ROUTES INTERFERE, REPRESENTATIONS TRANSFER: PARAMETER-EFFICIENT ADAPTATION OF PROMPTABLE SEGMENTATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Promptable segmentation models increasingly expose several prompt interfaces over one shared backbone: a text-grounded concept route, a click-driven geometric route, and a visual-exemplar route. Adapting such a model to a specialist domain raises a question that cannot arise for single-interface models – should the routes share one adaptation? We answer it on SAM3 in a setting that needs all three routes at once, model-assisted annotation of printed circuit boards, where the pretrained model is not merely weak but silent, returning no confident detection at all for two of its seven concepts. First, *routes interfere*: folding geometric prompts into concept training costs 20.5 classification-gated F1 at an identical parameter budget, because the objectives ask the shared features for opposing properties – category discrimination against prompt-conditioned figure–ground separation – so the damage falls on recognition, not localisation, exactly as that account predicts. One small adapter set per route removes the conflict: 0.10% of the parameters raise classification-gated F1 from 38.0 to 55.9 ($22.3 \rightarrow 51.2$ on a board-disjoint held-out set), and a disjoint 0.011% in the mask decoder lifts one-click IoU from 0.79 to 0.89 while calibrating its own quality estimate for review triage. Second, *representations transfer across routes*: adapters trained on text alone, never exposed to an exemplar, improve the untrained exemplar route by 45% relative, with the largest gains on the fine-grained concepts. Low-rank adaptation thus changes what the backbone represents rather than how language is matched to it – which also explains why placement dominates capacity: quadrupling the trainable budget drops the model *below* zero-shot. We release the adapters, code, and a 330-image, $6{,}968$-mask benchmark.

1 INTRODUCTION

A promptable segmentation model is no longer a single mapping from an image and a prompt to a mask. SAM (Kirillov et al., 2023) accepted points and boxes; its successors accept far more. SEEM (Zou et al., 2023), T-Rex2 (Jiang et al., 2024), and DINO-X (Ren et al., 2024) combine heterogeneous prompts, and SAM3 (Carion et al., 2026) exposes three interfaces over one shared backbone: a *concept* route grounding a noun phrase in a set detector, a *geometric* route in which clicks drive a SAM-style decoder to one object, and a *visual-exemplar* route encoding a selected region into tokens the detector consumes where text would otherwise enter.

This changes what adaptation means. For a single-interface model there is one objective and the only open question is where to place the adapter (He et al., 2022). For a multi-route model a prior question appears: **how many adaptations should there be, and which routes should share one?** The default answer — one low-rank residual trained on a mixture of the available prompt types — is the cheapest option and what the multi-task tradition recommends (Caruana, 1997). We show it is the wrong answer here, and that the reason is more useful than the recipe.

The question has practical bite because real workflows use several routes at once. Ours is model-assisted annotation, where dense instance labelling dominates the cost of a corpus — one Cityscapes image took over $1.5\,\mathrm{h}$ (Cordts et al., 2016) — and a promptable model turns it into per-concept prompting: the concept route pre-annotates a category from one phrase, the geometric route corrects

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.



"capacitor" "varistor" "film capacitor" (a) zero-shot SAM3 (b) with Δ_{txt}

Figure 1: **A domain where the pretrained model is silent, not merely weak.** Left: the three mutually exclusive categories a general model conflates. Right: one crop of an unseen board under "film capacitor" — zero-shot SAM3 returns nothing above its 0.5 threshold (a), adapting 0.10% of the parameters recovers all three (b).

what it gets wrong, the exemplar route carries a component the annotator cannot name across the remaining images. Buying accuracy on one route by degrading another is not a trade such a loop can absorb.

A diagnostic domain. Measuring interference needs a domain where the pretrained model genuinely fails, so the training signal rather than a lucky threshold decides the outcome. Fine-grained printed-circuit-board (PCB) components provide it (Fig. 1): zero-shot SAM3 localises common parts competently but produces *no confident detection at all* for film capacitors or varistors, confusing them with the capacitors beside them. Localisation survives the domain shift, recognition does not.

Finding 1: routes interfere. Training one adapter set on a mixture of text and geometric prompts costs 20.5 classification-gated F1 against the identical budget trained on text alone, and the loss falls on image-level recognition (IL_MCC $0.870 \rightarrow 0.584$), not localisation. The routes make opposing demands on shared features: the concept route needs them category-discriminative, so "capacitor" does not attract a varistor, while the geometric route needs them category-*agnostic*, so a click on any object yields that object's boundary. One residual optimised for both is pulled in two directions. Route-disjoint adaptation — one adapter set per interface, each trained on its own objective — removes the conflict at no inference cost.

Finding 2: representations transfer across routes. Our concept adapters see text prompts exclusively and never an exemplar token, yet they improve the purely *visual* exemplar route by 45% relative across images, with gains falling on precisely the fine-grained categories the text route had to learn. Had adaptation merely re-aligned words with pre-existing features, a route that uses no words could not benefit. The adapters change what the backbone represents — which also explains why *where* the residual acts dominates how large it is: quadrupling the budget drops performance below zero-shot, while $884{,}736$ well-placed parameters give the largest gain we measure.

Contributions. We formalise *route-disjoint adaptation* and quantify the *route interference* that makes it necessary (Sec. 6.1); we report *cross-route transfer* as evidence about what low-rank adaptation learns (Sec. 6.3), which with the placement-over-capacity ablations favours a representational over an alignment account; and we show the principle repairs a domain SAM3 cannot serve at 0.114% of its parameters. We release the benchmark (330 images, $6{,}968$ masks), adapter weights and code.

2 Related Work

From one prompt interface to many. Interactive segmentation predicts one object from clicks: RITM (Sofiiuk et al., 2022) established iterative training with mask feedback, refined by localised updates (Chen et al., 2022b) and plain vision transformers (Liu et al., 2023), and SAM (Kirillov et al., 2023) scaled it to a billion masks. Open-vocabulary work moved in parallel from aligned image–text embeddings (Radford et al., 2021) to phrase grounding (Li et al., 2022; Liu et al., 2024b) and unified detection and segmentation (Zhang et al., 2023). The lines have since converged onto shared backbones (Tab. 1): how many adaptations a model needs is a question only for the lower block, and answerable today only for the row that ships adaptable weights. Where these models have been compared head to head the margin is wide, if on a benchmark SAM3's own authors built: it reports 54.1 cgF1 on open-vocabulary SA-Co/Gold against 24.6 for the strongest baseline (72.8 is

Table 1: **Where the question applies.** Prompt routes over a *shared* backbone: *concept* (a phrase retrieves every instance), *geometric* (clicks refine one object), *exemplar* (a marked region retrieves every match). Interference needs two to coexist, ruling out the top block. This is *scope*, not a results comparison: we measure on SAM3, the only model with all three routes that also emits masks and ships adaptable weights. [a]referring/open-vocabulary, not exhaustive retrieval; [b]exemplar-style, not iterative refinement; [c]pseudo-label head; [d]Kirillov et al. (2023); Ravi et al. (2025); Li et al. (2024).

	Concept	Geometric	Exemplar	Masks	Weights
SAM, SAM2, Semantic-SAM[d]	—	✓	—	✓	open
OpenSeeD (Zhang et al., 2023)	✓	—	—	✓	open
SEEM (Zou et al., 2023)	(✓)[a]	✓	✓	✓	open
T-Rex2 (Jiang et al., 2024)	✓	—[b]	✓	—	API
DINO-X (Ren et al., 2024)	✓	—[b]	✓	(✓)[c]	API
SAM3 (Carion et al., 2026)	✓	✓	✓	✓	open

human), and $+18.3$ AP over T-Rex2 on COCO under one-exemplar prompting (Carion et al., 2026). We take it as the strongest available instance of the multi-route design, not as a contribution of ours.

Parameter-efficient adaptation. Adapters learn a small residual pathway beside frozen weights (Houlsby et al., 2019), visual prompt tuning prepends learnable tokens (Jia et al., 2022), and LoRA expresses the update as two low-rank matrices (Hu et al., 2022), motivated by the low intrinsic dimensionality of fine-tuning (Aghajanyan et al., 2021). DoRA (Liu et al., 2024a), AdaptFormer (Chen et al., 2022a) and ConvPass (Jie & Deng, 2024) vary the mechanism; He et al. (2022) unify them in a design space where *placement* is a first-class choice, and low-rank updates forget less than full fine-tuning (Biderman et al., 2024). Adaptations of SAM and SAM2 (Zhang & Liu, 2023; Wu et al., 2025; Chen et al., 2023; Ke et al., 2023) touch a *single* route, since those models expose only a geometric interface. Concurrent work does adapt SAM3 itself with low-rank residuals — medical concepts (Liu et al., 2025), surgical scenes at 0.98% of parameters with the backbone frozen (Liu et al., 2026), structural defects at 0.12–1.34% (Malaisree et al., 2026) — but each trains one route on one prompt type; none asks how many adaptations a multi-route model needs.

Interference, negative transfer, and modularity. That jointly trained objectives can hurt one another is long established (Caruana, 1997): groupings matter more than task count (Standley et al., 2020), conflicting gradients sometimes need surgical separation (Yu et al., 2020), and negative transfer is ordinary when objectives are misaligned (Wang et al., 2019; Kumar et al., 2022). The parameter-efficient literature answers with modularity (Pfeiffer et al., 2021; Ilharco et al., 2023; Huang et al., 2024). Our contribution is a sharper axis along which this interference appears — not between downstream *tasks*, but between *prompt interfaces of one model on one task and one dataset* — and the finding that the resulting modules are not merely non-destructive but mutually beneficial (Sec. 6.3). Public PCB resources (Pramerdorfer & Kampel, 2015; Lu et al., 2020; Arbash et al., 2024; Nayak, 2021) provide coarse groups or box-level labels rather than the fine-grained instance masks this study needs, which is why we annotate a new corpus (Sec. 5.1).

3 Prompt Routes and Route-Disjoint Adaptation

Multi-route promptable models. We write such a model as a shared backbone with K prompt routes, $\mathcal{M} = (f_\theta, \{g_k\}_{k=1}^K)$ with $f_\theta : \mathcal{I} \to \mathcal{Z}$ and $g_k : \mathcal{P}_k \times \mathcal{Z} \to \mathcal{O}_k$, where f_θ maps an image to token features z and route k turns a prompt from its own space into an output. All routes read the same z and may share parameters. SAM3 (Carion et al., 2026) instantiates this with $K = 3$. The **concept route** embeds a noun phrase, fuses it with z in a DETR-style encoder–decoder (Carion et al., 2020), and emits $N = 200$ instance queries $(\hat{p}_i, \hat{b}_i, \hat{m}_i)$ plus an image-level presence logit gating their scores. The **geometric route** feeds clicks, boxes, and the previous mask into a SAM-style two-way decoder returning one mask and a predicted-IoU token. The **exemplar route** encodes a selected region into tokens injected into the concept route's fusion encoder where the text embedding would go: not an independent head but a second prompt encoder feeding the concept head, which is what makes transfer between the two measurable (Sec. 6.3).

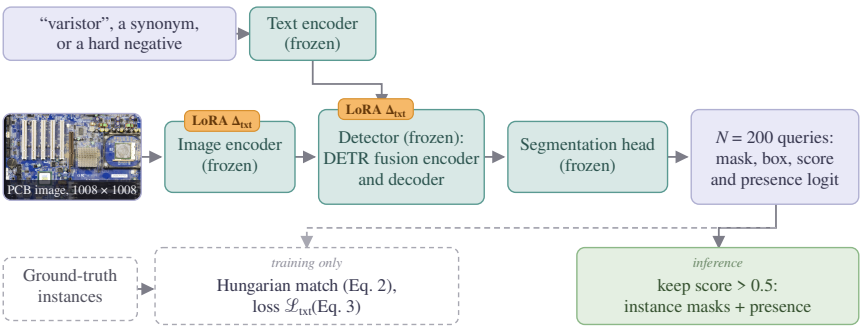


Figure 2: **Concept-route adaptation (Δ_{txt}).** Teal marks a frozen SAM3 module, lavender a tensor, dashes a training-only path; the orange badges are the rank-8 LoRA residuals, which wrap the attention projections of the last eight vision blocks and of every detector layer. Every SAM3 weight stays frozen. In training the $N = 200$ queries are matched one-to-one to the ground-truth instances of the prompted category (Eq. (2)) and supervised by $\mathcal{L}_{\text{txt}}$ (Eq. (3)); at inference those above 0.5 are returned.

Two adaptation regimes. Let Δ be low-rank residuals on selected linear maps and $\mathcal{L}_k$ route k's objective. *Shared adaptation* trains one Δ on a mixture, $\min_\Delta \mathbb{E}_{k\sim q}[\mathcal{L}_k(\Delta)]$: cheapest in parameters, and what the multi-task tradition recommends (Caruana, 1997). *Route-disjoint adaptation* instead trains a family $\{\Delta_k\}$ with disjoint trainable supports in a fixed order π, each on its own objective with all earlier sets frozen but *active*:

$$\Delta_{\pi(t)} \leftarrow \arg\min_{\Delta}\ \mathcal{L}_{\pi(t)}\big(\Delta \,\big|\, \Delta_{\pi(1)}, \ldots, \Delta_{\pi(t-1)} \text{ frozen}\big)\,, \tag{1}$$

so route $\pi(t)$ is optimised on top of the representation its predecessors produced. A route omitted from the schedule we call *free-riding*: evaluated with the other adapters active, it contributes no objective and no parameters. We use $\pi = (\text{txt}, \text{geo})$ with the exemplar route free-riding, writing the trained sets Δ_{txt} and Δ_{geo}; because LoRA residuals are additive this costs nothing at inference beyond the route actually invoked.

Why the routes should be expected to conflict. Shared adaptation asks one residual to serve two opposing demands on the same features. The concept objective is a *selectivity* requirement: the presence logit for "capacitor" must respond to capacitors and not to the varistors beside them, so z must expose fine distinctions between look-alike categories. The geometric objective is an *invariance* requirement in the same variable: one decoder must turn a click into that object's boundary whatever the object is. Category identity is the signal for one route and the nuisance for the other. The two are not formally incompatible, but we doubt a low-rank residual on a few attention projections can arrange the separation while optimising both losses, and we expect the cost to fall on selectivity, since the pretrained features already satisfy invariance. The prediction is specific: under shared adaptation image-level *recognition* should degrade while per-object *localisation* stays intact.

4 Instantiation on SAM3

We adapt SAM3 under Eq. (1) with the schedule $\pi = (\text{txt}, \text{geo})$. Every residual is the standard low-rank update (Hu et al., 2022) $W = W_0 + \frac{\alpha}{r}BA$ with $r = 8$, $\alpha = 16$ and B zero-initialised; only placement and objective differ between routes.

4.1 Concept route: Δ_{txt}

Placement. LoRA wraps the query, key, value and output projections in the last eight of 32 vision blocks, the self- and image-cross-attention of all six fusion-encoder layers, and all attention of the six DETR decoder layers (Fig. 2). The text encoder, early vision blocks, every MLP and normalisation layer and the 2.3M-parameter segmentation head stay frozen, leaving 46 adapted modules holding 884,736 parameters, 0.103% of SAM3's 860.5M. The placement targets the two shifts we expect and no others: late vision blocks carry the appearance gap, detector attention the grounding gap of which

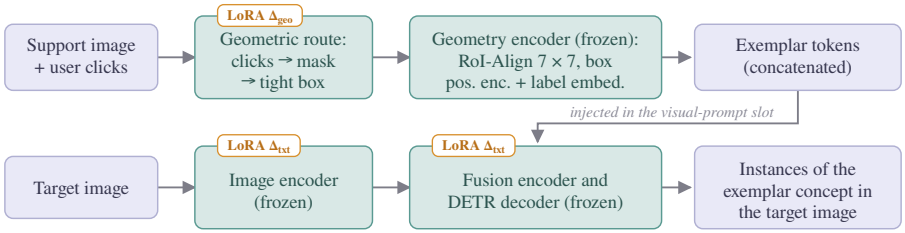


Figure 3: **Exemplar route: free-riding, extended across images.** Outlined badges are adapter sets loaded and *active* but not trained. The geometry encoder runs on the *support* image's features and its tokens enter a *different*, target image's fusion encoder where a text embedding would go. No parameter here is trained, which makes the route a clean probe of what Δ_{txt} changed (Sec. 6.3).

evidence a component name should attract. Freezing the segmentation head is a deliberate constraint, not an economy — unfreezing it is actively harmful (Sec. 6.1).

Instance-level supervision. We assign the $N = 200$ returned queries one-to-one to the M ground-truth instances of the prompted category with the Hungarian algorithm (Kuhn, 1955) under

$$\mathcal{C}_{ij} = -0.25\,\hat{p}_i + \mathcal{C}_{ij}^{\text{Dice}} + \|\hat{b}_i - b_j\|_1 + 1 - \text{gIoU}(\hat{b}_i, b_j), \tag{2}$$

with boxes in normalised centre–size coordinates and generalised IoU (Rezatofighi et al., 2019); unmatched queries are supervised as background. This is not a formality: supervising the *union* of same-category masks attains high pixel overlap but collapses instance AP to 0.005, because nothing pushes the queries apart (Sec. H.1).

Objective and prompts. The concept loss is

$$\mathcal{L}_{\text{txt}} = \mathcal{L}_{\text{BCE}} + 1.5\,\mathcal{L}_{\text{Dice}} + 0.1\,\mathcal{L}_{\text{box}} + 0.2\,\mathcal{L}_{\text{obj}} + 0.2\,\mathcal{L}_{\text{pres}}, \tag{3}$$

where mask BCE and Dice (Milletari et al., 2016) with box ℓ_1 and gIoU losses act on matched queries, object BCE supervises every query with unmatched negatives down-weighted by 0.1, and presence BCE trains SAM3's image-level presence logit. That last term is decisive rather than auxiliary: presence gates every detection score, so a model that cannot decide a concept is *absent* fires on look-alikes. Accordingly, with probability 0.3 a sample's category is replaced by an absent, preferably confusable one paired with an empty target set; categories are mildly rebalanced ($w_c \propto n_c^{-0.15}$) and each prompt is drawn afresh each epoch from the canonical name and a small synonym set. We train with AdamW at 10^{-4}, effective batch size two, and early stopping within 40 epochs (Secs. A and H.4).

4.2 Geometric route: Δ_{geo}

Following Eq. (1), the concept-trained vision adapters stay frozen and *active*, so this route inherits the adapted features, and rank-8 LoRA is inserted into all attention projections of the two-way mask decoder — $94{,}208$ further parameters, 0.011% of the model. Prompt encoder, backbone and neck stay frozen, so image features are computed once and reused across passes. Training rolls out a simulated correction session from one positive click inside a target mask, re-sampled each epoch; each pass receives every click so far and the previous pass's mask logits as a dense prompt, and one corrective click is sampled inside the larger of the remaining leak and miss regions provided it exceeds 100 pixels and 5% of the instance area. Each of the P passes contributes to

$$\mathcal{L}_{\text{geo}} = \frac{1}{P}\sum_{p=1}^{P}\Big(\mathcal{L}_{\text{focal}}^{(p)} + \mathcal{L}_{\text{Dice}}^{(p)} + \big(\hat{u}^{(p)} - u^{(p)}\big)^2\Big), \tag{4}$$

with sigmoid focal loss (Lin et al., 2017), Dice loss, and a squared error between the decoder's predicted-IoU token $\hat{u}$ and the realised mask IoU u — the term that turns the model's own quality estimate into a calibrated review signal (Sec. 6.4).

4.3 EXEMPLAR ROUTE: FREE-RIDING, EXTENDED ACROSS IMAGES

The exemplar route receives no adapters and no objective; we extend only its *scope*, since SAM3's exemplars are a within-image interaction while an annotator working through an image set needs propagation across images. Fig. 3 gives the pipeline. It works because both images pass through the same adapted backbone, and it retrieves instances exactly as text-prompted operation does, presence gating included.

5 EXPERIMENTAL PROTOCOL

5.1 A FINE-GRAINED BENCHMARK WITH TWO ZERO-SCORE CATEGORIES

We assembled 330 PCB images from three public datasets (Pramerdorfer & Kampel, 2015; Arbash et al., 2024; Nayak, 2021) and our own captures, and annotated 6,968 instance masks from scratch over seven mutually exclusive categories, so a film capacitor returned under the "capacitor" prompt scores as a false positive. The two categories zero-shot SAM3 never detects are the two defined by a fine-grained contrast, and both are among the rarest (Tab. 4). The corpus splits image-wise into 264 training and 66 validation images, rotations and flips expanding training threefold. Because the in-house photographs share one physical board, we reserve 44 boards from a separate industrial corpus as a strictly *board-disjoint* test set, used for neither training nor any selection or threshold decision and evaluated once, at the end. All inputs use SAM3's native 1008×1008 resolution (Secs. A and G). Corpus, adapters and code are released; the board-disjoint images are a partner's production data.

Metrics. Pairing the 66 validation images with all seven canonical labels gives 462 queries, 180 positive; the board-disjoint set gives 308 (138 positive). We adopt the three metrics of Carion et al. (2026). Counting only predictions above 0.5 and matching them one-to-one to the ground truth, *positive micro F1* (pmF1) is the F1 over positive queries averaged across mask-IoU thresholds 0.50–0.95 (localisation); *image-level Matthews correlation* (IL_MCC) scores the binary presence decision over all queries (recognition); and *classification-gated F1* is their product, cgF1 $= 100 \times$pmF1$\times$IL_MCC. COCO mask AP uses unfiltered predictions, which is why a category can hold non-zero AP while producing no detection above threshold. For the geometric route we add NoC@90, the mean clicks to 0.90 IoU.

Baselines. We compare against zero-shot SAM3 and three adapter families — DoRA (Liu et al., 2024a), AdaptFormer (Chen et al., 2022a), ConvPass (Jie & Deng, 2024) — each at its best reached configuration under the same data, prompts, matching and loss. A closed-set specialist such as Mask2Former (Cheng et al., 2022) is excluded by construction: with a fixed label set and no prompt interface, none of our four protocols can be run on it.

6 RESULTS

6.1 FINDING 1: ROUTES INTERFERE

The base model is silent, not merely weak. Out of the box, SAM3 fails in a way that is easy to attribute: it localises PCB parts competently (pmF1 0.635) while recognition collapses to IL_MCC 0.599, with no detection above threshold at all for film capacitors and varistors (Tab. 4 and Fig. 4). The benchmark therefore resolves the two factors Sec. 3 predicts a shared adaptation will trade off.

Mixing the two objectives into one residual. Tab. 2 places the prompt mixture beside the two other axes of the design space. The comparison is budget-matched: the final configuration and the mixed run wrap the same 46 attention modules at rank 8 — the identical 884,736 parameters — under the same matching cost and loss weights, differing only in the prompt distribution. Mixing 40% geometric prompts costs 20.5 cgF1; the pair came from development runs, not a pre-registered single-factor experiment (Sec. H.3).

Three properties matter more than the size. The effect is *localised as predicted* — pmF1 falls 5.6% relative, IL_MCC 32.9%: the mixed model still finds plausible components but can no longer decide

Table 2: **The concept-route design space** (462-query validation protocol). All rows share the corpus, matching cost and evaluation, varying along three axes: *where* the residual acts, *how much* of it there is, *which prompts* train it. The most damaging change is the prompt mixture, at the same sites, rank and parameter count as the final configuration: it keeps pmF1 while losing IL_MCC, the decomposition Sec. 3 predicts. The capacity rows are not budget-matched by construction (Secs. H.2 and H.3).

Concept-route design	Trainable	pmF1	IL_MCC	cgF1	AP
Placement					
Detector attention only	491,520	**0.697**	0.710	49.52	0.624
+ **last eight vision blocks (final Δ_{txt})**	**884,736**	0.643	**0.870**	**55.94**	**0.747**
Capacity					
+ unfrozen segmentation head	3,450,117	0.377	0.816	30.73	0.492
+ instance-contrastive auxiliary loss	884,736	0.560	0.833	46.69	0.644
Prompt mixture (shared instead of route-disjoint adaptation)					
60% text + 40% geometric prompts	884,736	0.607	0.584	35.43	—

Table 3: **Main concept-route results.** Adapter families at their best reached configuration under an identical recipe; every family repairs recognition, only LoRA also preserves localisation. The board-disjoint set is evaluated once, after all model selection.

Model	pmF1	IL_MCC	cgF1	AP
Validation split — 66 images, 462 queries				
SAM3 zero-shot	0.635	0.599	37.99	0.522
AdaptFormer (Chen et al., 2022a)	0.384	0.816	31.33	0.441
ConvPass (Jie & Deng, 2024)	0.390	0.843	32.88	0.466
DoRA (Liu et al., 2024a)	0.405	0.858	34.75	0.451
Ours (Δ_{txt})	**0.643**	**0.870**	**55.94**	**0.747**
Board-disjoint industrial test — 44 boards, 308 queries				
SAM3 zero-shot	0.465	0.480	22.31	0.482
Ours (Δ_{txt})	**0.705**	**0.725**	**51.16**	**0.731**

whether the concept is present, re-creating the zero-shot failure in a model that had already fixed it. It is monotone in the mixture (Sec. H.3), and large in this paper's units — vision adapters are worth $+6.4$ cgF1, so interference destroys three times their benefit. Two controls isolate the cause: a second objective *without* a second prompt type (an instance-contrastive term) costs only 9.3 cgF1, and the geometric route, trained separately on the mask decoder, leaves the concept route bit-identical. The problem is not extra objectives, but two competing for one residual.

Placement dominates capacity. The other two axes of Tab. 2 trace one line. Detector-only LoRA reaches 49.52 cgF1; adding vision adapters — $1.8\times$ the parameters — raises it to 55.94, so the appearance shift is not absorbed by grounding layers alone, though those checkpoints differ in minor training settings too. Capacity in the wrong place is another matter: unfreezing the 2.3M-parameter segmentation head nearly quadruples the budget and drops the model *below* zero-shot (Sec. H.2).

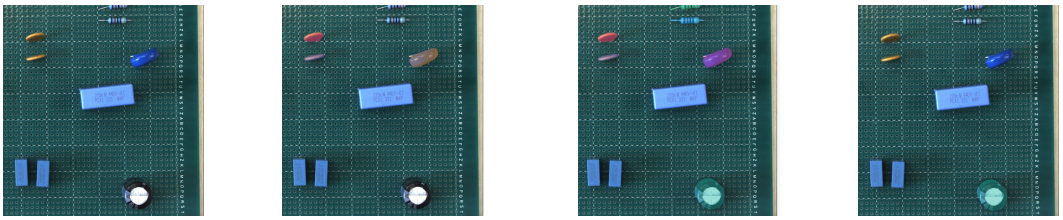
6.2 Route-disjoint adaptation repairs the concept route

With the routes separated, 0.103% of the parameters take cgF1 from 37.99 to 55.94 (Tab. 3). *Every* adapter family repairs recognition regardless of mechanism, so presence calibration is driven by the training signal — hard negatives and the presence term — not by adapter mechanics. Only LoRA also preserves localisation: the others fall below zero-shot pmF1, hitting the object but not its boundary (IoU 0.80: LoRA 0.700, DoRA 0.277; Sec. C). These are best reached configurations rather than a matched sweep (Sec. B), so this is a property of the checkpoints, not the mechanism.

Significance and decomposition. A paired bootstrap over the 462 queries (5,000 resamples, identical indices for both models) gives $+17.95$ cgF1, 95% CI $[11.9, 24.5]$, $p < 0.001$. Decomposing it repeats Sec. 6.1 with the sign reversed: the recognition gain is large and significant (IL_MCC $+0.271$, CI $[0.197, 0.348]$), the localisation change is not (pmF1 $+0.009$, $p = 0.68$). Adaptation repairs *whether* the correct concept fires, not pixel quality.

Table 4: **Per-category composition and results.** Top: what the benchmark contains and how many queries each category contributes (a query is one (image, category) pair, positive when the image contains that category). Bottom: zero-shot SAM3 against ours, better of each pair in bold. The two categories the base model never detects are the two defined by a fine-grained contrast and among the rarest; both become the strongest after adaptation. mIoU is the overlap between unions of predicted and target masks on positive queries, so a category never detected scores 0.

	Film cap.	Capac.	Connect.	Resistor	Screw	Varistor	Wire	**All**
Instances	329	2,853	2,100	729	604	199	154	**6,968**
Positive validation queries	25	45	37	26	13	20	14	**180**
Positive test queries	23	27	31	16	15	18	8	**138**
cgF1, zero-shot	0.00	44.06	25.97	44.83	73.00	0.00	64.27	37.99
cgF1, ours	**84.07**	**50.93**	**41.52**	**47.76**	**84.83**	**54.08**	**73.76**	**55.94**
Mask AP, zero-shot	0.200	0.614	0.603	0.508	0.937	0.029	**0.766**	0.522
Mask AP, ours	**0.895**	**0.692**	**0.673**	**0.509**	**0.964**	**0.767**	0.730	**0.747**
Mask mIoU, zero-shot	0.000	0.698	0.397	0.561	0.937	0.000	0.721	0.461
Mask mIoU, ours	**0.854**	**0.826**	**0.589**	**0.697**	**0.949**	**0.630**	**0.769**	**0.745**



(a) SAM3 (b) ours (c) SAM3 (d) ours

Figure 4: **Fine-grained recognition, in both error directions.** One scene, four prompted passes; a tint marks each predicted instance. (a, b) prompted "varistor"; (c, d) prompted "capacitor". The board carries three disc varistors (two gold upper left, one blue upper right), a film capacitor (large blue block) and an electrolytic capacitor (black disc). The base model returns nothing for "varistor" (a) where ours tints all three discs (b), and under "capacitor" it claims those same varistors (c) where ours does not (d).

Where the gain lives, and whether it generalises. Tab. 4 localises the gain and Fig. 4 shows both error directions on one scene: under-firing on categories the base model cannot name, over-firing on their look-alikes. On the board-disjoint boards, which informed neither selection nor threshold tuning, the relative gain is larger ($+129\%$) than on validation ($+47\%$), the zero-shot model degrading further out of domain than ours.

6.3 Finding 2: Representations transfer across routes

Nothing on the exemplar route is trained and no exemplar token was ever shown to Δ_{txt} (Fig. 3), so whatever it gains when Δ_{txt} is switched on must come from a change in the shared representation. It gains a great deal (Tab. 5, top): $+0.176$ pmF1 within-image and $+45\%$ relative across images, where exemplars from two support images per class are applied to disjoint targets. Its absolute level stays below the text route, but the per-category rows are decisive: the largest gains fall on exactly the categories the *text* route learned to discriminate. An alignment-only account cannot produce this, because the route uses no language.

One might object that this is a scoring problem any crop classifier could repair. The second row of Tab. 5 is that control: frozen DINOv3 (Siméoni et al., 2025) features, which separate these classes almost perfectly in isolation, repair much of the *zero-shot* model but add at most $+0.008$ after adaptation, across a full sweep (Sec. F).

6.4 Composition, retention, and cost

The geometric route composes with the concept route. The lower block of Tab. 5 carries the composition result: with a deterministic first click and up to seven error-driven corrections, Δ_{geo} cuts the residual first-click error by 47% and the clicks to 0.90 IoU from 1.50 to 1.16. The middle

Table 5: **The two routes not trained on their own objective.** Columns say which adapter sets are *active*, not which were trained for that row; dashes mark undefined cells. Δ_{geo} trains on a frozen, active Δ_{txt}, so its middle column separates the two contributions.

	SAM3	$+\,\Delta_{\text{txt}}$	$+\,\Delta_{\text{txt}} + \Delta_{\text{geo}}$
Exemplar route — free-riding, no trained parameters (pmF1)			
Within-image	0.518	**0.694**	—
Within-image + DINOv3 re-scoring	0.618	0.702	—
Cross-image, seven-category mean	0.337	**0.490**	—
film capacitor	0.350	**0.814**	—
varistor	0.257	**0.646**	—
capacitor	0.205	**0.477**	—
Geometric route — 1,291 validation instances, simulated clicks			
IoU after 1 click	0.792	0.842	**0.890**
Mean IoU over the sequence	0.831	—	**0.902**
Final IoU	0.866	—	**0.911**
NoC@90 ↓	1.50	—	**1.16**
IoU-token AUROC (flag true IoU <0.75)	0.795	0.783	**0.922**

column is the attribution — of that $+0.098$, the inherited representation contributes $+0.050$, the $94{,}208$ decoder parameters $+0.048$. Trained in sequence and kept disjoint, the two adapter sets compose rather than compete: the same pair of objectives that destroyed recognition when mixed into one residual. The predicted-IoU token agrees — only Δ_{geo} makes it trustworthy — so the loop can auto-accept confident masks and send only uncertain ones to a human.

Retention, robustness, and cost. Specialisation is neither brittle nor destructive of the general vocabulary: over five prompt styles ours averages 43.26 cgF1 against 22.98 zero-shot, on 100 COCO images it gives up 1.3 cgF1, and it costs $+5.7\%$ query latency ($+1.4\%$ merged) at under 4 MB per adapter set (Secs. D and K). It remains a seven-category specialist.

7 Discussion and Conclusion

What we claim, and how far. We claim a mechanism with a testable scope, not a magnitude that transfers: route interference should appear whenever two prompt interfaces make *opposing* demands on the same latent variable — here category identity, which the concept route must expose and the geometric route must ignore. Tab. 1 says where that is falsifiable: the single-route models should show no conflict under any prompt mixture, the three multi-route ones should show it. The second finding is evidence of a different kind: the free-riding route shows the residual has not merely re-aligned language with features that were adequate all along.

Limitations. Every measurement here is on SAM3: Tab. 1 argues the *question* generalises, not that the effect does. The interference claim rests on a single budget-matched pair from development runs rather than a pre-registered sweep, and the further mixed runs of Sec. H.3 are a trend only. We report no full fine-tuning baseline — 860M weights on 264 images exceeded our compute. The model is a single seed with bootstrap rather than cross-seed uncertainty, and validation drove early stopping, so the board-disjoint set carries the generalisation claim. The binding technical limitation is resolution: at 1008^2 small and thin parts occupy few backbone tokens and validation recall falls from 0.96 to 0.53 across size bands, with tiled inference and mask refinement (Ke et al., 2023) the natural remedies. Finally, the vocabulary covers seven categories and not unseen component types, the exemplar route rejects absent concepts less reliably, and one person annotated the corpus.

Conclusion. A model with several prompt interfaces over one backbone should be adapted route by route: mixing their prompts into one residual costs 20.5 cgF1 at a matched budget, while a route that is never trained gains 45% from text-only adapters.

AI Use Statement

In this work, we used generative AI tools (large language model assistants) for *implementation* and for *data analysis and presentation*: suggesting, refactoring, and debugging training and evaluation code,

and producing plotting and table-formatting scripts for results computed by that code. We also used them for tasks in the recommended-disclosure category: drafting, restructuring, and copy-editing the text of this paper, and preparing figures and tables. We did *not* use generative AI to generate synthetic data, to produce or modify any experimental result, to formulate the research questions or hypotheses, to design the adaptation method or the evaluation protocol, or to write or verify any image annotation; theoretical-model and mathematical-claim generation is not applicable to this work. All AI-assisted code was read, tested, and corrected by the authors, and every number reported in this paper was produced by executing that code on the stated data and re-checked against the raw experiment logs. The authors reviewed all AI-assisted text and take full responsibility for the final content of this work, including its text, claims, and artefacts.

Ethics Statement

This work uses images of printed circuit boards and contains no human subjects, no personally identifiable information, and no demographic attributes. Of the four public image sources we build on (Pramerdorfer & Kampel, 2015; Arbash et al., 2024; Nayak, 2021), we reuse only the raw imagery, under the respective original licences, and contribute entirely new annotations. The 44 board-disjoint test images and the boards used for the annotation observation are an industrial partner's production data, used with permission and not redistributable; this is stated wherever those results appear (Secs. 5.1, 6.2 and L). The base model's weights are gated by their authors and are not redistributed with our release. We see two foreseeable risks. First, an annotation assistant that is confidently wrong can propagate label errors at scale; we address this directly by calibrating the model's own quality estimate for review triage (Sec. 6.4) and by reporting where the adapted model is unreliable (Sec. 7). Second, adaptation narrows a general model onto a specific taxonomy; we measure rather than assume the retention of its prior open-vocabulary behaviour (Sec. 6.4) and report that specialisation on unseen component types is not achieved. The application domain — inspection, repair, and electronics recycling — carries no dual-use concern we are aware of.

Reproducibility Statement

The adaptation is specified in full in Sec. 4: adapter placement and rank in Secs. 4.1 and 4.2, the matching cost in Eq. (2), both training objectives in Eqs. (3) and (4), and the prompt-construction and optimisation settings in the text of Sec. 4.1. The evaluation protocol, including the split construction, the fixed operating threshold, and the definition of every metric, is given in Sec. 5. Sec. B reports the exact trainable-parameter counts per configuration, Sec. I the training curves and selected epochs, and Secs. F to H and J the supporting studies behind claims made in the main text. The 330-image corpus with all $6{,}968$ masks, the adapter weights for both routes, and the training and evaluation code are publicly released; the links are withheld from this anonymous submission and will be included in the camera-ready version. Results on the 44 board-disjoint industrial images (Tab. 3, bottom) are the one set that cannot be reproduced from public data, for the licensing reason stated in Sec. 5.1; every ablation, both central findings, and all in-distribution results use the released corpus only.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7319–7328, 2021.

Elias Arbash, Margret Fuchs, Behnood Rasti, Sandra Lorenz, Pedram Ghamisi, and Richard Gloaguen. PCB-Vision: A multiscene RGB-hyperspectral benchmark dataset of printed circuit boards. *IEEE Sensors Journal*, 24(10):17140–17158, 2024.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research (TMLR)*, 2024.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. SAM-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3367–3375, 2023.

Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1300–1309, 2022b.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *Conference on Language Modeling (COLM)*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 709–727, 2022.

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 (ECCV)*, pp. 38–57, 2024.

Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, 2024.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-SAM: Segment and recognize anything at any granularity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

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 (CVPR)*, pp. 10965–10975, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Anglin Liu, Rundong Xue, Xu R. Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. MedSAM3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025.

Changjing Liu, Yiming Huang, Beilei Cui, Liangjing Shao, Long Bai, Yanheng Li, Haoxuan Che, and Hongliang Ren. Parameter-efficient adaptation of SAM3 for prompt-driven surgical concept segmentation. *arXiv preprint arXiv:2607.23694*, 2026.

Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22290–22300, 2023.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, 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 (ECCV)*, pp. 38–55, 2024b.

Hangwei Lu, Dhwani Mehta, Olivia P. Paradis, Navid Asadizanjani, Mark Tehranipoor, and Damon L. Woodard. FICS-PCB: A multi-modal image dataset for automated printed circuit board visual inspection. *Cryptology ePrint Archive, Paper 2020/366*, 2020.

P. Malaisree, S. Youwai, S. Janrungautai, D. Amorndechaphon, P. Rojanavasu, and W. Songkitti. SAM3-LoRA: Parameter-efficient adaptation of a concept-promptable foundation model for multi-class structural defect segmentation. *arXiv preprint arXiv:2609.00469*, 2026.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

Animeshkumar Nayak. Printed circuit board (PCB) component detection. Kaggle dataset, 2021. https://www.kaggle.com/datasets/animeshkumarnayak/pcb-fault-detection.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 487–503, 2021.

Christopher Pramerdorfer and Martin Kampel. A dataset for computer-vision-based PCB analysis. In *Proceedings of the 14th IAPR International Conference on Machine Vision Applications (MVA)*, pp. 378–381, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li, Kent Yu, and Lei Zhang. DINO-X: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Konstantin Sofiiuk, Ilya A. Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 3141–3145, 2022.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 9120–9132, 2020.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11293–11302, 2019.

Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 102:103547, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1020–1031, 2023.

Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

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 (NeurIPS)*, 2023.

A Benchmark Details

Splits and per-category validation counts. The 330 base images split image-wise into 264 training and 66 validation images; offline rotation and flipping expand the training partition to 792 images, and validation is unaugmented. The board-disjoint industrial test set comprises 44 boards captured under production conditions at roughly 2,000 px on the long edge and annotated from scratch in the same taxonomy. No board in it appears in any training, selection, or case-study data. Per-category query counts for both splits are in Tab. 4.

Annotation quality control. A single annotator labelled all 6,968 instances and enforced consistency by repeated review rather than multi-annotator consensus: after an initial pass, random samples of 40–50 cropped instances per category were re-inspected on at least three separate occasions, with mislabelled or imprecisely masked instances corrected — particularly for the visually similar capacitor, film-capacitor, and varistor classes that carry the fine-grained contrast. The absence of an inter-annotator agreement study is recorded in Sec. 7.

Prompt synonym sets. Training draws each positive sample's prompt uniformly from the canonical name and the synonyms in Tab. 6, re-sampled every epoch, so one visual concept is paired with several surface forms without enlarging the dataset. Validation always uses the canonical name; the "unseen synonym" row of Tab. 8 draws from plausible alternatives held out from this table, and partly conflates wording changes with genuine concept drift ("ceramic capacitor" names a component type absent from the corpus), so its low scores reflect a shifted target as well as brittleness.

Table 6: Synonym sets used for on-the-fly prompt augmentation during training.

Category	Training synonyms
film capacitor	plastic film capacitor, box film capacitor
capacitor	capacitor component
connector	pin connector, pin header
resistor	through hole resistor, axial resistor
screw	mounting screw
varistor	mov, metal oxide varistor, disc varistor
wire	lead wire, jumper wire

B Trainable-Parameter Breakdown

Tab. 7 counts the trainable parameters of every configuration evaluated in Sec. 6, taken directly from the saved checkpoints. The full SAM3 model, including the interactive predictor, comprises 860,542,648 parameters. The two route-disjoint sets together account for 978,944 parameters, 0.114% of the model. The table also makes explicit why the adapter-family comparison of Tab. 3 is a between-configuration comparison and not a matched sweep: the DoRA and ConvPass runs additionally fine-tuned detection-head tensors, and the DoRA run omitted vision-encoder adapters, so those rows differ from ours in more than the adapter mechanism.

C Threshold-Resolved Localisation Across Adapter Families

The adapter-family comparison of Tab. 3 reports pmF1 averaged over IoU thresholds 0.50–0.95, which hides where the alternatives fail. Resolved by threshold, at IoU 0.50 all adapted models are comparable — LoRA 0.751, ConvPass 0.706, DoRA 0.699 — but at IoU 0.80 LoRA retains 0.700 while DoRA falls to 0.277 and ConvPass to 0.273. The alternatives therefore hit the object but produce masks that are too coarse at tight thresholds. A plausible explanation is that bottleneck branches and DoRA's magnitude rescaling perturb the frozen features consumed by the untouched

Table 7: Trainable parameters for the best configuration of each family. † These runs additionally fine-tuned detection-head tensors on top of their adapters, which is why their totals exceed ours despite comparable or fewer adapter sites.

Configuration	Adapted modules	Parameters	Share
Δ_{txt} (concept route)	vision encoder, 16 modules	393,216	
	DETR encoder/decoder, 30 modules	491,520	
	deployed total	**884,736**	**0.103%**
Δ_{geo} (geometric route)	mask decoder, 28 modules	94,208	0.011%
DoRA†	DETR transformer only, 30 modules	2,821,121	0.34%
ConvPass†	parallel convolutional bypasses	3,220,657	0.38%
AdaptFormer	parallel MLP bottlenecks	≈1.45M	0.17%
Unfrozen segmentation head	seg. head, + auxiliary adapter	3,450,117	0.41%

segmentation head more aggressively than LoRA's zero-initialised low-rank residual does, trading boundary fidelity for recognition. Since the configurations also differ in whether detection-head tensors were trained (Sec. B), we report this as a property of the checkpoints obtained rather than an isolated effect of the adapter mechanism.

D Prompt Robustness in Detail

Tab. 8 reports cgF1 under five prompt sets spanning the plausible range of operator inputs, with everything else identical to the main protocol. Three patterns emerge. Trivial variation is free: capitalisation leaves our model unchanged, and the "MOV" abbreviation is handled without loss — though, unlike pure capitalisation, "MOV" also appears among the training synonyms, so its robustness partly reflects exposure rather than invariance. Synonym augmentation pays off as intended: on synonyms seen in training our model retains 95% of its canonical score while zero-shot SAM3 loses about 40%. Both models degrade on inputs outside the trained vocabulary, though ours stays roughly triple the zero-shot score throughout. Averaged over the five sets, ours reaches mean cgF1 43.26 (worst case 23.18) against 22.98 (worst case 7.80). The last row of Tab. 8 is the retention check: on a 100-image sample of COCO `val2017` (Lin et al., 2014) scored with the same protocol, cgF1 moves from 53.82 to 52.53. That rules out gross forgetting (Biderman et al., 2024) but, on a sample this small and without confidence intervals, it cannot establish parity — it shows only that adaptation does not grossly degrade natural-image performance.

Table 8: cgF1 under prompt variation and on a 100-image COCO retention sample. Specialisation is bought with 1.3 points of natural-image cgF1.

Prompt / evaluation set	Zero-shot SAM3	Ours
Canonical PCB labels	37.99	**55.94**
Capitalisation / abbreviation	37.08	**55.94**
Seen training synonym	23.36	**52.95**
Descriptive phrase	8.66	**28.29**
Unseen synonym	7.80	**23.18**
Mean over the five PCB sets	22.98	**43.26**
COCO natural-image concepts (100 images)	**53.82**	52.53

E Calibration of the Predicted-IoU Token

Tab. 9 decomposes the calibration result of Sec. 6.4 across the three configurations, on a single deterministic click per instance over the 1,291 validation instances. Two effects separate cleanly. The concept adapters improve mask quality but barely change the token's *ranking* ability (Spearman $0.69 \rightarrow 0.69$, review AUROC essentially flat); they reduce calibration error only by moving the true IoU closer to the already-high predicted values. It is the mask-decoder adapters that make the token reliable: correlation with true IoU rises sharply, calibration error falls to 0.016, and the AUROC for

flagging masks below 0.75 true IoU reaches 0.922. At that AUROC, a mask that genuinely needs review receives a lower predicted IoU than an acceptable one in 92% of random review-versus-accept pairs — a ranking statistic, not a per-mask accuracy.

Table 9: Calibration of the interactive predictor's IoU token. ECE: expected calibration error (lower is better). $\text{AUROC}_{\text{review}}$: flagging masks with true IoU below 0.75 (higher is better).

Configuration	Pearson	Spearman	ECE ↓	$\text{AUROC}_{\text{review}}$ ↑
Unmodified SAM3	0.54	0.69	0.066	0.795
+ concept adapters Δ_{txt}	0.52	0.69	0.025	0.783
+ mask-decoder adapters Δ_{geo}	**0.73**	**0.83**	**0.016**	**0.922**

F Exemplar Route: Controls and Sweeps

The two exemplar protocols. *Within-image*: for each of the 180 positive validation image–category pairs, the largest ground-truth instance of that category supplies the exemplar box and the model must retrieve all instances of the category in the same image. *Cross-image*: the ground-truth boxes of two support images per category supply the exemplars, and the category is retrieved on held-out target images that are never used as support. Every configuration sees the identical exemplar set, so differences are attributable to the model alone. The two protocols are not comparable with each other or with the text-prompted results of Tab. 3: the prompt modality, the pair set (positive-only here), and the prompt-encoding path all differ.

DINOv3 is discriminative on this domain in isolation. Embedding $2{,}503$ ground-truth component crops with a frozen DINOv3 ViT-S (Siméoni et al., 2025) yields clearly separated class clusters under t-SNE, class-centroid cosine similarities at or below 0.70 off-diagonal, and a nearest-neighbour probe that retrieves a same-class crop as top-1 for 98.4% of crops. If the exemplar route's residual errors after adaptation were recognition errors, re-scoring with these features should repair them.

It does not, once the adapters are active. Tab. 10 reports both integration mechanisms across their full sweeps. On the frozen base model, hard similarity filtering helps substantially ($0.518 \rightarrow 0.618$), confirming that zero-shot exemplar scores are semantically unreliable here. With Δ_{txt} active, hard filtering never exceeds the unfiltered baseline by more than $+0.004$ at any threshold and degrades steadily as the threshold rises, while soft score fusion $s = (1-\alpha)s_{\text{SAM3}} + \alpha\,\text{sim}_{\text{DINO}}$ peaks at a marginal $+0.008$ at $\alpha = 0.4$ and collapses for larger weights. The class semantics an external encoder supplies are redundant with what the adapters injected; the residual errors are localisation errors that a crop-level similarity cannot repair. The final system therefore uses SAM3 exemplar tokens alone, avoiding a second backbone, a second forward pass per candidate crop, and a fusion hyperparameter.

Table 10: DINOv3 re-scoring of the exemplar route (pmF1, 180 shared within-image exemplar prompts). Left: hard similarity-threshold filtering. Right: soft score fusion. The unfiltered Δ_{txt} baseline is 0.6938.

Hard filter threshold	0.35	0.40	0.45	0.50	0.55	0.60
pmF1 with Δ_{txt}	0.6974	0.6962	0.6918	0.6763	0.6444	0.5948

Fusion weight α	0.1	0.2	0.3	**0.4**	0.5	0.7	0.9
pmF1 with Δ_{txt}	0.6963	0.6989	0.6991	**0.7016**	0.6983	0.6004	0.4667

Per-category cross-route transfer. The cross-image protocol of Sec. 6.3 uses ground-truth boxes from two support images per category as exemplars and retrieves that category on disjoint target images never used as support. Averaged over the seven categories, exemplar-only pmF1 rises from 0.337 to 0.490. The gains concentrate on the fine-grained categories — film capacitor $0.350 \rightarrow 0.814$, varistor $0.257 \rightarrow 0.646$, capacitor $0.205 \rightarrow 0.477$ — mirroring the per-category pattern of the text route (Tab. 4) even though no exemplar token was ever used in training. Absolute accuracy stays

below the text route because appearance differs across boards and because the exemplar pass proposes candidates even when the concept is absent; image-level presence rejection is the route's weak point, which is why it is best invoked after presence has been confirmed.

G HIERARCHICAL VERSUS FLAT CAPACITOR SCORING

The mutually exclusive taxonomy of Sec. 5.1 scores a film capacitor returned for the generic "capacitor" prompt as a false positive. Tab. 11 quantifies the alternative hierarchical convention, in which film capacitors also count as valid capacitor detections, by re-scoring the same predictions against the union of the capacitor and film-capacitor ground truth with everything else identical. For our model the hierarchical reading *lowers* cgF1 by 9.0 points, and the drop is localised to image-level presence (IL_MCC $0.76 \rightarrow 0.65$) rather than mask quality (pmF1 $0.67 \rightarrow 0.64$): on boards containing only film capacitors the adapted model correctly stays silent under "capacitor," which the hierarchical convention then penalises as a miss. Zero-shot SAM3, which conflates the classes, instead *gains* 1.4 points. The mutually exclusive convention therefore mirrors behaviour the adapted model actually learned, and relaxing it would reward the look-alike confusion the adaptation was designed to remove.

Table 11: Flat (mutually exclusive) versus hierarchical scoring of the "capacitor" prompt.

Model	Scoring	pmF1	IL_MCC	cgF1
SAM3 zero-shot	flat (mutually exclusive)	0.685	0.643	44.06
	hierarchical (film $\subseteq$ cap.)	0.654	0.695	45.46
Δ_{txt} (ours)	flat (mutually exclusive)	0.671	0.760	50.93
	hierarchical (film $\subseteq$ cap.)	0.643	0.652	41.93

H NEGATIVE RESULTS AND SUPPORTING ABLATIONS

This appendix collects the dead ends encountered while developing the recipe of Sec. 4. They are exploratory rather than strictly controlled: each run typically differs from the final configuration in more than one hyperparameter, and we report them as the lessons that shaped the design, not as isolated one-factor studies. Unless noted, all values use the validation protocol of Sec. 5, and the reference is the final Δ_{txt} (cgF1 55.94, AP 0.747, IL_MCC 0.870).

H.1 UNION-MASK SUPERVISION COLLAPSES INSTANCE SEPARATION

The earliest concept objective supervised, for each category prompt, the *union* of all instances of that category as one binary mask under pixel-wise BCE and Dice. This optimises pixel overlap but gives the model no signal to separate or count instances, and the result is a metric trap (Tab. 12): union-trained models reach mask mIoU up to 0.80 and Dice up to 0.89 while instance-level mask AP collapses to 0.005 for the detector-only variant. Even the best-tuned union run — which enabled the vision-encoder adapters, unfroze the segmentation head, and added offline augmentation and class balancing, roughly 3.2M trainable parameters — plateaued at AP 0.564, below what per-instance matching achieves with far fewer parameters and a frozen head. This is why Sec. 5 reports cgF1 and mask AP rather than pixel mIoU, and why Eq. (2) supervises instances.

Table 12: Union-mask supervision inflates pixel metrics while instance AP collapses. The per-instance model is not trained against a union target, so union-style pixel metrics are not comparable for it (—); the decisive column is mask AP.

Objective	Trainable	mIoU	Dice	AP
Union mask, detector only	491,520	0.789	0.882	0.005
Union mask, + vision enc. + seg. head	$\approx$3.2M	0.802	0.890	0.564
Per-instance matching (Δ_{txt})	884,736	—	—	**0.747**

H.2 More trainable capacity does not help

Two attempts to add capacity on top of Δ_{txt} both regressed (Tab. 13). A higher-capacity variant unfroze the segmentation head, initialised from the Δ_{txt} checkpoint. That head is where essentially all the added capacity sits: 2.30M of the 2.57M parameters this run trains on top of Δ_{txt}, the remainder being an auxiliary multi-scale adapter (266,500 parameters) carried over from the same experiment. At 3.45M trainable parameters, about four times our budget, it dropped cgF1 from 55.94 to 30.73 and AP from 0.747 to 0.492: a trainable segmentation head destabilises the calibrated localisation and presence behaviour that the frozen pretrained head otherwise supplies. An auxiliary instance-contrastive loss — pulling same-instance query embeddings together and pushing different instances apart — also failed to improve on plain matching, reaching cgF1 46.69 / AP 0.644; the set-prediction objective already supplies sufficient instance separation. Together with Sec. H.1, where unfreezing the head was necessary merely to reach AP 0.564, these runs place the lean adapter-only recipe at or beyond the useful capacity sweet spot for this data scale.

Table 13: Higher-capacity variants of the concept route.

Variant	Trainable	pmF1	IL_MCC	cgF1	AP
Δ_{txt} (reference)	884,736 (0.10%)	**0.643**	**0.870**	**55.94**	**0.747**
+ unfrozen segmentation head	3,450,117 (0.41%)	0.377	0.816	30.73	0.492
+ instance-contrastive loss	884,736 (0.10%)	0.560	0.833	46.69	0.644

H.3 Prompt mixture: the full set of runs

Tab. 2 in the main text reports the matched comparison; Tab. 14 records the full group of runs from which it was drawn, and how that group was found. A recurring failure during development was image-level recognition collapsing while per-positive localisation stayed intact. Tracing it back identified a single dominant cause: whether the concept path was trained on text prompts alone or on a mixture of text and geometric (box and negative-box) prompts. Several runs interleaved geometric prompts into concept batches — only 60–70% of samples used the category text — on the intuition that multi-prompt training would improve robustness. Because promptable concept segmentation is evaluated with text, this instead diluted the text-conditioned signal and miscalibrated the presence decision. Secondary choices compounded the collapse in individual runs, and Tab. 14 names them per row: these four are exploratory development runs rather than a designed sweep, which is why the quantitative claim in the main text rests on the matched pair (rows 1 and 3) alone. The group is nonetheless consistent and monotone — every text-only run reaches IL_MCC ≥ 0.71, every mixed run falls below 0.60.

Table 14: Effect of mixing geometric prompts into concept training. Rows 1 and 3 are matched in parameter count and adapter placement; rows 4–6 each additionally vary the secondary setting named. Localisation (pmF1) stays non-trivial throughout while recognition (IL_MCC) collapses.

Concept-training prompts	Secondary setting	IL_MCC	cgF1	pmF1
Text only	final placement (Δ_{txt})	**0.870**	**55.94**	0.643
Text only	detector attention only	0.710	49.52	**0.697**
60% text + geometric	—	0.584	35.43	0.607
60% text + geometric	presence weight 0.15	0.256	9.35	0.365
70% text + geometric	+ box-regression penalty	0.291	11.05	0.380
70% text + geometric	rank 4, last 2 blocks	0.246	8.06	0.328

H.4 Batch size and class balancing

Two further choices were fixed by direct comparison on an earlier version of the pipeline (union objective), so pixel mIoU is reported alongside cgF1 and AP; the qualitative conclusions carried over. Tab. 15 reports both comparisons. Halving the effective batch size from 4 to 2 with two-step gradient accumulation improved every metric substantially (mIoU $0.446 \rightarrow 0.802$, cgF1 $35.3 \rightarrow 58.5$): with a few hundred training images and a DETR-style set-prediction loss, smaller and more frequent

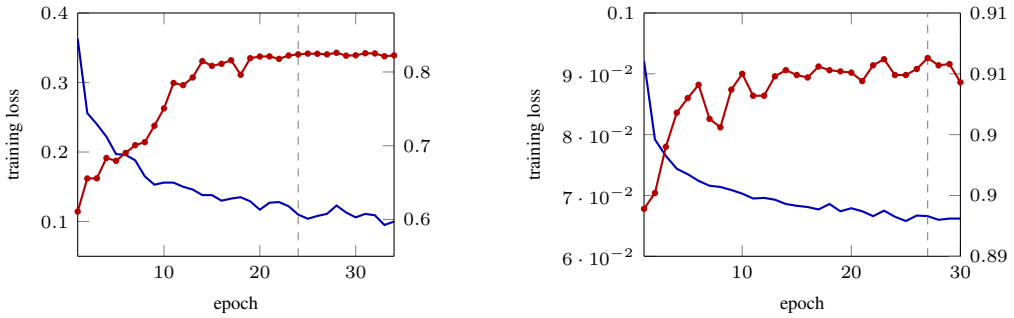
updates were markedly more stable. For the class-balance power $w_c \propto (1/n_c)^p$, milder balancing was better ($p = 0.3$ over $p = 0.6$: cgF1 $46.6 \rightarrow 58.5$), since aggressively upsampling rare categories starved the frequent ones; the final recipe softens this further to $p = 0.15$. The learning rate (10^{-4}, AdamW with weight decay 0.01, one warm-up epoch, cosine decay to $0.2\times$) was fixed early, as 2×10^{-4} trained less stably at this data scale.

Table 15: Batch-size and class-balance ablations on the earlier union-objective pipeline; each block varies one setting with the others held fixed.

Ablated setting	mIoU	cgF1	AP
Effective batch size (class-balance power 0.3)			
batch size 4	0.446	35.26	0.452
batch size 2 (chosen)	**0.802**	**58.52**	**0.564**
Class-balance power (batch size 2)			
$p = 0.6$	0.762	46.62	0.454
$p = 0.3$ (better here; final recipe uses 0.15)	**0.802**	**58.52**	**0.564**

I Training Curves and Checkpoint Selection

Fig. 5 shows per-epoch training loss and validation quality for both routes. Concept adaptation saturates after roughly 20–25 epochs, so the 40-epoch budget is generous and early stopping selects a checkpoint well inside it (epoch 24). Geometric fine-tuning starts from an already strong baseline — the concept-adapted trunk — and plateaus from about epoch 15, with the best checkpoint at epoch 27 of a 30-epoch budget. Both budgets were fixed in advance to lie clearly beyond the observed plateaus.



(a) Concept route Δ_{txt} (40-epoch budget)

(b) Geometric route Δ_{geo} (30-epoch budget)

Figure 5: Per-epoch training loss (left axis, blue) and validation selection score (right axis, red); the dashed line marks the checkpoint kept. The selection score is the mean per-class mask IoU on the validation split for the concept route (a) and the mean instance IoU for the geometric route (b).

J Augmentation Policy

Three offline-augmentation variants were trained and evaluated in a secondary experiment on an extended corpus combining the 330-image corpus with 392 further industrial images under its own split (504 validation image–prompt pairs; aggregate numbers only). That combined-corpus model is separate from the model reported in Sec. 6, which is trained on the 330-image corpus alone. Tab. 16 shows that adding random scaling, or scaling plus HSV colour jitter, on top of rotation and flipping changes results only in the third decimal. Geometric augmentation is also the principled choice for this domain: boards have no canonical orientation, so rotations and flips generate valid views, whereas strong colour jitter risks blurring exactly the hue and texture cues that distinguish look-alike components. The final recipe therefore uses rotation and flipping only.

Table 16: Offline-augmentation variants on the extended corpus. Scale and HSV jitter add no measurable benefit over rotation and flipping.

Offline augmentation	pmF1	IL_MCC	cgF1	AP
rotation + flip	0.740	0.849	62.85	0.742
rotation + flip + scale	0.742	0.851	62.91	0.742
rotation + flip + scale + HSV	0.741	0.849	62.91	0.740

K Inference Cost

Measured on one NVIDIA RTX 4090 (24 GB) at 1008×1008 with fp16 autocast, averaged over 50 image–prompt queries after 10 warm-up queries, each configuration benchmarked in a separate process for an isolated peak-memory reading. A concept query — one image encoding plus one text-prompted concept pass — takes 195.2 ms for the frozen base model and 206.3 ms with active adapters ($+5.7\%$; throughput 5.12 against 4.85 queries per second), while peak GPU memory rises from 12.6 to 15.3 GiB. Both overheads come from the adapter wrappers bypassing the base model's fused attention kernel, not from the adapter arithmetic: two rank-8 matrix products per projection are negligible. Because LoRA is purely additive, the learned deltas can be folded into the frozen projections after training (Hu et al., 2022); the merged model runs a concept query in 197.8 ms, within 1.4% of the base model. Merging does *not* reclaim the memory, which stays at 15.5 GiB, since the attention path still does not fall back on the most memory-efficient kernel — which is also why a small latency residual remains. In the interactive regime an image is encoded once and queried with several concepts: encoding dominates a single query (143 ms) while each additional concept on the cached image costs about 51 ms, so amortised over seven concepts the effective cost falls to 71 ms per concept, a $2.7\times$ reduction over re-encoding the board for every query.

L Annotation in Practice: An Uncontrolled Observation

The application that motivates this work (Sec. 1) was exercised end to end. Thirty industrial motherboards — a separate set from the 44 board-disjoint test boards, and from every training and selection image — were fully instance-annotated with the adapted model in the loop: text-prompted pre-annotation through the concept route, then click-based correction of individual masks through the geometric route, with the predicted-IoU token of Sec. E deciding which masks were opened for correction at all. Most boards carried 35–60 components and took roughly 10–20 minutes each; a few sparse boards with fewer than ten components went faster. As a rough yardstick, manual polygon annotation for COCO has been reported at about 80 s per instance (Lin et al., 2014), which puts a 35–60 component board at roughly 47–80 minutes of manual work, so the observed times correspond to something on the order of a $2\times$–$8\times$ reduction.

We report this as an observation about the workflow, not as a result. It is one annotator, who also built both the model and the tool, with no control condition, no second annotator, and no independent audit of the masks produced; the per-instance COCO figure comes from a different domain, a different taxonomy, and a different interface, so the ratio above should be read as an order of magnitude and nothing finer. Establishing an annotation speed-up properly needs a controlled study with several annotators, a held-out quality audit, and a manual-only control arm, which we have not run. Nothing in Sec. 6 depends on this paragraph: it is here because it is the use case the adaptation was built for, and because a reader is entitled to know what happened when it was used.

M Route Workflows and Additional Qualitative Results

Fig. 6 illustrates the geometric route that Sec. 4 describes in text, and Fig. 7 shows cross-image propagation on an unseen industrial board series. In detail, the exemplar pipeline works as follows. On a support image the user clicks a component; the adapted interactive predictor converts the clicks into an instance mask whose tight bounding box becomes the exemplar. That box is encoded with SAM3's geometry encoder *on the support image's own backbone features*: the box region is RoI-aligned into a 7×7 feature patch and projected, a positional encoding of the normalised box coordinates and a positive-label embedding are added, and a small prompt transformer contextualises

the sequence into exemplar tokens. Tokens from any number of exemplars, across any number of support images, are concatenated. To search a target image, the cached tokens are injected into its detection pass as visual prompt embeddings — the same input slot SAM3 uses for native same-image exemplars, with zero positional embedding and a placeholder text prompt — so that the fusion encoder conditions the target's image features on the exemplar appearance and the DETR decoder retrieves all matching instances exactly as in text-prompted operation, presence gating and the 0.5 confidence threshold included. The concept adapters remain active throughout.

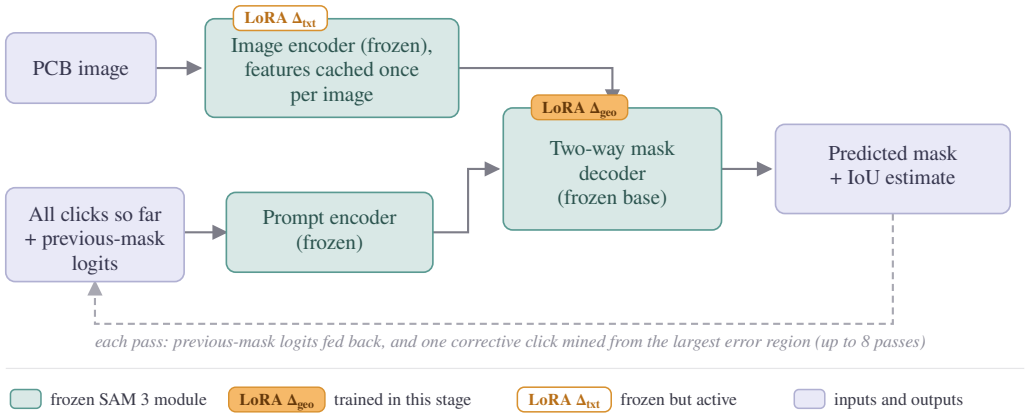
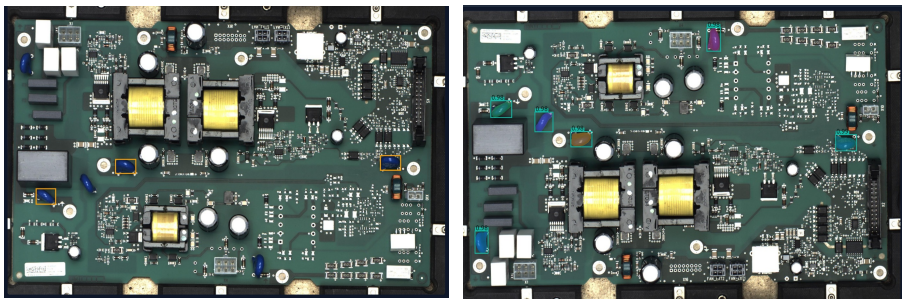


Figure 6: **Geometric-route training (Δ_{geo}).** Image features are computed once per image by the frozen backbone with the concept-route adapters Δ_{txt} loaded and active but not updated (outlined badge), and reused across all refinement passes. The two-way mask decoder receives every click accumulated so far together with the previous pass's mask logits, and the next click is mined from the larger of the remaining miss and leak regions. Only the mask-decoder residual Δ_{geo} (filled badge) receives gradients: $94{,}208$ parameters, 0.011% of the model.



(a) Support board, three clicked exemplars (b) Different capture, retrieved instances

Figure 7: **Cross-image exemplar propagation on an unseen industrial board series.** Three exemplar boxes from clicks on the support board (a) retrieve the matching components on a re-oriented capture (b) at ≥ 0.98 confidence, with no text prompt and no per-series training. Held-out cross-image numbers are more modest (Tab. 5); this is a favourable case.