madsen and Toma Susi. abtem: Ab initio transmission electron microscopy image simulation. *Microscopy and Microanalysis*, 26(S2):448–450, 2020.

[33] Nazar Mashtalir, Hiroshi Suzuki, Daniel P Farrell, Akshay Sankar, Jie Luo, Martin Filipovski, Andrew R D'Avino, Roodolph St Pierre, Alfredo M Valencia, Takashi Onikubo, et al. A structural model of the endogenous human baf complex informs disease mechanisms. *Cell*, 183(3):802–817, 2020.

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[35] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[36] Mateusz Olek, Kevin Cowtan, Donovan Webb, Yuriy Chaban, and Peijun Zhang. Icebreaker: Software for high-resolution single-particle cryo-em with non-uniform ice. *Structure*, 30(4):522–531, 2022.

[37] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[38] Micha Pfeiffer, Isabel Funke, Maria Ruxandra Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Ross, Matthew J. Clarkson, Kurinchi S. Gurusamy, Brian R. Davidson, Lena Maier-Hein, Carina Riediger, Thilo Welsch, Jürgen Weitz, and Stefanie Speidel. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

[39] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting, 2023.

[40] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10609–10619, 2021.

[41] 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.

[42] Alexis Rohou and Nikolaus Grigorieff. Ctffind4: Fast and accurate defocus estimation from electron micrographs. *Journal of structural biology*, 192(2):216–221, 2015.

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[45] Hans Rullgård, L-G Öfverstedt, Sergey Masich, Bertil Daneholt, and Ozan Öktem. Simulation of transmission electron microscope images of biological specimens. *Journal of microscopy*, 243(3):234–256, 2011.

[46] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.

[47] Shayan Shekarforoush, David B Lindell, David J Fleet, and Marcus A Brubaker. Residual multiplicative filter networks for multiscale reconstruction. *arXiv preprint arXiv:2206.00746*, 2022.

[48] Dimitry Tegunov and Patrick Cramer. Real-time cryo–em data pre-processing with warp. *Nature methods*, 16:1146 – 1152, 2018.

[49] Brian Teixeira, Vivek Singh, Terrence Chen, Kai Ma, Birgi Tamersoy, Yifan Wu, Elena Balashova, and Dorin Comaniciu. Generating synthetic x-ray images of a person from the surface geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[50] Miloš Vulović, Raimond BG Ravelli, Lucas J van Vliet, Abraham J Koster, Ivan Lazić, Uwe Lücken, Hans Rullgård, Ozan Öktem, and Bernd Rieger. Image formation modeling in cryo-electron microscopy. *Journal of structural biology*, 183(1):19–32, 2013.

[51] Thorsten Wagner, Felipe Merino, Markus Stabrin, Toshio Moriya, Claudia Antoni, Amir Apelbaum, Philine Hagel, Oleg Sitsel, Tobias Raisch, Daniel Prumbaum, Dennis Quentin, Daniel Roderer, Sebastian Tacke, Birte Siebolds, Evelyn Schubert, Tanvir R. Shaikh, Pascal Lill, Christos Gatsogiannis, and Stefan Raunser. Sphire-cryolo is a fast and accurate fully automated particle picker for cryo-em. *Communications Biology*, 2, 2019.

[52] Weilun Wang, Wen gang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14000–14009, 2021.

[53] David B Williams, C Barry Carter, David B Williams, and C Barry Carter. *The transmission electron microscope*. Springer, 1996.

[54] Wilson W. Wong, Xiao chen Bai, Alan Brown, Israel S. Fernández, Eric Hanssen, Melanie M Condron, Yan hong Tan, Jake Baum, and Sjors H. W. Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *eLife*, 3, 2014.

 https://doi.org/10.52202/079017-2019

[55] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022.

[56] Chentianye Xu, Xueying Zhan, and Min Xu. Cryomae: Few-shot cryo-em particle picking with masked autoencoders. *arXiv preprint arXiv:2404.10178*, 2024.

[57] Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. *ArXiv*, abs/2310.13165, 2023.

[58] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Jerry L. Prince, and Zongben Xu. Unsupervised mr-to-ct synthesis using structure-constrained cyclegan. *IEEE Transactions on Medical Imaging*, 39(12):4249–4261, 2020.

[59] Hangping Yao, Yutong Song, Yong Chen, Nanping Wu, Jialu Xu, Chujie Sun, Jiaxing Zhang, Tianhao Weng, Zheyuan Zhang, Zhigang Wu, et al. Molecular architecture of the sars-cov-2 virus. *Cell*, 183(3):730–738, 2020.

[60] Lin Yao, Ruihan Xu, Zhifeng Gao, Guolin Ke, and Yuhang Wang. Boosted ab initio cryo-em 3d reconstruction with ace-em. *ArXiv*, abs/2302.06091, 2023.

[61] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[63] Yue Zhang, R Tammaro, Peter J Peters, and RBG Ravelli. Could egg white lysozyme be solved by single particle cryo-em? *Journal of Chemical Information and Modeling*, 60(5):2605–2613, 2020.

[64] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022.

[65] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16402–16412, 2021.

[66] 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, 2021.

[67] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021.

[68] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[70] Jasenko Zivanov, Takanori Nakane, Björn O Forsberg, Dari Kimanius, Wim J. H. Hagen, Erik Lindahl, and Sjors H. W. Scheres. Relion-3: new tools for automated high-resolution cryo-em structure determination. *bioRxiv*, 2018.

—Supplementary Material— CryoGEM: Physics-Informed Generative Cryo-Electron Microscopy

A Additional results.

We include supplementary results in Figure 9. These serve as an extension to Figure 4 from our main paper. Here, we demonstrate CryoGEM's ability to generate synthetic micrographs with control over the particle mask, defocus value, and ice gradient. Additionally, we explore CryoGEM's zero-shot capability, showcasing its proficiency in generating authentic noise patterns on unseen datasets. These results show CryoGEM's potential in creating an extensive synthetic dataset in a wide range of particle and noise pattern combinations.

B Dataset Details

We evaluate CryoGEM across five diverse cryo-EM datasets. We source the **Proteasome**, **Ribosome**, and **Integrin** datasets from the Electron Microscopy Public Image Archive (EMPIAR) [19], a global public resource offering raw cryo-EM micrographs and selected particle stacks for constructing high-resolution 3D molecular maps. The **PhageMS2** and **HumanBAF** datasets were downloaded from the cryoPPP dataset [11] with provided manual particle picking annotations.

Proteasome Thermoplasma acidophilum 20S proteasome [6] (EMPAIR-10025) contains 196 real micrographs and 49,954 manually filtered particles to achieve a high-resolution result at 2.8 Å. The T20S proteasome forms a 700kDa complex that contains 14 α-helices and 14 β-sheets organized by D7 symmetry. The captured micrographs comprise particles with high symmetry, random rotational orientations, and mutual occlusions.

Ribosome Human 80S ribosome [54] (EMPIAR-10028) contains 1,081 real micrographs and 105,417 manually filtered particles to achieve a high-resolution result at 3.6 Å. It is a large assembly, incorporating numerous RNA chains and proteins.

Integrin Asymmetric αVβ8 integrin [5] (EMPAIR-10345) contains 1,644 real micrographs and 84,266 manually filtered particles to achieve a high-resolution result at 3.3 Å. The αVβ8 integrin is a complex of the human αVβ8 ectodomain with porcine L-TGF-β1, showing significant movement in its leg in the captured micrographs.

PhageMS2 Bacteriophage MS2 [25] (EMPAIR-10075) contains 300 real micrographs and 12682 manually filtered particles [11] to achieve a high-resolution result at 8.7 Å. PhageMS2 consists of 178 copies of the coat protein, a single copy of the A-protein, and the RNA genome. Note that its size is significantly larger than others, making it a challenge for the particle picking model.

HumanBAF Endogenous human BAF complex [33] (EMPAIR-10590) contains 300 real micrographs and 62493 manually filtered particles [11] to achieve a high-resolution result at 7.8 Å.

In Figure 10, we demonstrate datasets in a wide variety of noise patterns, particle scales, and types. These characteristics present significant challenges to the accuracy of particle picking and pose estimation, which are critical for automated high-resolution cryo-EM reconstruction. Notably, CryoGEM can generate similar visual effects such as the ring of defocus (Ribosome) and crystalline ice(Integrin). We analyze the distribution of particle counts and defocus values across micrographs for each dataset, calculating their mean and standard deviation. Based on these analyses, we generate two Gaussian distributions to simulate particle count and defocus value accurately in CryoGEM, as illustrated in Figures 11 and 12. During the training phase, we select 100 micrographs from each real dataset as training samples for CryoGEM, alongside other generative models like CUT [37] and CycleGAN [68]. To ensure a fair comparison, we exclude these micrographs from our evaluation.

https://doi.org/10.52202/079017-2019

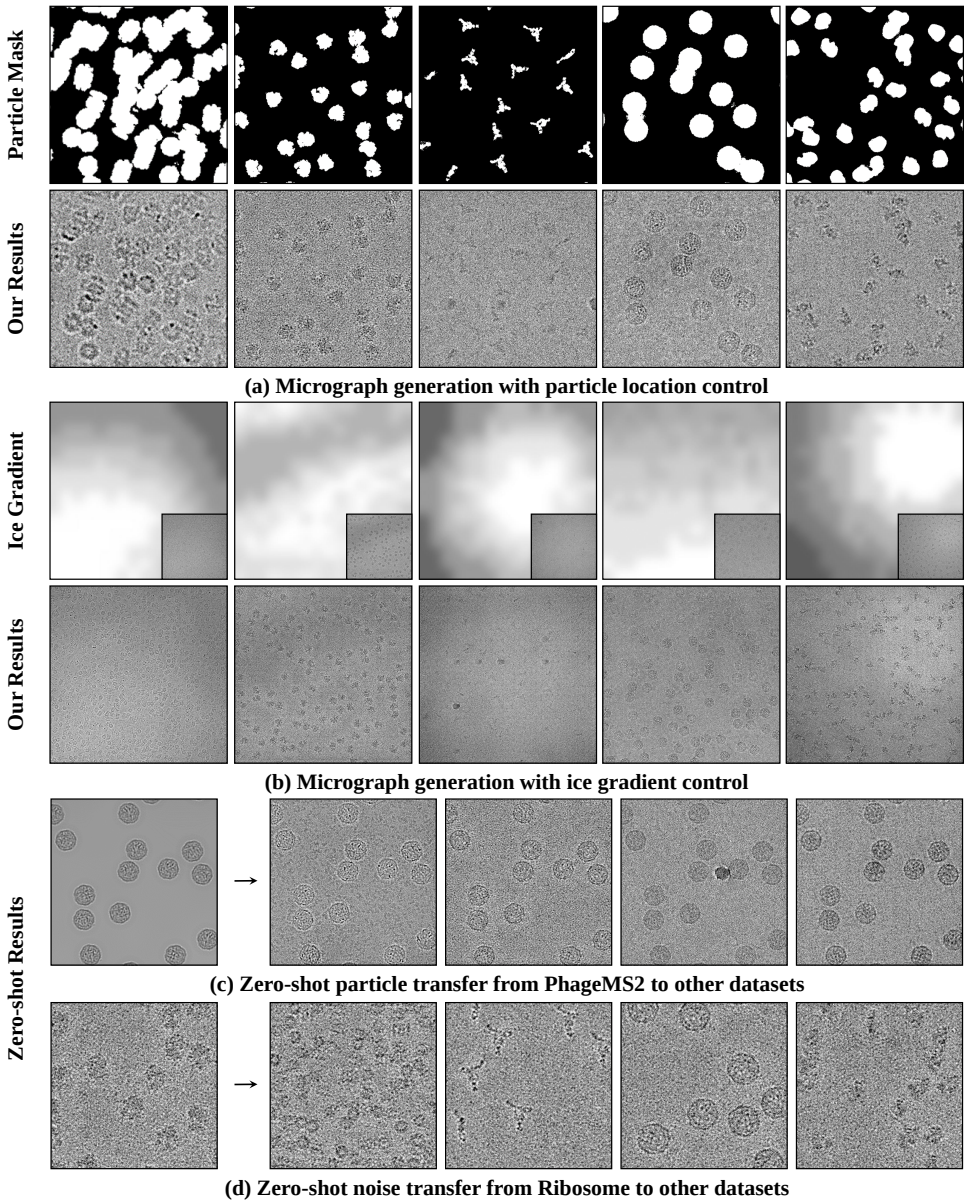

Figure 9: **(a) Location-controlled generation.** CryoGEM uses a particle mask to guide the sampling process in contrastive noise modeling. This method allows for precise control over the placement of particles in synthetic micrographs. **(b) Ice-gradient generation.** CryoGEM calculates the ice gradient in real micrographs, as indicated in the lower right corner of each ice gradient image. This calculated ice gradient is then incorporated into the final images. **(c, d) Zero-shot transfer between particle and noise.** We show that CryoGEM can generate convincing results on previously unseen datasets without the need for additional training. This demonstrates CryoGEM's ability for zero-shot transfer between different types of particles and noise.

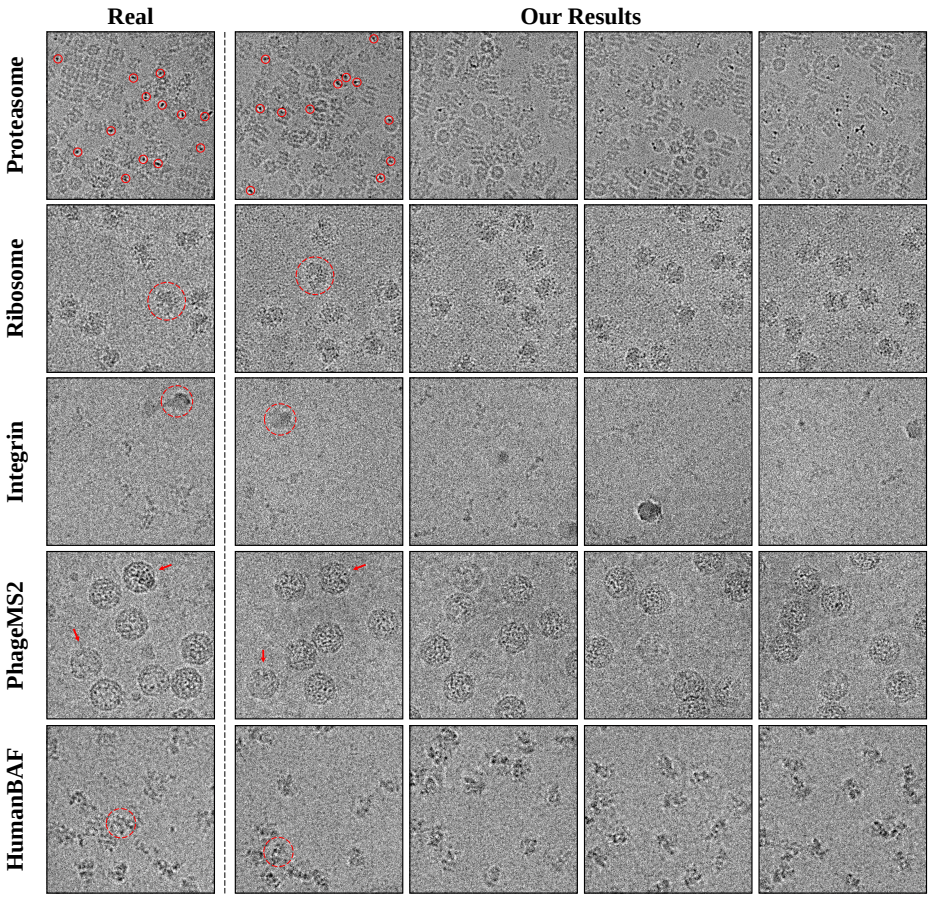

Figure 10: **Additional visual comparison with real datasets.** Our evaluation of five complex real datasets demonstrates that CryoGEM can accurately reproduce noise patterns, preserve structural details, and replicate specific anomalies like crystalline ice in Integrin. For each dataset, we highlight the visual characteristics adeptly captured by CryoGEM with red circles or arrows.

Dataset preprocessing. To improve the quality of the training datasets, we meticulously remove "dirty" micrographs, which are defined as micrographs captured outside the intended target area, those exhibiting substantial jittering artifacts, and those only containing background ice. To improve the training efficiency, we downsample the images to 1024×1024 resolution. We enhance the contrast of the electron microscope images by adjusting the mean and standard deviation of the image intensity values to 150 and 40 (with a maximum value of 255), respectively. All these pre-processing steps are crucial for a stable and high-performance training process.

Initial 3D result. We reconstruct a low-resolution 3D volume and simultaneously estimate the poses of input particle images using an *ab*-initio reconstruction of cryoSPARC [41] with its default settings. To generate an initial neural volume with conformational changes, we employ an advanced neural method, cryoDRGN [66] for continuous heterogeneous (dynamic) reconstruction, given the particle stack with estimated poses. We train the cryoDRGN 50 epochs by particles at a resolution aligned with the low-resolution 3D volume. Note that for homogeneous (static) datasets in our experiments, we directly utilize the 3D initial volume from cryoSPARC, as these datasets exhibit only negligible motions.

Ice gradient estimation. In Figure 1 of the main paper, we show that ice gradients result in intensity variations across micrographs, leading to signal-to-noise ratio (SNR) disparities in local areas. Images from thicker specimen regions often exhibit a lower SNR than those from thinner regions. To mimic this physical effect, we employ IceBreaker [36], an established method that addresses the challenge

https://doi.org/10.52202/079017-2019

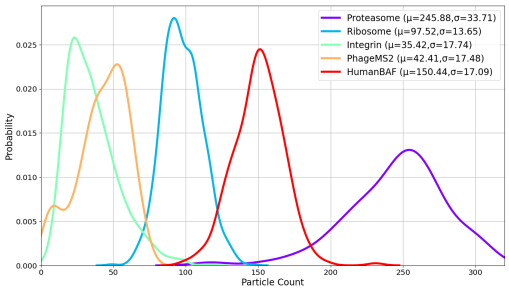

Figure 11: Probability curve of particle count in cryo-EM datasets.

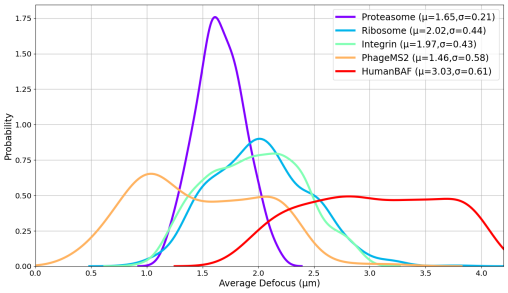

Figure 12: Probability curve of defocus value in cryo-EM datasets.

by treating the estimation of uneven ice thickness as a clustering problem. We have developed a GPU-accelerated version of IceBreaker for a more efficient estimation of average brightness per class. By dividing the outcome by its maximum value, we achieve a normalized weight map. Then, to replicate the gradual variations in ice thickness, we apply a broad Gaussian kernel.

Optical aberration estimation. In cryo-EM, a high-energy electron beam interacts with the specimen, goes through a sophisticated optical imaging process, and is then captured by the detector as projection images. This optical modulation can be described by the Contrast Transfer Function (CTF), with the objective defocus being a particularly crucial parameter. Experimentally, biologists can modify the defocus to enhance contrast or capture higher-frequency signals. We directly incorporate CTF by applying it in our physical simulation process. Similar to CTFFIND4 [42], CTF can be expressed as:

$$\begin{aligned}&\mathrm{CTF}(w,\lambda,\mathbf{g},\Delta f,C_s,\Delta\varphi)=\\&-\sqrt{1-w^2}\sin[\chi(\lambda,|\mathbf{g}|,\Delta f,C_s,\Delta\varphi)]\\&\quad -w\cos[\chi(\lambda,|\mathbf{g}|,\Delta f,C_s,\Delta\varphi)],\end{aligned}\tag{8}$$

where

$$\chi\left(\lambda,|\mathbf{g}|,\Delta f,C_s,\Delta\varphi\right)=\pi\lambda|\mathbf{g}|^2\left(\Delta f-\frac{1}{2}\lambda^2|\mathbf{g}|^2C_s\right)+\Delta\varphi.\tag{9}$$

w is a relative phase contrast factor, χ is the frequency-dependent phase shift function with the inputs including electron wavelength λ, the spatial frequency vector $\mathbf{g}$, the objective defocus Δf, the spherical aberration C_s and the phase shift $\Delta\varphi$. $w,\lambda,C_s,\Delta\varphi$ are the cryo-EM hardware parameters.

We utilize CTFFIND4 to estimate the range of the objective defocus, Δf. Subsequently, ResidualMFN [47] is used to generate a specific 2D CTF image for any given defocus value. As depicted in Figure 12, we estimate the defocus value distribution in real cryo-EM datasets. For the physical simulation phase, defocus values are randomly chosen from simplified Gaussian distributions, based on previously calculated means and standard deviations.

C Evaluation Details

C.1 Baseline Details

Traditional noise simulations. These baselines represent traditional simulation methods. The **Poisson** noise emulates the high-level noise caused by the stochastic nature of the detector under limited dosage conditions. Zero-mean **Gaussian** noise, as suggested by Vulovic et al. [50], is used to simulate a complex noise distribution, including readout noise, dark current noise, shot noise, and structural noise. We also combine these to create a **Poi-Gau** noise generation baseline. In all our experiments, we set the SNR to 0.1 for images generated by these baselines.

CycleGAN [69, 18]. We use UNet [44] as the generator and PatchGAN [9] as the discriminator. The complete CycleGAN objective comprises five components: synthetic-to-real GAN loss, real-to-synthetic GAN loss, two-cycle consistency losses, and identity mapping loss. To maintain training stability, we assign their relative importance as 1, 1, 10, 10, and 0.5, respectively.

CUT [37]. We employ UNet [44] as the generator and PatchGAN [9] as the discriminator. The complete objective of CUT includes three components: synthetic-to-real GAN loss, NCE loss on synthetic and generated images, and NCE loss on real and real-identity images. We follow the original paper, setting their relative importance to 1, 1, and 1.

CycleDiffusion [55, 35]. CycleDiffusion necessitates the use of diffusion models for both the source and target domains. Consequently, we train two Denoising Diffusion Probabilistic Models (DDPMs) for each dataset: one for the synthetic image domain and one for the real image domain. Each model is trained on 100 images at a resolution of 512×512, with a batch size of 2, over 10,000 steps. For CycleDiffusion inference, we use the default settings. The associated code and models will be made available.

C.2 Particle Picking Details

Topaz [4] is a cryo-EM particle picking model that outputs the probability for each pixel belonging to a particle. With a threshold, it can filter out detection results. The training data for Topaz consists of complete micrographs and the central coordinates of particles within them. In practice, we select pre-trained resnet8_u32 as baseline **Topaz** model without fine-tuning. When we fine-tune Topaz by synthetic datasets from different baselines, we use 10 annotated micrographs to fine-tune the pre-trained Topaz for 20 epochs and then test it on the real evaluation datasets. Note that Topaz has already been pre-trained on the Proteasome and Ribosome datasets, which have been utilized in our experiments. Therefore, for these datasets, we infer the picking results at the pre-trained resolution of 512×512. For other datasets, we evaluate it at a resolution of 1024×1024, aligning with the fine-tuned resolution.

Evaluation in terms of metric AUPRC. We retain the original particle picking results for comparison against the ground truth labels, without any thresholding. For a fair comparison of particle picking results, we prioritize our particle picks based on their confidence scores. Specifically, we select the highest-ranking 50,000 picks from each method for further filtering. For PhageMS2, however, due to the limited number of particles present in the micrographs, we only select the top 10,000 picks. By applying different thresholds divided into n intervals by $\{\tau_i\}_{i=1}^{n-1}$, we can filter out varying prediction results and calculate corresponding precision and recall metrics against the true labels. Finally, we can compute the area under the precision-recall curve (AUPRC) as expressed by

$$\sum_{k=1}^{n} \mathrm{Pr}(k)(\mathrm{Re}(k) - \mathrm{Re}(k-1)), \tag{10}$$

where the precision $\mathrm{Pr}(k)$ represents the proportion of true positives in the prediction results with a probability greater than or equal to τ_k, the recall $\mathrm{Re}(k)$ means the proportion of true positives in the prediction results with a probability greater than or equal to τ_k out of the total number of true positives.

Evaluation in terms of metric Res(Å). We further assess the particle picking results by evaluating the final reconstruction resolution achieved using cryoSPARC [41]. The filtered picked results are fed into cryoSPARC for the subsequent reconstruction process. For static proteins such as Proteasome, Ribosome, PhageMS2, and HumanBAF, the reconstruction process involves a 1-class *ab initio* reconstruction followed by homogeneous refinement. For dynamic proteins like Integrin, the reconstruction process includes a 5-class *ab initio* reconstruction, heterogeneous refinement, and homogeneous refinement for each class. We select the best final resolution value as the reconstruction resolution. Importantly, since no 2D classification or particle filtering operations are performed before reconstruction, the resolution of the resolved structures directly reflects the precision of particle picking and the quality of the particles. This method ensures that no human bias is introduced.

 https://doi.org/10.52202/079017-2019

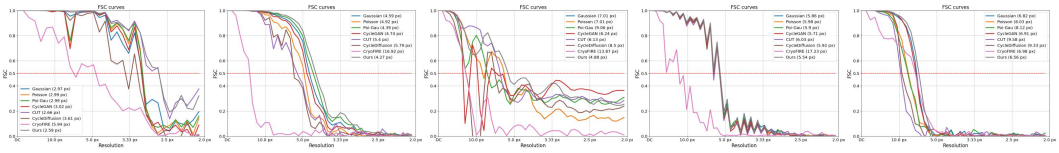

Figure 13: **Quantitative comparison of pose estimation using FSC curves.** Our approach demonstrates superior performance in the Res(px) metric. Notably, we present CryoFIRE, the sole neural-based reconstruction method that operates without ground truth particle poses.

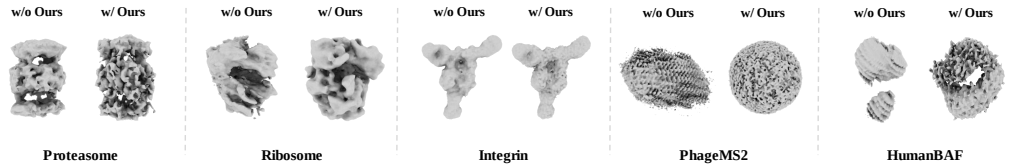

Figure 14: **Qualitative evaluation of performance improvement on pre-trained CryoFIRE.** The figure illustrates the detailed improvements in ab initio volumes. (Left: original CryoFIRE, Right: pre-trained CryoFIRE using CryoGEM's data).

C.3 Pose Estimation Details

To evaluate whether CryoGEM can help with pose estimation tasks, we utilize the state-of-the-art *ab*-initio reconstruction algorithm, cryoFIRE [28], from its official Github page. We employ synthetic datasets that provide a direct supervision mechanism through pairs of synthetic particle images and their corresponding true poses. In all experiments, we generate 100,000 particle images per baseline, each with a ground-truth orientation $\{R_i\}_{i=1}^N$ and translation $P = \{(R_i, T_i)\}_{i=1}^N$. To train the pose prediction module, we adopt the loss function for direct pose supervision by the loss introduced from ACE-EM [60].

$$\mathcal{L}_{\text{pose}} = \frac{1}{B}\sum_{i=1}^{B}\left[\frac{1}{9}\|R_i^{\text{gt}} - R_i^{\text{pred}}\|_F^2 + \frac{1}{2}\|T_i^{\text{gt}} - T_i^{\text{pred}}\|_1\right], \tag{11}$$

where B represents the training batch size, while R_i^{gt}, R_i^{pred}, T_i^{gt}, and T_i^{pred} respectively represent the ground truth and network-predicted rotation and translation. We set B=256 in our experiments. For various baselines, we train the pose estimation module for 200 epochs using only particles and their ground-truth poses. We then test the fine-tuned cryoFIRE's pose estimation module on the real datasets. Notably, we include the original **cryoFIRE** as an additional baseline in the pose estimation task. To determine the final reconstruction resolution, we randomly split the real data into two equal parts for separate reconstructions. The final reconstruction resolution is calculated by thresholding the Fourier Shell Curve (FSC) of the two reconstructions to 0.5.

Evaluation in terms of metric Res(px). We evaluate the pose estimation module on real datasets by using the estimated poses as training labels. For this purpose, we reconstruct two independent volumes from equally split particle stacks using the filter back-projection (FBP) method. We choose this traditional reconstruction algorithm over contemporary neural methods to directly assess pose accuracy without the bias of reconstruction algorithms. Due to the characteristics of FBP, some reconstructions' Fourier Shell Curves (FSC) do not meet the standard threshold of FSC=0.143. Therefore, we use a correlation coefficient of 0.5 for the Fourier shells as the metric Res(px) for pose estimation reconstruction resolution.

We showcase the performance of the original CryoFIRE, the only neural-based reconstruction method in our comparison that does not rely on ground truth particle poses. Figure 13 displays the FSC curves for structures resolved during pose estimation. Additionally, we present the enhanced performance of pre-trained CryoFIRE when supplemented with CryoGEM's pose-labeled particle images, illustrating these improvements in Figure 14. The formula for the Fourier Shell Correlation (FSC) is provided

below:

$$FSC(r) = \frac{\sum_{r_i \in r} F_1(r_i) \cdot F_2(r_i)^*}{\sqrt{\sum_{r_i \in r} \|F_1(r_i)\|^2 \cdot \sum_{r_i \in r} \|F_2(r_i)\|^2}} \tag{12}$$

where F_1, F_2 are the Fourier transforms of the FBP reconstructed volumes on equally split particle stacks, respectively. r represents all three-dimensional frequency components shown in a one-dimensional form.

Evaluation in terms of metric Rot(rad). In addition to evaluating the resolution of resolved structures in pose estimation, we also report posing errors for all real images. For each dataset, we obtain the estimated ground truth pose from the cryoSPARC reconstruction pipeline. Due to a coordinate system misalignment issue, a rigid 6D-body alignment is applied to the poses predicted by the pose estimation module trained on generated data. This alignment ensures that the resolved structure is in the same coordinate system as the ground truth coarse volume. We then compute the mean angular error between the aligned estimated rotations $\hat{R}_i^{\text{pred}}$ and the ground truth rotations R_i^{gt}, in radians.

$$\text{Rot(rad)} = \frac{180}{\pi n_{\text{rots}}} \sum_{i=1}^{n_{\text{rots}}} \arccos\left(\frac{R_i^{\text{gt}} \cdot v}{\|R_i^{\text{gt}} \cdot v\|} \cdot \frac{\hat{R}_i^{\text{pred}} \cdot v}{\|\hat{R}_i^{\text{pred}} \cdot v\|}\right) \tag{13}$$

where v represents a unit vector $(0, 0, 1)$.

D Societal Impacts

The CryoGEM technique, a novel physics-informed generative cryo-electron microscopy (cryo-EM) tool, has significant societal implications. we outline the potential positive and negative impacts, acknowledge uncertainties, and emphasize the broader implications of this technology.

D.1 Positive Impacts

Improvement in biomedical fields: By improving particle picking and pose estimation, CryoGEM enhances 3D reconstructions of proteins, potentially accelerating the discovery and understanding of biomolecular structures, crucial for drug development and disease treatment.

Reduction in labor-intensive tasks: CryoGEM automates the generation of annotated datasets, reducing the burden on human experts.

D.2 Negative Impacts

Ethical and Misuse Concerns: The ability to generate realistic synthetic data raises ethical concerns about misuse, such as manipulating research outcomes or creating misleading scientific evidence. Establishing guidelines for ethical use is crucial. There are several ways to prevent it from misusing including 1) developing and enforcing comprehensive ethical guidelines, 2) ensuring transparency in data generation, and forcing the users to claim the usage of CryoGEM in their research projects.

D.3 Conclusion

CryoGEM offers transformative opportunities for cryo-EM and related fields, with significant benefits for biomedical research. However, careful consideration of ethical implications, potential misuse, and long-term societal impacts is essential. By fostering a balanced and reflective approach, we can maximize the positive outcomes of CryoGEM and mitigate potential risks, steering the technology in a beneficial direction for society.

https://doi.org/10.52202/079017-2019

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We believe that the main contributions of our method are appropriately described in our abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We clearly discuss the limitations of our method in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: There is no theory assumptions or proofs in this paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have clearly described the implementation details including network architecture, loss functions and training details in Section 5 and Appendix B, C.

   Guidelines:

https://doi.org/10.52202/079017-2019

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not provide open-source code and dataset in this submission. However, we will release our code and dataset upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe the implementation details in Section 5 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not provide error bars for our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computing resources in Section 5 (a single RTX 3090 GPU).

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-2019

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research are all conducted in line with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impact of our method in Section 1 and Appendix D.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

Justification: We do not describle safeguards in our paper, but we will consider to prevent our model and data from misuse when we release our code and dataset.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All of code and datasets we built upon are in public with proper licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not provide new assets in this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

https://doi.org/10.52202/079017-2019

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not have crowdsourcing experiments in our paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not have crowdsourcing experiments in our paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.