GaugeAnything: Promptable Quantitative Inspection for Industrial Micro-Vision

Masks In, Millimeters Out

Hyunwoo Joo
Falcon Eyes Inc.
github.com/falcons-eyes/GaugeAnything

Abstract

Vision foundation models stop at perception: Segment Anything yields masks, Depth Anything relative depth, counting models counts—but industrial inspection decisions are made in *physical units*: "crack width 0.42 mm ± 0.05, condition Fair." As of mid-2026 no promptable model emits such measurements, and—strikingly—no public dataset pairs defect photos with physically measured widths. We introduce GaugeAnything, a promptable quantitative inspection pipeline that couples a concept-promptable segmentation backbone (SAM 3) with a *metrology core*: tilt-robust pixel-to-millimeter scale resolvers, mask-geometry measurement, a boundary-regime router (sharp/fuzzy/boundaryless), and a signal-based width estimator. Our central finding is a decomposition we call **"mask for *where*, signal for *how wide*"**: reading width from binary mask geometry costs 4–6× accuracy, whereas using the mask only to localize the centerline and regressing width from the raw intensity profile attains **23.2 µm median error** against manually measured physical ground truth (krkCMd, 19,098 profiles)—within 2× of the dataset authors' specialized supervised model. On mm-accurate industrial CAD ground truth (T-LESS) the full promptable pipeline measures part dimensions at 2.5% **median error**, statistically indistinguishable from a perfect-mask ceiling of 2.83%. We further report the first peer-comparable SAM 3 crack-segmentation IoU (0.442 ± 0.011 crack-only, 2.44× classical baselines), a calibration ladder in which a six-parameter logit-threshold + quantile scheme (0.437 rel. err.) beats a 1.9M-parameter neural refiner (0.564), and a per-source conformal audit of the ladder: 90% intervals transfer to held-out sources only in their non-adaptive form—the efficient adaptive variants (normalized conformal, CQR) collapse to 21%/11% coverage on the worst source, whose failure no feature-based difficulty signal detects (a concept shift, not a covariate shift). Beyond static scenes, gated handheld RGB-D measurement holds 1.06% median error (TUM) and oracle-gated egocentric multiview fusion 8.7% median 3-D dimension error (Aria Digital Twin, 229 objects), while an ROI-only control collapses to 316%—localization, again, is the binding constraint. We also report honest negative results: zero-shot counting collapses on dense touching parts regardless of prompt wording, and an alpha-matting head that wins 20× on synthetic data initially failed to transfer until the synthesis was made directionally realistic (real-fray IoU $0.483 \rightarrow 0.949$). All code, audited benchmarks, and task heads are released.

1 Introduction

Industrial inspection is a measurement discipline. Whether a concrete crack requires repair depends on whether its width exceeds a codified threshold (e.g. 0.3 mm); whether an assembly passes depends on bolt counts and spacings; maintenance scheduling depends on graded severity. Yet the modern "promptable anything" toolbox—Segment Anything [1–3], Depth Anything [4], class-agnostic counting [5]—stops one step short of the quantity that drives the decision. The model that

tells an inspector *how many millimeters* does not exist, and the gap is structural: our survey (§2) finds that the closest prior work validates promptable measurement on a single object at "±10%" [6], and that public crack datasets provide pixel masks only—none pairs images with physically measured widths.

GaugeAnything fills this gap with a deliberately decomposed design (§3): a frozen concept-promptable backbone (SAM 3) proposes instances from a noun-phrase prompt; a *metrology core* converts instances into *Inspection Atoms* {mask, class, count, size in mm $\pm\sigma$, condition grade, confidence}. The core contains (i) *scale resolvers*—fiducial markers, known-dimension objects, and a tilt-robust homography (*PlaneScale*) that reduces a 19.3% scale error at 50° camera tilt to 0.7%; (ii) *mask geometry*—skeleton + Euclidean distance transform width profiles, equivalent diameters, spacings; (iii) a *boundary-regime router* that dispatches sharp-boundary defects to binary segmentation, fuzzy boundaries (fray) to alpha matting, and boundaryless fields (uneven/mura) to illumination-residual modeling; and (iv) a *signal-based width estimator* that is the paper's central contribution.

Central finding: mask for *where*, signal for *how wide*. A binary mask quantizes a gradual intensity transition into a hard boundary, destroying sub-pixel evidence. We show this is the dominant error source for thin-structure measurement: on scanner imagery with manually measured physical widths (krkCMd [7]), widths read from SAM 3 mask geometry err by 144–186 µm, while a 1-D CNN regressing width from the raw 501-pixel brightness profile—using the mask *only* to localize the profile position—achieves 39.9 µm MAE and 23.2 µm median where localization succeeds (§4.8). The information was in the image all along; the binary mask was discarding it.

Contributions.

1. **Task & system**: Promptable Quantitative Inspection—image + noun phrase → Inspection Atoms—implemented as an open, audited pipeline on a frozen SAM 3 backbone with a license-clean metrology core.
2. **Signal-based metrology**: the *where/how-wide* decomposition, validated on physical ground truth: 23.2 µm median (localization-gated) vs. 144–186 µm for mask geometry; on T-LESS CAD ground truth, promptable part measurement at 2.5% median ≈ the 2.83% perfect-mask ceiling.
3. **Calibration ladder with an uncertainty audit**: a systematic, honestly-reported progression for width bias—raw 0.730 → neural refiner 0.564 → 5-number quantile calibration 0.480 → a tiny learned head 0.472 → logit-threshold + quantile 0.437 (held-out sources)—in which the simple method beats the large learned one, and we say so; plus a per-source conformal coverage audit (90% target) showing that only non-adaptive intervals survive source shift.
4. **First numbers**: the first peer-comparable SAM 3 crack IoU (0.442 ± 0.011 crack-only; the field's supervised SOTA reports two-class mIoU, a metric trap we quantify); the first promptable physical crack-width MAE; per-class soft-map results on Magnetic-Tile fray/uneven, which prior work reports only in aggregate; gated dynamic-scene measurement curves (TUM handheld 1.06%; ADT egocentric oracle bound 8.7% with a 316% ROI-only control).
5. **Honest negatives, released**: prompt-synonym collapse and its ensemble rescue; zero-shot dense counting failure (six prompts, 0% acc@10%) with a partial SAHI recovery; a synthetic-to-real matting failure and its directional-synthesis fix; an equivalent-width hypothesis rejected

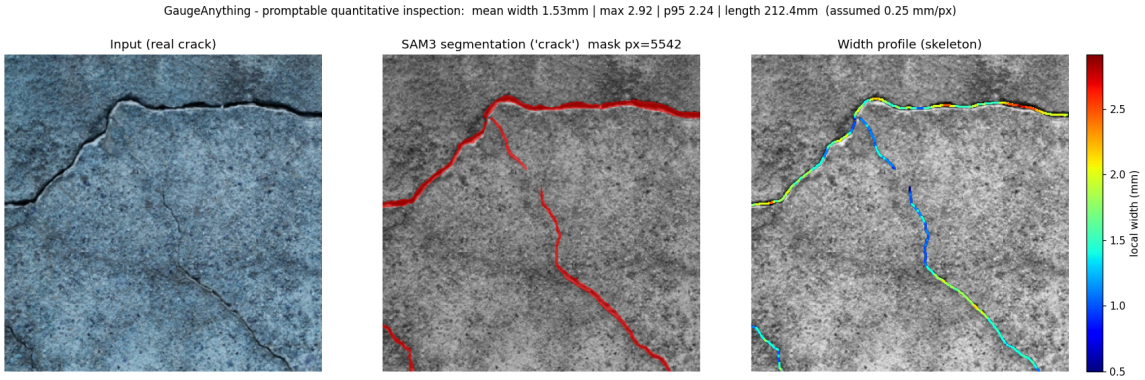

Figure 1: GaugeAnything on a real concrete surface: prompt "crack" → SAM 3 instances → skeleton+EDT width profile → physical width (assumed scale shown; absolute scale comes from markers or known-dimension objects).

by textured backgrounds; adaptive conformal intervals and three difficulty signals all missing a worst-source concept shift.

2 Related Work

Promptable segmentation. SAM [1], SAM 2 [2] and SAM 3 [3] established promptable mask generation; SAM 3 adds concept prompts (noun phrases) with cgF1 54.1 on SA-Co/Gold and strong counting (CountBench MAE 0.12). Its authors note weak generalization to "fine-grained out-of-domain concepts in niche visual domains"—precisely the industrial regime we probe. No peer-reviewed SAM 3 crack-segmentation pixel IoU exists prior to this work.

Crack segmentation. Supervised SOTA on CrackSeg9k [8] reports *two-class* mIoU (background + crack): DeepLabV3+ 0.7599, HrSegNet-B48 80.32 [9], CrackMamba 81.75 [10]; because background IoU is ∼0.95, these are not comparable to crack-only IoU. SAM-adaptation work reports crack-only IoU: CrackSAM (LoRA) 0.6416 in-domain [11], SAC (LayerNorm-only tuning) 44.13 after training on OmniCrack30k [12]; OmniCrack30k [13] finds nnU-Net beats all specialized crack networks (clIoU_{4px} 64%). Our zero-shot 0.442 equals SAC's *fine-tuned* crack IoU.

Crack width measurement. RGB-only skeleton/orthogonal methods report r=0.962, RMSE 0.24 mm on synthetic cracks [14]; depth-assisted systems reach < 0.05 mm with a LiDAR camera [15]; laser-projection systems 0.02–0.57 mm [16]. krkCMd [7] releases 19,098 cross-crack brightness profiles with manual widths—to our knowledge the only public physical width GT—whose authors' deep model (DLM) attains 11.1 µm MAE. We position GaugeAnything as RGB-only, single-image, and promptable, between the RGB-only and sensor-assisted regimes.

Promptable measurement. Measure Anything [6] (SAM-2 + stereo) reports stem-diameter automation and a single "±10%" bottle validation, with no caliper MAE; agricultural stereo systems are similar. A quantitative, physically grounded promptable-metrology table has been empty; we contribute its first entries.

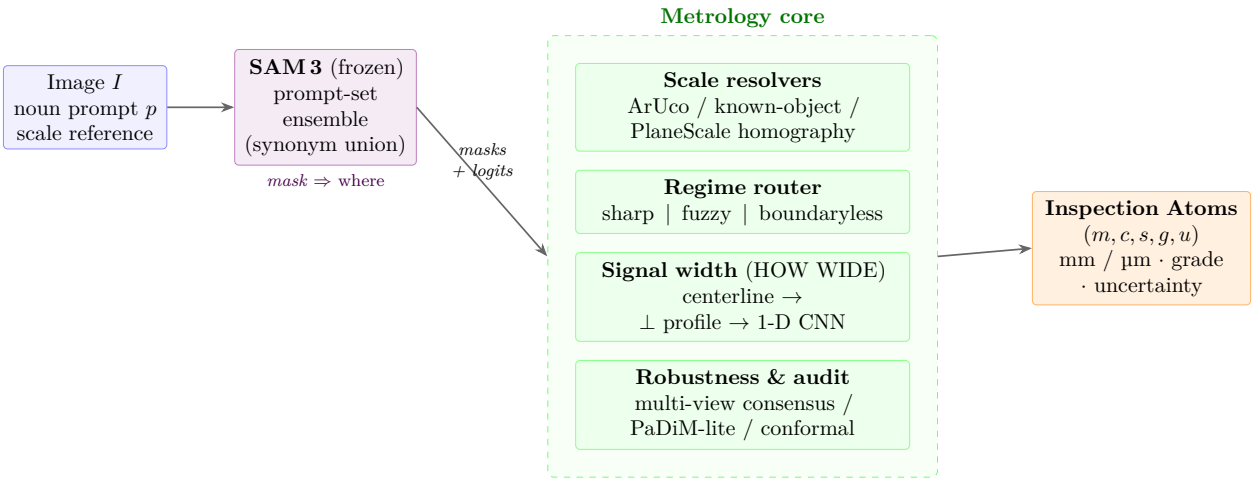

Figure 2: GaugeAnything pipeline. A frozen SAM 3 backbone (wrapped in a synonym prompt-set ensemble) turns an image and a noun prompt into instance masks and soft logits; the *metrology core* converts pixels to physical units. The central decomposition is **"mask for *where*, signal for *how wide*"**: the mask localizes the centerline, but width is regressed from the raw intensity *signal*, not mask geometry. Scale resolution (incl. tilt-robust PlaneScale), a boundary-regime router, multi-view consensus, anomaly modeling, and conformal intervals make the output an audited measurement: an *Inspection Atom* (m, c, s, g, u).

Counting and anomaly detection. FSC-147 SOTA: CountGD test MAE 6.75 [5], GeCo 7.91 [17]; supervised rebar counting reaches count-accuracy 86.27% [18]. Zero-shot text-prompted anomaly detection (AnomalyCLIP [19], pixel-AUROC 91.1 on MVTec) localizes but does not measure. On Magnetic-Tile defects [20], supervised mIoU reaches 86.5 [21]; per-class fray/uneven numbers are rarely tabulated.

3 Method

3.1 Task: Promptable Quantitative Inspection

Given an image I, a noun-phrase prompt p (e.g. "crack", "hex bolt"), and an optional scale reference (fiducial marker size or a known-dimension object class), GaugeAnything returns a set of *Inspection Atoms*: $a_i = (m_i, c_i, s_i, g_i, u_i)$—instance mask, class, size measurements in physical units, condition grade, and confidence/uncertainty. Counting is $|\{a_i\}|$; spacing is pairwise centroid distance in mm.

Pipeline overview (Fig. 2). The system is a frozen perception front-end followed by an explicit metrology back-end. (1) The prompt p is expanded to a synonym set and run through SAM 3, whose instances are unioned to a mask set $\{m_i\}$ with per-instance soft logits ℓ_i. (2) A *scale resolver* fixes a pixel-to-millimeter map from the scale reference. (3) Each instance is routed by boundary regime to the matching measurement operator—mask geometry, alpha matting, or illumination-residual modeling—and thin structures additionally invoke the *signal-based width* head. (4) Robustness and audit modules (multi-view consensus, anomaly modeling, conformal intervals) turn raw readings into audited Inspection Atoms. Every stage is deterministic and inspectable; only SAM 3 and four small task heads contain learned parameters, none of which touch the scale arithmetic.

3.2 Backbone and prompt robustness

We use SAM 3 frozen, via its image concept-segmentation interface. Because concept prompts are brittle—"fracture" and "pit" score 0.000 where "crack" and "hole" score 0.374/0.352—we wrap the backbone in a *prompt-set ensemble*: a curated synonym set per defect family, instances unioned with IoU-0.5 deduplication keeping the highest-scoring instance. The raw model also exposes per-query soft mask logits (range $[-67, 10]$ pre-sigmoid), which we exploit in §4.4 and for uncertainty.

3.3 Metrology core

Scale. Pixel-to-mm conversion comes from (i) ArUco/ChArUco fiducials; (ii) known-dimension objects (bolt-head across-flats, coin diameters); (iii) *PlaneScale*. A naive global factor $s_0 = D_{\text{ref}}/d_{\text{ref}}^{\text{px}}$ (true size over pixel size of the reference) is correct only fronto-parallel: under tilt it mis-scales points away from the reference. PlaneScale instead estimates the full plane-to-image homography H from the four fiducial corners $x_k \leftrightarrow X_k$ (known metric coordinates) by the DLT, and maps each pixel back to the metric plane, $\tilde{X} = H^{-1}\tilde{u}$ (homogeneous $\tilde{u} = (u, v, 1)^\top$). Measurements are then taken in plane coordinates; equivalently, the *local* scale at an instance centroid u is the linearization of H^{-1},

$$s(u) \;=\; \sqrt{\left|\det J_{H^{-1}}(u)\right|} \quad [\text{mm/px}], \qquad J_{H^{-1}}(u) = \frac{\partial\, \pi(H^{-1}\tilde{u})}{\partial u}, \tag{1}$$

with π the perspective divide. This per-pixel scale absorbs foreshortening: at 50° tilt, the naive s_0 errs 19.3% while (1) errs 0.7% (§4.11). For 3-D parts a single plane is only an approximation; the residual is quantified as a ceiling in §4.6 and reduced by multi-view consensus (§3.5). **Geometry.** Thin structures: skeletonization + Euclidean distance transform; width = 2·EDT at skeleton points, yielding a width profile (mean, P95, max). Blobs: equivalent diameter $\sqrt{4A/\pi}$. The thin/blob choice is automatic via elongation. **Regime router.** Defects differ in boundary physics: sharp (crack on smooth concrete), fuzzy (fray—frayed bristle-like boundaries), boundaryless (uneven/mura—a gradual field). A statistics-based router dispatches to binary masks, alpha matting (trained on directional synthetic compositing), or illumination-residual modeling (2-D polynomial fit + ISO-25178-style roughness), respectively. Soft outputs are measured by area = $\sum \alpha$ with uncertainty Var $\approx \sum \alpha(1-\alpha)$.

3.4 Signal-based width: mask for where, signal for how wide

For thin-structure width we *do not* read the mask boundary. A binary mask thresholds a gradual intensity transition into a hard edge, and the threshold's exact location—hence the measured width—moves with contrast, illumination, and the segmenter's operating point. We therefore use the mask only for *localization* and read width from the raw signal.

Localization. Tiled SAM 3 inference yields a crack mask; we keep the longest skeleton component (margin-band pixels removed) as the centerline. At each station t along the centerline we sample the image intensity along the unit normal n_t,

$$P_t(k) \;=\; I\Big(c_t + \tfrac{k}{K}\, w\, n_t\Big), \quad k = -\tfrac{K}{2}, \ldots, \tfrac{K}{2}, \tag{2}$$

giving a fixed-length profile of $K{+}1{=}501$ samples over a $\pm w$ window. The crack is a valley in P_t; we re-center the window on the valley minimum (snap-to-valley) so the model sees a canonical position, and z-normalize P_t so it sees *shape*, not absolute brightness—the property that makes the estimate transfer across imaging conditions (§4.8).

Regression head. A 1-D CNN (three Conv-BN-ReLU blocks, channels 16→32→32, kernels 7/5/3, two length-2 poolings, global average pool, linear; ∼0.2M parameters) maps P_t to physical width in micrometers. It is trained on krkCMd's 14,424 training profiles (CC BY 4.0, the one corpus pairing profiles with manual physical widths) under an L_1 loss with a *group* split (no station leaks between train and test).

Gate. A correlation-based confidence test rejects stations whose profile is not a trustworthy single-valley crack section (multiple cracks, texture, occlusion); rejected stations are reported as *not measurable* rather than guessed. We treat this gate as part of the instrument's contract and always report *coverage* (fraction measurable) alongside accuracy, so a low-coverage/high-accuracy regime is never mistaken for full coverage.

3.5 Robustness and audit modules

Four modules turn a raw reading into an audited one; each targets a failure mode surfaced by our own evaluation (§4).

Multi-view consensus. A single image fixes one depth plane, so Eq. (1) mis-scales 3-D parts whose extent leaves that plane, and image projection only ever *shortens* a tilted chord. When the same physical instance is seen in V views (e.g. a part imaged from many angles), let $\hat{d}_v$ be its per-view plane-scale dimension estimate. Because foreshortening is a one-sided (downward) bias, the robust consensus is the upper tail rather than the median:

$$\hat{d} \;=\; \frac{1}{|\mathcal{T}|}\sum_{v\in\mathcal{T}} \hat{d}_v, \qquad \mathcal{T} = \{\, v : \hat{d}_v \geq Q_{0.8}(\{\hat{d}_v\}) \,\}, \tag{3}$$

the trimmed mean of the top quintile (the most fronto-parallel, best-conditioned views); a high quantile $Q_{0.9}$ behaves similarly. This selects the views in which the measured axis is least foreshortened without any depth sensor.

Anomaly modeling for diffuse defects. Boundaryless/fuzzy defects (uneven, fray) have no edge for a mask to find, and we show hand-crafted residual maps stall near chance. We instead model the distribution of *defect-free* appearance: patch features f from a frozen ImageNet backbone (ResNet-18, layers 1–2) on normal tiles define a global Gaussian (μ, Σ), and a test patch is scored by its Mahalanobis distance

$$A(f) \;=\; \sqrt{(f-\mu)^\top \Sigma^{-1} (f-\mu)}, \tag{4}$$

a PaDiM/PatchCore-style detector requiring no defect labels. Anomalous regions score high; the continuous map is thresholded only for reporting.

Dense counting by density regression. Where instances touch and merge, detection under-counts structurally. A small density head (0.11 M params) regresses a map whose integral is the count; to counter the dense-image under-count we weight the count loss by object frequency, $\mathcal{L} = \sum_i \sqrt{g_i}^{\,\omega} \, |\hat{n}_i - n_i|$ (g_i the GT count, ω a deployment knob trading overall MAE for dense recovery, §4.12).

Conformal uncertainty. The size estimate is wrapped in a split/cross-conformal interval calibrated to 90% coverage. We deliberately use the *non-adaptive* variant: §4.5 shows that the efficient adaptive variants (normalized conformal, CQR) collapse on the worst held-out source, a concept shift that no feature-based difficulty signal detects, so we keep the interval whose coverage actually transfers.

Table 1: Evaluation corpora. GT type: *phys.* = physically measured units (µm/mm), *CAD* = pose-projected CAD millimeters, *mask*/*count* = annotation only. †research-only (upstream license unstated or mixed).

Dataset	Quantity / GT	Size	Holdout	License
CrackSeg9k	crack mask (13 src.)	∼9k img	source	research†
krkCMd	crack width, *phys.* µm	19,098 prof./4 ser.	group + cross-series	CC BY 4.0
M2 cache	mask-derived width	1,560 prof.	source	research†
T-LESS	part dims, *CAD* mm	20 scenes	per-view + multi-view	BOP
coins	diameter, known-object mm	22 photos	leave-one-out	self
ROI-1555	rebar *count*, polygons	1,260 lab.	count-stratified	research†
Magnetic-Tile	defect mask + normal tiles	6 classes	normal-fit / pixel	research†
SmartDoc15-CH1	document A4, *phys.* mm	video frames	per-frame	research
TUM RGB-D	checkerboard, *phys.* mm	398 frames	motion-gated	research
Aria Digital Twin	3-D dims, oracle pose	2 seq./229 obj.	oracle gate	research

4 Experiments

All numbers were produced on a single NVIDIA GB10 (aarch64); protocols enforce val/test separation (configuration selected on validation only), held-out test *sources*, multi-seed reporting where applicable, and crack-only metrics with empty-GT images scored separately. Datasets and licenses are catalogued in the repository (`paper/DATASETS.md`); deprecated pre-audit numbers are retained in the results log. All headline numbers are pinned to canonical result files and re-verified by a release gate (`benchmark/`); learned krkCMd width heads and SAM-localized end-to-end numbers carry ∼1 µm of GPU run-to-run variance, which the gate tolerances reflect.

4.1 Datasets, metrics, and protocols

Table 1 lists every evaluation corpus, the physical quantity it grounds, and its split. We span the axes that matter for field metrology: planar vs. 3-D, static vs. moving, mask GT vs. *physical* GT (manual micrometers, CAD millimeters, known-object diameters), and sharp vs. diffuse defects.

Metrics. Measurement uses relative error $|\hat{q} - q|/q$ (median, mean, and pass@5/10%) and, for width, MAE in µm; segmentation uses *crack-only* IoU with empty-GT images scored separately (a two-class mIoU is not comparable, §4.2); diffuse defects use pixel ROC-AUC of the continuous score against the GT mask (threshold-free); counting uses MAE and a dense-bin (GT ≥ 40) breakdown; uncertainty uses empirical coverage of the 90% interval and its median relative width. **Splits.** Selection is on validation only; test labels are never used for model or hyper-parameter choice. We hold out by the strictest unit available: *source* (CrackSeg9k, M2), *group*/station (krkCMd profiles), *cross-series* condition (krkCMd width×age, §4.8), *count-stratified* frames (ROI-1555), and *normal-only fit* with defect tiles unseen (Magnetic-Tile anomaly). **Gating.** Where localization or depth can fail (signal width, TUM, ADT), frames failing an a-priori gate are reported as not-measurable and *coverage* is always given, so accuracy is never inflated by silent rejection. **Reproducibility.** Each headline number is pinned to a canonical `experiments/results/*.json` and re-checked by the release gate (`benchmark/`); the 2026-06 audit that drove §4.6, §4.9, §4.8, and §4.12 is archived under `eval_runs/`.

4.2 Zero-shot promptable segmentation (Gauge-Bench)

On CrackSeg9k (crack-only IoU, empty-GT excluded, 3 seeds): frangi 0.115 ± 0.005, adaptive threshold 0.181 ± 0.006, **SAM 3 zero-shot** 0.442 ± 0.011 (2.44× the best classical), with a non-crack clean rate of 0.68 vs. 0.00 for the adaptive baseline. We emphasize the *metric trap*: supervised

Table 2: Calibration ladder for crack width on held-out sources (n=219). Lower is better. "Params" counts learned/selected quantities.

Method	Params	Rel. err. ↓	Bias
Raw SAM 3 mask width (θ=0.5)	0	0.730	+0.680
Global affine calibration	2	0.679	+0.633
M2 neural refiner (UNet)	1.9M	0.564	+0.503
Quantile-ratio calibration	5	0.480	+0.411
GaugeHead-Tiny (ExtraTrees, 19 statistics)	$\sim 10^5$	0.472	+0.456
Logit iso-level θ^*=0.7	1	0.493	+0.406
θ^* + quantile	**6**	**0.437**	+0.367

crack SOTA reports two-class mIoU (0.77–0.82), which is not comparable; in crack-only terms our zero-shot equals SAC's fine-tuned 44.13 [12]. Cross-domain, the same prompt-driven pipeline transfers from concrete to magnetic tile (crack 0.454; blowhole via "hole" 0.429 with automatic blob→diameter dispatch).

4.3 Segmentation rank ≠ measurement rank

On width fidelity (GT 11.3 px), the best-IoU method is not the best measure: adaptive (IoU 0.181) errs 43.5% relative, SAM 3 (IoU 0.442) errs 62.9%. This dissociation motivates everything that follows.

4.4 The calibration ladder

With three crack sources held out entirely (cfd, cracktree200, deepcrack; n=219), Table 2 traces width relative error. The 1.9M-parameter measurement-aware neural refiner (M2) improves raw SAM 3 but *loses* to a five-number quantile calibration fit on training sources; exploiting the exposed mask logits—selecting the binarization iso-level θ^*=0.7 on train/val—and stacking quantile calibration reaches 0.437 with zero learned parameters. Width bias is domain-dependent (train sources −0.17, held-out +0.68), so a single global correction cannot finish the job: the residual is a property of the mask, which is what §4.8 removes.

4.5 An owned measurement head and its per-source conformal audit

Two questions follow the ladder: can *any* learned component overtake the simple rungs, and can the instrument state its own uncertainty per domain? First, a tiny tabular head ("GaugeHead-Tiny": ExtraTrees over 19 mask/image statistics, family selected on train-source validation only, refit on train+val) becomes the first learned rung to pass quantile calibration on held-out sources (0.472 vs. 0.480)—though it still trails the logit-level scheme (0.437), and the worst source (CrackTree200, 0.720 rel. err.) is unchanged; we claim a path, not a victory. Second, we calibrate 90% intervals around this head four ways—absolute-residual and log-residual split conformal, difficulty-normalized conformal with a learned $\sigma(x)$, and CQR [22]—selecting by validation efficiency only. The non-adaptive log-residual interval covers every held-out source (0.97/1.00/0.98); its 5-fold cross-conformal variant keeps the 0.472 point error with per-source coverage 0.91/1.00/0.95. The honest negative: the *efficient* adaptive intervals collapse on the worst source (0.21 and 0.11 coverage at the nominal 0.90), and no feature-based difficulty signal we audited (learned $\sigma(x)$, ensemble variance, kNN feature distance; thresholds fixed on validation) flags that source—each fires more on the *easiest* source. CrackTree200 fails through a shifted width–label relationship at similar features (concept shift), not

through a shifted feature distribution, so exchangeability-based adaptivity [23] should not be trusted across inspection domains without exactly this kind of per-source audit. The deployed checkpoint therefore ships the non-adaptive interval, and we record the adaptive collapse as a finding.

4.6 Promptable part metrology on CAD ground truth (T-LESS)

BOP-format CAD models are metrically exact: projecting model vertices through the annotated pose yields, for the maximal projected chord, an exact mm distance. With *perfect* (ground-truth) visible masks and a single pose-depth plane scale, measurement attains median 2.83% (94% within 10%; worst-case 34.6% on depth-elongated objects)—a methodological ceiling for single-depth plane scaling on 3-D parts. Replacing perfect masks with SAM 3 zero-shot (concrete-noun prompts: "electrical component", "plastic part", "white object"): match rate 100%, mask IoU 0.94, and **measurement median** 2.5%—the segmentation step costs essentially nothing. The abstract prompt "industrial part" matches 0%: synonym collapse reappears in a third domain, reinforcing the prompt-ensemble design.

Breaking the single-depth ceiling with multi-view consensus. The worst-case 34.6% is foreshortening: a single depth plane mis-scales chord endpoints that lie off the object center, and projection only ever *shortens* the true chord. Since T-LESS scenes photograph a fixed arrangement from hundreds of viewpoints, the same physical instance is seen many times; aggregating its per-view plane-scale estimates with a foreshortening-aware consensus (trimmed mean of the top quintile, which selects the most fronto-parallel views) recovers the true caliper dimension. On the object-caliper task (124 instances, median 34 views each), single-view scaling gives mean 5.27% / worst 36.8% (reproducing the ceiling), while **multi-view consensus gives mean** 1.13% **/ median** 0.49% **/ worst** 11.5% **(**99% **within** 10%**)**. The depth-elongated parts that drove the worst case recover the most: obj. 24 8.57% → 0.82%, obj. 13 0.82% → 0.11%. Multiple views, not depth sensing, close the single-image gap.

4.7 Real-photo consistency (coins)

On 22 real photos (8–60 same-denomination euro coins each), leave-one-out known-object scaling (each coin measured using the others as reference) yields mean error 1.74%, 100% within 5%. This validates the segmentation→diameter chain on real imagery; the absolute marker chain is validated synthetically (ArUco 0.38%; e2e 5.6%).

4.8 Physical crack width: from profiles to the promptable chain

Profile level. On krkCMd's manual widths (group-split test 4,674 profiles), a deterministic valley rule with linear calibration attains 25.9 µm (5-fold 27.8 ± 2.5; leave-one-series worst 46.7), matching the authors' classical AED (26.5); their supervised DLM anchors at 11.1 µm. **Image level, mask geometry.** Reconstructing every profile's image position from the released ROI coordinates (correlation-verified, 0.95 gate), widths read from SAM 3 mask geometry err 144–186 µm even after θ^* transfer (which itself improves the default by ∼30%, showing the logit knob generalizes across datasets and into physical units). **Image level, signal-based (ours).** The 1-D profile CNN reaches 18.6 µm on the table test split; applied at *oracle* positions in the image, 26.2 µm; applied at SAM 3-localized, valley-snapped positions, **39.9 µm MAE / 23.2 µm median** on the 46–66% of stations passing the localization gate (failed stations: 186 µm, reported as not-measurable). A systematic deterministic-loop study (continuity gating, zoom re-acquisition, crack-likeness scoring, Viterbi path selection) then revealed that this coverage ceiling is largely an *annotation-matching artifact*: the scenes contain more than one genuine crack, and a single-path selector cannot know

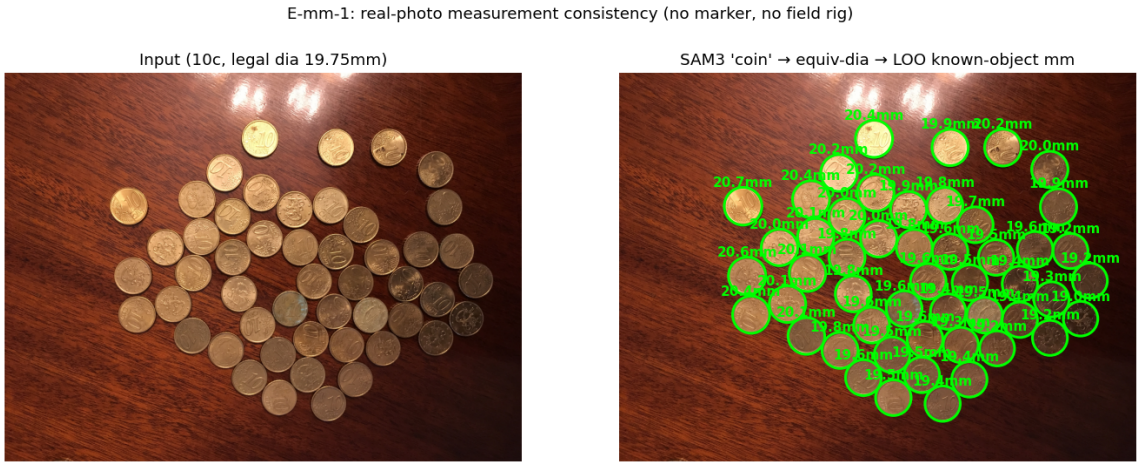

Figure 3: Real-photo known-object consistency (E-mm-1): each coin is measured using the other same-denomination coins as scale reference; mean leave-one-out error 1.74% over 22 scenes.

Table 3: Physical crack-width error on krkCMd (µm). "Gated" = stations passing the localization confidence gate (coverage 46–66%).

Level	Method	MAE
Profile (table)	authors' DLM (supervised)	11.1
Profile (table)	**ours: 1-D CNN**	**18.6**
Profile (table)	valley rule + linear cal. (5-fold)	27.8 ± 2.5
Image, oracle pos.	ours: 1-D CNN	26.2
Image, promptable (gated)	**ours: SAM 3 locate + 1-D CNN**	**39.9 (median 23.2)**
Image, promptable	SAM 3 mask geometry (θ^*, best)	144–186

which one the annotator chose. When the instrument reports *all* crack instances—each measured along its own centerline, as our Inspection Atoms formulation prescribes—the annotated crack is covered at 93% **recall** with 29.8 **µm MAE (**15.7 **µm median)** on the matched instance. Width information survives imaging; binary masks were the bottleneck; and the remaining open items are the 7% unrecovered stations and per-station confidence gating, not path selection. **Cross-source generalization.** krkCMd spans four conditions (crack width 0.23/0.28 mm × age 2/20 months) with independent manual GT. Training and calibrating on a *single* condition (3,936 profiles) and testing on the other three held-out conditions (unseen width *and* age), the signal representation transfers where mask thresholding does not: held-out macro MAE is 51.5 µm for a global-threshold mask width versus 24.0 **µm for our 1-D signal CNN** and 17.7 µm for the deterministic DLM rule—a 2–3× cross-condition advantage for the signal view. Only the hardest held-out condition (both width *and* age shifted) degrades the signal models to ∼32 µm, an honest domain-shift tail.

4.9 Boundary regimes: fuzzy and boundaryless defects

Binary segmentation collapses to chance on Magnetic-Tile fray/uneven (AUC ≈0.50). With val/test protocol: illumination-residual modeling recovers uneven to 0.669 test AUC (DRAEM-lite: 0.636); for fray, a matting head trained on *directional* synthetic compositing achieves real-fray preservation IoU 0.949 vs. guided-filter 0.860—after an honest failure: the blob-synthesis v1 won 20× synthetically yet transferred at 0.483. Synthesis realism, not architecture, was the fix.

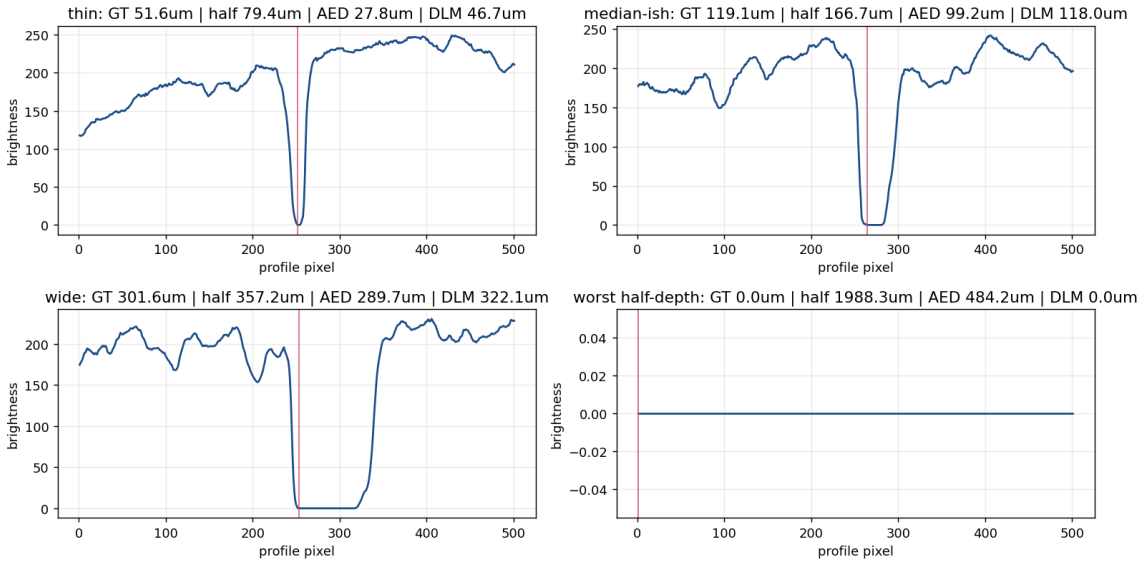

Figure 4: krkCMd cross-crack brightness profiles (501 samples, 3.97 µm/px) with manual width GT—the physical ground truth behind Table 3.

Hand-crafted maps hit a method-class ceiling; normal-distribution modeling breaks it. Eight hand-crafted continuous detectors (global polynomial residual, multi-scale band-pass, local-contrast normalization, robust-IRLS polynomials of order 1–3, Hessian ridge, and their fusion) all *fail to beat* the simple order-2 residual (uneven 0.609, fray 0.681): these defects are low-frequency, so band-pass removes the signal and contrast-normalization suppresses the defect's own texture. Treating the problem as MVTec-style anomaly detection instead—a global Gaussian over ImageNet patch features (ResNet-18 layers 1–2) fit on 250 defect-free tiles, scored by patch Mahalanobis distance—raises fray to 0.791 AUC (+11 pt) and, more importantly, generalizes across defect types the mask pipeline could not touch at all: **blowhole** 0.993**, crack** 0.944**, break** 0.939. Uneven, the broadest and lowest-contrast field, remains hard (0.645): a genuinely open case.

4.10 Dynamic scenes: handheld and egocentric metrology

Field capture is not a tripod, so we ask whether the metric signal survives camera motion. On TUM handheld RGB-D [24], after gating out depth-desynchronized and anisotropic frames, checkerboard-square measurement over 160 frames attains 1.06% **median** / 2.60% **p90** relative error, with every 0.1–1.0 m/s motion bin near 1%: gated handheld measurement is essentially static-grade. On Aria Digital Twin egocentric walkthroughs [25] (2 sequences, 480 frames, 229 objects), multiview RGB-D fusion under an *oracle* GT-pose/volume gate attains 8.7% **median** 3-D dimension error (9.1% in the 0.5 m/s+ bin), while an ROI-only control without the gate collapses to 316%. Motion does not destroy the metric signal; *localization* does—consistent with the static-scene finding. We report the ADT number explicitly as an upper bound: replacing the oracle gate with promptable masks, and reporting the gate failure rate, is the open problem.

4.11 Metrology rigor

Two silent measurement killers, quantified: camera tilt (50°: 19.3% → 0.7% via PlaneScale) and prompt synonym collapse (0.000 → 0.374/0.352 via the prompt-set ensemble). A 14-test metrology self-check (width ±10%, scale ±5%, e2e ±15% on synthetic GT) runs without any model and gates every release.

4.12 Honest negative: dense counting

Zero-shot rebar counting fails: best prompt MAE 13.1 (GT mean 22.5), acc@10% 5%, and six prompt variants all fail—a capability gap, not vocabulary (the same model detects 60 separated coins reliably). SAHI tiling recovers to MAE 7.4–8.9 but dense touching ends remain undercounted (e.g. GT 81 → 40). Counting here is a representation problem, not a prompt one: a small owned density head (0.11M parameters, ROI-1555, count-stratified holdout) brings MAE to **7.0**, below the zero-shot SAHI bar, and is excellent mid-range (GT 61 → 60.8). Up-weighting dense images in the count loss (by $\sqrt{\mathrm{GT}}$) exposes a clean operating-point trade-off between overall MAE and dense recovery: at weight 0.75 the dense undercount bias improves from −27.2 to −20.2 ($\mathrm{GT} \geq 40$) *while* overall MAE drops to 6.56; pushing to weight 1.5 cuts the dense bias further to −12.2 and dense MAE from 29.3 to **19.8**, at a small overall cost (7.2). It does not close the regime: a post-hoc monotonic calibration helps on validation but does not transfer to the few (n=16) dense test images, and near-empty images still over-count—a regression-to-mean floor whose root cause is the scarcity of densely labeled frames in ROI-1555, not the loss. Promptable detection sets the gate; an owned density regressor is the honest path past it, but the densest crowding remains data-limited.

5 Limitations

(1) *Residual localization gaps*: with multi-instance reporting the annotated crack is recovered at 93%; the remaining 7% of stations and the per-station confidence gate are open engineering items, and selecting a *specific* crack among several is delegated to point prompts by design. (2) *Field captures*: our physical-unit evidence comes from CAD ground truth, known-dimension objects, and scanner profiles; an ArUco+caliper field protocol (released with a printable board and capture checklist) awaits real deployments. (3) *Single image, single depth*: a single plane scale degrades on depth-elongated objects (single-view worst 36.8%); multi-view consensus mitigates this to worst 11.5% where repeated views exist, but the genuinely single-shot case still wants depth-aware measurement. (4) *Transfer of the profile CNN*: now validated *across krkCMd conditions* (held-out crack width and age, 24.0 µm macro MAE versus 51.5 for mask), but transfer to a second independent physical-GT crack corpus remains a data-moat gap. (5) Some task heads are trained on datasets with unstated licenses and are released for research use only, clearly marked. (6) *Uncertainty*: the conformal intervals that survive source shift are wide (median relative width 1.39, roughly ±70%); their coverage on the worst source is the product of conservatism over 19 samples, not of detection—flagging concept-shifted domains from the image evidence itself remains open. (7) *Dynamic scenes*: the ADT result is an oracle-gated upper bound; promptable masks have not yet replaced the gate.

6 Conclusion

GaugeAnything turns promptable segmentation into promptable *measurement*. The decisive move is a decomposition—masks for localization, raw signal for quantity—validated on physical ground truth at 23.2 µm median crack width and 2.5% industrial part dimensions, extended to moving cameras

(1.06% handheld; 8.7% egocentric oracle bound), and audited down to its own uncertainty (per-source conformal coverage, with the adaptive collapse reported): every number audited, every negative reported, and every artifact released: code and benchmarks (GitHub), task heads (HuggingFace), and a live project page (falcons-eyes.github.io/GaugeAnything).

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

[2] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[3] Meta AI. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[4] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *NeurIPS*, 2024.

[5] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. CountGD: Multi-modal open-world counting. In *NeurIPS*, 2024.

[6] Yongkyu Yu, Pravin Sivakumar, et al. Measure anything: Real-time, multi-stage vision-based dimensional measurement using segment anything. *arXiv preprint arXiv:2412.03472*, 2024.

[7] Jacek Jakubowski and Kamil Tomczak. Deposition of data for developing deep learning models to assess crack width and self-healing progress in concrete (krkCMd). *Scientific Data*, 2025. Zenodo doi:10.5281/zenodo.14568398, CC BY 4.0.

[8] Shreyas Kulkarni, Shreyas Singh, Dhananjay Balakrishnan, Siddharth Sharma, Saipraneeth Devunuri, and Sai Chowdeswara Rao Korlapati. CrackSeg9k: A collection and benchmark for crack segmentation datasets and frameworks. In *ECCV Workshops*, 2022.

[9] Yongshang Li, Ronggui Ma, Han Liu, and Gaoli Cheng. Real-time high-resolution neural network with semantic guidance for crack segmentation. *Automation in Construction*, 156: 105112, 2023.

[10] Shengyu Zuo et al. CrackMamba: A state space model for crack segmentation. *arXiv preprint arXiv:2410.19894*, 2024.

[11] Kang Ge, Chen Wang, Yutao Guo, Yansong Tang, Zhenzhong Hu, and Hongbing Chen. Fine-tuning vision foundation model for crack segmentation in civil infrastructures. *Structural Health Monitoring*, 2024. arXiv:2312.04233.

[12] Mohammadsadegh Rostami, Sumanth Iyer, Ebrahim Eslami, and Farnoud Fahimi. Segment any crack: Deep semantic segmentation adaptation for crack detection. *ASCE Journal of Computing in Civil Engineering*, 40(3), 2026. arXiv:2504.14138.

[13] Christian Benz and Volker Rodehorst. OmniCrack30k: A benchmark for crack segmentation and the reasonable effectiveness of transfer learning. In *CVPR Workshops*, 2024.

[14] Zhang et al. Crack width measurement based on skeleton-tangent orthogonal neighborhood shortest distance. *Engineering Structures*, 326:119519, 2025.

[15] Muhammad Hussain et al. Multi-modality cross-fusion for crack segmentation and quantification with RGB-D sensing. *Construction and Building Materials*, 487:141961, 2025.

[16] Lee et al. Crack width measurement with laser-beam metric reference. *Applied Sciences*, 13(8): 4981, 2023.

[17] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation. In *NeurIPS*, 2024.

[18] Wang et al. UAV-based rebar counting with shearing augmentation. *Scientific Reports*, 2025. PMC12480898.

[19] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

[20] Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36:85–96, 2020.

[21] Liu et al. CINFormer: Transformer network with multi-stage CNN feature injection for surface defect segmentation. *arXiv preprint arXiv:2309.12639*, 2023.

[22] Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *NeurIPS*, 2019.

[23] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[24] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012.

[25] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *ICCV*, 2023.