Supplementary Information

EM Generalist: A physics-driven diffusion foundation model for electron microscopy

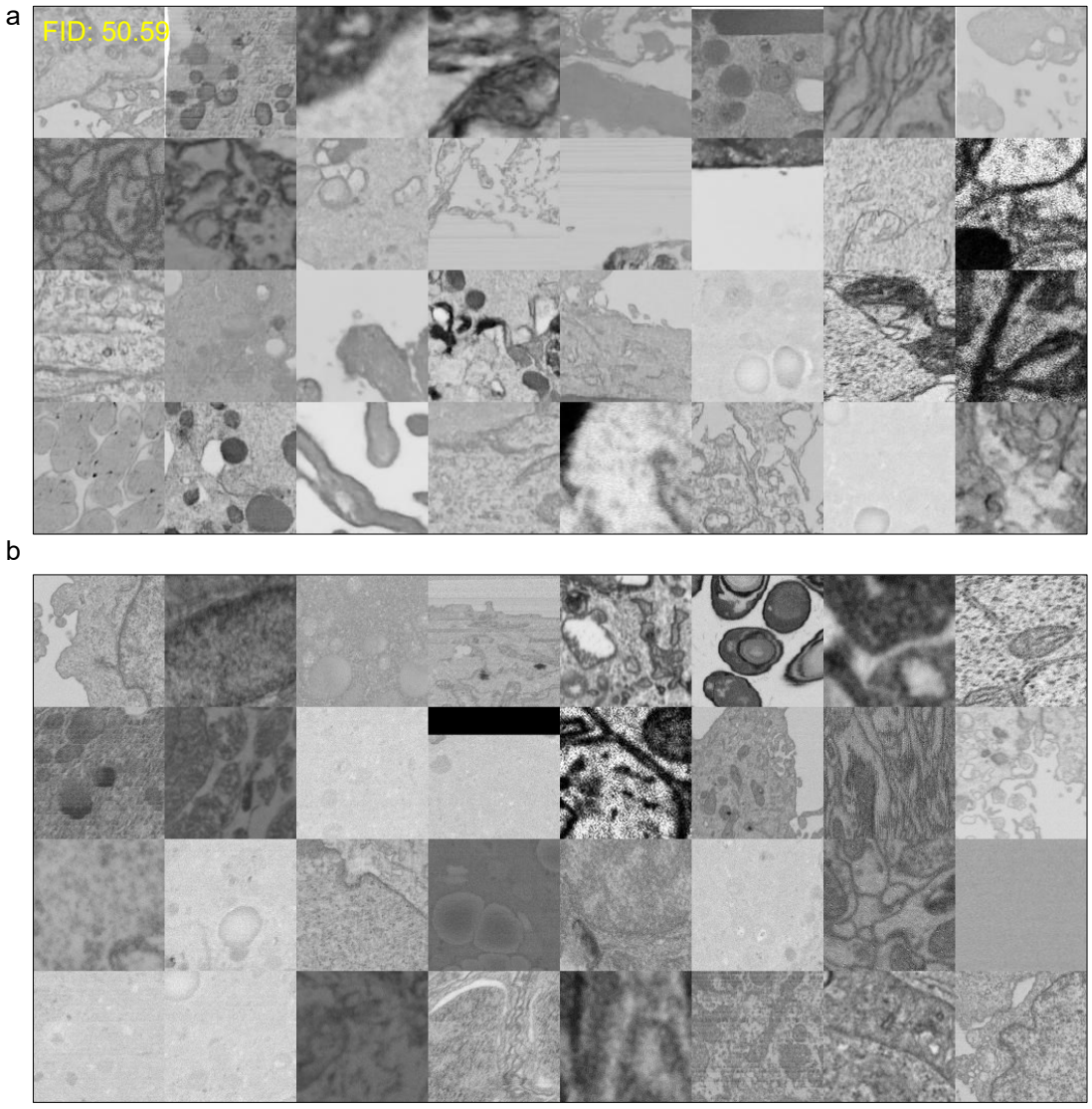
Enze Ye[1,2,†], Chixiang Lu[3,†], Zixuan Jiang[4,†], Weimin Bai[1,2,5], Shaochi Ren[1,2], Chenyu Wang[1,2], Jinyang Zhang[1,2,6], Ruohua Shi[2,7,8], Lei Ma[2,7,8], Kun Song[1,2,*], Xiaojuan Qi[9,*], Haibo Jiang[3,*], He Sun[1,2,5,*]

1 College of Future Technology, Peking University, Beijing, China
2 National Biomedical Imaging Center, Peking University, Beijing, China
3 Department of Chemistry, The University of Hong Kong, Hong Kong, China
4 College of Engineering, Peking University, Beijing, 100871, China
5 Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
6 College of Informatics, Huazhong Agricultural University, Wuhan, China
7 National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China
8 National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, Beijing, China
9 Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China

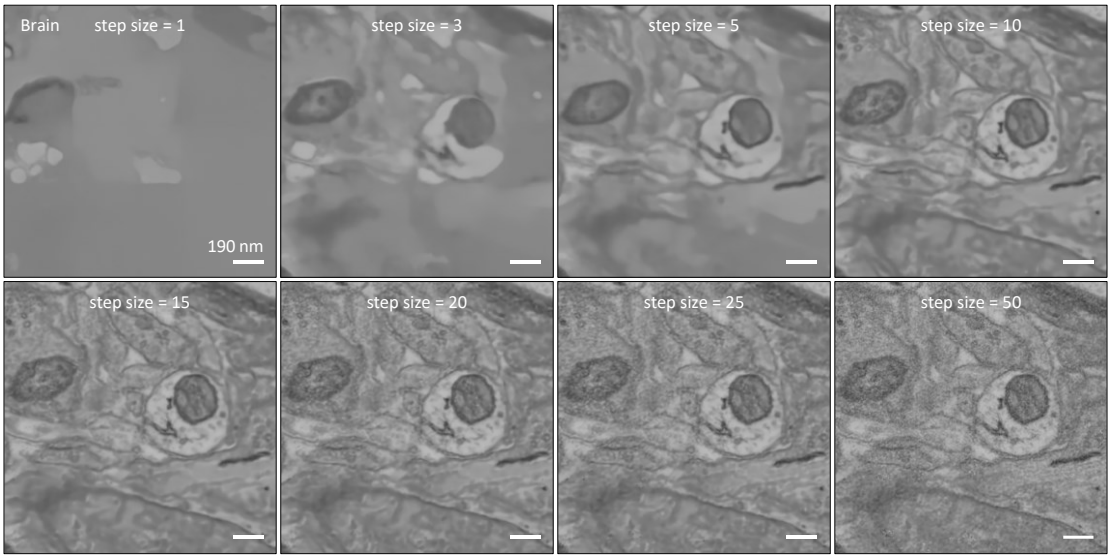
† Contributed equally to this work: Enze Ye, Chixiang Lu, and Zixuan Jiang.

* Corresponding authors: K. Song (kun.song@pku.edu.cn), X. Qi (xjqi@eee.hku.hk), H. Jiang (hbjiang@hku.hk), and H. Sun (hesun@pku.edu.cn)

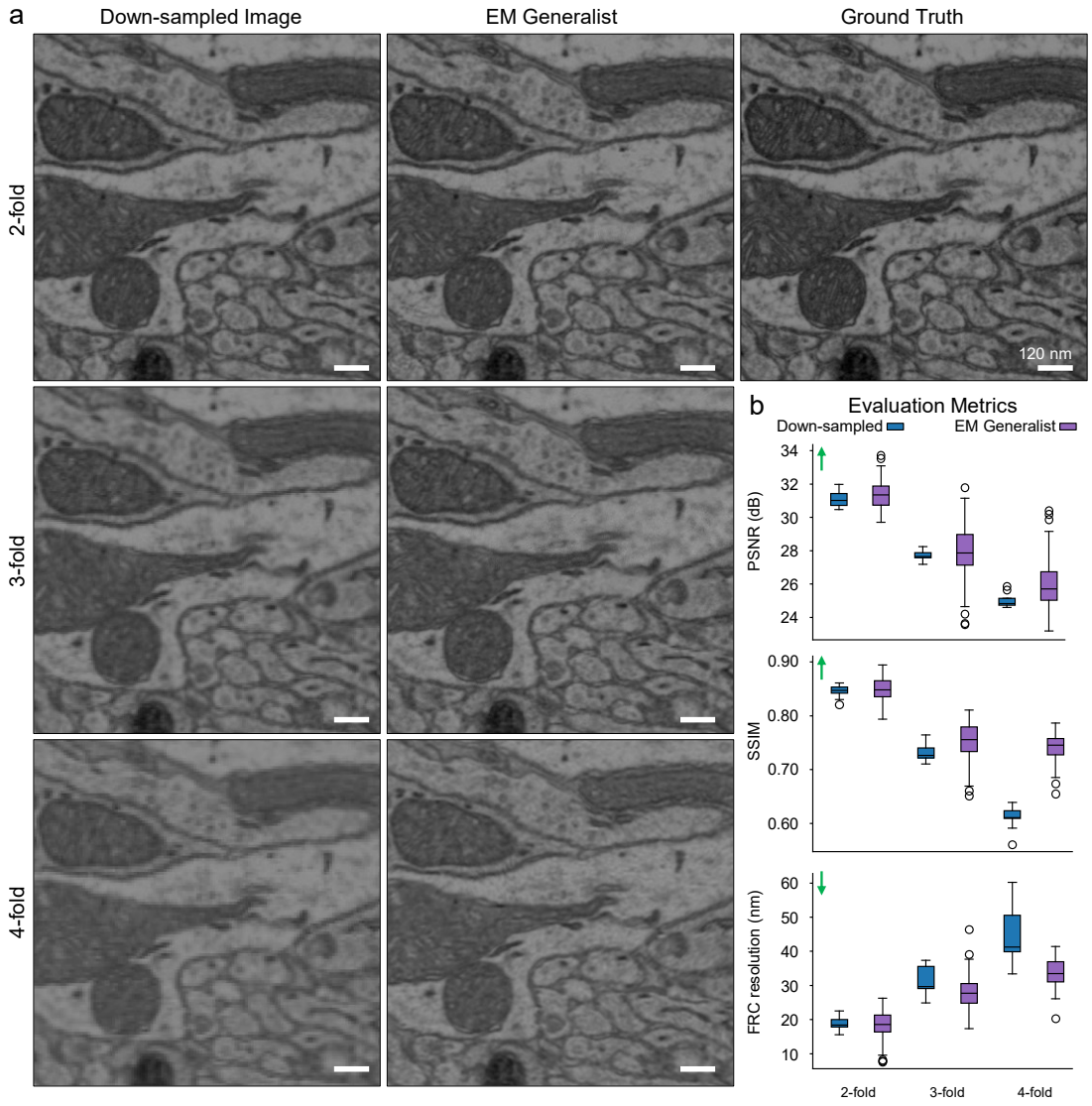
Editorial corresponding author: H. Sun (hesun@pku.edu.cn)



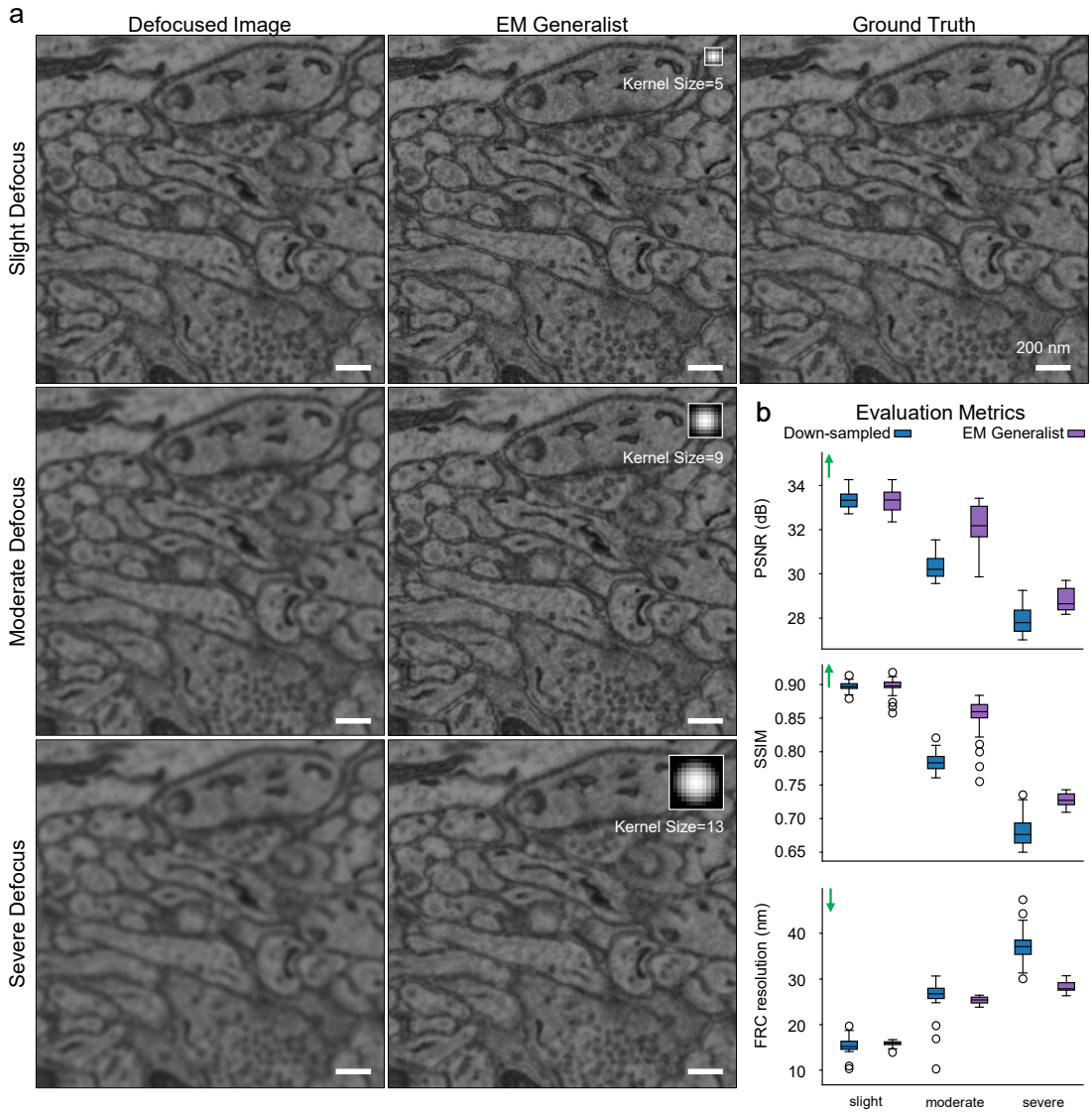
SI Fig. 1. (a) Uncurated EM images generated by the diffusion foundation model, alongside the Fréchet Inception Distance (FID) value calculated between 2500 training images and 2500 generated images. **(b)** Randomly selected EM images from training dataset (1.7M patches).



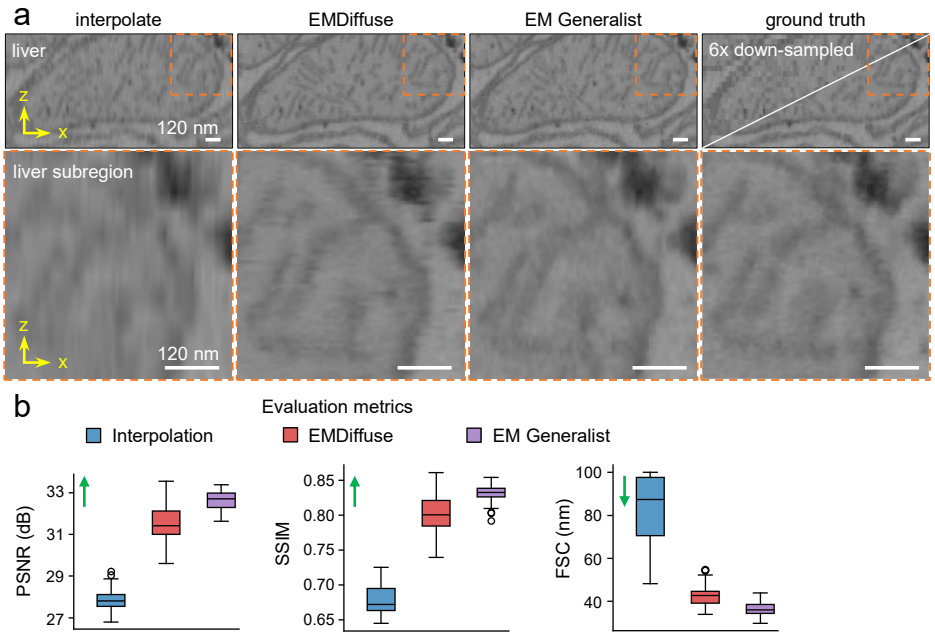
SI Fig. 2. Visualization of the effect of the parameter *step size* (gradient scaling coefficient in the likelihood term) on denoising reconstruction results. The results show that a small step size (e.g., 1 or 3) leads to overly smoothed and blurred images, while a large step size (e.g., 50) introduces noise and artifacts. Optimal step size values (e.g., 15) balance the image prior from the diffusion model and the data consistency, yielding high-quality results.



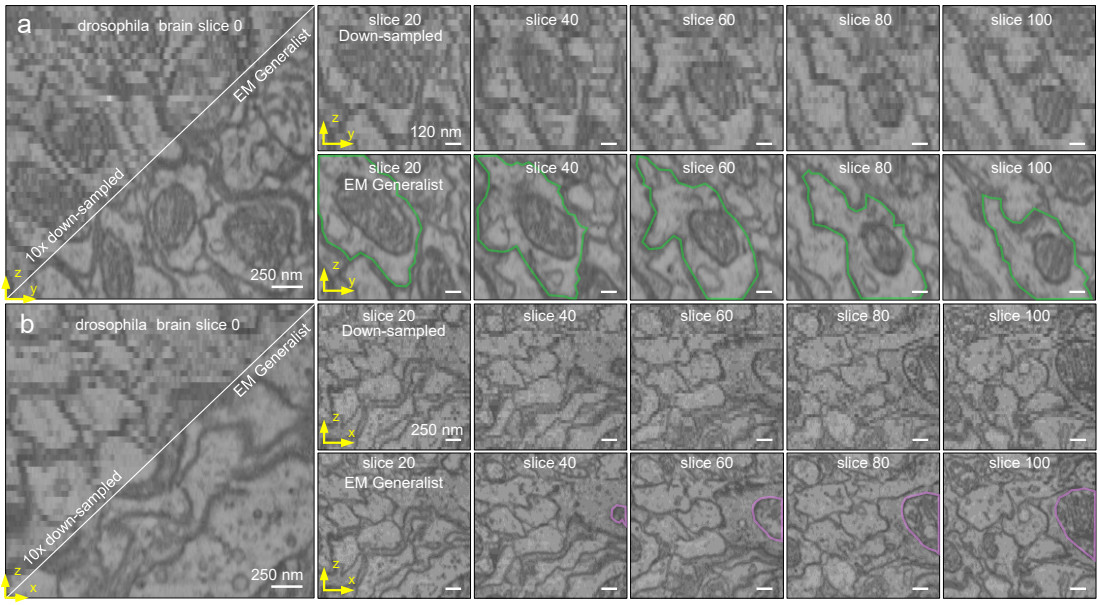
SI Fig. 3. (a) Super-resolution results of EM Generalist applied to mouse brain images down-sampled by factors of 2-, 3-, and 4-fold, alongside comparisons with the ground truth. **(b)** Quantitative evaluation of the reconstruction results. As the downsampling factor increases and effective information is progressively lost, the reconstruction quality gradually declines. Nevertheless, even under 4-fold downsampling, EM Generalist achieves a FRC resolution below 35 nm (pixel size = 4.0 nm), with a PSNR of 26 and a SSIM of 0.75, demonstrating its ability to significantly recover the original image details.



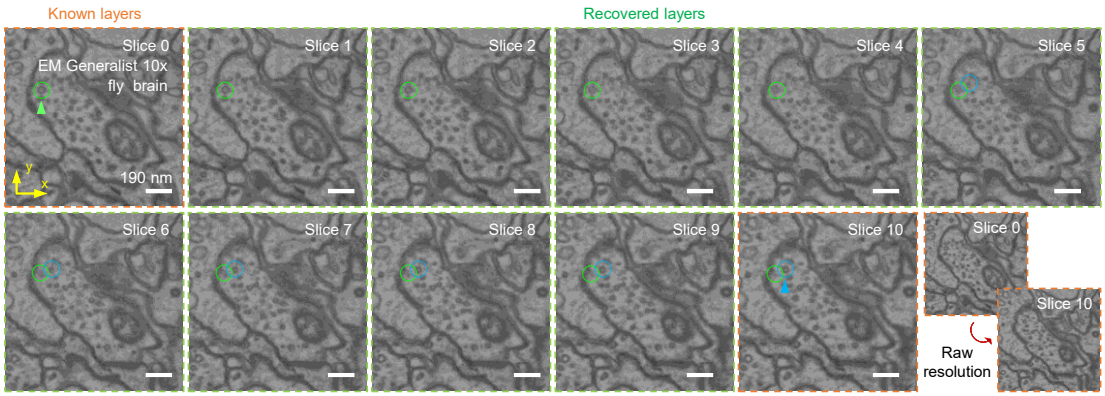
SI Fig. 4. (a) Recovery results of EM Generalist for mouse brain EM images with **slight defocus**, **moderate defocus**, and **severe defocus**, alongside comparisons with the ground truth. **(b)** Quantitative evaluation of the reconstruction results. For images with **severe defocus**, EM Generalist significantly mitigates the impact of defocus, achieving a PSNR of up to 29 and a SSIM of 0.72, with a FRC resolution below 30 nm (pixel size = 4.0 nm), demonstrating its robustness in challenging restoration scenarios.



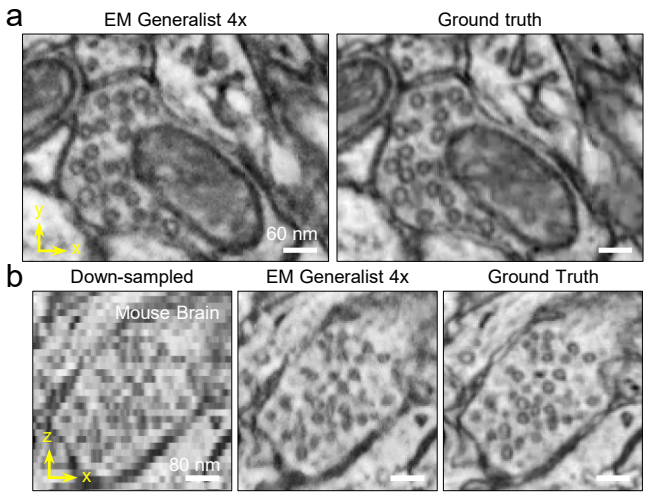
SI Fig. 5. (a) Comparison of 6-fold super-resolution results (x-z plane, mouse liver dataset) between interpolation, EMDiffuse and EM Generalist. EM Generalist demonstrates superior reconstruction quality as visualized by the highlighted regions (orange dashed boxes). **(b)** Quantitative metrics (PSNR, SSIM, FRC resolution) of EM Generalist reconstruction also surpass those of interpolation and EMDiffuse.



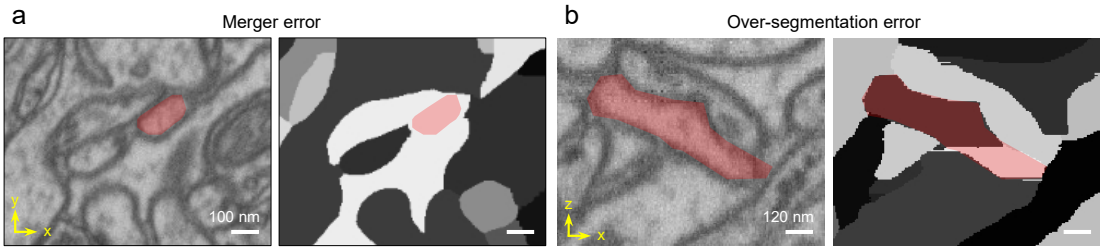
SI Fig. 6. EM Generalist achieves 10-fold super-resolution on raw anisotropic fly brain vEM data (resolution: 4×4×40 nm). Selected slices (0, 20, 40, …, 100) of raw images and reconstructions along yz-plane **(a)** and xz-plane **(b)** highlight the restoration of fine structures. Green borderlines in **(a)** and pink borderlines in **(b)** show that EM Generalist recovers clear boundaries of neuronal cells and mitochondria in different layers.



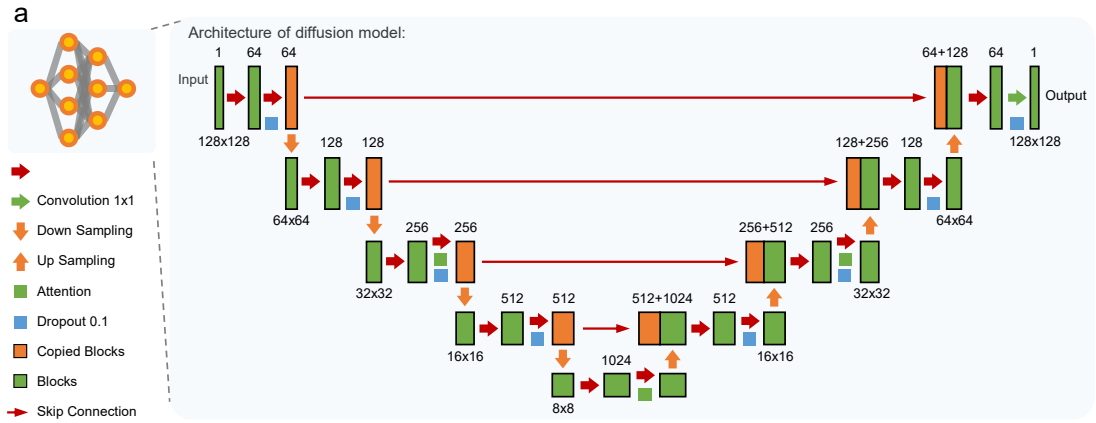
SI Fig. 7. Isotropic reconstructions of synaptic vesicles in fly brain data at 10-fold super-resolution using EM Generalist. Orange dashed boxes denote known anisotropic slices, while other green dashed are reconstructed slices. Since the vesicles marked in green are present in both the upper and lower known planes, they are correctly and consistently recovered across all intermediate reconstructed slices. In contrast, vesicles marked in blue, observed only in the lower plane, appear predominantly in the latter half of the reconstructed volume. The reconstructed images successfully approximate the spherical morphology of these vesicles across different reconstructed planes, effectively recovering their 3D structures.



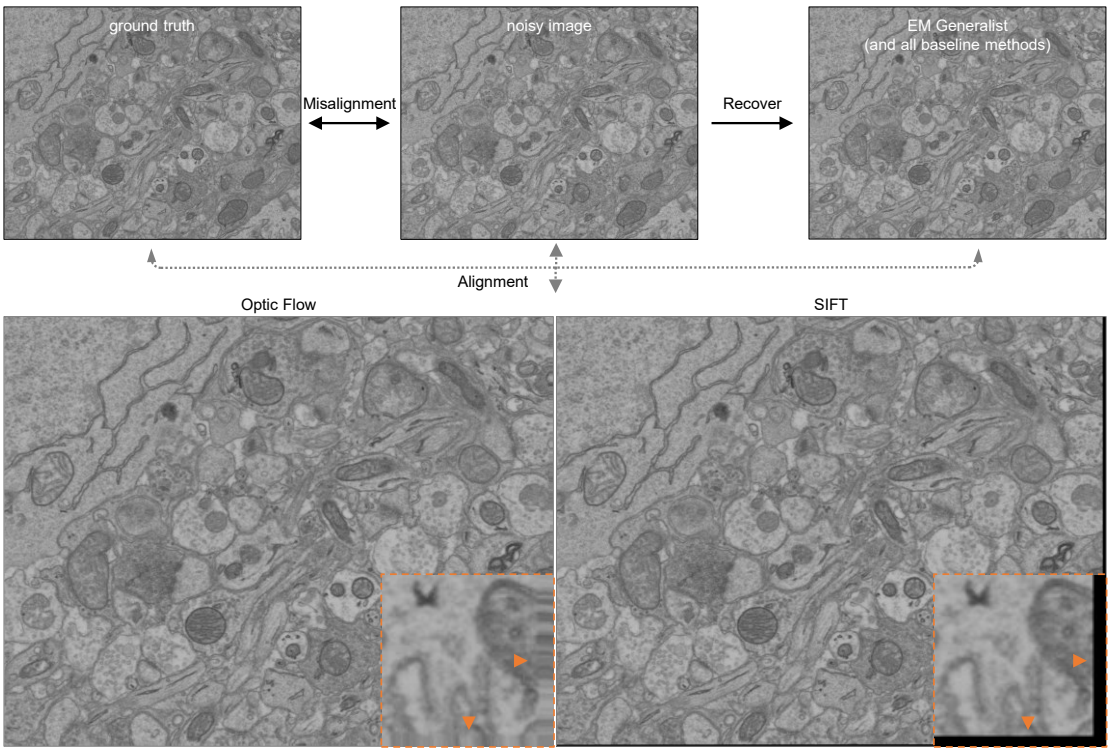
SI Fig. 8. Reconstruction of another mouse brain dataset (4×4×4 nm resolution, down-sampled to 16 nm axial resolution, from https://elifesciences.org/articles/25916/figures#videos). EM Generalist accurately recovers vesicles in both **(a)** x-y (interpolated slice) and **(b)** x-z planes, matching the number and location in the ground truth.



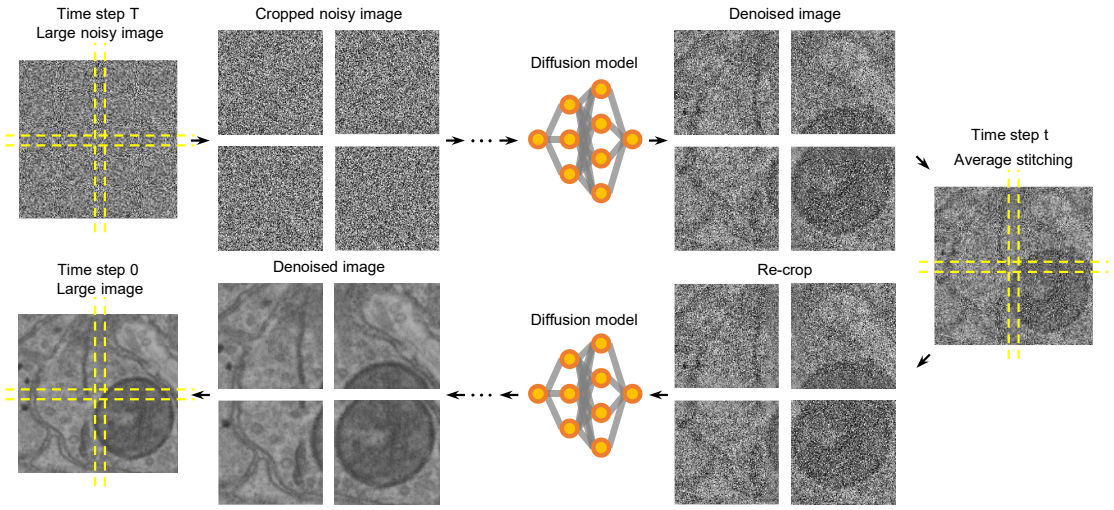
SI Fig. 9. (a) Merger error: The red masks in vEM image (left) and segmentation map (right) highlight an example of a merger error, where an individual dendritic spine is incorrectly segmented together with adjacent regions. (b) Over-segmentation error: The red masks in vEM image (left) and segmentation map (right) highlight a complete neural process within the vEM image, which has been erroneously divided into separate segments by the segmentation result.



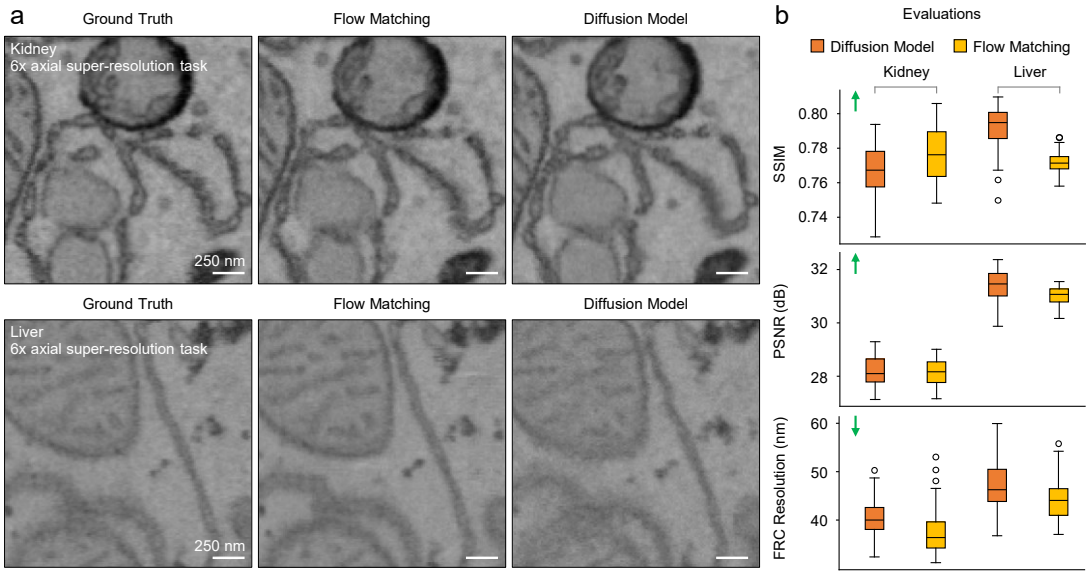
SI Fig. 10. The architecture of the diffusion foundation model for EM Generalist is a U-net model, comprising an encoder and a decoder (hidden layer dimension: 128 ,64, 32, 16, 8; layers: 64, 128, 256, 512, 1024). Attention blocks are integrated into the third and final layers to enhance model performance.



SI Fig. 11. To address the misalignment issues of the ground truth image and degraded image in real paired imaging datasets, we use preprocessing procedures offered by the alignment method in the supervised baseline, EMDiffuse. Both degraded images and the results from EM Generalist or other baseline methods were aligned using optical flow and SIFT-based registration before metric computation. We performed alignment on the recovered images rather than on the raw images, as the alignment procedures alter the noise distribution of the raw data, thereby degrading the denoising performance of EM Generalist. The bottom row showcases optical flow and SIFT alignment results, with the orange dashed box highlighting regions of significant mismatch. This alignment ensures fair and accurate performance evaluations for different methods.



SI Fig. 12. The large-scale image generation process using a diffusion model trained on 128×128 patches. During each denoising step, the large image is divided into overlapping 128×128 patches with a 16-pixel overlap margin. After denoising, the patches are reassembled into the full image, with overlapping regions blended by averaging.



SI Fig. 13. Flow matching, a novel variant of diffusion model, is used to accelerate the reconstruction process in EM Generalist framework. **(a)** A comparison between the ground truth image, and the recovered results by flow matching and diffusion model for 6-fold axial super-resolution tasks on kidney and liver vEM test data. **(b)** Evaluations for the flow matching and diffusion model results. Flow matching can recover images over 20 times faster, and hold a comparable evaluation result with the diffusion model-based approach.

a Training data

	Pixel size	Equipment	Source
CEM500K	2-20 nm	multiple	https://elifesciences.org/articles/65894
Tobacco Leaf Chloroplast	3.6 nm	SBF-SEM	https://www.ebi.ac.uk/empiar/EMPIAR-11831/
Human hepatocellular carcinoma cell	4 nm	FIB-SEM	https://www.ebi.ac.uk/empiar/EMPIAR-11849/
Myelin	5 nm	FIB-SEM	https://www.ebi.ac.uk/empiar/EMPIAR-11214/
Human brain	4 nm	ssSEM	https://h01-release.storage.googleapis.com/data.html
Microphage	4 nm	FIB-SEM	https://openorganelle.janelia.org/datasets/jrc_macrophage-2
Mouse brain	3.3 nm	FEI Verios SEM	https://zenodo.org/records/10205819
Mouse Heart	4.4 nm	FEI Verios SEM	https://zenodo.org/records/10205819
Mouse Liver	4.4 nm	FEI Verios SEM	https://zenodo.org/records/10205819
Hela cell	4 nm	FIB-SEM	https://openorganelle.janelia.org/datasets/jrc_hela-3
T cell	8 nm	FIB-SEM	https://openorganelle.janelia.org/datasets/jrc_ctl-id8-4
Mouse Kidney (3D)	8 nm	FIB-SEM	https://openorganelle.janelia.org/datasets/jrc_mus-kidney
Mouse Liver (3D)	8 nm	FIB-SEM	https://openorganelle.janelia.org/datasets/jrc_mus-liver
Mouse Skin (3D)	8 nm	FIB-SEM	https://openorganelle.janelia.org/datasets/jrc_mus-skin-1

b 3D test data

	Pixel size	Equipment	Source
Mouse Brain CA1 hippocampus region	5 × 5 × 5 nm	SEM	https://www.epfl.ch/labs/cvlab/data/data-em
Fly brain	4 × 4 × 40 nm	SEM	https://www.janelia.org/project-team/flyem/manc-connectome
Mouse Brain	4 × 4 × 4 nm	FIB-SEM	https://elifesciences.org/articles/25916/figures#videos

SI Table 1: Lists of training datasets **(a)** and 3D test datasets **(b)**, including their sample types, pixel sizes, imaging equipment and source links.

Sample Preparation and Imaging Details

The experimental validation samples for denoising task (plant stigma, kidney, HeLa cells, heart, and mouse oocytes) were collected from Electron Microscopy Platform, School of Life Sciences, Peking University, originating from experiments conducted by multiple independent research groups. And for super-resolution and deblurring tasks, we prepared the mouse brain cortex sample.

(1) Stigma, kidney, human HeLa cells, heart, and mouse oocytes: Imaging was performed using a Helios dual-beam scanning electron microscope at an accelerating voltage of 2 kV. Images were acquired with dwell times of 500 ns, 1 μs, 3 μs, and 10 μs, where the first three were used to produce noisy images, and the images captured at 10 μs served as ground truth. Three magnifications were applied: 15,000× (pixel size 8.98 nm) for stigma and kidney tubules, 20,000× (pixel size 6.73 nm) for kidney Glomerulus, mouse oocytes, and human HeLa cells, and 40,000× (pixel size 3.36 nm) for imaging mitochondria in kidney cells. Subsequently, by manually adjusting the defocus and stigmator (X and Y axes) settings, we acquired blurred-clear paired images at a dwell time of 3 μs, ensuring that the imaging area matched the noisy region, and the pixel size remained consistent with that of the noisy images for each category.

(2) Mouse brain cortex data: Animals were anesthetized and perfused with 15 mL PBS followed by 30 mL fixative mixture containing 2% PFA (EMS), 2.5% glutaraldehyde (EMS), 2.1% sucrose (EMS), and 0.1 M sodium cacodylate buffer (pH=7.4) (Sigma-Aldrich). Mouse brain was cut into small sizes and fixed with fixative 48-72 h. Then, the samples were fixed with 2% OsO4 aqueous solution (EMS) and 1.5% potassium ferrocyanide (EMS) in 0.1 M sodium cacodylate for 1h at 4oC. Subsequently, the samples were incubated with 1% thiocarbohydrazide (EMS) for 20 min at RT, 2% OsO4 aqueous solution for 30 min at RT, and 2% Neodymium (Sigma-Aldrich) at 4oC overnight. Tissues were washed three times (5 min each) using H2O between each step. On the next day, the samples were dehydrated through a graded ethanol series (30, 50, 75, 85, 95, 100%, 7 min each, all cooled at 4°C) followed by immersion into 1:1 and 2:1 mixtures of acetone and EMbed 812 embedding kits (EMS) at room temperature for 1h and pure Embed 812 resin overnight on a rotator. Immersed samples were then incubated in pure resin and placed in embedding moulds (Ted Pella) in a pre-warmed oven (60°C) for 48 to 72 h. After polymerization, the resin-embedded tissue was cut into thick sections of 100 nm. The sections were mounted on a silicon wafer.

Imaging was performed using a Zeiss GeminiSEM 360 scanning electron microscope at an accelerating voltage of 2 kV. The ground truth pixel size was 4 nm, and the dwell time was 1.6 μs. For each task, only one imaging condition was altered. For super-resolution tasks, the dwell time was kept constant at 1.6 μs, while the pixel size was varied to 16 nm, 12 nm, and 8 nm, respectively. For defocus tasks, the focus was manually adjusted until the images appeared blurred. For ultra-large images, the pixel

size remained 4 nm, with the total number of pixels reaching approximately 320 megapixels.

Publicly Available EM Datasets Collection

To clearly document the sources and characteristics of various public EM datasets used in this paper, we provide detailed information in Supplementary Information (SI) Table 1. Both the training data and test data are included.

Model architecture and implementations for EM Generalist

We employed the 'UNet2DModel' from the Hugging Face 'diffusers' library as the basic architecture of our diffusion model, As depicted in SI Fig. 10, the model consists a down-sampling pathway and an up-sampling pathway[1]. Since EM images are grayscale, the model's input and output channels are set to 1. The architecture consists of four stages of down-sampling and four corresponding up-sampling stages. Each down-sampling block consists of two convolution layers per block, both with batch normalization applied. The down-sampling pathway consists of two down-sampling modules, followed by a self-attention module, and another down-sampling module, allowing the model to focus on significant features at a coarser resolution.

The channel dimensions are progressively increased as 64, 128, 256, 512. In attention mechanism[2], the input feature map is first transformed through three linear projections into the 'Query' (Q), 'Key' (K), and 'Value' (V) embedding spaces. Each embedding typically has the shape $\mathbb{R}^{H\times W\times C}$, where H and W represent the spatial height and width of the feature map, and C denotes the feature dimension. The attention weights are computed using the dot product between the Query and Key, normalized to prevent large values, thus we can derive the formula as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{C}})V.$$

In this equation QK^T represents the dot product between Query and Key, calculating the similarity between different positions. $\sqrt{C}$ is a scaling factor that prevents the dot product from growing too large when the feature dimension is high. The softmax operation normalizes the computed similarities into a probability distribution, ensuring the attention weights sum to 1.

The up-sampling pathway mirrors the down-sampling pathway, employing symmetric block types, including three up-sampling modules and a self-attention module. The input to each up-sampling block differs slightly from that of the down-sampling blocks, as it combines the output of the previous up-sampling block with the output from the corresponding down-sampling block at the same level. This skip connection facilitates the fusion of shallow and deep features, thereby enhancing the model's performance. A dropout rate of 0.1 is used for regularization to mitigate overfitting.

Evaluation metrics

Structural Similarity Index (SSIM)[3] is a widely used evaluation metric for measuring the similarity between two images. SSIM is particularly useful in tasks such as image reconstruction, compression, and denoising, as it assesses perceived visual quality by comparing structural information, luminance, and contrast between two images. Unlike traditional pixel-wise metrics like mean squared error (MSE), SSIM aligns with human visual perception of image quality. Given two images x and y, the SSIM index is defined as:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where μ_x and μ_y are the local means of images x and y representing luminance, σ_x^2 and σ_y^2 are the local variances of x and y, representing contrast, σ_{xy} is the covariance between x and y, representing structural similarity, and C_1 and C_2 are constants that stabilize the computation by preventing division by zero. SSIM values range from [-1, 1], where 1 indicates perfect similarity, values close to 0 indicate low similarity, and negative values suggest significant differences between the images.

Peak Signal-to-Noise Ratio (PSNR) is a widely used evaluation metric to measure the similarity between a processed image and a reference image. It is based on pixel-wise error, i.e., MSE, and expresses the ratio of signal strength to noise in a logarithmic scale. Given a reference image x and a processed image y, PSNR is defined as:

$$\mathrm{PSNR}(x,y) = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}(x,y)}\right),$$

where MAX is the maximum possible pixel value of the image. For example, for 8-bit images, $\mathrm{MAX}=255$. $\mathrm{MSE}(x,y)$ is the Mean Squared Error between the two images, defined as:

$$\mathrm{MSE}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2,$$

where x_i and y_i are the pixel values of images x and y, and N is the total number of pixels in the image. The PSNR value is expressed in decibels (dB) and reflects the ratio between signal strength and noise. A higher PSNR indicates a higher similarity between the images, with less noise.

Besides PSNR, and SSIM, we also use **Fourier ring Correlation (FRC)** and **Fourier Shell Correlation (FSC)** to estimate the resolution of the recovered image/volume[4]. FRC is based on the Fourier transform of 2D images, and it calculates the correlation

between the frequency components of the recovered and ground-truth 2D images in annular frequency bands. The formula for FRC is:

$$\mathrm{FRC}(f) = \frac{\sum_{k\in \mathrm{ring}(f)} F_1(k)F_2^*(k)}{\sqrt{\sum_{k\in \mathrm{ring}(f)} |F_1(k)|^2 \sum_{k\in \mathrm{ring}(f)} |F_2(k)|^2}},$$

Where $F_1(k)$ and $F_2(k)$ are the Fourier transforms of two images, $F_2^*(k)$ is the complex conjugate of $F_2(k)$, and $\mathrm{ring}(f)$ refers to the frequency band at frequency f in the Fourier space. FRC evaluates the similarity of 2D images in the frequency domain by analyzing correlations at specific spatial frequencies.

Similarly, we have FSC for 3D volumes as follows:

$$\mathrm{FSC}(f) = \frac{\sum_{k\in \mathrm{shell}(f)} F_1(k)F_2^*(k)}{\sqrt{\sum_{k\in \mathrm{shell}(f)} |F_1(k)|^2 \sum_{k\in \mathrm{shell}(f)} |F_2(k)|^2}},$$

where $F_1(k)$ and $F_2(k)$ are the Fourier transforms of the two 3D volumes, and $\mathrm{shell}(f)$ refers to the spherical shell at frequency f in the 3D Fourier space. Further, we derive the FRC/FSC resolution from FRC/FSC curve. The structural resolution was determined using FRC/FSC analysis. Following standard cryo-EM protocols, we applied the 0.5 correlation threshold to identify the spatial frequency at which two independent reconstructions maintain structural consistency, ensuring the measured resolution reflects authentic biological features rather than stochastic noise.

In addition to the evaluation metrics calculated through comparison with ground truth, the article also employs several no-reference image quality metrics to assess the quality of the reconstructed images, particularly in the absence of ground truth images. These methods include parameter-free resolution and NIQE[5,6]. Parameter-free resolution is a metric designed to assess the resolution of an image without requiring any prior knowledge or reference parameters. It is based on the analysis of the image's spatial frequencies, evaluating how well fine structures are resolved. It is particularly useful when assessing images where high-frequency components indicate sharper details. The metric does not require the specification of any parameters such as thresholds or scaling factors, making it inherently robust to various imaging conditions. For details, please refer to Descloux's work[5].

The NIQE index is a widely used no-reference image quality assessment metric. It is designed to predict perceptual quality without the need for reference images. NIQE operates by first extracting a set of natural scene statistics from the image and then comparing these statistics to a pre-established model of natural image features. The

model captures typical image structures, such as textures and edges, that are naturally expected in high-quality images. The NIQE score is calculated as:

$$\text{NIQE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{f}_i - \mathbf{m})^T \mathbf{C}^{-1}(\mathbf{f}_i - \mathbf{m})}$$

where $\mathbf{f}_i$ represents the extracted features from the image, $\mathbf{m}$ is the mean of the natural image feature distribution, and $\mathbf{C}$ is the covariance matrix of these features. A lower NIQE score indicates higher image quality.

During the evaluation process, to fix the slight misalignment between the real degraded image and the ground truth image caused by separate acquisition processes, Scale-Invariant Feature Transform (SIFT)[7] and Optical Flow[8] algorithms were employed to register the reconstructed image with the ground truth image (SI Fig. 11).

Confidence intervals: To quantify the precision of detection performance, we computed 95% confidence intervals (CIs) using the Wilson score method for binomial proportions[9]. This approach provides robust interval estimates that account for sampling variability while maintaining accuracy across extreme probability ranges. The interval is defined as:

$$\text{CI} = \frac{p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

where p is the observed proportion (e.g., detection rate), n is the total sample size, and z=1.96 corresponds to the 95% confidence level in the standard normal distribution. The Wilson score method was selected over asymptotic approximations (e.g., Wald intervals) due to its superior coverage properties, particularly for small sample sizes or near-boundary proportions (e.g., >99% detection rates). This formulation ensures statistically rigorous uncertainty quantification in classification tasks.

Baseline settings

The baseline methods used in this study include two unsupervised approaches, SCUnet[10] and ZS-deconvnet[11], as well as the state-of-the-art supervised learning-based method EMDiffuse[12].

SCUnet integrates Swin Transformer blocks with convolutional layers to establish hierarchical feature learning for EM denoising, where we specifically adopted its noise-level=50 grayscale model (https://replicate.com/cszn/scunet) which provides the best reconstructions. **ZS-DeconvNet** implements zero-shot microscopy enhancement via Noise2Noise-based training on in-domain EM data, deployed through its ImageJ plugin

with 250-epoch CPU training (3-4 days per task). **EMDiffuse** is trained on paired EM datasets (Zenodo:10205819) to implement supervised diffusion processes through its open-source framework (https://github.com/Luchixiang/EMDiffuse/), iteratively refining structural features specific to EM imaging.

Stitching for Large EM Image During Reconstruction

To effectively recover large degraded images using our model—which is trained on small 128 × 128 patches—we propose a stitching strategy to align adjacent image patches seamlessly (SI Fig. 12). Specifically, large images are divided into overlapping 128 × 128 patches, each overlapping its neighbors by 12.5% (16 pixels). Initially, we generate a Gaussian noise image matching the degraded image size and similarly divide it into overlapping patches. During each subsequent iteration of the reconstruction, overlapping regions between neighboring patches are averaged, creating a cohesive intermediate image. This intermediate image is again segmented into overlapping patches for the next step. This method reduces artifacts and ensures smooth transitions between patches, significantly enhancing the overall reconstruction quality.

Flow Matching-based inference time acceleration

The computational efficiency of EM Generalist is constrained by two key limitations: (1) the inherent 1000-step diffusion sampling process, and (2) the computationally intensive gradient calculation required for conditional score function estimation. To alleviate these constraints, we adopt a diffusion model architecture based on flow matching[13,14]. Flow matching is a technique where a continuous velocity field is learned to deterministically transport a simple base distribution (such as Gaussian noise) to the target data distribution. Rather than following the stochastic reverse process of typical diffusion models, flow matching directly learns the dynamics of this transformation via an ordinary differential equation (ODE). Specifically, if we denote the latent variable at time step t as $\mathbf{x}_t$, the generative process is governed by:

$$\frac{\partial \mathbf{x}_t}{\partial \mathrm{t}} = v_t(\mathbf{x}_t),$$

where $v_t(\mathbf{x}_t)$ represents the instantaneous velocity field at time t. This velocity field essentially characterizes the "speed" and "direction" with which $\mathbf{x}_t$ should be adjusted to morph into a sample from the data distribution. This ODE formulation enables high-quality sample generation with significantly fewer steps than the iterative denoising used by conventional diffusion models.

The iterative sampling procedure based on flow matching can be derived via a simple Euler integration. Let $\mathbf{x}_t$ denote the latent variable at time t and consider two successive time points, the update equation then becomes:

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \Delta t v_\theta(\mathbf{x}_t, t) = \mathbf{x}_t + \Delta t \frac{\partial \hat{\mathbf{x}}_0(\mathbf{x}_t, t)}{\partial t}.$$

Here, $v_\theta(\mathbf{x}_t, t)$ denotes the estimated velocity field parameterized by θ—equivalent to the derivative of the estimated denoised latent variable $\hat{\mathbf{x}}_0(\mathbf{x}_t, t) = \mathbf{x}_t - t\, v_\theta(\mathbf{x}_t, t)$—and Δt represents the discretized time interval. This update rule is analogous to the reverse diffusion update used in DDPM but replaces the expensive 1000-step-iteration with much fewer step iterations (as few as 100 steps) by the learned velocity field. Consequently, the sampling process becomes considerably faster while preserving the model's overall performance.

Follow the predefined posterior sampling approach, we can formulate the posterior sampling approach for flow matching as follows:

$$\mathbf{x}_{t-1} = \mathbf{x}_t + \Delta t \frac{\partial \hat{\mathbf{x}}_0(\mathbf{x}_t, t)}{\partial \mathbf{x}_t} - (1-t)^{\alpha} \nabla_{\hat{\mathbf{x}}_0} \| \, y - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{x}_t, t)) \, \|^2.$$

Here, the first term represents the flow matching sampling process, and the second term enforced data consistency. By taking the gradient for $\hat{\mathbf{x}}_0$ rather than $\mathbf{x}_t$, the approximation of the conditional term can further accelerate the inference process by skipping the heavy calculation for gradient with respect to the diffusion model. The integration of flow matching with the new gradient calculation formulation reduced reconstruction time of an isotropic 128×128×128-pixel volume from 40 minutes to 1.5 minutes on a single NVIDIA A800 GPU, achieving a 26.7× speedup without performance degradation (SI Fig. 13).

Our implementation for flow matching utilizes the same model architecture the diffusion model, featuring a U-Net structure enhanced with residual blocks, channel multiplication, and self-attention mechanisms. The training protocol comprised 200,000 optimization steps executed across multiple GPUs with a batch size of 256, requiring approximately 24 hours of computation on NVIDIA A40 hardware. We employed the Adam optimizer with an initial learning rate of 10^{-4}, incorporating a 1,000-step warmup phase to ensure training stability. Furthermore, we applied Exponential Moving Average (EMA) to model parameters with a decay rate of 0.999 to enhance model convergence and robustness during training. During reconstruction, the hyper-parameter α can be selected from the range [0, 1] to impose data consistency constraints of varying strengths (normally 0.3), where 1 indicates no constraint and 0 corresponds to full enforcement of data consistency. Based on empirical evaluation, α is typically set to 0.3, and users are encouraged to manually adjust this hyper-parameter for different datasets to optimize reconstruction quality.

References

1. Ronneberger, O., Fischer, P. & Brox, T. 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* 234–241 (2015).
2. Vaswani, A. *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 30, (2017).
3. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 600–612 (2004).
4. Saxton, W. O. & Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein. *J Microsc* 127, 127–138 (1982).
5. Descloux, A., Grußmayer, K. S. & Radenovic, A. Parameter-free image resolution estimation based on decorrelation analysis. *Nat Methods* 16, 918–924 (2019).
6. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal Process Lett* 20, 209–212 (2012).
7. Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *Int J Comput Vis* 60, 91–110 (2004).
8. Lucas, B. D. & Kanade, T. An iterative image registration technique with an application to stereo vision. in *Proceedings of the 7th International Joint Conference on Artificial Intelligence* vol. 2 674–679 (1981).
9. Casella, G. & Berger, R. L. *Statistical Inference*. (Duxbury/Thomson Learning, 2002).
10. Zhang, K. *et al.* Practical blind image denoising via Swin-Conv-UNet and data synthesis. *Machine Intelligence Research* 20, 822–836 (2023).
11. Qiao, C. *et al.* Zero-shot learning enables instant denoising and super-resolution in optical fluorescence microscopy. *Nat Commun* 15, 4180 (2024).
12. Lu, C. *et al.* Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nat Commun* 15, 4677 (2024).
13. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).
14. Martin, S., Gagneux, A., Hagemann, P. & Steidl, G. PnP-Flow: Plug-and-play image restoration with flow matching. *arXiv preprint arXiv:2410.02423* (2024).