Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability of atomic-column localization in scanning transmission electron microscopy under noise, annotation variation, and domain shift

Gourab Datta
University of Oklahoma
Sarah Sharif
University of Oklahoma
Yaser Banad
bana@ou.edu
University of Oklahoma

Article

Keywords: Scanning transmission electron microscopy, atomic-column localization, atom segmentation, deep learning, noise robustness, domain adaptation

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11485423/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Reliability of atomic-column localization in scanning transmission electron microscopy under noise, annotation variation, and domain shift

Gourab Datta[1,2], Sarah Sharif[1,2], Yaser Banad[1,2*]

[1]Intelligent Neuromorphic and Quantum Understanding for Innovative Research and Engineering (INQUIRE) Lab, University of Oklahoma, Norman, 73019, Oklahoma, U.S.A.
[2]School of Electrical and Computer Engineering, University of Oklahoma, Norman, 73019, Oklahoma, U.S.A.

*Corresponding author(s). E-mail(s): bana@ou.edu;
Contributing authors: gourab.datta-1@ou.edu; s.sh@ou.edu;

Abstract

Reliable atomic-column coordinates are essential for quantitative structural analysis in scanning transmission electron microscopy (STEM), yet strong performance on clean images does not ensure reliable localization under noise or across imaging domains. Localization errors can also propagate into coordinate-based structural measurements. Here, we evaluate automated atomic-site localization using Hybrid-STEMSeg, a compact convolutional neural network (CNN)–Transformer segmentation model, alongside established segmentation and microscopy-specific baselines. We test parent-held-out Sm-doped $BiFeO_3$ images under controlled corruption, examine how localization errors affect nearest-neighbor spacing, compare annotation targets, and assess transfer to external STEM datasets with and without target-domain adaptation. Model rankings depend on corruption, annotation, and imaging domain, showing that clean-image scores alone do not identify the most reliable model. Zero-shot transfer is limited, whereas target-domain adaptation improves held-out localization. This evaluation framework helps determine whether a localization model is suited to the imaging conditions and structural analyses for which it is intended.

Keywords: Scanning transmission electron microscopy, atomic-column localization, atom segmentation, deep learning, noise robustness, domain adaptation

1 Introduction

Atomic-resolution scanning transmission electron microscopy (STEM) connects local atomic structure to material behavior by resolving atomic columns and their spatial arrangement [1–5]. In complex oxides, small changes in atomic positions and local symmetry can accompany ferroic order, phase coexistence, and domain-wall behavior [6–8]. Studies of Sm-doped $BiFeO_3$ (Sm-BFO), for example, use atomic-scale structural information to examine the relationships among composition, polarization, and competing mechanisms [9, 10]. Accurate atomic-site coordinates are therefore the starting point for quantitative structural analysis.

Recovering these coordinates becomes difficult when noise, drift, scan distortion, contamination, or changes in acquisition conditions alter atomic-column contrast [5, 11]. Thresholding and other conventional image-processing methods provide useful starting points [12], but heterogeneous experimental images motivate learned localization methods [13, 14]. Fully convolutional networks and U-Net-style models support dense prediction [15, 16], while atrous-convolution and attention-based architectures extend the available spatial context [17–21]. Within electron microscopy, learned methods have enabled atom identification, defect analysis, and structure recognition [22–24]. TEMImageNet/AtomSegNet and AtomAI further provide practical tools for atom-level image analysis and coordinate extraction [25, 26].

A remaining difficulty is deciding when the recovered coordinates are reliable. Clean-image mask overlap, detection under synthetic noise, and transfer to a new material answer different questions. A broad predicted mask can overlap the target poorly while preserving its center; conversely, an apparently plausible segmentation can omit or displace sites needed for structural measurements. Robustness to common corruptions must also be evaluated directly rather than inferred from clean performance [27]. For quantitative STEM, these distinctions require an assessment that connects detection quality to coordinate errors and downstream structural descriptors.

Domain shift adds another constraint. Changes in material, contrast, atomic density, and annotation convention can invalidate a model selected on its source domain. Domain adaptation and limited-label learning offer established approaches to this problem [28–32], but their practical value depends on performance on target images excluded from model selection. The annotation itself also matters: changing the coordinate set used to construct a segmentation target changes the localization task, even when the Gaussian target width and matching tolerance remain fixed.

The localization-and-segmentation workflow of Eliasson and Erni links noise-controlled validation to experimental nanoparticle analysis [14]. Here, we address the complementary model-selection question: whether conclusions about localization remain valid when corruption, annotation, or imaging domain changes. We develop a reliability-oriented evaluation framework using Hybrid-STEMSeg, a compact CNN–Transformer segmentation model, alongside conventional and microscopy-specific baselines. The workflow (Fig. 1) connects mask prediction to atomic-site extraction and structural measurements. We first establish source-domain robustness with independent parent-held-out tests, then examine structural-error propagation, architecture cost, and dependence on annotation and imaging domain. Finally, held-out Foundry and Pt/Fe aberration-corrected STEM (AC-STEM) experiments test whether limited

target-domain supervision can recover localization after zero-shot transfer fails. This design connects three levels of evidence—site recovery, structural-measurement error, and held-out target performance—so that a model can be selected for its intended microscopy task rather than by clean mask overlap alone.

2 Results and discussion

2.1 From STEM images to atomic-site measurements

The workflow takes a STEM image to a probability map, extracts connected-component centroids, and evaluates both site detection and coordinate-based structural measurements (Fig. 1). A common post-processing rule is used across the primary mask benchmarks: a probability threshold of 0.5, no morphological cleanup, and greedy one-to-one matching within 6 pixels. Centroid-F1 measures recovered sites, while intersection-over-union (IoU) measures mask agreement; their definitions and sensitivity checks are given in Methods and Supplementary Table S16.

Sm-BFO provides the source-domain task, with Gaussian targets constructed from the center-of-mass (COM) atomic-site annotations (Fig. 2). The dataset contains 14 parent frames (full-field STEM images, each indexed by a composition key), which are the independent experimental units; all patches are extracted from these frames. Three repeated 9/2/3 parent train/validation/test splits provide the primary robustness estimate: test parents never contribute to training or checkpoint selection. A separate 11/3 train/validation split provides 4400 training patches and a fixed 300-patch validation pool for descriptive corruption and structural analyses. These complementary experiments distinguish independent testing from characterization of an already selected model.

Hybrid-STEMSeg combines convolutional feature extraction and decoding with a reduced-resolution spatial Transformer bottleneck (Fig. 3). It is trained for dense mask prediction; centroids are extracted afterward rather than optimized through a coordinate loss. U-Net, DeepLabV3+, and SegFormer provide architectural comparisons, and a matched Hybrid-NoTransformer model bypasses attention computation. AtomAI-UNet and AtomSegNet are included in the common-annotation benchmark. Full model and optimization details are given in Methods.

2.2 Localizing atomic sites under image degradation

The first test asks whether comparable clean detection performance implies comparable robustness. The primary Sm-BFO robustness result is based on three repeated parent-disjoint train/validation/test splits across the 14 parent keys. Each split uses 9 parents for training, 2 distinct parents for checkpoint selection, and 3 test parents that are never used for training or model selection. The split seeds are 42, 101, and 202. Table 1 reports mean centroid-F1 $\pm$ standard deviation across the three repeated splits. Clean performance is similar across architectures, whereas the hybrid variants retain markedly higher F1 than U-Net and DeepLabV3+ at severe Gaussian corruption. At $\sigma = 1.5$, Hybrid-STEMSeg achieves 0.637 ± 0.203 and Hybrid-NoTransformer 0.601 ± 0.213,

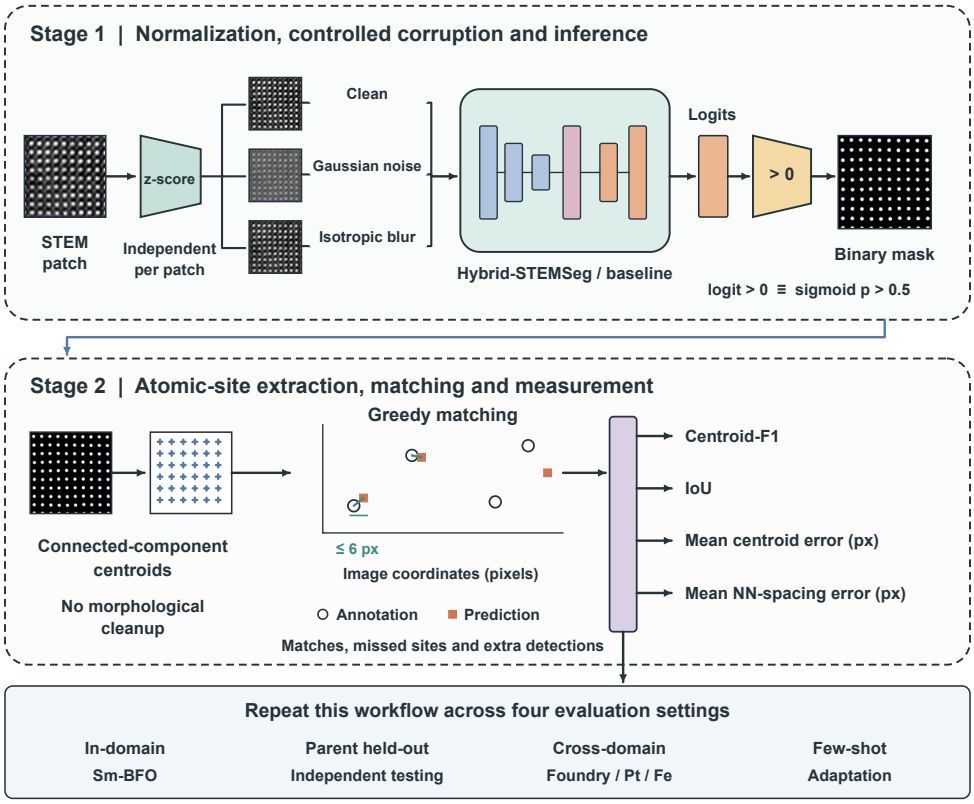
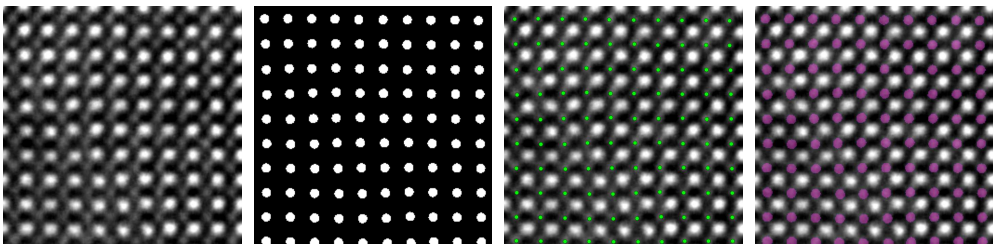


Fig. 1: Reliability-oriented evaluation workflow for atomic-site localization. STEM patches are independently z-score normalized and optionally subjected to Gaussian, Poisson, or mixed noise, or isotropic Gaussian blur before inference with Hybrid-STEMSeg or a comparison baseline. Model logits are binarized at zero (equivalent to a sigmoid probability threshold of 0.5), connected-component centroids are extracted without morphological cleanup, and predicted centroids are greedily matched to annotated sites within a 6-pixel tolerance. Performance is assessed using centroid-F1, IoU, mean centroid error, and mean nearest-neighbor spacing error across in-domain Sm-BFO evaluation, independent parent-held-out testing, cross-domain Foundry/Pt/Fe evaluation, and few-shot adaptation.



(a) Input STEM patch (b) Gaussian target mask (c) Centroid annotation (d) Model prediction

Fig. 2: Example Sm-BFO preprocessing and prediction workflow. Atomic-site coordinate annotations are converted into Gaussian target masks for dense segmentation training. During evaluation, predicted masks are post-processed to extract atomic-site centroids, which are compared with the ground-truth coordinates using centroid-F1.

compared with 0.415 ± 0.329 for SegFormer, 0.167 ± 0.082 for U-Net, and 0.151 ± 0.066 for DeepLabV3+.

Table 1: Independent parent-held-out Sm-BFO robustness. Values are mean centroid-F1 ± standard deviation across three repeated 9/2/3 parent train/validation/test splits (seeds 42, 101, and 202). Test parents are excluded from both training and checkpoint selection.

σ	Hybrid-STEMSeg	Hybrid-NoTransformer	U-Net	DeepLab V3+	SegFormer
0.0	0.752 ± 0.099	0.750 ± 0.098	0.757 ± 0.088	0.770 ± 0.123	0.769 ± 0.100
0.3	0.760 ± 0.093	0.760 ± 0.095	0.776 ± 0.105	0.760 ± 0.124	0.768 ± 0.104
0.6	0.762 ± 0.099	0.755 ± 0.095	0.634 ± 0.203	0.396 ± 0.233	0.755 ± 0.117
1.0	0.750 ± 0.114	0.731 ± 0.116	0.215 ± 0.147	0.146 ± 0.053	0.636 ± 0.231
1.5	0.637 ± 0.203	0.601 ± 0.213	0.167 ± 0.082	0.151 ± 0.066	0.415 ± 0.329
2.0	0.453 ± 0.310	0.395 ± 0.343	0.157 ± 0.071	0.151 ± 0.065	0.341 ± 0.323

These independent tests support the source-domain robustness conclusion. The lower clean F1 values in the repeated 9/2/3 splits relative to the fixed 11/3 validation pool are consistent with training on fewer parent frames and with between-parent composition heterogeneity. The complete independent-split severity summary is also provided in Supplementary Table S1 and Supplementary Fig. S1. The complete 300-patch validation pool remains useful as a matched descriptive analysis because it provides dense severity curves, structural-descriptor calculations, and direct continuity with the other Sm-BFO experiments reported below.

For the fixed 11-train/3-validation split, all 300 validation patches from the three held-out parents were evaluated with identical samples and corruption realizations across models. The descriptive trend is consistent with the independent parent-held-out

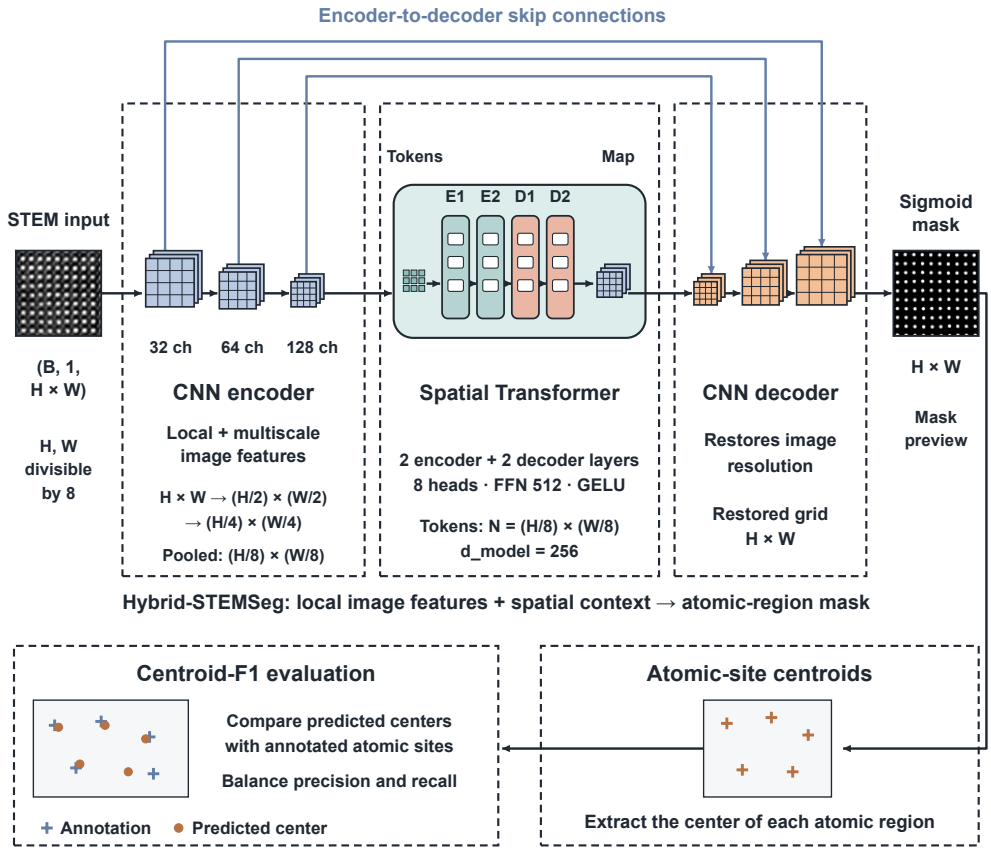


Fig. 3: Architecture of the Hybrid-STEMSeg segmentation model. An input STEM patch of size $H \times W$ is first processed by a convolutional encoder to extract local and multiscale image features. The resulting bottleneck representation is converted into spatial tokens of size $(H/8) \times (W/8)$ and passed through a compact spatial Transformer with two encoder layers and two decoder layers. A convolutional decoder then restores the spatial resolution and produces a sigmoid mask prediction, from which atomic-site centroids are extracted for centroid-F1 evaluation.

analysis: Hybrid-STEMSeg and Hybrid-NoTransformer remain stable through moderate-to-severe Gaussian corruption, whereas U-Net and DeepLabV3+ degrade sharply after $\sigma = 0.3$ and SegFormer declines at the highest severities. At $\sigma = 1.5$, the validation-pool F1 values are 0.8605, 0.8438, 0.4450, 0.0933, and 0.0934 for Hybrid-STEMSeg, Hybrid-NoTransformer, SegFormer, U-Net, and DeepLabV3+, respectively.

The evaluation-pool sensitivity analysis is provided in Supplementary Table S2 and Supplementary Fig. S2, while integrated corruption summaries and the complete validation-pool severity curves are reported in Supplementary Table S3 and Supplementary Fig. S3. Gaussian corruption in this pipeline is added after per-patch z-score normalization. For Hybrid-STEMSeg, F1 increases from 0.8835 on clean data to 0.9043 at $\sigma = 0.6$, accompanied by precision increasing from 0.8707 to 0.8926 and recall from 0.9029 to 0.9220. These values all come from the same single-stream Gaussian realization (seed 123). The increase is interpreted as an operating-point response of the fixed post-normalization inference pipeline rather than evidence that added noise improves the physical microscopy signal.

Poisson and mixed-corruption experiments were repeated on the same complete 300-patch Sm-BFO validation set. Supplementary Table S4 shows that the strong Gaussian robustness of the hybrid variants is not identical across corruption families. Hybrid-NoTransformer slightly exceeds Hybrid-STEMSeg under both reported non-Gaussian conditions, while SegFormer remains competitive under Poisson corruption but is lower under mixed corruption. U-Net and DeepLabV3+ degrade substantially.

These results reinforce that robustness should be interpreted as a property of a model–target–corruption combination rather than as a single transferable model attribute. The full Poisson and mixed-corruption results are also tabulated in Supplementary Table S4. As a complementary physically interpretable stress test, isotropic Gaussian blur was used to approximate progressive probe or point-spread-function (PSF) broadening. Across blur widths up to 2 pixels, centroid-F1 changed only modestly on the complete validation pool (for example, Hybrid-STEMSeg 0.8835 $\rightarrow$ 0.8848 and SegFormer 0.9730 $\rightarrow$ 0.9641), showing that the severe degradation observed under additive-noise conditions does not automatically extend to PSF-like broadening. The full blur sweep is reported in Supplementary Table S5 and Supplementary Fig. S4.

2.3 Preserving coordinate-based structural measurements

To determine whether localization robustness affects quantitative analysis, predicted centroids were propagated into localization and nearest-neighbor spacing measurements. Because instrument-specific spatial calibration is unavailable in the supplied manifests, the primary errors are reported in pixels. Table 2 summarizes the frame-averaged results across the three held-out Sm-BFO parent frames.

At Gaussian $\sigma = 0.6$, Hybrid-STEMSeg, Hybrid-NoTransformer, and SegFormer preserve sub-pixel centroid and nearest-neighbor errors, whereas U-Net and DeepLabV3+ show much larger geometric errors. Figure 4 traces this propagation over the Gaussian severity range. The complete structural-error table, including Poisson and mixed corruptions and the lattice-descriptor error, is provided in Supplementary Table S6.

These errors connect the detection benchmark to quantities used in lattice analysis. They do not establish calibrated strain or polarization accuracy: the matching threshold

Table 2: Downstream structural errors on the three held-out Sm-BFO parent frames. Values are frame-averaged means in pixels. Loc. denotes localization error and NN denotes nearest-neighbor spacing error.

Model	Clean	Loc. 0.6	Loc. 1.0	NN 0.6	NN 1.0
Hybrid-STEMSeg	0.29	0.56	0.87	0.32	0.40
Hybrid-NoTransformer	0.33	0.49	0.63	0.31	0.29
SegFormer	0.31	0.53	0.81	0.33	0.84
U-Net	0.29	1.83	2.17	3.63	25.78
DeepLabV3+	0.33	2.47	3.26	29.24	25.86

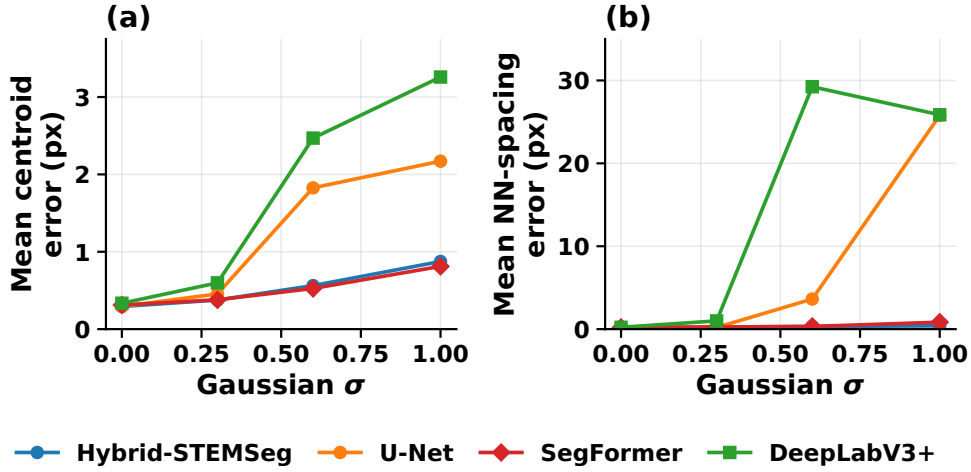


Fig. 4: Effect of additive Gaussian noise on atomic-site localization and nearest-neighbor spacing in the held-out Sm-BFO frames. (a) Mean centroid localization error. (b) Mean nearest-neighbor spacing error. Curves show Hybrid-STEMSeg, U-Net, DeepLabV3+, and SegFormer, as indicated in the legend. Each point is the frame-averaged error at the indicated Gaussian noise standard deviation, σ, added after patch-wise z-score normalization. Errors are reported in pixels because instrument-specific spatial calibration is unavailable; lower values indicate closer agreement with the reference coordinates and spacings.

and descriptor errors are expressed in pixels, and a validated physical calibration and reference lattice would be needed for those applications.

2.4 Assessing the contribution and cost of attention

Hybrid-NoTransformer closely tracks Hybrid-STEMSeg in both the independent parent-held-out analysis and the complete validation-pool experiment. In the independent splits, the mean F1 difference at $\sigma = 1.5$ is modest relative to split-to-split variability (0.637 versus 0.601). In the complete validation pool, the difference is 0.0168. The measured median GPU latency is 5.07 ms for Hybrid-STEMSeg versus 2.42 ms for Hybrid-NoTransformer, a relative penalty of about 109%. The NoTransformer variant also has lower centroid and nearest-neighbor errors at Gaussian $\sigma = 1.0$ and slightly higher F1 under the reported Poisson and mixed corruptions. The robustness–efficiency tradeoff is reported in Supplementary Table S11 and Supplementary Fig. S8, with complete model-size, latency, and memory measurements in Supplementary Table S12. These results argue against treating the Transformer bottleneck as the sole source of robustness.

Both implementations register 4,287,425 parameters, but the NoTransformer forward path uses only 471,617 because its registered attention modules are bypassed (Supplementary Table S12). The matched comparison therefore tests the cost and effect of attention computation, rather than comparing two physically pruned parameter inventories. Hybrid-STEMSeg has fewer registered parameters than U-Net (24,430,097), but its measured latency is higher. Model size, active computation, and inference speed should consequently be reported separately.

2.5 Testing annotation and imaging-domain dependence

To test whether the Sm-BFO robustness ranking depended on target representation, all models were evaluated under a common `xy_atms` atomic-site label definition on the same 300-patch Sm-BFO validation pool. This common-target experiment also enabled a fairer comparison with AtomAI-UNet and AtomSegNet within the `xy_atms` setting. Table 3 shows a qualitatively different robustness ordering from the COM-target benchmark. AtomAI-UNet has the highest clean F1, whereas SegFormer retains the highest F1 at $\sigma = 0.6$, 1.0, and 1.5. Hybrid-STEMSeg decreases from 0.8756 on clean data to 0.0682 at $\sigma = 1.5$. The repeated value 0.0867 at $\sigma = 1.5$ for DeepLabV3+ and AtomSegNet is not a duplicated table cell: the 300-patch validation pool contains 26 empty masks, so a model that predicts no foreground anywhere receives a score of $26/300 \approx 0.0867$ from those empty-mask cases alone. The small non-monotonic change for U-Net between $\sigma = 0.6$ and 1.0 occurs within this same failure-floor regime and reflects the underlying CSV values rather than rounding.

The COM-target and `xy_atms` coordinate keys are not identical: the audit found no parent for which the two coordinate sets were identical, and site counts differ for multiple parents. The ranking reversal therefore reflects sensitivity to target definition and annotation semantics and cannot be attributed to rasterization geometry alone. Both settings use a Gaussian rasterization width of $\sigma_{\mathrm{mask}} = 3$ pixels and the same

Table 3: Equal-label Sm-BFO comparison using the `xy_atms` target representation. All models are evaluated with the same target definition and centroid-matching protocol.

Model	Clean F1	Clean IoU	$\sigma = 0.6$	$\sigma = 1.0$	$\sigma = 1.5$
Hybrid-STEMSeg	0.8756	0.7584	0.5415	0.0578	0.0682
U-Net	0.8737	0.7304	0.0623	0.0843	0.0116
DeepLabV3+	0.8840	0.7444	0.1827	0.0717	0.0867
SegFormer	0.8810	0.7369	**0.8375**	**0.7198**	**0.6434**
AtomAI-UNet	**0.9147**	**0.8371**	0.3981	0.0165	0.0215
AtomSegNet	0.9014	0.8201	0.5919	0.0723	0.0867

6-pixel centroid-matching tolerance. The full common-target results and curves are provided in Supplementary Table S7 and Supplementary Fig. S5.

The 299-image Foundry Wei atom-localization benchmark was next evaluated [33]. A stratified 209/45/45 train/validation/test split (seed 42) was used for full target-domain training, with the test split used only for final reporting. Foundry target masks were generated by placing a Gaussian $\exp(-r^2/(2\sigma^2))$ at each annotated atom with $\sigma = 2.5$ pixels and binarizing at 0.25. Under full Foundry training, U-Net achieved the highest held-out test F1 (0.9036), followed by Hybrid-STEMSeg (0.7389), DeepLabV3+ (0.6774), and SegFormer (0.6131). This ordering differs from the Sm-BFO corruption ranking; the full target-domain training summary is provided in Supplementary Table S8.

Figure 5 places the descriptive Sm-BFO COM-target sweep, the common-target `xy_atms` experiment, and the Foundry full-training noise sweep side by side. For the Foundry panel, the model trained with the fixed 209/45/45 split (split seed 42) was evaluated on the 45-image held-out test split over Gaussian $\sigma \in \{0, 0.3, 0.6, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0\}$. For each severity, the evaluation code reseeds one generator to 42 and then draws noise sequentially across the 45 test images. The comparison makes clear that no single architecture ordering persists across all three regimes.

2.6 Adapting to independent experimental domains

Having established that full-training rankings depend on domain, we next tested whether source-trained models could be adapted with limited target labels. A separate experiment initialized each model from its Sm-BFO-trained checkpoint and adapted it with $K \in \{1, 3, 5, 10\}$ labeled Foundry training frames. Table 4 reports held-out test performance over seeds 101, 202, and 303. Strict zero-shot performance is weak for all models. At $K = 10$, Hybrid-STEMSeg reaches 0.575 ± 0.044 F1, U-Net 0.502 ± 0.013, DeepLabV3+ 0.435 ± 0.020, and SegFormer 0.351 ± 0.089. Very small label budgets show substantial split sensitivity, particularly at $K = 1$. The complete Foundry few-shot grid and adaptation curves are provided in Supplementary Table S9 and Supplementary Fig. S6.

The Pt/Fe AC-STEM dataset provides a materially and experimentally distinct external test. A fixed five-frame test subset was reserved before adaptation; for each

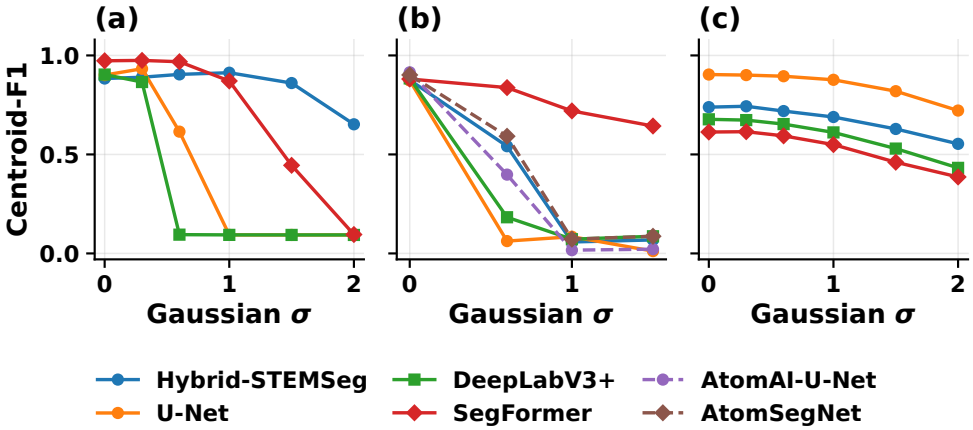


Fig. 5: Domain- and target-dependent robustness. (a) Sm-BFO COM-target evaluation on all 300 held-out patches, (b) common-target Sm-BFO `xy_atms` evaluation, and (c) Foundry evaluation after full target-domain training. The model ordering changes across regimes.

Table 4: Sm-BFO-to-Foundry cross-domain adaptation. $K = 0$ denotes strict zero-shot transfer; $K \geq 1$ values are mean $\pm$ standard deviation across seeds 101, 202, and 303 on the 45-image held-out Foundry test set.

K	Metric	Hybrid	U-Net	DeepLab	SegFormer
0	F1	0.205	0.162	0.121	0.182
	IoU	0.074	0.066	0.054	0.076
1	F1	0.389±0.335	0.179±0.120	0.065±0.070	0.362±0.344
	IoU	0.371±0.114	0.299±0.058	0.250±0.026	0.315±0.062
3	F1	0.398±0.105	0.147±0.131	0.051±0.038	0.233±0.056
	IoU	0.394±0.037	0.285±0.028	0.258±0.027	0.306±0.042
5	F1	0.430±0.060	0.200±0.179	0.108±0.133	0.273±0.042
	IoU	0.402±0.031	0.316±0.077	0.287±0.064	0.344±0.022
10	F1	**0.575±0.044**	0.502±0.013	0.435±0.020	0.351±0.089
	IoU	**0.507±0.025**	0.464±0.018	0.441±0.024	0.390±0.057

seed and label budget, adaptation and checkpoint selection used only the remaining non-test frames. Under strict zero-shot evaluation, all four Sm-BFO-trained models have centroid-F1 and IoU equal to 0.000 to three decimals. Few-shot target-domain adaptation recovers performance on the independent test subset, as summarized in Table 5. At $K = 10$, Hybrid-STEMSeg reaches 0.636 ± 0.012 test F1, compared with 0.588 ± 0.052 for DeepLabV3+, 0.574 ± 0.010 for U-Net, and 0.533 ± 0.017 for SegFormer. The full held-out-test grid is provided in Supplementary Table S10 and Supplementary Fig. S7.

Table 5: Sm-BFO-to-Pt/Fe AC-STEM adaptation evaluated on a fixed five-frame held-out test subset. $K = 0$ denotes strict zero-shot transfer; $K \geq 1$ values are mean $\pm$ standard deviation across seeds 42, 101, and 202. Adaptation and checkpoint selection use only non-test frames.

K	Metric	Hybrid	U-Net	DeepLab	SegFormer
0	F1	0.000	0.000	0.000	0.000
	IoU	0.000	0.000	0.000	0.000
1	F1	0.340 ± 0.187	0.250 ± 0.109	0.072 ± 0.064	0.204 ± 0.147
	IoU	0.130 ± 0.104	0.077 ± 0.041	0.036 ± 0.039	0.046 ± 0.037
3	F1	0.578 ± 0.038	0.353 ± 0.141	0.320 ± 0.205	0.405 ± 0.051
	IoU	0.259 ± 0.006	0.126 ± 0.055	0.151 ± 0.101	0.133 ± 0.026
5	F1	0.593 ± 0.012	0.406 ± 0.135	0.548 ± 0.032	0.472 ± 0.011
	IoU	0.259 ± 0.010	0.155 ± 0.057	0.251 ± 0.017	0.204 ± 0.022
10	F1	0.636 ± 0.012	0.574 ± 0.010	0.588 ± 0.052	0.533 ± 0.017
	IoU	0.292 ± 0.006	0.239 ± 0.005	0.255 ± 0.022	0.240 ± 0.009

The two external adaptation experiments are compared in Fig. 6. They show a consistent qualitative pattern: strict zero-shot transfer is weak, while target-domain supervision recovers substantial localization performance, with high variance at the smallest Foundry label budgets.

Representative Pt/Fe predictions before and after five-shot adaptation are shown in Fig. 7. On the illustrated frame, the zero-shot Hybrid-STEMSeg sigmoid map has a maximum value of 0.377, so no pixel reaches the 0.5 probability threshold and the corresponding binary prediction is empty. Adaptation recovers visible detections.

Full target-domain training and few-shot transfer thus represent different deployment choices. U-Net is strongest after full Foundry training, whereas Hybrid-STEMSeg has the highest mean F1 at the larger adaptation budgets evaluated here. The latter observation does not establish universal label efficiency: only three adaptation seeds were evaluated, the smallest Foundry budgets have large variance, and target validation labels are also used for checkpoint selection. Here, K counts optimization frames, not the total target annotation effort.

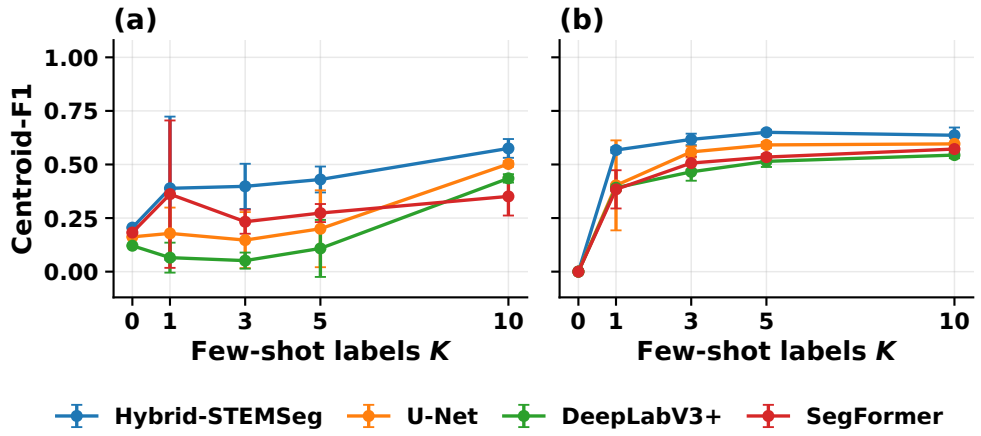
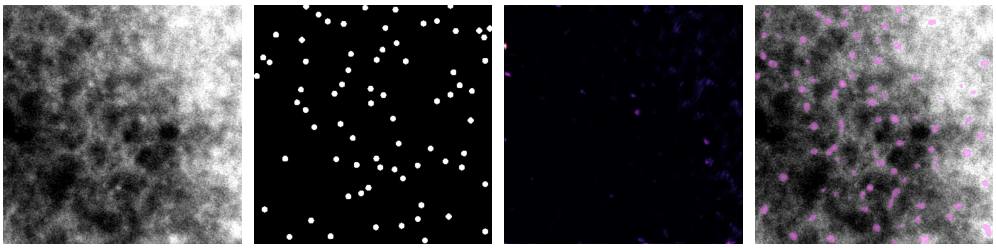


Fig. 6: Cross-domain few-shot adaptation from Sm-BFO checkpoints to (a) the Foundry held-out test set and (b) the fixed Pt/Fe AC-STEM held-out test subset. Error bars denote standard deviation across the reported random seeds for $K \geq 1$.



(a) External input (b) Expert annotation (c) Zero-shot sigmoid map (d) After 5-shot adaptation

Fig. 7: Representative external Pt/Fe AC-STEM prediction before and after five-shot adaptation. Panel (c) shows the pre-threshold zero-shot Hybrid-STEMSeg sigmoid map for the same frame as panel (a); its color scale spans 0 to the observed maximum of 0.377. Because no pixel reaches the 0.5 probability threshold, the corresponding binary zero-shot prediction is empty. Panel (d) shows the nonempty prediction after five-shot adaptation.

2.7 Scope and implications for quantitative STEM

The experiments support a sequence of deployment checks: establish detection on independent source data, quantify corruption response and coordinate errors, test annotation and post-processing dependence, and evaluate transfer on held-out target images. The main practical outcome is that a favorable score in one of these settings cannot replace evaluation in the others. In particular, source-domain noise robustness neither guarantees zero-shot transfer nor fixes the architectural ranking under a different annotation definition.

Several limitations constrain the generality of the conclusions. The Sm-BFO dataset contains only 14 parent composition keys. The complete-pool 11-train/3-validation analysis has no separate test partition, so the headline robustness claim is instead supported by three repeated 9/2/3 parent train/validation/test splits with independent test parents. Within this protocol, only three repeated splits were feasible, and parent-level uncertainty remains substantial. Controlled Gaussian, Poisson, and mixed corruptions do not reproduce all experimental failure modes such as nonstationary drift, scan distortion, contamination, thickness variation, or microscope-specific transfer functions. Because the archived Sm-BFO shards are already normalized, a raw-intensity pre-normalization Gaussian twin experiment was not available; the Gaussian diagnostic therefore characterizes the fixed post-normalization inference pipeline. Spatial calibration was not available in the supplied manifests, so downstream structural errors are not instrument-calibrated physical displacements. Pt/Fe AC-STEM and Foundry provide materially distinct external tests, but broader validation across microscopes and material classes remains necessary. Finally, the adaptation studies use three random seeds, supporting effect-size and variability reporting more strongly than formal claims of statistical superiority.

The common-target experiment further shows that changes in target definition and annotation semantics can change model ranking. Broad statements that one model is the most robust atom detector are therefore inappropriate without specifying the domain, annotation target, preprocessing, and evaluation protocol.

The resulting framework provides a reproducible way to qualify atomic-site localization for a stated imaging task. It supports selecting a model on the coordinates and structural quantities needed in that task, with independent target evaluation determining whether adaptation is sufficient.

3 Methods

3.1 Sm-BFO STEM dataset and annotation representation

The primary source-domain dataset consists of atomic-resolution Sm-BFO STEM images with atomic-site coordinate annotations. The data contain 14 parent frames (full-field STEM images, each indexed by a composition key), which are the independent experimental units; all patches are extracted from these frames. Splitting is performed at the parent-frame/composition-key level before patch extraction: 11 parent keys contribute 4400 training patches and three disjoint parent keys contribute 300 validation patches. The corresponding annotated site counts are 290,813 for training and 65,289

for validation. This descriptive 11/3 split has no dedicated test partition; the primary robustness analysis instead uses the three separate 9/2/3 splits described above. Split-integrity checks found no parent-key overlap and no exact cross-split duplicate patches.

Coordinate annotations are converted into dense Gaussian target masks centered at the annotated sites for the primary COM-target segmentation task. A second common-target analysis uses the `xy_atms` annotations so that the generic models and the AtomAI/AtomSegNet baselines are evaluated against the same `xy_atms` definition. Both COM and `xy_atms` masks use a Gaussian rasterization width of $\sigma_{\text{mask}} = 3$ pixels, the same split seed, loss, logit threshold, and 6-pixel centroid-matching tolerance. However, the coordinate sets themselves are not identical: no parent has identical COM and `xy_atms` coordinates, and site counts differ for multiple parents. Consequently, comparisons between the two tasks probe target-definition and annotation-semantic sensitivity, not rasterization width alone. The model is trained as a segmentation system but evaluated at the centroid level after post-processing.

For the Sm-BFO experiments, patches of size 256×256 pixels are normalized independently by z-score,

$$I_{\text{norm}} = \frac{I - \mu_I}{\sigma_I + \epsilon}, \tag{1}$$

where I is the input patch, μ_I and σ_I are the patch mean and standard deviation, and ϵ is a small constant. The stored evaluation shards are already z-scored. In the primary Gaussian-noise implementation, additive noise is applied after this normalization step, in normalized intensity units. Accordingly, nominal Gaussian σ should be interpreted as a stress-test parameter of the fixed inference pipeline rather than a calibrated physical detector-noise level.

Figure 2 illustrates the target construction and prediction stages.

3.2 Hybrid-STEMSeg architecture

The evaluated model, referred to as Hybrid-STEMSeg, is a compact hybrid CNN–Transformer segmentation model for dense atomic-column mask prediction. The architecture is illustrated in Fig. 3. It consists of a convolutional encoder–decoder with a small spatial Transformer module inserted at the bottleneck. The convolutional encoder uses three stages with 32, 64, and 128 channels and max-pooling to reach the 1/8-resolution bottleneck. The bottleneck is linearly projected to tokens with $d_{\text{model}} = 256$ and processed by two Transformer encoder layers and two decoder layers, each using 8 attention heads, feed-forward dimension 512, and GELU activations. A linear projection maps the tokens back to a feature map, and the convolutional decoder restores spatial resolution with skip connections for single-channel mask prediction. This design is related to existing hybrid CNN–Transformer segmentation architectures [20, 21], and is used here as a practical model for evaluating robustness rather than as a fundamentally new segmentation architecture.

For an input image of size $H \times W$, where H and W are divisible by 8, the bottleneck token grid has size $(H/8) \times (W/8)$. Thus, for a 256×256 Sm-BFO patch, the Transformer operates on a 32×32 grid, corresponding to 1024 spatial tokens. For variable-size external frames, the same rule is applied after cropping the spatial dimensions to be

divisible by 8. In the present implementation, the maximum sequence length is capped at 4096 tokens, imposing $(H/8)(W/8) \leq 4096$ at the 1/8 bottleneck resolution; for square inputs this corresponds to 512×512 pixels.

The Transformer bottleneck contains two encoder layers and two decoder layers. Attention is therefore applied only at the reduced-resolution bottleneck rather than throughout the full image-resolution feature hierarchy. This choice limits the computational cost of attention while retaining convolutional encoding and decoding for local feature extraction and spatial recovery. Hybrid-STEMSeg is trained using the segmentation objective described below, and the matched Hybrid-NoTransformer variant isolates the engineering tradeoff associated with the bottleneck attention module.

The final output is a single-channel probability map,

$$\hat{M} = \sigma(f_\theta(I_{\text{norm}})), \tag{2}$$

where f_θ denotes the segmentation network and $\sigma(\cdot)$ is the sigmoid function. Atomic-site centroids are extracted from the predicted probability map using a fixed post-processing procedure applied consistently across all learned segmentation models. Thus, centroid-level information is used during evaluation and post-processing, while the model itself is trained as a dense segmentation network.

3.3 Baseline and ablation models

Hybrid-STEMSeg was compared with U-Net, DeepLabV3+, and SegFormer, representing convolutional encoder–decoder segmentation, atrous-convolution semantic segmentation, and Transformer-based segmentation, respectively [16, 17, 21]. The matched Hybrid-NoTransformer variant bypasses the spatial Transformer bottleneck during the forward pass while retaining the convolutional encoder–decoder path, providing a direct architecture tradeoff comparison. For the common-target `xy_atms` experiment, AtomAI-UNet and AtomSegNet were additionally evaluated under the shared target definition. The independent parent-held-out comparison is reported in Table 1, with model-specific training details documented in the Supplementary Information.

3.4 Training protocol

Within each Sm-BFO experiment, learned models used the same parent-level data split and evaluation samples. The primary segmentation objective combined binary cross-entropy and Dice loss,

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}. \tag{3}$$

where $\mathcal{L}_{\text{BCE}}$ is binary cross-entropy with positive-class weight 18 and $\lambda_{\text{Dice}} = 0.5$ weights the Dice term [34]. The principal Sm-BFO mask benchmarks use this segmentation objective consistently across the compared models.

For the complete-pool 11/3 Sm-BFO split, model selection was performed using the three held-out validation parents. For the independent robustness analysis, each of the

three repeated 9/2/3 parent splits used only its two validation parents for checkpoint selection, and the three test parents were evaluated only after model selection. Across the primary mask benchmarks, Hybrid-STEMSeg, Hybrid-NoTransformer, U-Net, DeepLabV3+, and SegFormer used Adam with learning rate 2×10^{-4}, batch size 8, up to 100 epochs, early-stopping patience 15, BCE+Dice loss with positive-class weight 18 and Dice weight 0.5, logit threshold 0, and a 6-pixel matching tolerance. Architecture-specific initialization/backbones were retained (including ImageNet ResNet34 encoders for U-Net/DeepLabV3+ and MiT-B0 pretraining for SegFormer). Unless otherwise stated, reported uncertainty reflects variability across the evaluation units specified for each experiment, such as random seeds, image samples, or few-shot splits. For experiments performed with a single trained checkpoint, results are reported without implying multi-seed training variability. The fixed run configuration and software environment are documented in Supplementary Tables S13 and S14, and the per-model fairness and hyperparameter record is summarized in Supplementary Table S15.

3.5 Evaluation metrics

Two complementary metrics were used: pixel-level intersection-over-union (IoU) and centroid-based F1. IoU measures overlap between the predicted binary mask $\hat{B}$ and the ground-truth binary mask B:

$$\mathrm{IoU} = \frac{|\hat{B} \cap B|}{|\hat{B} \cup B|}. \tag{4}$$

IoU is useful for assessing mask morphology, but it can be sensitive to differences in predicted atomic-column thickness.

Centroid-F1 was used as the primary task-specific metric because atomic-site coordinates are the key output for downstream crystallographic analysis. Models output binary-segmentation logits; the primary mask path binarizes at logit 0 (sigmoid probability 0.5). Connected components are obtained with two-dimensional connectivity 2 using `skimage.measure.label`; no size filter, morphological opening/closing, hole filling, or border-component removal is applied. Component centroids are computed using `regionprops(...).centroid`. Predicted and ground-truth centroids are matched greedily by nearest-neighbor distance using a pairwise distance matrix, with each ground-truth site matched at most once. The default tolerance is 6 pixels for Sm-BFO, Pt/Fe, and Foundry. Unmatched predictions are false positives and unmatched ground-truth sites are false negatives; merged or split components receive no additional repair beyond the connected-component and matching rules. Because the tolerance is fixed in pixels rather than calibrated physical units, cross-dataset scores are treated as task-specific detection metrics rather than equivalent physical-displacement thresholds. A probability-threshold and matching-tolerance sensitivity analysis is reported in Supplementary Table S16. That grid tests ranking stability under a separate per-patch Gaussian realization, reseeded to $123 + i$ for patch i; its default-point scores are not replacements for the single-stream severity results. Let TP, FP, and FN denote the number of matched predictions, unmatched predictions, and unmatched ground-truth

sites, respectively. Precision, recall, and centroid-F1 are defined as

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{F1} = \frac{2PR}{P + R}. \tag{5}$$

Centroid-F1 therefore evaluates coordinate-level recovery rather than pixel-level mask overlap. This distinction is important because a model may correctly recover atomic-site centers while producing masks with different thickness, shape, or Gaussian target width. For the complete 300-patch Gaussian severity sweep and the post-processing sensitivity grid, the reported F1 is the arithmetic mean of per-patch F1 values. Consequently, it need not equal the harmonic mean of the reported mean precision and mean recall. A sensitivity analysis varying the probability threshold from 0.3 to 0.7 and the matching tolerance from 2 to 10 pixels shows that the qualitative ordering at Gaussian $\sigma = 1.0$ is stable across thresholds 0.3–0.7 and tolerances 4–10 pixels; only the strictest 2-pixel tolerance swaps the two hybrid variants while leaving both above SegFormer, U-Net, and DeepLabV3+.

3.6 Corruption-robustness evaluation

To evaluate robustness beyond clean validation performance, the held-out Sm-BFO validation images were corrupted using controlled evaluation-time noise and degradation settings. The main corruption family was additive Gaussian noise with increasing standard deviation. Additional corruption families included Poisson noise and mixed corruption settings. These corruptions were applied only during evaluation; the models were not retrained or fine-tuned separately for each corruption level. The evaluation shards are stored after per-patch z-score normalization, and Gaussian corruption is added afterward in this normalized space. The robustness results should therefore be interpreted as properties of the fixed preprocessing, prediction, thresholding, and centroid-extraction pipeline rather than as architecture-only or physically calibrated noise responses.

For each corruption severity, centroid-F1 and IoU were computed on matched corrupted inputs. In the complete 300-patch Gaussian sweep, one generator is seeded once at 123 and advanced across patches. The post-processing sensitivity grid instead reseeds the generator to $123+i$ for each patch i, with the same checkpoints, aggregation, default probability threshold, and matching tolerance. At $\sigma = 1.0$, the severity-sweep F1 values are 0.9127 (Hybrid-STEMSeg) and 0.9096 (Hybrid-NoTransformer); the separate grid gives 0.9136 and 0.9019. The two realizations are kept separate throughout the analysis. The headline Gaussian result uses the repeated 9/2/3 parent train/validation/test protocol described above. The complete 300-patch validation pool is retained as a descriptive analysis with dense severity sampling and structural metrics, and a 100-patch single-parent subset is used only for evaluation-pool sensitivity. In addition, isotropic Gaussian blur with $\sigma_{\mathrm{blur}} \in \{0, 0.5, 1.0, 1.5, 2.0\}$ pixels is used as a PSF-broadening stress test. Because the independent unit is the parent frame, patch counts are not treated as independent experimental replications.

3.7 Foundry benchmark and few-shot protocol

The independent Foundry Wei benchmark contains 299 labeled STEM images [33]. Images were stratified into 209 training, 45 validation, and 45 held-out test images using seed 42. Foundry target masks were generated by placing a Gaussian $\exp(-r^2/(2\sigma^2))$ at each annotated atom with $\sigma = 2.5$ pixels and binarizing the resulting mask at 0.25. For full target-domain training, models were optimized on the training split for up to 100 epochs with validation-based early stopping and were evaluated once on the held-out test set. The auxiliary Foundry Gaussian-noise sweep in Fig. 5(c) uses this 45-image held-out test split and $\sigma \in \{0, 0.3, 0.6, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0\}$. For each severity, the evaluation code reseeds one generator to 42 and then draws noise sequentially across the 45 test images.

For cross-domain adaptation, each model was initialized from its Sm-BFO-trained checkpoint. Target-domain label budgets of $K \in \{1, 3, 5, 10\}$ were sampled only from the Foundry training split using seeds 101, 202, and 303. For a given seed and K, the same target frames were used across models. The validation split was used for checkpoint selection, and the 45-image test set was reserved for final evaluation. A 6-pixel centroid-matching tolerance was used for this benchmark.

3.8 External Pt/Fe AC-STEM evaluation

To evaluate transfer beyond the Sm-BFO source domain, the trained checkpoints were tested on a real Pt/Fe AC-STEM dataset with expert annotations from a public single-atom detection study [13]. This dataset differs from the Sm-BFO data in material system, contrast statistics, object morphology, imaging context, and annotation semantics. It therefore provides a useful single-dataset test of external-domain behavior, but it should not be interpreted as evidence of broad cross-material generalization.

The external images were used at their available frame sizes after cropping the spatial dimensions to be divisible by 8. Expert point annotations were converted into Gaussian target masks for fine-tuning and evaluation using the same mask-generation procedure applied to the source-domain data. In the few-shot adaptation experiments, a Gaussian label width of $\sigma_{\text{mask}} = 2.5$ pixels was used for the target-domain masks.

The external evaluation reserves a fixed five-frame test subset before any target-domain adaptation. Strict zero-shot transfer applies each Sm-BFO-trained checkpoint directly to this test subset. For few-shot adaptation, label budgets $K \in \{1, 3, 5, 10\}$ are sampled from the remaining 19 non-test frames using seeds 42, 101, and 202. For each seed and K, K frames are used for adaptation and the remaining non-test frames are used for checkpoint selection (18, 16, 14, and 9 validation frames for $K = 1, 3, 5, 10$, respectively). The fixed five test frames are never used for optimization or checkpoint selection and are used only for final reporting.

Fine-tuning was initialized from the Sm-BFO-trained checkpoint for each model. Batch size was set to one to support variable frame sizes, and batch-normalization layers were kept frozen during fine-tuning. Freezing batch normalization avoided unstable running-statistic updates under very small batch sizes and reduced failures in architectures with small intermediate feature maps. The external zero-shot and few-shot

adaptation results are summarized in Table 5 and Fig. 6; representative predictions are shown in Fig. 7.

3.9 Statistical analysis

The independent source-domain unit for Sm-BFO is the parent frame rather than an individual patch or atom. The headline robustness analysis reports mean and standard deviation across three repeated parent-disjoint 9/2/3 train/validation/test splits. These repeated partitions can reuse parent frames across runs and are not three independent datasets. Three splits remain a small sample, so the manuscript emphasizes effect sizes, consistency of qualitative trends, and descriptive uncertainty rather than formal claims of statistical superiority. The three-parent validation-pool analysis is retained for dense severity-response and structural-descriptor characterization, but its patches are not treated as independent experimental replications.

For Pt/Fe AC-STEM and Foundry few-shot adaptation, results are reported as mean ± standard deviation across random seeds for each label budget and are evaluated on held-out test data. Pt/Fe uses seeds 42, 101, and 202 with a fixed five-frame test subset; Foundry uses seeds 101, 202, and 303 with its fixed 45-image test split. The small number of seeds exposes split sensitivity at very low K but does not justify strong significance claims. Exploratory frame-level uncertainty estimates compare mean centroid-localization errors, not median errors, for the complete Sm-BFO validation pool. They are provided in Supplementary Table S17 and are not used for formal superiority claims.

3.10 Reproducibility

Checkpoints, split identifiers, corruption settings, random seeds, and metric outputs were preserved for all reported experiments. Sm-BFO split integrity was checked at the parent-key level before patch extraction and by exact patch hashes. The independent robustness analysis uses three repeated 9/2/3 parent train/validation/test splits (seeds 42, 101, and 202), with test parents excluded from model selection. The complete 11/3 validation-pool analysis uses identical evaluation samples across models. Foundry uses a fixed 209/45/45 train/validation/test split (seed 42); Pt/Fe uses a fixed five-frame held-out test subset for the few-shot analysis. The software environment and major package versions are reported in Supplementary Table S14.

Data Availability

The Sm-BFO STEM images and associated atom-parameterization data underlying the COM and `xy_atms` annotations are from the datasets associated with the cited Ziatdinov et al. studies [9, 10] and are publicly available at 10.5281/zenodo.4555978. The external Pt/Fe AC-STEM data are from the public dataset associated with the cited single-atom detection study [13]. The Foundry Wei benchmark is publicly available under DOI 10.18126/e73h-3w6n [33]. Processed split identifiers and analysis outputs needed to reproduce the reported results are provided through the project repository.

Code Availability

Custom training, corruption, centroid-extraction, cross-domain adaptation, statistical-analysis, and figure-generation code is available at https://github.com/INQUIRELAB/Hybrid-STEMSeg-Centroid-Robust-Atomic-Site-Localization-in-Noisy-STEM-Images.

Author Contributions

G.D. designed the study, developed the Hybrid-STEMSeg framework, and implemented the atomic-site localization pipeline for STEM image analysis. He conducted all computational experiments, including Sm-BFO preprocessing, Gaussian target-mask generation, baseline benchmarking, corruption-robustness evaluation, ablation studies, statistical analysis, zero-shot external-domain testing, common-target domain-specific benchmarking, Foundry full-training evaluation, and few-shot adaptation on the Pt/Fe AC-STEM and Foundry datasets. He curated and analyzed the experimental outputs, prepared the figures and tables, and wrote the initial manuscript draft. S.S. and Y.B. supervised the research and provided scientific guidance throughout the development and evaluation of the framework. They contributed to the interpretation of the results, methodological refinement, editorial feedback, manuscript revision, and review of the final analyses. Y.B. additionally oversaw the overall project direction, ensured methodological rigor, and contributed to securing the resources and support that enabled the research.

Competing interests

The authors declare no competing interests.

Funding

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-25-1-0322.

References

[1] Crewe, A.V., Wall, J., Langmore, J.: Visibility of single atoms. Science **168**(3937), 1338–1340 (1970)

[2] Pennycook, S., Jesson, D.: High-resolution Z-contrast imaging of crystals. Ultramicroscopy **37**(1-4), 14–38 (1991)

[3] Pennycook, S.J., Nellist, P.D. (eds.): Scanning Transmission Electron Microscopy: Imaging and Analysis. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-7200-2

[4] Urban, K.W.: Studying atomic structures by aberration-corrected transmission electron microscopy. Science **321**(5888), 506–510 (2008)

[5] Ophus, C.: Quantitative scanning transmission electron microscopy for materials science: Imaging, diffraction, spectroscopy, and tomography. Annual Review of Materials Research **53**(1), 105–141 (2023)

[6] Fujino, S., Murakami, M., Anbusathaiah, V., Lim, S.-H., Nagarajan, V., Fennie, C., Wuttig, M., Salamanca-Riba, L., Takeuchi, I.: Combinatorial discovery of a lead-free morphotropic phase boundary in a thin-film piezoelectric perovskite. Applied Physics Letters **92**(20) (2008)

[7] Borisevich, A.Y., Eliseev, E., Morozovska, A., Cheng, C.-J., Lin, J.-Y., Chu, Y.-H., Kan, D., Takeuchi, I., Nagarajan, V., Kalinin, S.V.: Atomic-scale evolution of modulated phases at the ferroelectric–antiferroelectric morphotropic phase boundary controlled by flexoelectric interaction. Nature Communications **3**(1), 775 (2012)

[8] Nelson, C.T., Vasudevan, R.K., Zhang, X., Ziatdinov, M., Eliseev, E.A., Takeuchi, I., Morozovska, A.N., Kalinin, S.V.: Exploring physics of ferroelectric domain walls via Bayesian analysis of atomically resolved STEM data. Nature Communications **11**(1), 6361 (2020)

[9] Ziatdinov, M., Nelson, C.T., Zhang, X., Vasudevan, R.K., Eliseev, E., Morozovska, A.N., Takeuchi, I., Kalinin, S.V.: Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution scanning transmission electron microscopy data. npj Computational Materials **6**(1), 127 (2020)

[10] Ziatdinov, M., Creange, N., Zhang, X., Morozovska, A., Eliseev, E., Vasudevan, R.K., Takeuchi, I., Nelson, C., Kalinin, S.V.: Predictability as a probe of manifest and latent physics: The case of atomic scale structural, chemical, and polarization behaviors in multiferroic Sm-doped $BiFeO_3$. Applied Physics Reviews **8**(1) (2021)

[11] Kalinin, S.V., Ophus, C., Voyles, P.M., Erni, R., Kepaptsoglou, D., Grillo, V., Lupini, A.R., Oxley, M.P., Schwenker, E., Chan, M.K., *et al.*: Machine learning in scanning transmission electron microscopy. Nature Reviews Methods Primers **2**(1), 11 (2022)

[12] Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)

[13] Mitchell, S., Parés, F., Faust Akl, D., Collins, S.M., Kepaptsoglou, D.M., Ramasse, Q.M., Garcia-Gasulla, D., Pérez-Ramírez, J., López, N.: Automated image analysis for single-atom detection in catalytic materials by transmission electron microscopy. Journal of the American Chemical Society **144**(18), 8018–8029 (2022)

[14] Eliasson, H., Erni, R.: Localization and segmentation of atomic columns in supported nanoparticles for fast scanning transmission electron microscopy. npj Computational Materials **10**(1), 168 (2024)

[15] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[17] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*: An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

[20] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[21] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)

[22] Ziatdinov, M., Dyck, O., Maksov, A., Li, X., Sang, X., Xiao, K., Unocic, R.R., Vasudevan, R., Jesse, S., Kalinin, S.V.: Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. ACS Nano **11**(12), 12742–12752 (2017)

[23] Maksov, A., Dyck, O., Wang, K., Xiao, K., Geohegan, D.B., Sumpter, B.G., Vasudevan, R.K., Jesse, S., Kalinin, S.V., Ziatdinov, M.: Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS_2. npj Computational Materials **5**(1), 12 (2019)

[24] Leitherer, A., Yeo, B.C., Liebscher, C.H., Ghiringhelli, L.M.: Automatic identification of crystal structures and interfaces via artificial-intelligence-based electron microscopy. npj Computational Materials **9**(1), 179 (2023)

[25] Lin, R., Zhang, R., Wang, C., Yang, X.-Q., Xin, H.L.: 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)

[26] Ziatdinov, M., Ghosh, A., Wong, C.Y., Kalinin, S.V.: AtomAI framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy. Nature Machine Intelligence **4**(12), 1101–1112 (2022)

[27] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)

[28] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)

[29] Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018)

[30] Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[31] Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: PANet: Few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)

[32] Catalano, N., Matteucci, M.: Few shot semantic segmentation: a review of methodologies, benchmarks, and open challenges. arXiv preprint arXiv:2304.05832 (2023)

[33] Wei, J., Blaiszik, B., Morgan, D., Voyles, P.: Benchmark Dataset for Locating Atoms in STEM images. Materials Data Facility. Dataset (2021). https://doi.org/10.18126/e73h-3w6n

[34] Milletari, F., Navab, N., Ahmadi, S.-A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). IEEE

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CVSupplementary1.pdf