Supplementary Information

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

1 Independent parent-held-out Sm-BFO robustness

To separate checkpoint selection from final Sm-BFO evaluation, three repeated parent-disjoint 9/2/3 train/validation/test splits were constructed using seeds 42, 101, and 202. Test parents were never used for training or checkpoint selection. Table S1 summarizes centroid-F1 across the independent splits, and Fig. S1 visualizes the severity curves.

Table S1: Independent parent-held-out Sm-BFO centroid-F1. Values are mean ± standard deviation across three repeated 9/2/3 parent train/validation/test splits.

σ	Hybrid-STEMSeg	Hybrid-No Transformer	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

2 Sensitivity to evaluation-pool composition

To quantify the effect of evaluation-pool composition on robustness estimates, Table S2 compares a 100-patch subset drawn from one held-out parent with the complete 300-patch validation pool spanning all three held-out parents.

Table S2: Sensitivity of Gaussian centroid-F1 to the evaluation pool. The 100-patch set is the first held-out parent only; the 300-patch set includes all three held-out parents.

	100-patch single-parent			300-patch all-parent		
Model	$\sigma = 0$	$\sigma = 1.0$	$\sigma = 1.5$	$\sigma = 0$	$\sigma = 1.0$	$\sigma = 1.5$
Hybrid-STEMSeg	0.9854	0.9729	0.9156	0.8835	0.9127	0.8605
Hybrid-NoTransformer	0.9825	0.9713	0.8907	0.8823	0.9096	0.8438
U-Net	0.9852	0.0002	0.0000	0.9008	0.0937	0.0933
DeepLabV3+	0.9844	0.0006	0.0000	0.9035	0.0937	0.0934
SegFormer	0.9880	0.9279	0.4330	0.9730	0.8710	0.4450

Figure S2 visualizes the same sensitivity. Because evaluation-pool composition affects both absolute scores and some rank positions, the complete 300-patch validation pool is used for dense

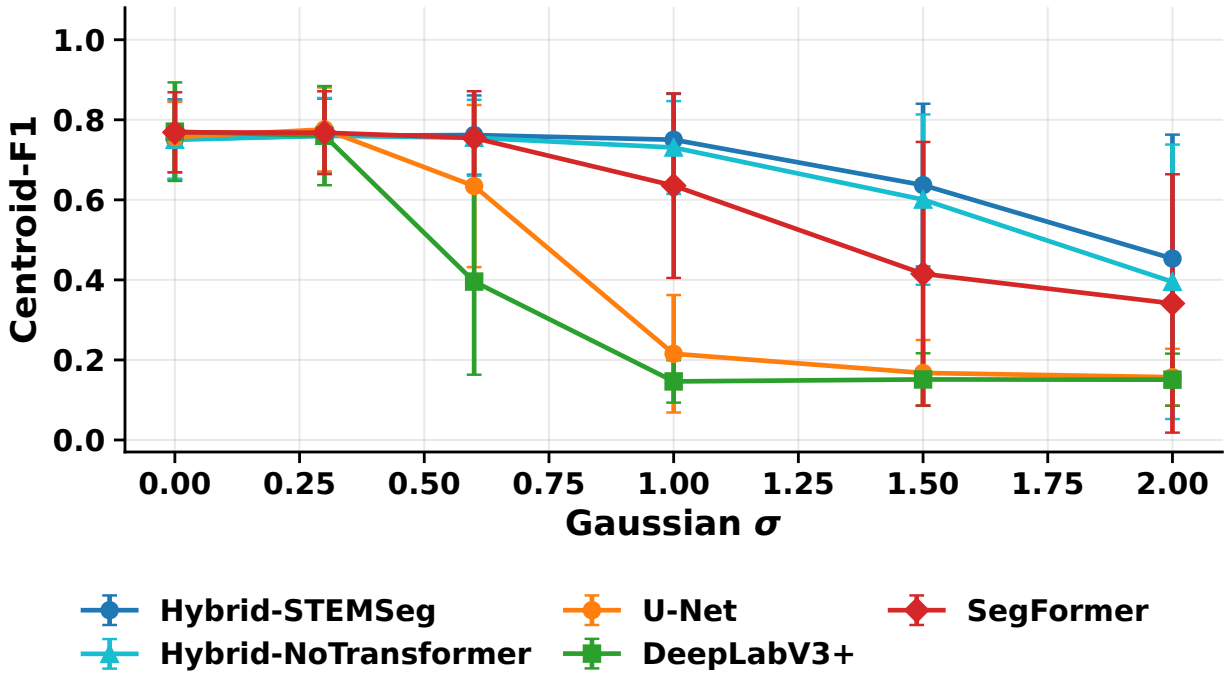


Figure S1: Independent parent-held-out Gaussian robustness. Error bars denote standard deviation across the three repeated 9/2/3 parent train/validation/test splits.

descriptive analyses, while the main manuscript uses repeated parent-disjoint train/validation/test splits for its headline Sm-BFO robustness claim.

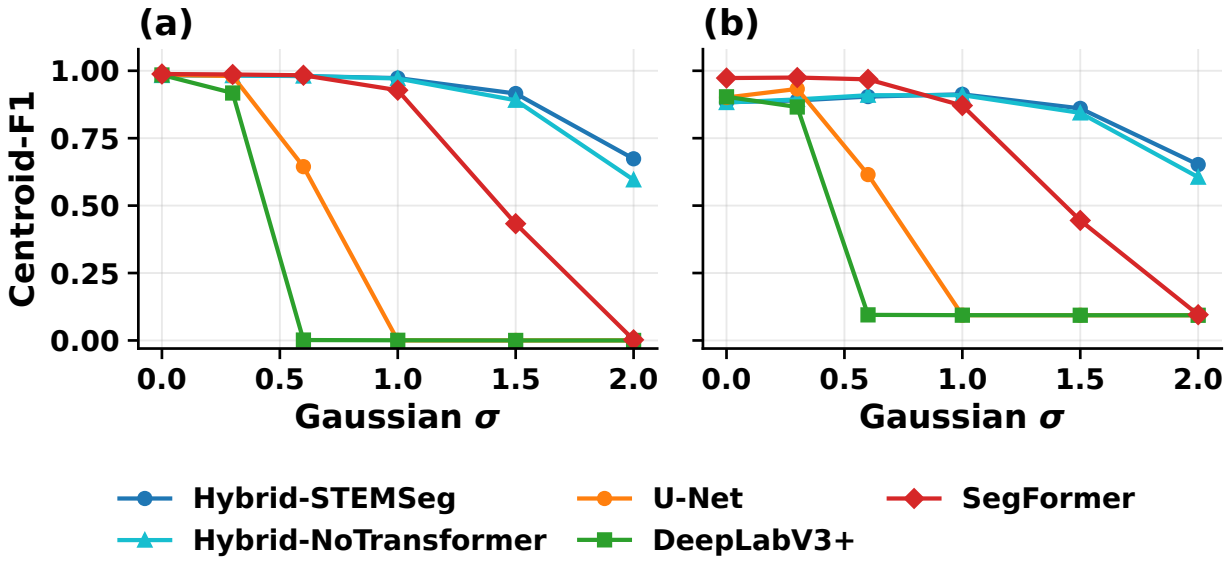


Figure S2: Sensitivity of Gaussian robustness to evaluation-pool composition, comparing a 100-patch single-parent subset with the complete 300-patch three-parent validation pool.

3 Gaussian robustness on the complete validation pool

Integrated summaries of the complete 300-patch Gaussian sweep are reported in Table S3. To avoid ambiguity in the term AURC, both integrated corruption risk and mean F1 over the tested severity range are reported.

Table S3: Integrated Gaussian robustness summaries on all 300 Sm-BFO validation patches. Integrated corruption risk is $\int (1 - F1(\sigma))\, d\sigma$ over $\sigma \in [0, 2]$ and is lower when robustness is better. Mean F1 is the area under the F1 curve divided by the tested severity range. Normalized loss integrates $1 - F1(\sigma)/F1(0)$.

Model	F1(0)	Risk	Mean F1	Norm. loss	F1(1.5)/F1(0)	F1(2.0)/F1(0)
Hybrid-STEMSeg	0.8835	0.2800	0.8600	0.0531	0.9740	0.7382
Hybrid-NoTransformer	0.8823	0.2987	0.8506	0.0716	0.9564	0.6862
U-Net	0.9008	1.2576	0.3712	1.1758	0.1036	0.1038
DeepLabV3+	0.9035	1.4594	0.2703	1.4016	0.1034	0.1033
SegFormer	0.9730	0.5844	0.7078	0.5451	0.4574	0.0981

Figure S3 shows the complete Gaussian severity curves underlying Table S3.

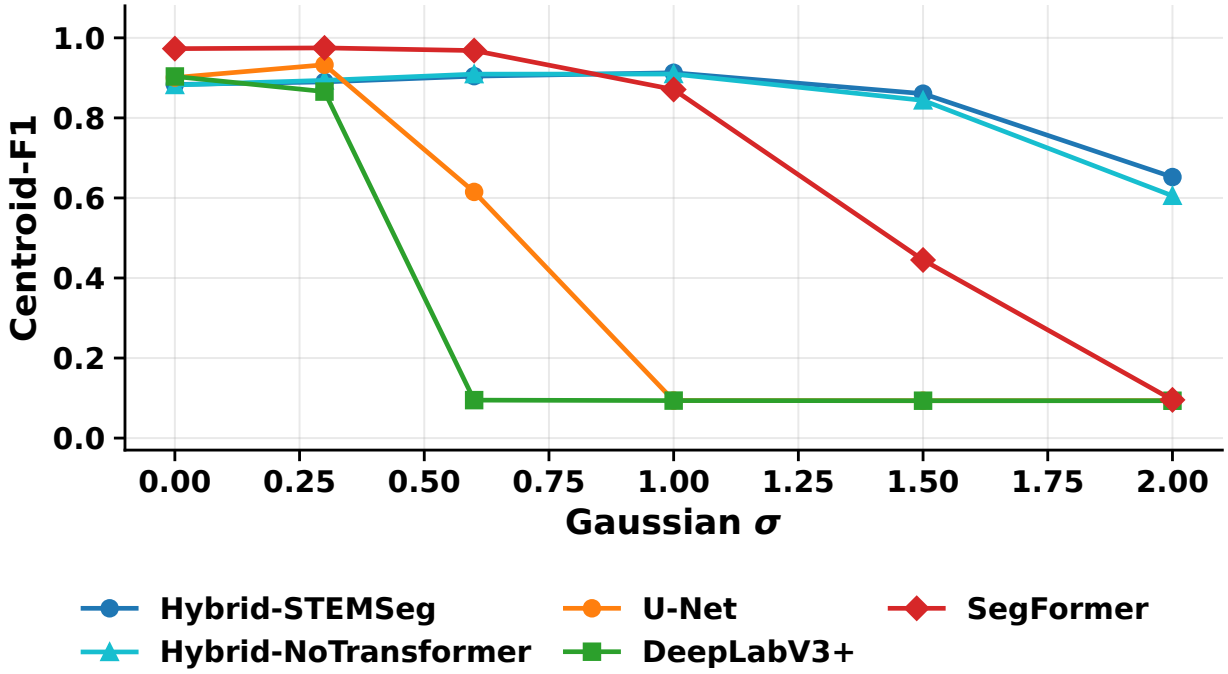


Figure S3: Centroid-F1 across Gaussian-noise severity on all 300 Sm-BFO validation patches.

4 Additional corruption families

Poisson and mixed-corruption results on the complete validation pool are provided in Table S4.

5 Gaussian diagnostic and PSF-like blur stress test

The primary Gaussian evaluation adds noise after per-patch z-score normalization. For Hybrid-STEMSeg, the validation-pool F1 increases from 0.8835 at $\sigma = 0$ to 0.9043 at $\sigma = 0.6$, accompanied

Table S4: Poisson and mixed-corruption performance on all 300 Sm-BFO validation patches.

Model	**Poisson F1**	**Mixed F1**	**Poisson IoU**	**Mixed IoU**
Hybrid-STEMSeg	0.9231	0.8878	0.7348	0.6509
Hybrid-NoTransformer	0.9293	0.9019	0.7576	0.6847
U-Net	0.2731	0.0938	0.1044	0.0933
DeepLabV3+	0.0803	0.0933	0.0800	0.0933
SegFormer	0.8796	0.7657	0.6869	0.5557

by precision increasing from 0.8707 to 0.8926 and recall from 0.9029 to 0.9220. The effect is parent-dependent: one held-out parent decreases slightly, one increases substantially, and one is nearly unchanged. It is therefore reported as a fixed-pipeline operating-point effect rather than evidence that physical image quality improves with added noise.

As an additional experimentally interpretable stress test, isotropic Gaussian blur approximates progressive probe/PSF broadening. Table S5 and Fig. S4 report centroid-F1 for blur widths from 0 to 2 pixels on the complete 300-patch validation pool.

Table S5: Centroid-F1 under PSF-like Gaussian blur on all 300 Sm-BFO validation patches.

Model	$\sigma_b = 0$	0.5	1.0	1.5	2.0
Hybrid-STEMSeg	0.8835	0.8838	0.8846	0.8849	0.8848
Hybrid-NoTransformer	0.8823	0.8818	0.8813	0.8781	0.8781
U-Net	0.9008	0.9005	0.8967	0.8933	0.8973
DeepLabV3+	0.9035	0.8900	0.8863	0.8900	0.8972
SegFormer	0.9730	0.9731	0.9734	0.9738	0.9641

6 Downstream structural errors

Table S6 reports centroid-localization, nearest-neighbor (NN), and lattice-descriptor errors for the complete frame-level analysis. All quantities are in pixels because instrument-specific spatial calibration was unavailable. Dashes indicate that a descriptor could not be computed reliably because valid matched detections were unavailable.

7 Common-target annotation benchmark

The common-target `xy_atms` comparison is given in Table S7. It demonstrates that the robustness ranking can change materially when the target definition and annotation semantics change.

The repeated value 0.0867 at $\sigma = 1.5$ for DeepLabV3+ and AtomSegNet is not a duplicated cell. The 300-patch `xy_atms` 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 non-monotonic U-Net values at $\sigma = 0.6$ and 1.0 are also present in the source CSV and occur within this same failure-floor regime rather than arising from rounding.

Figure S5 shows the corresponding corruption curves.

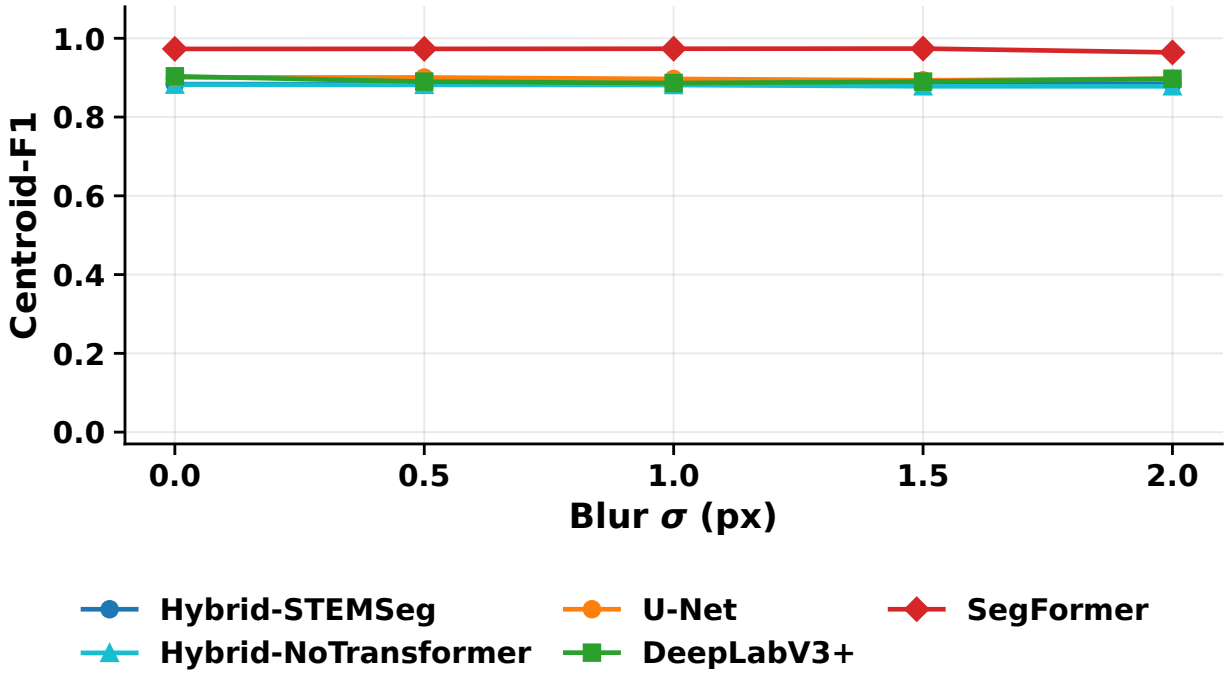


Figure S4: Centroid-F1 under isotropic Gaussian blur used as a PSF-broadening stress test.

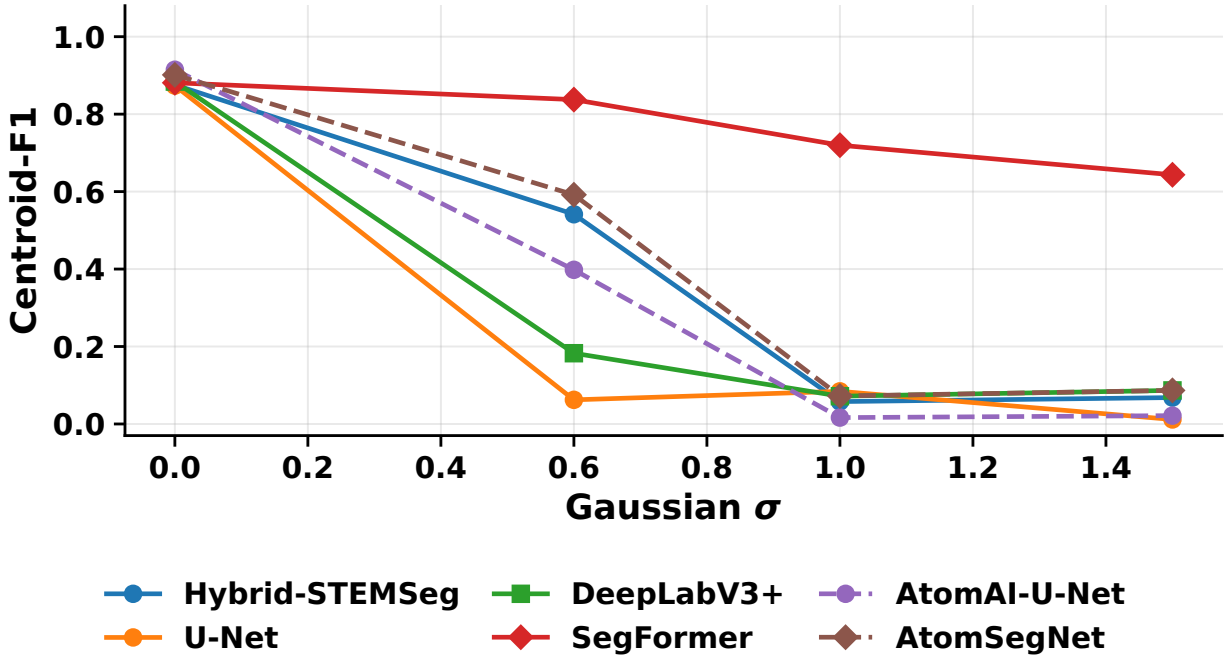


Figure S5: Common-target `xy_atms` Gaussian robustness curves.

Table S6: Frame-level structural errors across the three held-out Sm-BFO parent frames. Loc. denotes localization error and NN denotes nearest-neighbor spacing error. Dashes indicate that a descriptor could not be computed reliably because valid matched detections were unavailable.

Model	Condition	Med. loc.	Mean loc.	Med. NN	Mean NN	Lattice err.
Hybrid-STEMSeg	Clean	0.2250	0.2926	0.1546	0.2052	0.2238
	Gaussian 0.3	0.3188	0.3762	0.1920	0.2508	0.2414
	Gaussian 0.6	0.5378	0.5626	0.2539	0.3220	0.2744
	Gaussian 1.0	0.8413	0.8746	0.4308	0.3966	0.4770
	Poisson 10	0.7965	0.7952	0.2695	0.4873	0.6344
	Mixed 1.0	1.0909	1.1030	1.3927	1.0927	1.5882
Hybrid-NoTransformer	Clean	0.2747	0.3296	0.1613	0.2132	0.2404
	Gaussian 0.3	0.3140	0.3639	0.1918	0.2297	0.2415
	Gaussian 0.6	0.4574	0.4916	0.2901	0.3095	0.2845
	Gaussian 1.0	0.6029	0.6305	0.2892	0.2891	0.5117
	Poisson 10	0.5738	0.6132	0.5230	0.5221	0.9284
	Mixed 1.0	0.7704	0.7916	0.8789	1.5221	2.5426
U-Net	Clean	0.2487	0.2928	0.1544	0.2095	0.2094
	Gaussian 0.3	0.3964	0.4552	0.2233	0.2028	0.2297
	Gaussian 0.6	1.7484	1.8261	3.6190	3.6290	6.4101
	Gaussian 1.0	2.1418	2.1708	25.9774	25.7750	–
	Poisson 10	1.8669	1.8451	10.0562	13.3865	30.3997
	Mixed 1.0	1.9244	1.9140	25.9917	25.8822	–
DeepLabV3+	Clean	0.2754	0.3345	0.1738	0.2217	0.2434
	Gaussian 0.3	0.5924	0.5982	0.8860	0.9950	2.3520
	Gaussian 0.6	2.5257	2.4690	25.8012	29.2413	84.6437
	Gaussian 1.0	3.5731	3.2583	25.8138	25.8609	–
	Poisson 10	2.4976	2.7271	25.1863	24.7143	–
	Mixed 1.0	–	–	–	–	–
SegFormer	Clean	0.2336	0.3119	0.1513	0.2067	0.2161
	Gaussian 0.3	0.3111	0.3791	0.1920	0.2536	0.2049
	Gaussian 0.6	0.4654	0.5258	0.2586	0.3263	0.2859
	Gaussian 1.0	0.7616	0.8112	0.4665	0.8377	2.3000
	Poisson 10	0.6670	0.7064	0.3493	0.6574	2.1552
	Mixed 1.0	0.8890	0.8866	0.7277	1.2972	1.2957

Table S7: Equal-label `xy_atms` Sm-BFO benchmark on 300 validation patches.

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

8 Target-definition audit

The COM and `xy_atms` tasks use the same source file, split seed, Gaussian rasterization width ($\sigma = 3$ pixels), segmentation loss, logit threshold, and 6-pixel centroid-matching tolerance. However, the underlying coordinate keys differ. No parent has identical COM and `xy_atms` coordinates, and site counts differ for multiple parents. Therefore the ranking reversal between the two tasks cannot be attributed to rasterization geometry alone; it reflects target-definition and annotation-semantic differences.

9 Foundry full-training and few-shot benchmarks

Full target-domain training on Foundry is summarized in Table S8. Foundry target masks use a Gaussian width of $\sigma = 2.5$ pixels and are binarized at 0.25.

Table S8: Foundry Wei full-training benchmark using the 209/45/45 train/validation/test split with seed 42.

Model	Test F1	Test IoU	Val F1	Best epoch
U-Net	0.9036	0.7974	0.9618	21
Hybrid-STEMSeg	0.7389	0.6683	0.8214	37
DeepLabV3+	0.6774	0.6323	0.6774	14
SegFormer	0.6131	0.6119	0.5936	10

The auxiliary Foundry Gaussian corruption sweep uses the same fixed 45-image held-out test split with 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 complete Sm-BFO-to-Foundry few-shot grid is reported in Table S9.

Table S9: Sm-BFO-to-Foundry few-shot adaptation. $K = 0$ denotes zero-shot evaluation; $K \geq 1$ values are mean ± standard deviation over seeds 101, 202, and 303.

(a) Centroid-F1.

K	Hybrid-STEMSeg	U-Net	DeepLab V3+	SegFormer
0	0.205	0.162	0.121	0.182
1	0.389 ± 0.335	0.179 ± 0.120	0.065 ± 0.070	0.362 ± 0.344
3	0.398 ± 0.105	0.147 ± 0.131	0.051 ± 0.038	0.233 ± 0.056
5	0.430 ± 0.060	0.200 ± 0.179	0.108 ± 0.133	0.273 ± 0.042
10	0.575 ± 0.044	0.502 ± 0.013	0.435 ± 0.020	0.351 ± 0.089

(b) Mask IoU.

K	Hybrid-STEMSeg	U-Net	DeepLab V3+	SegFormer
0	0.074	0.066	0.054	0.076
1	0.371 ± 0.114	0.299 ± 0.058	0.250 ± 0.026	0.315 ± 0.062
3	0.394 ± 0.037	0.285 ± 0.028	0.258 ± 0.027	0.306 ± 0.042
5	0.402 ± 0.031	0.316 ± 0.077	0.287 ± 0.064	0.344 ± 0.022
10	0.507 ± 0.025	0.464 ± 0.018	0.441 ± 0.024	0.390 ± 0.057

Figure S6 visualizes the Foundry F1 adaptation curves.

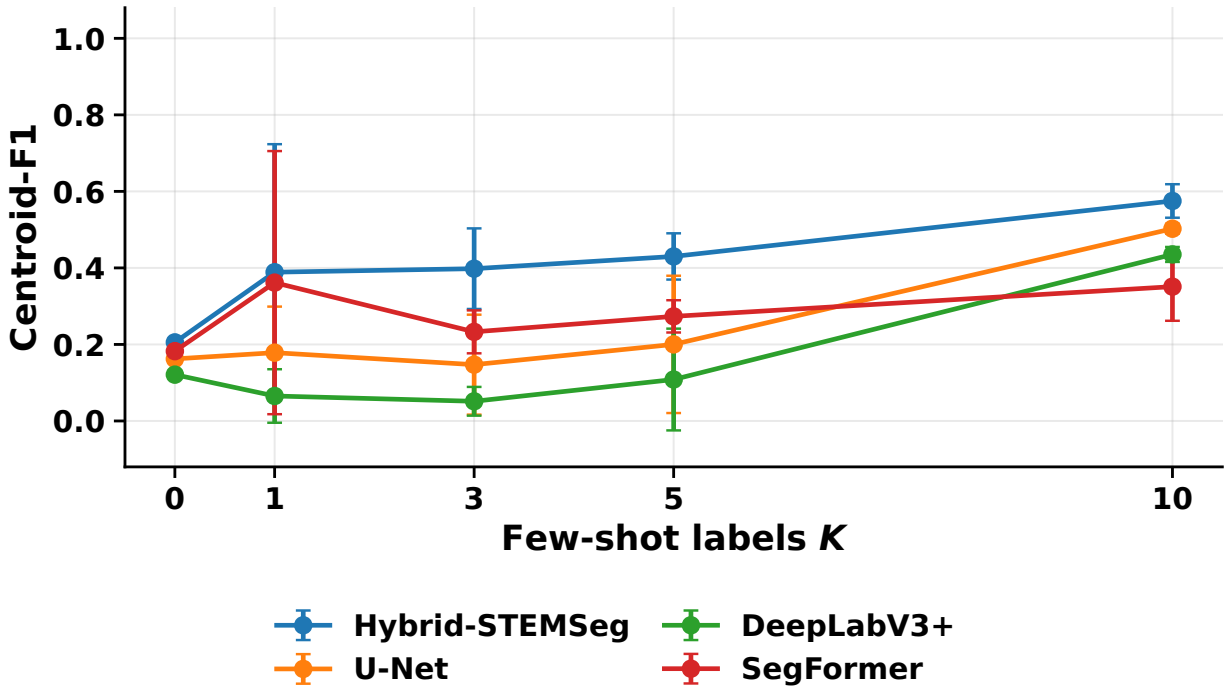


Figure S6: Sm-BFO-to-Foundry few-shot adaptation. Error bars show standard deviation across seeds for $K \geq 1$.

10 Pt/Fe AC-STEM held-out-test adaptation

The Pt/Fe protocol reserves a fixed five-frame test subset before adaptation. Table S10 reports held-out-test performance; adaptation and checkpoint selection use only the remaining non-test frames.

Table S10: Pt/Fe AC-STEM zero-shot and few-shot adaptation on a fixed five-frame held-out test subset. $K \geq 1$ values are mean ± standard deviation over seeds 42, 101, and 202.

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

Figure S7 shows the corresponding held-out-test F1 curves.

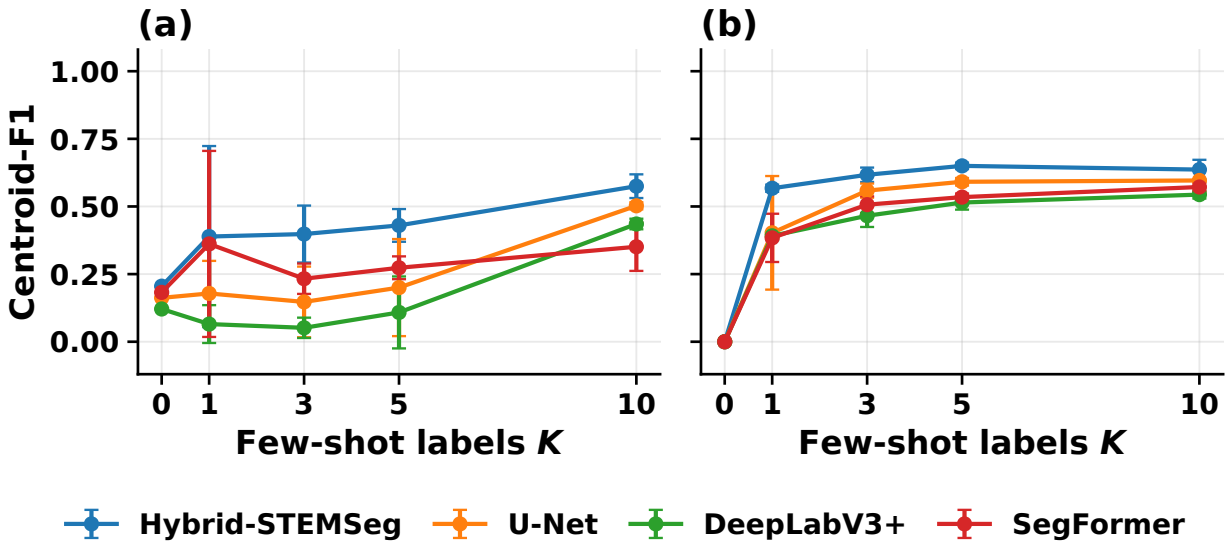


Figure S7: Cross-domain few-shot adaptation. Panel (a) shows Sm-BFO-to-Foundry held-out-test F1 and panel (b) shows Sm-BFO-to-Pt/Fe held-out-test F1. Error bars show standard deviation across seeds for $K \geq 1$.

11 Hybrid-STEMSeg versus NoTransformer engineering tradeoff

The direct robustness–efficiency comparison is summarized in Table S11.

Table S11: Hybrid-STEMSeg versus Hybrid-NoTransformer. Negative Δ risk indicates lower integrated corruption risk for Hybrid-STEMSeg; positive latency Δ indicates higher latency.

Metric	Hybrid	NoTransformer	Δ
Clean F1	0.8835	0.8823	+0.0012
F1, Gaussian 1.0	0.9127	0.9096	+0.0031
F1, Gaussian 1.5	0.8605	0.8438	+0.0168
F1, Poisson 10	0.9231	0.9293	-0.0061
F1, Mixed 1.0	0.8878	0.9019	-0.0141
Integrated Gaussian risk	0.2800	0.2987	-0.0188
Mean centroid error, Gaussian 1.0	0.8746	0.6305	+0.2441
Mean NN error, Gaussian 1.0	0.3966	0.2891	+0.1075
Parameters	4,287,425	4,287,425	0
Median latency (ms)	5.0734	2.4231	+2.6504
p95 latency (ms)	5.3609	2.9892	+2.3717
Peak GPU memory (MB)	94.74	91.24	+3.50

Figure S8 provides a compact visual summary of the corruption and latency tradeoff.

12 Efficiency

Inference efficiency is summarized in Table S12. Latency was measured with CUDA at batch size 1 on 256×256 inputs after 20 warm-up iterations and 100 timed iterations.

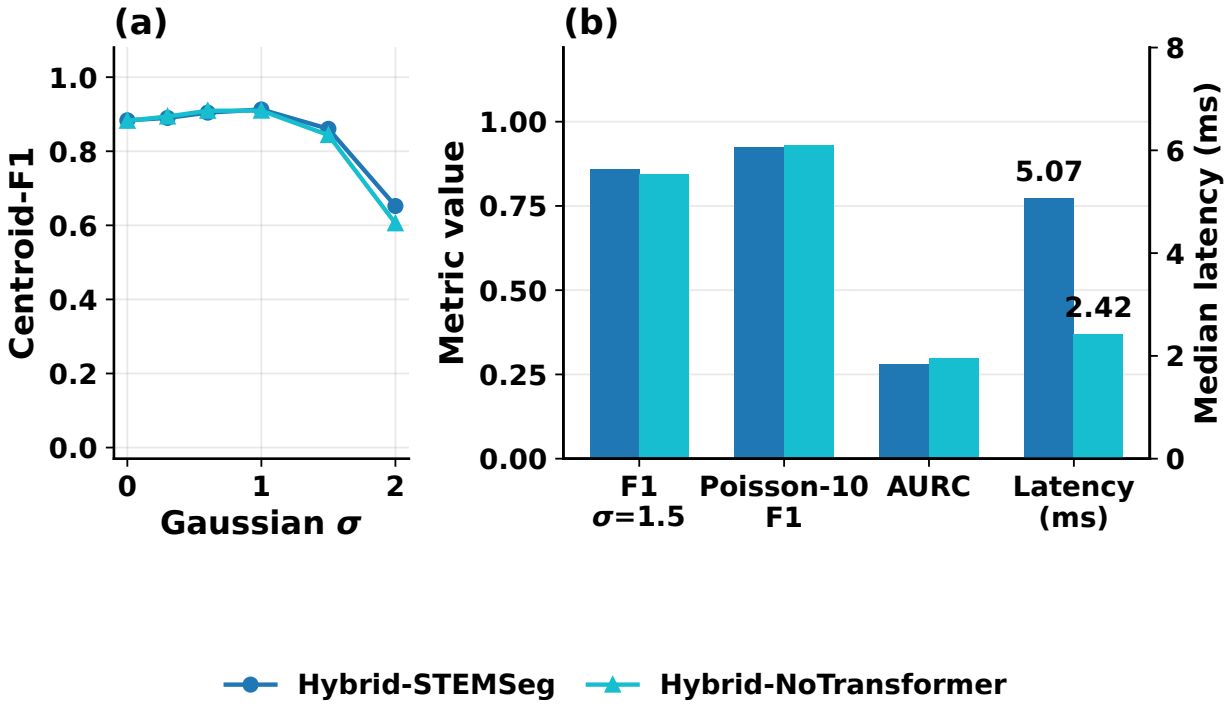


Figure S8: Selected robustness metrics for Hybrid-STEMSeg and Hybrid-NoTransformer together with their measured median inference latency.

Table S12: Model size, GPU latency, and peak GPU memory.

Model	Parameters	Median ms	p95 ms	Peak MB
Hybrid-STEMSeg	4,287,425	5.0734	5.3609	94.74
Hybrid-NoTransformer	4,287,425	2.4231	2.9892	91.24
U-Net	24,430,097	4.1700	5.1378	127.87
DeepLabV3+	22,431,185	4.0527	6.2221	120.31
SegFormer	3,711,265	6.4343	9.2235	60.53

The NoTransformer ablation retains the Transformer modules as registered parameters but bypasses them during forward computation. Consequently, both implementations contain 4,287,425 registered parameters, whereas only 471,617 parameters participate in the NoTransformer computational graph. The ablation therefore removes attention computation without physically removing its parameter tensors.

13 Reproducibility, data integrity, and software environment

The fixed evaluation configuration and Sm-BFO split-integrity checks are summarized in Table S13. The headline Sm-BFO robustness analysis uses repeated 9/2/3 parent train/validation/test splits with independent test parents, while the complete 300-patch 11/3 validation pool is retained for dense descriptive analyses. Foundry few-shot experiments use seeds 101, 202, and 303, whereas Pt/Fe few-shot experiments use seeds 42, 101, and 202.

Table S13: Evaluation configuration and Sm-BFO split/data-integrity summary.

Item	Value
Sm-BFO parent keys	14
Complete-pool split (train / validation / test parents)	11 / 3 / 0
Complete-pool train / validation patches	4400 / 300
Annotated sites, train / validation	290,813 / 65,289
Independent robustness split	9 / 2 / 3 parents, repeated for 3 seeds
Independent split seeds	42, 101, 202
Headline robustness evaluation	Test parents excluded from training and checkpoint selection
Parent-key overlap	0
Exact train–validation duplicate patches	0
Descriptive robustness pool	300 validation patches, 3 parents
Gaussian corruption seed	123
Foundry split	209 / 45 / 45, seed 42
Foundry few-shot seeds	101, 202, 303
Pt/Fe few-shot seeds	42, 101, 202
Primary task metric	Centroid-F1
Secondary mask metric	IoU

The recorded software environment is given in Table S14.

14 Benchmark fairness and training settings

Table S15 summarizes the principal model-specific and shared training settings. Hybrid-STEMSeg, Hybrid-NoTransformer, U-Net, DeepLabV3+, and SegFormer use the same Sm-BFO mask-benchmark optimizer and loss settings; architecture-specific backbones and pretrained initialization are retained where applicable. No dedicated hyperparameter grid search was used, and no Sm-BFO test data were used for tuning.

Shared mask-benchmark settings are Adam optimization, a constant learning rate of 2×10^{-4} (no scheduler), zero weight decay, batch size 8, up to 100 epochs, early-stopping patience 15, random cropping at preprocessing time without additional online augmentation, per-patch z-score normalization clipped at ± 6, and BCE-with-logits plus Dice loss with positive-class weight 18 and

Table S14: Major software dependencies recorded for reproducibility.

Package	**Version**
PyTorch	`2.11.0`
TorchVision	`0.26.0`
MONAI	`1.5.2`
segmentation-models-pytorch	`0.5.0`
Transformers	`5.7.0`
timm	`1.0.26`
NumPy	`2.2.6`
SciPy	`1.15.3`
scikit-image	`0.25.2`
Matplotlib	`3.10.9`
Pillow	`12.2.0`

Table S15: Benchmark-fairness summary. LR denotes learning rate, BS batch size, ES early-stopping patience, Thr. output threshold, and Match the centroid-matching tolerance in pixels. For Hybrid-STEMSeg, Hybrid-NoTransformer, U-Net, DeepLabV3+, SegFormer, and AtomAI-UNet, Thr.=0.0 denotes a logit threshold; for AtomSegNet, Thr.=0.5 denotes a sigmoid-probability threshold. These are equivalent operating points because sigmoid(0) = 0.5. A dash indicates that the parameter count was not recorded in the final audit table.

Model	Params	LR	BS	Epochs	ES	Thr.	Match	Initialization / backbone
Hybrid-STEMSeg	4,287,425	0.0002	8	100	15	0.0	6	From scratch; custom U-Net + Transformer bottleneck
Hybrid-NoTransformer	4,287,425	0.0002	8	100	15	0.0	6	From scratch; custom U-Net without Transformer
U-Net	24,430,097	0.0002	8	100	15	0.0	6	ImageNet-pretrained ResNet34 encoder
DeepLabV3+	22,431,185	0.0002	8	100	15	0.0	6	ImageNet-pretrained ResNet34 encoder
SegFormer	3,711,265	0.0002	8	100	15	0.0	6	MiT-B0 pretrained initialization
AtomAI-UNet	–	0.0002	8	100	15	0.0	6	AtomAI U-Net; common-target `xy_atms` run
AtomSegNet	–	0.0002	8	100	15	0.5	6	AtomSegNet U-Net path; common-target `xy_atms` run

Dice weight 0.5. Checkpoints are selected by best validation loss. No dedicated architecture-specific hyperparameter grid search was performed; the comparison uses repository defaults and the shared benchmark configuration, while architecture-specific backbones and pretrained initialization are retained. The independent 9/2/3 robustness experiment keeps test parents separate from checkpoint selection.

15 Centroid post-processing and sensitivity

The primary mask path thresholds model logits at 0 (sigmoid probability 0.5), labels two-dimensionally connected components with connectivity 2, applies no component-size filtering or morphology, retains border components, and computes centroids using `regionprops`. Matching is greedy nearest-neighbor with one match per ground-truth site and a default tolerance of 6 pixels. A sensitivity grid varying the probability threshold from 0.3 to 0.7 and the match tolerance from 2 to 10 pixels shows that the qualitative ordering at Gaussian $\sigma = 1.0$ is stable for thresholds 0.3–0.7 and tolerances 4–10 pixels. At the strictest 2-pixel tolerance, Hybrid-NoTransformer slightly exceeds Hybrid-STEMSeg, while both remain above SegFormer, U-Net, and DeepLabV3+.

Table S16 reports the corresponding centroid-F1 values at Gaussian $\sigma = 1.0$. The threshold scan holds the matching tolerance at 6 pixels, while the tolerance scan holds the probability threshold at 0.5.

Table S16: Centroid post-processing sensitivity at Gaussian $\sigma = 1.0$. Threshold rows use a 6-pixel matching tolerance; tolerance rows use a probability threshold of 0.5.

Sweep	Setting	Hybrid-STEMSeg	Hybrid-NoTransformer	U-Net	DeepLabV3+	SegFormer
Threshold	0.3	0.8987	0.8968	0.1123	0.0946	0.8756
Threshold	0.4	0.9144	0.9038	0.0971	0.0939	0.8738
Threshold	0.5	0.9136	0.9019	0.0936	0.0937	0.8723
Threshold	0.6	0.9134	0.9005	0.0933	0.0934	0.8707
Threshold	0.7	0.9144	0.9138	0.0933	0.0934	0.8690
Tolerance (px)	2	0.8888	0.8913	0.0935	0.0934	0.8591
Tolerance (px)	4	0.9119	0.9017	0.0936	0.0936	0.8722
Tolerance (px)	6	0.9136	0.9019	0.0936	0.0937	0.8723
Tolerance (px)	8	0.9136	0.9020	0.0936	0.0937	0.8723
Tolerance (px)	10	0.9136	0.9020	0.0936	0.0937	0.8723

16 Exploratory frame-level uncertainty

Because the Sm-BFO validation set contains only three independent held-out parent frames, the frame-level comparisons in Table S17 are exploratory. They are reported as paired mean differences, bootstrap confidence intervals, and standardized paired effects; they are not used to support claims of formal statistical superiority.

Table S17: Exploratory frame-level comparisons of mean centroid error. Differences are Hybrid-STEMSeg minus the comparator; lower centroid error is better.

Comparator	**Condition**	**Δ error [95% CI]**	**Cohen's d_z**
DeepLabV3+	clean	-0.0419 $[-0.0533, -0.0273]$	-3.1515

Comparator	**Condition**	**Δ error [95% CI]**	**Cohen's d_z**
Hybrid-NoTransformer	clean	-0.0370 $[-0.0526, -0.0213]$	-2.3643
SegFormer	clean	-0.0192 $[-0.0653, 0.0162]$	-0.4610
U-Net	clean	-0.0002 $[-0.0237, 0.0245]$	-0.0081
DeepLabV3+	Gaussian 0.6	-1.9064 $[-3.2000, -0.6829]$	-1.5130
Hybrid-NoTransformer	Gaussian 0.6	0.0710 $[0.0651, 0.0804]$	8.5910
SegFormer	Gaussian 0.6	0.0368 $[0.0142, 0.0724]$	1.1777
U-Net	Gaussian 0.6	-1.2635 $[-1.3984, -1.1815]$	-10.7343
DeepLabV3+	Gaussian 1.0	-2.3837 $[-2.9948, -1.3543]$	-2.6584
Hybrid-NoTransformer	Gaussian 1.0	0.2441 $[0.2323, 0.2614]$	15.8948
SegFormer	Gaussian 1.0	0.0634 $[0.0224, 0.0880]$	1.7746
U-Net	Gaussian 1.0	-1.2962 $[-1.4643, -1.1240]$	-7.6172
Hybrid-NoTransformer	Mixed 1.0	0.3114 $[0.2951, 0.3205]$	22.0309
SegFormer	Mixed 1.0	0.2164 $[0.1103, 0.2923]$	2.2856
U-Net	Mixed 1.0	-0.8110 $[-1.2976, -0.2102]$	-1.4676
DeepLabV3+	Poisson 10	-1.9319 $[-4.1897, 0.0013]$	-0.9137
Hybrid-NoTransformer	Poisson 10	0.1820 $[0.1230, 0.2227]$	3.4791
SegFormer	Poisson 10	0.0888 $[0.0315, 0.1725]$	1.1973
U-Net	Poisson 10	-1.0499 $[-1.2126, -0.9607]$	-7.4413