# OPENMORPHOQUERY: FROM VISUAL MORPHOLOGY PROMPTS TO TARGET-GROUNDED QUERIES FOR UNSEEN INDUSTRIAL DEFECT LOCALIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

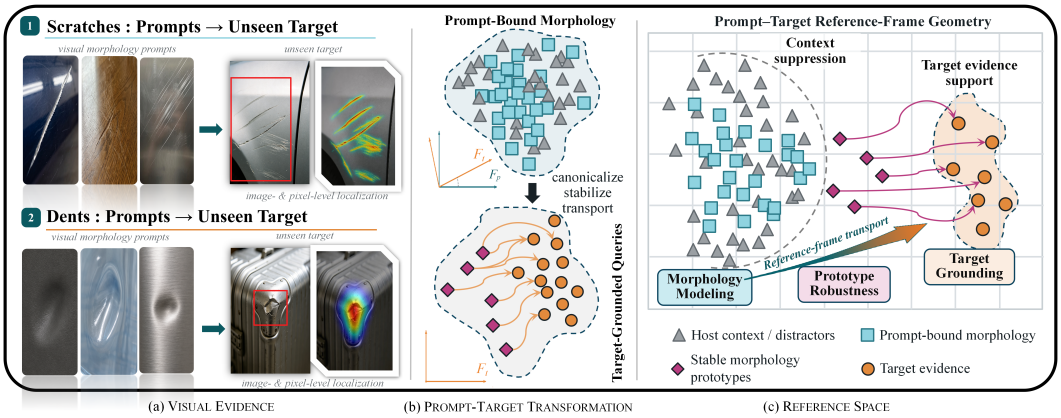
Under a category-disjoint open-set setting, localizing unseen industrial defects from scarce visual morphology prompts requires transferring fine-grained morphology revealed by exemplars to target images. Such transfer characterizes morphology-defined openness, whose central challenge stems from a reference-frame mismatch in morphological representation. Within prompt RoIs, discriminative morphology is encoded relative to material texture, product context, and annotation boundaries rather than as an independently transferable category attribute. Conventional aggregation preserves this sample-dependent anchoring, yielding prototypes that characterize the intended morphology but cannot establish reliable correspondence with its target-image manifestation. We propose OpenMorphoQuery (OMQ), a morphology-guided framework that addresses this mismatch through progressive reference-frame transformation. The Normalized Morphology Dictionary Encoder re-expresses context-bound RoI features relative to shared morphology bases through normalized residual coding, establishing a transferable morphology frame. Perturbation-Consistent Prototype Experts retain relations stable under appearance, geometry, and boundary perturbations, further stabilizing this frame against incidental variations. Prototype-Transport Query Initialization then transports the stabilized prototypes into a target-conditioned frame through optimal-transport alignment, yielding spatially grounded decoder queries. Experiments on a unified four-dataset benchmark of over 160,000 images establish state-of-the-art aggregate performance in multi-granular morphology grounding among visual-prompt methods. Ablations further quantify the incremental contribution of each component.

## 1 INTRODUCTION

How should an unseen industrial defect be specified? Natural-image recognition typically casts a novel category as a stable semantic concept grounded through language, but industrial defects expose the limits of this assumption. Defects sharing a coarse label such as scratch can vary substantially in structure, texture, scale, and boundary across metals, circuit boards, and leather, whereas visually similar abnormalities may denote different defects across products and manufacturing processes. Recent studies likewise reveal strong dependence on domain-specific appearance and context (Ni et al., 2026; Zhang et al., 2025). Industrial defects are therefore better understood as fine-grained morphological events embedded in product-specific backgrounds than as context-free semantic labels. We term the need to generalize through transferable morphology rather than fixed semantic vocabularies morphology-defined openness.

Vision-language detectors extend recognition beyond closed vocabularies by grounding names and phrases in images (Cheng et al., 2024; Li et al., 2022; Liu et al., 2024a; Minderer et al., 2022), while promptable segmentation models broaden target specification through spatial or multimodal prompts (Kirillov et al., 2023; Zou et al., 2023). These paradigms remain fragile for unseen industrial defects. Labels such as scratch, crack, pit, stain, and contamination collapse texture, boundary sharpness, orientation, scale, contrast, and product-dependent structure into coarse semantics, while points or boxes indicate locations without defining transferable defect morphology. Neither captures the fine-grained cues needed to distinguish defects from normal structures or neighboring categories.

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.




(a) Visual Evidence (b) Prompt-Target Transformation (c) Reference Space

Figure 1: Overview. Addressing the prompt–target reference-frame mismatch for unseen-defect localization. OpenMorphoQuery canonicalizes prompt-bound visual morphology and transports it into target-grounded queries supported by target evidence.

RoI-annotated visual exemplars expose structures obscured by ambiguous labels and offer a more faithful specification interface. Visual in-context prompting and open-set defect recognition demonstrate their potential beyond fixed categories (Kim et al., 2026; Li et al., 2024). Beyond prompt scarcity, transferring exemplar-specific morphology to target images presents a distinct difficulty, which we formulate as a morphology reference-frame mismatch. Within prompt RoIs, discriminative morphology is encoded relative to material texture, product context, local contrast, and annotation boundaries rather than as a transferable category attribute. Here, a morphology reference frame denotes the relational basis for representing defect features, not physical image coordinates. Conventional aggregation compresses these observations without changing their sample-dependent anchoring, yielding prototypes that characterize the intended morphology but cannot establish reliable correspondence with its target-image manifestation. Appearance, geometry, and boundary variations further induce prototype drift. This perspective motivates a progressive transformation from a sample-dependent prompt frame, through a stable shared morphology frame, into a target-conditioned frame.

To this end, we propose OpenMorphoQuery (OMQ), a morphology-guided framework that progressively resolves this mismatch by transforming sparse prompt observations into morphology-aware, perturbation-consistent, and target-grounded decoder queries, as illustrated in Figure 1. A shared encoder extracts target-image and prompt-RoI tokens. The Normalized Morphology Dictionary Encoder (NMDE) canonicalizes prompt morphology through normalized residual coding relative to shared morphology bases, producing compact prototypes in a transferable frame. Perturbation-Consistent Prototype Experts (PCPE) preserve relations stable under appearance, geometry, and boundary perturbations, stabilizing this frame against incidental variations. Prototype-Transport Query Initialization (PTQI) aligns the stabilized prototypes with target tokens via optimal transport and transports them into a target-conditioned frame to yield spatially grounded queries. A query-based decoder then predicts instance extents and fine-grained boundaries. Experiments on a category-disjoint benchmark of over 160,000 images across four public industrial datasets establish state-of-the-art performance among evaluated visual-prompt methods in both spatial localization and fine-grained boundary delineation. Our contributions are summarized below:

- We revisit unseen industrial defect localization through morphology-defined openness and characterize its central transfer barrier as a reference-frame mismatch in morphological representation.
- We propose OpenMorphoQuery, whose NMDE, PCPE, and PTQI progressively canonicalize prompt morphology, stabilize its shared frame, and transport it into a target-conditioned frame to initialize spatially grounded queries.
- Experiments on a unified large-scale industrial benchmark demonstrate state-of-the-art multi-granular morphology grounding, with ablations quantifying the incremental contribution of each component.

## 2 Related Work

**Category-Bound and Anomaly-Based Industrial Inspection.** CNN- and Transformer-based detectors, often coupled with multi-scale representations, have been applied to metal surfaces (Zhong et al., 2024), fabrics (Liu et al., 2020), glass (Chen et al., 2023), and component inspection (Zhang et al., 2024). Advances in texture representation (Zhou et al., 2024), contextual modeling (Shen et al., 2024), and attention mechanisms (Drir et al., 2026) improve the localization of small and irregular defects in complex backgrounds. Data augmentation, generative synthesis, few-shot learning (Shi et al., 2023), and self-supervised learning (Yildizli et al., 2026) further alleviate data scarcity and annotation costs. In parallel, anomaly detection explores reconstruction, anomaly generation, prompt learning, structural refinement, and defect synthesis (Wu et al., 2026; Tong et al., 2026; Lee et al., 2026; Jiang et al., 2026; Li et al., 2026a; Song et al., 2025). Despite substantial progress, these methods generally assume predefined defect categories or model deviations from normality, rather than localizing unseen categories specified by defect-positive exemplars.

**Open Industrial Defect Localization from Semantics to Morphology.** Open-vocabulary detectors—GLIP (Li et al., 2022), OWL-ViT (Minderer et al., 2022), Grounding DINO (Liu et al., 2024a), and YOLO-World (Cheng et al., 2024)—align visual regions with language. Industrial extensions SFR-Net (Chen et al., 2026) and RTVP (Zhang et al., 2025), together with IMDD-1M (Ni et al., 2026) and MAU-GPT (Wang et al., 2026), broaden this paradigm, but language underrepresents fine-grained, product-dependent morphology. Visual-prompt methods such as T-Rex2 (Jiang et al., 2024), DINOv (Li et al., 2024), and UniSpector (Kim et al., 2026) specify novel targets through visual exemplars, exposing fine-grained appearances beyond class names. However, they do not explicitly re-express context-entangled prompt morphology in a shared frame and transport it into target-grounded queries. OpenMorphoQuery addresses this gap through normalized morphology coding, perturbation-consistent calibration, and target-conditioned prototype transport. FastRef (Li et al., 2026b) applies query statistics and optimal transport (Cuturi, 2013) to normal prototypes, but addresses normality-based anomaly detection rather than exemplar-specified unseen categories. Thus, morphology re-expression, perturbation-consistent stabilization, and target-conditioned transport remain underexplored.

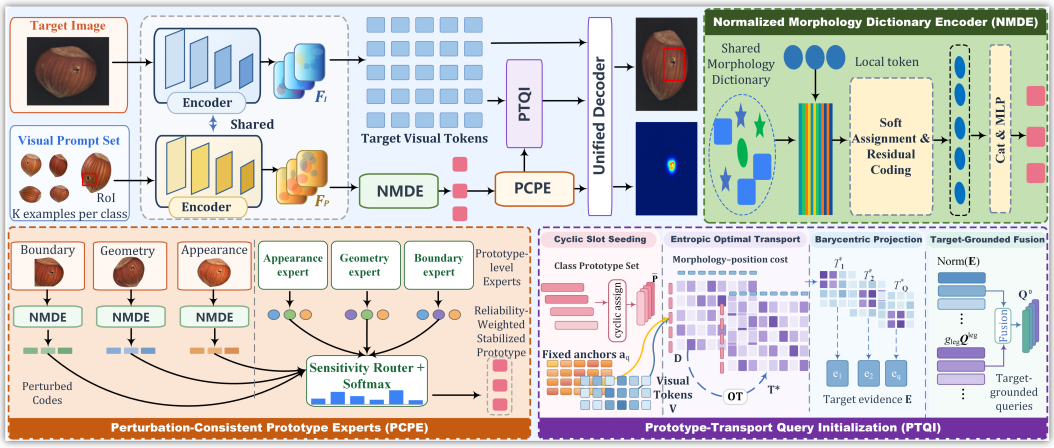



Figure 2: The OpenMorphoQuery pipeline. Given a target image and visual prompts, NMDE re-expresses support-bound morphology in a shared reference frame (Sec. 3.2), PCPE stabilizes its prototypes against perturbations (Sec. 3.3), and PTQI transports them into the target-image frame to initialize grounded queries (Sec. 3.4).

## 3 Methodology

### 3.1 Problem Formulation and Overview

We study unseen industrial defect localization with disjoint training and test categories, $\mathcal{C}_{\text{tr}} \cap \mathcal{C}_{\text{te}} = \varnothing$. Given a target image $I^q$ and prompt sets $\{\mathcal{P}_c\}_{c=1}^{C}$, where $\mathcal{P}_c = \{(I^p_{c,k}, R^p_{c,k})\}_{k=1}^{K}$ for $C$ unseen

categories, the model jointly predicts box- and pixel-level localizations for instances matching the prompted morphologies. Prompt RoIs entangle defect structure with product texture, context, and annotation boundaries, yielding sample-dependent prototypes without target grounding. A shared encoder and RoIAlign extract target tokens $\mathbf{V} \in \mathbb{R}^{N_t \times D}$ and a set of prompt RoI token matrices $\{\mathbf{X}_{c,k} \in \mathbb{R}^{N_p \times D}\}_{k=1}^{K}$. NMDE independently re-expresses each exemplar in a shared morphology frame, PCPE calibrates and aggregates the resulting exemplar prototypes, and PTQI couples the category representation with target-image tokens to initialize decoder queries $\mathbf{Q}^0$.

### 3.2 Normalized Morphology Dictionary Encoder

Visual morphology prompts reveal category-defining defect structures, but these structures are encoded relative to prompt-specific product textures, local contexts, and RoI observations, placing different exemplars in mismatched, sample-dependent reference frames. Directly pooling such features therefore mixes transferable morphology with incidental prompt context. NMDE re-expresses each prompt as residual deviations from shared morphology bases, transforming heterogeneous prompt observations into a common morphology reference frame. For category $c$, the $k$-th prompt RoI yields the local-token matrix $\mathbf{X}_{c,k} = [\mathbf{x}_{c,k,n}]_{n=1}^{N_p} \in \mathbb{R}^{N_p \times D}$, $k = 1, \ldots, K$. We then introduce a learnable morphology dictionary $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_M]^\top \in \mathbb{R}^{M \times D}$, where each basis captures a category-agnostic morphology pattern. Each prompt token is softly assigned to the shared bases according to normalized similarity:

$$A_{c,k,n,m} = \frac{\exp\big(\langle \bar{\mathbf{x}}_{c,k,n}, \bar{\mathbf{b}}_m \rangle / \tau\big)}{\sum_{j=1}^{M} \exp\big(\langle \bar{\mathbf{x}}_{c,k,n}, \bar{\mathbf{b}}_j \rangle / \tau\big)}. \tag{1}$$

Here, $\bar{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$ and $\bar{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|_2$ are normalized features, and $\tau$ is a temperature parameter. NMDE then aggregates the normalized residuals relative to each basis:

$$\mathbf{z}_{c,k,m} = \frac{\sum_{n=1}^{N_p} A_{c,k,n,m}(\mathbf{x}_{c,k,n} - \bar{\mathbf{b}}_m)}{\sum_{n=1}^{N_p} A_{c,k,n,m} + \epsilon}, \qquad m = 1, \ldots, M. \tag{2}$$

Each $\mathbf{z}_{c,k,m} \in \mathbb{R}^D$ describes the category-specific deviation of the $k$-th prompt from the $m$-th shared morphology basis. The dictionary-wise residuals are concatenated and projected through a lightweight bottleneck as $\mathbf{z}_{c,k}^{\mathrm{cat}} = [\mathbf{z}_{c,k,1} \oplus \cdots \oplus \mathbf{z}_{c,k,M}] \in \mathbb{R}^{MD}$ and $\mathbf{p}_{c,k} = \mathrm{Norm}\Big(\mathrm{LN}\Big(\mathbf{W}_\uparrow \phi(\mathbf{W}_\downarrow \mathbf{z}_{c,k}^{\mathrm{cat}}) + \mathbf{b}_\uparrow\Big)\Big)$. The resulting exemplar codes are stacked as $\mathbf{P}_c = [\mathbf{p}_{c,k}]_{k=1}^{K} \in \mathbb{R}^{K \times D}$, where $\phi(\cdot)$ denotes the activation function, $\mathrm{Norm}(\cdot)$ denotes $\ell_2$ normalization, LN denotes layer normalization, and $\oplus$ denotes concatenation. The parameters satisfy $\mathbf{b}_\uparrow \in \mathbb{R}^D$, $\mathbf{W}_\downarrow \in \mathbb{R}^{d_h \times MD}$, $\mathbf{W}_\uparrow \in \mathbb{R}^{D \times d_h}$, and $d_h = 64$. Rather than averaging prompt features, NMDE independently encodes each exemplar relative to a shared morphology vocabulary. The resulting prototype set $\mathbf{P}_c$ is subsequently calibrated by PCPE.

### 3.3 Perturbation-Consistent Prototype Experts

Although NMDE re-expresses prompt features in a shared morphology reference frame, prototypes estimated from few exemplars may remain sensitive to local appearance, geometric configuration, and RoI boundaries. We therefore introduce Perturbation-Consistent Prototype Experts (PCPE) to calibrate these uncertainties across controlled perturbation views. PCPE canonicalizes the $7 \times 7$ RoIAlign feature map $\bar{\mathbf{G}}_{c,k}$ as $\mathbf{G}_{c,k} = \mathrm{Pool}_{4\times4}(\bar{\mathbf{G}}_{c,k})$ and tokenizes it as $\mathbf{X}_{c,k} = \mathrm{Tok}(\mathbf{G}_{c,k}) \in \mathbb{R}^{16 \times D}$. It then constructs appearance, geometry, and boundary perturbations before NMDE encoding:

$$\mathcal{T}_{\mathrm{app}}(\mathbf{X}_{c,k}) = (1 - \gamma_{\mathrm{app}})\mathbf{X}_{c,k} + \gamma_{\mathrm{app}} \mathbf{1}_{16} \boldsymbol{\mu}_{c,k}^\top, \mathcal{T}_{\mathrm{bnd}}(\mathbf{X}_{c,k}) = \mathrm{Tok}\big[\mathbf{G}_{c,k} \odot \big(\mathbf{1} - \gamma_{\mathrm{bnd}} \mathbf{M}_{\mathrm{edge}}\big)\big]$$
$$\mathcal{T}_{\mathrm{geo}}(\mathbf{X}_{c,k}) = \mathrm{Tok}\big[(1 - \gamma_{\mathrm{geo}})\mathbf{G}_{c,k} + \gamma_{\mathrm{geo}} \,\mathrm{Rot}_{90^\circ}\big(\mathrm{AvgPool}_{3\times3}(\mathbf{G}_{c,k})\big)\big], \mathbf{p}_{c,k}^{e} = f_{\mathrm{NMDE}}(\mathcal{T}_e(\mathbf{X}_{c,k})), e \in \mathcal{E}. \tag{3}$$

Here, $\boldsymbol{\mu}_{c,k}$ is the token mean, $\mathbf{1}_{16}$ denotes broadcasting, $\mathrm{Tok}(\cdot)$ retokenizes feature maps, and $\mathcal{E} = \{\mathrm{app}, \mathrm{geo}, \mathrm{bnd}\}$. The coefficients $\gamma_{\mathrm{app}}, \gamma_{\mathrm{geo}}, \gamma_{\mathrm{bnd}}$ control the corresponding perturbation strengths. $\mathbf{M}_{\mathrm{edge}} \in \{0, 1\}^{4\times4}$ equals one at the outermost RoI positions and zero elsewhere. A shared NMDE encodes all views in the same morphology frame, with the original prototype $\mathbf{p}_{c,k} = f_{\mathrm{NMDE}}(\mathbf{X}_{c,k})$. PCPE measures perturbation sensitivity as $d_{c,k,e} = 1 - \cos(\mathbf{p}_{c,k}, \mathbf{p}_{c,k}^{e})$ and predicts $\boldsymbol{\alpha}_{c,k} = \mathrm{Softmax}(R([d_{c,k,e}]_{e\in\mathcal{E}}))$, where $R(\cdot)$ is a lightweight router and $\alpha_{c,k,e}$ is

normalized over $\mathcal{E}$. Each perturbation expert $E_e : \mathbb{R}^D \to \mathbb{R}^D$ is implemented as a bottleneck MLP and provides a residual correction:

$$\mathbf{p}^r_{c,k} = \mathrm{Norm}\left(\mathbf{p}_{c,k} + \sum_{e\in\mathcal{E}} \alpha_{c,k,e} E_e(\mathbf{p}_{c,k})\right),\ \mathbf{p}^{r,e}_{c,k} = \mathrm{Norm}\left(\mathbf{p}^e_{c,k} + \sum_{e'\in\mathcal{E}} \alpha_{c,k,e'} E_{e'}(\mathbf{p}^e_{c,k})\right). \quad (4)$$

The residual connection preserves the morphology abstracted by NMDE, while the sensitivity-conditioned mixture adapts prototype calibration to the perturbation profile of each exemplar. PCPE further estimates exemplar reliability from its average agreement across perturbations and uses it to aggregate the $K$ calibrated prototypes:

$$r_{c,k} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \cos(\mathbf{p}_{c,k}, \mathbf{p}^e_{c,k}), \quad b_{c,k} = \frac{\exp(r_{c,k}/\tau_r)}{\sum_{k'=1}^K \exp(r_{c,k'}/\tau_r)}, \quad \widehat{\mathbf{p}}^r_c = \mathrm{Norm}\left(\sum_{k=1}^K b_{c,k}\mathbf{p}^r_{c,k}\right). \quad (5)$$

The aggregated representation is adaptively fused with $\mathbf{p}^{\mathrm{leg}}_c$, the category prototype produced by the original visual-prompt pathway from the same support RoIs:

$$\bar{r}_c = \sum_{k=1}^K b_{c,k} r_{c,k}, \quad g_c = \sigma(g_0 + g_1(1-\bar{r}_c)), \quad \bar{\mathbf{p}}^r_c = \mathrm{Norm}\big(\widehat{\mathbf{p}}^r_c + g_c\,\mathrm{Norm}(\mathbf{p}^{\mathrm{leg}}_c)\big). \quad (6)$$

Here, $g_0$ and $g_1$ are learnable scalars, making the contribution of the legacy prompt pathway reliability-adaptive. Finally, we impose consistency between the calibrated original and perturbed prototypes:

$$\mathcal{L}_{\mathrm{pc}} = \frac{1}{|\mathcal{I}|\,|\mathcal{E}|}\sum_{(c,k)\in\mathcal{I}}\sum_{e\in\mathcal{E}}\Big[1 - \cos(\mathbf{p}^r_{c,k}, \mathbf{p}^{r,e}_{c,k})\Big], \quad (7)$$

where $\mathcal{I}$ denotes the prompt instances in the training batch. This objective penalizes perturbation-induced prototype deviations after expert calibration. The resulting category prototypes are subsequently coupled with target-image tokens by PTQI to initialize target-conditioned decoder queries.

### 3.4 Prototype-Transport Query Initialization

The calibrated prototypes capture category-discriminative morphology but remain spatially ungrounded with respect to the target image. Directly replicating them across decoder slots preserves content symmetry, which can cause multiple queries to concentrate on the same salient host structure rather than complementary defect evidence. PTQI addresses these limitations by grounding the prototypes in target-image evidence under a global budget of $Q$ localization queries. Within this budget, the $C$ category prototypes cyclically initialize the query contents via $c(q) = 1 + ((q-1) \bmod C)$ and $\widetilde{\mathbf{p}}_q = \bar{\mathbf{p}}^r_{c(q)}$, for $q = 1, \ldots, Q$. Queries initialized from the same prototype remain differentiated by their fixed spatial anchors and distinct legacy decoder queries.

Let $\mathbf{V} = [\mathbf{v}_j]_{j=1}^{N_t}$ be the target tokens and $\bar{\mathbf{v}}_j = \mathrm{Norm}(\mathbf{v}_j)$. Using the fixed query anchor $\mathbf{a}_q$ and normalized target position $\mathbf{z}_j$, PTQI defines the morphology–position transport cost as:

$$D_{qj} = -\widetilde{\mathbf{p}}_q^\top \bar{\mathbf{v}}_j + \lambda_{\mathrm{pos}} \left\|\mathbf{a}_q - \mathbf{z}_j\right\|_2^2, \quad (8)$$

where $\lambda_{\mathrm{pos}}$ balances morphology similarity and spatial compatibility. PTQI measures target relevance by $s_j = \max_c (\bar{\mathbf{p}}^r_c)^\top \bar{\mathbf{v}}_j$, and uses the uniform query marginal $\pi_q = 1/Q$ and content-weighted target marginal $\boldsymbol{\nu} = \mathrm{Softmax}(\mathbf{s}/\tau_{\mathrm{mass}})$. The entropy-regularized transport plan is:

$$\mathbf{T}^\star = \underset{\mathbf{T}\in\mathcal{U}(\boldsymbol{\pi},\boldsymbol{\nu})}{\arg\min} \langle \mathbf{D}, \mathbf{T}\rangle + \varepsilon_{\mathrm{ot}} D_{\mathrm{KL}}\big(\mathbf{T} \,\|\, \boldsymbol{\pi}\boldsymbol{\nu}^\top\big), \quad (9)$$

where $\mathcal{U}(\boldsymbol{\pi}, \boldsymbol{\nu})$ contains non-negative transport matrices with row and column sums $\boldsymbol{\pi}$ and $\boldsymbol{\nu}$. These constraints coordinate query assignments while concentrating target mass on category-relevant tokens.

The transport plan is computed through differentiable Sinkhorn iterations with element-wise division:

$$\mathbf{G} = \exp(-\mathbf{D}/\varepsilon_{\mathrm{ot}}), \quad \boldsymbol{\xi}^{(\ell+1)} = \frac{\boldsymbol{\pi}}{\mathbf{G}\boldsymbol{\zeta}^{(\ell)}}, \quad \boldsymbol{\zeta}^{(\ell+1)} = \frac{\boldsymbol{\nu}}{\mathbf{G}^\top\boldsymbol{\xi}^{(\ell+1)}}, \quad \mathbf{T}^\star \approx \mathrm{Diag}(\boldsymbol{\xi})\mathbf{G}\,\mathrm{Diag}(\boldsymbol{\zeta}). \quad (10)$$

For barycentric aggregation, let $Z_q = \sum_{j=1}^{N_t} T^\star_{qj} + \epsilon$ and $\mathbf{e}_q = Z_q^{-1}\sum_{j=1}^{N_t} T^\star_{qj}\mathbf{v}_j$. PTQI then initializes $\mathbf{q}^0_q = \mathrm{LN}(\mathrm{Norm}(\mathbf{e}_q) + g_{\mathrm{leg}}\mathbf{q}^{\mathrm{leg}}_q)$, where $g_{\mathrm{leg}} = \sigma(\beta_{\mathrm{leg}})$, $\epsilon$ is a numerical stabilizer, and $\beta_{\mathrm{leg}}$ is learnable. Stacking all slots gives $\mathbf{Q}^0 \in \mathbb{R}^{Q\times D}$. Implementation values are reported in Table A2. The additional $C$ category-content tokens are appended separately and are not counted in the $Q$ localization queries. During decoding, every localization query is scored against all $C$ category-content tokens, forming a shared query pool for multi-category prediction.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets and Category-Disjoint Split.** Industrial defects are fine-grained morphological events embedded in specific products, materials, and manufacturing contexts rather than concepts defined solely by stable semantic labels. To evaluate generalization beyond the relatively concentrated distributions of individual datasets, we integrate four public datasets—**GC10-DET** (Lv et al., 2020), **Magnetic Tile** (Huang et al., 2018), **MVTec AD** (Bergmann et al., 2019), and **Real-IAD** (Wang et al., 2024)—into a unified benchmark. We adopt a strict category-disjoint split comprising **147** seen training categories and **52** unseen test categories, with $\mathcal{C}_{\mathrm{tr}} \cap \mathcal{C}_{\mathrm{te}} = \varnothing$. The benchmark has **38,583** training and **121,447** test images; **88.0%** of test images are defect-free. This cross-dataset, defect-sparse search space requires localizing and delineating unseen defects amid extensive negatives, providing a stringent and practically relevant evaluation of morphology-defined openness. Per-dataset statistics and split details are provided in Appendix A.2.

**Evaluation Metrics and Implementation Details.** We evaluate multi-granular morphology grounding at the *box-* and *mask-level* using standard COCO protocols. OpenMorphoQuery employs a Swin-T encoder initialized from the DINOv pretrained checkpoint and a MaskDINO-style (Li et al., 2023) decoder. NMDE, PCPE, and PTQI transform prompt RoI tokens into target-grounded queries. Models are trained at $640 \times 640$ resolution for 50 epochs using AdamW, with a learning rate of $2 \times 10^{-5}$ and a total batch size of 32, on 8 GPUs. We train on 147 seen categories and evaluate on 52 disjoint unseen categories without using their training annotations. Full implementation and evaluation details are provided in Appendix A.3.

Table 1: Box/mask localization performance (%) on the unified morphology-defined open-set benchmark. GDINO, YLE, and YW denote Grounding DINO, YOLOE, and YOLO-World; S/M/L denote model scales. Prompts: *Text*, category names; *Vis.*, visual prompts; *Sp.*, points, boxes, or scribbles; *Ref. Mas.*, reference images with binary masks; *Ref. Ann.*, annotated reference images; and *Vis. Morph.*, morphology-focused exemplars from annotated RoIs. † denotes five fixed, distinct visual exemplars per category. Dashes indicate unavailable mask results. Best and second-best results are bold and underlined.

| | | **Average Precision** | | | **AP by Scale** | | | **Average Recall** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Methods** | **Prompt** | AP | $AP_{75}$ | $AP_{50}$ | $AP_S$ | $AP_M$ | $AP_L$ | $AR_1$ | $AR_{10}$ | $AR_{100}$ |
| GDINO (ECCV'24) | *Text* | 16.3/– | 18.2/– | 24.5/– | 14.8/– | 19.4/– | 17.0/– | 36.2/– | 53.5/– | <u>59.2</u>/– |
| YLE-S (ICCV'25) | *Text/Vis.* | 12.6/6.7 | 13.3/5.2 | 20.8/15.3 | 10.9/2.9 | 13.2/7.8 | 12.6/10.2 | 23.7/14.2 | 28.2/16.3 | 28.4/16.4 |
| YLE-M (ICCV'25) | *Text/Vis.* | 16.2/8.5 | 17.5/6.5 | 25.7/19.5 | 15.9/4.2 | 16.7/10.1 | 16.2/13.6 | 26.0/15.6 | 30.0/17.0 | 30.0/17.1 |
| YLE-L (ICCV'25) | *Text/Vis.* | 16.7/8.6 | 18.2/6.4 | 26.6/19.7 | 16.3/4.2 | 18.1/10.9 | 18.3/14.9 | 25.7/15.2 | 29.5/16.7 | 29.5/16.7 |
| YW-S (CVPR'24) | *Text* | 11.5/– | 12.7/– | 17.8/– | 10.6/– | 13.4/– | 12.9/– | 25.2/– | 28.8/– | 29.4/– |
| YW-M (CVPR'24) | *Text* | 17.8/– | 19.5/– | 27.5/– | 17.4/– | 19.8/– | 17.9/– | 32.6/– | 38.8/– | 38.8/– |
| GLEE (CVPR'24) | *Text* | 10.0/7.7 | 11.0/7.7 | 16.1/14.4 | 9.7/5.6 | 11.6/10.0 | 14.2/15.2 | 30.1/22.4 | 39.8/28.7 | 40.9/29.3 |
| SEEM (NeurIPS'23) | *Text/Sp.* | 3.7/3.2 | 3.0/2.3 | 8.0/7.9 | 3.1/1.4 | 5.5/4.2 | 3.0/6.9 | 12.8/10.8 | 19.2/14.7 | 21.6/15.8 |
| SINE† (NeurIPS'24) | *Ref. Mas.* | 0.2/0.1 | 0.2/0.1 | 0.5/0.4 | 0.1/0.0 | 0.4/0.2 | 0.2/0.2 | 17.2/15.4 | 19.6/17.3 | 19.6/17.3 |
| UniSpector† (CVPR'26) | *Ref. Ann.* | 3.5/2.8 | 3.7/2.9 | 5.6/4.9 | 3.1/2.1 | 4.5/4.0 | 5.7/4.7 | 35.4/27.1 | 44.0/32.4 | 44.2/32.6 |
| DINOv† (CVPR'24) | *Ref. Ann.* | <u>22.0</u>/<u>16.2</u> | <u>22.6</u>/<u>15.1</u> | <u>36.3</u>/<u>31.5</u> | <u>18.1</u>/<u>12.2</u> | <u>20.5</u>/<u>16.7</u> | <u>21.2</u>/<u>18.1</u> | <u>40.3</u>/<u>31.3</u> | <u>53.7</u>/<u>40.7</u> | 56.9/43.0 |
| OpenMorphoQuery† | *Vis. Morph.* | **28.5/21.5** | **31.3/21.6** | **43.6/39.1** | **21.2/13.8** | **29.2/23.0** | **26.6/23.3** | **44.5/34.7** | **56.2/43.6** | **59.4/45.2** |

## 4.2 MAIN RESULTS

**Benchmarking Morphology-Defined Openness.** We jointly evaluate models trained on 147 seen categories against 52 unseen categories pooled from four datasets. By excluding train–test category overlap while retaining variations in morphology, products, materials, backgrounds, and imaging conditions, this protocol evaluates novel-category generalization across heterogeneous industrial contexts and is more demanding than isolated dataset-wise testing. Table 1 compares OpenMorphoQuery with representative methods, including GDINO (Liu et al., 2024a), YOLOE (Wang et al., 2025), YOLO-World (Cheng et al., 2024), SEEM (Zou et al., 2023), SINE (Liu et al., 2024b), GLEE (Wu et al., 2024), UniSpector (Kim et al., 2026), and DINOv (Li et al., 2024). Details of our paper-guided UniSpector reimplementation are provided in Appendix A.3. SINE, UniSpector, DINOv, and OpenMorphoQuery use the same five fixed, mutually distinct exemplars. UniSpector retains **44.2** $AR_{100}$

but only **3.5** AP in this defect-sparse search space, revealing a pronounced recall–precision gap under extensive negative exposure. OpenMorphoQuery ranks first in all 18 entries, attaining box/mask AP, $AP_{50}$, and $AP_{75}$ of **28.5%/21.5%**, **43.6%/39.1%**, and **31.3%/21.6%**, and surpassing DINOv by **+6.5/+5.3**, **+7.3/+7.6**, and **+8.7/+6.5** points, respectively. The pronounced $AP_{75}$ gains indicate more precise morphology-to-target grounding under strict overlap.

**Source-Stratified Generalization across Industrial Contexts.** We further evaluate compared methods separately on the category-disjoint splits of GC10-DET, Magnetic Tile, MVTec AD, and Real-IAD. Together, these source-specific subsets cover diverse products, materials, defect morphologies, manufacturing settings, and acquisition conditions, enabling a finer-grained assessment across heterogeneous industrial contexts. This analysis instead examines whether the performance gains persist across different industrial contexts represented within each source. As shown in Table 2, OpenMorphoQuery achieves the highest box/mask AP on Magnetic Tile, MVTec AD, and Real-IAD, outperforming DINOv by **+1.8/+2.0**, **+7.6/+8.2**, and **+7.5/+4.6** points, respectively. These consistent improvements across three substantially different sources support the robustness of morphology-guided prompting under source-dependent context variation.

Table 2: Bounding-box/pixel-level localization performance (%) from separate category-disjoint evaluations on four industrial datasets. Other notation follows Table 1.

| Method (Box/Seg) | GC10-DET | | | Magnetic Tile | | | MVTec AD | | | Real-IAD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{75}$ | $AP_{50}$ | AP | $AP_{75}$ | $AP_{50}$ | AP | $AP_{75}$ | $AP_{50}$ | AP | $AP_{75}$ | $AP_{50}$ |
| GDINO (ECCV'24) | 0.1/– | 0.1/– | 0.2/– | 2.8/– | 1.9/– | 7.6/– | 8.4/– | 8.8/– | 13.1/– | 24.2/– | 27.5/– | 35.8/– |
| YLE-S (ICCV'25) | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 5.5/0.6 | 2.7/0.0 | 13.5/4.2 | 6.7/5.9 | 6.5/6.9 | 11.5/10.6 | 18.4/8.1 | 20.0/4.8 | 29.9/20.7 |
| YLE-M (ICCV'25) | 0.2/0.3 | 0.0/0.1 | 0.9/1.0 | 4.7/0.6 | 3.8/0.0 | 9.3/3.4 | 6.3/5.7 | 6.1/5.7 | 10.7/10.9 | 25.5/11.6 | 28.0/7.9 | 39.6/28.3 |
| YLE-L (ICCV'25) | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 10.0/1.7 | **7.1**/0.0 | 25.0/10.2 | 7.1/6.8 | 7.3/6.9 | 11.7/11.9 | 25.7/11.1 | 28.3/7.0 | 40.1/27.7 |
| YW-S (CVPR'24) | 0.0/– | 0.0/– | 0.0/– | 0.0/– | 0.0/– | 0.0/– | 6.0/– | 6.7/– | 9.4/– | 17.1/– | 18.9/– | 26.5/– |
| YW-M (CVPR'24) | 0.0/– | 0.0/– | 0.0/– | 1.4/– | 1.1/– | 3.0/– | 11.0/– | 11.6/– | 16.4/– | 25.2/– | 27.9/– | 39.2/– |
| GLEE (CVPR'24) | 0.0/0.0 | 0.0/0.0 | 0.1/0.1 | 2.3/0.8 | 0.5/0.0 | 14.1/5.0 | 2.1/1.7 | 1.9/1.4 | 4.5/3.8 | 17.0/13.1 | 19.1/13.2 | 26.1/23.9 |
| SEEM (NeurIPS'23) | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 1.0/0.2 | 0.7/0.0 | 2.5/1.0 | 1.1/1.8 | 1.0/1.4 | 2.3/4.1 | 6.0/4.7 | 4.9/3.4 | 13.1/11.7 |
| SINE† (NeurIPS'24) | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 0.1/0.0 | 0.0/0.0 | 0.2/0.1 | 0.1/0.1 | 0.1/0.1 | 0.3/0.4 | 0.3/0.2 | 0.3/0.1 | 0.7/0.5 |
| UniSpector† (CVPR'26) | 1.3/0.9 | 0.4/0.2 | 3.5/2.5 | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | 0.1/0.1 | 6.3/5.0 | 6.8/5.4 | 10.0/8.7 |
| DINOv† (CVPR'24) | **13.1/8.9** | **5.6**/4.6 | **34.2/31.3** | 20.7/8.5 | 3.8/**0.2** | **59.1**/42.9 | 17.7/12.6 | 17.5/11.4 | 29.7/24.5 | 26.0/19.7 | 28.7/19.6 | 40.5/36.1 |
| OpenMorphoQuery† | 5.6/4.6 | 3.9/**3.1** | 13.5/12.7 | **22.5/10.5** | 6.5/**0.2** | 57.1/**44.9** | **25.3/20.8** | **27.8/22.4** | **37.0/34.3** | **33.5/24.3** | **37.6/23.8** | **51.0/45.2** |

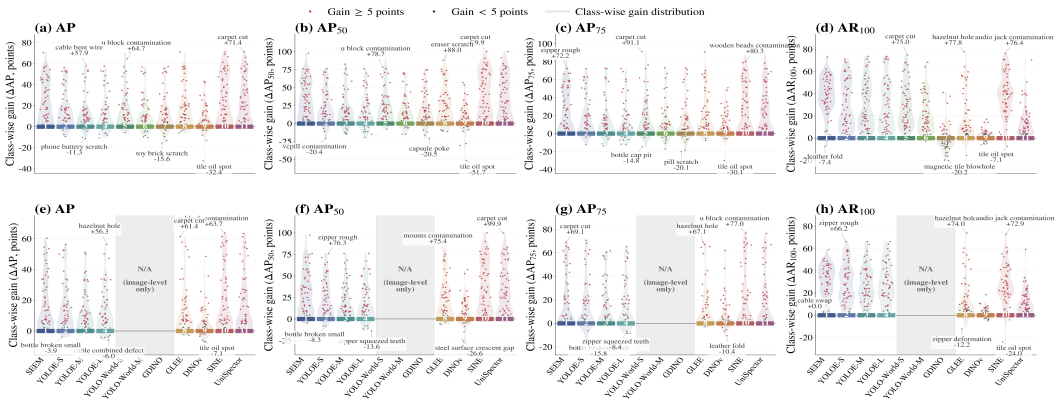



Figure 3: Class-wise performance gains of OpenMorphoQuery over competing methods across 52 unseen industrial defect categories. Panels (a)–(d) and (e)–(h) report AP, $AP_{50}$, $AP_{75}$, and $AR_{100}$ for box and mask prediction, respectively. Each point denotes one category–baseline pair, shaded contours summarize the gain distribution, and red points indicate gains of at least 5 points.

**Distributional Class-wise Gain Analysis.** Beyond aggregate results, we quantify OpenMorphoQuery's category-wise gains as $\Delta m_{b,c} = m_{\mathrm{OMQ},c} - m_{b,c}$. Figure 3 reports box/mask gain distributions in panels (a)–(d)/(e)–(h), respectively. Each point denotes a category–baseline pair, density contours summarize gain distributions, red markers highlight gains above 5 points, and positive values indicate that OpenMorphoQuery outperforms the corresponding baseline. The distributions lie

predominantly above zero, with isolated category-specific regressions. OpenMorphoQuery especially surpasses SINE on *audio jack contamination* by **+76.4/+72.9** box/mask $AR_{100}$ points. Complete category-wise box and mask results are reported in Tables A4–A11.

**Ablation Studies.** Table 3 reports cumulative component ablations while keeping the backbone, training protocol, and fixed, mutually distinct 5-shot support set unchanged. NMDE provides the largest incremental gain, raising box/mask AP by **4.2/3.1** points and $AP_{75}$ by **6.0/3.8** points. This supports encoding prompt morphology relative to shared bases instead of leaving it tied to a support-specific reference frame. PTQI then adds **1.6/1.3** box/mask AP points, indicating the additional benefit of anchoring the shared representation to target-image evidence. Adding PCPE further improves box/mask AP by 0.7/0.9 points, with gains across all six precision metrics, indicating that prototype calibration provides an additional benefit beyond shared-frame encoding and target anchoring. Overall, OMQ exceeds DINOv by **6.5/5.3** box/mask AP points and **8.7/6.5** $AP_{75}$ points. The larger gains under strict overlap, together with the defect-centered responses in Table 5, provide complementary evidence that OMQ reduces the mismatch between support-bound morphology and its grounding in the target image.

Table 3: Component-wise ablation. Components are added cumulatively, each triplet reports $AP/AP_{75}/AP_{50}$. Best results are shown in bold.

| Stage | Box | Mask |
|---|---|---|
| | $AP/AP_{75}/AP_{50}$ | |
| DINOv | 22.0/22.6/36.3 | 16.2/15.1/31.5 |
| + NMDE | 26.2/28.6/40.3 | 19.3/18.9/35.8 |
| + PTQI | 27.8/30.4/42.2 | 20.6/20.6/37.7 |
| + PCPE (OMQ) | **28.5/31.3/43.6** | **21.5/21.6/39.1** |

Table 4: Visual-prompt scaling. Each cell stacks box (B) and mask (M) $AP/AP_{75}$ under identical fixed support sets. Best results are shown in bold.

| Method | 1 shot | 3 shot | 5 shot |
|---|---|---|---|
| SINE | *B*: 0.0/0.0<br>*M*: 0.0/0.0 | *B*: 0.0/0.0<br>*M*: 0.0/0.0 | *B*: 0.2/0.2<br>*M*: 0.1/0.1 |
| DINOv | *B*: 18.5/19.3<br>*M*: 13.6/13.0 | *B*: 21.4/21.7<br>*M*: 15.5/14.5 | *B*: 22.0/22.6<br>*M*: 16.2/15.1 |
| OMQ | *B*: **20.9/22.6**<br>*M*: **15.8/15.4** | *B*: **26.5/28.9**<br>*M*: **19.8/19.7** | *B*: **28.5/31.3**<br>*M*: **21.5/21.6** |

**Visual-Prompt Scaling and Sample Efficiency.** Table 4 examines how SINE, DINOv, and OMQ scale with the number of support exemplars, where each entry reports $AP/AP_{75}$ for box and mask localization. For each shot budget, the support images are selected once before evaluation, are mutually distinct within each category, do not appear in the evaluation query set, and are identical across all methods. OMQ consistently outperforms both competing methods across all shot budgets. Compared with the stronger DINOv baseline, OMQ improves box $AP/AP_{75}$ by **+2.4/+3.3**, **+5.1/+7.2**, and **+6.5/+8.7** points under the three shot settings, respectively. The corresponding mask gains are **+2.2/+2.4**, **+4.3/+5.2**, and **+5.3/+6.5** points. The advantage of OMQ widens as additional exemplars become available. From 1 to 5 shots, its box and mask AP increase by **7.6** and **5.7** points, whereas DINOv improves by only **3.5** and **2.6** points. This widening margin indicates that OMQ converts additional exemplars into complementary morphological evidence rather than redundant prompt observations, thereby strengthening category-level morphology estimation and target grounding.

Table 5: Matched prompt–target response alignment on 14,320 GT-positive queries from 52 unseen categories. Scores are category-macro averages. $\Delta$ denotes OMQ minus DINOv. Confidence intervals are obtained from 10,000 paired bootstrap resamples over categories. Win rate is the percentage of categories with positive gain. All metrics are higher-is-better.

| Metric ↑ | DINOv | OMQ | $\Delta$ | 95% CI | Win rate |
|---|---|---|---|---|---|
| Pixel AUROC | 0.9592 | **0.9722** | +0.0129 | [0.0047, 0.0238] | 76.9% |
| Response-mass precision | 0.3703 | **0.6218** | +0.2515 | [0.2066, 0.2978] | 96.2% |
| Top-1% precision | 0.4508 | **0.4921** | +0.0413 | [0.0140, 0.0798] | 75.0% |
| Top-5% precision | 0.2308 | **0.2542** | +0.0234 | [0.0034, 0.0537] | 71.2% |

**Prompt–Target Reference-Frame Diagnostic.** We test whether prompt–target reference-frame mismatch drives the transfer difficulty through an observable prediction. If prompt morphology

remains anchored to support-specific appearance and context, category-conditioned responses should spread over similar host-product structures. Reanchoring should instead preserve defect morphology while relocating evidence to its target position and extent. We examine this prediction on 14,320 GT-positive queries from 52 unseen categories using identical fixed 5-shot supports. Figure 4 shows the response geometry, and Table 5 quantifies spatial ranking, mass allocation, and peak localization.

The results are consistent with this prediction. Relative to DINOv, OMQ improves category-macro pixel AUROC, response-mass precision, and top-1%/top-5% precision by **+1.29**, **+25.15**, and **+4.13/+2.34** percentage points, respectively. All paired bootstrap confidence intervals exclude zero, with positive category-wise gains on **71.2%–96.2%** of the categories. The rank-based gains indicate improved defect-pixel ordering, while the response-mass gain indicates greater concentration of category-conditioned evidence within annotated defect regions. Because these metrics are invariant to uniform positive rescaling of the final response map, the improvements cannot be attributed to response-magnitude amplification alone. With support exemplars, evaluation queries, backbone family, decoder architecture, and evaluation protocol matched across the compared systems, the alignment gains, component ablations, and qualitative response patterns provide convergent evidence for the reference-frame account of morphology transfer. Further details are provided in Appendix A.6.

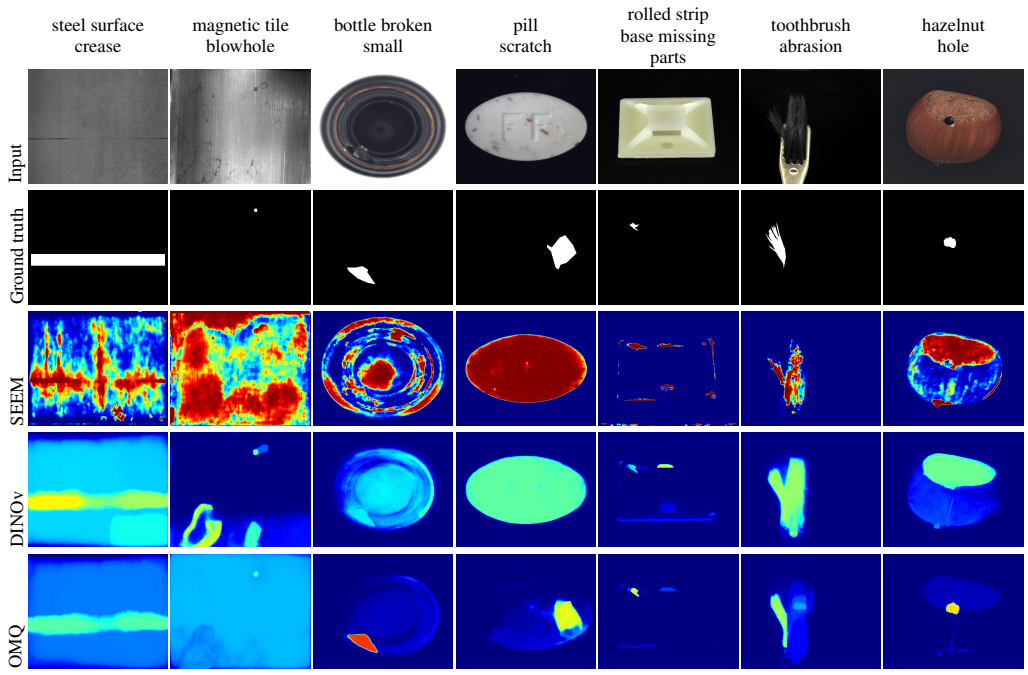



Figure 4: Class-conditioned response heatmaps across seven unseen defect categories. Rows show the input, ground truth, and responses from SEEM, DINOv, and OpenMorphoQuery; warmer colors indicate stronger localization responses.

## 5 Conclusion

We present OpenMorphoQuery for unseen industrial defect localization under morphology-defined openness. It stems from identifying reference-frame mismatch between context-bound prompt morphology and target-image evidence as a source of localization errors. NMDE builds a shared morphology representation, PCPE stabilizes its prototypes across perturbations, and PTQI forms target-grounded queries through optimal transport. Across category-disjoint evaluations of over 160,000 images from four datasets, OpenMorphoQuery achieves state-of-the-art localization performance among visual-prompt methods. Ablations and alignment analyses support the design and its reference-frame interpretation. This formulation offers a new perspective for open-world industrial inspection and motivates flexible, automatically acquired prompts.

### AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval, language editing, and manuscript presentation. All retrieved references, technical descriptions, numerical results, and scientific claims were independently checked by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

### REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3–4 and Appendix A document the benchmark construction and category-disjoint split, fixed 1-/3-/5-shot support protocols, model architecture and module formulations, optimization and inference settings, global query allocation, evaluation metrics, and computational cost. Appendix A further provides complete implementation hyperparameters, pseudocode, per-category results for all 52 unseen categories, and the construction and aggregation procedure of the alignment diagnostic. Where applicable, visual-prompt methods are evaluated using identical category splits, fixed support exemplars, and COCO evaluation protocols.

## REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.

Yongqi Chen, Jiawei Pan, Jiayu Lei, Deyu Zeng, Zongze Wu, and Changsheng Chen. EEE-Net: Efficient edge enhanced network for surface defect detection of glass. *IEEE Transactions on Instrumentation and Measurement*, 72:1–13, 2023.

Zhao-Min Chen, Xinjian Huang, Yisu Ge, and Yu Li. SFR-Net: Steering-Fusion-Refining network in multi-label zero-shot sewer defect detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41636–41645, 2026.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-Time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Nadia Drir, Adel Mellit, Nicola Blasuttigh, and Alessandro Massi Pavan. Enhancing defect detection in photovoltaic cells through advanced cascade CNN with attention mechanisms on electroluminescence images. *Energy Conversion and Management*, 348:120636, 2026.

Yibin Huang, Congying Qiu, Yue Guo, Xiaonan Wang, and Kui Yuan. Surface defect saliency of magnetic tile. In *Proceedings of the IEEE 14th International Conference on Automation Science and Engineering*, pp. 612–617, 2018.

Le Jiang, Yan Huang, Zhen Xu, Yong Xu, Hau-San Wong, and Si Wu. Defect cue-preserved structural feature refinement for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35607–35616, 2026.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-Rex2: Towards generic object detection via text-visual prompt synergy. In *Proceedings of the European Conference on Computer Vision*, pp. 38–57, 2024.

Geonuk Kim, Minhoi Kim, Kangil Lee, Minsu Kim, Hyeonseong Jeon, Jeonghoon Han, Hyoungjoon Lim, and Junho Yim. UniSpector: Towards universal open-set defect recognition via spectral-contrastive visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6261–6270, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Yujin Lee, Sewon Kim, Daeun Moon, Seoyoon Jang, and Hyunsoo Yoon. Bidirectional multimodal prompt learning with scale-aware training for few-shot multi-class anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35577–35586, 2026.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3041–3050, 2023.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Visual in-context prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12861–12871, 2024.

Huimin Li, Boxuan Hu, Yulin Zhang, Xiuzhuang Zhou, and Junlin Hu. Hyperbolic defect feature synthesis for few-shot defect classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19602–19612, 2026a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Yufei Li, Long Tian, Yuyang Dai, Wenchao Chen, Liang Bao, and Xiyang Liu. FastRef: Fast prototype refinement for few-shot industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43040–43049, 2026b.

Juhua Liu, Chaoyue Wang, Hai Su, Bo Du, and Dacheng Tao. Multistage GAN for fabric defect detection. *IEEE Transactions on Image Processing*, 29:3388–3400, 2020.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55, 2024a.

Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25095–25119, 2024b.

Xiaoming Lv, Fajie Duan, Jia-jia Jiang, Xiao Fu, and Lin Gan. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6):1562, 2020.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 728–755, 2022.

Tsai-Ching Ni, Cheng-Chi Chen, and Yuan-Fu Yang. Towards open-vocabulary industrial defect understanding with a large-scale multimodal dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13059–13068, 2026.

Xiangkai Shen, Jinhai Liu, Lin Jiang, Xiaoyuan Liu, and Huaguang Zhang. A novel weld defect detection method for intelligent magnetic flux leakage detection system via contextual relation network. *IEEE Transactions on Industrial Electronics*, 71(6):6304–6314, 2024.

Xiangwen Shi, Shaobing Zhang, Miao Cheng, Lian He, Xianghong Tang, and Zhe Cui. Few-shot semantic segmentation for industrial defect recognition. *Computers in Industry*, 148:103901, 2023.

Jaewoo Song, Daemin Park, Kanghyun Baek, Sangyub Lee, Jooyoung Choi, Eunji Kim, and Sungroh Yoon. DefectFill: Realistic defect generation with inpainting diffusion model for visual inspection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18718–18727, 2025.

Xuan Tong, Yuxuan Lin, Junxiong Lin, Xinji Mai, Haoran Wang, Zeng Tao, Yang Yao, Ruofan Wang, and Wenqiang Zhang. CADiff: Context-Aware diffusion for controllable anomaly generation in anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12): 9557–9565, 2026.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-Time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24591–24602, 2025.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22883–22892, 2024.

Zhuonan Wang, Zhenxuan Fan, Siwen Tan, Yu Zhong, Yuqian Yuan, Haoyuan Li, Hao Jiang, Wenqiao Zhang, Feifei Shao, Hongwei Wang, and Jun Xiao. MAU-GPT: Enhancing multi-type industrial anomaly understanding via anomaly-aware and generalist experts adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(31):26787–26795, 2026.

Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3783–3795, 2024.

Rongcheng Wu, Hao Zhu, Shiying Zhang, Mingzhe Wang, Zhidong Li, Hui Li, Jianlong Zhou, Jiangtao Cui, Fang Chen, Pingyang Sun, Qiyu Liao, and Ye Lin. RcAE: Recursive reconstruction framework for unsupervised industrial anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13):10736–10744, 2026.

Tugba Yildizli, Tianlong Jia, Jeroen Langeveld, and Riccardo Taormina. Self-supervised learning for multi-label sewer defect classification. *Automation in Construction*, 182:106751, 2026.

Ke Zhang, Ruiheng Zhou, Jiacun Wang, Yangjie Xiao, Xiwang Guo, and Chaojun Shi. Transmission line component defect detection based on UAV patrol images: A self-supervised HC-ViT method. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(11):6510–6521, 2024.

Zekai Zhang, Qinghui Chen, Maomao Xiong, Shijiao Ding, Zhanzhi Su, Xinjie Yao, Yiming Sun, Cong Bai, and Jinglin Zhang. Zero-shot learning in industrial scenarios: New large-scale benchmark, challenges and baseline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (10):10357–10366, 2025.

Hao Zhong, Daxing Fu, Ling Xiao, Fang Zhao, Jie Liu, Youmin Hu, and Bo Wu. STFE-Net: A multi-stage approach to enhance statistical texture feature for defect detection on metal surfaces. *Advanced Engineering Informatics*, 61:102437, 2024.

Xin Zhou, Shihua Zhou, Yongchao Zhang, Zhaohui Ren, Zeyu Jiang, and Hengfa Luo. GDALR: Global dual attention and local representations in transformer for surface defect detection. *Measurement*, 229:114398, 2024.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19769–19782, 2023.

# Supplementary Material

# OpenMorphoQuery: From Visual Morphology Prompts to Target-Grounded Queries for Unseen Industrial Defect Localization

## A Additional Experimental Details and Results

### A.1 OpenMorphoQuery Query-Generation Pipeline

Algorithm A1 summarizes the morphology-to-query transformation over all $C$ prompted categories in a single forward pass. The decoder uses $Q = 300$ localization queries in total. The cyclic index determines the prototype used to seed each localization query. The additional $C$ category-content tokens are appended separately and are not counted in $Q$. The resulting queries are decoded with the target features and scored against all $C$ category-content tokens, forming a shared query pool for box and mask prediction. During training, the prototype-consistency term $\mathcal{L}_{\mathrm{pc}}$ supplements the standard localization objective.

### A.2 Benchmark Composition and Category-Disjoint Split

Table A1 summarizes the unified category-disjoint benchmark constructed from GC10-DET (Lv et al., 2020), Magnetic Tile (Huang et al., 2018), MVTec AD (Bergmann et al., 2019), and Real-IAD (Wang et al., 2024). These datasets are integrated into a common train–test space. The training split contains 38,583 images, 43,842 annotated defect instances, and 147 seen categories. The evaluation split contains 121,447 images and 17,100 annotated instances from 52 category-disjoint unseen categories, with $\mathcal{C}_{\mathrm{tr}} \cap \mathcal{C}_{\mathrm{te}} = \varnothing$. Among these images, 14,580 contain annotated defects and 106,867 are defect-free, making 88.0% of the evaluation pool negatives. This large, negative pool casts unseen-defect localization as cross-dataset, defect-sparse search, jointly evaluating morphology transfer and false-positive control across products, materials, and backgrounds.

For traceability, our 147/52 seen–unseen category assignment coincides with the class partition released by InsA with seed 42 (Kim et al., 2026). We fix this category partition and use the same five-shot support set across all methods. The five support images within each category are mutually distinct and excluded from the query pool. The category configuration and seen-category training split remain unchanged, while defect-free images are retained during evaluation. This extends defect-focused evaluation to large-scale defect-sparse search under substantially greater false-positive pressure.

Table A1: Dataset-level statistics of the unified industrial benchmark and its category-disjoint split. Img., Inst., and Cat. denote the numbers of images, annotated defect instances, and defect categories, respectively. D/F denotes defective and defect-free test images.

| Dataset | Train Img./Inst./Cat. | Test Img./Inst./Cat. | Test D/F |
|---|---|---|---|
| GC10-DET | 1,862 / 2,698 / 7 | 420 / 568 / 3 | 418 / 2 |
| Magnetic Tile | 273 / 398 / 4 | 1,071 / 116 / 1 | 115 / 956 |
| MVTec AD | 887 / 352 / 53 | 4,467 / 484 / 20 | 371 / 4,096 |
| Real-IAD | 35,561 / 40,394 / 83 | 115,489 / 15,932 / 28 | 13,676 / 101,813 |
| **Total** | **38,583 / 43,842 / 147** | **121,447 / 17,100 / 52** | **14,580 / 106,867** |

### A.3 Additional Implementation Details

Unless otherwise specified, all model variants use identical data partitions, support exemplars, data augmentations, and optimization schedules. We employ a Swin-T backbone initialized from the DINOv pretrained checkpoint together with a MaskDINO-style decoder. Table A2 summarizes the principal architecture, optimization, and model-size settings required for reproduction. All OpenMorphoQuery-specific hyperparameters are fixed across shot settings and are never selected or adjusted using the 52 unseen evaluation categories.

Table A2: Core implementation, optimization, and efficiency settings of OpenMorphoQuery.

| **Setting** | **Configuration** |
|---|---|
| Input and optimization | Input resolution $640 \times 640$, AdamW optimizer, 50 training epochs, initial learning rate $2 \times 10^{-5}$, weight decay $0.05$, global batch size 32 on eight NVIDIA RTX 3090 GPUs. |
| Shared architecture | Swin-T backbone, MaskDINO-style decoder, and feature dimension $D = 256$. A global budget of $Q = 300$ localization queries is distributed across all $C$ prompted categories. The additional $C$ category-content tokens are appended separately and are not counted in $Q$. |
| Prompt RoI processing | RoIAlign produces $7 \times 7$ prompt features, which are adaptively pooled to $4 \times 4$ and tokenized into 16 tokens per exemplar for PCPE. |
| NMDE | $M = 8$ shared morphology bases, assignment temperature $\tau = 0.07$. |
| PCPE | Three $256 \to 32 \to 256$ GELU experts, $3 \to 12 \to 3$ GELU router with standard softmax, full soft expert mixture, $\tau_r = 0.10$, $\lambda_{\mathrm{pc}} = 0.10$, $\gamma_{\mathrm{app}} = 0.10$, $\gamma_{\mathrm{geo}} = 0.25$, and $\gamma_{\mathrm{bnd}} = 0.15$. |
| PTQI | Global cyclic assignment of $Q = 300$ localization slots, fixed spatial anchors from the first 300 locations of an $18 \times 18$ grid, positional weight $\lambda_{\mathrm{pos}} = 0.15$, uniform query marginal $\pi_q = 1/Q$, content-weighted target marginal with $\tau_{\mathrm{mass}} = 0.07$, $\varepsilon_{\mathrm{ot}} = 0.10$ and $L_{\mathrm{sk}} = 5$ Sinkhorn iterations, barycentric aggregation and query initialization by $\mathrm{LN}\big(\mathrm{Norm}(\mathbf{E}) + g_{\mathrm{leg}}\mathbf{Q}^{\mathrm{leg}}\big)$. |
| Model efficiency | 54.315M trainable parameters in total. NMDE, PCPE, and PTQI contain 150,273, 50,104, and 513 trainable parameters, respectively, totaling 200,890 module-specific parameters. |

The legacy embedding $\mathbf{p}_c^{\mathrm{leg}}$ is the category prototype produced by the original DINOv visual-prompt pathway from the same support RoI features before OpenMorphoQuery refinement. It therefore introduces neither external semantic information nor additional support exemplars. For every shot setting, the support images within each category are mutually distinct, selected once before evaluation, shared across all applicable visual-prompt methods, and completely excluded from the query test set. This protocol prevents support–query leakage while ensuring that performance differences cannot be attributed to different exemplar selections.

All trainable OpenMorphoQuery components are optimized jointly with the underlying detector, while the downstream DINOv decoding and prediction procedures remain unchanged. The reported parameter counts include all trainable model components. Additionally, OpenMorphoQuery requires 78.5ms per image on a single NVIDIA RTX 3090 GPU.

UniSpector reimplementation. Because a complete public implementation of UniSpector was unavailable at the time of evaluation, we reimplemented it according to the architecture and training specifications reported in the original paper (Kim et al., 2026). Our implementation uses a Swin-T visual backbone, a six-layer multiscale deformable encoder, and a nine-layer MaskDINO-style decoder. SSPE fuses the spatial representation of each prompt RoI with a 256-bin radial spectrum, CPE improves within-class prompt consistency through an angular-margin objective, and PQS selects 300 prompt-relevant queries from the target-image features. Table A3 reports the complete configuration used for evaluation. All UniSpector results are obtained using this implementation under the same category split, fixed non-repeated five-exemplar support sets, and evaluation protocol as DINOv and OpenMorphoQuery.

Table A3: Configuration of our paper-guided UniSpector reimplementation.

| Component | Configuration |
|---|---|
| Implementation basis | Paper-guided reimplementation of UniSpector in a MaskDINO-style detection and segmentation framework. |
| Backbone | Swin-T initialized from `model_swinT.pth`, with four-scale `res2–res5` features. |
| Feature encoder | Six-layer multiscale deformable encoder with hidden dimension 256 and four feature levels. |
| Decoder | Nine layers, eight attention heads, and FFN dimension 2048. |
| Queries | 300 generic queries, with one content token generated for each visual prompt. |
| SSPE | Each prompt RoI is converted to grayscale and resized to $256 \times 256$. An exact GPU `rFFT2` with orthonormal normalization produces 256 radial frequency bins. The L2-normalized spectrum is fused with the spatial prompt feature. |
| SSPE engineering | Gradients are disabled through the FFT path, at most eight RoIs are processed per chunk, and the cuFFT plan cache is limited to 32. |
| CPE | ArcFace-style angular contrastive learning with scale 30, margin 0.5, and loss weight 1.0. The loss is computed after gathering prompts across GPUs. |
| PQS | Cosine similarity is computed between visual tokens and category prototypes. Gumbel–Softmax uses temperature 1.0, followed by hard top-300 selection with straight-through gradient estimation. The original four-dimensional reference points are retained. |
| Denoising training | Segmentation denoising with 100 denoising queries and noise scale 0.4. |

### A.4 Evaluation Protocol and Complete Category-Wise Results

We evaluate all methods on the same 52 unseen categories using the COCO evaluation protocol. AP averages precision over IoU thresholds from 0.50 to 0.95, $AP_{50}$ and $AP_{75}$ report precision at fixed IoU thresholds, and $AR_{100}$ reports average recall with at most 100 predictions per image. All entries are percentages rounded to one decimal place. UniSpector, DINOv, SINE, and OpenMorphoQuery use the same fixed, non-repeated five-exemplar support set for each category, while methods that do not require visual exemplars follow their native prompting protocol. For every category, the best result is boldfaced and the second-best distinct result is underlined. Ties receive the same emphasis. A dash indicates that the corresponding method does not provide a pixel-level result.

All numerical comparisons follow a unified evaluation protocol. We reimplement UniSpector (Kim et al., 2026) according to the architecture and training specifications reported in the original paper and evaluate it using the same category split and fixed, non-repeated five-exemplar support sets as the other visual-prompt methods. Details of the reimplementation are provided in Appendix A.3. For compactness, method names are abbreviated consistently across all tables. YLE-S/M/L denote YOLOE-S/M/L, YW-S/M denote YOLO-World-S/M, GDN denotes Grounding DINO, UniSp. denotes UniSpector, and OMQ denotes OpenMorphoQuery. Tables A4–A7 report exhaustive bounding-box results, whereas Tables A8–A11 report the corresponding pixel grounding results. Categories are grouped by their source dataset to expose cross-domain behavior without filtering or category selection.

**Algorithm A1** Multi-Category Morphology-to-Query Generation in OpenMorphoQuery

**Require:** Target image $I^q$, prompt sets $\{\mathcal{P}_c\}_{c=1}^{C}$ with $\mathcal{P}_c = \{(I^p_{c,k}, R^p_{c,k})\}_{k=1}^{K}$, dictionary $\mathbf{B}$, legacy category embeddings $\{\mathbf{p}_c^{\text{leg}}\}_{c=1}^{C}$, legacy decoder queries $\mathbf{Q}^{\text{leg}} = [\mathbf{q}_q^{\text{leg}}]_{q=1}^{Q}$, fixed anchors $\{\mathbf{a}_q\}_{q=1}^{Q}$, $Q = 300$, $\tau, \tau_r, \tau_{\text{mass}}, \lambda_{\text{pos}}, \varepsilon_{\text{ot}}, L_{\text{sk}}, \epsilon$

**Ensure:** Localization queries $\mathbf{Q}^0 \in \mathbb{R}^{Q\times D}$, category prototypes $\mathbf{P}^r \in \mathbb{R}^{C\times D}$, training loss $\mathcal{L}_{\text{pc}}$

$(\mathbf{F}^q, \mathbf{V}, \mathbf{Z}) \leftarrow$ EncodeTarget$(I^q)$ // $\mathbf{Z} = [\mathbf{z}_j]_{j=1}^{N_t}$

$\bar{\mathbf{V}} \leftarrow \mathrm{RowNorm}(\mathbf{V})$

$\bar{\mathbf{B}} \leftarrow \mathrm{RowNorm}(\mathbf{B})$

**Normalized morphology encoding and prototype calibration**

**for** $c = 1$ to $C$ **do**

  **for** $k = 1$ to $K$ **do**

    $(\mathbf{G}_{c,k}, \mathbf{X}_{c,k}) \leftarrow$ PromptTokens$(I^p_{c,k}, R^p_{c,k})$ // $7\times7 \to 4\times4 \to 16$ tokens

    $\boldsymbol{\mu}_{c,k} \leftarrow 16^{-1}\sum_{n=1}^{16} \mathbf{x}_{c,k,n}$

    $\bar{\mathbf{X}}_{c,k} \leftarrow \mathrm{RowNorm}(\mathbf{X}_{c,k})$

    $\mathbf{A}_{c,k} \leftarrow \mathrm{RowSoftmax}\big(\bar{\mathbf{X}}_{c,k}\bar{\mathbf{B}}^\top/\tau\big)$

    **for** $m = 1$ to $M$ **do**

      $\mathbf{z}_{c,k,m} \leftarrow \frac{\sum_{n=1}^{16} A_{c,k,n,m}(\mathbf{x}_{c,k,n} - \bar{\mathbf{b}}_m)}{\sum_{n=1}^{16} A_{c,k,n,m} + \epsilon}$

    **end for**

    $\mathbf{p}_{c,k} \leftarrow$ ProjectNorm$(\mathbf{z}_{c,k,1} \oplus \cdots \oplus \mathbf{z}_{c,k,M})$

    **for all** $e \in \mathcal{E} = \{\text{app}, \text{geo}, \text{bnd}\}$ **do**

      $\mathbf{X}^e_{c,k} \leftarrow \mathcal{T}_e(\mathbf{X}_{c,k}; \mathbf{G}_{c,k}, \boldsymbol{\mu}_{c,k})$

      $\mathbf{p}^e_{c,k} \leftarrow \mathrm{NMDE}(\mathbf{X}^e_{c,k}; \mathbf{B}, \tau)$

      $d_{c,k,e} \leftarrow 1 - \cos\big(\mathbf{p}_{c,k}, \mathbf{p}^e_{c,k}\big)$

    **end for**

    $\boldsymbol{\alpha}_{c,k} \leftarrow \mathrm{Softmax}(R([d_{c,k,e}]_{e\in\mathcal{E}}))$

    $\mathbf{p}^r_{c,k} \leftarrow \mathrm{Norm}\big(\mathbf{p}_{c,k} + \sum_{e\in\mathcal{E}} \alpha_{c,k,e} E_e(\mathbf{p}_{c,k})\big)$

    $\mathbf{p}^{r,e}_{c,k} \leftarrow \mathrm{Norm}\big(\mathbf{p}^e_{c,k} + \sum_{e'\in\mathcal{E}} \alpha_{c,k,e'} E_{e'}(\mathbf{p}^e_{c,k})\big), \quad \forall e \in \mathcal{E}$

    $r_{c,k} \leftarrow |\mathcal{E}|^{-1}\sum_{e\in\mathcal{E}} \cos\big(\mathbf{p}_{c,k}, \mathbf{p}^e_{c,k}\big)$

  **end for**

  $\mathbf{b}_c \leftarrow \mathrm{Softmax}\big([r_{c,k}/\tau_r]_{k=1}^{K}\big)$

  $\widehat{\mathbf{p}}^r_c \leftarrow \mathrm{Norm}\Big(\sum_{k=1}^{K} b_{c,k}\mathbf{p}^r_{c,k}\Big)$; $\bar{r}_c \leftarrow \sum_{k=1}^{K} b_{c,k} r_{c,k}$

  $g_c \leftarrow \sigma(g_0 + g_1(1 - \bar{r}_c))$; $\bar{\mathbf{p}}^r_c \leftarrow \mathrm{Norm}\big(\widehat{\mathbf{p}}^r_c + g_c\,\mathrm{Norm}(\mathbf{p}_c^{\text{leg}})\big)$

**end for**

$\mathbf{P}^r \leftarrow$ Stack$(\bar{\mathbf{p}}^r_1, \ldots, \bar{\mathbf{p}}^r_C)$

**if** training **then**

  $\mathcal{L}_{\text{pc}} \leftarrow$ ConsistencyLoss$(\{\mathbf{p}^r_{c,k}, \mathbf{p}^{r,e}_{c,k}\}_{c,k,e})$ // prototype-consistency objective

**else**

  $\mathcal{L}_{\text{pc}} \leftarrow 0$

**end if**

**Prototype-Transport Query Initialization**

**for** $q = 1$ to $Q$ **do**

  $c(q) \leftarrow 1 + ((q-1) \bmod C)$; $\widetilde{\mathbf{p}}_q \leftarrow \bar{\mathbf{p}}^r_{c(q)}$ // cyclic prototype seeding

**end for**

$\widetilde{\mathbf{P}} \leftarrow$ Stack$(\widetilde{\mathbf{p}}_1, \ldots, \widetilde{\mathbf{p}}_Q)$ // $Q = 300$ global localization queries

$\mathbf{D} \leftarrow \big[-\widetilde{\mathbf{p}}_q^\top \bar{\mathbf{v}}_j + \lambda_{\text{pos}}\|\mathbf{a}_q - \mathbf{z}_j\|_2^2\big]_{q,j}$

$s_j \leftarrow \max_{1\le c\le C} (\bar{\mathbf{p}}^r_c)^\top \bar{\mathbf{v}}_j, \quad j = 1, \ldots, N_t$

$\boldsymbol{\pi} \leftarrow Q^{-1}\mathbf{1}_Q$; $\boldsymbol{\nu} \leftarrow \mathrm{Softmax}\big([s_j/\tau_{\text{mass}}]_{j=1}^{N_t}\big)$

$\mathbf{T}^\star \leftarrow$ Sinkhorn$(\mathbf{D}, \boldsymbol{\pi}, \boldsymbol{\nu}, \varepsilon_{\text{ot}}, L_{\text{sk}})$

$g_{\text{leg}} \leftarrow \sigma(\beta_{\text{leg}})$

**for** $q = 1$ to $Q$ **do**

  $\mathbf{e}_q \leftarrow \dfrac{\sum_{j=1}^{N_t} T^\star_{qj}\mathbf{v}_j}{\sum_{j=1}^{N_t} T^\star_{qj} + \epsilon}$

  $\mathbf{q}^0_q \leftarrow \mathrm{LN}\big(\mathrm{Norm}(\mathbf{e}_q) + g_{\text{leg}}\mathbf{q}_q^{\text{leg}}\big)$

**end for**

$\mathbf{Q}^0 \leftarrow$ Stack$(\mathbf{q}^0_1, \ldots, \mathbf{q}^0_Q)$

*All decoded localization queries are scored against all $C$ category-content tokens.*

**return** $\mathbf{Q}^0, \mathbf{P}^r, \mathcal{L}_{\text{pc}}$ // $C$ category tokens are separate from $Q$

Table A4: Complete category-level bounding-box AP (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 3.8 | **8.9** | 0.0 | 7.7 |
| steel surface crescent gap | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.1 | **13.4** | 0.0 | 0.5 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **16.8** | 0.0 | 8.6 |
| magnetic tile blowhole | 1.0 | 5.5 | 4.7 | 10.0 | 0.0 | 1.4 | 2.8 | 2.3 | 0.0 | 20.7 | 0.1 | **22.5** |
| bottle broken small | 8.1 | 56.2 | 43.0 | 51.2 | 53.1 | **63.5** | 41.2 | 9.8 | 0.0 | 22.9 | 0.1 | 53.1 |
| cable bent wire | 0.0 | 1.3 | 0.0 | 0.0 | 0.7 | 7.4 | 0.6 | 0.0 | 0.0 | 42.2 | 0.0 | **57.9** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 0.2 | 4.9 | 5.0 | **12.6** | 3.7 | 0.0 | 1.9 | 5.7 | 0.0 | 0.2 | 0.0 | 5.5 |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | 1.3 | 0.4 | 0.2 | 0.0 | 0.0 | 0.8 | 0.0 | **2.4** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.0 | 0.0 | 0.0 | 0.1 | 10.1 | 0.0 | **16.4** |
| capsule poke | 0.0 | 0.0 | 0.0 | 0.0 | 1.3 | 0.6 | **21.6** | 0.0 | 0.0 | 15.5 | 0.0 | 12.1 |
| capsule scratch | 0.0 | 2.9 | 0.6 | 0.0 | 11.6 | 18.4 | 12.7 | 0.0 | 0.1 | 21.4 | 0.0 | **43.8** |
| carpet cut | 0.8 | 32.0 | 24.4 | 22.9 | 0.6 | 18.7 | 18.9 | 0.1 | 0.0 | 46.1 | 0.0 | **71.5** |
| hazelnut hole | 8.4 | 10.0 | 22.4 | 1.7 | 29.6 | 22.3 | 3.7 | 0.0 | 0.0 | 11.8 | 0.3 | **55.0** |
| leather cut | 1.1 | 11.6 | 9.0 | 11.9 | 13.6 | 1.7 | 1.1 | 0.0 | 0.2 | 15.1 | 0.1 | **19.3** |
| leather fold | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 0.9 | 0.0 | 0.0 | **6.5** | 0.0 | 4.8 |
| metal nut bent | 0.3 | 1.2 | 0.5 | 14.4 | 4.6 | **21.0** | 0.7 | 2.4 | 0.1 | 15.5 | 0.0 | 17.1 |
| metal nut scratch | 0.0 | 0.0 | 0.1 | 0.3 | 0.0 | 0.8 | 0.0 | 0.3 | 0.0 | 0.1 | 0.2 | **6.3** |
| pill scratch | 0.0 | 0.1 | 0.6 | 0.9 | 0.0 | 23.9 | **47.8** | 0.0 | 0.0 | 20.4 | 0.8 | 34.7 |
| tile oil spot | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.0 | **43.8** | 0.0 | 11.4 |
| wood combined defect | 2.4 | 12.2 | 11.6 | 16.7 | 0.1 | 12.6 | 8.5 | 0.0 | 0.0 | 7.6 | 0.0 | **20.6** |
| zipper broken teeth | 0.0 | 0.8 | 9.3 | 0.0 | 0.0 | 7.3 | 3.7 | 13.7 | 0.0 | 11.2 | 0.3 | **16.6** |
| zipper rough | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 20.6 | 0.9 | 2.1 | 0.0 | **55.8** | 0.4 | 52.8 |
| zipper squeezed teeth | 0.0 | 0.1 | 0.1 | **8.4** | 0.0 | 0.1 | 3.5 | **8.4** | 0.3 | 7.2 | 0.0 | 5.6 |
| audio jack contamination | 11.6 | 9.7 | 54.3 | 62.8 | 29.8 | 36.3 | **64.3** | 6.5 | 54.1 | 58.5 | 0.0 | 61.4 |
| bottle cap pit | 8.4 | 22.3 | 36.1 | 42.8 | **46.5** | 41.5 | 36.9 | 23.9 | 1.2 | 41.9 | 0.5 | 40.8 |
| button battery contamination | 17.0 | 34.8 | 37.5 | 37.3 | 20.2 | 28.7 | 49.5 | 41.9 | 20.6 | 30.6 | 0.0 | **50.5** |
| end cap damage | 0.8 | 5.0 | **15.4** | 9.2 | 0.0 | 10.1 | 6.0 | 2.1 | 3.5 | 4.7 | 0.0 | 11.2 |
| end cap missing parts | 0.9 | 0.7 | 1.6 | 8.3 | 7.5 | 15.5 | 7.9 | 4.1 | 6.2 | 15.1 | 0.0 | **25.4** |
| eraser scratch | 25.3 | 52.0 | 6.5 | 21.0 | 57.4 | 51.0 | **61.4** | 3.2 | 0.8 | 45.6 | 5.1 | 60.5 |
| mint foreign objects | 0.7 | 1.9 | 34.0 | 24.4 | 3.0 | 36.3 | 33.9 | 0.7 | 9.5 | **46.4** | 0.0 | 45.5 |
| mounts contamination | 0.1 | 3.0 | 5.7 | 14.6 | 8.1 | 4.5 | 0.6 | 0.0 | 0.5 | **59.6** | 0.1 | 56.0 |
| mounts missing parts | 0.2 | 2.9 | 22.5 | 1.7 | 15.6 | 18.4 | 0.8 | 0.2 | 0.6 | 8.7 | 0.6 | **29.3** |
| mounts pit | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.4 | **2.0** | 0.1 | **2.0** |
| pcb missing parts | 0.2 | 2.9 | 4.5 | **8.3** | 1.7 | 3.0 | 0.2 | 0.0 | 0.1 | 2.5 | 0.0 | 2.0 |
| phone battery scratch | 1.5 | 34.0 | **35.9** | 35.2 | 16.0 | 32.8 | 27.9 | 2.8 | 0.8 | 17.3 | 0.1 | 24.6 |
| rolled strip base missing parts | 11.8 | 15.7 | 30.0 | 25.6 | 33.5 | 28.5 | **39.5** | 13.5 | 1.2 | 22.4 | 1.3 | 33.2 |
| tape missing parts | 7.1 | 40.1 | 40.1 | 30.4 | 17.7 | 27.6 | 49.3 | **59.7** | 1.6 | 52.2 | 0.2 | 54.0 |
| toothbrush abrasion | 0.9 | 9.7 | 13.1 | **13.2** | 3.8 | 9.5 | 4.7 | 0.0 | 4.4 | 5.5 | 0.0 | 11.6 |
| toothbrush contamination | 0.7 | 2.5 | 2.4 | 3.5 | 0.9 | 2.6 | 3.9 | 3.4 | 2.2 | **4.0** | 0.0 | 2.3 |
| toy contamination | 3.8 | 27.1 | 39.4 | 39.2 | 7.0 | 38.7 | 53.1 | 50.3 | 34.8 | 52.4 | 0.1 | **58.2** |
| toy brick scratch | 8.3 | 35.1 | 43.0 | 29.0 | 20.2 | 38.1 | **48.4** | 27.9 | 0.2 | 3.2 | 0.0 | 32.8 |
| transistor missing parts | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | **0.4** |
| u block abrasion | 0.0 | 39.8 | 41.8 | 36.3 | 27.1 | **57.7** | 15.7 | 25.1 | 0.7 | 5.8 | 0.4 | 47.9 |
| u block contamination | 10.1 | 13.8 | 50.8 | 59.8 | 2.0 | 9.9 | 52.3 | 55.4 | 0.5 | 66.4 | 0.0 | **66.7** |
| usb adapter contamination | 0.1 | 9.7 | 23.2 | 27.6 | 6.1 | **28.5** | 1.8 | 9.9 | 0.4 | 26.7 | 0.0 | 27.8 |
| usb adapter scratch | 0.6 | 25.8 | 46.3 | **51.2** | 41.8 | 46.8 | 3.6 | 49.2 | 0.0 | 25.3 | 0.0 | 45.2 |
| vcpill contamination | 28.5 | 44.9 | 39.6 | **45.5** | 31.0 | 42.0 | 24.8 | 18.7 | 2.3 | 24.5 | 0.0 | 37.4 |
| vcpill missing parts | 1.7 | 28.0 | 23.7 | 22.7 | 20.9 | 32.7 | 23.0 | 11.3 | 19.7 | **38.8** | 0.6 | **38.8** |
| wooden beads contamination | 27.0 | 52.2 | 64.7 | 67.3 | 57.9 | 62.3 | 68.4 | 65.3 | 7.5 | 64.7 | 0.0 | **69.1** |
| zipper damage | 0.0 | 1.8 | 0.9 | 1.0 | 0.8 | 0.7 | 0.2 | 0.0 | 0.6 | **2.8** | 0.1 | 2.0 |
| zipper deformation | 0.1 | 0.0 | 0.0 | 0.4 | 0.0 | 0.5 | 0.0 | 0.1 | 1.0 | 1.2 | 0.0 | **1.6** |

**Results and discussion.** Across the 52 categories, OMQ attains a macro-average of 28.55 AP, improving over the matched DINOv baseline by 6.57 points and over UniSpector's 3.47 AP by 25.08 points. It is best (including ties) on 22 categories and ranks in the top two on 37. The gain over DINOv is positive on 37 categories and reaches at least five points on 24. UniSpector obtains 54.1 AP on audio jack contamination but remains near zero across most GC10-DET and MVTec AD categories, so its low macro AP reflects severe cross-category instability rather than uniformly weak localization. In contrast, the large OMQ gains for hazelnut hole (+43.2) and u block abrasion (+42.1) show that morphology-guided queries transfer distinctive structural evidence more consistently across contexts. The mean change on GC10-DET is $-7.45$ points, and tile oil spot drops by 32.4 points, revealing continued sensitivity to subtle, low-contrast texture variations.

Table A5: Complete category-level bounding-box $AP_{50}$ (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.1 | 9.6 | **21.0** | 0.0 | 17.2 |
| steel surface crescent gap | 0.0 | 0.0 | 2.7 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 | 0.6 | **32.0** | 0.0 | 1.5 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.2 | **49.6** | 0.0 | 21.8 |
| magnetic tile blowhole | 2.5 | 13.5 | 9.3 | 25.0 | 0.0 | 3.0 | 7.6 | 14.1 | 0.0 | **59.1** | 0.2 | 57.1 |
| bottle broken small | 21.4 | 77.1 | 54.5 | 74.6 | **78.0** | 75.8 | 48.4 | 10.8 | 0.0 | 29.8 | 0.1 | 70.0 |
| cable bent wire | 0.1 | 2.0 | 0.0 | 0.0 | 1.3 | 11.3 | 0.9 | 0.0 | 0.0 | 63.5 | 0.0 | **70.0** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 0.4 | 6.5 | 8.6 | **15.1** | 4.0 | 0.0 | 2.6 | 6.7 | 0.0 | 0.2 | 0.0 | 7.1 |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | 3.4 | 0.7 | 0.4 | 0.0 | 0.0 | 3.0 | 0.0 | **4.4** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 0.2 | 0.8 | 0.0 | 0.0 | 0.0 | 0.4 | 31.7 | 0.0 | **40.5** |
| capsule poke | 0.0 | 0.0 | 0.0 | 0.0 | 1.7 | 1.2 | **37.9** | 0.0 | 0.0 | 26.0 | 0.1 | 17.3 |
| capsule scratch | 0.1 | 12.8 | 1.0 | 0.0 | 22.9 | 34.6 | 17.5 | 0.0 | 0.1 | 34.8 | 0.0 | **59.8** |
| carpet cut | 2.2 | 66.9 | 52.3 | 50.4 | 2.0 | 29.2 | 42.1 | 0.3 | 0.1 | 67.4 | 0.1 | **100.0** |
| hazelnut hole | 13.5 | 12.3 | 32.2 | 2.0 | 41.7 | 30.2 | 7.2 | 0.0 | 0.1 | 14.7 | 0.5 | **68.4** |
| leather cut | 3.8 | 27.3 | 19.1 | 23.6 | 25.8 | 3.1 | 2.1 | 0.0 | 0.4 | **37.9** | 0.2 | 32.1 |
| leather fold | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 2.1 | 0.0 | 0.1 | **20.3** | 0.0 | 9.7 |
| metal nut bent | 0.8 | 2.8 | 0.7 | 16.8 | 5.4 | **27.0** | 0.9 | 3.1 | 0.2 | 25.7 | 0.1 | 23.2 |
| metal nut scratch | 0.0 | 0.0 | 0.1 | 0.4 | 0.0 | 1.1 | 0.0 | 0.4 | 0.0 | 0.4 | 0.7 | **10.0** |
| pill scratch | 0.0 | 0.1 | 0.8 | 1.4 | 0.0 | 38.4 | **65.3** | 0.0 | 0.1 | 35.5 | 1.3 | 53.7 |
| tile oil spot | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 | **74.4** | 0.1 | 22.8 |
| wood combined defect | 3.0 | 17.9 | 17.2 | 24.1 | 0.1 | 19.9 | 13.4 | 0.0 | 0.0 | 10.7 | 0.0 | **32.0** |
| zipper broken teeth | 0.0 | 3.5 | 26.3 | 0.0 | 0.0 | 21.6 | 10.5 | **38.7** | 0.1 | 21.1 | 0.7 | 28.5 |
| zipper rough | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 32.6 | 1.5 | 4.1 | 0.0 | **78.0** | 1.4 | 76.3 |
| zipper squeezed teeth | 0.0 | 0.7 | 0.2 | 25.7 | 0.0 | 0.2 | 9.2 | **25.9** | 0.4 | 18.9 | 0.1 | 14.6 |
| audio jack contamination | 20.3 | 13.7 | 71.6 | **81.7** | 39.7 | 48.8 | 79.8 | 8.7 | 73.0 | 72.8 | 0.0 | 76.6 |
| bottle cap pit | 22.8 | 41.0 | 65.6 | 82.2 | 84.3 | 87.4 | 78.0 | 60.3 | 2.6 | **87.6** | 1.5 | 85.9 |
| button battery contamination | 34.9 | 54.5 | 56.7 | 57.0 | 32.4 | 43.2 | 70.7 | 61.3 | 28.6 | 43.0 | 0.1 | **71.3** |
| end cap damage | 2.7 | 9.4 | **31.7** | 19.2 | 0.0 | 18.8 | 10.6 | 4.5 | 8.0 | 15.6 | 0.1 | 22.6 |
| end cap missing parts | 3.5 | 1.4 | 3.5 | 16.9 | 14.2 | 26.0 | 14.7 | 9.0 | 11.9 | 25.9 | 0.0 | **40.9** |
| eraser scratch | 59.3 | 78.5 | 10.2 | 35.6 | 86.9 | 78.2 | 88.6 | 5.0 | 1.3 | 75.0 | 8.4 | **92.9** |
| mint foreign objects | 2.3 | 3.4 | 57.7 | 42.2 | 5.2 | 59.2 | 57.0 | 1.3 | 18.2 | **75.4** | 0.0 | 75.1 |
| mounts contamination | 0.3 | 4.2 | 7.8 | 21.3 | 11.0 | 6.7 | 0.9 | 0.0 | 0.8 | **79.5** | 0.3 | 75.7 |
| mounts missing parts | 0.5 | 4.4 | 33.6 | 2.4 | 22.7 | 25.1 | 1.2 | 0.3 | 1.0 | 11.8 | 1.4 | **38.3** |
| mounts pit | 0.0 | 0.0 | 0.0 | 0.0 | 2.2 | 2.5 | 0.0 | 0.0 | 0.8 | 4.3 | 0.1 | **4.6** |
| pcb missing parts | 0.9 | 8.2 | 12.0 | **20.5** | 4.1 | 8.8 | 0.5 | 0.1 | 0.1 | 6.0 | 0.0 | 5.7 |
| phone battery scratch | 3.8 | 55.6 | **59.4** | 55.8 | 27.1 | 50.5 | 44.2 | 4.7 | 1.7 | 31.0 | 0.2 | 44.4 |
| rolled strip base missing parts | 38.5 | 28.6 | 50.4 | 42.6 | 56.4 | 44.7 | **67.3** | 23.4 | 2.4 | 40.6 | 3.2 | 58.2 |
| tape missing parts | 24.7 | 67.3 | 63.5 | 47.6 | 28.4 | 42.2 | 72.0 | **90.4** | 2.8 | 79.6 | 0.3 | 80.6 |
| toothbrush abrasion | 4.7 | 22.9 | 28.0 | 26.9 | 7.4 | 18.3 | 10.6 | 0.0 | 10.1 | 14.6 | 0.0 | **28.3** |
| toothbrush contamination | 1.3 | 4.6 | 4.5 | 6.5 | 1.3 | 4.1 | 6.4 | 7.7 | 5.7 | **8.8** | 0.0 | 6.1 |
| toy contamination | 6.8 | 41.2 | 55.3 | 54.4 | 10.7 | 54.1 | 73.4 | 71.6 | 51.4 | 71.5 | 0.2 | **81.9** |
| toy brick scratch | 17.1 | 52.0 | 61.3 | 45.2 | 29.6 | 54.6 | **64.8** | 43.9 | 0.5 | 6.6 | 0.1 | 48.7 |
| transistor missing parts | 0.0 | 0.1 | 0.1 | 0.3 | 0.4 | 0.1 | 0.0 | 0.0 | 0.0 | 0.4 | 0.0 | **1.6** |
| u block abrasion | 0.0 | 59.1 | 60.0 | 51.0 | 39.6 | **80.4** | 20.2 | 37.7 | 1.3 | 8.6 | 0.8 | 65.6 |
| u block contamination | 15.7 | 17.4 | 62.2 | 72.6 | 2.6 | 12.5 | 65.3 | 71.1 | 0.7 | **81.5** | 0.0 | 81.3 |
| usb adapter contamination | 0.3 | 18.4 | 43.1 | **51.0** | 10.4 | 50.6 | 3.2 | 17.1 | 0.9 | 49.3 | 0.1 | 48.2 |
| usb adapter scratch | 1.6 | 38.9 | 65.9 | **76.4** | 57.7 | 64.4 | 5.8 | 73.3 | 0.1 | 37.9 | 0.0 | 64.6 |
| vcpill contamination | 49.0 | **75.1** | 68.3 | 74.1 | 48.8 | 69.4 | 38.3 | 29.8 | 4.3 | 39.3 | 0.0 | 54.7 |
| vcpill missing parts | 5.1 | 51.8 | 43.1 | 41.0 | 37.8 | 57.1 | 38.8 | 20.8 | 33.2 | 62.9 | 1.2 | **64.0** |
| wooden beads contamination | 51.2 | 78.1 | 89.8 | **91.7** | 77.0 | 84.3 | 88.5 | 88.2 | 11.2 | 85.1 | 0.1 | 90.7 |
| zipper damage | 0.2 | 6.8 | 4.0 | 4.1 | 3.1 | 2.5 | 0.6 | 0.2 | 2.5 | **11.7** | 0.2 | 9.5 |
| zipper deformation | 0.4 | 0.3 | 0.0 | 1.6 | 0.0 | 2.2 | 0.0 | 0.5 | 3.7 | 7.0 | 0.0 | **9.8** |

**Results and discussion.** At the permissive IoU threshold, OMQ reaches 43.57 $AP_{50}$, exceeding DINOv by 7.24 points and UniSpector's 5.60 $AP_{50}$ by 37.97 points. It improves over DINOv by at least five points on 25 categories, with the strongest average gains on Real-IAD (+10.51) and MVTec AD (+7.32). Carpet cut reaches 100.0 and u block abrasion gains 57.0 points, demonstrating that target-grounded morphology substantially improves unseen-defect discovery and coarse placement. UniSpector's isolated 73.0 result on audio jack contamination contrasts with near-zero scores on most MVTec AD categories, showing that success on one visually salient morphology does not translate into stable category-disjoint transfer. The deficits on the three GC10-DET classes ($-20.69$ points on average) and tile oil spot ($-51.7$) already appear at IoU 0.50, which identifies missing or misdirected prompt–target correspondence rather than box refinement as the primary cause in these cases.

Table A6: Complete category-level bounding-box $AP_{75}$ (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 1.3 | 4.6 | 0.0 | **6.4** |
| steel surface crescent gap | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | **6.1** | 0.0 | 0.2 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **6.0** | 0.0 | 5.0 |
| magnetic tile blowhole | 0.7 | 2.7 | 3.8 | **7.1** | 0.0 | 1.1 | 1.9 | 0.5 | 0.0 | 3.8 | 0.0 | 6.5 |
| bottle broken small | 7.5 | 61.0 | 50.2 | 59.8 | 62.8 | **75.8** | 47.0 | 10.8 | 0.0 | 25.0 | 0.0 | 58.5 |
| cable bent wire | 0.0 | 2.0 | 0.0 | 0.0 | 0.1 | 6.7 | 0.6 | 0.0 | 0.0 | 38.8 | 0.0 | **58.0** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 0.2 | 4.9 | 3.6 | **14.4** | 4.0 | 0.0 | 2.2 | 6.2 | 0.0 | 0.2 | 0.0 | 7.1 |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.5 | 0.2 | 0.0 | 0.0 | 0.1 | 0.0 | **2.7** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 0.2 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 2.7 | 0.0 | **8.1** |
| capsule poke | 0.0 | 0.0 | 0.0 | 0.0 | 1.5 | 0.7 | **17.2** | 0.0 | 0.0 | 13.7 | 0.0 | 11.1 |
| capsule scratch | 0.0 | 1.7 | 0.4 | 0.0 | 11.8 | 20.6 | 14.9 | 0.0 | 0.1 | 17.5 | 0.0 | **46.3** |
| carpet cut | 0.0 | 28.0 | 18.2 | 22.2 | 0.0 | 17.8 | 17.8 | 0.0 | 0.0 | 60.6 | 0.0 | **91.1** |
| hazelnut hole | 9.1 | 11.7 | 19.4 | 2.0 | 34.3 | 22.6 | 2.1 | 0.0 | 0.0 | 14.5 | 0.3 | **62.5** |
| leather cut | 0.1 | 7.4 | 9.7 | 8.3 | 13.5 | 1.6 | 1.1 | 0.0 | 0.1 | 5.1 | 0.0 | **20.0** |
| leather fold | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.0 | 0.0 | **1.0** | 0.0 | 0.0 |
| metal nut bent | 0.2 | 0.9 | 0.7 | 16.4 | 5.1 | **21.8** | 0.7 | 2.4 | 0.1 | 11.3 | 0.0 | 18.0 |
| metal nut scratch | 0.0 | 0.0 | 0.1 | 0.3 | 0.0 | 1.0 | 0.0 | 0.4 | 0.0 | 0.0 | 0.1 | **5.8** |
| pill scratch | 0.0 | 0.1 | 0.7 | 0.8 | 0.0 | 26.1 | **57.5** | 0.0 | 0.0 | 17.9 | 1.0 | 37.4 |
| tile oil spot | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | **39.7** | 0.0 | 9.6 |
| wood combined defect | 3.0 | 12.1 | 12.9 | 17.0 | 0.1 | 13.1 | 8.0 | 0.0 | 0.0 | 9.9 | 0.0 | **24.0** |
| zipper broken teeth | 0.0 | 0.0 | 6.3 | 0.0 | 0.0 | 1.8 | 2.5 | 13.7 | 0.0 | 14.3 | 0.1 | **21.0** |
| zipper rough | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 21.9 | 0.9 | 2.3 | 0.0 | 71.8 | 0.3 | **72.2** |
| zipper squeezed teeth | 0.0 | 0.0 | 0.0 | 3.8 | 0.0 | 0.1 | 2.0 | 3.0 | 0.2 | **5.4** | 0.0 | 2.9 |
| audio jack contamination | 11.9 | 11.2 | 59.7 | **74.6** | 34.6 | 41.5 | 73.7 | 7.6 | 61.6 | 66.8 | 0.0 | 70.7 |
| bottle cap pit | 4.2 | 21.7 | 37.0 | 42.9 | **46.0** | 32.5 | 29.3 | 15.0 | 0.8 | 31.5 | 0.1 | 31.2 |
| button battery contamination | 15.2 | 38.8 | 41.7 | 41.8 | 21.8 | 30.0 | 56.6 | 48.2 | 24.6 | 34.4 | 0.0 | **58.2** |
| end cap damage | 0.2 | 5.5 | **13.9** | 8.6 | 0.0 | 10.4 | 6.3 | 1.7 | 2.7 | 2.1 | 0.0 | 11.2 |
| end cap missing parts | 0.2 | 0.8 | 1.4 | 8.0 | 6.9 | 16.2 | 7.9 | 3.2 | 6.0 | 16.0 | 0.0 | **28.0** |
| eraser scratch | 17.6 | 59.9 | 7.3 | 21.7 | 65.9 | 59.8 | **73.4** | 3.7 | 1.0 | 49.8 | 5.6 | 69.8 |
| mint foreign objects | 0.1 | 1.7 | 36.6 | 26.5 | 3.1 | 40.6 | 36.8 | 0.6 | 9.2 | **53.2** | 0.0 | 51.6 |
| mounts contamination | 0.1 | 3.8 | 5.9 | 17.1 | 9.4 | 5.2 | 0.8 | 0.0 | 0.5 | **72.5** | 0.1 | 67.2 |
| mounts missing parts | 0.0 | 3.1 | 25.9 | 2.0 | 18.6 | 22.6 | 1.0 | 0.2 | 0.7 | 10.4 | 0.3 | **35.7** |
| mounts pit | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.6 | 0.0 | 0.0 | 0.3 | 1.5 | 0.1 | **1.6** |
| pcb missing parts | 0.0 | 1.6 | 2.7 | **3.6** | 1.2 | 1.5 | 0.1 | 0.0 | 0.1 | 1.6 | 0.0 | 0.3 |
| phone battery scratch | 1.3 | 39.3 | **40.8** | 39.6 | 16.3 | 37.6 | 33.7 | 3.1 | 0.7 | 19.1 | 0.0 | 27.1 |
| rolled strip base missing parts | 2.7 | 15.2 | 31.7 | 28.8 | 35.5 | 33.0 | **44.5** | 14.3 | 1.1 | 22.7 | 0.7 | 35.1 |
| tape missing parts | 2.5 | 42.5 | 43.7 | 34.4 | 19.1 | 31.9 | 57.3 | **69.2** | 1.6 | 60.7 | 0.1 | 61.5 |
| toothbrush abrasion | 0.1 | 6.9 | 10.5 | **12.2** | 3.6 | 8.6 | 3.7 | 0.0 | 3.0 | 3.1 | 0.0 | 8.2 |
| toothbrush contamination | 1.0 | 2.4 | 1.9 | 3.6 | 0.8 | 2.8 | **3.9** | 2.5 | 1.3 | 3.2 | 0.0 | 1.4 |
| toy contamination | 3.8 | 31.0 | 48.9 | 48.8 | 8.2 | 47.6 | 65.4 | 62.4 | 42.7 | 65.0 | 0.1 | **70.0** |
| toy brick scratch | 7.1 | 40.6 | 49.8 | 33.1 | 22.5 | 44.4 | **56.7** | 31.7 | 0.2 | 3.2 | 0.0 | 38.4 |
| transistor missing parts | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **0.2** |
| u block abrasion | 0.0 | 44.8 | 47.1 | 42.0 | 31.0 | **65.6** | 18.9 | 29.4 | 0.4 | 6.8 | 0.4 | 57.7 |
| u block contamination | 11.6 | 16.3 | 58.9 | 69.2 | 2.5 | 12.0 | 62.1 | 66.7 | 0.6 | 78.8 | 0.0 | **78.9** |
| usb adapter contamination | 0.1 | 9.1 | 23.6 | 25.5 | 5.8 | **29.2** | 1.4 | 11.3 | 0.3 | 26.3 | 0.0 | 29.1 |
| usb adapter scratch | 0.3 | 29.0 | 53.1 | **58.4** | 51.0 | 55.7 | 3.5 | 56.9 | 0.0 | 29.3 | 0.0 | 51.3 |
| vcpill contamination | 29.9 | 48.1 | 42.0 | **50.8** | 34.6 | 46.9 | 28.0 | 20.3 | 2.3 | 27.7 | 0.0 | 44.1 |
| vcpill missing parts | 0.3 | 27.4 | 24.2 | 22.9 | 20.4 | 35.1 | 25.2 | 11.5 | 20.0 | **42.3** | 0.6 | **42.3** |
| wooden beads contamination | 26.0 | 59.7 | 75.6 | 76.5 | 68.0 | 70.4 | **80.5** | 75.0 | 8.1 | 74.6 | 0.0 | 80.4 |
| zipper damage | 0.0 | 0.2 | 0.1 | 0.2 | 0.3 | 0.3 | 0.1 | 0.0 | 0.0 | **0.5** | 0.0 | 0.3 |
| zipper deformation | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | **0.1** | 0.0 | 0.0 |

**Results and discussion.** Under strict localization, OMQ obtains 31.27 $AP_{75}$, exceeding DINOv by 8.70 points and UniSpector's 3.69 $AP_{75}$ by 27.58 points. Its margin over DINOv widens from 7.24 points at $AP_{50}$ to 8.70 points at $AP_{75}$, indicating that target-conditioned transport improves spatial agreement after relevant instances are found. OMQ improves on DINOv for 37 categories, ranks in the top two for 38, and loses at least five points in only three cases. The gains for hazelnut hole (+48.0), u block abrasion (+50.9), and carpet cut (+30.5) therefore reflect more accurate box placement rather than only stronger discovery. UniSpector reaches 61.6 on audio jack contamination but remains weak at the macro level, revealing that precise localization on isolated classes is insufficient without category-stable morphology transfer. The principal exception is tile oil spot ($-30.1$), where target grounding cannot compensate for a prompt prototype that does not isolate the relevant appearance change.

Table A7: Complete category-level bounding-box $AR_{100}$ (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.4 | 0.6 | 0.0 | 0.0 | 3.1 | 0.3 | 23.6 | 27.7 | 30.5 | **39.5** | 5.6 | 33.5 |
| steel surface crescent gap | 0.0 | 0.0 | 4.6 | 0.0 | 3.4 | 0.0 | 32.3 | 0.0 | 12.8 | **49.8** | 3.7 | 44.7 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 12.3 | 21.3 | 5.2 | 39.8 | 0.6 | **43.8** |
| magnetic tile blowhole | 14.3 | 25.7 | 37.9 | 20.9 | 4.3 | 39.5 | **62.8** | 17.4 | 0.0 | 35.8 | 23.2 | 42.5 |
| bottle broken small | 33.5 | 61.2 | 52.7 | 63.1 | 61.2 | 70.8 | **81.9** | 69.2 | 58.5 | 69.0 | 28.1 | 72.5 |
| cable bent wire | 47.5 | 10.0 | 0.0 | 0.0 | 11.9 | 49.4 | **75.0** | 58.8 | 70.0 | 70.0 | 38.8 | 73.6 |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 37.5 | 0.0 | 0.0 | 25.7 | 0.0 | **40.0** |
| cable combined defect | 19.5 | 18.0 | 16.3 | 15.8 | 4.8 | 0.0 | 49.0 | 29.8 | 25.0 | 60.0 | 3.0 | **64.2** |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | 39.6 | 19.1 | 50.0 | 0.9 | 28.9 | 47.8 | 6.1 | **53.3** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 2.7 | 22.3 | 0.0 | 0.0 | 38.2 | 26.5 | 33.5 | 1.8 | **41.2** |
| capsule poke | 3.6 | 0.0 | 0.0 | 0.0 | 9.5 | 21.8 | **72.7** | 0.0 | 28.2 | 53.5 | 19.1 | 56.5 |
| capsule scratch | 10.8 | 21.3 | 8.8 | 0.0 | 54.6 | 54.2 | **81.7** | 17.5 | 60.0 | 70.0 | 2.5 | 73.2 |
| carpet cut | 21.7 | 55.0 | 58.3 | 57.2 | 1.7 | 32.8 | 73.3 | 54.4 | 55.0 | 73.3 | 52.8 | **76.7** |
| hazelnut hole | 50.0 | 54.5 | 55.5 | 37.5 | 62.0 | 63.0 | 81.0 | 4.5 | 66.9 | **82.3** | 43.0 | **82.3** |
| leather cut | 57.0 | 41.0 | 48.0 | 37.5 | 27.0 | 13.5 | **73.0** | 5.0 | 52.9 | 52.9 | 43.0 | 60.0 |
| leather fold | 42.4 | 0.0 | 0.0 | 0.0 | 0.0 | 5.3 | **54.7** | 18.8 | 18.3 | 37.5 | 20.0 | 35.0 |
| metal nut bent | 34.8 | 40.6 | 41.6 | 48.7 | 32.9 | 54.8 | 78.7 | 63.2 | 61.5 | 64.6 | 29.0 | **78.8** |
| metal nut scratch | 25.8 | 1.8 | 11.0 | 14.5 | 2.0 | 30.2 | 24.5 | 50.8 | 44.2 | 45.2 | 16.0 | **62.3** |
| pill scratch | 20.0 | 24.8 | 18.0 | 5.2 | 10.4 | 62.8 | **83.2** | 16.8 | 66.5 | 75.0 | 37.2 | 76.5 |
| tile oil spot | 53.9 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 55.6 | 62.2 | 51.5 | 66.9 | **75.6** | 68.5 |
| wood combined defect | 13.9 | 43.2 | 30.5 | 23.9 | 22.0 | 50.7 | 69.3 | 25.6 | 2.9 | 63.6 | 2.4 | **73.2** |
| zipper broken teeth | 0.0 | 11.1 | 32.6 | 0.0 | 1.5 | 30.0 | **66.3** | 39.3 | 46.8 | 44.7 | 31.5 | 50.5 |
| zipper rough | 14.0 | 0.0 | 0.0 | 0.0 | 3.6 | 74.0 | 68.8 | 67.6 | 48.8 | **75.6** | 47.6 | 71.3 |
| zipper squeezed teeth | 5.9 | 0.6 | 10.0 | 32.9 | 0.0 | 5.3 | **62.4** | 48.8 | 44.2 | 50.8 | 23.5 | 55.0 |
| audio jack contamination | 46.3 | 26.7 | 72.0 | 74.1 | 45.0 | 70.3 | **82.4** | 79.3 | 74.1 | 80.6 | 3.8 | 80.2 |
| bottle cap pit | 35.0 | 57.3 | 54.6 | 55.2 | 58.8 | 54.6 | **68.1** | 50.7 | 53.9 | 54.9 | 35.4 | 60.3 |
| button battery contamination | 37.8 | 57.2 | 54.5 | 54.7 | 40.5 | 54.5 | **73.7** | 64.4 | 68.3 | 68.5 | 8.3 | 70.2 |
| end cap damage | 10.6 | 40.3 | 40.9 | 25.4 | 27.0 | 47.4 | **67.2** | 53.8 | 48.7 | 56.8 | 6.6 | 61.8 |
| end cap missing parts | 19.7 | 4.3 | 17.0 | 34.9 | 37.1 | 49.0 | **70.3** | 56.1 | 51.7 | 62.7 | 2.8 | 65.8 |
| eraser scratch | 58.2 | 69.9 | 42.1 | 60.7 | 69.5 | 71.1 | **78.7** | 70.6 | 68.4 | 72.6 | 49.4 | 74.0 |
| mint foreign objects | 14.0 | 13.2 | 53.5 | 49.9 | 41.1 | 57.5 | 64.7 | 57.6 | 56.8 | 65.5 | 0.1 | **66.3** |
| mounts contamination | 15.4 | 12.7 | 6.9 | 25.6 | 41.7 | 32.1 | **75.7** | 8.6 | 69.6 | 73.2 | 30.4 | 72.8 |
| mounts missing parts | 18.4 | 5.4 | 29.0 | 3.5 | 63.8 | 67.3 | 66.6 | 65.4 | 65.3 | **77.1** | 34.8 | 74.4 |
| mounts pit | 15.6 | 0.2 | 0.0 | 0.0 | 27.9 | 30.0 | 47.5 | 0.0 | 40.5 | 49.8 | 12.2 | **54.2** |
| pcb missing parts | 6.2 | 13.0 | 17.7 | 22.3 | 18.4 | 19.7 | 22.1 | 7.4 | 13.3 | 31.4 | 1.1 | **32.9** |
| phone battery scratch | 27.1 | 58.3 | 60.0 | 60.1 | 50.0 | 50.6 | **64.1** | 52.7 | 57.1 | 59.2 | 24.4 | 59.9 |
| rolled strip base missing parts | 26.3 | 58.8 | 57.1 | 59.5 | 56.5 | 58.6 | **70.0** | 63.1 | 55.9 | 61.8 | 41.1 | 65.9 |
| tape missing parts | 25.9 | 60.7 | 55.2 | 44.8 | 62.4 | 65.1 | **75.1** | 72.4 | 68.1 | 73.5 | 42.1 | 72.7 |
| toothbrush abrasion | 11.8 | 29.8 | 32.1 | 31.2 | 14.9 | 30.0 | **51.0** | 12.0 | 34.8 | 43.3 | 0.8 | 43.1 |
| toothbrush contamination | 11.0 | 17.7 | 10.0 | 11.7 | 9.1 | 14.3 | **45.7** | 28.4 | 26.9 | 33.9 | 0.1 | 32.5 |
| toy contamination | 35.3 | 45.5 | 46.5 | 51.7 | 57.3 | 64.2 | **76.0** | 69.4 | 68.9 | 72.0 | 23.3 | 71.0 |
| toy brick scratch | 41.4 | 60.8 | 59.2 | 58.6 | 51.8 | 60.5 | **71.1** | 59.5 | 39.9 | 60.6 | 13.1 | 63.4 |
| transistor missing parts | 0.8 | 1.5 | 0.9 | 3.4 | 3.7 | 4.7 | 0.0 | 5.3 | 3.2 | **7.5** | 0.0 | **7.5** |
| u block abrasion | 0.0 | 64.1 | 57.5 | 65.1 | 66.2 | 71.9 | **79.0** | 70.3 | 66.9 | 75.2 | 32.6 | 73.1 |
| u block contamination | 47.7 | 65.1 | 78.6 | 77.1 | 66.4 | 63.3 | **83.0** | 80.7 | 78.9 | 82.1 | 20.8 | 82.7 |
| usb adapter contamination | 3.9 | 40.7 | 43.8 | 42.8 | 17.4 | 43.9 | 57.1 | 54.0 | 45.6 | 59.6 | 3.7 | **61.6** |
| usb adapter scratch | 10.3 | 65.1 | 65.6 | 69.1 | 48.2 | 59.2 | **72.5** | 69.6 | 19.1 | 69.7 | 3.9 | 69.7 |
| vcpill contamination | 40.8 | 56.5 | 50.4 | 57.2 | 55.4 | 56.9 | **72.7** | 67.3 | 63.0 | 68.9 | 0.8 | 71.7 |
| vcpill missing parts | 18.2 | 47.6 | 41.3 | 42.7 | 42.1 | 50.3 | **70.4** | 54.8 | 53.8 | 61.4 | 32.6 | 60.9 |
| wooden beads contamination | 52.2 | 73.4 | 76.7 | 76.5 | 72.3 | 77.5 | **81.6** | 80.6 | 77.3 | 79.7 | 21.3 | 81.0 |
| zipper damage | 15.0 | 18.1 | 12.4 | 13.3 | 6.6 | 7.6 | 23.7 | 9.5 | 16.9 | **25.4** | 17.7 | 21.0 |
| zipper deformation | 9.4 | 2.3 | 0.1 | 4.3 | 0.3 | 9.4 | 0.0 | **23.3** | 5.9 | 5.2 | 2.6 | 7.9 |

**Results and discussion.** OMQ records 59.37 $AR_{100}$, improving over DINOv by 2.47 points and over UniSpector's 44.19 recall by 15.19 points, while ranking in the top two on 39 categories. UniSpector retains nontrivial recall on MVTec AD (42.83) and Real-IAD (49.75) despite its 3.47 box AP. This large recall–precision gap indicates that its predictions often reach relevant instances but are not ranked, localized, or filtered reliably under the benchmark's defect-sparse setting. Grounding DINO exhibits the same distinction from another direction, with 59.22 recall that nearly matches OMQ but only 16.34 AP. OMQ's main benefit therefore lies in the quality and spatial agreement of retrieved instances rather than the number of emitted candidates. Its recall gains over DINOv concentrate on MVTec AD (+5.12), whereas Real-IAD is closer to saturation (+0.95), leaving rare low-signal categories as the main coverage bottleneck.

Table A8: Complete category-level pixel grounding AP (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 2.7 | **7.5** | 0.0 | 7.4 |
| steel surface crescent gap | 0.0 | 0.0 | 0.9 | 0.0 | – | – | – | 0.0 | 0.1 | **7.2** | 0.0 | 0.3 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **12.0** | 0.0 | 6.1 |
| magnetic tile blowhole | 0.2 | 0.6 | 0.6 | 1.7 | – | – | – | 0.8 | 0.0 | 8.5 | 0.0 | **10.5** |
| bottle broken small | 10.4 | 52.9 | 42.3 | **53.2** | – | – | – | 8.5 | 0.0 | 23.2 | 0.3 | 48.9 |
| cable bent wire | 0.0 | 0.4 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 19.8 | 0.0 | **33.6** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 0.2 | 2.9 | 3.8 | **10.6** | – | – | – | 5.0 | 0.0 | 0.2 | 0.0 | 4.7 |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 1.6 | 0.0 | **2.1** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 0.1 | – | – | – | 0.0 | 0.0 | 5.1 | 0.0 | **6.8** |
| capsule poke | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 9.3 | 0.1 | **9.6** |
| capsule scratch | 0.0 | 0.5 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 12.1 | 0.0 | **26.7** |
| carpet cut | 1.0 | 32.7 | 22.7 | 24.7 | – | – | – | 0.1 | 0.0 | 39.7 | 0.0 | **61.4** |
| hazelnut hole | 21.4 | 9.9 | 25.5 | 1.5 | – | – | – | 0.0 | 0.1 | 9.2 | 0.4 | **57.7** |
| leather cut | 0.5 | 6.9 | 4.0 | **10.0** | – | – | – | 0.0 | 0.1 | 5.4 | 0.0 | 7.0 |
| leather fold | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **7.2** | 0.0 | 4.7 |
| metal nut bent | 0.2 | 1.0 | 0.4 | 12.6 | – | – | – | 2.0 | 0.1 | 11.9 | 0.0 | **16.3** |
| metal nut scratch | 0.0 | 0.0 | 0.1 | 0.3 | – | – | – | 0.2 | 0.0 | 0.1 | 0.5 | **4.8** |
| pill scratch | 0.0 | 0.1 | 0.5 | 0.9 | – | – | – | 0.0 | 0.0 | 20.3 | 0.8 | **33.8** |
| tile oil spot | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **12.8** | 0.0 | 5.7 |
| wood combined defect | 2.1 | 10.0 | 9.2 | 12.4 | – | – | – | 0.0 | 0.0 | 7.8 | 0.0 | **20.9** |
| zipper broken teeth | 0.0 | 0.3 | 5.6 | 0.0 | – | – | – | 8.6 | 0.0 | 9.1 | 0.2 | **15.6** |
| zipper rough | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 1.9 | 0.0 | **52.0** | 0.4 | 50.0 |
| zipper squeezed teeth | 0.0 | 0.1 | 0.1 | **9.3** | – | – | – | 7.5 | 0.2 | 5.7 | 0.0 | 5.2 |
| audio jack contamination | 12.8 | 6.4 | 35.0 | 38.5 | – | – | – | 6.0 | 52.9 | 56.4 | 0.0 | **58.8** |
| bottle cap pit | 9.2 | 7.9 | 9.9 | 10.0 | – | – | – | 24.8 | 0.7 | **38.7** | 0.5 | 35.2 |
| button battery contamination | 16.9 | 28.8 | 32.0 | 31.3 | – | – | – | 37.8 | 20.4 | 27.4 | 0.0 | **45.5** |
| end cap damage | 0.3 | 1.2 | 3.4 | 2.5 | – | – | – | 1.1 | 1.6 | 1.5 | 0.0 | **7.8** |
| end cap missing parts | 0.5 | 0.4 | 0.5 | 2.1 | – | – | – | 1.9 | 3.0 | 8.6 | 0.0 | **13.6** |
| eraser scratch | 7.3 | 4.8 | 1.1 | 2.3 | – | – | – | 0.6 | 0.3 | 16.9 | 0.9 | **20.8** |
| mint foreign objects | 0.0 | 0.0 | 0.0 | 0.1 | – | – | – | 0.1 | 0.1 | **8.1** | 0.0 | **8.1** |
| mounts contamination | 0.2 | 2.2 | 3.7 | 7.6 | – | – | – | 0.0 | 0.4 | **56.7** | 0.1 | 54.8 |
| mounts missing parts | 0.1 | 2.6 | 17.4 | 1.7 | – | – | – | 0.2 | 0.3 | 6.5 | 0.6 | **21.9** |
| mounts pit | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.4 | **1.9** | 0.1 | **1.9** |
| pcb missing parts | 0.1 | 1.4 | 2.6 | **5.1** | – | – | – | 0.0 | 0.0 | 1.5 | 0.0 | 1.3 |
| phone battery scratch | 0.1 | 4.9 | 6.4 | 5.9 | – | – | – | 0.8 | 0.2 | 4.3 | 0.0 | **7.3** |
| rolled strip base missing parts | 9.7 | 10.4 | 18.8 | 16.0 | – | – | – | 10.0 | 0.9 | 18.5 | 0.7 | **27.2** |
| tape missing parts | 2.7 | 24.6 | 24.7 | 17.6 | – | – | – | **49.4** | 1.2 | 46.1 | 0.1 | 46.6 |
| toothbrush abrasion | 0.6 | 1.8 | **3.0** | 2.8 | – | – | – | 0.0 | 1.0 | 1.0 | 0.0 | 2.2 |
| toothbrush contamination | 0.5 | 1.9 | 1.4 | 2.6 | – | – | – | **3.2** | 2.0 | **3.2** | 0.0 | 1.6 |
| toy contamination | 4.9 | 16.9 | 22.9 | 21.9 | – | – | – | 48.8 | 31.3 | 49.7 | 0.1 | **55.8** |
| toy brick scratch | 2.7 | 3.4 | 4.5 | 1.8 | – | – | – | 4.7 | 0.0 | 0.6 | 0.0 | **7.0** |
| transistor missing parts | 0.0 | 0.0 | 0.0 | 0.1 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | **0.2** |
| u block abrasion | 0.0 | 27.5 | 28.4 | 24.6 | – | – | – | 21.5 | 0.4 | 4.8 | 0.5 | **40.6** |
| u block contamination | 10.2 | 7.5 | 26.0 | 30.4 | – | – | – | 53.7 | 0.4 | 63.3 | 0.0 | **63.8** |
| usb adapter contamination | 0.3 | 4.7 | 10.0 | 11.5 | – | – | – | 8.4 | 0.3 | 21.4 | 0.0 | **22.3** |
| usb adapter scratch | 0.3 | 1.0 | 2.9 | 2.6 | – | – | – | 13.8 | 0.0 | 7.1 | 0.0 | **14.1** |
| vcpill contamination | 28.2 | 20.9 | 18.5 | 20.8 | – | – | – | 13.9 | 1.8 | 18.4 | 0.0 | **29.0** |
| vcpill missing parts | 1.2 | 18.7 | 16.0 | 15.1 | – | – | – | 9.6 | 16.1 | 31.5 | 0.7 | **32.2** |
| wooden beads contamination | 23.0 | 27.3 | 35.4 | 35.7 | – | – | – | 57.7 | 4.5 | 55.3 | 0.0 | **58.1** |
| zipper damage | 0.0 | 0.2 | 0.2 | 0.2 | – | – | – | 0.0 | 0.1 | **0.6** | 0.0 | 0.5 |
| zipper deformation | 0.0 | 0.2 | 0.0 | 0.7 | – | – | – | 0.1 | 0.5 | **2.0** | 0.0 | 1.4 |

**Results and discussion.** For pixel-level localization, OMQ achieves 21.52 macro AP, exceeding DINOv by 5.37 points and UniSpector's 2.77 AP by 18.76 points. It obtains the best result (including ties) on 34 of 52 categories and ranks in the top two on 45. UniSpector reaches 52.9 on audio jack contamination and 31.3 on toy contamination but remains near zero across most MVTec AD categories. This concentration on a few salient cases shows that transferable mask quality is not secured by occasional strong category responses. OMQ instead improves on DINOv for 38 categories, including hazelnut hole (+48.5), u block abrasion (+35.8), and bottle broken small (+25.7), which confirms that morphology-conditioned queries support pixel-level delineation as well as box discovery. Its mask AP remains 7.03 points below box AP, and the negative GC10-DET mean change ($-4.33$) identifies exact recovery of thin or weakly contrasted surface patterns as the remaining cross-domain limitation.

Table A9: Complete category-level pixel grounding $AP_{50}$ (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.0 | 0.1 | 0.0 | 0.0 | – | – | – | 0.1 | 7.2 | **21.0** | 0.0 | 16.3 |
| steel surface crescent gap | 0.0 | 0.0 | 3.1 | 0.0 | – | – | – | 0.0 | 0.3 | **27.8** | 0.0 | 1.2 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.1 | 0.2 | **45.0** | 0.0 | 20.6 |
| magnetic tile blowhole | 1.0 | 4.2 | 3.4 | 10.2 | – | – | – | 5.0 | 0.0 | 42.9 | 0.1 | **44.9** |
| bottle broken small | 39.5 | **78.3** | 62.3 | 77.9 | – | – | – | 10.8 | 0.0 | 37.1 | 0.4 | 70.0 |
| cable bent wire | 0.0 | 2.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 38.9 | 0.0 | **58.0** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 0.6 | 6.5 | 8.6 | **15.1** | – | – | – | 6.6 | 0.0 | 0.5 | 0.0 | 7.1 |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 2.9 | 0.0 | **4.5** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 0.6 | – | – | – | 0.0 | 0.3 | 19.4 | 0.0 | **25.2** |
| capsule poke | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **16.1** | 0.2 | 15.4 |
| capsule scratch | 0.0 | 1.4 | 0.2 | 0.0 | – | – | – | 0.0 | 0.0 | 28.7 | 0.0 | **50.7** |
| carpet cut | 2.4 | 66.9 | 52.3 | 50.4 | – | – | – | 0.2 | 0.1 | 67.4 | 0.1 | **100.0** |
| hazelnut hole | 32.1 | 13.9 | 36.3 | 2.0 | – | – | – | 0.0 | 0.1 | 16.5 | 0.7 | **69.2** |
| leather cut | 2.4 | 18.1 | 20.2 | 24.5 | – | – | – | 0.0 | 0.2 | **25.6** | 0.1 | 24.2 |
| leather fold | 0.2 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **18.6** | 0.0 | 9.8 |
| metal nut bent | 0.6 | 2.3 | 0.7 | 17.3 | – | – | – | 3.2 | 0.2 | **25.8** | 0.0 | 24.7 |
| metal nut scratch | 0.1 | 0.0 | 0.1 | 0.3 | – | – | – | 0.4 | 0.0 | 0.4 | 2.5 | **9.1** |
| pill scratch | 0.1 | 0.1 | 0.8 | 1.4 | – | – | – | 0.0 | 0.1 | 35.5 | 1.4 | **53.7** |
| tile oil spot | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **24.3** | 0.1 | 9.8 |
| wood combined defect | 3.0 | 19.7 | 17.2 | 23.7 | – | – | – | 0.0 | 0.0 | 11.3 | 0.0 | **32.0** |
| zipper broken teeth | 0.0 | 1.9 | 19.5 | 0.0 | – | – | – | **34.3** | 0.1 | 27.3 | 0.7 | **34.3** |
| zipper rough | 0.1 | 0.0 | 0.0 | 0.0 | – | – | – | 3.2 | 0.0 | **78.0** | 1.2 | 76.3 |
| zipper squeezed teeth | 0.0 | 0.7 | 0.2 | **25.8** | – | – | – | 17.8 | 0.4 | 15.8 | 0.1 | 12.1 |
| audio jack contamination | 21.7 | 13.4 | 67.7 | 75.9 | – | – | – | 8.6 | 72.9 | 72.7 | 0.0 | **76.6** |
| bottle cap pit | 28.3 | 29.5 | 43.7 | 45.8 | – | – | – | 60.3 | 2.2 | **87.4** | 1.6 | 85.4 |
| button battery contamination | 36.0 | 53.5 | 55.6 | 55.5 | – | – | – | 60.1 | 28.8 | 44.9 | 0.1 | **69.7** |
| end cap damage | 1.5 | 5.1 | 13.6 | 8.7 | – | – | – | 3.6 | 5.1 | 5.5 | 0.1 | **17.3** |
| end cap missing parts | 2.2 | 0.8 | 1.7 | 8.8 | – | – | – | 5.8 | 9.9 | 23.4 | 0.0 | **37.3** |
| eraser scratch | 34.0 | 24.0 | 4.4 | 11.2 | – | – | – | 2.5 | 0.9 | 53.9 | 4.1 | **69.6** |
| mint foreign objects | 0.1 | 0.0 | 0.1 | 0.3 | – | – | – | 0.5 | 0.5 | 36.7 | 0.0 | **37.3** |
| mounts contamination | 0.4 | 4.2 | 7.1 | 17.3 | – | – | – | 0.0 | 0.8 | **79.9** | 0.4 | 75.4 |
| mounts missing parts | 0.5 | 4.4 | 33.1 | 2.4 | – | – | – | 0.3 | 0.8 | 11.1 | 1.5 | **37.2** |
| mounts pit | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.8 | 4.3 | 0.2 | **4.7** |
| pcb missing parts | 0.9 | 4.2 | 7.2 | **14.8** | – | – | – | 0.1 | 0.1 | 4.9 | 0.0 | 4.2 |
| phone battery scratch | 0.5 | 23.9 | **32.4** | 27.3 | – | – | – | 3.0 | 0.6 | 15.7 | 0.1 | 30.8 |
| rolled strip base missing parts | 35.4 | 26.9 | 46.8 | 39.8 | – | – | – | 21.7 | 2.1 | 39.0 | 2.7 | **55.2** |
| tape missing parts | 7.3 | 55.1 | 53.7 | 38.1 | – | – | – | **85.0** | 2.7 | 79.8 | 0.3 | 78.9 |
| toothbrush abrasion | 2.0 | 7.9 | **13.0** | 12.4 | – | – | – | 0.0 | 4.8 | 4.0 | 0.0 | 10.5 |
| toothbrush contamination | 1.4 | 4.2 | 4.3 | 6.6 | – | – | – | **7.6** | 5.6 | **7.6** | 0.0 | 4.9 |
| toy contamination | 10.0 | 39.7 | 53.3 | 51.2 | – | – | – | 72.6 | 51.6 | 71.5 | 0.2 | **81.8** |
| toy brick scratch | 11.7 | 16.1 | 21.1 | 9.9 | – | – | – | 20.2 | 0.2 | 3.0 | 0.0 | **26.7** |
| transistor missing parts | 0.0 | 0.1 | 0.0 | 0.2 | – | – | – | 0.0 | 0.0 | 0.2 | 0.0 | **0.5** |
| u block abrasion | 0.0 | 58.1 | 57.1 | 49.1 | – | – | – | 37.6 | 0.7 | 8.7 | 1.1 | **66.0** |
| u block contamination | 16.7 | 15.7 | 53.8 | 66.5 | – | – | – | 70.8 | 0.6 | 81.1 | 0.0 | **81.5** |
| usb adapter contamination | 0.6 | 14.1 | 33.0 | 36.4 | – | – | – | 17.1 | 0.7 | **47.6** | 0.1 | 47.3 |
| usb adapter scratch | 1.1 | 5.1 | 13.2 | 13.9 | – | – | – | **52.7** | 0.0 | 26.3 | 0.0 | 49.4 |
| vcpill contamination | 58.6 | 62.1 | 56.8 | **65.4** | – | – | – | 29.2 | 4.0 | 38.3 | 0.0 | 54.0 |
| vcpill missing parts | 5.1 | 46.7 | 40.5 | 38.4 | – | – | – | 20.9 | 34.3 | 62.7 | 1.5 | **63.3** |
| wooden beads contamination | 51.7 | 63.1 | 76.8 | 77.4 | – | – | – | 87.3 | 9.0 | 84.8 | 0.1 | **90.2** |
| zipper damage | 0.1 | 1.3 | 0.9 | 0.9 | – | – | – | 0.0 | 1.0 | **3.8** | 0.1 | 2.7 |
| zipper deformation | 0.0 | 0.9 | 0.0 | 2.7 | – | – | – | 0.6 | 2.7 | **11.2** | 0.1 | 8.0 |

**Results and discussion.** At IoU 0.50, OMQ reaches 39.14 mask $AP_{50}$, improving over DINOv by 7.67 points and over UniSpector's 4.86 $AP_{50}$ by 34.29 points, while ranking in the top two on 42 categories. Twenty-five categories gain at least five points over DINOv. U block abrasion (+57.4), hazelnut hole (+52.8), bottle broken small (+32.9), and carpet cut (+32.6) provide the clearest evidence that prompt-derived morphology recovers the correct target region despite category novelty. UniSpector's low $AP_{50}$ despite its substantially higher mask recall shows that broad candidate coverage does not by itself produce calibrated category-conditioned masks. The contrast with GC10-DET remains sharp ($-18.54$ points on average), where extended or low-contrast cues impede even coarse mask recovery. The matching failure pattern in box and mask $AP_{50}$ locates this weakness in representation transfer rather than in the mask head alone.

Table A10: Complete category-level pixel grounding $AP_{75}$ (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.5 | 5.1 | 0.0 | **7.5** |
| steel surface crescent gap | 0.0 | 0.0 | 0.2 | 0.0 | – | – | – | 0.0 | 0.0 | **0.3** | 0.0 | 0.0 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **2.2** | 0.0 | 1.7 |
| magnetic tile blowhole | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **0.2** | 0.0 | **0.2** |
| bottle broken small | 0.3 | **71.4** | 55.8 | 64.0 | – | – | – | 10.8 | 0.0 | 25.6 | 0.3 | 55.6 |
| cable bent wire | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 13.4 | 0.0 | **33.3** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 0.2 | 2.4 | 1.5 | **12.9** | – | – | – | 5.8 | 0.0 | 0.2 | 0.0 | 6.2 |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **2.0** | 0.0 | 1.8 |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| capsule poke | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 7.5 | 0.0 | **12.2** |
| capsule scratch | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 7.9 | 0.0 | **23.4** |
| carpet cut | 0.0 | 35.1 | 11.3 | 16.3 | – | – | – | 0.0 | 0.0 | 41.3 | 0.0 | **69.1** |
| hazelnut hole | 23.6 | 12.4 | 30.9 | 1.8 | – | – | – | 0.0 | 0.0 | 5.0 | 0.4 | **67.1** |
| leather cut | 0.0 | 6.2 | 0.4 | **8.4** | – | – | – | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 |
| leather fold | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **10.7** | 0.0 | 0.2 |
| metal nut bent | 0.1 | 0.9 | 0.6 | 12.5 | – | – | – | 2.3 | 0.1 | 8.1 | 0.0 | **20.2** |
| metal nut scratch | 0.0 | 0.0 | 0.1 | 0.3 | – | – | – | 0.2 | 0.0 | 0.0 | 0.1 | **4.4** |
| pill scratch | 0.0 | 0.1 | 0.7 | 1.4 | – | – | – | 0.0 | 0.0 | 17.8 | 1.3 | **41.7** |
| tile oil spot | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **10.1** | 0.1 | 4.9 |
| wood combined defect | 3.0 | 9.0 | 11.7 | 9.5 | – | – | – | 0.0 | 0.0 | 9.1 | 0.0 | **26.2** |
| zipper broken teeth | 0.0 | 0.0 | 1.2 | 0.0 | – | – | – | 1.0 | 0.0 | 4.4 | 0.1 | **16.9** |
| zipper rough | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 2.4 | 0.0 | **63.5** | 0.2 | 61.8 |
| zipper squeezed teeth | 0.0 | 0.0 | 0.0 | **11.3** | – | – | – | 5.9 | 0.2 | 0.9 | 0.0 | 2.9 |
| audio jack contamination | 14.6 | 4.3 | 32.3 | 34.0 | – | – | – | 7.0 | 64.9 | 69.7 | 0.0 | **72.5** |
| bottle cap pit | 3.1 | 1.3 | 0.6 | 0.4 | – | – | – | 16.4 | 0.2 | **27.6** | 0.1 | 21.0 |
| button battery contamination | 13.9 | 29.6 | 33.7 | 32.0 | – | – | – | 43.0 | 24.0 | 27.8 | 0.0 | **48.5** |
| end cap damage | 0.1 | 0.1 | 0.4 | 0.3 | – | – | – | 0.5 | 0.6 | 0.7 | 0.0 | **6.6** |
| end cap missing parts | 0.0 | 0.5 | 0.1 | 0.3 | – | – | – | 0.5 | 0.9 | 4.2 | 0.0 | **7.0** |
| eraser scratch | 0.2 | 0.5 | 0.5 | 0.3 | – | – | – | 0.1 | 0.0 | **3.9** | 0.0 | 3.5 |
| mint foreign objects | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **0.2** | 0.0 | 0.1 |
| mounts contamination | 0.2 | 2.0 | 4.1 | 4.9 | – | – | – | 0.0 | 0.3 | **68.4** | 0.0 | 64.8 |
| mounts missing parts | 0.0 | 2.7 | 16.8 | 2.0 | – | – | – | 0.2 | 0.3 | 6.9 | 0.4 | **22.6** |
| mounts pit | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.3 | **1.6** | 0.1 | 1.4 |
| pcb missing parts | 0.0 | 0.1 | 1.1 | **1.4** | – | – | – | 0.0 | 0.1 | 0.3 | 0.0 | 0.1 |
| phone battery scratch | 0.0 | 0.1 | 0.0 | 0.2 | – | – | – | 0.1 | 0.0 | 0.7 | 0.0 | **1.1** |
| rolled strip base missing parts | 3.0 | 4.3 | 8.4 | 7.7 | – | – | – | 8.3 | 0.5 | 16.2 | 0.1 | **24.1** |
| tape missing parts | 1.6 | 18.1 | 20.7 | 14.4 | – | – | – | 49.9 | 0.9 | 48.6 | 0.1 | **51.1** |
| toothbrush abrasion | 0.0 | 0.0 | **0.1** | **0.1** | – | – | – | 0.0 | 0.0 | **0.1** | 0.0 | **0.1** |
| toothbrush contamination | 0.2 | 1.4 | 0.5 | 1.5 | – | – | – | **2.2** | 1.1 | **2.2** | 0.0 | 0.6 |
| toy contamination | 3.7 | 7.8 | 10.2 | 11.5 | – | – | – | 59.6 | 36.0 | 61.3 | 0.1 | **66.8** |
| toy brick scratch | 0.5 | 0.5 | 0.6 | 0.2 | – | – | – | 0.3 | 0.0 | 0.0 | 0.0 | **1.0** |
| transistor missing parts | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| u block abrasion | 0.0 | 19.1 | 25.9 | 20.8 | – | – | – | 24.6 | 0.3 | 5.4 | 0.4 | **52.2** |
| u block contamination | 11.5 | 6.2 | 22.3 | 20.5 | – | – | – | 63.8 | 0.4 | **77.5** | 0.0 | 77.0 |
| usb adapter contamination | 0.1 | 1.5 | 3.3 | 2.7 | – | – | – | 6.9 | 0.1 | 15.1 | 0.0 | **16.9** |
| usb adapter scratch | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.9 | 0.0 | 0.8 | 0.0 | **1.9** |
| vcpill contamination | 23.9 | 6.0 | 3.7 | 5.1 | – | – | – | 12.1 | 1.4 | 16.5 | 0.0 | **28.1** |
| vcpill missing parts | 0.3 | 10.9 | 10.4 | 8.2 | – | – | – | 7.5 | 13.7 | **28.7** | 0.5 | 28.1 |
| wooden beads contamination | 17.3 | 18.0 | 26.4 | 27.0 | – | – | – | 66.5 | 3.9 | 64.9 | 0.0 | **68.1** |
| zipper damage | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **0.1** | 0.0 | 0.0 |
| zipper deformation | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | **0.1** | 0.0 | 0.0 |

**Results and discussion.** OMQ obtains 21.59 mask $AP_{75}$, improving over DINOv by 6.50 points and over UniSpector's 2.90 $AP_{75}$ by 18.68 points. It ranks in the top two on 46 categories and achieves 31 category-wise best results when ties are counted. The strongest gains over DINOv occur for hazelnut hole (+62.1), u block abrasion (+46.8), bottle broken small (+30.0), and carpet cut (+27.8). The average gain on MVTec AD rises to 11.03 points at this stricter threshold, which supports the view that target-conditioned transport improves fine spatial alignment rather than only foreground activation. UniSpector reaches 64.9 on audio jack contamination while retaining a 2.90 macro score, again exposing highly localized success without category-stable boundary transfer. The remaining failures are concentrated in leather fold, bottle cap pit, and tile oil spot, where boundary appearance varies too strongly across exemplars.

Table A11: Complete category-level pixel grounding $AR_{100}$ (%) over all 52 unseen defect categories. † denotes the fixed, non-repeated 5-shot visual-prompt setting. The best and second-best results in each row are shown in bold and underlined, respectively. – denotes an unavailable result.

| Category | SEEM | YLE-S | YLE-M | YLE-L | YW-S | YW-M | GDN | GLEE | UniSp.† | DINOv† | SINE† | OMQ† |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| steel surface crease | 0.6 | 0.4 | 0.0 | 0.0 | – | – | – | 23.6 | 27.8 | **34.3** | 2.1 | 29.2 |
| steel surface crescent gap | 0.0 | 0.0 | 5.8 | 0.0 | – | – | – | 0.0 | 9.0 | **31.0** | 0.2 | 26.9 |
| steel surface welding line | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 17.0 | 4.4 | 32.8 | 0.0 | **34.8** |
| magnetic tile blowhole | 4.2 | 3.9 | 9.1 | 3.4 | – | – | – | 4.6 | 0.0 | 17.0 | 17.8 | **23.5** |
| bottle broken small | 28.8 | 58.5 | 53.8 | 60.4 | – | – | – | 62.7 | 54.5 | 62.0 | 59.2 | **64.0** |
| cable bent wire | 19.4 | 3.8 | 0.0 | 0.0 | – | – | – | 38.1 | 42.7 | 44.5 | 22.5 | **47.3** |
| cable swap | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cable combined defect | 15.7 | 12.0 | 12.5 | 13.8 | – | – | – | 21.5 | 19.2 | 45.8 | 5.8 | **50.0** |
| capsule crack | 0.0 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 24.4 | 40.6 | 6.5 | **41.7** |
| capsule faulty imprint | 0.0 | 0.0 | 0.0 | 1.8 | – | – | – | 20.9 | 17.6 | **23.5** | 5.0 | 21.2 |
| capsule poke | 1.4 | 0.0 | 0.0 | 0.0 | – | – | – | 0.0 | 23.5 | 45.9 | 31.4 | **50.0** |
| capsule scratch | 2.9 | 6.7 | 0.4 | 0.0 | – | – | – | 2.5 | 31.1 | 46.3 | 2.1 | **48.9** |
| carpet cut | 27.8 | 57.2 | 52.8 | 56.7 | – | – | – | 54.4 | 55.8 | **67.5** | 48.3 | 65.0 |
| hazelnut hole | 62.5 | 63.0 | 61.5 | 30.5 | – | – | – | 4.5 | 71.5 | 74.6 | 58.5 | **78.5** |
| leather cut | 24.0 | 20.5 | 22.0 | 22.0 | – | – | – | 2.0 | 25.0 | 27.1 | 19.0 | **30.0** |
| leather fold | 21.8 | 0.0 | 0.0 | 0.0 | – | – | – | 5.9 | 10.0 | 23.3 | 4.7 | **34.2** |
| metal nut bent | 29.0 | 36.5 | 35.2 | 44.8 | – | – | – | 52.3 | 58.5 | 62.3 | 27.7 | **68.8** |
| metal nut scratch | 11.8 | 1.0 | 7.8 | 13.3 | – | – | – | 38.3 | 35.5 | 41.6 | 18.0 | **50.0** |
| pill scratch | 24.0 | 21.6 | 16.0 | 5.2 | – | – | – | 12.4 | 59.5 | 68.0 | 32.8 | **69.0** |
| tile oil spot | 57.2 | 0.0 | 0.0 | 0.0 | – | – | – | 24.4 | 23.8 | 39.2 | **81.7** | 57.7 |
| wood combined defect | 22.0 | 30.0 | 22.9 | 16.8 | – | – | – | 20.0 | 4.6 | 56.8 | 7.6 | **63.9** |
| zipper broken teeth | 0.0 | 7.0 | 20.7 | 0.0 | – | – | – | 28.1 | 42.6 | 38.4 | 30.7 | **44.7** |
| zipper rough | 16.4 | 0.0 | 0.0 | 0.0 | – | – | – | 61.6 | 41.9 | **70.0** | 46.4 | 66.3 |
| zipper squeezed teeth | 5.3 | 0.6 | 10.0 | 33.5 | – | – | – | 43.5 | 38.3 | 44.2 | 30.0 | **51.7** |
| audio jack contamination | 46.0 | 18.5 | 48.1 | 50.2 | – | – | – | 71.2 | 68.9 | 74.8 | 2.7 | **75.6** |
| bottle cap pit | 27.2 | 24.0 | 20.9 | 20.0 | – | – | – | 50.8 | 39.7 | 52.1 | 33.9 | **62.4** |
| button battery contamination | 33.5 | 45.4 | 45.8 | 45.7 | – | – | – | 56.5 | 62.6 | 61.8 | 8.4 | **63.5** |
| end cap damage | 5.5 | 14.0 | 13.7 | 9.2 | – | – | – | 35.2 | 29.2 | 43.5 | 3.6 | **48.4** |
| end cap missing parts | 10.4 | 1.5 | 7.3 | 14.2 | – | – | – | 37.2 | 33.6 | 40.6 | 2.0 | **42.9** |
| eraser scratch | 20.9 | 15.0 | 10.3 | 11.9 | – | – | – | 24.8 | 31.1 | **35.2** | 12.3 | 34.6 |
| mint foreign objects | 1.8 | 0.6 | 1.2 | 1.2 | – | – | – | 16.5 | 2.4 | 19.8 | 0.0 | **20.9** |
| mounts contamination | 20.0 | 9.7 | 4.9 | 15.3 | – | – | – | 8.2 | 56.9 | 68.1 | 25.7 | **68.5** |
| mounts missing parts | 17.6 | 4.8 | 22.2 | 3.1 | – | – | – | 53.1 | 53.9 | **63.3** | 34.2 | 62.8 |
| mounts pit | 16.6 | 0.3 | 0.0 | 0.0 | – | – | – | 0.0 | 37.7 | 48.0 | 14.5 | **54.2** |
| pcb missing parts | 2.1 | 8.4 | 12.7 | 16.3 | – | – | – | 3.5 | 11.2 | 24.1 | 0.8 | **25.3** |
| phone battery scratch | 3.7 | 16.7 | 17.2 | 16.7 | – | – | – | 21.8 | 21.6 | 24.7 | 6.7 | **25.4** |
| rolled strip base missing parts | 18.8 | 41.4 | 38.6 | 40.4 | – | – | – | 49.4 | 44.3 | 52.4 | 25.1 | **56.2** |
| tape missing parts | 11.4 | 39.9 | 38.9 | 31.2 | – | – | – | 58.6 | 56.5 | **64.8** | 32.5 | 62.5 |
| toothbrush abrasion | 3.2 | 6.4 | 7.7 | 7.3 | – | – | – | 2.6 | 9.9 | 12.1 | 0.0 | **12.2** |
| toothbrush contamination | 8.5 | 12.0 | 6.5 | 7.3 | – | – | – | 20.7 | 21.5 | **26.7** | 0.1 | 24.6 |
| toy contamination | 32.8 | 30.7 | 29.9 | 33.1 | – | – | – | 65.9 | 61.8 | **67.2** | 23.8 | 67.0 |
| toy brick scratch | 15.0 | 12.6 | 12.5 | 10.7 | – | – | – | 19.5 | 13.3 | 22.1 | 2.7 | **24.3** |
| transistor missing parts | 0.1 | 1.2 | 0.6 | 2.7 | – | – | – | 3.0 | 2.5 | 3.0 | 0.0 | **3.1** |
| u block abrasion | 0.0 | 41.7 | 39.2 | 43.1 | – | – | – | 54.4 | 53.7 | **59.0** | 35.5 | **59.0** |
| u block contamination | 48.2 | 40.2 | 46.2 | 46.5 | – | – | – | 72.6 | 65.3 | 75.3 | 15.3 | **76.2** |
| usb adapter contamination | 4.1 | 20.6 | 20.9 | 20.1 | – | – | – | 42.7 | 31.2 | 48.0 | 2.4 | **48.1** |
| usb adapter scratch | 1.6 | 7.5 | 7.2 | 7.3 | – | – | – | **26.1** | 2.8 | 25.5 | 0.3 | 24.8 |
| vcpill contamination | 37.0 | 30.4 | 25.4 | 29.9 | – | – | – | 55.3 | 51.3 | 54.7 | 0.6 | **58.2** |
| vcpill missing parts | 9.4 | 33.0 | 29.8 | 31.2 | – | – | – | 45.2 | 42.7 | **48.7** | 30.1 | 48.5 |
| wooden beads contamination | 46.4 | 45.2 | 45.4 | 44.3 | – | – | – | **69.6** | 58.9 | 68.5 | 16.9 | 69.4 |
| zipper damage | 4.4 | 5.4 | 3.5 | 4.2 | – | – | – | 3.1 | 5.7 | 11.0 | **12.4** | 8.3 |
| zipper deformation | 0.0 | 4.9 | 0.2 | 5.3 | – | – | – | **20.1** | 1.9 | 5.7 | 0.9 | 7.9 |

**Results and discussion.** Mask recall is the most category-stable result. OMQ reaches 45.23 $AR_{100}$, improving over DINOv by 2.20 points and over UniSpector's 32.56 recall by 12.67 points. It ranks in the top two on 50 of 52 categories and improves on DINOv in 38, with only steel surface crease losing more than five points. UniSpector is substantially closer to OMQ in recall than in mask AP, where the gap is 18.76 points. This contrast reinforces that candidate coverage is not the main limitation and instead points to category ranking, false-positive control, and boundary agreement. Tile oil spot provides a parallel within-category example, as recall rises by 18.5 points over DINOv while mask AP falls by 7.1. The resulting precision–recall mismatch identifies prototype ambiguity and spatial calibration, rather than proposal scarcity, as the remaining failure mode.

**Cross-metric synthesis.** Taken together, Tables A4–A7 and Tables A8–A11 reveal a consistent separation between candidate coverage and morphology-grounded localization. UniSpector retains box and mask $AR_{100}$ of 44.19 and 32.56 but obtains only 3.47 and 2.77 AP, exposing the limitations of candidate coverage alone under the stronger false-positive pressure created by the 106,867 defect-free images retained in our evaluation but excluded from InsA. Relative to DINOv, OMQ gains 6.57 and 5.37 points in box and mask AP, compared with only 2.47 and 2.20 points in $AR_{100}$, while retaining substantial improvements at $AP_{75}$. The larger gains in AP than in AR indicate that OMQ's improvements are concentrated in the ranking quality and spatial agreement of retrieved instances, rather than in simply expanding candidate coverage. NMDE re-expresses prompt morphology relative

to shared bases, PCPE stabilizes it under appearance, geometry, and boundary variations, and PTQI reanchors the resulting representation to target evidence. The resulting gains across boxes, masks, and most unseen categories support the central reference-frame account: unseen-defect transfer succeeds when morphology is re-expressed independently of its support context and subsequently grounded at the target location and extent. The remaining failures on GC10-DET and tile oil spot mark the current limit of this mechanism, as low-contrast morphology that is not sufficiently isolated before transport remains difficult to reanchor.

### A.5 Extended Qualitative Analysis

Figures A1–A10 present 50 OpenMorphoQuery response maps. Each group follows the same three-row organization: the input image, its ground-truth mask, and the OMQ response map. The cases cover compact contamination and holes, elongated scratches and creases, missing components, boundary abrasion, and compound deformation across markedly different products and materials.

**Cross-case behavior.** To examine how OMQ transfers prompt morphology across unseen products and materials, Figures A1–A10 compare the input image, ground-truth region, and OMQ response for 50 additional cases. Across these cases, the response geometry adapts to diverse defect structures rather than following a fixed object template. Compact responses are observed for localized blowholes, holes, pits, and contamination, as illustrated by Samples 4, 14–16, 22–24, and 30–31. Elongated or boundary-oriented responses appear for capsule scratch (Sample 13), toothbrush abrasion (Samples 35–38), U-block abrasion (Samples 43–44), USB-adapter scratch (Sample 46), and zipper deformation (Sample 50). Structural and missing-part defects, including cable combined defect (Sample 10) and the missing-part cases in Samples 27, 32–34, 40–42, and 48, produce responses that adapt to their component-level layouts. Despite substantial changes in product identity, material, scale, and viewpoint, the response topology remains associated with the prompted defect structure while its spatial support adapts to each target image.

**Host-context suppression and reanchoring.** In Figure 4, baseline responses frequently spread over recognizable host-product structures rather than remaining confined to the prompted defect morphology. This pattern makes reference-frame misalignment directly observable in the response maps. When the morphology representation remains anchored to the observation frame of a support exemplar, matching inherits its product appearance and spatial structure, hindering direct alignment with defects in the target image. In contrast, OMQ suppresses host-structure responses and refocuses activation on the locations and extents of target defects. Figures A1–A10 further show that this behavior persists across products, retaining defect-relevant morphology while adapting to each target's spatial configuration. Together with the 25.15-point improvement in response-mass precision and the component ablations, these results indicate that NMDE and PCPE weaken support-specific contextual anchors, while PTQI reanchors the resulting morphology to the target reference frame, thereby alleviating prompt–target reference-frame misalignment.

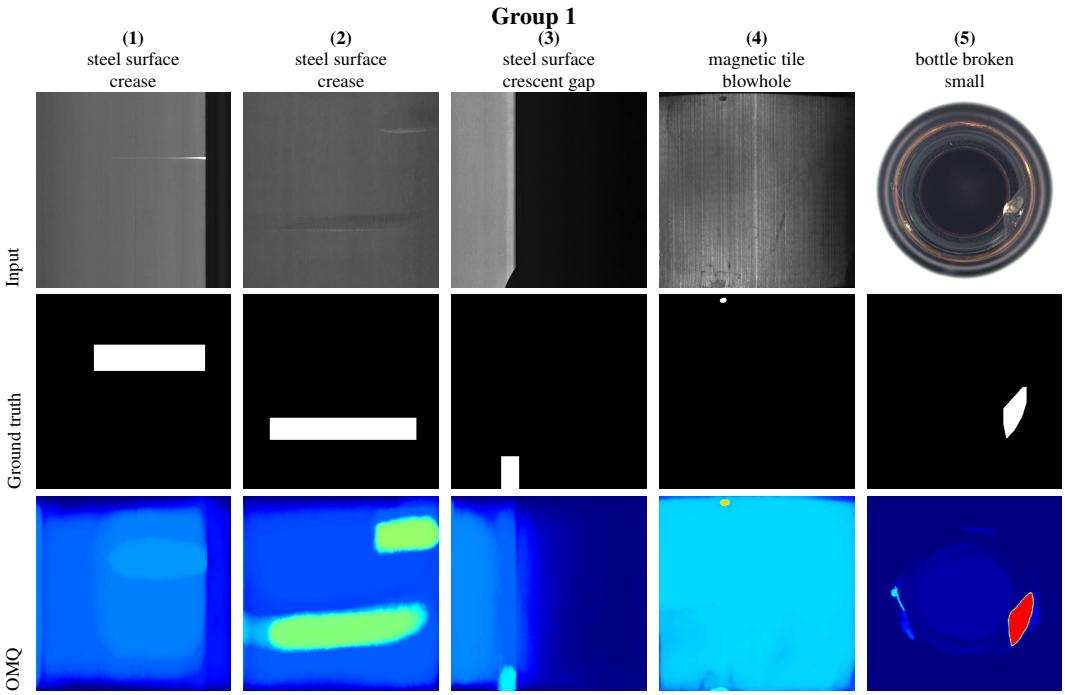



Figure A1: Additional qualitative results, Group 1 (Samples 1–5). Rows show the input, ground truth, and OMQ response map from top to bottom.

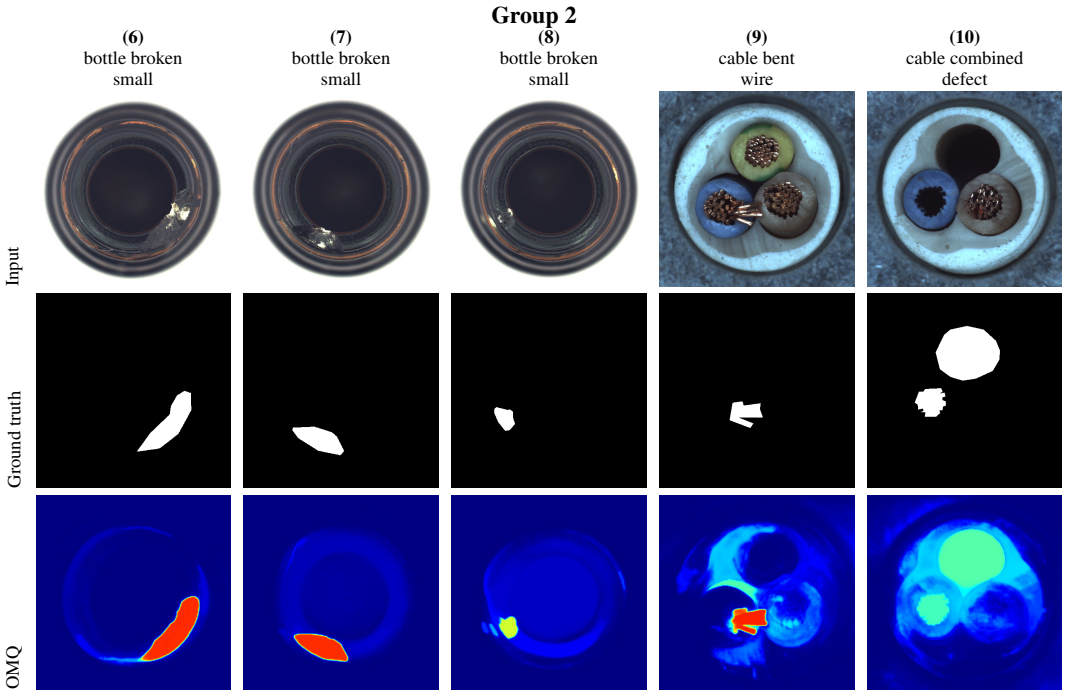



Figure A2: Additional qualitative results, Group 2 (Samples 6–10). Rows show the input, ground truth, and OMQ response map from top to bottom.

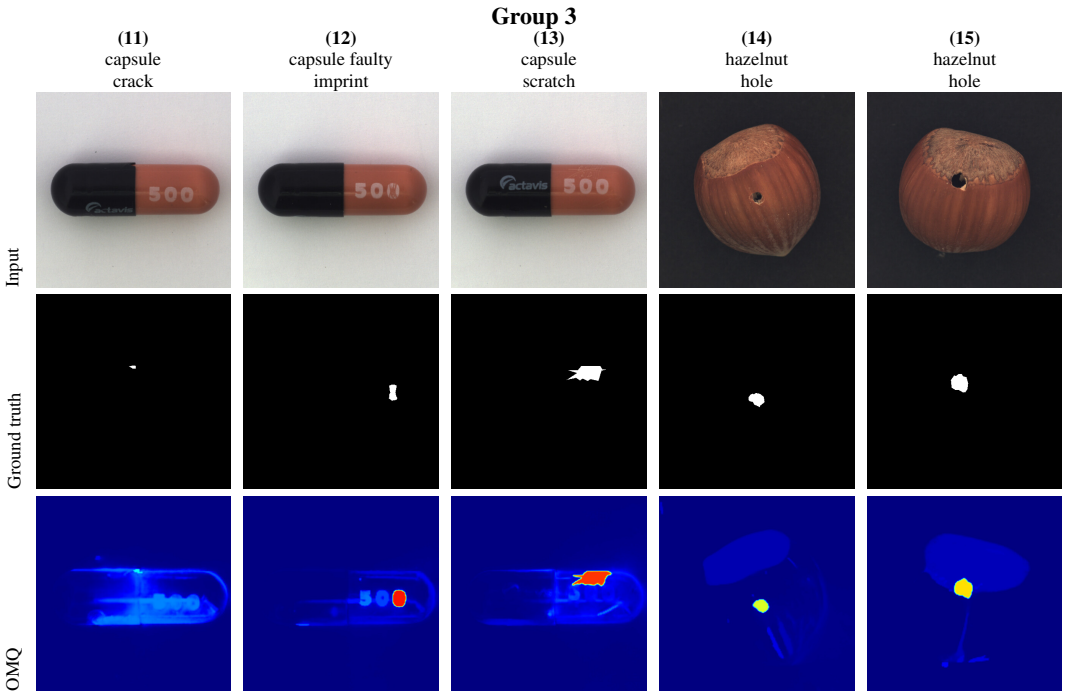



Figure A3: Additional qualitative results, Group 3 (Samples 11–15). Rows show the input, ground truth, and OMQ response map from top to bottom.

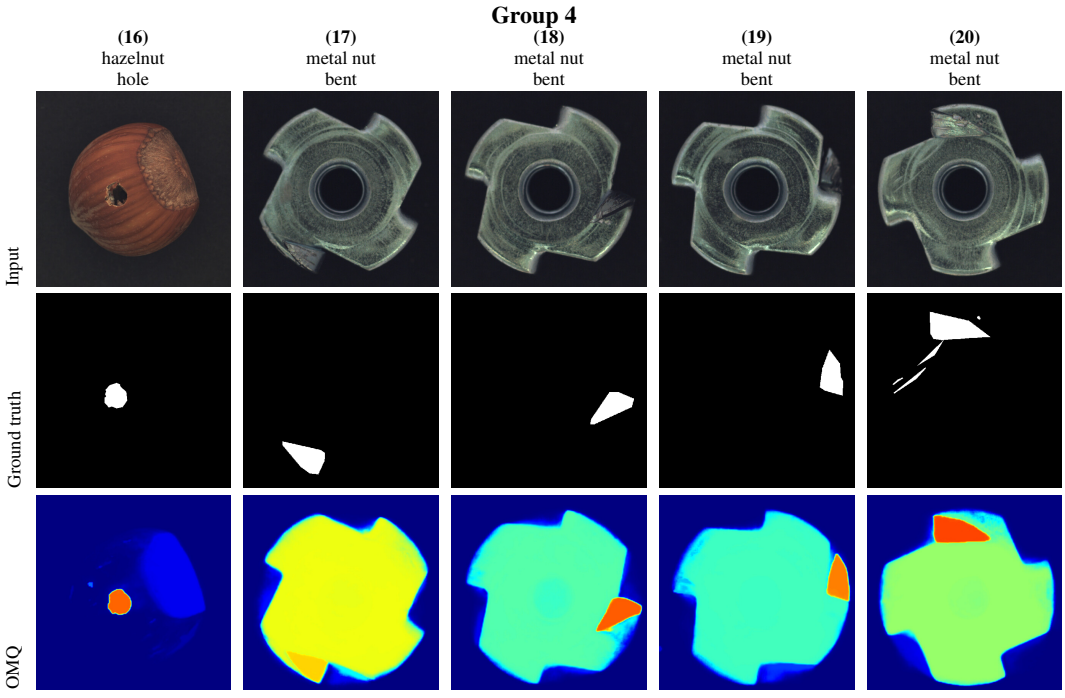



Figure A4: Additional qualitative results, Group 4 (Samples 16–20). Rows show the input, ground truth, and OMQ response map from top to bottom.

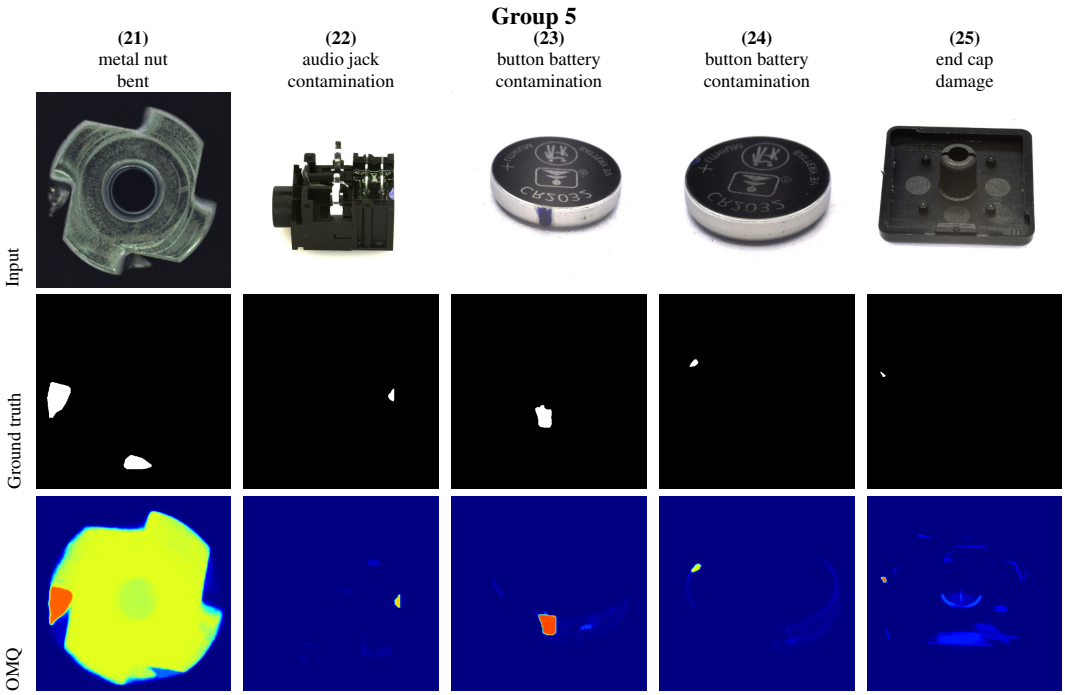



Figure A5: Additional qualitative results, Group 5 (Samples 21–25). Rows show the input, ground truth, and OMQ response map from top to bottom.

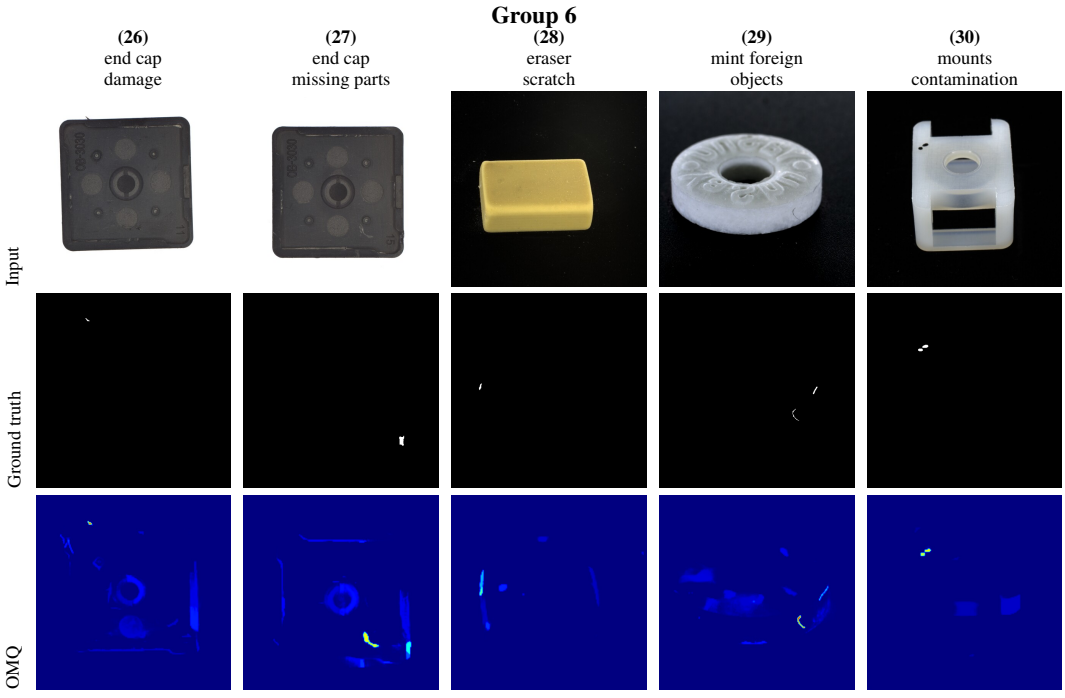



Figure A6: Additional qualitative results, Group 6 (Samples 26–30). Rows show the input, ground truth, and OMQ response map from top to bottom.

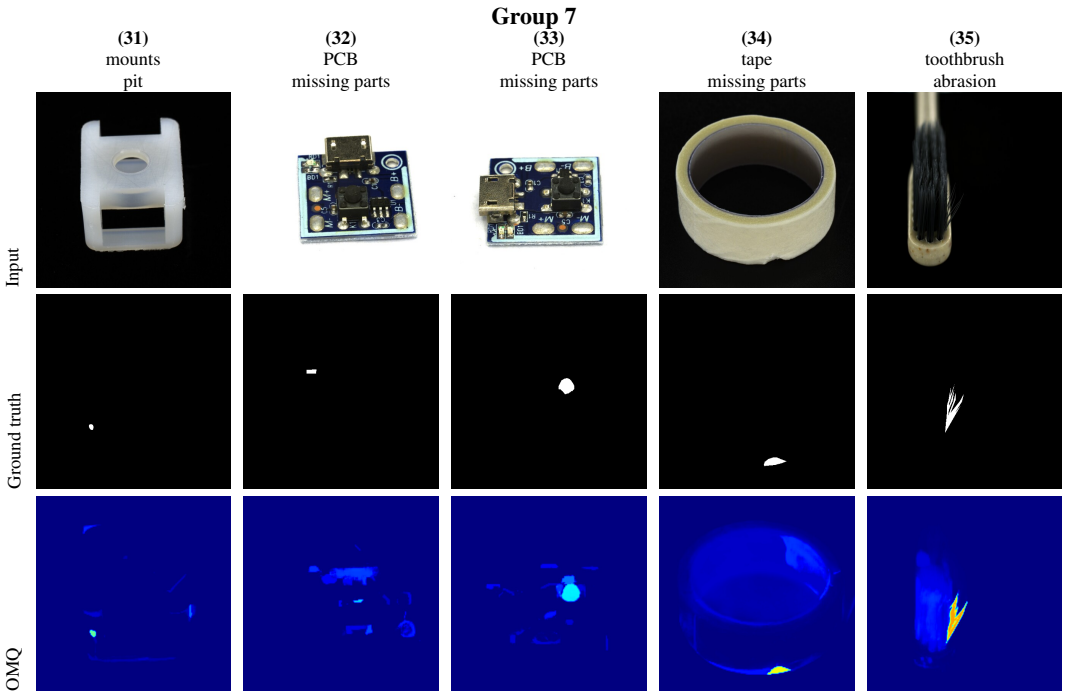



Figure A7: Additional qualitative results, Group 7 (Samples 31–35). Rows show the input, ground truth, and OMQ response map from top to bottom.

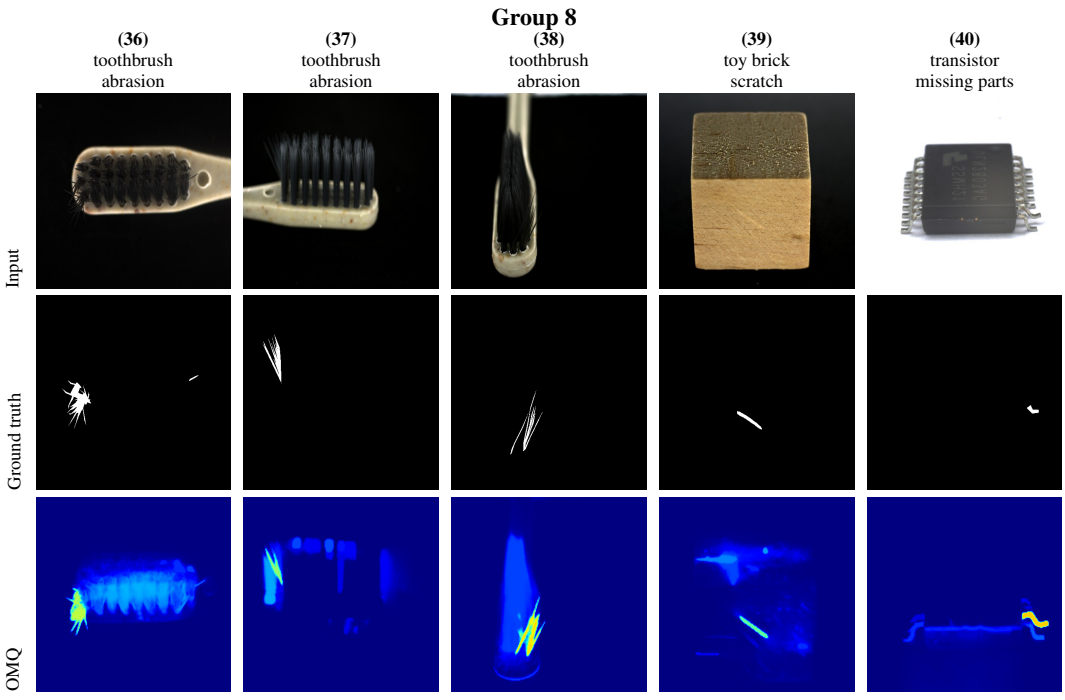



Figure A8: Additional qualitative results, Group 8 (Samples 36–40). Rows show the input, ground truth, and OMQ response map from top to bottom.

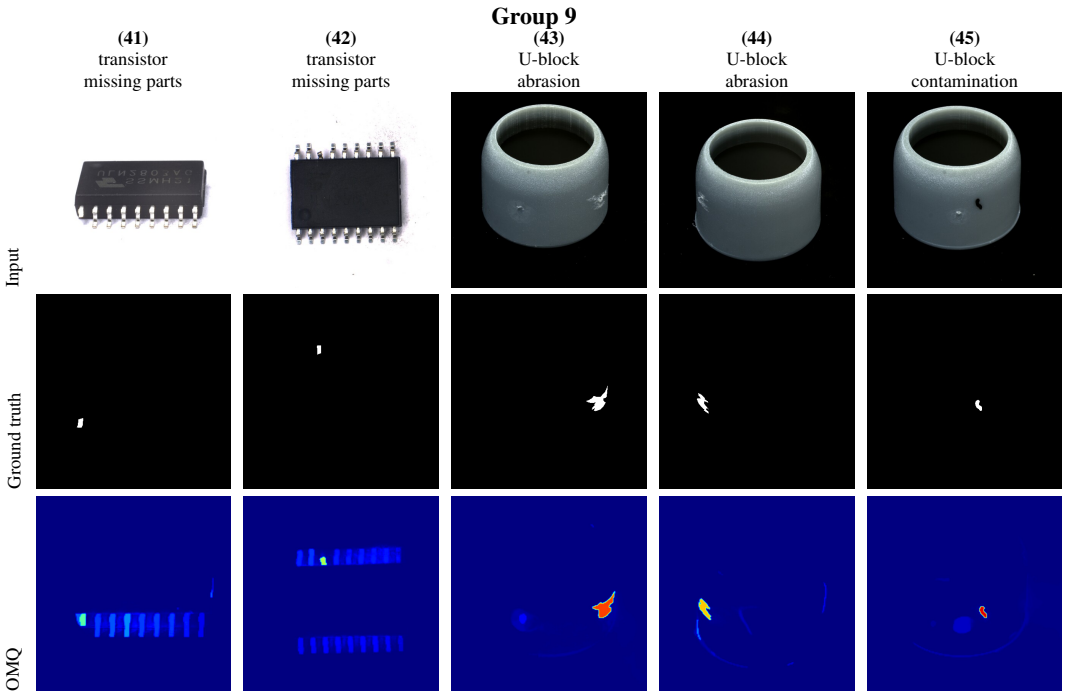



Figure A9: Additional qualitative results, Group 9 (Samples 41–45). Rows show the input, ground truth, and OMQ response map from top to bottom.

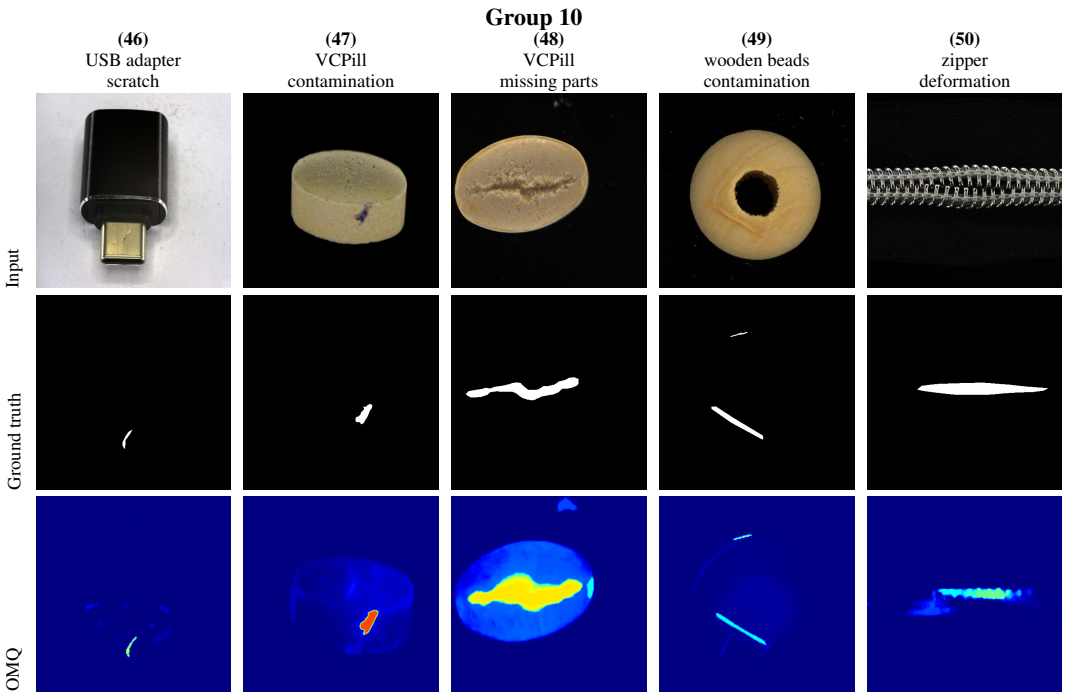



Figure A10: Additional qualitative results, Group 10 (Samples 46–50). Rows show the input, ground truth, and OMQ response map from top to bottom.

**Challenge diagnosis.** The qualitative results show that OMQ generally preserves morphology-consistent responses across variations in product appearance, material, scale, and viewpoint, with most activations concentrated around the prompted defect structures. A shared limitation nevertheless appears in a small number of cases. Low-contrast surfaces in Samples 1–4 weaken the defect-

conditioned residual available to NMDE, repetitive structures in Samples 17–21 and 40–42 create multiple locations with nearly equivalent geometry, and the tiny defects in Samples 24, 28, and 31 can be mixed with normal context within a feature token's receptive field. These conditions make neighboring or structurally similar target tokens receive comparable PTQI matching costs, producing broader, displaced, or partially aligned responses. The activations still remain close to the relevant surface or component, indicating successful retrieval of the correct morphological neighborhood but uncertainty in its precise spatial extent.

This interpretation is consistent with the quantitative results. OMQ improves mask AP over DINOv by 5.37 points, compared with a 2.20-point gain in $AR_{100}$, showing that its main benefit lies in response concentration, ranking, and spatial agreement rather than increased proposal coverage. Steel surface crease is also the only category whose recall decreases by more than five points, suggesting that this limitation is localized rather than systematic. Finer-resolution encoding, geometry-aware transport, and stronger boundary calibration may further improve these cases.

### A.6 Prompt–Target Alignment Diagnostic

**Diagnostic construction.** We further examine whether the performance gains of OpenMorphoQuery are accompanied by stronger spatial correspondence between matched visual prompts and target defects. For decoder query $q$, let $M_q(u)$ denote its mask logit at spatial location $u$, and let $S_{q,c}$ denote its classification logit for the matched prompt category $c$. We construct the category-conditioned response as:

$$R_c(u) = \max_q \sigma(M_q(u))\, \sigma(S_{q,c})\,. \tag{11}$$

This response is computed before thresholding and mask post-processing. Ground-truth masks are projected onto each model's native response grid using the corresponding resizing and padding geometry, with padded positions excluded. Ground truth is used only for diagnostic evaluation and does not enter model inference.

**Metrics and aggregation.** Pixel AUROC measures whether response scores rank GT pixels above non-GT pixels. Response-mass precision and top-$\rho$ precision are defined as:

$$P_{\text{mass}} = \frac{\sum_{u\in G} R_c(u)}{\sum_{u\in\Omega} R_c(u)}, \qquad P_\rho = \frac{\left|\text{Top}_\rho(R_c) \cap G\right|}{\left|\text{Top}_\rho(R_c)\right|}, \tag{12}$$

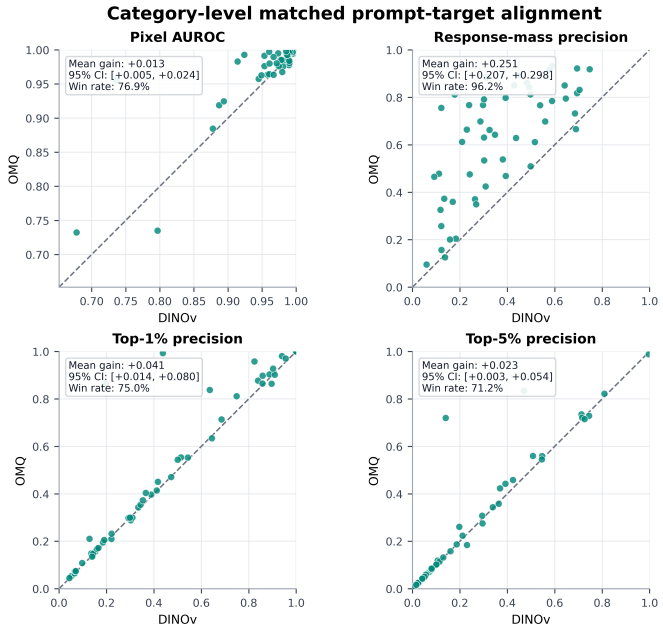
where $G$ is the projected GT region, $\Omega$ is the valid response grid, and $\rho \in \{0.01, 0.05\}$. Metrics are first averaged within each category and then macro-averaged over the 52 predefined unseen categories. Uncertainty is estimated using 10,000 paired bootstrap resamples over categories.

**Category-level behavior.** Figure A11a compares paired category-level scores. OMQ outperforms DINOv on 40/52 categories for pixel AUROC, 50/52 for response-mass precision, 39/52 for top-1% precision, and 37/52 for top-5% precision. The response-mass improvement is particularly pronounced, increasing from 0.3703 to 0.6218, with a 95% confidence interval of [0.2066, 0.2978] for the paired gain. The positive AUROC and top-$k$ gains further show that OMQ improves spatial ranking and peak localization, rather than merely changing the global magnitude of its decoder responses.

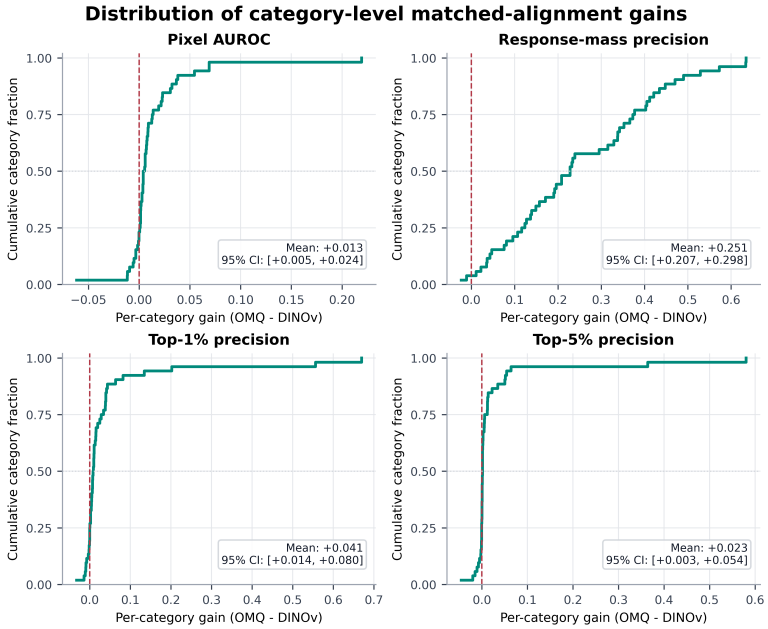
**Reference-frame interpretation.** This diagnostic makes prompt–target reference-frame mismatch observable through decoder responses. When prompt morphology remains anchored to support-specific texture and context, category-conditioned evidence spreads onto visually similar host-product structures. Generic foreground matching would likewise produce broad object-level activation. The paired results in Figures A11a and A11b show the opposite behavior. OMQ preserves category-relevant morphology while reallocating response mass toward annotated defect regions, yielding simultaneous gains in spatial ranking, peak localization, and response-mass precision. Because these changes are measured on matched prompts and queries, they establish that the transfer barrier lies in how prompt morphology is anchored across support and target contexts, rather than in a global change in response magnitude.

Together with the component ablations and qualitative cross-context responses in Figure 4, this diagnostic identifies prompt–target reference-frame mismatch as a concrete and measurable source of

transfer error. NMDE and PCPE suppress prompt-bound appearance and contextual variation, while PTQI reanchors the stabilized morphology to the target location and extent. The resulting transition from diffuse host-structure activation to concentrated defect evidence is the predicted signature of successful reference-frame transformation. These convergent results validate the central premise of OpenMorphoQuery and demonstrate that its gains arise from morphology re-expression, stabilization, and target-conditioned grounding.




(a) Paired category-level alignment between DINOv and OMQ.




(b) ECDFs of category-level alignment gains (OMQ minus DINOv).

Figure A11: Matched prompt–target alignment across 52 unseen categories. In (a), each point denotes one category, the diagonal indicates equal performance, and points above it favor OMQ. Insets report category-macro gains, 95% bootstrap confidence intervals, and win rates. In (b), the red dashed line marks zero gain; the positive displacement indicates improvements distributed across categories.