Data Generation for Benchmarking Deep Learning on Materials Images via Noise Injection and CycleGAN

Masato Suzuki[1,2*] **Yasuhiko Igarashi**[1,3,4,5†]
[1]Institute of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
[2]Research & Advanced Development Division, Materials Analysis Laboratory, The Yokohama Rubber Co. Ltd, 2-1 Oiwake, Hiratsuka, Kanagawa 254-8601, Japan
[3]Tsukuba Institute for Advanced Research (TIAR), University of Tsukuba, Ibaraki 305-8577, Japan
[4]Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Ibaraki 305-0003, Japan
[5]Japan Synchrotron Radiation Research Institute (JASRI), Hyogo 679-5198, Japan

Abstract

Manual annotation of material microscopy images is time-consuming, costly, and requires domain expertise. This annotation bottleneck limits model training and fair benchmarking. Prior cycle-consistent generative adversarial network (CycleGAN)-based data generation, despite being promising, often relied on computationally expensive simulations and struggled to capture the diverse noise characteristics, making it task-specific. In this study, we introduce an automated pipeline which simplifies dataset generation and improves generality by combining parametric simulations, diverse modality-specific noise injection, and CycleGAN-based texture transfer while preserving the ground-truth masks. Case studies on rubber materials with stripe-like noise in optical microscopy highlight its versatility. This pipeline was evaluated on a public transmission electron microscopy (TEM) nanoparticle dataset to obtain a quantitative comparison with manual annotations. Our results show that the segmentation accuracy approached that of human-labeled data while also reproducing characteristic imaging artifacts. This framework reduces dataset cost, explicitly addresses noise diversity, and enables customized, reproducible, and noise-aware benchmarks aligned with real experimental settings.

1 Introduction

Machine learning (ML), particularly deep learning (DL), promises substantial gains in materials science [1, 2], yet data acquisition is costly and slow, often requiring expert operations on advanced instruments [3, 4, 5, 6]. Unlike general computer-vision tasks with large-scale labeled datasets [7], and in contrast to a few specialized tasks where large-scale datasets exist [8], materials datasets are far smaller [9, 10, 11, 12, 13]. However, a core problem is the annotation bottleneck: Creating a task-specific ground truth (e.g., phase boundaries or defect regions) requires advanced expertise that varies across material classes such as polymers, ceramics, and metals. Consequently, training data remain scarce and standardized benchmark datasets for fair, reproducible comparisons are lacking. Compounding this, advanced imaging often contains noise from both instrumentation and sample physics, including Gaussian noise and structured artifacts [14, 15, 16, 17, 18, 19, 20, 21]. While targeted removal methods exist [20, 21, 22], the scarcity of clean/noisy pairs hinders supervised training and fair benchmarking. Existing attempts to mitigate this gap using data generation with cycle-consistent generative adversarial networks (CycleGAN) [23] have shown promise [24, 25, 26]; however, they often depend on computationally expensive high-accuracy simulations. Incorporating diverse noise into such simulations remains difficult, making them task-specific. Extreme mismatch

*masato.suzuki@y-yokohama.com
†Corresponding author: igayasu1219@cs.tsukuba.ac.jp

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat Themed Track – “Benchmarking in AI for Materials Design".

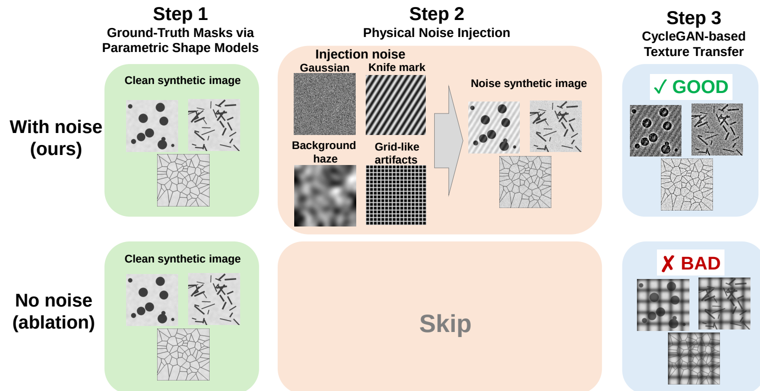

Figure 1: Overview of our proposed pipeline. Step 1 — Simulation: parametric shape models generate clean structural images and binary ground-truth masks (circles, rods, Voronoi). Step 2 — Physical noise injection: mimics acquisition to inject artifacts according to sample preparation and imaging conditions, such as knife-mark stripes from cutting damage, and Gaussian noise arising from electron microscopy observation. Step 3 — CycleGAN-based texture transfer: the style of unlabeled real images is learned and transferred to the noisy simulations while preserving the original masks. By injecting modality-specific noise beforehand, the simulated images become closer to the target domain, making it easier for CycleGAN to adapt and produce more realistic textures – paired datasets of CycleGAN-adapted images with accurate labels for supervised training.

in entropy or high-frequency content between the synthetic and real domains may cause CycleGAN to embed hidden signals, making the generated images appear realistic but unstable [27, 28]. Injecting modality-specific noise into synthetic data helps reduce this mismatch and improves stability. In this study, we introduce an automated pipeline that complements existing methods, simplifies the data generation process, and improves robustness under degraded conditions while enhancing generality. We created annotation-free datasets via simulations, noise injection, and CycleGAN-based texture transfer, and validated them on rubber and TEM nanoparticle data. Our pipeline transfers experiment-like noise onto simulated structures (e.g., fillers and particles), thereby generating training datasets for segmentation, detection, and regression on materials microscopy images. This framework lowers the cost of dataset creation, increases generalizability, and enables customized, reproducible benchmarks aligned with real experimental conditions, thereby fostering a closer integration of ML and materials science.

2 Methods

Our framework generates structural images together with ground truth masks from parametric models to replace manual annotation, further adapting them to obtain experimental realism. This study aims to reproduce the noise statistics of experimental images by first performing noise injection on simulated structural images to add acquisition-like artifacts such as random noise and illumination inhomogeneity, and then applying CycleGAN-based domain transfer to emulate the contrast and noise characteristics associated with instrument-induced electron microscopy noise (e.g., drift and detector noise) and knife-mark artifacts introduced during sample sectioning [24], allowing combinations of multiple noise sources to reproduce composite experimental noise. To better match the real data, diverse types of noise are injected. An overview of this process is shown in Figure 1. First, we generated clean images and masks by following specified parameters for shapes and size distributions. By adjusting the modeling parameters, on-demand datasets tailored to specific materials and analysis tasks can be generated. Importantly, this method supports deliberate abstraction, which is a standard in materials practice. For example, a complex polygonal agglomerate may be represented as an equivalent-area circle [29] and treated as the ground truth. This aligns the learning target with the analyst's intended abstraction. In materials research, complex real-world shapes are often modeled as simplified, task-relevant representations for quantitative analysis [30, 31, 32]. To reproduce experimental artifacts, we injected diverse types of physical noise into clean simulations while keeping the masks unchanged. Depending on the imaging modality, these artifacts include knife-mark stripes from sample preparation [17, 18] and Gaussian or Poisson noise from electron microscopy [15, 16].

Explicitly encoding these noise processes brings the synthetic images closer to the real domain and better captures the variability observed in practice. This allows the synthetic images to reflect the injected noise and become closer to the experimental data; however, differences in texture remain. Finally, to address this texture gap, we used CycleGAN, a style-transfer model that learns from noisy synthetic and unlabeled real images. Although GANs (Generative Adversarial Networks)[33] have been used to generate materials microscopy data [34, 35], they typically require paired training data, whereas CycleGAN operates on unpaired data and produces realistic images aligned with ground-truth masks, yielding ideal training pairs. The key point is that as long as the authentic microscope images fed to CycleGAN are identical to those that will later be analyzed, the spatial coverage of the images produced by CycleGAN exactly matches the region of data required for the downstream task and is therefore sufficient in terms of both quantity and quality. Moreover, because our pipeline is annotation-free, any future changes in sample type or imaging conditions can be accommodated by simply adding new unlabeled images and retraining the CycleGAN. Our approach starts with simple simulations and injects noise together with CycleGAN-based adaptation, to produce images that resemble those in real experiments while preserving ground-truth masks. This makes it possible to add annotations to originally unlabeled datasets and generate reproducible, noise-aware benchmarks for the fair evaluation of machine learning methods in materials science.

3 Results and Discussions

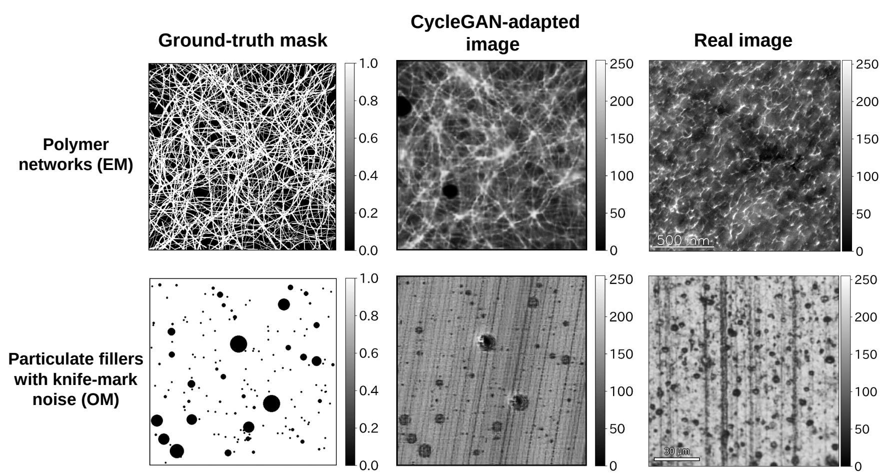

Figure 2: Case studies of training data generation. Top row: Nanoscale polymer networks observed by electron microscopy. Ground-truth image created by random-walk simulation was converted into realistic textures, providing paired datasets for model development. Bottom row: Micron-scale particulate fillers with knife-mark noise observed using optical microscopy. Circular particle masks were simulated and transformed into realistic noisy images, yielding datasets that reproduce characteristic experimental artifacts.

3.1 Extending Versatility: to Multiscale, Multigeometry, and Multinoise Scenarios

To demonstrate the adaptability of the proposed framework to a wide range of material structures, we conducted two case studies on rubber materials with different geometric configurations and observation scales. First, we created training data for the nanoscale polymer networks observed using electron microscopy[36]. Specifically, we target high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) of vapor-phase $\mathrm{OsO_4}$-stained ultrathin rubber sections acquired at $200\,\mathrm{kV}$ and $60{,}000\times$ with 1024×1024 pixels at $1.52\,\mathrm{nm/pixel}$ (field of view $\approx 1.56\,\mu\mathrm{m} \times 1.56\,\mu\mathrm{m}$). String-like structures were simulated using a random-walk model and then transformed into realistic textures, yielding training datasets paired with ground-truth masks. Notably, the synthetic data also reproduced the characteristic features of real HAADF-STEM images, such as faint appearances and blurred structures caused by staining. In addition to conventional binary segmentation, there are also enhancement tasks where filamentous or string-like structures are modeled with continuous intensity values rather than binary 0/1 labels. Such formulations aim to emphasize the visibility of networks under noisy conditions, instead of enforcing strict segmentation boundaries. Next, we focused on micron-scale particulate fillers observed in optical microscopy[37], with particular

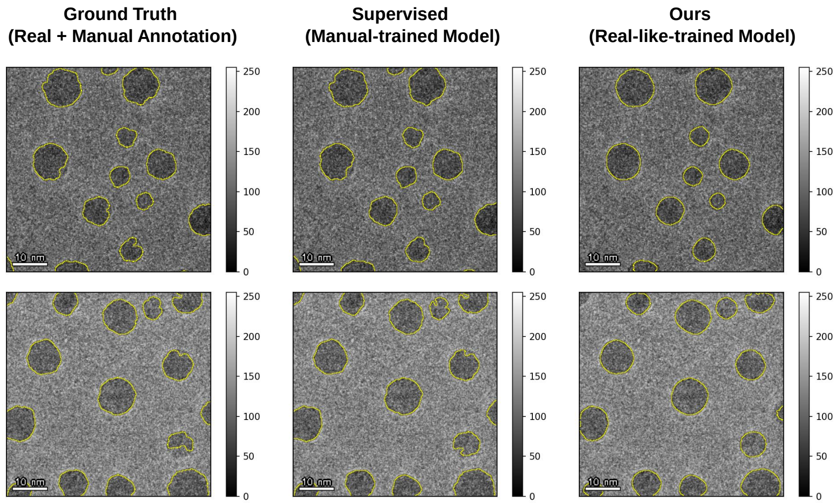

Figure 3: Comparison of segmentation results on real microscopy images. Left: Ground truth, i.e., real captured images with manual annotations shown as yellow contour overlays. Center: Predictions from a supervised model trained on manually annotated real data (yellow contour overlays). Right: Predictions from our method trained solely on real-like synthetic data (yellow contour overlays).

attention paid to the stripe-like noise caused by knife marks. Circular particles were simulated and texture transfer was applied to generate realistic noisy images. Importantly, the synthetic images successfully reproduced the characteristic knife-mark noise. Combined, these results demonstrate the versatility of the proposed approach.

3.2 Quantitative Evaluation: Comparison with Existing Benchmarks

To objectively evaluate the datasets generated by our framework, we adopted the publicly available high-resolution TEM nanoparticle datasets introduced by Horwath et al. [38]. Using our method, we created a large-scale synthetic dataset without manual annotation that mimicked the statistical and visual characteristics of the Horwath dataset. We then compared the segmentation accuracy (measured by the IoU and Dice coefficient) on unseen test images between two models: (i) Model A, trained on the original Horwath manual labels, and (ii) Model B, trained on 2,000 automatically generated image–mask pairs from our framework. We adhered to a strict image-level, group-aware split (70/30), preventing specimen/session leakage. CycleGAN training, as well as U-Net training and validation, were conducted within this 70% subset, and all metrics (IoU, Dice) are reported on the held-out 30% that was never used for CycleGAN/U-Net training or model selection. In addition to the quantitative evaluation, we qualitatively observed that the synthetic images faithfully reproduced the characteristics of key imaging artifacts of the TEM nanoparticle micrographs. Specifically, our "real-like" images captured both global background intensity inhomogeneities and edge-related bright fringes, commonly referred to as edge contrast effects, which closely resemble those seen in real TEM images. The IoU reached 0.884 with our method, which corresponds to about 95% of the IoU obtained with human annotations (0.931).

4 Conclusion

To address the annotation bottleneck in materials science, we propose a simple pipeline that integrates simulations, noise injection, and CycleGAN-based texture transfer to generate labeled datasets without manual effort. Case studies on rubber materials and a TEM benchmark confirmed that injecting realistic noise improves the fidelity of synthetic images. The versatility of the approach was further demonstrated across different structures and modalities. This method is effective for simple simulable structures but struggles with irregular geometries, requires domain-specific tuning, and performance may degrade on unseen images. These results motivate broadening to additional materials/modality pairs and challenging shape regimes in future releases. Nevertheless, by providing a practical route for generating reliable datasets, our study contributes to the development of fair and reproducible benchmarks, aligning with the goals of AI4Mat, to advance meaningful evaluation in materials science.

References

[1] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, April 2022.

[2] Ankit Agrawal and Alok Choudhary. Deep materials informatics: Applications of deep learning in materials science. *MRS Communications*, 9(3):779–792, September 2019.

[3] Michelle A. Smeaton, Patricia Abellan, Steven R. Spurgeon, Raymond R. Unocic, and Katherine L. Jungjohann. Tutorial on In Situ and Operando (Scanning) Transmission Electron Microscopy for Analysis of Nanoscale Structure–Property Relationships. *ACS Nano*, 18(52):35091–35103, December 2024.

[4] Linda E. Franken, Kay Grünewald, Egbert J. Boekema, and Marc C. A. Stuart. A Technical Introduction to Transmission Electron Microscopy for Soft-Matter: Imaging, Possibilities, Choices, and Technical Developments. *Small*, 16(14):1906198, 2020.

[5] Giovanni Fevola, Peter S. Jørgensen, Mariana Verezhak, Azat Slyamov, Andrea Crovetto, Zoltan I. Balogh, Christian Rein, Stela Canulescu, and Jens W. Andreasen. Resonant x-ray ptychographic nanotomography of kesterite solar cells. *Physical Review Research*, 2(1):013378, March 2020. Publisher: American Physical Society.

[6] M. Álvarez Murga, J. P. Perrillat, Y. Le Godec, F. Bergame, J. Philippe, A. King, N. Guignot, M. Mezouar, and J. L. Hodeau. Development of synchrotron X-ray micro-tomography under extreme conditions of pressure and temperature. *Journal of Synchrotron Radiation*, 24(1):240–247, January 2017. Publisher: International Union of Crystallography.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919.

[8] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, July 2017. ISSN: 1063-6919.

[9] Batuhan Yildirim and Jacqueline M. Cole. Bayesian Particle Instance Segmentation for Electron Microscopy Image Quantification. *Journal of Chemical Information and Modeling*, 61(3):1136–1149, March 2021. Publisher: American Chemical Society.

[10] Shih-Lin Lin. Research on tire crack detection using image deep learning method. *Scientific Reports*, 13(1):8027, May 2023. Publisher: Nature Publishing Group.

[11] D. J. Groom, K. Yu, S. Rasouli, J. Polarinakis, A. C. Bovik, and P. J. Ferreira. Automatic segmentation of inorganic nanoparticles in BF TEM micrographs. *Ultramicroscopy*, 194:25–34, July 2018.

[12] Lech Staniewicz, Thomas Vaudey, Christophe Degrandcourt, Marc Couty, Fabien Gaboriaud, and Paul Midgley. Electron tomography provides a direct link between the Payne effect and the inter-particle spacing of rubber composites. *Scientific Reports*, 4(1):7389, December 2014. Publisher: Nature Publishing Group.

[13] Daniil A. Boiko, Evgeniy O. Pentsak, Vera A. Cherepanova, and Valentine P. Ananikov. Electron microscopy dataset for the recognition of nanoscale ordering effects and location of nanoparticles. *Scientific Data*, 7:101, March 2020.

[14] Thien B. Nguyen-Tat, Tran Quang Hung, Pham Tien Nam, and Vuong M. Ngo. Evaluating pre-processing and deep learning methods in medical imaging: Combined effectiveness across multiple modalities. *Alexandria Engineering Journal*, 119:558–586, April 2025.

[15] M. Vulovic, B. Rieger, L. J. van Vliet, A. J. Koster, and R. B. G. Ravelli. A toolkit for the characterization of CCD cameras for transmission electron microscopy. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):97–109, January 2010. Publisher: International Union of Crystallography.

[16] Christian Zietlow and Jörg K. N. Lindner. An applied noise model for scintillation-based CCD detectors in transmission electron microscopy. *Scientific Reports*, 15:3815, January 2025.

[17] Trivikrama Raju C, Jakeer Hussain S, Yedukondalu G, and Murahari Kolli. An effective investigation of chatter prediction system on Al6061 alloy in an end milling process. *Journal of Engineering and Applied Science*, 71(1):150, July 2024.

[18] D. A. Hemsley, editor. *Applied Polymer Light Microscopy*. Dordrecht, 1989.

[19] Takehito Seki, Yuichi Ikuhara, and Naoya Shibata. Theoretical framework of statistical noise in scanning transmission electron microscopy. *Ultramicroscopy*, 193:118–125, October 2018.

[20] Manuel Hüpfel, Andrei Yu Kobitski, Weichun Zhang, and G. Ulrich Nienhaus. Wavelet-based background and noise subtraction for fluorescence microscopy images. *Biomedical Optics Express*, 12(2):969–980, February 2021. Publisher: Optica Publishing Group.

[21] R. Rana, V. Singh, A. Jain, D.R. Bednarek, and S. Rudin. Anti-scatter grid artifact elimination for high resolution x-ray imaging CMOS detectors. *Proceedings of SPIE–the International Society for Optical Engineering*, 9412:941243, 2015.

[22] Xiaobin Wu, Liangliang Zheng, Chunyu Liu, Tan Gao, Ziyu Zhang, and Biao Yang. Single-Image Simultaneous Destriping and Denoising: Double Low-Rank Property. *Remote Sensing*, 15(24):5710, December 2023. Publisher: Multidisciplinary Digital Publishing Institute.

[23] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[24] Abid Khan, Chia-Hao Lee, Pinshane Y. Huang, and Bryan K. Clark. Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images. *npj Computational Materials*, 9(1):85, May 2023.

[25] Bastian Rühle, Julian Frederic Krumrey, and Vasile-Dan Hodoroaba. Workflow towards automated segmentation of agglomerated, non-spherical particles from electron microscopy images using artificial neural networks. *Scientific Reports*, 11(1):4942, March 2021.

[26] Trushal Sardhara, Christian J. Cyron, Martin Ritter, and Roland Aydin. Generating ideal synthetic data for 3D reconstruction of FIB tomography data using generative adversarial networks. In *AI for Accelerated Materials Design - NeurIPS 2024*, November 2024.

[27] Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a Master of Steganography. In *NIPS 2017 Workshop on Machine Deception*, Long Beach, CA, USA, December 2017. _eprint: 1712.02950.

[28] Jonas Utz, Tobias Weise, Maja Schlereth, Fabian Wagner, Mareike Thies, Mingxuan Gu, Stefan Uderhardt, and Katharina Breininger. Focus on Content not Noise: Improving Image Generation for Nuclei Segmentation by Suppressing Steganography in CycleGAN. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3858–3866, 2023.

[29] Fabian Coupette, Long Zhang, Björn Kuttich, Andrei Chumakov, Stephan V. Roth, Lola González-García, Tobias Kraus, and Tanja Schilling. Percolation of rigid fractal carbon black aggregates. *The Journal of Chemical Physics*, 155(12):124902, September 2021.

[30] G. Beaucage. Approximations Leading to a Unified Exponential/Power-Law Approach to Small-Angle Scattering. *Journal of Applied Crystallography*, 28(6):717–728, December 1995. Publisher: International Union of Crystallography.

[31] Xingshuai Zheng, Tengfei Sun, Jixing Zhou, Rupeng Zhang, and Pingmei Ming. Modeling of Polycrystalline Material Microstructure with 3D Grain Boundary Based on Laguerre–Voronoi Tessellation. *Materials*, 15(6):1996, March 2022.

[32] J. C. Halpin and J. L. Kardos. The Halpin-Tsai equations: A review. *Polymer Engineering & Science*, 16(5):344–352, 1976. _eprint: https://4spepublications.onlinelibrary.wiley.com/doi/pdf/10.1002/pen.760160512.

[33] Ian J. 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, December 2014.

[34] Kevin de Haan, Zachary S. Ballard, Yair Rivenson, Yichen Wu, and Aydogan Ozcan. Resolution enhancement in scanning electron microscopy using deep learning. *Scientific Reports*, 9(1):12050, August 2019.

[35] Jeffrey M. Ede and Richard Beanland. Partial Scanning Transmission Electron Microscopy with Deep Learning. *Scientific Reports*, 10(1):8332, May 2020.

[36] Masato Suzuki and Yasuhiko Igarashi. Application of a Hessian-Based Image-Processing Method for Enhanced Visualization of Nanoscale Rubber Cross-Linked Network Structures from Electron Microscopy Images. *ACS Applied Nano Materials*, 8(1):112–120, January 2025.

[37] Masato Suzuki and Yasuhiko Igarashi. Domain-Specific Simulated Data Enhances Knife-Mark Noise Suppression in Microscopy Images of Materials. *Microscopy*, 00(00):1–14, 2025. in press.

[38] James P. Horwath, Dmitri N. Zakharov, Rémi Mégret, and Eric A. Stach. Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images. *npj Computational Materials*, 6(1):108, July 2020. Publisher: Nature Publishing Group.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 234–241. Cham, May 2015. Series Title: Lecture Notes in Computer Science.

A Technical Appendices and Supplementary Material

A.1 Compute Resources

We report the hardware and operating system used for all experiments to support reproducibility.

- **Operating System:** Microsoft Windows 11 Pro
- **Workstation Model:** Dell Precision 3680
- **CPU:** Intel Core i9-14900K, 24 cores
- **Memory:** 64 GB RAM
- **GPU:** NVIDIA RTX 4500 Ada Generation, 24 GB VRAM
- **Typical training time:** Training a CycleGAN model for 20 epochs took approximately 30 min on the RTX 4500 GPU.

A.2 Hyperparameter Details

Common setup. We enable mixed precision (`mixed_float16`) and use TensorFlow Addons InstanceNormalization in the generators and discriminators.

Table 1: Shared configuration across experiments.

Item	Value
Image size	256×256 (grayscale, single channel)
Augmentation (real images)	random crop=256, rotations $\{0, 90, 180, 270\}$,
Normalization for GAN	map uint8 $[0, 255] \to [-1, 1]$
Normalization for U-Net	map to $[-1, 1]$

Table 2: CycleGAN (training; dataset-dependent).

Item	Value
Training images (per domain)	*dataset-dependent* (Sim: 1000 or 2000 generated; Real: augmented to match)
Epochs / Batch size	*dataset-dependent* (e.g., rubber: 50)
Optimizer / LR	Adam (β_1=0.5); LinearDecay from $1 - 2 \times 10^{-4}$
Loss (generator)	Adv (MSE; LSGAN) + Cycle (L1) + Identity (L1)
	+ FFT amplitude (L1) [+ Brightness (L1)], weights: dataset-dependent
`gan_dim` / `n_blocks`	32 / 6
Normalization / Aug.	$[-1, 1]$ norm; random jitter+crop+flip

Table 3: U-Net (training; dataset-dependent).

Item	Value
Epochs / Batch size	20 / 32
Optimizer / LR	Adam / 1×10^{-4}
Loss / Metric	binary cross-entropy / accuracy
Output activation	sigmoid
Input normalization	map to $[-1, 1]$
Image size	256×256 (grayscale, single channel)

Table 4: Comparison of segmentation performance (IoU) under different training conditions (mean and standard deviation across $N = 196$ test images; gold session).

Training condition	IoU (mean)	Std.
Otsu binarization	0.027	0.011
U-Net (Synthetic-only (no adaptation))	0.880	0.013
U-Net (CycleGAN-adapted synthetic)	0.770	0.018
U-Net (Proposed pipeline)	0.884	0.011
U-Net (Human annotation baseline)	0.931	0.013

Table 5: Comparison of segmentation performance (Dice) under different training conditions (mean and standard deviation across $N = 196$ test images; gold session).

Training condition	Dice (mean)	Std.
Otsu binarization	0.053	0.021
U-Net (Synthetic-only (no adaptation))	0.936	0.007
U-Net (CycleGAN-adapted synthetic)	0.870	0.011
U-Net (Proposed pipeline)	0.938	0.006
U-Net (Human annotation baseline)	0.964	0.007

A.3 Evaluation for Table 4, 5 (mean & std)

We report IoU (and Dice) means and standard deviations computed *across test images* for each training condition. Test images and masks are resized to 256×256. For U-Net[39], inputs are per-image standardized; Otsu uses OpenCV's global threshold. For each model, we predict on the test set, compute per-image IoU/Dice, and aggregate mean and std. All values are the results reported in Sec. 3.2 (gold session; $N = 196$ images).

Overview of evaluated methods. We evaluated the models based on U-Net[39]. U-Net is a segmentation architecture originally proposed for biomedical image analysis and has since been widely used across various segmentation tasks. In our setting, the model is trained on pairs of synthetic images and ground-truth masks, and then directly applied to predict particle regions in experimental images. To ensure a fair comparison, only the training dataset was varied across methods, while the network architecture, hyperparameters, and training procedures were kept identical.

(i) **Synthetic-only (no adaptation):** the U-Net is trained on pairs of synthetic images and ground-truth masks, then directly applied to predict particle regions in experimental images. (ii) **CycleGAN-adapted synthetic:** synthetic clean images without noise are first transformed into experiment-like style by a CycleGAN; the adapted images and corresponding masks are then used to train a U-Net, which is subsequently evaluated on experimental inputs. (iii) **Proposed pipeline:** training data are generated by our simulation– noise injection–CycleGAN pipeline, yielding paired images and masks that reflect realistic imaging conditions; a U-Net trained on this dataset is used to segment experimental images. (iv) **Human annotation baseline:** the U-Net is trained on manually annotated experimental images and masks, and tested on experimental inputs.

A.4 Rubber (Knife-mark Noise) Dataset: Acquisition & Preprocessing

Sample preparation. The base is an SBR compound (phr: SBR 100, HAF carbon black 61, ZnO 3, S 1.4, CBS 1.7, DPG 1.5). Sheets were vulcanized at 160 °C for 20 min and cut into 3×3 cm specimens with a thickness of 2–3 mm. Knife-mark stripes produced during cutting are intentionally retained in the real images.

Imaging (real images). Optical microscopy at $100\times$ magnification; effective pixel size $\approx 0.8\,\mu$m/px. For learning/evaluation, images were converted to 8-bit grayscale and cropped to 256×256 px.

Simulated data (training/evaluation). All simulated images are 256×256 px. Filler agglomerates are modeled as disks whose radii follow a power-law distribution. Based on the measured maximum agglomerate radius of 19 px in real images, the simulation upper bound was set to 28.5 px ($= 19 \times 1.5$) to define the *ground-truth* masks. Knife-mark noise is synthesized by superimposing multiple straight lines of width 1–3 px at random angles θ, with randomness in density, thickness, and slant.

A.5 Rubber (TEM/HAADF-STEM) Dataset: Acquisition & Preprocessing

Sample preparation. Cross-linked isoprene rubber (IR) was compounded with ZnO, sulfur (soluble or insoluble), and various accelerators (CBS, MBTS, DPG, TMTD, HMTA) to prepare eight compositions with different cross-link densities (see Table 1 for formulations and properties). Samples were vulcanized at 160 °C for 30 min, then cut from the vulcanizates, swollen in styrene, embedded, trimmed, and sectioned into ultrathin slices with an ultramicrotome. Sections were vapor-stained with OsO_4.

Imaging (real EM data). High-angle annular dark-field scanning TEM (HAADF-STEM) was performed at 200 kV and 60,000$\times$ magnification. Images used for analysis were recorded at 1024×1024 px with an effective sampling of 1.52 nm/px (field of view $\sim$1.56 μm $\times$ 1.56 μm) 14 images.

A.6 Code Availability

All codes and data generation scripts have been released at: `https://github.com/fanfanfuzzy/Noise2CycleGAN-Benchmark`

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 follow 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:Justification: The main claims in the abstract and introduction are supported by results (Sec. 2–4, Fig. 1–3, Table 1).

   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:Limitations such as irregular geometries and unseen imaging conditions are discussed (Sec. 4).

   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:The paper contains no formal theoretical results or proofs.

   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: Methods, datasets, and evaluation metrics are described in detail (Sec. 2–3, Table 1).

   Guidelines:

   - 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: [Yes]

   Justification: To ensure reproducibility, we will release part of the dataset generation code before the final camera-ready version. TEM dataset are taken from publicly available datasets.

   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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:We specify datasets and splits, preprocessing/augmentation (e.g., 256×256, per-image standardization), model/loss/optimizer and schedule (epochs, batch, LR), and the evaluation protocol in Sec. 2–3; consolidated hyperparameter tables are provided in the Appendix (CycleGAN/U-Net; dataset-dependent settings).

   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: [Yes]

   Justification: We report mean±std for IoU (and Dice) across N=196 test images (Sec. 3.2; Table 4 5); no formal hypothesis tests are performed.

   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 Appendix (Compute Resources) specifies OS/CPU/RAM/GPU–VRAM and a typical training time.

   Guidelines:

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

- 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: The work uses public data and simulations without ethical risks.

   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: Positive impacts on reducing annotation cost are discussed; potential misuse risks are limited.

    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: No high-risk pretrained models or datasets are released.

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: Existing datasets such as Horwath et al. TEM are properly cited.

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: No new datasets or code are released.

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**

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: No human subjects or crowdsourcing were involved.

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: Human subjects research was not conducted.

    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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were not part of the core methodology, only for writing assistance.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.