SITE-AIT: SITE-BOUND LATENT TOKENIZATION AND JOINT OPTICAL INFERENCE FOR DENSE QUBIT READOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense fluorescence images of large qubit registers violate the independence assumption behind per-site thresholding: neighboring point-spread functions overlap, emitter coordinates drift, and local observations depend on unknown neighboring states. Moreover, pixel-based representations and generic object queries fail to maintain a persistent one-to-one correspondence with dense physical sites, hindering high-fidelity readout. We propose Site-AIT, a representation that binds one latent token to each dense site and resolves coupled decisions via register-wide self-attention. Tokens are constructed via offset-corrected, point-spread-function-weighted Top-16 selection from a 13×13 candidate field, reducing local features from 2,704 to 256 dimensions. When labeled data are scarce, a detector-aware noise prior initializes the image backbone, while a differentiable optical reconstruction objective guides training without adding inference overhead. Evaluated on a 2,000-frame benchmark from a 300-site trapped-ion register, Site-AIT achieves a per-ion accuracy of 0.9978 ± 0.0004, a localization error of 0.0152 ± 0.0004 pixels, and an FPGA latency of 0.31 ± 0.01 ms. Extensive ablations and physical perturbation experiments confirm that aligning the inference unit with the physical measurement unit provides an effective design principle for calibrated, densely overlapping emitter arrays.

1 INTRODUCTION

Recognizing many small targets in one image becomes a structured inference problem when the targets have persistent identities and their measurements overlap. The decision at one location cannot be separated from the unknown states of its neighbors. This regime arises directly in fluorescence readout of large qubit registers, where a camera frame must be converted into one state estimate per physical site for downstream syndrome extraction and feedback. Trapped-ion systems with hundreds of sites make the representation choice especially consequential: site-resolved measurements must be accurate, calibrated, and completed within a short feedback cycle (Guo et al., 2024; Wolf et al., 2024; Liu et al., 2025).

Conventional readout often starts from photon counts, local thresholds, matched filters, or a separate classifier at each site (Burrell et al., 2010; Seif et al., 2018; Jeong et al., 2023; Phuttitarn et al., 2024; Kent et al., 2026). These approaches are effective when local evidence is sufficiently isolated or when only a few ions are present. However, in a dense register, fluorescence from one bright emitter spreads into neighboring regions. Treating this signal as local noise discards a useful dependency because the source of the spillover is another unknown state. The problem is therefore not merely to extract stronger local features; it is to preserve physical identity while exchanging evidence across all sites. Figure 1 contrasts joint site-indexed inference with independent local decisions.

Context-aware detection and segmentation offer alternatives, but their output organization differs from dense-register readout. Segmentation models produce masks or label fields (Ronneberger et al., 2015; Chen et al., 2024; Zheng et al., 2021; Kerssies et al., 2025; Fu et al., 2025), which our protocol aggregates at dense sites. Query-based detectors predict object sets using learned queries or spatial anchors (Carion et al., 2020; Zhu et al., 2021; Wang et al., 2022; Sun et al., 2023; Huang et al., 2025); our detector adaptations recover site correspondence through output-to-site assignment.

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.

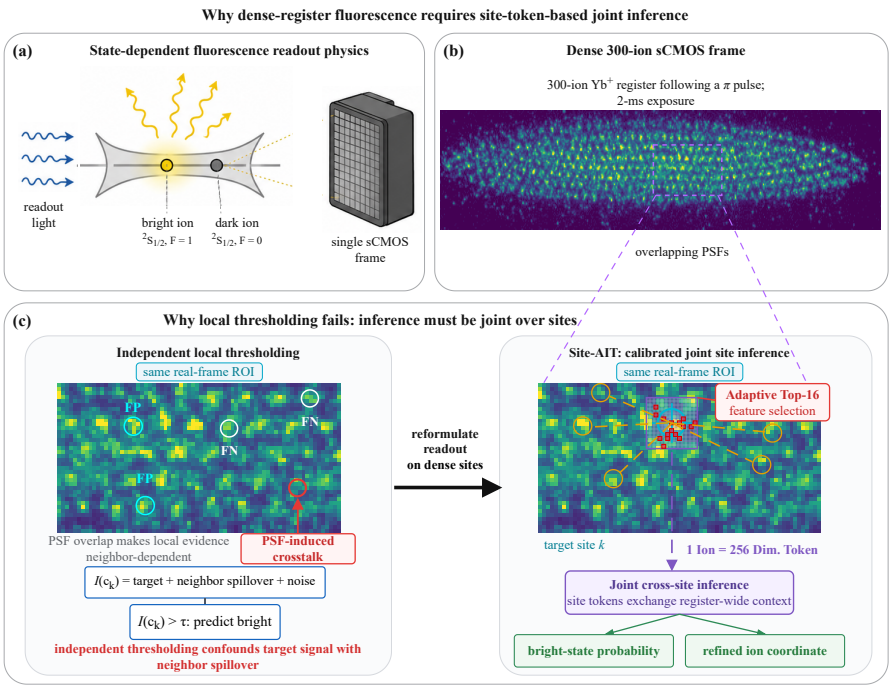


Figure 1: **Why dense-register fluorescence requires site-token-based joint inference.** (a) Bright and dark ions produce distinct fluorescence. (b) A dense 300-ion frame contains overlapping PSFs. (c) Local thresholding confounds target signal with neighboring spillover; Site-AIT instead exchanges evidence among dense site tokens to predict states and coordinates. FP and FN denote false-bright and missed-bright decisions.

Their native formulations do not enforce a fixed physical-site identity for each latent unit. This is a difference in inductive bias, not an inability to use context or calibrated coordinates. Tokenizer analysis examines the trade-off in information retention (Qian et al., 2023); quadrangle attention adapts sampling geometry (Zhang et al., 2024). These mechanisms address information retention, spatial flexibility, or efficiency, but do not by themselves impose one persistent token per dense site. For dense-register readout, evidence selection must adapt without dropping, merging, or reassigning physical-site identities, including dark sites with little visible signal.

We ask how to estimate a bright-state probability and a refined coordinate for every one of K coarse sites in a single forward pass. Four difficulties must be addressed together. First, nominal coordinates are imperfect, so a narrowly centered patch may miss displaced evidence, whereas a wide patch admits irrelevant or contaminated features. Second, processing all features in a wide candidate field conflicts with compact, low-latency readout. Third, accurately labeled scientific frames are expensive, even though the model must learn to handle detector noise and weak-signal cases. Noise modeling and data-centric denoising improve the construction or use of noisy training data for image restoration (Feng et al., 2024; Lu et al., 2025; Feng et al., 2026). Their restoration objectives do not directly supervise site-state discrimination; the relevant question is whether detector-aware pretraining transfers to readout with fewer expert labels. Fourth, ordinary classification and coordinate losses do not require the joint predictions to explain the observed, PSF-overlapped image. Physics-informed methods address reconstruction and denoising in photon-limited imaging (Cai et al., 2026; Yang & Yang, 2026), but the objective here is not image restoration: it is to constrain the predicted states and coordinates through image formation while leaving the deployed readout path unchanged.

Our key idea is to make the model's unit of inference coincide with the instrument's unit of measurement. Site-AIT binds one persistent latent token to every dense site. A bounded offset recenters a local candidate field; a learned selector retains 16 PSF-weighted features; and the resulting site tokens exchange evidence through register-wide self-attention. Detector-aware noise-prior pretraining reduces dependence on labeled frames. During supervised training, a differentiable PSF model

reconstructs fluorescence from predicted states and coordinates, but this branch is removed for deployment.

1.1 CONTRIBUTIONS

(i) **Qubit Tokenization (§§2–3, §4.3):** We bind one persistent latent token to each ion (qubit) and combine offset-corrected adaptive Top-16 sampling with register-wide self-attention for joint state and coordinate estimation.

(ii) **Detector-aware pretraining (§3.4, §4.3):** Self-supervised pretraining with a measured-detector noise chain reduces dependence on expert labels. With 4,500 labeled frames, the resulting model reaches an F_1 of 0.9969, within 0.0010 of the 20,000-frame pretrained reference.

(iii) **Training-only optical consistency (§3.5, §4.3):** A differentiable, measured-PSF-initialized renderer couples state and coordinate predictions during training and is removed at inference. Ablations test the reconstruction term, PSF-width learning, and its weight schedule.

(iv) **Large-register benchmark and resources (Reproducibility statement):** On a held-out 2,000-frame benchmark from one 300-site trapped-ion platform, Site-AIT achieves an accuracy of 0.9978 ± 0.0004 under the common evaluation protocol. We release the Real1600–Hard400 benchmark, training data, protocols, code, and model weights.

2 PROBLEM FORMULATION

Let $I : \Omega \to \mathbb{R}$ be one fluorescence frame and let $\mathcal{C}^0 = \{\boldsymbol{c}_k^0\}_{k=1}^K$ denote the calibrated nominal coordinates of K candidate sites. The image origin is at the upper-left corner, with x increasing to the right and y increasing downward. Site indices $k = 1, \ldots, K$ follow lexicographic order in (y_k^0, x_k^0): increasing vertical coordinate is the primary key, and increasing horizontal coordinate breaks ties. This nominal-lattice ordering defines persistent physical-site identities across frames. Each site has an unknown bright/dark state $y_k \in \{0, 1\}$ and a frame-specific coordinate $\boldsymbol{c}_k$. A compact observation model is

$$I(\boldsymbol{p}) = b(\boldsymbol{p}) + \sum_{k=1}^{K} a_k(y_k)\psi(\boldsymbol{p} - \boldsymbol{c}_k) + n(\boldsymbol{p}), \tag{1}$$

where b is the background, ψ is the measured PSF, a_k is the state-dependent fluorescence amplitude, and n collects detector and photon noise. When neighboring PSFs overlap, pixels near site k contain contributions from multiple unknown states y_j. Consequently, $p(y_k \mid I)$ cannot generally be inferred from an isolated count without losing information carried by neighboring sites.

The calibrated coordinates are informative but not exact in every frame. We write $\boldsymbol{c}_k = \boldsymbol{c}_k^0 + \boldsymbol{\delta}_k$ and ask the model to infer the bright-state probability $\hat{p}_k \approx p(y_k = 1 \mid I, \mathcal{C}^0)$, a refined coordinate $\hat{\boldsymbol{c}}_k$, and an auxiliary site-existence probability $\hat{e}_k$. The desired output is the ordered set

$$\widehat{\mathcal{Y}} = \{(\hat{p}_k, \hat{\boldsymbol{c}}_k, \hat{e}_k)\}_{k=1}^K, \tag{2}$$

with the kth output retaining the identity of the kth physical site. The auxiliary existence output and its fixed-register limitations are described in Sec. 3.3.

These challenges motivate a site-indexed representation with adaptive, fixed-budget evidence collection, detector-aware learning, and training-only optical consistency for deterministic, batch-one inference.

3 ANCHOR-TOKEN INFERENCE

3.1 OVERVIEW

Figure 2 summarizes the computation and separates offline pretraining, deployed inference, and training-only supervision. Let F denote the 16-channel image feature map produced by DW-UNet from frame I. The following sections define local token construction, register-wide self-attention (Vaswani et al., 2017), the readout heads, and the training objectives.

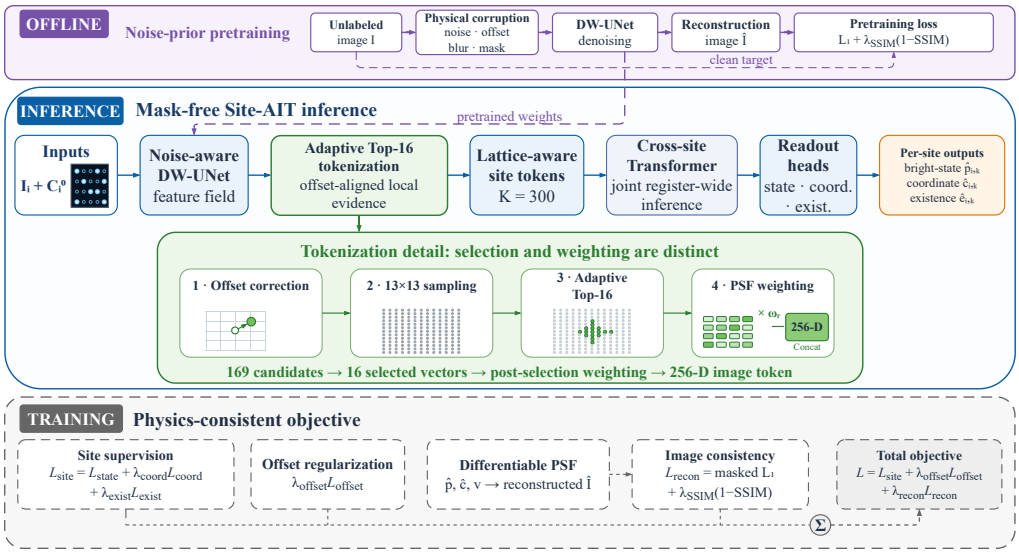


Figure 2: **Site-AIT pipeline.** DW-UNet encodes frame I_i; each dense site recenters a 13×13 field, selects and PSF-weights 16 vectors, and forms one 256-dimensional token. The $K = 300$ site tokens exchange context and feed the state, coordinate, and existence heads. Noise-prior pretraining initializes the encoder, while PSF reconstruction supplies training-only supervision. Mask-free inference requires no dense segmentation output.

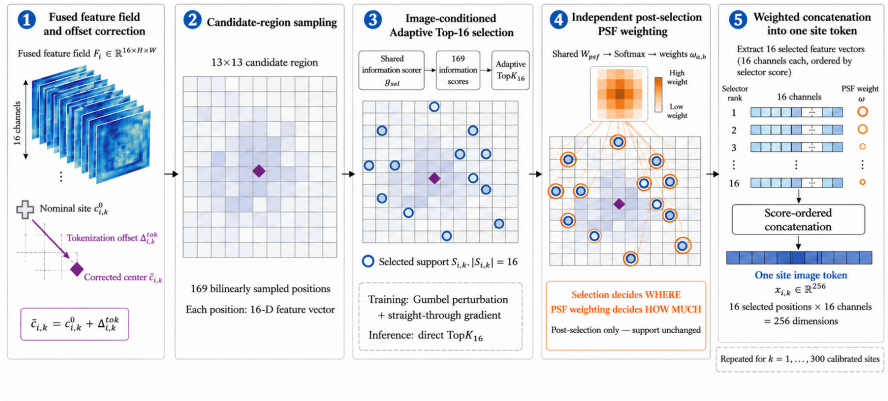


Figure 3: **Offset-corrected adaptive Top-16 tokenization for site k.** A bounded offset recenters a 13×13 field. Top-16 selects 16 of 169 sampled feature vectors; PSF weighting and ordered concatenation form the 256-dimensional token $\boldsymbol{x}_k$. Training uses a straight-through Gumbel surrogate; inference is deterministic.

3.2 Offset-Corrected Adaptive Top-16 Tokenization

Figure 3 illustrates the tokenizer defined below. For site k, a pooled feature around $\boldsymbol{c}_k^0$ is passed through a shared multilayer perceptron g_Δ. A bounded displacement

$$\Delta \boldsymbol{c}_k = r \tanh\big(g_\Delta(F, \boldsymbol{c}_k^0)\big), \quad \bar{\boldsymbol{c}}_k = \boldsymbol{c}_k^0 + \Delta \boldsymbol{c}_k \tag{3}$$

recenters the local field with $r = 10$ pixels per axis. The candidate window extends ± 6 pixels per axis around $\bar{\boldsymbol{c}}_k$, where bilinear sampling yields $N = 13^2 = 169$ candidate features $\{\boldsymbol{f}_{k,j} \in \mathbb{R}^{16}\}_{j=1}^{N}$.

A shared scorer g_s assigns logits $s_{k,j} = g_s(\boldsymbol{f}_{k,j})$. During training, Gumbel perturbations and a straight-through relaxation permit learning a discrete subset (Jang et al., 2017; Kool et al., 2019);

during inference, deterministic Top-16 selection is used. If $\pi_k(1), \ldots, \pi_k(M)$ are the selected indices in score order and $M = 16$, a learned nonnegative PSF mask supplies normalized weights $w_{k,j}$. The site token is

$$\boldsymbol{x}_k = \mathrm{concat}_{m=1}^{M} \left[w_{k,\pi_k(m)} \boldsymbol{f}_{k,\pi_k(m)} \right] \in \mathbb{R}^{256}. \tag{4}$$

The pre-projection width is reduced from $169 \times 16 = 2{,}704$ to $16 \times 16 = 256$. The resulting representation uses K fixed-budget 256-dimensional tokens regardless of the number of image pixels; for the evaluated instrument, $K = 300$, and register-wide attention scales as $O(K^2 d)$ for transformer width d. All reported latency includes the complete deployed forward pass, with cross-engine results separated because precision and compiler effects cannot be attributed to any single module.

3.3 Persistent Site Tokens and Readout Heads

Each $\boldsymbol{x}_k$ is projected to the transformer width by W_x and combined with a site-identity encoding

$$\boldsymbol{r}_k = \boldsymbol{e}_k + \phi_{\mathrm{coord}}(\tilde{\boldsymbol{c}}_k^0), \qquad \boldsymbol{z}_k = W_x \boldsymbol{x}_k + \boldsymbol{r}_k, \tag{5}$$

where $\boldsymbol{e}_k$ is a learnable embedding of the persistent site index k, $\tilde{\boldsymbol{c}}_k^0$ is the calibrated coordinate normalized to $[0, 1]^2$, and ϕ_{coord} is a shared two-layer perceptron. The index embedding $\boldsymbol{e}_k$ attaches a fixed latent identity to each physical site across frames, while ϕ_{coord} supplies its lattice position. Self-attention operates on the ordered set of all K tokens. Its purpose is not to discover object identities; it lets the token for site k condition its state estimate on evidence gathered for any other site. This full-register exchange represents multi-body optical crosstalk without restricting interactions to a manually selected neighborhood.

The contextualized token $\boldsymbol{h}_k$ feeds three multilayer-perceptron heads: $\hat{p}_k = \sigma(g_y(\boldsymbol{h}_k))$ for state classification, $\hat{\boldsymbol{c}}_k = \boldsymbol{c}_k^0 + b \tanh(g_c(\boldsymbol{h}_k))$ for coordinate refinement, and $\hat{e}_k = \sigma(\ell_k^e)$ with $\ell_k^e = g_e(\boldsymbol{h}_k)$ for site existence. Here, $b = \sqrt{W^2 + H^2}$ is the image diagonal length in pixels, and W and H denote the image width and height in pixels, respectively. The last output is intended for future variable-occupancy settings. Here, all existence targets are one, so its supervision is positive-only; it is not used to reject sites, and missing-ion detection remains untested. All three outputs retain the identity of the same physical-site token.

3.4 Detector-Aware Noise-Prior Pretraining

Pretraining uses 100,000 laboratory-acquired CCD/sCMOS images spanning tens to hundreds of ions, including normal and faulty/noisy acquisitions. Before site-state and coordinate supervision, DW-UNet reconstructs the target I from the synthetically corrupted input in Fig. 4 using a weighted L_1–SSIM loss. Appendix C gives the corruption ranges and optimization settings. Only backbone weights are transferred; the task-specific modules are subsequently optimized under supervision, and the pretraining network is not deployed.

3.5 Training-Only PSF Reconstruction

Classification and coordinate regression supervise the outputs directly, but neither alone checks whether the combined predictions explain the observed fluorescence. We render an auxiliary image from the predicted probabilities and coordinates:

$$\hat{I}(\boldsymbol{p}) = \hat{b}(\boldsymbol{p}) + \sum_{k=1}^{K} \alpha_k \hat{p}_k \psi_\sigma(\boldsymbol{p} - \hat{\boldsymbol{c}}_k), \tag{6}$$

where α_k is a learnable site amplitude and $\hat{b}(\boldsymbol{p})$ is a learnable background, both parameterized independently of the learnable width of ψ_σ, which is initialized from the measured PSF. Saturated pixels are gated out of the comparison. The supervised objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_c \mathcal{L}_{\mathrm{coord}} + \lambda_e \mathcal{L}_{\mathrm{exist}} + \lambda_o \mathcal{L}_{\mathrm{offset}} + \lambda_r(t) \mathcal{L}_{\mathrm{recon}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{cls}}$ is binary cross-entropy, $\mathcal{L}_{\mathrm{coord}}$ is Smooth-L_1, $\mathcal{L}_{\mathrm{exist}}$ is binary cross-entropy on the existence logit, and $\mathcal{L}_{\mathrm{offset}}$ is the mean squared tokenizer displacement. We use $\lambda_c = 0.05$, $\lambda_e = 0.1$, and $\lambda_o = 0.01$; $\lambda_r(t)$ is zero for 10 epochs and then increases linearly over 5 epochs to 0.05. The reconstruction term connects state and coordinate errors through measured image formation. At deployment, Eq. (6) and $\mathcal{L}_{\mathrm{recon}}$ are discarded.

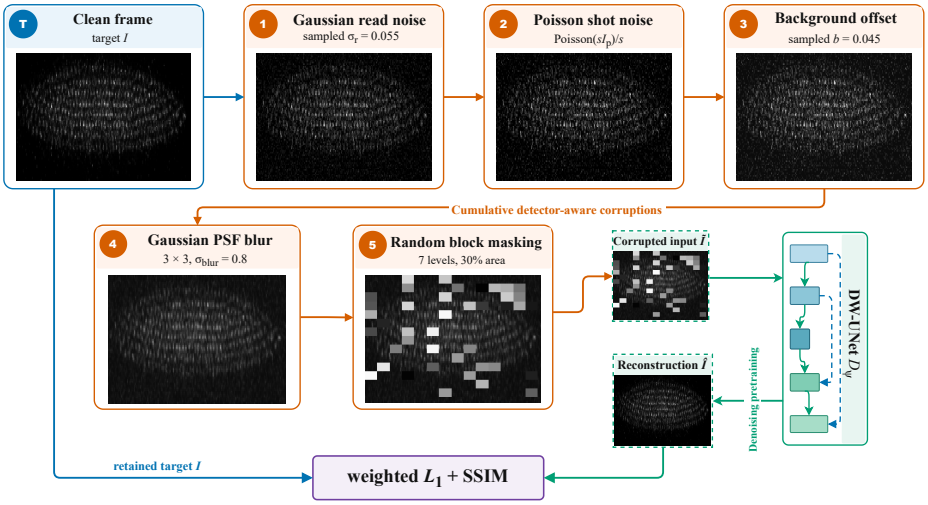


Figure 4: **Detector-aware noise-prior pretraining.** A normalized copy of the target I undergoes, in order, Gaussian read noise ($\sigma_r = 0.055$), pixelwise $\mathrm{Poisson}(sI_p)/s$ shot noise, background offset ($b = 0.045$), 3×3 Gaussian PSF blur ($\sigma_{\mathrm{blur}} = 0.8$ pixels), and seven-gray-level block masking (30%). Here, s is the Poisson scale; post-Gaussian I_p and the final input are clipped to $[0, 1]$; and the displayed σ_r and b are illustrative, not fixed. DW-UNet D_ψ reconstructs $\widehat{I}$ from $\widetilde{I}$ against the original measured target I, not a noise-free reference, using weighted L_1 and SSIM.

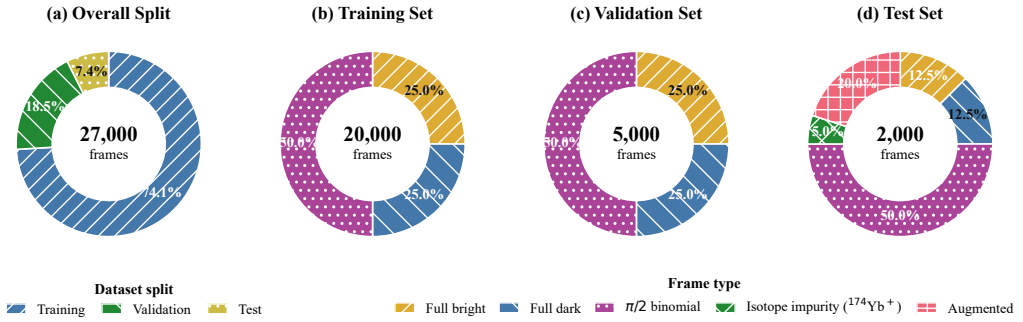


Figure 5: **Composition of the 27,000-frame dataset for a 300-site Yb^+ register.** (a) The 20,000/5,000/2,000 training/validation/test split. (b)–(d) Frame-type distributions within each partition; the center labels give frame counts, and the sectors give percentages. The held-out test set contains 1,000 mixed-state, 250 all-bright, 250 all-dark, and 100 impurity frames (Real1600), plus 400 frames generated from disjoint held-out sources using coordinate perturbation, isotropic and anisotropic blur, foreground attenuation with noise, hot pixels, and one-site displacement (Hard400).

4 Experiments

4.1 Data, Comparators, and Protocol

Figure 5 summarizes the 20,000/5,000/2,000-frame training/validation/test split. Each 456×88 sCMOS frame records a 2-ms exposure of the 300-site register.

We compare Site-AIT with a validation-tuned matched filter, ion–pixel mapping (IPM), DETR-style queries, a register-preserving D-FINE+MAL adaptation reported as DEIM, and five dense segmentation networks. Neural comparators use the same partitions, hardware environment, training budget, and seeds. Dense baselines undergo model-specific validation searches and then convert mask probabilities to site scores using a validation-selected disk and threshold. DETR-style (2020)

Table 1: Site-state performance on Real1600–Hard400. Classification metrics are computed by pooling confusion counts over all 2,000 test images. Neural results are three-seed means. ± denotes the sample standard deviation; deterministic estimators have no seed spread. Latency is measured with batch-one PyTorch FP32 inference on an RTX 4090. Boldface marks the best value.

Model	Acc.↑	Precision↑	Recall↑	F_1↑	Latency (ms)↓
Site-AIT (ours)	**0.9978** ±0.0004	**0.9966** ±0.0007	**0.9993** ±0.0001	**0.9979** ±0.0004	**4.50** ±0.15
Classical estimators					
Matched filter (2026) (Kent et al., 2026)	0.8913	0.8719	0.9282	0.8992	6.12
Ion–pixel mapping (2025) (Liu et al., 2025)	0.8765	0.8556	0.9185	0.8859	15.1
Query-based detection baselines					
DETR-style (2020) (Carion et al., 2020)	0.9792 ±0.0002	0.9653 ±0.0004	0.9960 ±0.0001	0.9804 ±0.0002	9.50 ±0.20
DEIM (2025) (Huang et al., 2025)	0.9677 ±0.0005	0.9853 ±0.0012	0.9524 ±0.0007	0.9686 ±0.0004	6.20 ±0.18
Dense segmentation with mask probabilities sampled at dense sites					
U-Net (2015) (Ronneberger et al., 2015)	0.8750 ±0.0004	0.8549 ±0.0021	0.9033 ±0.0005	0.8784 ±0.0012	5.10 ±0.12
TransUNet-style (2024) (Chen et al., 2024)	0.9470 ±0.0004	0.9054 ±0.0030	0.9983 ±0.0005	0.9496 ±0.0019	18.40 ±0.45
SegMAN-T (2025) (Fu et al., 2025)	0.9148 ±0.0018	0.9287 ±0.0020	0.8986 ±0.0018	0.9134 ±0.0014	7.10 ±0.25
SETR-style (2021) (Zheng et al., 2021)	0.8433 ±0.0020	0.7668 ±0.0027	0.9867 ±0.0020	0.8630 ±0.0021	24.60 ±0.60
EoMT (2025) (Kerssies et al., 2025)	0.8202 ±0.0010	0.7930 ±0.0031	0.8667 ±0.0050	0.8282 ±0.0010	21.30 ±0.55

uses unconstrained learned queries and full-image cross-attention. For protocol compatibility, it substitutes DW-UNet for ResNet, shares Site-AIT's state, coordinate, and existence heads, and uses Hungarian matching during training. The matched filter uses the eight nearest calibrated neighbors as a fixed design choice and tunes its remaining parameters per site on validation data. All final neural models were trained for 100 epochs with a batch size of 48 on one RTX 4090 using seeds $\{1, 2, 3\}$.

We report pooled per-ion accuracy, bright-class precision, recall, and F_1, AUROC, and 15-bin expected calibration error (ECE) (Pakdaman Naeini et al., 2015). Localization error is computed only for valid sites containing ground-truth bright emitters; it is the root mean square of their coordinate residuals after subtracting their pooled mean and is therefore a dispersion measure rather than absolute centroid error. Classification metrics are computed by pooling confusion counts over all 2,000 test images. Neural results are three-seed means; dispersions are sample standard deviations.

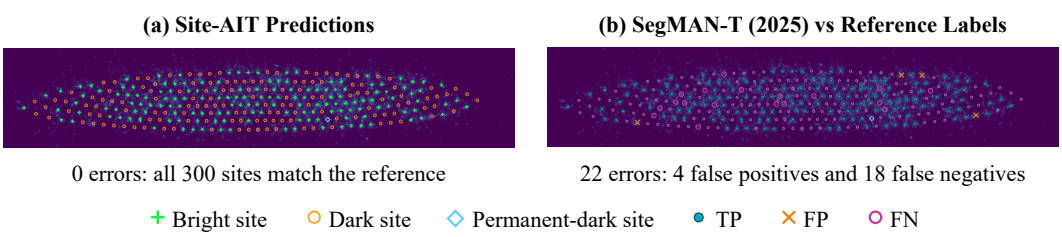
4.2 Readout Accuracy and Error Reduction

Table 1 reports the common benchmark. Site-AIT achieves the best accuracy and F_1 among the evaluated methods, reducing the error rate by a factor of 9.5 relative to the strongest baseline. DEIM favors precision over recall, while the matched filter improves on IPM but remains behind the learned query models. Dense segmentation methods lose further accuracy when their probability maps are reduced to one decision per dense site.

Figure 6 provides a single-frame qualitative comparison: Site-AIT matches all 300 reference site labels, whereas SegMAN-T makes 22 site-state errors on the same frame.

4.3 Ablation Study

Retrained architectural controls. Table 2 compares independently retrained controls under the matched protocol specified in its caption. The full model performs best across the reported classification and localization metrics. Replacing adaptive local tokenization with position-aware full-image dense cross-attention reduces accuracy and increases latency. This control uses $K = 300$ site-conditioned queries over all $H \times W$ DW-UNet features, retaining the downstream Transformer,



Method	TP	FP	FN	TN	F1
Site-AIT	155	0	0	145	**1.0000**
SegMAN-T	137	4	18	141	0.9257

Figure 6: **On this single frame, Site-AIT matches all 300 reference labels, whereas SegMAN-T makes 22 site errors.** (a) Site-AIT predictions coincide with the 155 bright and 145 dark reference sites. The blue diamond marks the permanent-dark $^{174}\mathrm{Yb}^{+}$ impurity. (b) SegMAN-T's error map relative to the same reference. The table reports the same single-frame confusion-count and F_1 definitions for both methods.

Table 2: Ablation study on Real1600–Hard400. Each variant is independently optimized over seeds $\{1, 2, 3\}$ with fixed data partitions, training budgets, and validation rules. All controls except the no-pretraining control start from the same detector-aware pretrained backbone; task-specific modules are initialized independently.

Variant	**Acc.** ↑	**Precision** ↑	**Recall** ↑	F_1 ↑	**Localization error (px)** ↓	**Latency (ms)** ↓
Site-AIT (full model)	**0.9978** ±0.0004	**0.9966** ±0.0007	**0.9993** ±0.0001	**0.9979** ±0.0004	**0.0152** ±0.0004	4.50 ±0.15
Architectural components						
adaptive Top-16 tokenizer → position-aware full-image cross-attn.	0.9082 ±0.0015	0.9035 ±0.0022	0.9140 ±0.0018	0.9087 ±0.0020	0.0245 ±0.0006	11.80 ±0.35
w/o cross-site exchange	0.8743 ±0.0020	0.8700 ±0.0025	0.8800 ±0.0022	0.8750 ±0.0022	0.0285 ±0.0007	3.80 ±0.10
w/o tokenization offsets	0.9547 ±0.0010	0.9500 ±0.0015	0.9600 ±0.0012	0.9550 ±0.0010	0.0256 ±0.0006	4.22 ±0.10
w/o learned PSF mask	0.9507 ±0.0012	0.9460 ±0.0016	0.9560 ±0.0014	0.9510 ±0.0012	0.0215 ±0.0005	4.48 ±0.12
w/o noise-prior pretraining	0.9295 ±0.0020	0.9240 ±0.0022	0.9360 ±0.0020	0.9300 ±0.0018	0.0238 ±0.0006	4.52 ±0.14
w/o coordinate supervision ($\lambda_{\mathrm{coord}} = 0$)	0.9894 ±0.0013	0.9931 ±0.0014	0.9857 ±0.0011	0.9894 ±0.0012	0.0482 ±0.0015	4.45 ±0.14
PSF reconstruction consistency loss (Sec. 3.5)						
Full $-\,\mathcal{L}_{\mathrm{recon}}$, no recon. loss	0.8956 ±0.0080	0.9123 ±0.0031	0.8753 ±0.0137	0.8934 ±0.0086	0.0278 ±0.0008	4.48 ±0.16
Full $+\,\mathcal{L}_{\mathrm{recon}}$, σ_{psf} frozen	0.9904 ±0.0022	0.9894 ±0.0042	0.9914 ±0.0003	0.9904 ±0.0020	0.0163 ±0.0003	4.49 ±0.15
Full $+\,\mathcal{L}_{\mathrm{recon}}$, no weight warm-up	0.9744 ±0.0064	0.9791 ±0.0011	0.9694 ±0.0127	0.9742 ±0.0065	0.0210 ±0.0007	4.51 ±0.14

readout heads, reconstruction objective, and model dimensions. For *w/o cross-site exchange*, a diagonal-only mask in every Transformer layer and head lets each site token attend only to itself. All other modules and objectives remain unchanged; the performance loss supports cross-site exchange without changing network depth or parameter count. The no-offset control fixes $\Delta \boldsymbol{c}_k = 0$ while retaining the rest of the tokenizer. The remaining controls support contributions from offset correction, learned PSF weighting, detector-aware pretraining, coordinate supervision, and PSF reconstruction. Freezing the PSF width has a smaller effect, whereas removing reconstruction or its weight warm-up causes larger degradation or seed variation. Figure 7 further compares pretrained and randomly initialized models across labeled-frame budgets.

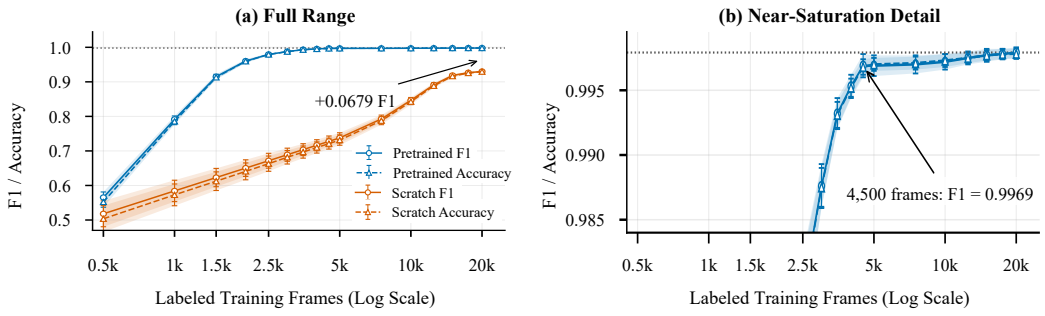


Figure 7: **Noise-prior pretraining reduces Site-AIT's labeled-frame budget.** Per-ion accuracy (dashed triangles) and bright-class F_1 (solid circles) on the fixed Real1600–Hard400 test set are plotted against the number of labeled frames for pretrained (blue) and scratch (orange) models. Error bars and bands show one sample standard deviation. (a) Full range; (b) near-saturation detail. The dotted line marks the 20,000-frame pretrained reference; annotations report the 4,500-frame result and full-data gain.

The dense-cross-attention and self-only controls separate local evidence selection from cross-site exchange, while the coordinate-supervision and reconstruction controls test how state and position are coupled through the measured PSF.

Initial-coordinate reference and perturbation. Without retraining, we perturb every coordinate-dependent input under either the calibrated lattice or 300 uniformly sampled coordinates sorted by the same vertical-then-horizontal rule. At 2.0 pixels, these references yield accuracy/F_1 of 0.9974 ± 0.0007 and 0.9972 ± 0.0007, with localization errors of 0.0161 ± 0.0009 and 0.0163 ± 0.0009 pixels, respectively. Mean accuracy decreases by approximately 0.0004 and 0.0005 from the corresponding unperturbed references. This supports limited sensitivity within the tested geometry and known target count, not robustness to arbitrary geometry or unknown target counts.

4.4 Additional Experimental Evidence

Confidence and calibration. Paired bootstrapping gives a 95% interval of $[0.0183, 0.0189]$ for the accuracy gain over DETR-style (2020), excluding zero. Site-AIT also obtains an AUROC of 0.9980 ± 0.0003 and an ECE of 0.0120 ± 0.0002, so the improvement extends to posterior ranking and calibration.

Deployment and controlled perturbations. Relative to native PyTorch FP32 on the RTX 4090, Zynq INT8 reduces batch-one inference latency from 4.50 ± 0.15 to 0.31 ± 0.01 ms with an accuracy of 0.9902 ± 0.0004; image acquisition is excluded. In the fixed-checkpoint 3:1 anisotropic-elongation stress test, Site-AIT retains an F_1 of 0.9623 versus 0.7897 for DETR-style (2020).

5 Conclusion

Site-AIT assigns one persistent token to every dense site, combines offset-corrected Top-16 sampling with register-wide attention, and uses detector-aware pretraining plus a training-only PSF objective. On the evaluated 300-site, 2K-frame benchmark, this design provides accurate, calibrated, and low-latency readout.

6 Limitations

Evaluation uses one 300-site trapped-ion platform with a fully occupied, spatially stable register and no absent-site examples. Transfer across platforms, imaging modalities, or target densities remains untested, as do missing-ion detection and variable-occupancy generalization.

AI USE STATEMENT

Generative AI (OpenAI GPT) was used to polish grammar and tone.

REPRODUCIBILITY STATEMENT

The main paper provides the problem definition, model equations, training objectives, dataset split, annotation quality-control procedure, evaluation metrics, validation-only model-selection rule, random-seed protocol, and inference-timing boundary in Sections 2–4. Appendices B–J provide the complete loss and operator definitions, pretraining settings, apparatus and imaging characterization, data-processing and Hard400 generation protocols, baseline search spaces, full quantitative and uncertainty results, architectural-ablation specifications, controlled-perturbation operators, and cross-platform deployment measurements. Anonymous source code together with pretrained and trained weights is available at `https://anonymous.4open.science/r/Site_Dia_4_Qubit_Readout`; the benchmark and training data are available at `https://huggingface.co/datasets/anonymousshare/300-Qubits-Register-Dataset`.

REFERENCES

Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QCrw0u9LQ7.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

A. H. Burrell, D. J. Szwer, S. C. Webster, and D. M. Lucas. Scalable simultaneous multiqubit readout with 99.99% single-shot fidelity. *Physical Review A*, 81:040302, Apr 2010. doi: 10.1103/PhysRevA.81.040302.

Siao Cai, Zhicheng Yu, Shaobing Gao, Zeyu Chen, and Yiguang Liu. Single-photon imaging in complex scenarios via physics-informed deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5827–5843, 2026. doi: 10.1109/TPAMI.2026.3654264.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. doi: 10.1016/j.media.2024.103280.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8b52c29d1a77e3fcc6c8f90fa0ac0e8e-Abstract-Conference.html.

Hansen Feng, Lizhi Wang, Yuzhi Wang, Haoqiang Fan, and Hua Huang. Learnability enhancement for low-light raw image denoising: A data perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):370–387, 2024. doi: 10.1109/TPAMI.2023.3301502.

Hansen Feng, Lizhi Wang, Yiqi Huang, Yuzhi Wang, Lin Zhu, and Hua Huang. Learning physics-informed noise models from dark frames for low-light raw image denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):3952–3969, 2026. doi: 10.1109/TPAMI.2026.3651447.

Shuang Fu, Wei Shi, Eugene A. Katrukha, Xi Chen, Yue Fei, Ke Fang, Ruixiong Wang, Tianlun Zhang, Donghan Ma, and Yiming Li. Aberration-aware 3D localization microscopy via self-supervised neural-physics learning. *Nature Communications*, 17:6504, 2026. doi: 10.1038/s41467-026-73045-9.

Yunxiang Fu, Meng Lou, and Yizhou Yu. SegMAN: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19077–19087, 2025. doi: 10.1109/CVPR52734.2025.01777.

S.-A. Guo, Y.-K. Wu, J. Ye, L. Zhang, W.-Q. Lian, R. Yao, Y. Wang, R.-Y. Yan, Y.-J. Yi, Y.-L. Xu, B.-W. Li, Y.-H. Hou, Y.-Z. Xu, W.-X. Guo, C. Zhang, B.-X. Qi, Z.-C. Zhou, L. He, and L.-M. Duan. A site-resolved two-dimensional quantum simulator with hundreds of trapped ions. *Nature*, 630(8017):613–618, 2024. doi: 10.1038/s41586-024-07459-0.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. DEIM: DETR with improved matching for fast convergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15162–15171, 2025. doi: 10.1109/CVPR52734.2025.01412.

Alexander Impertro, Julian F. Wienand, Sophie Häfele, Hendrik von Raven, Scott Hubele, Till Klostermann, Cesar R. Cabrera, Immanuel Bloch, and Monika Aidelsburger. An unsupervised deep learning algorithm for single-site reconstruction in quantum gas microscopes. *Communications Physics*, 6:166, 2023. doi: 10.1038/s42005-023-01287-w.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017. doi: 10.48550/arXiv.1611.01144. URL https://openreview.net/forum?id=rkE3y85ee.

Junho Jeong, Changhyun Jung, Taehyun Kim, and Dongil "Dan" Cho. Using machine learning to improve multi-qubit state discrimination of trapped ions from uncertain emccd measurements. *Optics Express*, 31(21):35113–35130, Oct 2023. doi: 10.1364/OE.491301.

Robert M. Kent, Linipun Phuttitarn, Chaithanya Naik Mude, Swamit Tannu, Mark Saffman, Gregory Lafyatis, and Daniel J. Gauthier. Efficient measurement of neutral-atom qubits with matched filters. *Physical Review Applied*, 25(4):044048, 2026. doi: 10.1103/x3h2-nz8y.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is secretly an image segmentation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25303–25313, June 2025. doi: 10.1109/CVPR52734.2025.02356. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kerssies_Your_ViT_is_Secretly_an_Image_Segmentation_Model_CVPR_2025_paper.html.

Ilia Khomchenko, Andrei Fionov, Artem Alekseev, Daniil Volkov, Ilya A. Semerikov, Nikolay N. Kolachevsky, and Aleksey K. Fedorov. Detection states of ions in a paul trap via conventional and quantum machine learning algorithms. arXiv preprint, December 2024.

Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aD7uesX1GF_.

Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-Top-k trick for sampling sequences without replacement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508, 2019. doi: 10.48550/arXiv.1903.06059. URL https://proceedings.mlr.press/v97/kool19a.html.

Ye Li, Chen Tang, Yuan Meng, Jiajun Fan, Zenghao Chai, Xinzhu Ma, Zhi Wang, and Wenwu Zhu. PRANCE: Joint token-optimization and structural channel-pruning for adaptive ViT inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):283–298, 2026. doi: 10.1109/TPAMI.2025.3605239.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. EViT: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn_.

Esther Y. H. Lin, Zhecheng Wang, Rebecca Lin, Daniel Miau, Florian Kainz, Jiawen Chen, Xuaner Zhang, David B. Lindell, and Kiriakos N. Kutulakos. Learning lens blur fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7383–7395, 2025. doi: 10.1109/TPAMI.2025.3578587.

Kaiji Liu, Zhipeng Gan, Shian Guo, Chenxi Wang, Chuanxin Huang, Xuanzeng Song, Bin Fan, Quanxin Mei, Yukai Wu, Binxiang Qi, and Luming Duan. High-Fidelity, Low-Latency State Detection for 300 Trapped Ion Qubits. In *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2025. doi: 10.1109/ISCAS56072.2025.11043463.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oMI9PjOb9Jl.

Binglei Lou, Gautham Duddi Krishnaswaroop, Filip Wojcicki, Ruilin Wu, Richard Rademacher, Zhiqiang Que, Wayne Luk, and Philip H. W. Leong. Low-latency FPGA control system for real-time neural network processing in CCD-based trapped-ion qubit measurement. arXiv preprint, 2025.

Liying Lu, Raphaël Achddou, and Sabine Süsstrunk. Dark noise diffusion: Noise synthesis for low-light image denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–11, 2025. doi: 10.1109/TPAMI.2025.3598330. Early Access.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 2901–2907, 2015. doi: 10.1609/aaai.v29i1.9602.

L. Phuttitarn, B. M. Becker, R. Chinnarasu, T. M. Graham, and M. Saffman. Enhanced measurement of neutral-atom qubits with machine learning. *Physical Review Applied*, 22:024011, 2024. doi: 10.1103/PhysRevApplied.22.024011.

Lewis R. B. Picard, Manfred J. Mark, Francesca Ferlaino, and Rick van Bijnen. Deep learning-assisted classification of site-resolved quantum gas microscope images. *Measurement Science and Technology*, 31(2):025201, 2020. doi: 10.1088/1361-6501/ab44d8.

Shengju Qian, Yi Zhu, Wenbo Li, Mu Li, and Jiaya Jia. What makes for good tokenizers in vision transformer? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13011–13023, 2023. doi: 10.1109/TPAMI.2022.3231442.

Yongming Rao, Zuyan Liu, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dynamic spatial sparsification for efficient vision transformers and convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10883–10897, 2023. doi: 10.1109/TPAMI.2023.3263826.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Alireza Seif, Kevin A. Landsman, Norbert M. Linke, Caroline Figgatt, C. Monroe, and Mohammad Hafezi. Machine learning assisted readout of trapped-ion qubits. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 51(17):174006, 2018. doi: 10.1088/1361-6455/aad62b.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=b9tUk-f_aG.

Xiaolu Su, Mingcheng Liang, Tengyu Zhang, Zhengran Zhao, Peng Yin, Yunkun Yang, Wenqing Dai, Xiaoqin Luo, Xiangliang Li, and Li You. Field-Programmable Gate Array–accelerated exposure-aware convolutional neural network for fluorescence-based qubit readout. *Physical Review Applied*, 25(5):054004, 2026. doi: 10.1103/vjzm-f13w.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Zehuan Yuan, and Ping Luo. Sparse R-CNN: An end-to-end framework for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15650–15664, 2023. doi: 10.1109/TPAMI.2023.3292030.

Julián Tachella, Mike Davies, and Laurent Jacques. UNSURE: Self-supervised learning with unknown noise level and Stein's unbiased risk estimate. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d36dfcdb14473a8526111c221660f2ab-Abstract-Conference.html.

Matthieu Terris, Samuel Hurault, Maxime Song, and Julián Tachella. Reconstruct anything model: A lightweight general model for computational imaging. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/babb3447efd22f411ec90db3bbd38d01-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. doi: 10.48550/arXiv.1706.03762.

Yancheng Wang and Yingzhen Yang. Efficient visual transformer by learnable token merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9597–9608, 2025. doi: 10.1109/TPAMI.2025.3588186.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor DETR: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2567–2575, 2022. doi: 10.1609/aaai.v36i3.20158.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Robert N. Wolf, Joseph H. Pham, Julian Y. Z. Jee, Alexander Rischka, and Michael J. Biercuk. Efficient site-resolved imaging and spin-state detection in dynamic two-dimensional ion crystals. *Physical Review Applied*, 21(5):054067, 2024. doi: 10.1103/PhysRevApplied.21.054067.

Hong Yang and Xianqiang Yang. Zero-shot learning for limited photon budget denoising in structured illumination microscopy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2): 1309–1320, 2026. doi: 10.1109/TPAMI.2025.3612886.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3608–3624, 2024. doi: 10.1109/TPAMI.2023.3347693.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890, 2021. doi: 10.1109/CVPR46437.2021.00681.

Yaoting Zhou, Weisen Wang, Zhuangzhuang Tian, Bin Huang, Huancheng Chen, Donghao Li, Zhongxiao Xu, Li Chen, and Heng Shen. Neural-network-assisted Bayesian qubit readout at the single-photon level for scalable atomic quantum processors. *Physical Review Letters*, 137:013601, 2026. doi: 10.1103/y222-kfxl.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. doi: 10.48550/arXiv.2010.04159.

SUPPLEMENTARY MATERIAL

ABSTRACT

This document provides extended related work, derivations, acquisition details, evaluation protocol, baseline adaptations, extended quantitative results, controlled-perturbation definitions, and deployment measurements supporting the main paper. Figure, table, and equation references without an appendix letter refer to the main paper. No additional cross-platform or cross-modality claim is introduced here.

A RELATED WORK

A.1 STRUCTURED LATENT UNITS AND ADAPTIVE TOKENIZATION

Tokenization controls retained evidence and latent identity. Qian *et al.* analyze its information trade-off (Qian et al., 2023); quadrangle attention learns sampling transformations (Zhang et al., 2024); and learnable token merging removes within-block redundancy (Wang & Yang, 2025). EViT retains attentive tokens and fuses the rest (Liang et al., 2022); dynamic spatial sparsification learns input-dependent computation (Rao et al., 2023); and PRANCE jointly allocates token and channel budgets (Li et al., 2026). ToMe merges similar tokens by bipartite soft matching (Bolya et al., 2023), whereas Iterative Patch Selection uses cross-attention to select patches for image-level recognition (Bergner et al., 2023). Site-AIT instead selects and concatenates local evidence before projection, retaining every site identity for contextual inference.

Detectors structure latent units as object queries or proposals: DETR uses fixed queries (Carion et al., 2020), deformable DETR learns sampling locations (Zhu et al., 2021), Anchor DETR associates queries with reference points (Wang et al., 2022), DAB-DETR refines box anchors (Liu et al., 2022), and Sparse R-CNN learns free proposals (Sun et al., 2023). DEIM improves matching for faster convergence (Huang et al., 2025). These unordered predictions require output-to-site matching; spatial anchoring alone does not define persistent dense-site correspondence.

Object-centric methods instead group evidence into slots. DINOSAUR reconstructs self-supervised visual features (Seitzer et al., 2023); conditional SAVi initializes slots from first-frame locations and propagates them through video with a recurrent predictor–corrector trained on optical flow (Kipf et al., 2022). Site-AIT assumes known site identities and processes frames independently, rather than discovering or tracking objects. Dense segmentation, including U-Net and transformer/context models (Ronneberger et al., 2015; Chen et al., 2024; Zheng et al., 2021; Kerssies et al., 2025; Fu et al., 2025), permits sampling label fields at calibrated positions, but pixel latents do not themselves maintain one-to-one target identity.

A.2 LEARNED READOUT FOR ATOMIC QUBIT ARRAYS

Multi-qubit readout has progressed from thresholded fluorescence to learned estimators of uncertain, crosstalk-affected measurements (Burrell et al., 2010; Seif et al., 2018; Jeong et al., 2023; Khomchenko et al., 2024). Site-resolved imaging includes dynamic ion crystals (Wolf et al., 2024) and quantum gas microscopes (Picard et al., 2020; Impertro et al., 2023); neutral-atom readout uses learning and matched filters to address optical crosstalk (Phuttitarn et al., 2024; Kent et al., 2026). Recent studies cover 300-ion readout (Liu et al., 2025), exposure-aware and low-latency FPGA processing (Su et al., 2026; Lou et al., 2025), and single-photon Bayesian readout in a distinct photon-limited regime (Zhou et al., 2026).

Joint/context-aware readout is established: neural classifiers infer multi-qubit states (Seif et al., 2018), CNN-array uses full-image features for site-specific outputs (Phuttitarn et al., 2024), and MF-array uses dense-site intensity summaries (Kent et al., 2026). The distinction here is adaptive spatial evidence collection for each persistent site latent, not joint readout or dense-site outputs alone.

A.3 Lattice-Indexed and Physics-Constrained Scientific Imaging

Low-light imaging addresses limited observations through learnability-oriented raw denoising (Feng et al., 2024), diffusion and physics-informed detector-noise synthesis (Lu et al., 2025; Feng et al., 2026), and latent denoising objectives for transferable representations (Chen et al., 2025). UNSURE extends Stein-based risk estimation to reconstruction without a known noise level (Tachella et al., 2025). Site-AIT uses an explicit detector-corruption chain for backbone initialization, not deployed restoration.

Physics-informed methods address single-photon reconstruction (Cai et al., 2026), limited-photon microscopy (Yang & Yang, 2026), and aberration-aware localization via self-supervised neural-physics learning (Fu et al., 2026). Optical blur can be learned as a device-specific continuous PSF field (Lin et al., 2025), while RAM conditions non-iterative reconstruction on the forward operator and noise (Terris et al., 2026).

Together, these precedents motivate a precise distinction: Site-AIT binds each token and its adaptive sampling field to a physical site, exchanges register-wide context, and jointly predicts states and sub-pixel coordinates without output-to-site matching. A measured-PSF-initialized renderer couples these outputs only during training; deployment requires neither restoration nor PSF reconstruction.

B Additional Formulation and Tokenization Optimization

The main paper defines one ordered output $(\hat{p}_k, \hat{\boldsymbol{c}}_k, \hat{e}_k)$ for each dense site. For a frame I_i, let $y_{i,k} \in \{0,1\}$ be the state label, $\boldsymbol{c}^*_{i,k}$ be the reference coordinate, $v_{i,k} \in \{0,1\}$ be the site-existence target, and $\ell^e_{i,k}$ be the existence logit with $\hat{e}_{i,k} = \sigma(\ell^e_{i,k})$. The implemented supervised terms are

$$\mathcal{L}_{\text{cls}} = \frac{\sum_{i,k} v_{i,k}\, \text{BCE}(\hat{p}_{i,k}, y_{i,k})}{\sum_{i,k} v_{i,k}}, \tag{B.1}$$

$$\mathcal{L}_{\text{coord}} = \frac{\sum_{i,k} v_{i,k}\, \text{SmoothL1}(\hat{\boldsymbol{c}}_{i,k}, \boldsymbol{c}^*_{i,k})}{\sum_{i,k} v_{i,k}}, \tag{B.2}$$

$$\mathcal{L}_{\text{exist}} = \frac{\sum_{i,k} v_{i,k}\, \text{BCEWithLogits}(\ell^e_{i,k}, v_{i,k})}{\sum_{i,k} v_{i,k}}, \tag{B.3}$$

$$\mathcal{L}_{\text{offset}} = \frac{1}{2NK} \sum_{i,k} \|\Delta \boldsymbol{c}_{i,k}\|_2^2. \tag{B.4}$$

Here, N is the batch size, and $\Delta \boldsymbol{c}_{i,k}$ is the two-dimensional tokenizer offset. The loss weights follow Eq. (7); Appendix C defines reconstruction. All $v_{i,k} = 1$, so existence supervision is positive-only and missing-ion detection remains untested. Rejecting absent sites requires negative existence labels and a loss that includes them.

B.1 Two-site crosstalk approximation

The need for joint site inference can be seen in a two-site approximation. If sites 1 and 2 are separated by $d = \|\boldsymbol{c}_1 - \boldsymbol{c}_2\|_2$, the intensity evaluated at site 1 is

$$I(\boldsymbol{c}_1) \approx b + s y_1 \psi(0) + s y_2 \psi(d) + \varepsilon, \tag{B.5}$$

where b is the background, s is the bright-state signal scale, ψ is the measured spatial response, and ε collects residual noise. The spillover term $s y_2 \psi(d)$ depends on the unknown neighboring state. An independent local classifier must absorb this term as input-dependent noise, whereas site-token self-attention can condition the estimate for site 1 on the evidence assigned to site 2. Equation (B.5) is an explanatory approximation rather than an additional fitted image-formation model.

B.2 Straight-through Top-16 selection

For site k, let $s_{k,j}$ be the selector logit for candidate j and let $g_{k,j}$ be a sampled Gumbel perturbation during training. A temperature-τ soft distribution is

$$q_{k,j} = \frac{\exp[(s_{k,j} + g_{k,j})/\tau]}{\sum_{j'} \exp[(s_{k,j'} + g_{k,j'})/\tau]}. \tag{B.6}$$

If $h_{k,j}$ denotes the hard 16-hot mask produced by the forward Top-16 operation, the straight-through gate is

$$\pi_{k,j}^{\mathrm{ST}} = h_{k,j} - \mathrm{sg}(q_{k,j}) + q_{k,j}, \tag{B.7}$$

where sg stops the gradient. The forward value is the discrete mask $h_{k,j}$; the backward pass follows the softmax surrogate. Because $\sum_j q_{k,j} = 1$ while $\sum_j h_{k,j} = 16$, the soft branch is a biased gradient surrogate rather than a continuous 16-hot relaxation. The temperature is annealed linearly from 0.8 to 0.1 during training. The selected, PSF-weighted feature is $\pi_{k,j}^{\mathrm{ST}} w_{k,j} \boldsymbol{f}_{k,j}$. Only the 16 hard-selected vectors are gathered and concatenated into the 256-dimensional token in Eq. (4). At inference, both the Gumbel perturbation and the soft branch are removed, and deterministic TopK_{16} is applied to the unperturbed logits.

The tokenization offset remains differentiable because the candidate field is sampled bilinearly. Within one interpolation cell, each sampled feature is piecewise linear in the horizontal and vertical sampling coordinates. Gradients therefore propagate through the sampled features to the bounded offset in Eq. (3), except at the measure-zero cell boundaries, where the interpolation stencil changes.

C Training, Noise-Prior Pretraining, and PSF Consistency

For the measured target I in Fig. 4, the corrupted input is

$$\begin{aligned}\widetilde{I} = \mathcal{C}_{[0,1]}\Big[&\mathcal{T}_{\mathrm{mask}} \circ \mathcal{T}_{\mathrm{blur}} \circ \mathcal{T}_{\mathrm{bg}} \\ &\circ \mathcal{T}_{\mathrm{Poisson}} \circ \mathcal{T}_{\mathrm{Gaussian}}(I)\Big].\end{aligned} \tag{C.1}$$

The final clip $\mathcal{C}_{[0,1]}$ follows all five stages. Gaussian noise is also clipped, $I_p^{\mathrm{G}} = \min\{1, \max\{0, I_p + \epsilon_p\}\}$ with $\epsilon_p \sim \mathcal{N}(0, \sigma_r^2)$, to ensure nonnegative rates for independent pixelwise sampling:

$$I_p^{\mathrm{P}} = \frac{\mathrm{Poisson}(s I_p^{\mathrm{G}})}{s}, \tag{C.2}$$

Before later corruptions and clipping, this preserves the mean of I_p^{G} and introduces signal-dependent variance. The pretraining network D_ψ minimizes

$$\mathcal{L}_{\mathrm{pre}} = \lambda_1 \|D_\psi(\widetilde{I}) - I\|_1 + \lambda_{\mathrm{ssim}}\big[1 - \mathrm{SSIM}(D_\psi(\widetilde{I}), I)\big]. \tag{C.3}$$

SSIM (Wang et al., 2004) uses an 11×11 Gaussian window ($\sigma = 1.5$) on $[0, 1]$ intensities. DW-UNet fuses the raw image and upsampled multiscale decoder features into the supervised model's 16-channel field; only its backbone weights are transferred. Table C.1 gives the corruption and optimization settings.

C.1 Saturation-gated PSF consistency

The image rendered by Eq. (6) is compared with the observed frame only at nonsaturated pixels. Let Ω_I denote the image grid. The saturation threshold and mask are

$$\begin{aligned}\mathcal{V}_{\mathrm{train}} &= \{I_i(\boldsymbol{p}) : i \in \mathcal{D}_{\mathrm{train}},\ \boldsymbol{p} \in \Omega_I\}, \\ T_{\mathrm{sat}} &= Q_{0.999}(\mathcal{V}_{\mathrm{train}}), \\ m_i(\boldsymbol{p}) &= \mathbf{1}[I_i(\boldsymbol{p}) < T_{\mathrm{sat}}].\end{aligned} \tag{C.4}$$

The percentile is computed once over all training pixels and frozen, not recomputed per frame or batch. The masked reconstruction loss is

$$\begin{aligned}\mathcal{L}_{\mathrm{recon}} = &\frac{\sum_{i,\boldsymbol{p}} m_i(\boldsymbol{p}) |\hat{I}_i(\boldsymbol{p}) - I_i(\boldsymbol{p})|}{\sum_{i,\boldsymbol{p}} m_i(\boldsymbol{p})} \\ &+ \frac{\lambda_{\mathrm{ssim}}}{N} \sum_i \big[1 - \mathrm{SSIM}_{m_i}(\hat{I}_i, I_i)\big].\end{aligned} \tag{C.5}$$

SSIM_{m_i} averages only windows whose pixels all satisfy $m_i = 1$, excluding saturation from both loss components. PSF kernels have a 6-pixel deposition radius around predicted coordinates; their learnable width is initialized from the measured profile.

Table C.1: Corruption settings and optimization hyperparameters for detector-aware pretraining. Every frame receives all five stages in order; stages 1–3 resample at the granularity listed. Figure 4 illustrates $\sigma = 0.055$ and $b = 0.045$.

Stage	Corruption	Setting
1	Gaussian read noise	$\sigma \sim \mathcal{U}(0.01, 0.08)$, drawn per batch
2	Pixel-wise Poisson shot noise	$s \sim \mathcal{U}(20, 60)$ per batch
3	Background-baseline offset	$b \sim \mathcal{U}(0, 0.06)$, drawn per frame
4	Gaussian PSF blur	3×3 kernel, $\sigma_{\mathrm{blur}} = 0.8$
5	Random block masking	7 gray levels, 30% of frame area; block sides 3–12% of H and W
Optimization		
Optimizer		Adam
Learning rate		1×10^{-3}
Schedule		Cosine annealing, $T_{\max} = 100$
Epochs		100
Batch size		48
Loss weights		$\lambda_1 = 0.8$, $\lambda_{\mathrm{ssim}} = 0.2$
Input size		456×88 pixels

For epoch t, the reconstruction weight follows

$$\lambda_{\mathrm{recon}}^{(t)} = \begin{cases} 0, & t < E_0, \\ \lambda_{\max}(t - E_0)/E_r, & E_0 \le t < E_0 + E_r, \\ \lambda_{\max}, & t \ge E_0 + E_r, \end{cases} \tag{C.6}$$

Here, $E_0 = 10$, $E_r = 5$, and $\lambda_{\max} = 0.05$; this schedule delays consistency until an initial supervised solution is learned. The branch is training-only (Sec. 3.5).

D Apparatus and Imaging Characterization

The experimental frames were acquired from a two-dimensional $^{171}\mathrm{Yb}^{+}$ register confined in a monolithic three-dimensional Paul trap on an alumina chip. The trap operates at a base temperature of $6.1\,\mathrm{K}$ and is driven at $\Omega_{\mathrm{rf}} = 2\pi \times 35.280\,\mathrm{MHz}$. Twenty-eight independently biased DC electrodes arranged in a 4×7 array shape the planar potential and suppress out-of-plane RF micromotion. Ions are loaded by $1030\,\mathrm{nm}$ pulsed laser ablation followed by resonant photoionization at $399\,\mathrm{nm}$. Global $370\,\mathrm{nm}$ beams provide Doppler cooling, electromagnetically induced transparency cooling, optical pumping, and fluorescence detection; a narrow-band $411\,\mathrm{nm}$ beam provides sideband cooling; and $935\,\mathrm{nm}$, $976\,\mathrm{nm}$, and $3432\,\mathrm{nm}$ repumpers clear population from metastable manifolds.

The qubit uses the clock states $|0\rangle \equiv |^2S_{1/2}, F = 0, m_F = 0\rangle$ and $|1\rangle \equiv |^2S_{1/2}, F = 1, m_F = 0\rangle$. After initialization in $|0\rangle$, a composite SK1 $\pi/2$ pulse prepares mixed bright/dark neighborhoods, a global π pulse generates the all-bright condition, and the identity operation generates the all-dark condition. One loading sequence contains a single $^{174}\mathrm{Yb}^{+}$ isotope impurity at a dense site; because it is detuned from the $370\,\mathrm{nm}$ detection resonance, it remains nonfluorescent.

Global $411\,\mathrm{nm}$ and $3432\,\mathrm{nm}$ pulses shelve the $|0\rangle$ population before a $2\,\mathrm{ms}$ fluorescence exposure, during which the unshelved $|1\rangle$ state is bright. An objective with a numerical aperture of 0.33 collects the fluorescence onto an sCMOS sensor using 2×2 binning; each recorded frame contains 456×88 pixels at an object-plane scale of $0.535\,\mu\mathrm{m/pixel}$.

The calibrated nearest-neighbor spacing is 7.80 ± 0.84 pixels, with a minimum of 6.38 pixels. Gaussian fits to 807 isolated bright ions give an empirical width of 1.36 pixels (FWHM 3.21 pixels). The diffraction scale $0.51\lambda/\mathrm{NA} = 0.57\,\mu\mathrm{m}$ is approximately 1.1 object-plane pixels, compared with the measured FWHM of 3.21 pixels. The stacked profile retains 12% of its peak at a radius of 4 pixels

and 1.6% at 7 pixels, compared with 1.3% and 2×10^{-6} for the fitted Gaussian core. Directional intensities differ by as much as a factor of eight at 5 pixels. These measurements describe the spatial overlap addressed by the site-indexed estimator; they are not used as evidence of transfer to another apparatus.

E Dataset, Annotation, and Evaluation Protocol

The data partitions follow Fig. 5. The training/validation partition was generated using the data-split seed 25001 and is held fixed across methods and training runs. This split seed is distinct from the model-training seeds $\{1, 2, 3\}$. Real1600 contains the 1,600 unmodified test frames. Hard400 contains 400 degraded frames generated from a disjoint held-out source pool; its source frames do not occur in the training partition, the validation partition, or Real1600.

E.1 Hard400 construction: stacked P1–P6 perturbations

Bright and dark source frames are sampled deterministically from the respective full-bright and full-dark PNG pools, after excluding frames in the corresponding unmodified test subsets. The $\pi/2$ sources are selected from the remaining acquisition stacks with their source state annotations retained; the 200 selected frames jointly contain 30,000 bright and 30,000 dark site labels. The per-frame perturbation seed is $12345 + i$, where $i = 0, \ldots, 399$ is the zero-based output index. The source-frame mapping, random seed, operator parameters, and output hashes are recorded in the generation manifest and per-frame metadata.

Hard400 contains 100 bright, 100 dark, and 200 $\pi/2$ frames. Sources are min–max normalized to $[0, 1]$; all six operators are applied to every frame as follows:

(1) P1 independently perturbs each input-lattice coordinate by Gaussian noise ($\sigma = 0.5$ px), without changing the image. P5 shifts one randomly selected site's estimated positive local fluorescence component by 0.5 px in a random direction and updates its reference coordinate.
(2) The image receives P2 isotropic blur ($\sigma = 0.2$ px), then randomly oriented P6 anisotropic blur ($\sigma_x = 1.0$ px, $\sigma_y = 1.2$ px).
(3) P3 transforms its input I using $B = Q_{0.20}(I)$, $F = \max(I - B, 0)$, and $I' = I - F + 0.8F + \epsilon$, with independent $\epsilon \sim \mathcal{N}(0, 0.2\,\mathrm{Var}(F))$. P4 then saturates 40 random pixels (approximately 0.1% of a 456×88 frame) to 1, or 255 in the saved 8-bit image.

Outputs are clipped to $[0, 1]$ and quantized to 8 bits; state and existence labels are unchanged. These are synthetic stress tests, not independent measurements or measured P5 trajectories. Unlike the operator-wise sweeps in Appendix I, Hard400 stacks all six perturbations; the later P6 1:1 reference adds no filter but retains these existing degradations.

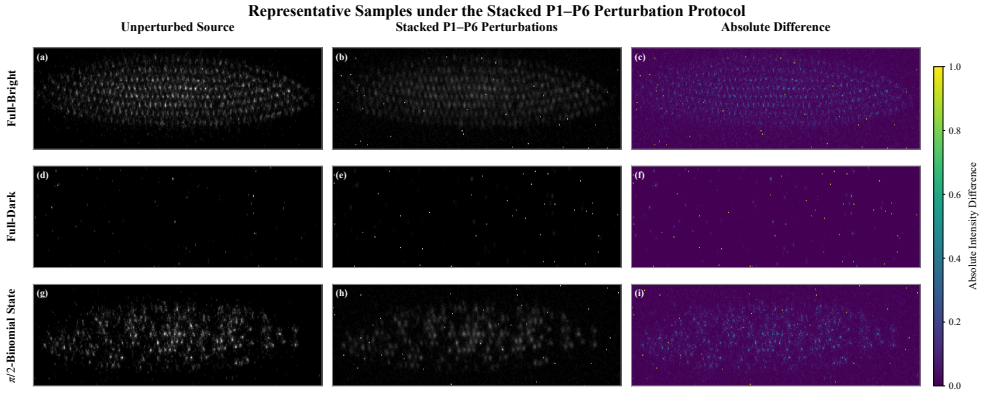


Figure E.1: **Stacked P1–P6 examples from Hard400.** Rows show full-bright (a–c), full-dark (d–f), and $\pi/2$-prepared (g–i) frames. Columns show the source, augmented image, and absolute pixelwise difference, all on a common $[0, 1]$ display range. Section E.1 defines the generation protocol.

E.2 Annotation and quality control

Automatic initialization removes small connected components associated with sparse stray-light artifacts, smooths each frame with a 7×7 Gaussian kernel of width $\sigma = 1.7$ pixels, and uses the 150 strongest local maxima as provisional bright-ion centers. Dark-site coordinates are initialized from the temporally nearest preceding frame in which the corresponding site is bright. Professional annotators then refine both states and centroid coordinates. The automatic initialization agrees with the final human-consensus annotation on 71% of site assignments; this percentage records annotation workload rather than physical-state fidelity.

A separate quality-control pass uses two independent professional annotators, neither of whom is involved in the initial annotation, who blindly re-annotate a random 20% of the test frames without consulting each other. Their state-label agreement is $P_o = 0.9991$ (0.0900% raw disagreement), with Cohen's $\kappa = 0.9982$. For matched sites, the centroid discrepancy is $d_j = \|\boldsymbol{c}_j^A - \boldsymbol{c}_j^B\|_2$; the resulting discrepancies satisfy MAE ≤ 0.30 pixels, RMSE ≤ 0.38 pixels, a 95th-percentile distance ≤ 0.48 pixels, and a maximum distance of less than one pixel. These values quantify inter-annotator consistency, not error relative to independently measured physical states; metric differences of comparable magnitude should therefore be interpreted cautiously rather than against zero-error ground truth.

E.3 Three-seed protocol

Every final learned model is evaluated with seeds $\{1, 2, 3\}$ on identical data partitions. Initialization, data-loader ordering, and stochastic training augmentations vary; the validation and test frames remain fixed. Reported dispersions are sample standard deviations across the three seeds. Deterministic IPM and matched-filter estimators have no seed spread.

E.4 Metric definitions

Bright is the positive class. Accuracy, precision, recall, and F_1 are computed from pooled site decisions. Localization error is computed only for valid sites containing ground-truth bright emitters, for which $v_{i,k}y_{i,k} = 1$. For the residual $\boldsymbol{\delta}_{i,k} = \hat{\boldsymbol{c}}_{i,k} - \boldsymbol{c}^*_{i,k}$ and its mean $\bar{\boldsymbol{\delta}}$ over those emitters, localization error is

$$\left(\frac{\sum_{i,k} v_{i,k}y_{i,k} \|\boldsymbol{\delta}_{i,k} - \bar{\boldsymbol{\delta}}\|_2^2}{\sum_{i,k} v_{i,k}y_{i,k}} \right)^{1/2}. \tag{E.1}$$

Subtracting the mean residual makes this a dispersion measure. It should not be interpreted as absolute accuracy beyond the uncertainty of the reference centroids. AUROC uses the pooled posterior scores. ECE partitions those scores into 15 bins and sums the bin-frequency-weighted absolute difference between empirical accuracy and mean confidence.

F Baseline Definitions and Validation-Only Selection

All normalization statistics, hyperparameters, decoding radii, thresholds, and assignment rules are selected from the training and validation partitions and frozen before evaluation on Real1600–Hard400.

F.1 Classical estimators

IPM grows a site-specific pixel set from a local peak in the mean bright calibration image. Its frozen settings are a contrast factor $\xi = 0.5$, a maximum of 90 pixels per set, and four boundary-growth iterations. The site decision uses the summed intensity and a calibration-derived threshold.

For the matched-filter array (Kent et al., 2026), the scalar mean of the training partition is subtracted from every frame, and every partition is divided by the intensity range of the centered training frames. Validation and test frames do not contribute to the normalization statistics. Site centers are initialized from the calibrated lattice and refined by two-dimensional Gaussian fits to the mean training frame, preserving site identity. For site k, the feature vector concatenates a flattened $s \times s$ local patch, the

mean boundary intensity around each of its $k_{\mathrm{nn}} = 8$ nearest calibrated neighbors, and a constant bias. With site features arranged as columns of X_k and labels in Y_k, the weights are

$$W_k = Y_k X_k^{\mathsf{T}} \left(X_k X_k^{\mathsf{T}} + \alpha I \right)^{-1}. \tag{F.1}$$

The neighbor count is fixed a priori. Patch width $s \in \{2, \ldots, 14\}$, threshold $t \in \{0.01, 0.02, \ldots, 0.99\}$, and $\alpha \in \{0, 10^{-6}, 10^{-4}\}$ are selected independently for each site based on validation classification fidelity and then frozen. For calibration metrics only, a positive-temperature, per-site logistic map centered on the selected threshold is fitted using validation negative log-likelihood. Because this map may reorder scores across sites, pooled AUROC is recomputed after calibration rather than assumed invariant. Inference uses the frozen linear filters and is deterministic.

F.2 Learned query and dense baselines

DETR-style (2020) follows DETR (Carion et al., 2020) with unconstrained learned object queries and full-image cross-attention. For protocol compatibility, DW-UNet replaces ResNet; its full-frame feature field is pooled at a fixed, non-tuned stride of 4, while the state, coordinate, and existence heads use the same MLP architectures as Site-AIT. Because the queries are not site-anchored, the coordinate head predicts normalized absolute image coordinates rather than a residual from a dense site. Training-time Hungarian matching uses state-BCE and normalized ℓ_1 coordinate costs with fixed weights 1 and 5, respectively; evaluation maps predictions to dense site indices by coordinate-only Hungarian assignment before computing site metrics. It does not initialize queries from calibrated coordinates or use a local window, tokenization offset, Top-16 selection, or post-selection PSF weighting.

The DEIM row uses D-FINE with the matchability-aware loss (Huang et al., 2025). Each bright ion is represented by a square target box whose side length is tied to the calibrated PSF, enabling Hungarian one-to-one matching and the matchability objective. Mosaic and mixup are disabled because they break correspondence with the dense register; the Dense O2O targets that depend on such compositing are therefore absent. Detections are assigned to their nearest dense sites, and unassigned sites are classified as dark. Consequently, the row is a register-preserving D-FINE + MAL adaptation, not the unmodified DEIM recipe.

For SegMAN-T (Fu et al., 2025), the official Tiny encoder is initialized from ImageNet-1K and the binary decoder is initialized randomly. Each 456×88 frame is reflection-padded to 480×96, and the output probability map is cropped back to the original size before mask-to-site evaluation. Its training mask, objective, validation-selected evaluation radius, and threshold follow the common dense protocol below.

For EoMT (Kerssies et al., 2025), the reproduction protocol specifies EoMT-B with the `vit_base_patch14_reg4_dinov2` backbone. The encoder is initialized with pretrained DINOv2 backbone weights, without loading a COCO-fine-tuned segmentation checkpoint. Each 456×88 frame is symmetrically reflection-padded, without resizing, to 462×98, the smallest resolution compatible with the 14×14 patch grid. This adds 3 pixels to each horizontal boundary and 5 pixels to each vertical boundary. The output probability map is bilinearly upsampled to the padded resolution and then center-cropped by removing the same margins, yielding the original 456×88 field before mask-to-site evaluation. Thus, SegMAN-T and EoMT follow the same no-rescaling padding-and-cropping policy, with architecture-compatible padded resolutions of 480×96 and 462×98, respectively. The number of segmentation queries is set to 300 as a task-specific adaptation, matching the query count of DETR-style (2020). Queries interact with image tokens in the last three Transformer blocks, following the official EoMT-B configuration. These architectural settings are fixed rather than included in the capacity search.

Dense networks retain binary segmentation training. A radius-3 disk around every bright reference coordinate defines the training mask. During evaluation, the probability map is averaged inside a validation-selected disk around each dense site. The disk radius and decision threshold are selected on validation data and then frozen. The validation-selected evaluation radius is $r_{\mathrm{eval}} = 3$ pixels for all five dense baselines. The resulting thresholds are 0.40, 0.43, 0.49, 0.50, and 0.59 for U-Net, TransUNet-style, SegMAN-T (Fu et al., 2025), SETR-style, and EoMT (Kerssies et al., 2025), respectively.

Table F.1: Mask-to-site threshold sensitivity on the **Real1600-Hard400** test set across thresholds $t \in [0.36, 0.66]$. Bold values indicate the optimal threshold selected strictly on the 5,000-frame validation split using Youden's index $J(t) = \mathrm{TPR}(t) - \mathrm{FPR}(t)$ and frozen for evaluation.

Threshold t	**Per-ion accuracy ↑**				
	U-Net	**TransUNet**	**SegMAN-T**	**SETR**	**EoMT**
0.36	0.6147 ±0.0013	0.6053 ±0.0011	0.6123 ±0.0020	0.6023 ±0.0025	0.6077 ±0.0018
0.39	0.8752 ±0.0004	0.8377 ±0.0011	0.7261 ±0.0020	0.6760 ±0.0024	0.6503 ±0.0018
0.40 (U-Net)	**0.8750** ±0.0004	0.8850 ±0.0010	0.7650 ±0.0019	0.7020 ±0.0024	0.6650 ±0.0017
0.42	0.8676 ±0.0013	0.9472 ±0.0004	0.8180 ±0.0019	0.7391 ±0.0024	0.6879 ±0.0017
0.43 (TransUNet)	0.8610 ±0.0013	**0.9470** ±0.0004	0.8450 ±0.0019	0.7580 ±0.0024	0.7010 ±0.0017
0.45	0.8521 ±0.0013	0.9360 ±0.0010	0.8838 ±0.0019	0.7898 ±0.0023	0.7222 ±0.0017
0.48	0.8313 ±0.0012	0.9136 ±0.0010	0.9152 ±0.0018	0.8266 ±0.0023	0.7529 ±0.0016
0.49 (SegMAN-T)	0.8200 ±0.0012	0.9010 ±0.0010	**0.9148** ±0.0018	0.8350 ±0.0021	0.7650 ±0.0016
0.50 (SETR)	0.8120 ±0.0012	0.8910 ±0.0010	0.9080 ±0.0018	**0.8433** ±0.0020	0.7710 ±0.0016
0.51	0.8040 ±0.0012	0.8809 ±0.0010	0.8982 ±0.0018	0.8430 ±0.0020	0.7778 ±0.0016
0.54	0.7729 ±0.0012	0.8415 ±0.0010	0.8640 ±0.0018	0.8272 ±0.0023	0.7990 ±0.0016
0.57	0.7365 ±0.0013	0.7938 ±0.0010	0.8149 ±0.0019	0.7916 ±0.0023	0.8130 ±0.0017
0.59 (EoMT)	0.7080 ±0.0013	0.7520 ±0.0011	0.7680 ±0.0019	0.7560 ±0.0024	**0.8202** ±0.0010
0.60	0.6969 ±0.0013	0.7402 ±0.0011	0.7543 ±0.0019	0.7426 ±0.0024	0.8205 ±0.0010
0.63	0.6525 ±0.0013	0.6809 ±0.0011	0.6830 ±0.0020	0.6822 ±0.0024	0.7491 ±0.0018
0.66	0.6048 ±0.0013	0.6148 ±0.0011	0.6018 ±0.0020	0.6103 ±0.0025	0.5983 ±0.0018

F.3 Search spaces and budgets

Searches use the fixed 20,000/5,000 training/validation split, batch size 48, seed 1, and a 100-epoch cap with early stopping on validation per-site accuracy; the test set is never consulted. Only selected configurations are retrained with seeds $\{1, 2, 3\}$ (Appendix E). The shared learning-rate grid is $\eta \in \{5 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$.

Model	**Search axes**	**Candidates**
Site-AIT	Shared η; $D \in \{128, 256, 512\}$	12
DETR-style (2020)	Shared η and D; decoder depth $L \in \{2, 4, 6\}$; heads $H \in \{4, 8\}$	72
D-FINE + MAL	Learning rate $\{10^{-4}, 2.5 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$; scale $\{\mathrm{S}, \mathrm{M}, \mathrm{L}\}$; $\gamma \in \{1.0, 1.5, 2.0\}$; box side $\{1.5\sigma_{\mathrm{PSF}}, 2\sigma_{\mathrm{PSF}}, 3\sigma_{\mathrm{PSF}}\}$; default/doubled matching-cost weight	216

DETR and Site-AIT share pretrained DW-UNet initialization and the above training/selection rules (`val_ion_acc`), but not total search budgets: DETR tests six architectures per common optimizer–dimension pair. Its full-image support has a fixed stride, with no template/window search. Site-AIT fixes its architecture; the selected configuration uses $\eta = 10^{-4}$, $D = 256$, $L = 3$ encoder layers, 8 heads, feed-forward width 512, and zero dropout. Checkpoint selection maximizes validation per-ion accuracy.

Dense searches run on one RTX 4090 with Adam, weight decay 10^{-5}, and ReduceLROnPlateau (factor 0.75). Register-preserving augmentation and radius-3 training masks are fixed. Capacity/output resolution is first screened at $\eta = 10^{-4}$ using equally weighted pixelwise BCE and Dice losses. Each selected capacity then crosses the shared four learning rates with three objectives: BCE+Dice, BCE alone, and positive-class-weighted BCE+Dice (4×3 candidates); class weights use training data only.

Within every dense candidate, decoding is evaluated across radii $r_{\text{eval}} \in \{2, 3, 4, 5\}$ pixels and thresholds $t \in \{0.30, 0.31, \ldots, 0.70\}$. For each radius, validation Youden's index selects the threshold; validation per-site accuracy then selects the radius–threshold pair. Architecture, objective, radius, and threshold are frozen before three-seed held-out evaluation.

G Extended Quantitative Results

This appendix expands the aggregate results in Table 1 without changing the evaluation set or decision rules.

G.1 Real1600 and Hard400

Table G.1: Confusion counts for Real1600 (480,000 decisions per row) and Hard400 (120,000). Learned-model counts are three-seed means rounded to integers; Real1600 metrics are averaged after per-seed computation. Deterministic estimators and the positive-class convention follow Appendix E. Hard400 metric means and dispersions appear in Table G.2. Boldface marks the best Real1600 metrics.

Model	TP	TN	FP	FN	Acc. ↑	Precision ↑	Recall ↑	F1 ↑
Real1600 (undegraded frames)								
Site-AIT (ours)	253,235	225,832	780	153	**0.9981**	**0.9969**	**0.9994**	**0.9981**
DETR-style (2020)	252,501	218,426	8,186	887	0.9811	0.9686	0.9965	0.9823
DEIM	242,895	223,385	3,227	10,493	0.9714	0.9869	0.9586	0.9725
Matched filter	237,222	195,444	31,168	16,166	0.9014	0.8839	0.9362	0.9093
IPM	234,992	191,154	35,458	18,396	0.8878	0.8689	0.9274	0.8972
Hard400 (augmented frames)								
Site-AIT (ours)	59,924	59,702	298	76	Metric means and dispersions: Table G.2			
DETR-style (2020)	59,638	56,957	3,043	362				
DEIM	55,573	58,768	1,232	4,427				
Matched filter	53,674	48,424	11,576	6,326				
IPM	52,854	46,883	13,117	7,146				

G.2 FIVE TEST SUBSETS

Table G.2: Five-subset site-state metrics under Appendix E: learned models show three-seed means ± sample standard deviations; deterministic estimators show single values. Dashes mark unreported all-dark metrics.

Model	Acc. ↑	Precision ↑	Recall ↑	F1 ↑
All-dark				
Site-AIT (ours)	0.9960 ± 0.0003	–	–	–
DETR-style (2020)	0.9630 ± 0.0002	–	–	–
DEIM	0.9853 ± 0.0003	–	–	–
Matched filter	0.8620	–	–	–
IPM	0.8420	–	–	–
Hard400				
Site-AIT (ours)	0.9969 ± 0.0005	0.9951 ± 0.0007	0.9987 ± 0.0002	0.9969 ± 0.0005
DETR-style (2020)	0.9716 ± 0.0003	0.9515 ± 0.0004	0.9940 ± 0.0002	0.9722 ± 0.0003
DEIM	0.9528 ± 0.0006	0.9783 ± 0.0012	0.9262 ± 0.0014	0.9515 ± 0.0006
Matched filter	0.8508	0.8226	0.8946	0.8571
IPM	0.8311	0.8012	0.8809	0.8391
$\pi/2$ pulse				
Site-AIT (ours)	0.9981 ± 0.0006	0.9970 ± 0.0010	0.9992 ± 0.0001	0.9981 ± 0.0006
DETR-style (2020)	0.9830 ± 0.0003	0.9704 ± 0.0005	0.9964 ± 0.0001	0.9832 ± 0.0003
DEIM	0.9717 ± 0.0007	0.9873 ± 0.0019	0.9557 ± 0.0008	0.9712 ± 0.0006
Matched filter	0.9105	0.8900	0.9367	0.9128
IPM	0.8987	0.8762	0.9285	0.9016
$^{174}Yb^{+}$				
Site-AIT (ours)	0.9997 ± 0.0003	0.9996 ± 0.0001	0.9998 ± 0.0002	0.9997 ± 0.0002
DETR-style (2020)	0.9715 ± 0.0004	0.9726 ± 0.0004	0.9980 ± 0.0000	0.9851 ± 0.0002
DEIM	0.9740 ± 0.0002	0.9925 ± 0.0002	0.9799 ± 0.0004	0.9862 ± 0.0001
Matched filter	0.9290	0.9921	0.9324	0.9613
IPM	0.9191	0.9907	0.9231	0.9557
All-bright				
Site-AIT (ours)	0.9998 ± 0.0000	1.0000 ± 0.0000	0.9998 ± 0.0000	0.9999 ± 0.0000
DETR-style (2020)	0.9965 ± 0.0000	1.0000 ± 0.0000	0.9965 ± 0.0000	0.9982 ± 0.0000
DEIM	0.9600 ± 0.0000	1.0000 ± 0.0000	0.9600 ± 0.0000	0.9796 ± 0.0000
Matched filter	0.9360	1.0000	0.9360	0.9669
IPM	0.9260	1.0000	0.9260	0.9616

Figure G.1 visualizes the accuracy and F_1 entries of Table G.2.

G.3 PAIRED UNCERTAINTY ANALYSIS

Frame identity is preserved during resampling. The hierarchical bootstrap first resamples frames and then resamples the three paired seed differences within each selected frame. The frame-cluster bootstrap first averages the paired seed differences and then resamples frames. Each interval uses 100,000 resamples. A strict frame win requires more correct site decisions after seed averaging; ties are not counted as wins.

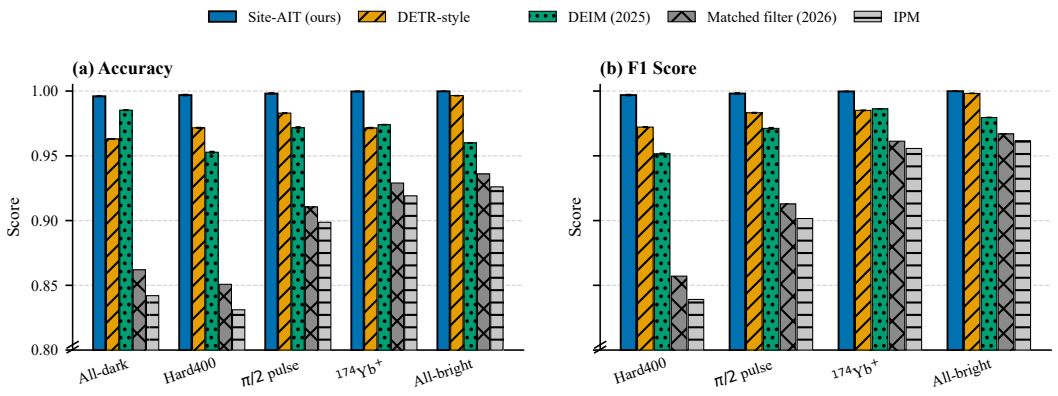


Figure G.1: **Site-AIT leads four comparators across five subsets of the 300-site test benchmark.** All-dark, all-bright, and $\pi/2$-pulse subsets contain homogeneous dark, homogeneous bright, and mixed states; the $^{174}\mathrm{Yb}^{+}$ subset contains one permanent-dark impurity per frame, and Hard400 contains 400 synthetically perturbed frames. (a) Pooled per-ion accuracy. (b) Bright-class F_1, the harmonic mean of precision and recall; it is not reported for the all-dark subset. Bars and error bars follow the means and sample standard deviations in Table G.2; deterministic estimators have no error bars. Both vertical axes start at 0.80, as indicated by the axis breaks.

Table G.3: Paired accuracy analysis on Real1600-Hard400. All differences are Site-AIT minus the indicated baseline.

Baseline	**Mean ΔAcc.**	**95% confidence interval**		**Win rate**
		Hierarchical	**Frame-cluster**	
DETR-style (2020)	$+0.0186$	$[+0.0183, +0.0189]$	$[+0.0184, +0.0188]$	100.0000%
DEIM	$+0.0301$	$[+0.0298, +0.0305]$	$[+0.0299, +0.0304]$	100.0000%

G.4 Posterior ranking and calibration

Table G.4: Calibration metrics on Real1600-Hard400. Values are reported as the mean $\pm$ standard deviation over three seeds.

Model	**AUROC ↑**	**ECE ↓**
Site-AIT (ours)	$\mathbf{0.9980} \pm 0.0003$	$\mathbf{0.0120} \pm 0.0002$
DETR-style (2020)	0.9940 ± 0.0006	0.0180 ± 0.0003
DEIM	0.9900 ± 0.0007	0.0220 ± 0.0004
U-Net	0.9100 ± 0.0003	0.0720 ± 0.0003
TransUNet-style	0.9680 ± 0.0002	0.0580 ± 0.0003
SegMAN-T	0.9350 ± 0.0004	0.0640 ± 0.0012
SETR-style	0.8850 ± 0.0017	0.1050 ± 0.0015
EoMT	0.8750 ± 0.0018	0.0920 ± 0.0020

H ARCHITECTURAL-ABLATION DEFINITIONS

The full model and two attention controls share the fused DW-UNet field $F_i = \{\boldsymbol{f}_{i,p}\}_{p=1}^{HW}$ ($\boldsymbol{f}_{i,p} \in \mathbb{R}^C$), site-conditioning vectors $\boldsymbol{r}_k$ (Eq. (5)), $K = 300$, and downstream Transformer width, head count, depth, and readout heads. Training objectives, data partitions, and seeds are matched; each control is independently optimized. The full model uses Eqs. (4) and (5), followed by $\{\boldsymbol{h}_{i,k}\}_{k=1}^{K} = \mathrm{Enc}(\{\boldsymbol{z}_{i,k}\}_{k=1}^{K})$.

Figure H.1 instead tests input-coordinate sensitivity with fixed checkpoints; Appendix I defines P1 and its two reference choices.

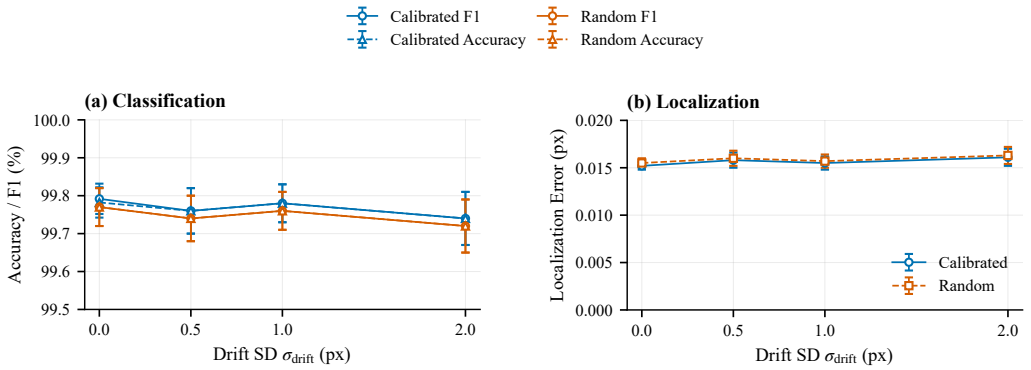


Figure H.1: **Fixed-checkpoint sensitivity to initial coordinates.** On Real1600–Hard400, P1 perturbs either the calibrated lattice or 300 uniformly sampled, site-index-assigned references (Appendix I); 0.0 pixels denotes the unperturbed reference for each condition. Central values follow Sec. 4.1, with sample standard deviations across seed means after three draws per seed. (a) Accuracy and bright-class F_1 nearly coincide on the 99.5–100.0% axis. (b) Localization error is the mean-subtracted RMS coordinate residual in pixels, not absolute centroid error.

H.1 UNCONSTRAINED FULL-IMAGE DENSE CROSS-ATTENTION

This control replaces the adaptive local tokenizer, not downstream site-token self-attention. Site- and coordinate-conditioned queries attend to all HW fused features. Normalized image coordinates $\widetilde{\boldsymbol{p}} \in [0,1]^2$ are encoded as $\boldsymbol{u}_p = \phi_{\text{imgcoord}}(\widetilde{\boldsymbol{p}})$ in the keys:

$$\boldsymbol{q}_k = W_Q \boldsymbol{r}_k, \tag{H.1}$$

$$\boldsymbol{k}_{i,p} = W_K(\boldsymbol{f}_{i,p} + \boldsymbol{u}_p), \qquad \boldsymbol{v}_{i,p} = W_V \boldsymbol{f}_{i,p}, \tag{H.2}$$

$$\alpha_{i,k,p} = \mathrm{softmax}_{p=1,\dots,HW}\left(\frac{\boldsymbol{q}_k^{\top}\boldsymbol{k}_{i,p}}{\sqrt{d}}\right), \tag{H.3}$$

$$\boldsymbol{x}_{i,k}^{\text{glob}} = W_O \sum_{p=1}^{HW} \alpha_{i,k,p}\boldsymbol{v}_{i,p}. \tag{H.4}$$

Equation (H.4) shows one head; the multi-head implementation matches the full model's head count and width. Tokens $\boldsymbol{z}_{i,k}^{\text{glob}} = \boldsymbol{x}_{i,k}^{\text{glob}} + \boldsymbol{r}_k$ enter the common encoder and heads. Absolute key positions do not restrict support: there is no local window, tokenizer offset, Top-16 selection, attention mask/bias, or post-selection learned PSF mask. Training-only PSF reconstruction is retained. Thus, this control changes evidence gathering rather than cross-site exchange.

H.2 Self-only site attention without cross-site exchange

This control retains the full tokenizer, persistent site/coordinate embeddings, and Transformer, changing only the attention topology:

$$M_{kk'} = \begin{cases} 0, & k = k', \\ -\infty, & k \neq k', \end{cases} \qquad \{\boldsymbol{h}_{i,k}\}_{k=1}^{K} = \mathrm{Enc}(\{\boldsymbol{z}_{i,k}\}_{k=1}^{K}; M). \tag{H.5}$$

Q/K/V projections, feed-forward blocks, normalization, depth, and parameter count are unchanged; the diagonal mask prevents cross-site information exchange while preserving the common heads and objectives.

H.3 Capacity, other controls, and timing

The full and self-only models each have 5,631,857 trainable parameters. Dense cross-attention adds projections separate from the matched downstream Transformer, yielding 5,964,581 parameters (5.9% more), not a smaller budget. The coordinate-supervision control keeps its head and reconstruction pathway but sets $\lambda_{\mathrm{coord}} = 0$. The no-weight-warm-up control sets $\lambda_{\mathrm{recon}} = \lambda_{\max}$ from epoch zero while retaining the early state-gradient gate, changing only the reconstruction-weight schedule. Latency follows the native PyTorch FP32 boundary in Appendix J.

I Controlled Perturbations and Additional Qualitative Evidence

Table I.1: Coverage audit comparing self-supervised pretraining corruptions with test-time physical robustness stress tests.

Phenomenon / stress operator	**Pretraining** (Appendix C)	**Evaluated in Fig. I.1 sweep**
Gaussian detector read noise	✓ (simulated channel)	Not evaluated
Poisson shot / photon noise	✓ (simulated channel)	Not evaluated
Uniform background offset	✓ (simulated channel)	Not evaluated
PSF broadening / defocus blur	✓ (simulated channel)	✓ (operator P2)
Anisotropic PSF elongation / astigmatism	Held out	✓ (operator P6)
Random block masking	✓ (spatial drop)	Not evaluated
Initial-coordinate perturbation	Held out	✓ (operator P1)
Foreground photon budget / SNR loss	Held out	✓ (operator P3)
Hot-pixel sensor injection	Held out	✓ (operator P4)
Single-site ion displacement / swap	Held out	✓ (operator P5)

Here, "Held out" is used only in the pretraining column to denote an operator absent from the pretraining corruption pipeline; "Not evaluated" instead denotes a corruption omitted from the quantitative sweeps in Figure I.1. P2 scales an isotropic-blur channel already represented during pretraining. Although the foreground-attenuation operation in P3 is held out, its low-SNR regime partially overlaps with the Gaussian read-noise and Poisson shot-noise channels used during pretraining. P2 and P3 therefore assess resilience to partially modeled noise families rather than strictly decoupled zero-shot transfer. By contrast, the anisotropic elongation in P6 is absent from pretraining and provides a strictly decoupled physical stress test.

Figure I.1 summarizes the fixed-checkpoint responses to the six P1–P6 perturbations, reporting classification metrics and localization error across severity levels.

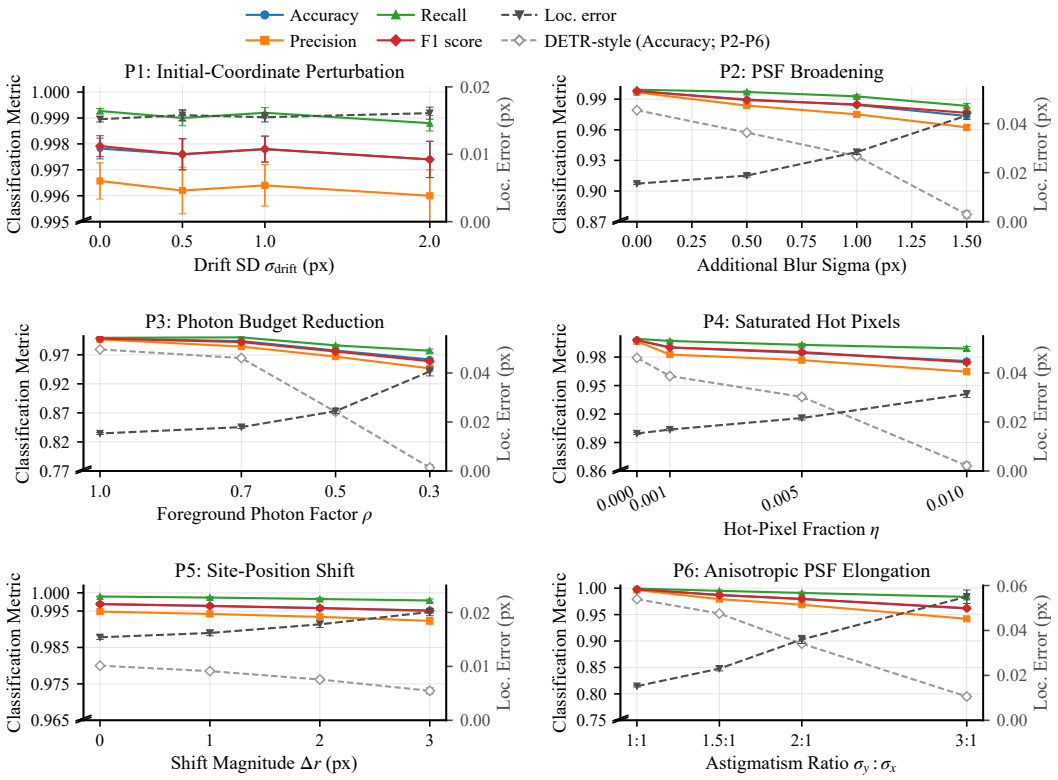


Figure I.1: **Fixed-checkpoint stress tests on Real1600–Hard400 (1,600 experimental and 400 augmented test frames).** P1 adds Gaussian offsets to calibrated input coordinates; P2 broadens the point-spread function (PSF) isotropically; P3 reduces foreground intensity with added Gaussian noise; P4 injects saturated hot pixels; P5 displaces one site per frame over $\Delta r \in \{0, 1, 2, 3\}$ pixels; and P6 applies anisotropic PSF blur. P6's 1:1 reference adds no filter; other ratios use $\sigma_x = 1.0$ pixel. Site-AIT's accuracy and bright-class precision, recall, and F_1 use the left axes; localization error uses the right axes and is the root-mean-square coordinate residual after subtracting its mean, in pixels. Dashed gray diamonds show DETR-style (2020) accuracy for P2–P6 only. Central values follow Sec. 4.1; error bars show one sample standard deviation across seeds; for P1, three perturbation draws are first averaged within each seed. The same checkpoints are retained across severity levels. P1 accuracy and F_1 are nearly coincident; P3's reversed axis places increasing severity from left to right.

In the updated P5 sweep, increasing Δr from 0 to 3 pixels changes Site-AIT accuracy from 0.9969 ± 0.0004 to 0.9951 ± 0.0006 and localization error from 0.0154 ± 0.0004 to 0.0201 ± 0.0007 pixels; at 3 pixels, DETR-style (2020) accuracy is 0.9731 ± 0.0008. These results support stability over the tested displacement range, not robustness beyond it.

I.1 Operator definitions

P1: initial-coordinate perturbation. On Real1600–Hard400, each initial reference coordinate receives an independent Gaussian offset with a standard deviation of $\sigma_{\text{drift}} \in \{0, 0.5, 1, 2\}$ pixels. The two reference conditions are the calibrated nominal lattice and 300 coordinates sampled uniformly over the full image. The random coordinates are sorted by the same vertical-first, horizontal-second rule as the calibrated nominal lattice; the kth sorted random coordinate is assigned to physical site index k, rather than using a random permutation of site identities. The target count remains fixed. In both conditions, the resulting coordinates replace all coordinate-dependent inputs: the tokenizer sampling centers, coordinate positional encodings, and coordinate-head reference centers.

The tokenizer window and correction bounds follow Eq. (3) and its accompanying definition. The coordinate head remains referenced to the input coordinate, with the per-axis image-diagonal bound defined in Sec. 3.3; the tokenizer correction does not enlarge this bound.

For each condition and perturbation level, let $m_{s,d}$ be the test-set metric for training seed $s \in \{1, 2, 3\}$ and random draw $d \in \{1, 2, 3\}$. We first average within each seed, $\bar{m}_s = \frac{1}{3}\sum_{d=1}^{3} m_{s,d}$, and then report $\bar{m} = \frac{1}{3}\sum_{s=1}^{3} \bar{m}_s$ and the sample standard deviation $\sqrt{\frac{1}{2}\sum_{s=1}^{3}(\bar{m}_s - \bar{m})^2}$. The error bars therefore describe variation across the three seed means, not the spread of nine independent training runs. Checkpoints are fixed across levels, with 0.0 pixels as the unperturbed reference for each coordinate choice; calibrated-reference classification reuses the combined-count full-model results. Figure I.1 shows calibrated references; Fig. H.1 compares both choices. Neither is a retrained architectural control.

P2: isotropic PSF broadening. The input is convolved with an additional Gaussian of width $\sigma_{\mathrm{blur}} \in \{0, 0.5, 1.0, 1.5\}$ pixels.

P3: photon-budget reduction. Each frame is decomposed into foreground $I_{\mathrm{fg}} = \max(0, I - b_{\mathrm{est}})$ and residual background, where b_{est} is the protocol's percentile-based background estimate. For $\rho \in \{1.0, 0.7, 0.5, 0.3\}$, the perturbed frame is

$$I_\rho = I - I_{\mathrm{fg}} + \rho I_{\mathrm{fg}} + \epsilon_\rho, \qquad \epsilon_\rho \sim \mathcal{N}(0, (1-\rho)\,\mathrm{Var}(I_{\mathrm{fg}}))\,. \tag{I.1}$$

The additional Gaussian term preserves a comparable marginal dynamic range while reducing the foreground signal-to-noise ratio.

P4: hot pixels. A fraction $\eta \in \{0, 0.001, 0.005, 0.01\}$ of pixels is selected uniformly and replaced by the frame saturation value.

P5: single-site displacement. One valid site per frame is shifted by $\Delta r \in \{0, 1, 2, 3\}$ pixels in a random image-plane direction while the supplied nominal lattice remains unchanged.

P6: anisotropic elongation. The 1:1 condition uses the original test frames without filtering and is the unperturbed reference. For an aspect ratio $r \in \{1.5, 2, 3\}$, an asymmetric Gaussian with $\sigma_x = 1.0$ pixel and $\sigma_y = r\sigma_x$ is applied without retraining.

Table I.2: Zero-shot model robustness to synthetic astigmatism on Real1600-Hard400. The 1:1 condition uses the original test frames without filtering and is the matched unperturbed reference. For aspect ratios $\sigma_y : \sigma_x \in \{1.5:1, 2:1, 3:1\}$, an asymmetric two-dimensional Gaussian blur with $\sigma_x = 1.0$ pixel is applied to the test frames. The same checkpoint is evaluated for each seed under every condition. The 1:1 values for Site-AIT, DETR-style (2020), and DEIM reuse the combined-count overall results; other values are means over seeds $\{1, 2, 3\}$.

$\sigma_y : \sigma_x$	Model	Acc. ↑	Precision ↑	Recall ↑	F1 ↑
Original (1:1)	**Site-AIT (ours)**	**0.9978**	**0.9966**	**0.9993**	**0.9979**
	DETR-style (2020)	0.9792	0.9653	0.9960	0.9804
	DEIM	0.9677	0.9853	0.9524	0.9686
	TransUNet	0.9470	0.9054	0.9983	0.9496
	U-Net	0.8750	0.8549	0.9033	0.8784
1.5:1	**Site-AIT (ours)**	**0.9870**	**0.9793**	**0.9950**	**0.9871**
	DETR-style (2020)	0.9520	0.9589	0.9445	0.9516
	DEIM	0.9450	0.9495	0.9400	0.9447
	TransUNet	0.8840	0.8647	0.9105	0.8870
	U-Net	0.7850	0.7591	0.8350	0.7952
2:1	**Site-AIT (ours)**	**0.9795**	**0.9687**	**0.9910**	**0.9797**
	DETR-style (2020)	0.8950	0.9072	0.8800	0.8934
	DEIM	0.8800	0.8878	0.8700	0.8788
	TransUNet	0.7950	0.7906	0.8025	0.7965
	U-Net	0.6800	0.6667	0.7200	0.6923
3:1	**Site-AIT (ours)**	**0.9615**	**0.9420**	**0.9835**	**0.9623**
	DETR-style (2020)	0.7950	0.8105	0.7700	0.7897
	DEIM	0.7800	0.7917	0.7600	0.7755
	TransUNet	0.6520	0.6573	0.6350	0.6460
	U-Net	0.5900	0.5865	0.6100	0.5980

I.2 Qualitative perturbation example

Figure I.2 compares SegMAN-T and Site-AIT site-level errors in the same region under uncorrupted, added-read-noise, and reduced-intensity conditions.

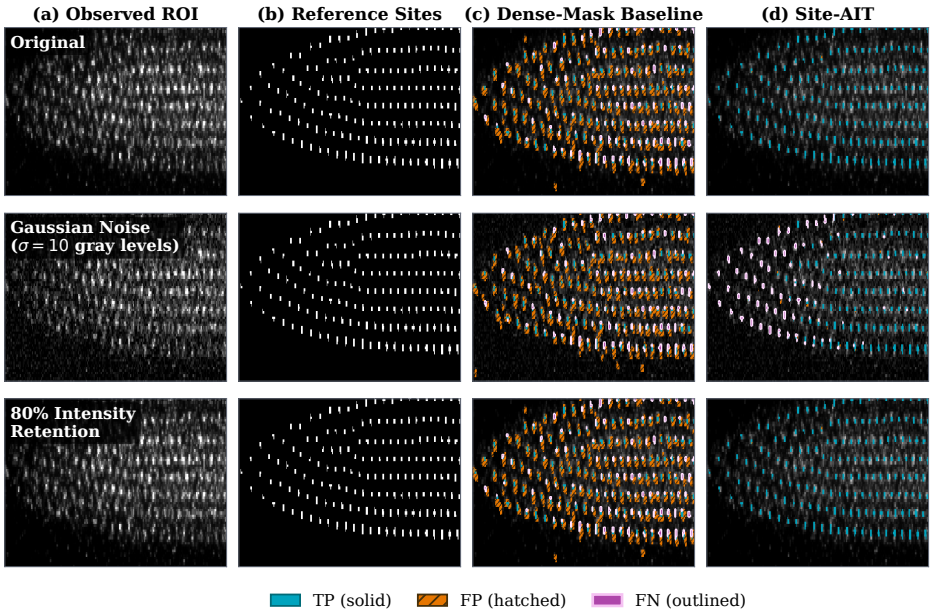


Figure I.2: **Qualitative site-level errors under three input conditions.** Columns (a–d) show the observed region of interest (ROI), reference bright sites, SegMAN-T (2025) (Fu et al., 2025) errors (labeled *Dense-Mask Baseline*), and Site-AIT errors. Rows show the same ROI without added corruption, with Gaussian read noise ($\sigma = 10$ gray levels), and with 80% intensity retention. With bright as the positive class and a 3-pixel matching radius, solid cyan marks true positives (TP), hatched orange false positives (FP), and outlined magenta false negatives (FN).

J Efficiency, Quantization, and Deployment

Native PyTorch FP32 timing uses batch size one, 50 warm-up iterations, and 1,000 timed runs on an RTX 4090. TensorRT and FPGA measurements use the matched trained checkpoints described in the main paper; compiler precision and hardware differ, so these values are not interpreted as module-level speedups.

Figure J.1 compares batch-one inference latency and per-ion accuracy for the neural baselines and for Site-AIT deployment variants, excluding image acquisition.

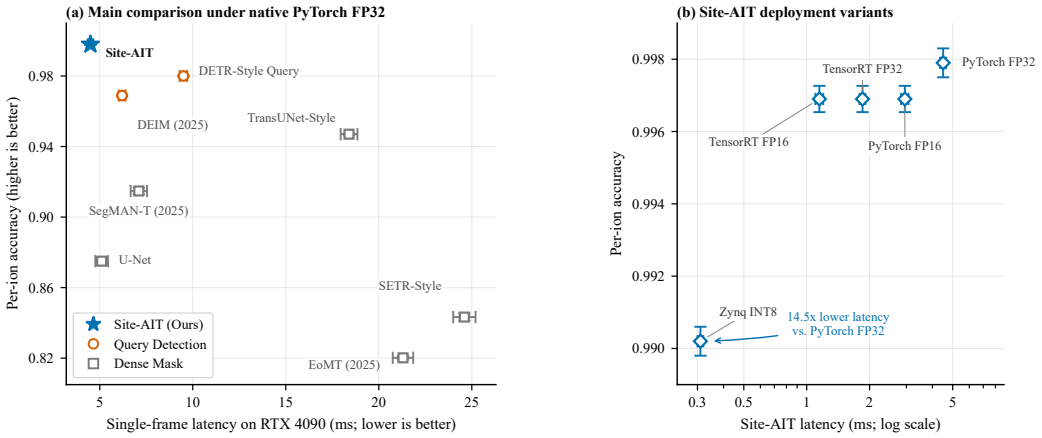


Figure J.1: **Batch-one model-inference latency versus per-ion accuracy on Real1600–Hard400 (1,600 experimental and 400 augmented test frames).** (a) Site-AIT and the neural comparators under the common RTX 4090 PyTorch FP32 protocol; classical estimators are not plotted. (b) Site-AIT deployment variants using matched checkpoints: PT denotes PyTorch, TRT TensorRT, FP32/FP16 32-/16-bit floating point, and INT8 8-bit integer inference on the Zynq XCZU9EG field-programmable gate array. GPU variants use the RTX 4090. Panel (b) uses a logarithmic latency axis. Points follow Sec. 4.1; error bars show one sample standard deviation across three seeds. Latency excludes image acquisition; the 14.5-fold annotation compares native PyTorch FP32 with Zynq INT8.

Table J.1: Cross-platform Site-AIT inference. Central values follow the main-paper reporting convention in Sec. 4.1; $\pm$ denotes the sample standard deviation over the same three checkpoints under each engine and precision setting.

Platform / engine	**Precision**	**Acc.** $\uparrow$	**F1** $\uparrow$	**Latency (ms)** $\downarrow$	**FPS** $\uparrow$
RTX 4090 PyTorch	FP32	0.9978 ± 0.0004	0.9979 ± 0.0004	4.50 ± 0.15	222
	FP16	0.9978 ± 0.0004	0.9979 ± 0.0004	2.95 ± 0.12	338
RTX 4090 TensorRT	FP32	0.9978 ± 0.0004	0.9979 ± 0.0004	1.85 ± 0.06	540
	FP16	0.9978 ± 0.0004	0.9979 ± 0.0004	1.15 ± 0.05	870
Zynq XCZU9EG custom IP	INT8	0.9902 ± 0.0004	0.9904 ± 0.0009	0.31 ± 0.01	3,226