CONTRASTIVE IMAGE-METADATA PRE-TRAINING FOR MATERIALS TRANSMISSION ELECTRON MICROSCOPY

Anonymous authors
Paper under double-blind review

ABSTRACT

The transmission electron microscope facilitates the highest-resolution imaging of any instrument ever created, and its limiting factor is no longer spatial resolution but dose efficiency. Low electron doses avoid sample damage but produce noisy images for which, unlike in classical computer vision, there is no ground truth. Autonomous materials experimentation poses a related problem, since closed-loop instruments need representations grounded in the microscope state at acquisition. Both demand representations grounded in how an image was acquired. We release 7,330 paired high-angle annular dark-field scanning-TEM (HAADF-STEM) images and their seven-dimensional acquisition metadata, and propose Contrastive Image-Metadata Pre-training (CIMP), a CLIP-style encoder that aligns the two modalities and reaches 84.4% Top-1 cross-modal retrieval on a held-out split. All seven parameters are individually recoverable from the frozen visual embedding through a linear probe, and we use the embedding to condition a metadata-conditioned style-transfer model that re-renders experimental images under different acquisition parameters. Retargeting low-dose images to higher-dose acquisition metadata turns this model into a physics-informed denoiser; in a blind user study, experimental microscopists prefer it over the current state-of-the-art denoiser for STEM imagery on 70.2% of trials.

1 INTRODUCTION

The transmission electron microscope facilitates the highest resolution imaging of any instrument ever created, with resolutions recorded of sub-0.5 ångströms (or 5×10^{-11} meters) (Erni et al., 2009; Sawada et al., 2015), sufficient to resolve individual atomic columns in most crystal lattices (Krivanek et al., 2015). Hence, the limiting factor of TEM is no longer spatial resolution, but instead dose efficiency, or minimising the number of electrons that hit the sample while retaining an interpretable signal. Limiting the electron dose is crucial to avoid sample damage and ensure physical relevance of observed phenomena, but naturally leads to very noisy images (Egerton, 2025; Reimer, 1984). Unlike in classical computer vision, there is no ground truth for these images, since acquiring a clean reference would itself degrade the sample. While there is optimism that machine learning workflows could mitigate the difficulties associated with low-dose imaging, a lack of open-source data means that current methods are typically trained from scratch for a specific noise level on a single or narrow set of samples. Beyond denoising specifically, the broader trajectory of materials-science instrumentation is towards autonomous scientific experimentation, in which microscopes operate in closed loops that collect, interpret, and respond to their own measurements with reduced human supervision (Abolhasani et al., 2026). The TEM's complexity, with many physically meaningful parameters tuned per acquisition, means that any such autonomous system needs a representation of each image grounded in the instrument state under which it was acquired. The same kind of representation also enhances interactive use, supplying operators with real-time tools that reason over the microscope's state during an experiment. In both modes, the missing piece is an encoder that aligns visual content with the physical parameters of the instrument.

Yet, for every published image, a massive volume of data, more than 90% by some estimates (Ivanova et al., 2025), remains unused, not because the quality of the data is bad but because TEM experiments require exploration and tuning to identify and record the golden data that eventually makes publication. Over time, TEM labs accumulate large volumes of leftover data from different users, materials samples, operating modes, and acquisition settings, and this data is rich in detail and variation, often

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

paired with automatically saved metadata that describe the state of the microscope at acquisition. Unfortunately, this data is almost never shared as the culture of data sharing is not established in the electron microscopy community (Ivanova et al., 2025).

In this work, we address these problems through four contributions. First, we release a paired image-metadata dataset of 7,330 high-angle annular dark-field scanning-TEM (HAADF-STEM) electron microscopy images, each associated with seven acquisition-metadata parameters: pixel size, dwell time, beam convergence angle, beam current, detector gain, detector offset, and inner collection angle (Figure 1). Second, we propose Contrastive Image-Metadata Pre-training (CIMP), a CLIP-style method (Radford et al., 2021) that learns a joint embedding space over HAADF-STEM images and their acquisition metadata, reaching Top-1 cross-modal retrieval of 84.4% on a held-out validation split, and we treat the resulting embeddings as a foundation layer for a variety of potential TEM-ML workflows rather than as infrastructure for any single downstream task. Third, linear probing shows that all seven acquisition parameters are individually recoverable from the frozen 128-d visual embedding through Ridge regression despite the contrastive objective never supervising any metadata value directly, and we use the embedding as both a perceptual loss and a conditioning signal for a generative metadata-conditioned style-transfer model that translates an experimental image into the style it would have had under different acquisition parameters. Finally, we use the style-transfer model as a physics-informed denoiser by retargeting low-dose images to higher-dose acquisition metadata. In a blind user study, experimental microscopists prefer this denoiser over the Thornley et al. baseline on 70.2% of trials.

2 Related Work

Open-Source STEM Data While petabytes of HAADF-STEM images are collected each year (Ivanova et al., 2025), very few are ever open-sourced. Microscopy publications typically focus on findings from the highest-quality images, and the culture of data sharing is not well-established in the field. Some larger experimental and synthetic datasets have nevertheless been released. The closest experimental precedent is from Sytwu et al. (2023), who released 407 HRTEM images of Au, Ag, and CdSe nanoparticles paired with acquisition metadata including magnification and electron dose. This release was much smaller in scale and in HRTEM rather than HAADF-STEM. Other large-scale open releases are simulated. Atomagined (Schwenker et al., 2020) provides PRISM-generated HAADF-STEM images post-processed to emulate noise and distortion, covering a wide range of ICSD structure prototypes. Eliasson & Erni (2025); Eliasson (2025); Eliasson et al. (2026) release a 250,000-image multislice-simulated HAADF-STEM corpus of Pt nanoparticles for size and morphology prediction. TEMImageNet (Lin et al., 2021) offers a simulated training library for atom-resolution STEM tasks and shows that simulation-trained models can transfer to experimental data. None of these match the scale of pairing between experimental HAADF-STEM images and rich instrument metadata that we target. Existing experimental datasets tend to be small, narrow in material coverage, and rarely include the acquisition metadata needed to train models that generalise across instrument settings.

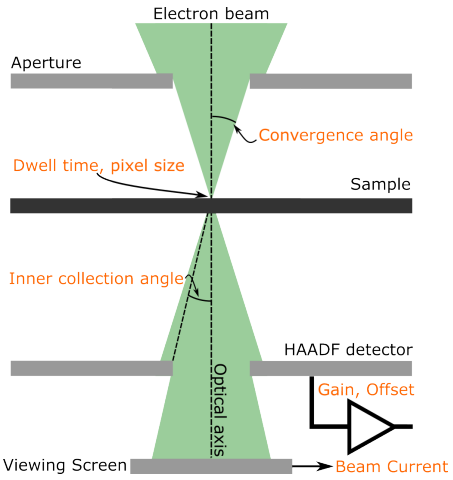


Figure 1: Diagram of a STEM with our acquisition parameters shown in orange text. Description of each acquisition parameter is provided in Appendix A.

Embedding Models CIMP adapts the CLIP image–text pretraining framework (Radford et al., 2021) to pairs of HAADF-STEM images and their acquisition metadata. The closest structural analogue in any microscopy domain is CLOOME (Sanchez-Fernandez et al., 2023), which contrastively aligns fluorescence cell-painting images with chemical-structure embeddings and demonstrates cross-modal retrieval across a large bioimaging database. ContIG (Taleb et al., 2022) is the closest conceptual parallel: a medical-imaging model that contrastively aligns retinal images with a structured

physical-annotation vector (SNP arrays, polygenic risk scores). To our knowledge, however, no prior work applies this or any similar framework to electron microscopy (EM).

Image-to-Image Translation in Electron Microscopy Prior work has applied image-to-image networks to EM data in two main directions: to bridge the simulation-to-experiment gap so that synthetic training data gains realism, and for denoising of experimental acquisitions. To close the simulation-to-experiment gap, Khan et al. (2023) use a CycleGAN (Zhu et al., 2017) with a reciprocal-space discriminator to add realistic high-frequency content to simulated STEM images so that a downstream segmentation network can transfer to real experiments. Eliasson et al. (2024) use a CycleGAN to translate between simulated and experimental HAADF-STEM of Pt nanoparticles, showcasing both denoising and realistic noise addition. And in another work, Lobato et al. (2024) train a GAN denoiser on synthetic image pairs with intricately modeled noise. These approaches are all one-to-one and agnostic to microscope settings (i.e., a dedicated network is trained per source/target domain pair, with no natural way to parameterise the target style). Simulating a different dwell time or detector gain with any previous method requires training a new CycleGAN with a new paired or unpaired corpus at that setting. The style-transfer network we present downstream of CIMP is instead one-to-many: a single generator is conditioned on continuous acquisition metadata through CIMP embeddings such that any target instrument setting can be queried at inference without retraining.

STEM Denoisers Microscopists are always pushing the limit of their detector in order to minimize the electron dose. With those ever diminishing electron doses, the need for appropriate denoisers to make low-dose data interpretable and easy to use, is constant. Self-supervised denoisers in the Noise2Noise / Noise2Void / UDVD family (Lehtinen et al., 2018; Krull et al., 2019; Sheth et al., 2021) are widely used for STEM imagery, but they are typically specialised to one noise profile, material, or acquisition sequence; Noise2Noise-style methods further require true time-series pairs whose content is unchanged. Case-specific self-supervised EM methods such as SHINE (Kim et al., 2025) similarly train or adapt to a target raw dataset whose noise statistics can be estimated directly. The most recent STEM-specific state of the art is the HAADF-STEM Noise2Void2 method of Thornley et al. (2026), which improves blind-spot denoising for atomic-resolution STEM and is therefore our primary learned baseline. GAN-based denoisers have also been explored: Noise2Atom (Wang et al., 2020) is an unsupervised GAN denoiser trained on noisy acquisitions of sub-nm catalytic clusters, and is, to our knowledge, the only STEM denoiser to release its training data. In contrast to these case- or domain-specific denoisers, CIMP is trained once on a heterogeneous archive and conditions on acquisition metadata at inference time, aiming to generalise across materials, fields of view, and microscope settings. As far as the authors are aware, there exist no prior *general* STEM denoisers that use acquisition settings and pixel size to produce denoised outputs across instrument settings.

3 Data Contribution

In this work, we introduce 7,330 HAADF-STEM images with paired metadata from a Titan microscope at an anonymised materials-research laboratory. The dataset covers more than 30 distinct materials systems, predominantly heterogeneous catalysts (Pt/CeO_2, Pd/ZrO_2, Ru/TiO_2, In_2O_3/ZrO_2, etc.), with more than 130 material samples synthesized by different techniques and with different metal loading, imaged with a variety of instrument settings. The full size images have variable sizes between 256×256 and 4096×4096, with 83% at 2048×2048. We show the distribution of metadata parameters in our dataset in Figure 4 in Appendix A, as well as a description of each of the metadata variables (also in Appendix A). We additionally solicited the creation of several *evaluation sets*, which are series of true experimental images captured of the same gold-on-carbon sample in quick succession while varying a single metadata parameter. These held-out series sweep dwell time, gain, and offset respectively. We exclude them from training and use them only for evaluation, including the qualitative sweeps in Figure 5 and the paired quantitative comparisons below. Outside of these sets, images were captured as the operator saw fit, so the corpus contains no paired or matched groupings, only a broad collection of samples spanning varied acquisition parameters. We will release this data (along with our models and code) with a `CC-by-4.0` license once the blind review period is concluded.

4 CONTRASTIVE IMAGE-METADATA PRETRAINING (CIMP)

We train a CLIP-style contrastive image-metadata embedding model on our dataset of 7,330 HAADF-STEM images and their paired metadata. We evaluate the performance of our embedding model, CIMP, by measuring retrieval accuracy at *Top-1*, *Top-5*, and *Top-10*. *Top-1* denotes the percentage of queries for which the correct match is ranked first among the retrieved results. *Top-5* and *Top-10* indicate the percentage of queries for which the correct match appears within the top 5 and top 10 results, respectively.

For training, each HAADF-STEM image is normalised from 16-bit unsigned integers to a floating-point tensor in $[0, 1]$ and a single random 256×256 crop is extracted per image per epoch, augmented with a horizontal flip (probability 0.5) and a random multiple-of-90° rotation. The accompanying metadata vector contains seven acquisition parameters. The four dimensions that span several orders of magnitude (pixel size, dwell time, convergence angle, beam current) are stored in $\log_{10}$ space, and all seven dimensions are z-scored using statistics computed across the training corpus before being passed to the metadata encoder.

To train CIMP, we split our dataset into training and validation splits of size 6,597 and 733, respectively, with a fixed seed. The split is stratified at the full-size image level such that all crops drawn from a given source image are assigned to the same split. Thus, retrieval metrics measure generalisation to unseen acquisitions rather than memorisation of within-image content (see Appendix B for details). We sample 1 random crop per image per training epoch, which yields effectively unlimited unique patches across our training.

CIMP is trained with the symmetric InfoNCE objective (van den Oord et al., 2018) as used in CLIP (Radford et al., 2021). For a batch of B paired samples, the visual encoder maps each image to an L_2-normalised embedding $f_i \in \mathbb{R}^{128}$ and the metadata encoder maps each metadata vector to an L_2-normalised embedding $g_j \in \mathbb{R}^{128}$. With pairwise cosine similarity $s_{ij} = f_i^\top g_j$, the loss is

$$\mathcal{L}_{\text{CIMP}} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\log \frac{\exp(\tau s_{ii})}{\sum_{j=1}^{B} \exp(\tau s_{ij})} + \log \frac{\exp(\tau s_{ii})}{\sum_{j=1}^{B} \exp(\tau s_{ji})} \right), \tag{1}$$

the average of two cross-entropies that pull each paired (f_i, g_i) together while pushing all unpaired samples in the batch apart. The temperature τ is learnable and initialised at 0.1.

To find the best architecture and hyper-parameters for CIMP, we compare four image-encoder backbones (ResNet-18 (He et al., 2016), ViT (Dosovitskiy et al., 2021) trained from scratch, ImageNet-pretrained ViT finetuned end-to-end, and ImageNet-pretrained ViT with the backbone frozen), four batch sizes ($128,\ 256,\ 512,\ 1024$), five metadata encoder sizes ($64, 128,\ 256,\ 512,\ 1024$), and four crop sizes ($128,\ 256,\ 512,\ 1024$). The results and discussion of all these studies is reported in Appendix D.

We find the best performance in and adopt going forward a ResNet-18 model trained from scratch with a single-channel input with a crop size of 256×256, a batch size of 512, and a 3-layer MLP metadata encoder of width 256. We train for 1000 epochs and achieve final Top-1 of 0.8438, Top-5 of 0.9688, and Top-10 of 0.9844 on our held-out validation set. Notably, an ImageNet-pretrained ViT finetuned end-to-end on our HAADF-STEM data (Top-1 82.8%) does not outperform our ResNet-18 trained from scratch (Top-1 84.4%) on this task (Appendix D), which conveys the need for HAADF-STEM image priors given the difference in distributions between natural images and atomic-scale images.

5 STYLE TRANSFER, DENOISING, AND METADATA RECOVERY

We evaluate CIMP in three stages, each reported as its own subsection below. We first characterise the learned representation by asking how much of the raw acquisition metadata can be recovered from the frozen visual embedding alone through linear probing. Then, we use the metadata embeddings as a conditioning signal and perceptual loss for a metadata-conditioned style-transfer model. As discussed in Section 1, noise in HAADF-STEM images increases with shorter dwell time and lower beam current, often used to preserve sample integrity, as fewer electrons impinge on the detector at each raster position of the STEM scan. To use our style-transfer model as a physics-informed denoiser, we retarget low-dose images to higher-dose acquisition metadata. We compare the fidelity

of our style transfer model to experimentally varying the acquisition parameters with our evaluation set. We evaluate its performance as a denoiser by comparing its output quantitatively and qualitatively against the CycleGAN-based Noise2Atom denoiser from Wang et al. (2020) and the state-of-the-art HAADF-STEM denoiser from Thornley et al. (2026), and we additionally validate it through a blind user study with experimental microscopists.

5.1 Metadata Recovery

For CIMP embeddings to serve as a foundation layer that downstream models can condition on, the representation must encode physically meaningful axes of microscope state, not just enough information to re-pair an image with its own metadata at retrieval time. To probe this, we extract 128-d visual embeddings from the frozen CIMP encoder for every image in the train and validation splits (6,597 / 733 images, the same split as CIMP pre-training). We then fit seven independent Ridge regressors ($\alpha = 1.0$) on the training embeddings, one per metadata dimension, each predicting a single z-scored target value. We report per-dimension R^2 and symmetric mean absolute percentage error (SMAPE) on the held-out validation split. R^2 measures explained variance relative to a constant mean predictor: 1 is perfect prediction, 0 matches the mean baseline, and negative values are worse than predicting the mean. SMAPE is a bounded percentage error in $[0, 200]\%$, where lower is better and 0% is exact recovery. We compute SMAPE in physical units after inverting both the z-score and the $\log_{10}$ transform applied during preprocessing. Since Ridge has a closed-form solution, the probe measures how much metadata information is linearly present in the embedding without additional nonlinear decoding capacity.

All seven metadata parameters are linearly decodable from the frozen 128-d embedding (Table 1). Gain, offset, inner-collection angle, and convergence angle are recovered with 5–13% SMAPE, ranging from detector gain (5.3%) to convergence angle (13.0%). Pixel size, dwell time, and beam current span wider dynamic ranges and are evaluated after inverting the $\log_{10}$ transform, so their SMAPE values are larger, from dwell time (28.4%) to pixel size (41.8%), while their R^2 values remain positive (0.65–0.78).

Table 1: Per-dimension linear-probe recovery from the frozen CIMP visual embedding. Values are mean ± standard error across 2000 bootstrap resamples; the Mean row reports a point estimate. Dimensions marked † are stored in $\log_{10}$ space.

Dimension	$R^2 \uparrow$	SMAPE↓
Pixel size†	0.72 ± 0.03	$41.8 \pm 1.2\%$
Dwell time†	0.78 ± 0.01	$28.4 \pm 0.9\%$
Convergence angle†	0.53 ± 0.04	$13.0 \pm 0.4\%$
Beam current†	0.65 ± 0.03	$34.6 \pm 1.1\%$
Gain	0.84 ± 0.01	$5.3 \pm 0.2\%$
Offset	0.80 ± 0.05	$9.2 \pm 1.2\%$
Inner coll. angle	0.49 ± 0.05	$9.5 \pm 0.4\%$
Mean	0.69	20.3%

The contrastive objective we used never supervises any metadata value directly; it only asks visual embeddings to align with the MLP projection of the metadata vector. Yet seven physically distinct acquisition parameters are independently recoverable from the frozen 128-d embedding through a closed-form linear probe, evidence that contrastive image-metadata alignment yields a factored, physics-aligned representation as a byproduct of retrieval training. Appendix E reports an additional offset-withholding experiment, where offset remains recoverable from the visual embedding even when omitted from the metadata objective during CIMP pretraining. To our knowledge, this is the first demonstration of such a factored representation for electron microscopy, and it is what makes CIMP embeddings useful as a conditioning signal beyond the retrieval task itself.

5.2 Style Transfer

In practice, the most useful denoiser for an experimental microscopist is one that runs while they are at the microscope, quicker than the frame rate of the acquisition, so that low-dose images can be interpreted in time to inform the next decision. This rules out methods whose inference cost is dominated by iterative refinement. We additionally cannot rely on paired source-target data, since acquiring matched low-dose and high-dose images of the same field would itself degrade the sample, and our training corpus of ∼7,000 images is small by computer-vision standards. These constraints

motivate a GAN-based architecture. GANs admit one-shot inference without iterative ODE solving, train reliably on small datasets, integrate naturally with the multi-task LPIPS-based objective we use, and allow us to jointly optimise for the correct metadata style, content preservation, and image realism without true image-metadata pairs. By contrast, Palette-style image-to-image diffusion (Saharia et al., 2021) requires paired source-target examples that we do not have, and flow matching (Lipman et al., 2022) is natively a noise-to-image generative method rather than an image-to-image translator that can be conditioned on metadata.

Our style transfer model takes an input image x along with a source embedding e_{id} of the image's original metadata and a target embedding e_{tgt} of the shuffled metadata we want the output to match. Both are produced by the frozen CIMP metadata encoder from raw acquisition-metadata vectors. The model then produces an image that preserves the content of x but matches the style associated with e_{tgt}. We implement this with a U-Net (Ronneberger et al., 2015) generator G whose convolutional blocks are conditioned on the embeddings via FiLM layers (Perez et al., 2018). This allows the metadata signal to modulate features at every scale, while the U-Net skip connections carry content information through unchanged.

We train G against four objectives. The adversarial loss $\mathcal{L}_{\text{adv}}$, implemented with the non-saturating binary-cross-entropy GAN objective, drives outputs toward the distribution of real micrographs. The remaining three components are defined as:

$$\mathcal{L}_{\text{cyc}} = \left\| G(G(x, [e_{\text{tgt}}; e_{\text{id}}]), [e_{\text{id}}; e_{\text{tgt}}]) - x \right\|_1, \tag{2}$$

$$\mathcal{L}_{\text{id}} = \text{LPIPS}\big(G(x, [e_{\text{id}}; e_{\text{id}}]),\ x\big), \tag{3}$$

$$\mathcal{L}_{\text{emb}} = \left\| e_{\text{tgt}} - \text{CIMP}_{\text{vis}}(G(x, [e_{\text{tgt}}; e_{\text{id}}])) \right\|_2^2, \tag{4}$$

$$\mathcal{L}_{\text{GAN_total}} = \mathcal{L}_{\text{adv}} + \lambda_1 \mathcal{L}_{\text{cyc}} + \lambda_2(t) \mathcal{L}_{\text{id}} + \lambda_3 \mathcal{L}_{\text{emb}}, \tag{5}$$

where $\mathcal{L}_{\text{cyc}}$ enforces content preservation by requiring a round trip (source to target, then target back to source) to recover x, following the cycle-consistency idea of Zhu et al. (2017) with an L_1 distance; $\mathcal{L}_{\text{id}}$ discourages G from modifying images whose target matches their source, using LPIPS (Zhang et al., 2018) as the distance metric; and $\mathcal{L}_{\text{emb}}$ anchors the output to e_{tgt} in the CIMP visual-embedding space, ensuring the generated image matches the target style. We set $\lambda_1 = 0.4$ and $\lambda_3 = 1.0$, and linearly decay $\lambda_2(t)$ from 2.0 to 0.1 over the first 10 epochs. We provide ablations over these parameters in Appendix J.

Table 2: Pixel-level fidelity on the three held-out evaluation series, averaged over all ordered pairs (x_i, x_j) within each series. CIMP GAN rows compare $G(x_{\text{src}}, e_{\text{tgt}}, e_{\text{id}})$ to the real reference x_j; Eval series rows compare the unmodified x_i to x_j.

Series	**Model**	SSIM↑	PSNR↑	LPIPS↓
Dwell	CIMP GAN	0.368	23.63	0.133
	Eval series	0.298	23.38	0.192
Gain	CIMP GAN	0.237	16.98	0.203
	Eval series	0.108	11.40	0.344
Offset	CIMP GAN	0.271	13.20	0.312
	Eval series	0.153	10.59	0.362
Overall	CIMP GAN	0.302	19.00	0.198
	Eval series	0.202	16.45	0.282

Table 3: Kernel Inception Distance (KID) between four pairs of image sets. Subset size 21 for the first three rows (bounded by the eval-set size) and 50 for the noise floor; the last three rows average across 5 disjoint training pools of 500 images with pinned seeds.

Comparison	**KID↓**
CIMP GAN vs. eval set	-0.015 ± 0.011
CIMP GAN vs. training	0.154 ± 0.006
Eval set vs. training (content floor)	0.171 ± 0.011
Training vs. training (noise floor)	0.0005 ± 0.0009

To verify that the generator follows its conditioning signal, we shuffle validation metadata, generate target-style images, and compare their CIMP visual embeddings with the target metadata embeddings. Correct image–metadata pairs exceed incorrect pairs by 0.531 cosine similarity (Appendix I), which confirms that the generator responds to the conditioning signal.

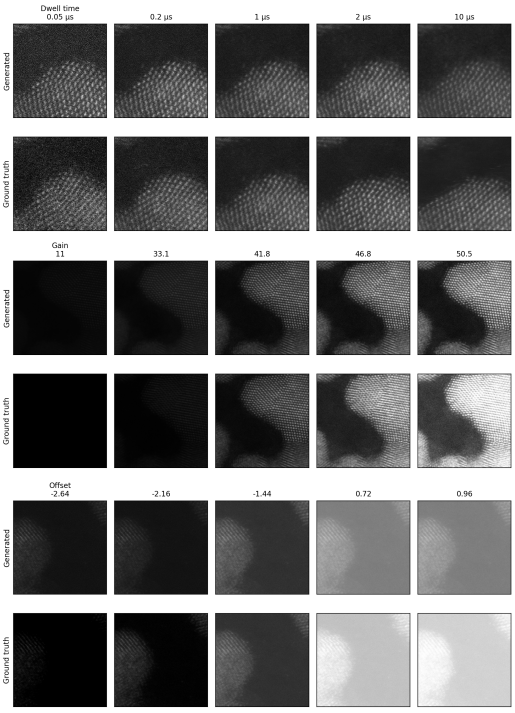


Figure 2: Virtual CIMP GAN sweeps (top rows) compared with experimental sweeps (bottom rows) for dwell time, gain, and offset. Scale bars are 2 nm. An enlarged version with full captions is shown in Appendix Figure 6.

qualitative outputs are provided in Appendix C.

We also evaluate target fidelity on held-out gold-on-carbon sweeps for dwell time, gain, and offset. For every ordered pair (x_i, x_j) within a sweep, CIMP GAN receives x_i and the metadata of x_j, and its output is compared with the real target image x_j. The eval-series baseline compares x_i directly with x_j with no model in the loop. Table 2 shows that CIMP GAN improves all three metrics over this baseline for all sweeps, which improves SSIM from 0.202 to 0.302, PSNR from 16.45 to 19.00, and LPIPS from 0.282 to 0.198.

Finally, we test distributional realism with Kernel Inception Distance (KID) (Bińkowski et al., 2018). CIMP GAN outputs are nearly indistinguishable from the real eval-set images (-0.015 ± 0.011), while their distance to the broader training corpus (0.154) is comparable to the eval-set content gap (0.171). Together, Tables 2 and 3 indicate that the generator matches both the target images and the narrow eval-series distribution.

Figure 2 compares virtual and experimental sweeps for dwell time, gain, and offset. CIMP GAN captures the main monotonic trends, especially for gain, but does not fully reproduce extreme detector boundary conditions such as clipping. Pixel size is excluded from qualitative style-transfer sweeps because changing it while keeping the image dimensions fixed has no physical basis (Appendix A). Additional per-parameter

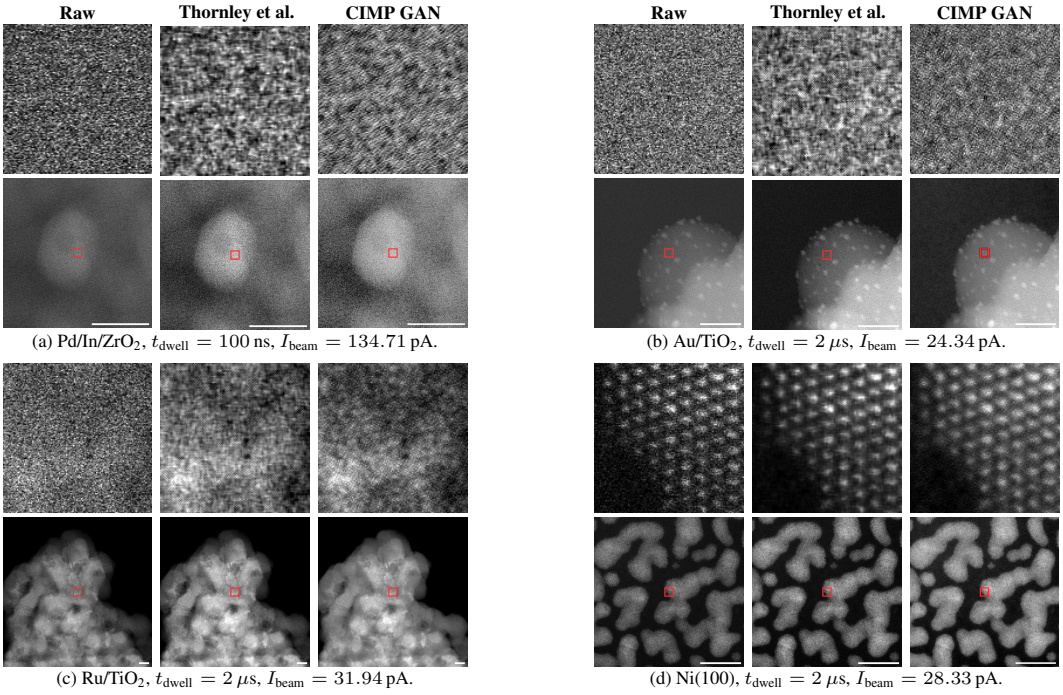
5.3 Denoising

Finally, we utilize the CIMP-conditioned GAN model as a denoiser by feeding it a target embedding e_{tgt} derived from higher-dose acquisition metadata. The qualitative examples in Figure 3, the reference-free quantitative metrics below, and the user study all use fixed high-dose target dwell time and beam current. We compare our model to the Noise2Void-based denoiser of Thornley et al. (2026) and to the GAN-based Noise2Atom denoiser (Wang et al., 2020), both retrained on our dataset for a best performance comparison.

Qualitatively, our CIMP GAN denoiser consistently outputs more realistic looking denoised images than the Thornley et al. baseline, which is overly aggressive in its denoising, suffers from checkerboard artifacts (Figure 3), and tends to hallucinate content in background regions by smoothing speckle noise into gaussians. Qualitative Noise2Atom outputs on four representative held-out frames are shown in Figure 12 (Appendix L). Noise2Atom fails on the broader denoising task, with complete content collapse on any input whose content is not a clean atomic-resolution lattice. We therefore do not consider Noise2Atom a comparable denoiser and exclude it from the user study below.

As a paired full-reference denoising benchmark, we use the dwell-time member of the held-out gold-on-carbon evaluation sets described above. Each method takes the lowest-dwell image as input and is compared with the corresponding highest-dwell experimental image from the same field of view. Table 4 compares CIMP GAN with Thornley et al., Noise2Atom, and tuned Gaussian and Fourier low-pass filters. CIMP GAN improves over the learned baselines on SSIM (0.663 vs. 0.379 for Thornley et al. and 0.058 for Noise2Atom), PSNR (27.77 vs. 24.72 and 15.58), and LPIPS (0.249 vs. 0.326 and 0.490). Low-pass filters tuned for SSIM or PSNR obtain higher values on those metrics, but have substantially higher LPIPS (0.596–0.599); filters tuned for LPIPS remain below CIMP GAN on both SSIM and LPIPS. The accompanying scatter plot summarizes these tradeoffs. Additional

paired metrics, including gradient correlation and edge-energy ratio, are reported in Appendix H, and representative low-pass outputs are shown in Appendix Figure 7.



(a) Pd/In/ZrO_2, $t_{\text{dwell}} = 100$ ns, $I_{\text{beam}} = 134.71$ pA.

(b) Au/TiO_2, $t_{\text{dwell}} = 2\ \mu$s, $I_{\text{beam}} = 24.34$ pA.

(c) Ru/TiO_2, $t_{\text{dwell}} = 2\ \mu$s, $I_{\text{beam}} = 31.94$ pA.

(d) Ni(100), $t_{\text{dwell}} = 2\ \mu$s, $I_{\text{beam}} = 28.33$ pA.

Figure 3: Four denoising examples spanning distinct material systems. Within each panel, columns are raw / Thornley et al. / CIMP-GAN; the top row is a 128×128 crop and the bottom row is the full 2048×2048 frame. CIMP-GAN target metadata is $t_{\text{dwell}} = 24.4\ \mu$s, $I_{\text{beam}} = 86.95$ pA in all four panels. Scale bars on the full-frame images represent 10 nm.

For the broader 100-frame validation-set evaluation, CIMP GAN preserves edge location best (gradient correlation 0.89) and retains more total edge magnitude (edge energy ratio 0.64) than the Thornley et al. baseline (0.69 and 0.36, respectively). The full reference-free metric set, including power-spectrum ratios and Noise2Atom, is reported in Appendix K.

Table 4: Gold-on-carbon low-to-high dwell-time denoising benchmark with tuned low-pass baselines. The Eval image pair row compares the lowest-dwell image directly with the highest-dwell reference. Gaussian and Fourier rows are tuned against the highest-dwell reference for the named objective. Right: SSIM–LPIPS tradeoff for the same rows.

Method	**Objective**	SSIM↑	PSNR↑	LPIPS↓
Eval image pair	–	0.197	21.69	0.369
Thornley et al.	–	0.379	24.72	0.326
Noise2Atom	–	0.058	15.58	0.490
CIMP GAN	–	**0.663**	**27.77**	**0.249**
Gaussian $\sigma = 8.0$	SSIM/PSNR	0.794	29.98	0.599
Gaussian $\sigma = 0.6$	LPIPS	0.351	24.28	0.339
Gaussian $\sigma = 5.0$	Gradient	0.783	29.42	0.728
Gaussian $\sigma = 1.75$	Edge	0.678	27.52	0.521
Gaussian $\sigma = 0.765$	PSD	0.434	25.21	0.370
Gaussian $\sigma = 0.77$	Thornley PSD	0.436	25.23	0.371
Fourier cutoff 0.04	SSIM/PSNR	0.793	29.89	0.596
Fourier cutoff 1.0	LPIPS/PSD	0.211	22.01	0.379
Fourier cutoff 0.10	Gradient	0.757	28.28	0.863
Fourier cutoff 0.20	Edge	0.679	27.19	0.744

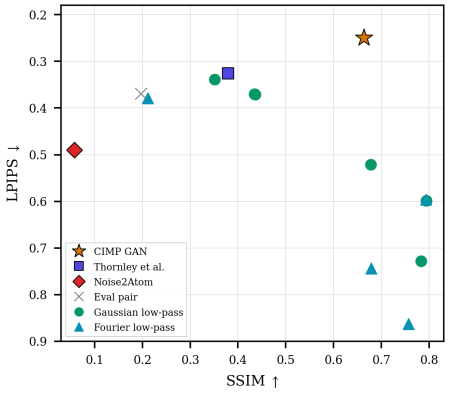


Table 5: Per-rater counts of CIMP GAN choices in the user study, with the Wilson 95% CI on the rate, and a two-sided binomial p-value against the $50/50$ null.

Rater	Judgments	CIMP GAN chosen (Wilson 95% CI)	Binomial p vs $50/50$
Rater 1	200	137 (68.5% [61.8, 74.5])	$< 10^{-3}$
Rater 2	101	77 (76.2% [67.1, 83.5])	$< 10^{-3}$
Rater 3	76	65 (85.5% [75.9, 91.7])	$< 10^{-3}$
Rater 4	73	37 (50.7% [39.5, 61.8])	1.00
Rater 5	36	25 (69.4% [53.1, 82.0])	0.03
Total	486	341 (70.2% [66.0, 74.1])	$< 10^{-3}$

To confirm the quality of CIMP-GAN as a denoiser, we conduct a blind user study with 5 independent experimental microscopists. The user study presented 200 random (raw, Thornley et al., CIMP-GAN) triplets from the validation set, where each image was a 128×128 image patch and the CIMP-GAN images were conditioned with target metadata of $t_{\text{dwell}} = 24.4\,\mu\text{s}$, $I_{\text{beam}} = 86.95\,\text{pA}$. The full instructions as well as the judgment overlap between raters are reported in Appendix M.

Across all 486 judgments, CIMP GAN is chosen on 70.2% (Wilson 95% CI $[66.0, 74.1]$) of trials, significantly above 50% (binomial $p < 10^{-3}$). Per-rater preference is uneven, four raters prefer CIMP at 68–86% while one is indistinguishable from chance at 50.7% (Table 5). Coverage is uneven across raters, ranging from 36 to 200 judgments per rater. A general impression from some of the raters was that "the one with cross artefacts" (Thornley et al.) generates sharper but more unrealistic images, and works better when the signal of the raw image is already quite high.

6 Limitations & Conclusion

We introduced CIMP, a CLIP-style contrastive framework that learns a joint embedding space over HAADF-STEM images and their acquisition metadata from 7,330 paired examples. The factored representation we obtain from contrastive image-metadata alignment is a foundation other workflows can condition on rather than infrastructure for any single application. It could anchor segmentation or analysis networks robust to varying instrument state, serve as a retrieval index over a metadata-tagged corpus, or act as a perceptual loss for a physics-informed denoiser, as shown.

Two principal limitations remain. Our dataset of 7,330 single-instrument images is modest by computer-vision standards, and the data underrepresents boundary conditions such as clipped detectors and saturated extremes that operators rarely record (full discussion in Appendix N). Future work will address both by extending the dataset across instruments, quantifying the denoiser against stronger baselines on paired data, and investigating how CIMP embeddings transfer to other analysis tasks.

More broadly, this work represents a step toward autonomous high-throughput microscopy, in which closed-loop instruments select, acquire, and interpret images with reduced human supervision. A factored representation grounded in physical microscope state is a foundational requirement for any such system, and the same representation extends to interactive workflows, which could provide operators with tools that reason over acquisition parameters in real time. However, the utility of such approaches scales with the diversity of paired corpora available to the field, and we hope this work motivates broader release of the metadata-rich data that remains underutilized in the electron microscopy community.

AI use statement

In this work, we used generative AI tools for SLURM job submission, bash scripting, syncing data between multiple remote computers, building the interactive user study interface, formatting figures and tables for the manuscript, reviewing text for clarity, and anonymizing our supplementary code. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results.

Ethics statement

This work includes a user study with experimental microscopists. Participation was voluntary, and raters evaluated anonymized image triplets without providing sensitive personal information. We are not aware of any ethics concerns associated with this work.

Reproducibility statement

One of the main contributions of this work is the HAADF-STEM image–metadata dataset, which will be released after double-blind review. The remaining experimental artifacts, including code, trained models, and evaluation scripts, are open-source and will be made available on GitHub and Hugging Face.

References

Milad Abolhasani, Mahshid Ahmadi, Sterling Baird, Curtis P. Berlinguette, Keith A. Brown, David Fenning, Ian T. Foster, Willi Gottstein, Sergio Pablo-Garcia, Gabe Gomes, Jason R. Hattrick-Simpers, Jason E. Hein, Taro Hitosugi, Sergei V. Kalinin, Benji Maruyama, Joshua Schrier, Martin Seifrid, Taylor D. Sparks, Shijing Sun, Brandon R. Sutherland, Nathan J. Szymanski, Dean Thomas, Lauren Walters, Jie Xu, Yan Zeng, and Charles Yang. On the need for autonomous science instruments: A call to action. ChemRxiv preprint, February 2026. URL https://chemrxiv.org/doi/full/10.26434/chemrxiv.10001836/v1. Posted 02 February 2026.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

R.F. Egerton. Two- and three-dimensional electron imaging of beam-sensitive specimens. *Micron*, 194: 103819, 2025. ISSN 0968-4328. doi: https://doi.org/10.1016/j.micron.2025.103819. URL https://www.sciencedirect.com/science/article/pii/S096843282500037X.

Henrik Eliasson. Synthetic data from: Improving nanoparticle size estimation from scanning transmission electron micrographs with a multislice surrogate model, 2025. URL https://doi.org/10.5281/zenodo.14608502. Dataset.

Henrik Eliasson and Rolf Erni. Improving nanoparticle size estimation from scanning transmission electron micrographs with a multislice surrogate model. *Nano Letters*, 25(6):2474–2479, 2025. doi: 10.1021/acs.nanolett.4c06025. URL https://doi.org/10.1021/acs.nanolett.4c06025.

Henrik Eliasson, Angus Lothian, Ivan Surin, Sharon Mitchell, Javier Pérez-Ramírez, and Rolf Erni. Precise size determination of supported catalyst nanoparticles via generative ai and scanning transmission electron microscopy. *Small Methods*, pp. 2401108, 2024. ISSN 2366-9608. doi: https://doi.org/10.1002/smtd.202401108. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/smtd.202401108.

Henrik Eliasson, Fangjinhua Wang, Xi Wang, Daniel Barath, Marc Pollefeys, and Rolf Erni. Morphology prediction of small nanoparticles in any orientation from single electron micrographs. *npj Computational Materials*, May 2026. ISSN 2057-3960. doi: 10.1038/s41524-026-02114-w. URL https://doi.org/10.1038/s41524-026-02114-w.

Rolf Erni, Marta D. Rossell, Christian Kisielowski, and Ulrich Dahmen. Atomic-resolution imaging with a sub-50-pm electron probe. *Phys. Rev. Lett.*, 102:096101, Mar 2009. doi: 10.1103/PhysRevLett.102.096101. URL https://link.aps.org/doi/10.1103/PhysRevLett.102.096101.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Nina M. Ivanova, Alexey S. Kashin, and Valentine P. Ananikov. Lost data in electron microscopy. *Chemistry*, 7(5), 2025. ISSN 2624-8549. doi: 10.3390/chemistry7050160. URL https://www.mdpi.com/2624-8549/7/5/160.

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. ISSN 2057-3960. doi: 10.1038/s41524-023-01042-3. URL https://doi.org/10.1038/s41524-023-01042-3.

Joodeok Kim, Jinho Rhee, Sungsu Kang, Mingyu Jung, Jihoon Kim, Miji Jeon, Junsun Park, Jimin Ham, Byung Hyo Kim, Won Chul Lee, Soung-Hun Roh, and Jungwon Park. Self-supervised machine learning framework for high-throughput electron microscopy. *Science Advances*, 11(14): eads5552, 2025. doi: 10.1126/sciadv.ads5552. URL https://www.science.org/doi/abs/10.1126/sciadv.ads5552.

O. L. Krivanek, M. F. Chisholm, N. Dellby, and M. F. Murfitt. Aberration-corrected STEM for atomic-resolution imaging and analysis. *Journal of Microscopy*, 259(3):165–172, 2015. doi: 10.1111/jmi.12254. URL https://doi.org/10.1111/jmi.12254.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void — learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2129–2137, 2019. doi: 10.1109/CVPR.2019.00223.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lehtinen18a.html.

Ruoqian Lin, Rui Zhang, Chunyang Wang, Xiao-Qing Yang, and Huolin L. Xin. Temimagenet training library and atomsegnet deep-learning models for high-precision atom segmentation, localization, denoising, and deblurring of atomic-resolution images. *Scientific Reports*, 11 (1):5386, 2021. doi: 10.1038/s41598-021-84499-w. URL https://doi.org/10.1038/s41598-021-84499-w.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

I. Lobato, T. Friedrich, and S. Van Aert. Deep convolutional neural networks to restore single-shot electron microscopy images. *npj Computational Materials*, 10(1):10, jan 2024. ISSN 2057-3960. doi: 10.1038/s41524-023-01188-0. URL https://doi.org/10.1038/s41524-023-01188-0.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Feature-wise linear modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ludwig Reimer. *Specimen Damage by Electron Irradiation*, pp. 421–453. Springer Berlin Heidelberg, Berlin, Heidelberg, 1984. ISBN 978-3-662-13553-2. doi: 10.1007/978-3-662-13553-2_10. URL https://doi.org/10.1007/978-3-662-13553-2_10.

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*, pp. 234–241. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24574-4_28.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2021. URL https://arxiv.org/abs/2111.05826.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023. doi: 10.1038/s41467-023-42328-w. URL https://doi.org/10.1038/s41467-023-42328-w.

Hidetaka Sawada, Naoki Shimura, Fumio Hosokawa, Naoya Shibata, and Yuichi Ikuhara. Resolving 45-pm-separated Si–Si atomic columns with an aberration-corrected STEM. *Microscopy*, 64(3): 213–217, 2015. doi: 10.1093/jmicro/dfv014. URL https://doi.org/10.1093/jmicro/dfv014.

Eric Schwenker, Fatih Sen, Chris Wolverton, Colin Ophus, and Maria K. Y. Chan. A simulated atomic-resolution haadf stem imaging dataset containing unique icsd structure prototypes, 2020. URL https://doi.org/10.18126/szeq-yde5. Dataset.

Dev Yashpal Sheth, Sreyas Mohan, Joshua Vincent, Ramon Manzorro, Peter A. Crozier, Mitesh M. Khapra, Eero P. Simoncelli, and Carlos Fernandez-Granda. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Katherine Sytwu, Luis Rangel DaCosta, and Mary Scott. Segmented high-resolution transmission electron microscopy images of nanoparticles, 2023. URL https://doi.org/10.7941/D1SP93. Dataset.

Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20876–20889. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02024. URL https://doi.org/10.1109/CVPR52688.2022.02024.

William Thornley, Sam Sullivan-Allsop, Rongsheng Cai, Nick Clark, Roman Gorbachev, and Sarah J. Haigh. Noise2void for denoising atomic resolution scanning transmission electron microscopy images. *npj Computational Materials*, 12(1):68, Jan 2026. ISSN 2057-3960. doi: 10.1038/s41524-025-01939-1. URL https://doi.org/10.1038/s41524-025-01939-1.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Feng Wang, Trond R. Henninen, Debora Keller, and Rolf Erni. Noise2atom: unsupervised denoising for scanning transmission electron microscopy images. *Applied Microscopy*, 50(1):23, 2020. ISSN 2287-4445. doi: 10.1186/s42649-020-00041-8. URL https://doi.org/10.1186/s42649-020-00041-8.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

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 (ICCV)*, pp. 2223–2232, 2017. doi: 10.1109/ICCV.2017.244.

A BRIEF EXPLANATION OF METADATA PARAMETERS

1. **Pixel Size:** This is the physical size of a pixel (typically around 15-50 pm for high resolution imaging), and the lateral step between probe positions in the STEM raster scan. Changing the pixel size is the result of changing magnification, which in normal operation would result in a larger field of view. It does not make physical sense to ask a model to change pixel size of the image while keeping the image size fixed. Therefore it is ignored and also not shuffled in the metadata shuffling within batches nor in the validation set for evaluation.
2. **Dwell Time:** Dwell time determines how long the electron beam stays at each pixel position. A longer dwell time leads to more electrons hitting the sample at each pixel position and consequently leads to a higher signal image. A shorter dwell time leads to reduced signal and a significant increase of poisson noise.
3. **Beam Convergence Angle:** The convergence angle of the beam is the incident angle of the convergent STEM probe as it hits the sample. The convergence angle is a key parameter of the diffraction-limited achievable resolution of the microscope. Changing the convergence angle can be done in multiple ways, but in this dataset, it was mostly controlled by the size of an aperture.
4. **Beam Current:** This is the current of the electron probe, as measured by the fluorescent viewing screen, which is placed after the sample and HAADF detector. While a larger beam current indicates that more electrons hit the sample, the effect on signal strength depends on how many electrons hit the HAADF detector which is affected by other metadata like inner collection angle described below. Beam current should in general be kept as low as possible to avoid damaging the sample and inducing phenomena that do not occur naturally.
5. **Detector Gain:** The "Gain" or "Contrast" is a parameter that determines how much the signal on the detector is amplified.
6. **Detector Offset:** The "Offset" or "Brightness" is a parameter that sets the black level of the detector reading.
7. **Inner Collection Angle:** The HAADF detector is annular, and the inner collection angle defines the minimum scattering angle of electrons that contribute to the recorded signal. This parameter is important both because it determines the fraction of the incident beam collected and because it governs the imaging contrast mechanism. At high scattering angles, the signal is dominated by incoherent Rutherford scattering, which forms the basis of HAADF-STEM contrast and gives rise to the approximate $Z^{1.7}$ dependence and mass–thickness contrast. At lower collection angles, coherently scattered electrons are also detected, introducing diffraction contrast and altering the imaging mechanism. For HAADF-STEM, including the majority of images in our dataset, the inner collection angle is typically chosen to be at least ~ 3 times the probe convergence semi-angle.

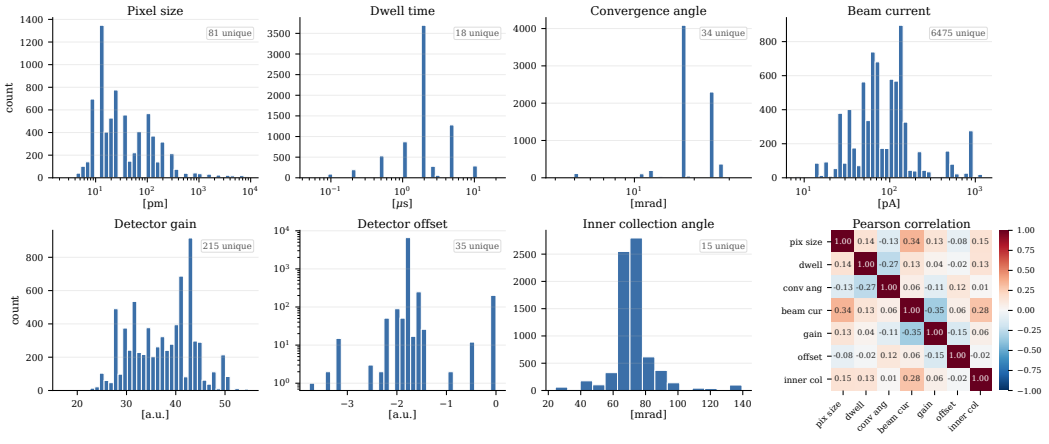


Figure 4: Distribution of acquisition metadata across 7,330 HAADF-STEM images.

B Dataset Construction & Contents

Figure 4 highlights several structural properties of the dataset that are relevant for downstream modelling. Pixel size, dwell time, and inner collection angle each take a small number of discrete values (81, 18, and 15 unique values respectively) as operators sweep these parameters through a fixed grid of magnifications, exposure presets, and detector geometries rather than varying them continuously. Beam current and detector gain are by contrast effectively continuous (thousands and hundreds of unique values). The many values for beam current originate from factors such as drift and degradation of the electron gun, lenses, and aperture placements. Convergence angle is bimodal, with two strongly populated values corresponding to the two aperture sizes routinely used during acquisition. The Pearson correlation panel (bottom-right) shows that no pair of parameters is strongly collinear, so each metadata dimension carries an independent supervisory signal for the contrastive loss.

Distribution across material systems. Material identity is inferred from the parent folder name of each source `.emd` file rather than an explicit metadata field. Matching the 130 folder names to catalyst-naming conventions assigns 5,017 of 7,330 acquisitions ($\sim 68\%$) to identifiable material families; the remaining $\sim 32\%$ come from non-material folders (e.g., alignment, selection, date, or QC folders) and cannot be resolved from filenames alone. Table 6 summarizes the resulting distribution. The dataset is dominated by heterogeneous catalysts, led by Pt/CeO_2 ($\sim 21\%$), with additional coverage of CO_2-conversion catalysts, CO-oxidation systems, single-atom catalysts, and bare oxide supports. Many classes are represented across fresh, pretreated, used, and in-situ states. Because these family assignments rely solely on operator-recorded folder names, they should be interpreted as indicative rather than curated labels.

Table 6: Heuristic material-family breakdown of the 7,330 acquisitions, derived by pattern-matching the parent folder name of each source `.emd` file against catalyst-naming conventions. The "Unspecified" row covers folders with alignment, QC, or date-only names that do not encode the material in the path.

Material family	**Imgs**	**%**
Pt/CeO_2	1511	20.6
Cr-oxide on CeO_2/Nb_2O_5	153	2.1
Pt/ZrO_2, Pt/TiO_2	37	0.5
Pd/(In, Zn, Zr)-Ox	538	7.3
In_2O_3-based (In/Zr-, In/Ti-, In/Hf-Ox)	445	6.1
Cu/Zn/Zr-, Cu/Zr-Ox (CZA family)	257	3.5
Zn/Zr-Ox	56	0.8
Ru/TiO_2	228	3.1
Ru/Ni & Ni/Ru bimetallic	98	1.3
Ru-based (other / unspec. support)	173	2.4
Au/TiO_2	330	4.5
Au/MnO_x	266	3.6
Au/CeO_2	46	0.6
Mn-/Ni- on CeO_2 or other oxides	110	1.5
Single- / few-atom on Al_2O_3 or CeO_2	96	1.3
Bare CeO_2 model supports	273	3.7
Bare TiO_2, HfO_2, aerogel supports	148	2.0
High-entropy oxide	1	0.0
Synthesis / pretreatment series	251	3.4
Unspecified (alignment, QC, date-only)	2311	31.5
Unmatched	2	0.0
Total	**7330**	**100.0**

Train/validation split strategy. The 6,597 / 733 split ($\approx 90/10$) is stratified at the source-acquisition level rather than at the patch level. Specifically, source acquisitions (one `.emd` file per record) are randomly shuffled with a fixed seed and partitioned into train and validation sets; all 256×256 training crops sampled from a given source acquisition are then assigned to the same split. Because each of the 7,330 records corresponds to a distinct full-resolution image, this is equivalent to a $90/10$ split over images and prevents patches from the same image from leaking from training into validation, which would otherwise inflate retrieval metrics through near-duplicate matching of within-image texture.

Images are stored in Thermo Fisher Scientific's Velox `.emd` (HDF5-based) format, and the seven acquisition-metadata fields are read directly from the per-image EMD record at dataset-construction time.

The four metadata dimensions that span several orders of magnitude (pixel size, dwell time, convergence angle, beam current) are stored in $\log_{10}$ space; all seven dimensions are additionally z-scored before being passed to the metadata encoder.

C QUALITATIVE ANALYSIS

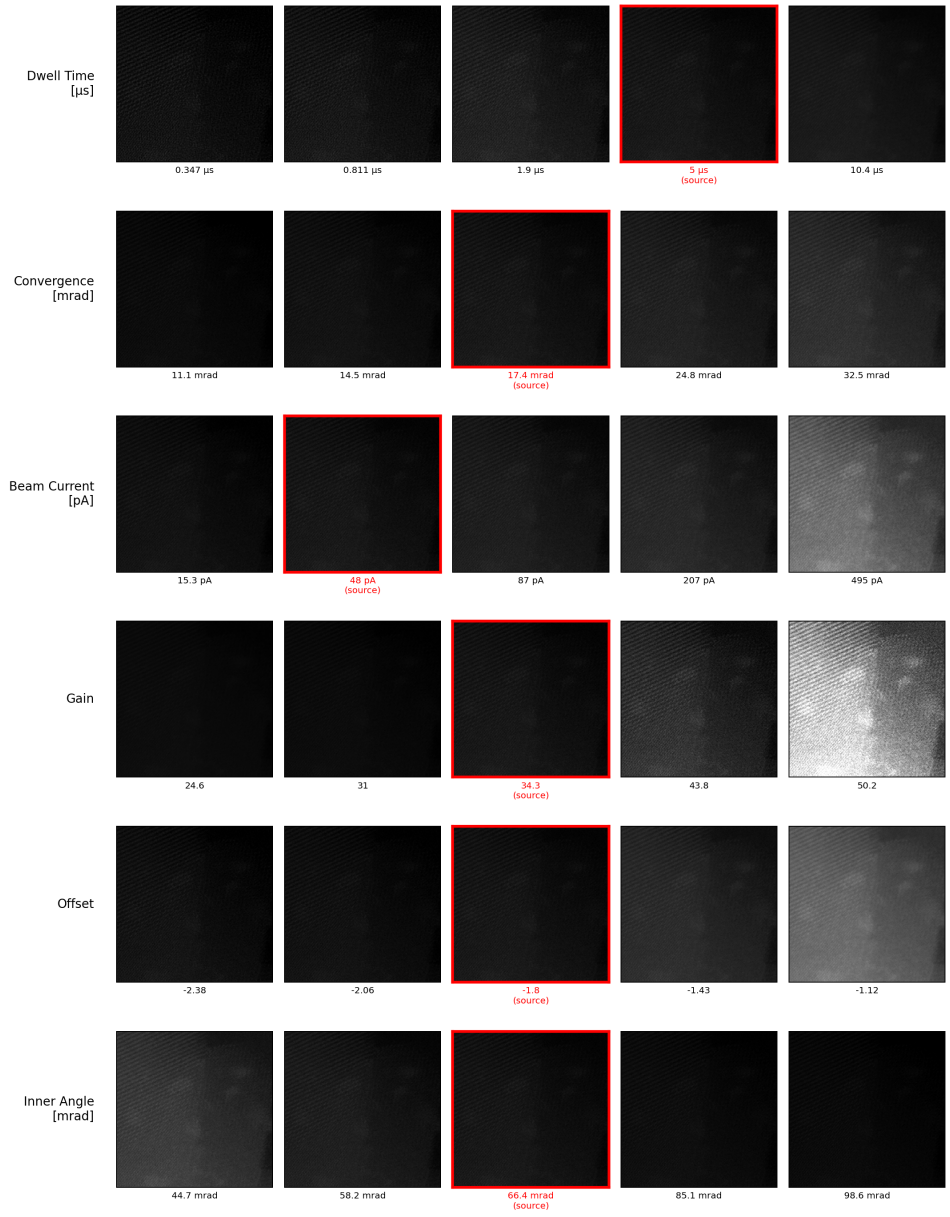
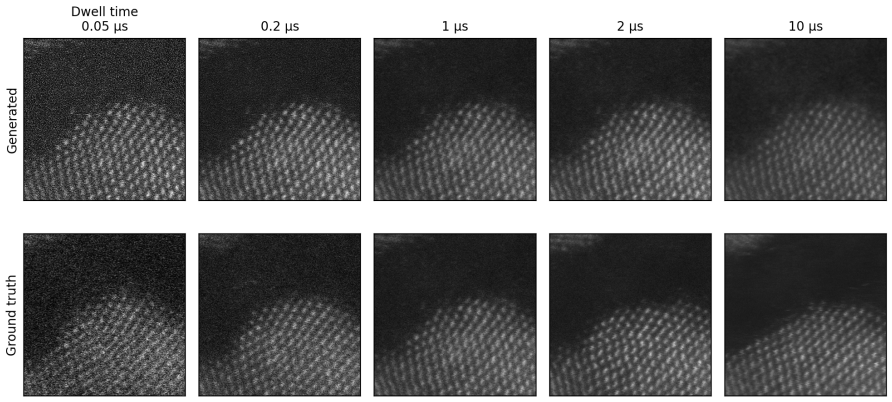
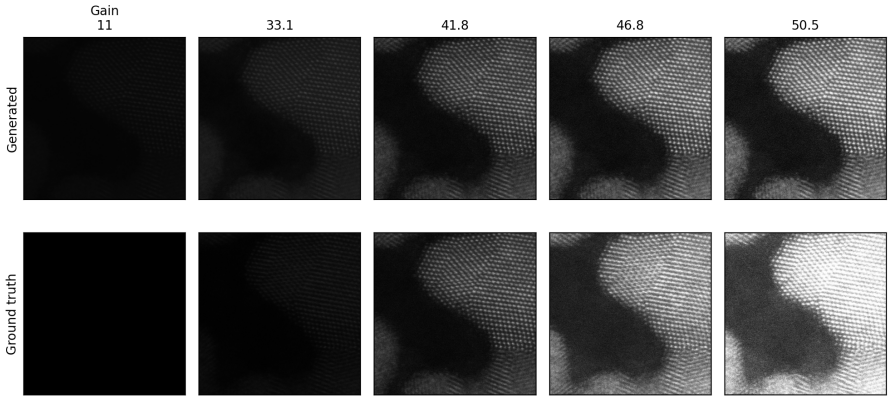


Figure 5: Our CIMP GAN style transfer model synthetically varying the metadata parameters of a validation image, one by one. Style-transferred images were generated at five target values (-2σ, -1σ, 0, $+1\sigma$, and $+2\sigma$ of the distribution of that metadata in the full dataset). The image in the red frame is the original input image and it replaces whichever of the generated images that it has the closest metadata to. Collecting the metadata of all images with a red frame gives you the original metadata. *Note: Gain and Offset are also denoted as Contrast and Brightness by some other STEM instruments/software, and their ranges can be different from Gain and Offset.* Scale bars are 2 nm.

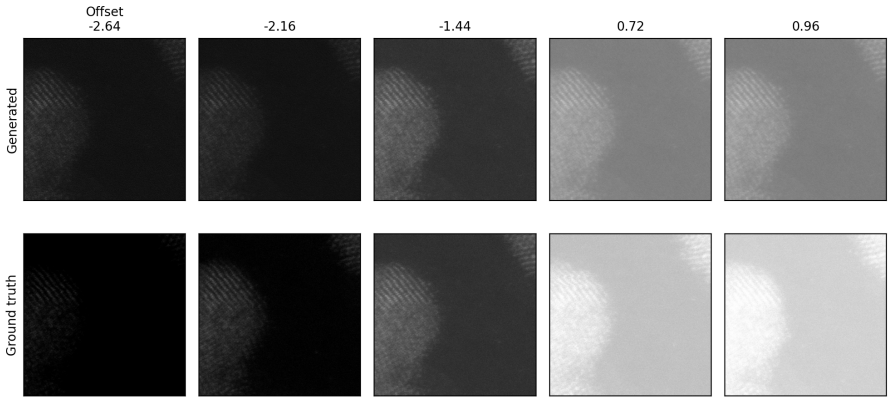
Per-parameter qualitative trends. For each parameter, Figure 5 shows the generator's output at five target values (-2σ, -1σ, 0, $+1\sigma$, and $+2\sigma$ of the distribution of that metadata in the entire dataset), with all other metadata held at the source image's original values. For dwell time, gain, and offset, the full qualitative comparisons are shown in Figures 6a, 6b, and 6c. The general observed trends for each metadata are as follows:



(a) Results of varying the dwell time synthetically (top row) with our style transfer model and experimentally (bottom row). The expected behaviour of increasing the dwell time is a smoother image with less poisson noise but with a maintained pixel intensity distribution. We observe that our model indeed makes the image smoother, but also alters the intensity distribution by making background slightly brighter and object slightly darker. Furthermore, reducing dwell time virtually does not add as much noise as we observe experimentally.



(b) Results of varying the gain (or contrast) synthetically (top row) with our style transfer model and experimentally (bottom row). The expected behaviour of increasing the gain is an amplification of both signal and noise and a brighter image. We observe that our model can successfully model this trend but fails to model the upper bound where the image would be clipped (further discussed in Appendix N).



(c) Results of varying the offset (or brightness) synthetically (top row) with our style transfer model and experimentally (bottom row). The expected behaviour of increasing the offset on this instrument is that the black level is lowered, pushing the visible content into a smaller part of the dynamic range, eventually clipping the image. Decreasing the offset leads to real signal being treated as black and parts of the object disappearing.

Figure 6: Full-size qualitative comparison between changing dwell time, gain, and offset virtually with CIMP GAN and experimentally on the microscope. Scale bars are 2 nm.

1. **Dwell time**: From low to high dwell time, the output image goes from noisy to smooth as expected.

2. **Convergence angle**: When virtually increasing the convergence angle, the resulting effect is a brighter image. This is not what you would expect in theory, and likely happens because of the way convergence angle was controlled in the recording of the dataset. For the vast majority of images (see Figure 4), the convergence angle was set to two different values, controlled by changing the size of an aperture. A smaller aperture gave a smaller convergence angle but also blocked more of the electron beam, leading to a lower beam current which leads to a darker image. This bias is likely what is picked up by the model.

3. **Beam Current**: For beam current, the trend is clear, more electrons lead to a brighter image. Eventually, this should also lead to clipping of the image, but as discussed in Appendix N, that boundary condition is not represented in the training data.

4. **Gain**: Increased gain also leads to a brighter image. However, in contrast to the beam current sweep, the noise is also amplified.

5. **Offset**: There is no clear trend for offset except that the background appears slightly darker for lower values.

6. **Inner collection angle**: Making the inner collection angle smaller results in a clear trend of increased image intensity. With a smaller inner collection angle, more of the scattered electrons end up on the HAADF detector, which should make the image brighter. However, electrons scattered to different angles differ, and with a much smaller collection angle we would expect the contrast mechanism to change as more Bragg-diffracted electrons reach the detector. Such intricacies are likely not learned by the model, especially since the vast majority of images in the dataset were recorded with large collection angles relative to beam convergence angle.

D PRETRAINING ABLATIONS

Table 7 summarises the architecture and hyperparameter sweeps used to select the CIMP encoder. ResNet-18 trained from scratch matches or exceeds the ImageNet-pretrained ViT, suggesting that HAADF-STEM image statistics differ enough from natural images that pretrained transformer features provide little transfer benefit. Crop sizes from 128–512 pixels give similar retrieval performance, which indicates that acquisition style is visible at small patch scales. Effective batch 256 already saturates retrieval performance, but we keep 512 as the default for the final encoder.

Table 7: Compact CIMP pretraining ablations. All entries report best-model retrieval accuracy on the held-out validation set; T1/T5/T10 denote Top-1/Top-5/Top-10. Unless varied in the corresponding panel, runs use 1000 epochs, effective batch 512, crop 256, and metadata encoder 256×3.

(a) Backbone

Backbone	T1	T5	T10
ResNet-18	**0.844**	**0.969**	0.984
ViT scratch	0.750	0.953	0.984
ViT pretrained	0.828	**0.969**	**1.000**
ViT frozen	0.234	0.438	0.641

(b) Crop size

Crop	Batch	T1	T5	T10
128	512	0.781	0.906	0.953
256	512	0.766	**0.969**	**0.984**
512	512	**0.844**	0.953	0.969
1024	512	0.738	0.913	0.950

(c) Metadata encoder

Hidden	Layers	T1	T5	T10
64	3	0.797	0.953	0.984
128	3	0.766	0.953	**1.000**
256	3	**0.828**	0.969	**1.000**
512	3	**0.828**	0.969	**1.000**
1024	3	**0.828**	**0.984**	**1.000**

(d) Effective batch, ResNet-18

Batch	T1	T5	T10
128	0.775	0.988	**1.000**
256	**0.859**	**1.000**	**1.000**
512	**0.859**	**1.000**	**1.000**

E HELD-OUT OFFSET METADATA PROBE

To test whether acquisition information can remain visually encoded even when it is not provided as a contrastive metadata target, we trained an otherwise identical CIMP encoder with detector offset omitted from the metadata vector. We then froze the visual encoder and fit the same Ridge linear probe used in Section 5.1 to predict offset from image embeddings.

Table 8: Offset recovery from visual embeddings when detector offset is included in or omitted from the CIMP metadata objective. Values are mean ± standard error across bootstrap resamples.

CIMP metadata objective	Offset $R^2 \uparrow$	Offset SMAPE↓
Offset included	0.80 ± 0.05	$9.2 \pm 1.2\%$
Offset omitted	0.696 ± 0.060	$10.38 \pm 1.17\%$

Offset remains recoverable even when it is omitted from the contrastive metadata vector, though performance decreases. This suggests that some acquisition information is visually encoded through its effect on image statistics rather than only reflected by the paired metadata embedding.

F GOLD-ON-CARBON ENDPOINT STYLE TRANSFER

Table 9: Gold-on-carbon endpoint style-transfer benchmark on the displayed evaluation images. For each held-out acquisition series, the first displayed image is the input. Eval image pair compares that input with the real image at the target setting; CIMP GAN compares the generated image at the same target setting with the real target image.

Series	**Target setting**	**Method**	SSIM↑	PSNR↑	LPIPS↓	Grad Corr	Edge Ratio
Dwell	1×10^{-5}	Eval image pair	0.2093	21.82	0.3374	0.1899	4.9296
		CIMP GAN	0.7407	29.73	0.0960	0.3453	0.8265
Gain	50.51	Eval image pair	0.0066	3.77	0.8219	0.0393	0.0017
		CIMP GAN	0.0765	9.79	0.2241	0.1059	0.9067
Offset	0.9598	Eval image pair	0.0083	1.40	0.7499	0.0888	0.3355
		CIMP GAN	0.4800	9.50	0.2178	0.0413	0.8520

G PHYSICAL GROUNDING CHECKS

CIMP is physically informed because it conditions on measured acquisition parameters and can be tested against held-out experimental parameter sweeps. It is not physically grounded in the stricter sense of embedding a multislice forward model, detector-response model, or explicit likelihood for electron-counting noise. To make this distinction concrete, we measured whether the metadata-conditioned generator reproduces simple dose and detector trends in flat background regions of the gold-on-carbon evaluation sweeps.

For dwell time, we selected low-gradient background regions, measured the coefficient of variation of the background intensity as a function of dose proxy, and fit a log-log slope. An ideal Poisson process has relative-noise slope -0.5 for the coefficient of variation. The real microscope sweep is shallower, with slope -0.385, while CIMP GAN gives a similar but slightly steeper slope of -0.426 (Table 10). This indicates that the generator captures the main dose-dependent noise trend, while not imposing an exact Poisson model.

Table 10: Dwell-time dose-law check in flat background regions. Slopes are fit on log-log curves across the held-out gold-on-carbon dwell sweep.

Metric	**Ideal slope**	**Real slope**	**CIMP slope**	**CIMP R^2**
Background CV	-0.500	-0.385	-0.426	0.985
Relative variance	-1.000	-0.770	-0.855	0.985
High-pass std	–	-0.380	-0.394	0.969

For gain and offset, we measured the mean background-intensity slope across the detector sweep. CIMP GAN captures the direction and linearity of the detector response, but underestimates the magnitude of the offset shift: the real offset slope is 0.245, whereas CIMP GAN has slope 0.079, or about 32% of the measured real slope (Table 11). This matches the qualitative limitation that the model handles monotonic detector changes better than extreme boundary conditions such as clipping.

Table 11: Detector-affine checks in flat background regions of the held-out gain and offset sweeps.

Sweep	**Metric**	**Real slope**	**CIMP slope**	**CIMP R^2**
Gain	Background mean	0.005	0.003	0.867
Gain	Background CV	0.007	0.006	0.869
Offset	Background mean	0.245	0.079	0.957
Offset	Background high-pass std	0.003	-0.001	0.725

H Paired Gold-on-Carbon Dwell Benchmark

Table 4 in the main text reports the primary SSIM, PSNR, and LPIPS comparison on the held-out aligned gold-on-carbon dwell series, together with tuned low-pass baselines. Table 12 reports the additional paired metrics for the learned methods and the raw low-to-high dwell image pair. Each method takes the lowest-dwell image as input and is compared with the corresponding highest-dwell experimental image from the same field of view.

Table 12: Gold-on-carbon low-to-high dwell-time denoising benchmark. The Eval image pair row compares the lowest-dwell image directly with the highest-dwell reference; Noise2Atom, Thornley et al., and CIMP GAN are evaluated against the same reference.

Method	SSIM↑	PSNR↑	LPIPS↓	Grad Corr↑	Edge Ratio
Eval image pair	0.1967	21.69	0.3691	0.1835	4.9947
Noise2Atom	0.0576	15.58	0.4901	0.1400	2.3107
Thornley et al.	0.3787	24.72	0.3258	0.1897	2.9361
CIMP GAN	0.6634	27.77	0.2488	0.2214	1.0583

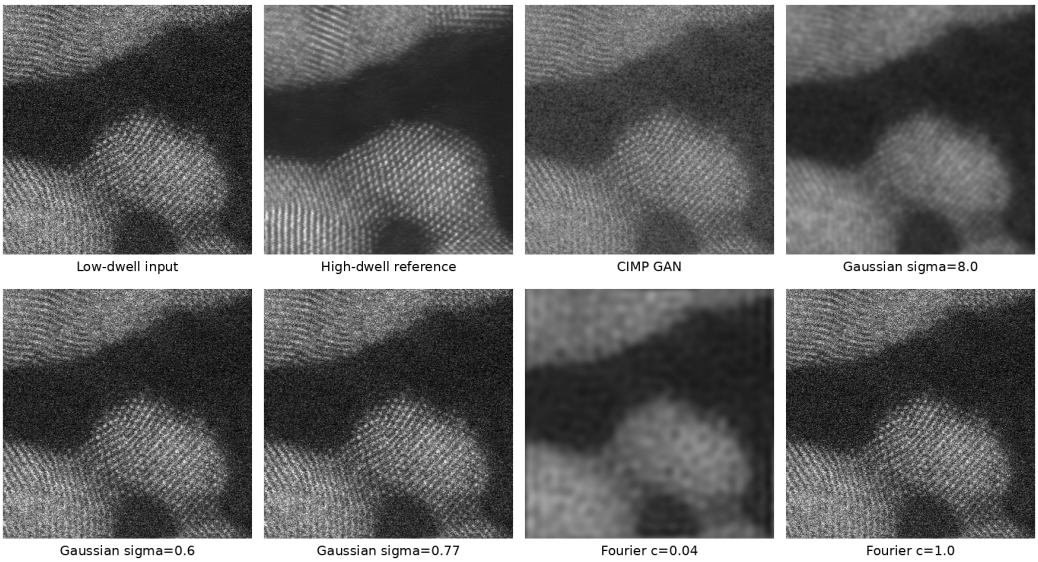


Figure 7: Representative outputs for the gold-on-carbon low-to-high dwell benchmark. The strongly smoothed Gaussian and Fourier settings optimize SSIM or PSNR but remove high-frequency structure, whereas weaker low-pass settings retain more texture while scoring worse on full-reference metrics than CIMP GAN in Table 4. All panels use shared grayscale normalization.

I Style Transfer Embedding Verification

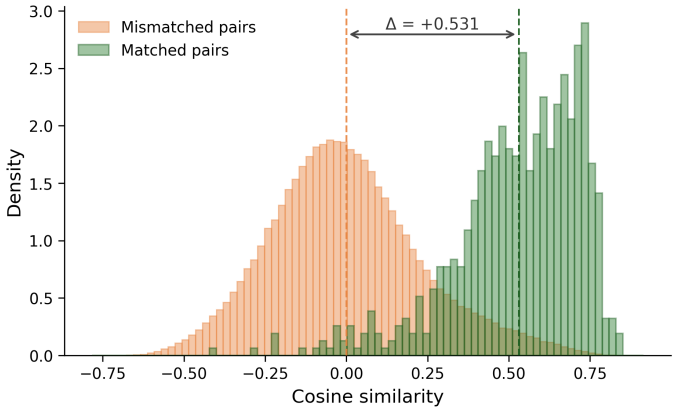


Figure 8: Cosine similarity distribution for style transfer. Style-transferred validation images are passed through the CIMP visual encoder, and the resulting embeddings are compared to their target metadata embedding (correct pairs) and to all other metadata embeddings in the validation batch (incorrect pairs). The mean similarity of correct pairs exceeds that of incorrect pairs by 0.531, indicating that the generator aligns image style with the provided conditioning embedding rather than producing arbitrary outputs.

J GAN Ablations

We swept loss weights, the cycle-consistency masking ratio, and a noise-augmentation toggle for the metadata-conditioned style-transfer GAN. All runs share the same CIMP encoder, U-Net architecture, and optimiser, and were trained for 100 epochs. Table 13 reports validation losses at the best composite-validation epoch within each run.

Table 13: GAN ablations, each run trained for 100 epochs. Validation embedding-alignment, identity, and cycle losses, and the best composite total validation loss within each run.

Run	λ_{emb}	λ_{GAN}	λ_{cyc}	λ_{id}	**mask**	**noise**	val_emb ↓	val_id ↓	val_cyc ↓	best total_val ↓
1	1.0	1.0	0.1	0.1	5%	off	0.00801	0.01337	0.00993	0.01005
2	1.0	1.0	0.2	0.1	5%	off	0.00906	0.01456	0.01619	0.01364
3	1.0	1.0	0.05	0.1	5%	off	0.00926	0.00321	0.00745	0.00978
4	1.0	1.0	0.1	0.05	5%	off	0.00778	0.01293	0.01103	0.00954
5	1.0	1.0	0.1	0.1	5%	on	0.00824	0.00277	0.00653	0.00904
6	1.0	1.0	0.05	0.05	5%	off	0.01011	0.00465	0.01137	0.01093
7	1.0	1.0	0.1	0.075	0%	off	0.00944	0.00321	0.00641	0.01017
8	1.0	1.0	0.1	0.1	15%	off	0.00759	0.00364	0.00813	0.00852
9	1.0	1.0	0.1	0.1	10%	off	0.00768	0.00281	0.00747	0.00859
10	2.0	1.0	0.1	0.1	5%	off	0.00768	0.01377	0.01902	0.01864
11	1.0	0	0.1	0.1	5%	off	0.00702	0.01921	0.00835	0.00978

Several runs achieved lower validation losses than Run 1, but qualitative inspection of their generated images revealed visually implausible outputs that were inconsistent with the target metadata and input content, so we selected Run 1 as the best-performing setting. In the final configuration reported in the main text, we trained for 200 epochs with $\lambda_{\text{emb}} = \lambda_{\text{GAN}} = 1$, $\lambda_{\text{cyc}} = 0.4$, and an identity weight annealed from 2.0 to 0.1 over the first 10 epochs.

Training Histories Figures 9 and 10 show the per-epoch training histories for the two learned models reported in the main text: the best CIMP encoder (ResNet-18, crop 256, metadata encoder 256×3, effective batch 512, 1000 epochs) and the metadata-conditioned style-transfer GAN trained on top of it (StyleUNet with FiLM conditioning, base filters 32, crop 256, batch 24, 200 epochs). Bold curves show a rolling mean (25 epochs for CIMP, 10 epochs for the GAN); raw per-epoch values are overlaid in faint.

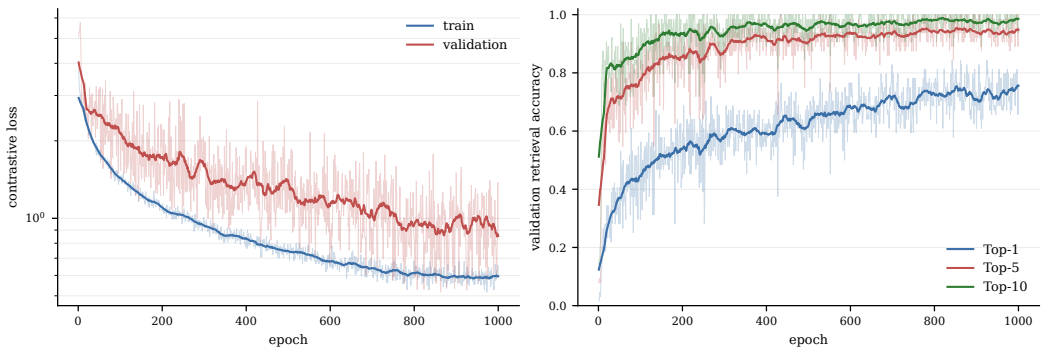


Figure 9: Training history of the best CIMP encoder. Left: train and validation contrastive loss (log scale). Right: validation retrieval accuracy (Top-1, Top-5, Top-10).

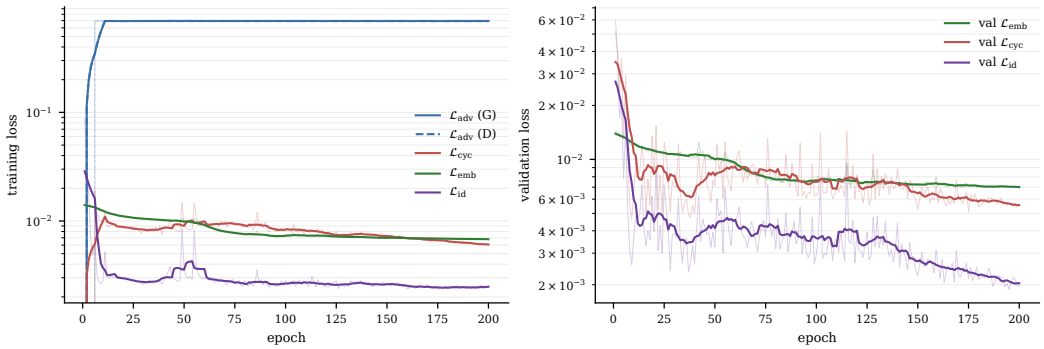


Figure 10: Training history of the style-transfer GAN. Left: training losses per term, including the adversarial generator / discriminator pair and the three non-adversarial components. Right: validation losses for the three non-adversarial components.

K Denoiser Metrics and Power Spectra

Table 14 reports the full set of reference-free denoising metrics for the 100 held-out validation frames discussed in the main text. Residual std and the SNR proxy are included for completeness. Both reward conservative behavior. A denoiser that returned the input unchanged would score perfectly on residual std and would have an unbounded SNR proxy, so we treat them as context rather than as primary measures of denoising quality.

Table 14: Full reference-free denoising metric set on 100 held-out validation frames. CIMP GAN uses fixed high-dose target metadata of $t_{\text{dwell}} = 24.4\,\mu$s and $I_{\text{beam}} = 86.95\,$pA. Values are mean $\pm$ standard error of the mean across frames. Residual std measures how much the denoiser changed the input. The SNR proxy is the variance ratio of denoised output to residual expressed in dB. Higher is better for gradient correlation and edge energy ratio. The bottom three rows report $\log_{10}$ ratios of raw-to-denoised radial power averaged within each frequency band, where positive values indicate suppression and negative values indicate added power.

Metric	CIMP GAN (ours)	Thornley et al.	Noise2Atom
Residual std	0.030 ± 0.003	0.015 ± 0.002	0.060 ± 0.005
Residual mean	-0.019 ± 0.007	0.0001 ± 0.00005	-0.088 ± 0.010
Gradient correlation ↑	0.89 ± 0.01	0.69 ± 0.01	0.24 ± 0.01
Edge energy ratio ↑	0.64 ± 0.04	0.36 ± 0.01	0.58 ± 0.07
SNR proxy (dB)	6.1 ± 0.6	10.1 ± 0.6	-2.9 ± 0.3
PSD ratio, low-freq ($\log_{10}$)	-0.23 ± 0.06	-0.012 ± 0.001	0.69 ± 0.06
PSD ratio, mid-freq ($\log_{10}$)	0.08 ± 0.05	0.44 ± 0.02	-0.09 ± 0.03
PSD ratio, high-freq ($\log_{10}$)	0.57 ± 0.05	1.08 ± 0.02	0.87 ± 0.07

Figure 11 shows the full mean radial power spectrum of each denoiser averaged over the same 100 frames. The three frequency-band $\log_{10}$ ratios reported in Table 14 are computed by averaging these curves within their respective bands.

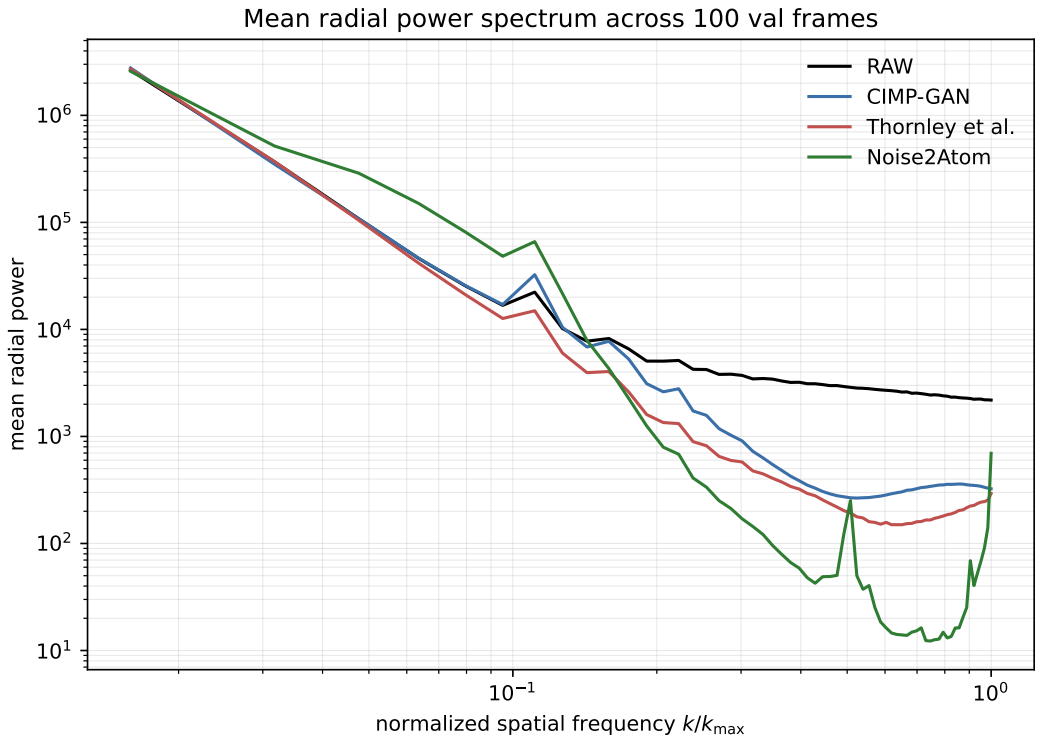
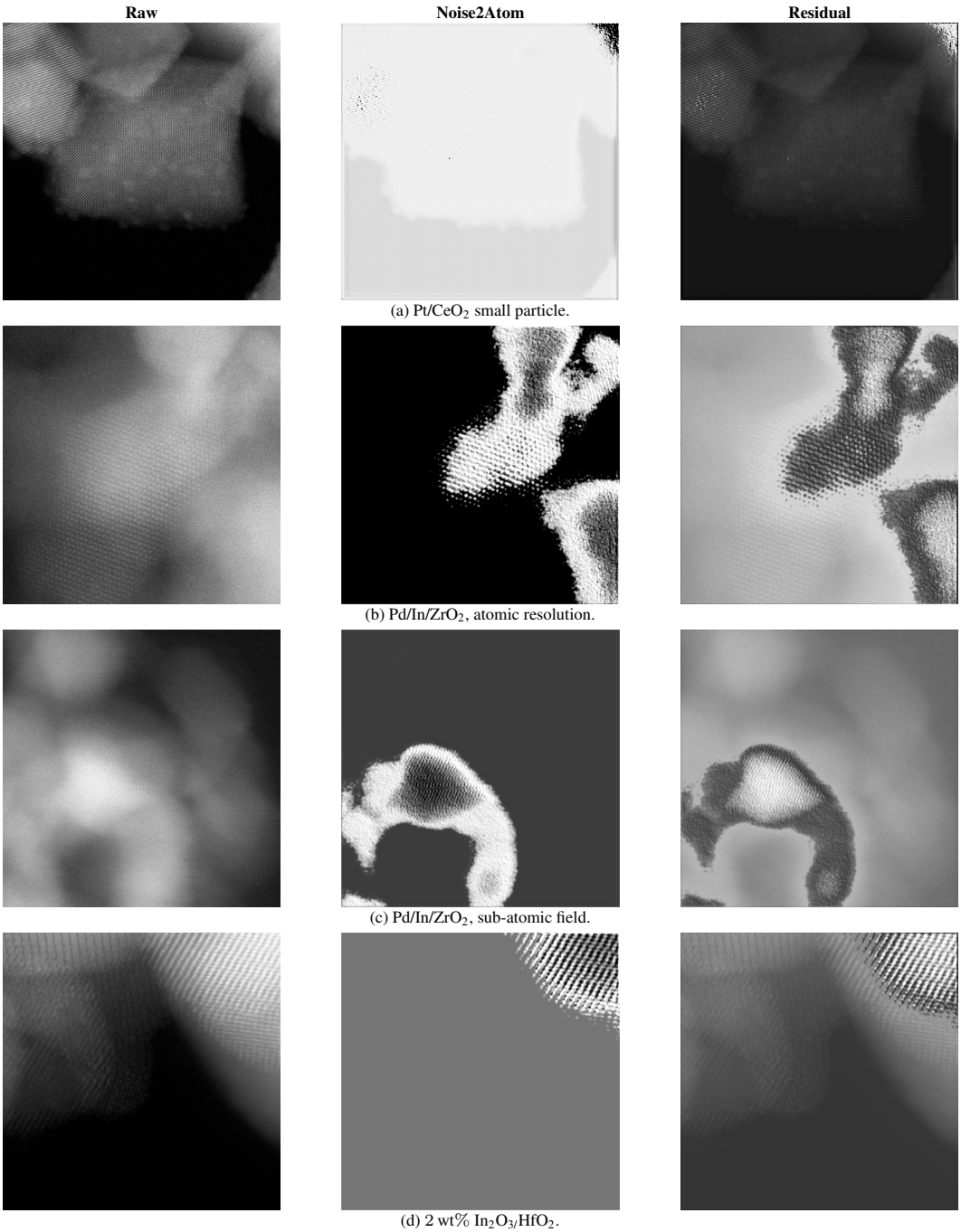


Figure 11: Mean radial power spectrum across 100 held-out validation frames for the raw input and each denoiser, plotted on log-log axes against normalized spatial frequency $k/k_{\max}$. The Thornley et al. baseline aggressively suppresses high-frequency power. CIMP GAN produces a more moderate spectral shift, with mild low-frequency amplification and less high-frequency attenuation than the Thornley et al. baseline. Noise2Atom adds mid-frequency power before crashing at high frequencies, with prominent peaks at the spatial frequencies of its learned atomic prior.

L NOISE2ATOM SAMPLE OUTPUTS

Figure 12 shows four representative held-out frames denoised by Noise2Atom, drawn from the same 100-frame validation set used for Table 14. Each row shows the raw acquisition, Noise2Atom output, and residual; panels a, c, and d show the non-atomic-content failure modes discussed in Section 5.3, while panel b shows the atomic-resolution regime for which Noise2Atom is best suited.



(a) Pt/CeO_2 small particle.

(b) $Pd/In/ZrO_2$, atomic resolution.

(c) $Pd/In/ZrO_2$, sub-atomic field.

(d) 2 wt% In_2O_3/HfO_2.

Figure 12: Four held-out validation frames denoised by Noise2Atom. Columns show raw acquisition, Noise2Atom output, and residual. Noise2Atom works best on clean atom-lattice content (b) and otherwise imprints periodic structure on non-atomic content (a, c) or collapses the image toward a near-uniform shade (d).

M USER STUDY

Five experimental microscopists submitted judgments, with uneven coverage across the 200 test triplets. The raters were not compensated in any way. Each rater independently chose between the Thornley et al. and CIMP GAN versions of every triplet they were shown, with left/right placement randomised per triplet. Their per-rater preference is reported in Table 5 of the main text, with Wilson 95% confidence intervals on each rate and a binomial p-value against the $50/50$ null. Here we focus on overlap and rater agreement, with corresponding confidence intervals and p-values.

Instructions to raters. Each rater was shown the following prompt at the start of the study:

> On each screen you'll see a raw noisy image at the top and two denoised versions of it below, labelled A and B. The two versions come from different denoising approaches, presented in a random left/right order so the comparison stays blind. Hold Shift and scroll over an image to zoom (up to $10\times$), drag to pan, double-click to reset.

Coverage. Every triplet has received at least one judgment. The distribution of raters per triplet is given in Table 15.

Table 15: Number of triplets covered by exactly k raters.

Raters per triplet	Triplets
1	29
2	79
3	70
4	21
5	1
Total	200

Pairwise agreement. For triplets judged by both members of a pair, the agreement rate is the fraction of overlapping triplets where the two raters chose the same denoiser. Cohen's κ corrects for chance agreement implied by each rater's marginal preference, so a κ near 0 indicates that the agreement rate is no better than would be expected from the marginals alone. We report κ with its asymptotic standard error and a one-sided p-value against the null $\kappa = 0$.

Table 16: Pairwise agreement on overlapping triplets, with Wilson 95% CI on agreement and Cohen's κ with its asymptotic standard error. The p-value is one-sided against $\kappa = 0$.

Pair	Overlap	Agreement (Wilson 95% CI)	Cohen's κ ($\pm$SE)	p vs κ=0
Rater 1, Rater 2	101	77 (76.2% [67.1, 83.5])	0.43 ± 0.10	0.01
Rater 1, Rater 3	76	59 (77.6% [67.1, 85.5])	0.33 ± 0.14	0.13
Rater 1, Rater 4	73	39 (53.4% [42.1, 64.4])	0.06 ± 0.12	0.36
Rater 1, Rater 5	36	22 (61.1% [44.9, 75.2])	0.17 ± 0.17	0.28
Rater 2, Rater 3	40	32 (80.0% [65.2, 89.5])	0.38 ± 0.19	0.18
Rater 2, Rater 4	33	18 (54.5% [38.0, 70.2])	-0.05 ± 0.20	0.56
Rater 2, Rater 5	14	9 (64.3% [38.8, 83.7])	0.19 ± 0.29	0.35
Rater 3, Rater 4	26	17 (65.4% [46.2, 80.6])	0.12 ± 0.24	0.38
Rater 3, Rater 5	13	12 (92.3% [66.7, 98.6])	0.63 ± 0.36	0.30
Rater 4, Rater 5	13	9 (69.2% [42.4, 87.3])	0.40 ± 0.25	0.18

Choice breakdown on overlapping triplets. Most disagreements involve one rater choosing the CIMP GAN where the other chose Thornley et al., rather than the symmetric reverse. The full 2×2 counts per pair are in Table 17, where columns are the second rater's choice and rows are the first rater's choice.

Table 17: Per-pair 2×2 counts of chosen denoiser on overlapping triplets. Rows are the first rater's choice, columns are the second rater's choice.

Pair	First \ Second	CIMP	Thornley
Rater 1, Rater 2	CIMP	60	7
	Thornley	17	17
Rater 1, Rater 3	CIMP	52	4
	Thornley	13	7
Rater 1, Rater 4	CIMP	27	24
	Thornley	10	12
Rater 1, Rater 5	CIMP	16	5
	Thornley	9	6
Rater 2, Rater 3	CIMP	28	2
	Thornley	6	4
Rater 2, Rater 4	CIMP	16	11
	Thornley	4	2
Rater 2, Rater 5	CIMP	7	2
	Thornley	3	2
Rater 3, Rater 4	CIMP	15	7
	Thornley	2	2
Rater 3, Rater 5	CIMP	11	1
	Thornley	0	1
Rater 4, Rater 5	CIMP	5	1
	Thornley	3	4

Unanimity at ≥ 3 raters. Of the 92 triplets judged by at least three raters, 47 (51.1%, Wilson 95% CI $[41.0, 61.1]$) showed unanimous agreement among the raters who saw them, with the remaining 45 split. A two-sided binomial test against the 50% null rate is not significant ($p = 0.92$), so unanimous agreement is no more or less common than a coin flip on the multiply-judged subset.

Four of the five raters individually prefer the CIMP GAN at 68–86% while one rater is indistinguishable from chance, and only the Rater 1 vs. Rater 2 pair has a κ that is significantly different from zero at $p < 0.05$. Other κ values are positive but small with confidence intervals straddling zero, so the per-triplet ratings are not interchangeable.

N LIMITATIONS

Boundary conditions are missing from the training data. Operators do not routinely record images with clipped intensities, saturated detectors, or other extreme boundary conditions, because such acquisitions are considered bad. As a result, the upper and lower bounds of each acquisition parameter are under-represented in the dataset, and the model has no examples from which to learn the corresponding image behaviours. This is visible in the gain sweep (Figure 2), where the model correctly reproduces the trend of increasing brightness and noise but fails to saturate at the high end, and in the offset sweep (Figure 2), where extreme values are poorly reproduced.

Scale and diversity relative to computer-vision datasets. To our knowledge, the 7,330 images released here constitute the largest paired image–metadata HAADF-STEM dataset published to date, and far exceed any comparable precedent in materials electron microscopy. It is nonetheless modest and narrow by computer-vision standards: ImageNet (Deng et al., 2009) contains 1.28 million natural images and has supported three orders of magnitude of model scaling beyond what our corpus can support. Future work will include amassing a higher volume and more diverse corpus of data from further research facilities.