PARTFACTOR: RECOVERING UNSEEN PARTS WITH DECOUPLED FOREGROUND PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary part segmentation remains challenging: models must distinguish visually similar parts, recover fine spatial boundaries and transfer to categories absent from task supervision. Existing approaches strengthen object–part correspondence and spatial aggregation, but a correctly ranked part can still disappear when recognition uncertainty suppresses foreground. We propose PartFactor, whose object-conditioned decoder first learns conditional part identity by combining structural features, image–text matching and relations among parts. With this decoder fixed, a compact visual head then learns foreground support. A background-logit reparameterization makes foreground probability depend on visual evidence alone, allowing the head to recover missing regions while preserving the decoder's conditional part distribution. Under the same full-image Pred-All evaluation, PartFactor improves unseen mIoU over a state-of-the-art baseline by 10.12 percentage points on PartImageNet and 23.79 points on PartImageNet-OOD. Fixed-decoder experiments trace the gains to recovery of correctly ranked parts, including at matched background false-positive rates. On PartImageNet, PartFactor needs only about 1/21 as many task-trainable parameters and achieves a 4.54× inference speedup over the same baseline.

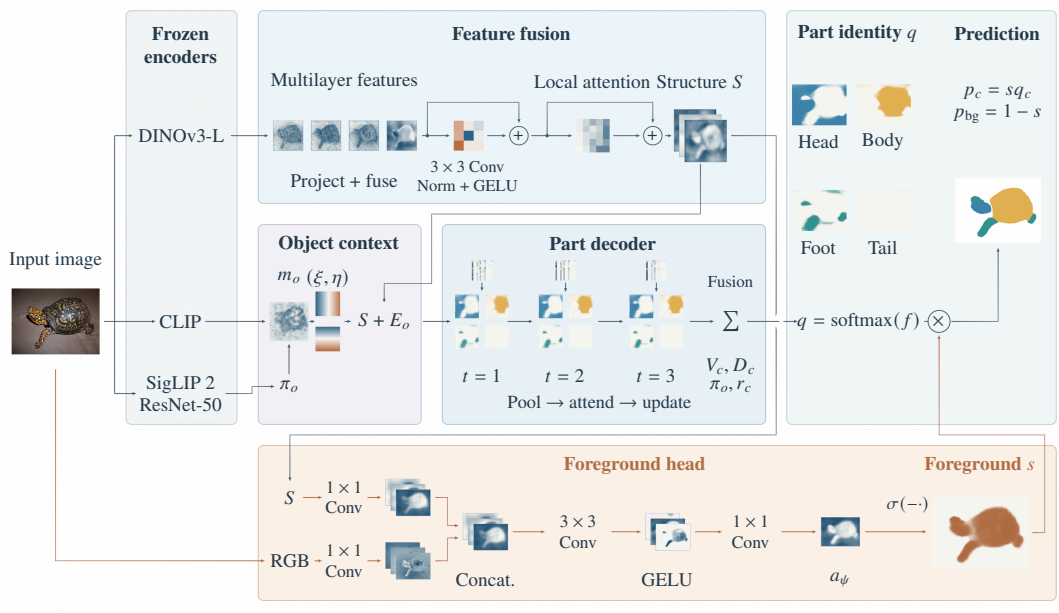


Figure 1: **PartFactor learns part identity and foreground support separately.** The object-conditioned decoder combines structural features, object context and three rounds of pooling, attention and part updates to produce q. The shared structural map S also supplies the foreground head. Once the decoder is trained, it is frozen together with the encoders; only the visual foreground head is trained in the second phase. Setting $b = A + a_\psi$ gives $s = \sigma(-a_\psi)$, and the outputs combine as $p_c = sq_c$ and $p_{\text{bg}} = 1 - s$. Feature and probability maps are measured on the same image; part order and colors are consistent across updates. Appendix J gives the layer and channel details.

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.

1 INTRODUCTION

Open-vocabulary part segmentation goes beyond recognizing an object: it must locate and name parts that may be absent from task annotations. A fish mask does not say where the head ends or which pixels belong to a fin. Vision–language models provide a vocabulary for this recognition problem (Radford et al., 2021), while part segmentation benchmarks require that vocabulary to be grounded in fine-grained image regions (He et al., 2022; Wei et al., 2023). Recent methods strengthen this correspondence through multi-granularity supervision (Sun et al., 2023; Choi et al., 2024), object–part hierarchies (Li et al., 2024; Miao et al., 2025), and spatial aggregation (Choi et al., 2025; Yin et al., 2026). HOPS further uses object attention and visual affinity to refine part predictions (Li et al., 2026).

A correct ranking among part names, however, does not guarantee a correct segmentation. The preferred part must also beat background. A fin can rank first among parts and still disappear from the output. This error differs from confusing a fin with a tail: the model has useful evidence for the correct identity but assigns too little probability to foreground. Figure 2 shows missing regions recovered without changing the part decoder. This distinction motivates learning conditional part identity and foreground support separately.

We propose PartFactor. First, an object-conditioned decoder learns a conditional distribution over part names. It combines structural image features and image–text matching in an object-relative coordinate frame, then iteratively updates part representations to capture their relationships. The pretrained image encoders remain frozen. This stage learns fine-grained part distinctions from source annotations while retaining the encoders' transferable visual and semantic features.

Next, a compact visual head learns foreground support while the trained part decoder remains fixed. The final distribution is $p_c = sq_c$, where q_c is the conditional probability of part c and s is foreground probability. To make s depend on visual evidence alone, we reparameterize the background logit to cancel the part-score contribution to foreground normalization. The head can then recover missing regions without changing the decoder's conditional part predictions. This step is essential: simply removing part scores from the head inputs leaves their influence through normalization. Both stages use source part annotations; Figure 1 shows the complete model.

PartFactor combines substantially better unseen-part segmentation with lower task-training and inference costs. Under the same full-image Pred-All evaluation, PI unseen mIoU rises from 35.18% with HOPS to 45.30% with PartFactor (+10.12 points). On OOD, it rises from 32.32% to 56.11% (+23.79 points).

On PI, PartFactor needs only 2.72M task-trainable parameters (HOPS: 56.83M), about $1/21$ as many, and delivers a 4.54× inference speedup. Task training takes 11.51 minutes. The OOD model uses 3.07M task-trainable parameters.

Controlled experiments explain how the foreground stage contributes. With the same trained decoder, the visual head raises PI unseen mIoU from 35.59% to 45.84% relative to a trained score-coupled head. At matched background false-positive rates, it recovers more correctly ranked unseen parts in all three seeds. Factorial experiments identify why head inputs and background parameterization must be considered together. OOD controls further separate visual evidence from nonlinear correction of part scores. The largest gains occur in larger unseen parts; image-level and part-area analyses show where the improvements are concentrated.

Contributions.

- **A factorized part segmentation model.** An object-conditioned decoder learns conditional part identity; a compact visual head then learns foreground support. Background-logit reparameterization removes the dependence of foreground probability on part scores.
- **An explanation of unseen-part recovery.** Fixed-decoder experiments distinguish background rejection from part confusion. Factorial controls and matched-error decision curves identify the roles of visual evidence, score inputs and normalization.
- **Unseen-part accuracy with efficient adaptation and inference.** PartFactor exceeds HOPS by 10.12 and 23.79 points in unseen mIoU on PI and OOD, respectively. On PI, it uses only about $1/21$ as many task-trainable parameters and runs 4.54× as fast.

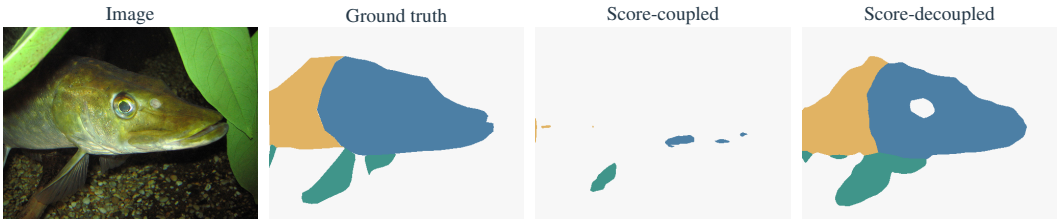


Figure 2: **Foreground prediction can discard useful part assignments.** Both outputs use the same trained part decoder. The score-coupled head retains only fragments of the tench body and fin; the score-decoupled head recovers their spatial extent. Part colors are consistent across columns. Figure 3 measures correct-part recovery over the full PI evaluation population.

2 Related work

Object–part correspondence. OV-PARTS studies transfer from open-vocabulary object recognition to fine-grained parts (Wei et al., 2023), while Going Denser learns dense correspondences from supervision at multiple granularities (Sun et al., 2023). PartCLIPSeg uses object context and attention control to reduce object confusion, overlapping masks and omitted parts (Choi et al., 2024). PartGLEE learns hierarchical object–part relationships through a query-forming module and separate object and part decoders (Li et al., 2024); LangHOPS grounds these hierarchies in a multimodal language model (Miao et al., 2025). PartCATSeg aggregates object and part matching costs with structural visual guidance (Choi et al., 2025). PBAPS progressively segments objects and parts and refines uncertain boundaries (Li et al., 2025), while HOPS uses object attention and visual affinity to filter and enhance part predictions (Li et al., 2026). These methods improve the evidence used to recognize and localize parts. PartFactor learns foreground support after part identity, so foreground training can recover regions while retaining the learned conditional part distribution.

Dense prediction from vision–language models. CLIP enables recognition through image–text matching (Radford et al., 2021). CLIPSeg adapts these features for dense prediction (Lüddecke & Ecker, 2022), and CAT-Seg aggregates matching costs while adapting CLIP (Cho et al., 2024). PCA-Seg separates spatial and class aggregation to reduce interference between contextual and semantic evidence (Yin et al., 2026). SAN learns mask proposals and attention biases alongside a frozen CLIP model, using its pretrained features for mask recognition (Xu et al., 2023). These approaches motivate efficient adaptation of pretrained representations. PartFactor combines frozen structural and vision–language features in an object-conditioned decoder and trains a separate, compact foreground head.

Foreground and part structure. DiSa uses saliency to disentangle foreground and background feature ensembles before refining and fusing them (Yao et al., 2026). Joint object–part representations address which object owns a part (de Geus & Dubbelman, 2024). PartFactor separates foreground probability from conditional part identity in the output distribution. Its background-logit reparameterization removes the part-score term from foreground normalization. Keeping the conditional decoder fixed then makes it possible to measure how foreground prediction changes the recovery of parts the decoder already recognizes.

3 Method

PartFactor first learns an object-conditioned part decoder, then trains a foreground head while keeping the decoder fixed (Fig. 1). We use *score-coupled* when foreground probability depends on part scores, and *score-decoupled* when it depends on visual features alone. This distinction includes both the head inputs and the background parameterization.

3.1 FACTORIZING THE SEGMENTATION DISTRIBUTION

Let x be an image, u a decoder-grid location and C the part vocabulary. Write $f_c(u)$ for the logit of part c and $b(u)$ for the background logit. The part log-normalizer is $A(u) = \log \sum_{c \in C} \exp f_c(u)$, and the conditional distribution over parts is $q_c(u) = \exp(f_c(u) - A(u))$. The categorical part–background distribution can be written exactly as

$$p_c(u) = s(u) q_c(u), \qquad p_{\text{bg}}(u) = 1 - s(u). \tag{1}$$

Here $s(u) = \sigma(A(u) - b(u))$ is the total probability assigned to parts. For $q_*(u) = \max_c q_c(u)$, the highest-ranked part strictly exceeds background when

$$s(u) > \frac{1}{1 + q_*(u)}. \tag{2}$$

For example, if $q_* = 0.1$, foreground probability must exceed 0.909. A correct ranking among parts is therefore not enough to produce the correct output label. Changing b with f fixed preserves every conditional ratio q_c / q_d on the decoder grid, while changing the total probability assigned to parts.

3.2 OBJECT-CONDITIONED PART DECODER

The decoder combines structural image features, text matching and object context (Fig. 1; Appendix J). Four DINOv3-L layers (Siméoni et al., 2025) are projected to 256 channels, spatially aligned and combined with learned softmax weights. A residual convolution followed by eight-head local attention over 5×5 neighborhoods produces the structural map S.

Dense CLIP features (Radford et al., 2021) provide matches to object and part text embeddings. SigLIP 2 (Tschannen et al., 2025) and an ImageNet-trained ResNet-50 supply an image-level object prior. For the four highest-scoring objects, soft masks define centered, scale-normalized coordinates. Encodings of the masks and coordinates are added to S, giving the part decoder an object-relative frame.

Projected part text initializes spatial part assignments within each object. Three iterations alternate soft-mask pooling, self-attention among part representations and updates to the spatial part scores. The decoder then combines CLIP matching scores, object log probabilities, conditional part log probabilities and text-matched structural features. It outputs part logits f and an initial background logit b_0; b denotes the background logit after foreground adaptation. Prediction compares all candidate parts with background.

We train the task-specific decoder on source part annotations while retaining the pretrained image encoders. This phase supervises annotated part pixels; foreground training subsequently includes pixels labeled as background (Appendix K). During foreground training, the encoders, object predictor and part decoder are all fixed. Each matched foreground comparison therefore uses the same conditional part distribution q before spatial interpolation.

3.3 SCORE-COUPLED SPATIAL FOREGROUND

A scalar shift of b_0 changes the global preference for foreground. A spatial head can also use visual features to distinguish locations with similar part scores:

$$\begin{aligned} b &= b_0 + g_\phi(S, R, e), \\ e &= [A - M,\ M - b_0,\ M,\ b_0], \qquad M = \max_c f_c. \end{aligned} \tag{3}$$

Here S is the structural map, R contains local RGB features and the four entries of e summarize the part scores and their relation to background. Learned 1×1 projections map S and R to 32 and 16 channels. After alignment, these features and e pass through a 3×3 convolution with 32 output channels, GELU and a scalar 1×1 output layer. The last layer is initialized to zero, so training starts at $b = b_0$. We train g_ϕ with categorical cross-entropy on source images; the scalar-shift control uses the same images and update count.

3.4 SCORE-DECOUPLED FOREGROUND BY LOGIT REPARAMETERIZATION

In Eq. 3, removing e would still leave foreground probability dependent on the part log-normalizer A. We remove the score inputs and set

$$b = A + a_\psi(S, R), \qquad s = \sigma\big(-a_\psi(S, R)\big). \tag{4}$$

The added A cancels the log-normalizer in $\sigma(A - b)$. At fixed visual features, changing the part logits can no longer change s, although it can change q. The final label still depends on both factors through Eq. 2. The visual head uses the same projections and convolutions as g_ϕ, without the four score inputs. Exact preservation of q applies on the decoder grid; interpolation of the joint posterior mixes neighboring distributions at the original resolution.

Parameterization and inputs. Let $h(u)$ denote the learned scalar correction, using either visual features and scores, $h = g_\phi(S, R, e)$, or visual features alone, $h = a_\psi(S, R)$. The two background parameterizations give

$$\begin{aligned} b = b_0 + h \quad &\Longrightarrow \quad s = \sigma(A - b_0 - h), \\ b = A + h \quad &\Longrightarrow \quad s = \sigma(-h). \end{aligned} \tag{5}$$

We call $b_0 + h$ the original background parameterization and $A + h$ the reparameterized form. Visual-only inputs leave the $A - b_0$ dependence in the first case. The second removes it, but remains score-coupled if h receives score inputs. The factorial experiments vary these choices separately.

We train a_ψ with binary foreground/background labels derived from source part annotations. With q fixed and identical pixel weights, categorical and binary cross-entropy give the same foreground-head gradients. The loss is computed on the decoder grid after resizing the scalar head output, but before joint-posterior interpolation (Appendix C). Uniform weighting averages valid-pixel losses. Balanced weighting gives foreground and background equal weight within each image, using whichever groups are present. We select balanced weighting on the PI source validation set and use uniform weights for the OOD parameterization and input comparisons.

Prediction at the original resolution. We compute Eq. 1 on the decoder grid, remove padding, bilinearly resize the probabilities to the original image dimensions and assign the highest-probability label. Appendix C gives the alignment, training and inference settings.

4 EXPERIMENTS

4.1 EVALUATION SETUP

We evaluate all 286 PartImageNet (PI) images, covering 92 seen and 55 unseen classes, and all 2,954 PartImageNet-OOD (OOD) images, covering 72 unseen classes (He et al., 2022; Li et al., 2026). Pred-All compares all 147 or 482 candidate parts, respectively, with background. Oracle-Obj restricts candidate parts to the true object category but retains background and full-image scoring. We pool confusion counts over original-resolution images, then average IoU over fixed class lists, retaining zero-IoU classes and false positives on background.

Each dataset uses its own source-trained part decoder. Table 1 compares source-trained systems under the same evaluation protocol. For PI, a separately trained PartFactor model is evaluated under both candidate sets; Appendix A reports the fixed-decoder foreground comparison. Three-seed experiments vary foreground-head initialization and image order at a fixed decoder. Appendices C and K specify training and model selection; published results under their original protocols are listed separately in Appendix E.

4.2 SEGMENTATION RESULTS AND FOREGROUND RECOVERY

PartFactor achieves the highest unseen mIoU among the compared methods on both datasets and under both candidate settings (Table 1). On PI, it reaches 45.30% under Pred-All and 47.68% under Oracle-Obj. The Pred-All result exceeds PartCATSeg by 9.60 points. On OOD, PartFactor reaches 56.11% under Pred-All and 56.27% under Oracle-Obj, exceeding HOPS by 23.79 points under Pred-All.

Table 1: **Full-image results under a shared evaluation protocol.** Values are percentages, with zero-IoU classes retained. Pred-All uses all candidate parts; Oracle-Obj uses parts of the true object category. Both retain background and use the same model within each method. PI has 92 seen and 55 unseen classes; OOD has 72 unseen classes. PartFactor uses score-decoupled foreground prediction on both datasets. PartFactor OOD values are means and sample standard deviations over three seeds. Bold marks the best value within each protocol.

PartImageNet (286 images)

Method	Pred-All					Oracle-Obj				
	Seen↑	Unseen↑	hIoU↑	mUSI↓	mOSI↓	Seen↑	Unseen↑	hIoU↑	mUSI↓	mOSI↓
PartCATSeg	43.50	35.69	39.21	40.34	26.69	48.36	44.64	46.43	35.43	14.62
HOPS	45.19	35.18	39.56	43.00	**20.93**	**51.03**	35.61	41.94	40.44	**9.68**
PartFactor	**49.52**	**45.30**	**47.32**	**27.23**	21.99	50.62	**47.68**	**49.11**	**25.65**	19.94

PartImageNet-OOD (2,954 images)

Method	Pred-All			Oracle-Obj		
	Unseen↑	mUSI↓	mOSI↓	Unseen↑	mUSI↓	mOSI↓
CAT-Seg	18.25	75.97	12.45	24.93	68.10	**6.21**
PartCATSeg	28.69	62.64	19.92	37.21	52.20	13.71
HOPS	32.32	61.92	10.80	36.19	51.03	8.40
PartFactor	**56.11** ±0.48	**31.40** ±1.08	**8.85** ±0.48	**56.27** ±0.48	**31.40** ±1.09	8.53 ±0.49

In the original PI comparison, replacing the trained score-coupled head with the source-selected score-decoupled head raises unseen mIoU from 35.59% to 45.84%. Uniform weighting already reaches 45.19%; foreground/background balancing adds 0.65 points (Appendix A). Unseen-part pixels assigned to background fall from 39.97% to 10.94%, while correctly labeled pixels rise from 51.11% to 76.31%. Wrong-part assignments also rise, from 8.92% to 12.75%: restoring foreground exposes both correct identities and remaining classification errors.

The 10.25-point gain has a paired-image 95% interval of [7.46, 12.36]. Among all images, 139 improve, 144 decline and three tie. Improvements average 13.83 points and losses average 4.23, giving a 4.59-point increase in mean image IoU; the median change is −0.02. Among the 111 images containing unseen parts, 78 improve, 30 decline and three tie. These image-weighted diagnostics differ from dataset mIoU, which pools pixels before averaging classes (Appendix H).

4.3 Recovering correct identities at matched background error

We vary the two sources of score dependence in Eq. 5: the offset $A - b_0$ and the head inputs. An independently trained PI decoder is fixed while parameterization, score inputs and pixel weights are crossed over three head seeds. With uniform weights, removing scores from $A + h$ raises unseen mIoU from 39.79% to 44.26%; removing them from $b_0 + h$ lowers it from 36.42% to 33.26%. The interaction between these effects is 7.63 points (95% paired-image interval [6.50, 9.17]). Balanced weighting gives the same interaction direction: 4.98 points [3.41, 6.49]. Table 10 reports all configurations and seed variation.

To identify predictions that are already correct among parts, let $r(u) = \arg\max_c q_c(u)$ and let $\mathcal{U}$ be the unseen-label set. We define the correct-part rejection rate on the fixed decoder grid as

$$D = \Pr[r(u) = y(u),\ \hat{y}(u) = \text{bg} \mid y(u) \in \mathcal{U}]. \tag{6}$$

Both D and correct unseen-part recall use all GT unseen-part pixels as their denominator. With $A + h$ and uniform weights, removing scores lowers D from 29.80% to 17.73% and raises correct recall from 54.19% to 66.26%, while background false-positive rate (FPR) falls from 5.41% to 5.23%. For every seed, conditional accuracy equals correct recall plus D. The three-seed means satisfy

$$\underbrace{83.99\%}_{\text{correct conditional identity}} = \underbrace{54.19\% + 29.80\%}_{\text{correct output + background rejection}} = 66.26\% + 17.73\%. \tag{7}$$

The visual head recovers 12.07 percentage points of correctly ranked unseen parts while reducing background FPR.

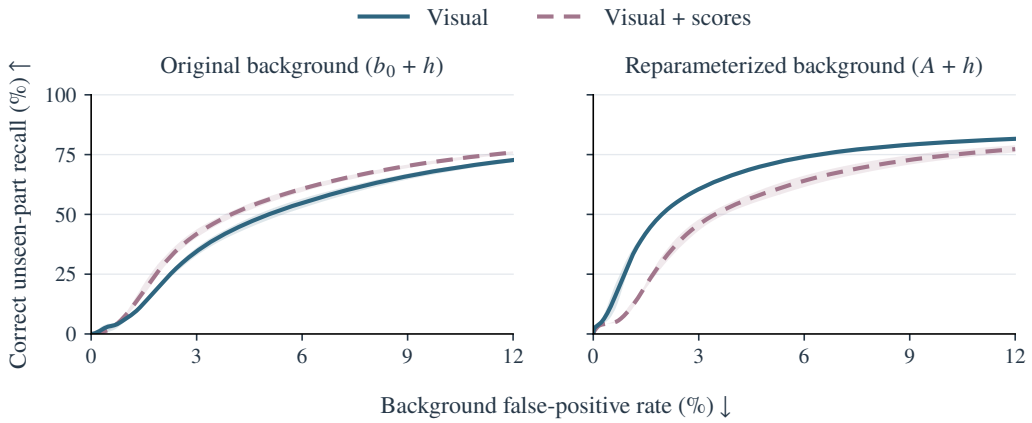


Figure 3: **Correct unseen-part recall at matched background false-positive rates.** The two background parameterizations reverse the effect of score inputs: visual-only inputs improve recall with $A + h$ but reduce it with $b_0 + h$. Curves evaluate all 286 PI images at original resolution. Lines show the three-seed mean; shading shows the seed range.

We also compare operating points at original resolution by scanning $\log \max_c p_c - \log p_{\text{bg}}$ after interpolation. The curves in Fig. 3 measure correct recall at matched background FPR. Removing scores from $A + h$ improves recall in every seed at all five tested budgets; removing them from $b_0 + h$ lowers recall at budgets of 1–10%. Appendix Fig. 8 gives each paired result. Model selection uses source validation or the fixed final epoch; threshold sweeps characterize the recall–error trade-off.

4.4 Visual evidence beyond local score correction

On OOD, we compare a global bias with spatial heads using score summaries, visual features or both, holding the decoder, source images and update budget fixed (Table 2). The compact four-score head reaches 30.82% unseen mIoU. Widening its hidden layer to 24,017 parameters raises this to 32.13%, whereas the 24,049-parameter visual-and-score head reaches 55.42%. The 23.29-point difference persists despite nearly equal parameter counts. Compared with the scalar, the visual-and-score head improves 2,751–2,810 of the 2,954 images across seeds and worsens 27–75 (Appendix K).

Table 2: **Visual inputs and source-based calibration on OOD.** Three head seeds share a frozen decoder, 1,024 source images and 20,480 updates. Unseen mIoU (%) is shown before and after background-bias calibration. Every offset is selected on the same 256 source-validation images and applied unchanged to the target set.

Background	Inputs	Parameters↓	Before↑	After↑
$b_0 + h$	Global scalar	**1**	24.80 ±0.07	33.47 ±0.01
$b_0 + h$	Scores (compact)	1,217	30.82 ±4.24	36.47 ±0.71
$b_0 + h$	Scores (parameter-matched)	24,017	32.13 ±6.07	36.92 ±0.39
$b_0 + h$	Visual + scores	24,049	55.42 ±0.79	56.59 ±0.28
$b_0 + h$	Visual	24,049	51.74 ±0.84	55.86 ±0.46
$A + h$	Visual + scores	24,049	55.99 ±0.71	**57.08** ±0.45
$A + h$	Visual	24,049	**56.11** ±0.48	56.90 ±0.72

The visual-and-score head also reduces both error types relative to the parameter-matched score head: background FPR falls from 11.32% to 1.76%, and true-part rejection falls from 42.85% to 12.88%. With $A + h$, visual-only and visual-and-score heads reach 56.11% and 55.99%. With $b_0 + h$, retaining scores is better: 55.42% versus 51.74%.

To compare source-selected operating points, we calibrate every head with a separate background offset chosen on the same 256 validation images by fixed seen-class mIoU. The offset is added before normalization and original-resolution interpolation, then applied unchanged to the target set.

Calibration raises the parameter-matched score head to 36.92% and the visual-and-score head to 56.59%, leaving a 19.67-point difference that is positive in all three seeds. Appendix K.5 gives the rule, offsets and complete error metrics.

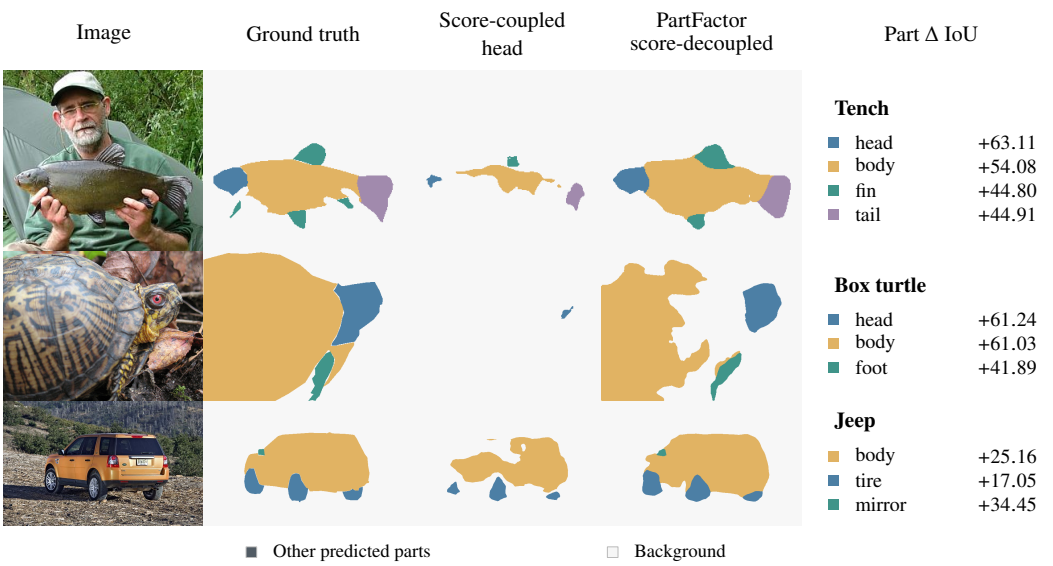


Figure 4: **Several parts can recover in the same image.** Complete-image predictions from score-coupled and score-decoupled heads sharing one part decoder. The right column gives every annotated part's IoU gain in percentage points, with matching color keys. Each example has at least three parts gaining 15 points. Appendix D specifies selection.

4.5 Foreground adaptation on other decoders

The source-trained PartCATSeg decoder starts at 35.69% unseen mIoU. Four balanced foreground configurations reach 35.15–36.04%, showing little change from further adaptation (Table 18).

HOPS OOD already has a trained background mechanism. We freeze its features and decoder, then compare a source-selected global bias with score-only, visual-only and visual-and-score heads. The global bias reaches 32.65% unseen mIoU; calibrated spatial heads reach 32.32–32.57%. At its default operating point, HOPS has $D = 0$: all correctly ranked unseen pixels remain foreground. Its correct-part recall also changes little at matched FPR (Table 20). The calibrated visual-only head has 35.13% background FPR, compared with 1.46% for the global bias (Table 17). The global bias therefore gives higher unseen mIoU and fewer background false positives. Here, foreground adaptation has no correctly ranked parts to recover from background rejection.

4.6 Where the gains occur

We group all 874 GT part regions by original-resolution pixel area and score-coupled IoU (Fig. 5). A region contains all pixels of one GT part class in an image; its IoU includes that class's false positives across the full image. Area bins use GT pixel counts; initial-IoU bins use the score-coupled prediction. Unseen regions smaller than 1,000 pixels gain just 0.40 IoU points on average, despite background error falling from 58.25% to 13.25%. Regions of at least 20,000 pixels gain 18.47 points and account for 76.45% of the net increase in correctly labeled unseen pixels; they contain 74.49% of unseen GT pixels. Larger unseen parts benefit most. For small parts, recovering foreground often leaves inaccurate boundaries or labels. Seen-region mean changes range from −1.26 to +0.51 points across bins. Table 7 gives counts and means; Appendix F also reports relative-area bins and false positives.

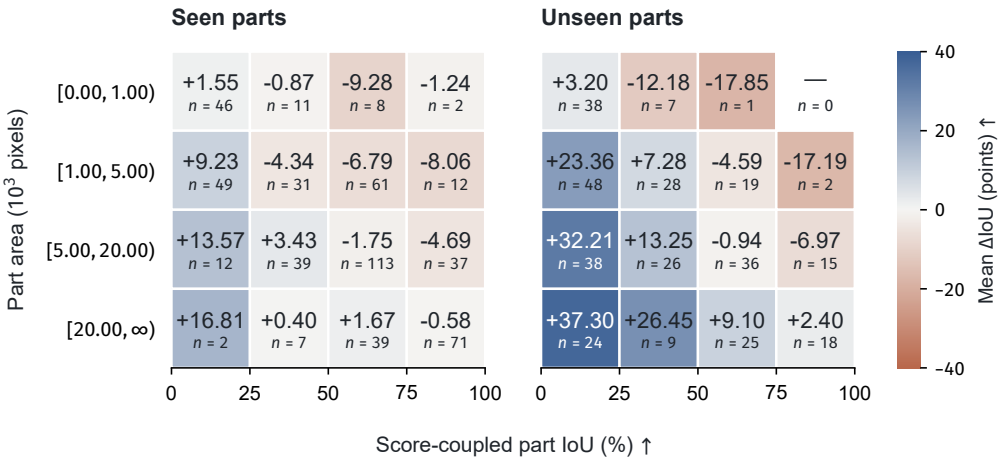


Figure 5: **Paired IoU gains by part area and initial IoU.** Cells show mean paired IoU change, grouped by GT part area and score-coupled IoU. Counts n refer to image–part regions; a dash marks an empty cell. Both groups use the same color scale. All 874 regions (540 seen, 334 unseen) are included, with full-image false positives counted in IoU. The final IoU interval includes 100%.

Recovery across several parts. Figure 4 shows three PI examples with at least three parts gaining 15 IoU points. Appendix D gives five additional examples and the selection rule. Figure 6 shows remaining identity errors.

4.7 Training and inference efficiency

On PI, PartFactor uses 2.72M task-trainable parameters (HOPS: 56.83M) and completes task training in 11.51 minutes. Inference averages 47.28 ms per image (HOPS: 214.54 ms), giving a 4.54× speedup. Including the frozen encoders, the inference model contains 846.32M parameters and requires 1,932.31 GFLOPs. Appendix G gives the measurement procedure and training-stage costs.

5 Conclusion

PartFactor learns conditional part identity first, then learns foreground support with a compact visual head. Background-logit reparameterization separates foreground probability from part scores. Compared with HOPS, unseen mIoU improves by 10.12 points on PI and 23.79 points on OOD. On PI, PartFactor uses only about 1/21 as many task-trainable parameters and runs 4.54× as fast. Matched-decoder experiments show that the visual head recovers more correctly ranked parts at the same background false-positive rates.

Outlook. A robot encountering an unfamiliar object still needs to decide where to grasp it, even if its part names are uncertain. Extending PartFactor to affordance learning could connect retained regions to the actions they support. Further directions include interactive editing, video segmentation and 3D scene understanding, with small parts and multi-object scenes as immediate challenges.

AI use statement

We used an LLM for code implementation and verification, language editing, and LaTeX formatting. Numerical claims were checked against model outputs and evaluation records. The authors take responsibility for the methods, code, results and manuscript.

REFERENCES

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiho Choi, Seonho Lee, Seungho Lee, Minhyun Lee, and Hyunjung Shim. Understanding multi-granularity for open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems*, 2024.

Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, and Hyunjung Shim. Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9782–9793, 2025.

Daan de Geus and Gijs Dubbelman. Task-aligned part-aware panoptic segmentation through joint object-part representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3174–3183, 2024.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pp. 128–145, 2022.

Junyi Li, Junfeng Wu, Weizhi Zhao, Song Bai, and Xiang Bai. PartGLEE: A foundation model for recognizing and parsing any objects. In *European Conference on Computer Vision*, 2024.

Xinlong Li, Di Lin, Shaoyiyi Gao, Jiaxin Li, Ruonan Liu, and Qing Guo. Open-vocabulary part segmentation via progressive and boundary-aware strategy. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xinlong Li, Di Lin, Shaoyiyi Gao, Yaxuan Liu, Jixian He, Jiaxin Li, Ruonan Liu, Qing Guo, Kairui Yang, and Wei Feng. HOPS: Hierarchical open-vocabulary part segmentation with attention-aware filtering and affinity-guided enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yang Miao, Jan-Nico Zaech, Xi Wang, Fabien Despinoy, Danda Pani Paudel, and Luc Van Gool. LangHOPS: Language grounded hierarchical open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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

Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15453–15465, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. OV-PARTS: Towards open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023.

Zhen Yao, Xin Li, Taotao Jing, Shuai Zhang, and Mooi Choo Chuah. DiSa: Saliency-aware foreground-background disentangled framework for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2601.20064*, 2026.

Jianjian Yin, Tao Chen, Yi Chen, Gensheng Pei, Xiangbo Shu, Yazhou Yao, and Fumin Shen. PCA-Seg: Revisiting cost aggregation for open-vocabulary semantic and part segmentation. *arXiv preprint arXiv:2603.17520*, 2026.

A ORIGINAL FOREGROUND COMPARISONS

Table 3 compares foreground heads sharing the original PI part decoder. The PI model in Table 1 was trained separately using the same two-stage procedure and reaches 45.30% unseen mIoU under Pred-All and 47.68% under Oracle-Obj.

Table 3: **Original fixed-decoder foreground comparisons.** Values are percentages; mIoU averages fixed class lists with zeros retained. Compared heads share source images and update budgets within each dataset. These are single-run results, distinct from the three-seed OOD experiment in Table 2. The part-training reference is measured before foreground adaptation.

(a) PartImageNet: 286 images, 147 candidate part classes

Foreground head	Seen↑	Unseen↑	hIoU↑	mUSI↓	mOSI↓
Part-training reference	17.54	19.49	18.47	**23.85**	75.71
Score-coupled spatial	50.09	35.59	41.62	38.25	**15.34**
Score-decoupled (uniform)	**50.82**	45.19	**47.84**	31.90	17.02
Score-decoupled (balanced)	49.41	**45.84**	47.56	27.01	22.08

(b) PartImageNet-OOD: 2,954 images, 482 candidate part classes

Foreground head	Unseen↑	mUSI↓	mOSI↓
Part-training reference	22.12	**20.35**	76.07
Scalar	25.16	67.35	35.38
Score-coupled spatial	**54.46**	34.46	**7.12**

B COMPLETE METRICS

Throughout the tables, ↑ and ↓ indicate the preferred direction. Bold marks the best value within a matched comparison, using the mean for repeated seeds. Area tables compare background error before and after adaptation within each row. Counts, pixel shares and image-level changes are descriptive statistics.

Let G_c, P_c and T_c denote the ground-truth (GT), predicted and correctly predicted pixel counts for class c. IoU is $T_c/(G_c + P_c - T_c)$, with an empty union assigned zero. Missed-part error is $(G_c - T_c)/G_c$ for GT-present classes. Seen and unseen mIoU average over fixed named class sets, retaining zero-IoU classes; their harmonic mean is $2su/(s + u)$. Table 3 uses these definitions for the original full-image component comparisons.

Dataset-level scoring. All main-text mIoU comparisons count predictions over the full image, including false positives on GT background. We pool integer confusion counts across images before averaging over fixed class lists. This class-macro metric differs from both pixel-weighted background error and image-macro IoU. PartCATSeg, HOPS and PartFactor use the same evaluation images, annotations and class lists. Published values and results under the author evaluation protocol appear separately in Appendix E.

Uncertainty. We draw 10,000 paired image-bootstrap samples (seed 20260906), rebuild aggregate class counts and recompute the fixed-class metrics for each draw. The 95% intervals measure evaluation-image uncertainty with the trained models fixed. Seed variation measures foreground-head training variability at a fixed part decoder.

mUSI and mOSI. For both datasets, Table 3 computes mUSI by averaging missed-pixel rates over GT-present labels, including background. For mOSI, fine labels are first collapsed to their semantic parent objects. We compute $(P_o - T_o)/P_o$ for each parent with both GT and prediction support and average over those parents, excluding background. This is a semantic parent error: different instances of the same parent category share one label.

Table 4: **Pixel outcomes for the original PI foreground comparison.** The first three rows partition all GT unseen-part pixels; the last uses GT-background pixels as its denominator. Rates are percentages and changes are percentage points. Both heads use the same part decoder on all 286 images.

Outcome	Score-coupled	Score-decoupled	Change
Unseen part → correct part ↑	51.11	**76.31**	+25.20
Unseen part → wrong part ↓	**8.92**	12.75	+3.83
Unseen part → background ↓	39.97	**10.94**	−29.03
Background → any part ↓	**4.87**	7.21	+2.34

C Training and inference settings

Image features and part logits are detached during foreground training. We resize the structural projection to the RGB-feature grid, then resize the head's scalar output to the decoder grid before computing probabilities.

Hardware. Experiments use an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU with 96 GB memory.

Source populations. We train a separate decoder for each dataset because their source and evaluation image sets overlap across dataset definitions. The OOD source set contains 16,517 annotated images after excluding six known exact RGB duplicates of evaluation images. Its part decoder is trained in nine stages on the corresponding source subsets, totaling 316,800 image presentations across decoder training, visual and semantic refinement, and regional prediction. All 256 OOD foreground-validation images appear in earlier decoder training but are excluded from foreground-head training. Public visual and text pretraining, including the external ImageNet object-identity model, is retained.

Foreground training and selection. The OOD scalar and spatial controls each use 1,024 source images for 20 epochs and 20,480 updates. The PI score-decoupled controls use 775 source images for 20 epochs and 15,500 updates. The part decoder uses AdamW. Foreground heads use Adam with learning rate 10^{-3}, zero weight decay, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and gradient-norm clipping at 5. Each update processes features from one source image; the two score-decoupled variants use identical images and order.

The PI foreground-validation set contains 200 images. Among variants that improve both source fine and background IoU, we select by source fine IoU. Balanced and uniform weighting give fine/background IoUs of 51.61/79.10 and 50.88/80.08; the score-coupled reference gives 43.37/75.93. All 200 validation images belong to the 2,981-image decoder-training set but are excluded from the 775-image foreground-training set. None of the 286 evaluation images overlaps the decoder-training set by filename, file content or decoded RGB pixels.

Head inputs and loss. The spatial head receives a 256-channel DINO structure map. Local RGB input is pooled three-channel RGB on PI and a 48-channel RGB stem on OOD. Learned projections produce 32 structural and 16 RGB channels. The score-coupled head adds the four summaries in Eq. 3; the score-decoupled head omits them and predicts a_ψ in Eq. 4. Balanced weighting gives foreground and background equal loss weight. The foreground probability used in Eq. 1 is $\sigma(-a_\psi)$.

Loss grid and target labels. For PartFactor, the loss is evaluated on the dense decoder grid at the processed image resolution. Images are resized to a 640-pixel short side, subject to a 768-pixel maximum side, and padded for the decoder. The scalar head output is bilinearly resized to this grid; part masks use nearest-neighbor resizing. No joint-posterior interpolation occurs before the loss.

Known source-part pixels supply foreground labels. Padding, unmapped labels, ignored annotations and conflicting masks do not contribute to the loss. Exterior pixels outside all annotated masks provide background supervision after invalid and conflicting regions are excluded. OOD maps 410 source labels into its 482-class inference vocabulary; PI maps 92 into 147. Training and evaluation

use the same semantic class mapping, while evaluation retains the complete original-resolution image. The PartCATSeg head uses its original-resolution score grid (Table 18).

Baseline training. PartCATSeg completes 20,000 updates with the paper settings. HOPS uses batch size 4, decoder learning rate 10^{-4} and CLIP-attention/DINO learning rates 10^{-6} for 20,000 updates. Table 1 uses its 15,000-update checkpoint, selected by validation hIoU computed with the HOPS evaluation protocol. Object and part loss weights are 0.4 and 0.6; augmentations are applied to images and annotations together. PartFactor uses DINOv3-L and an additional object-context predictor.

Original-resolution prediction. We form the part and background probabilities on the padded decoder grid, crop padding and bilinearly resize the posterior to the original image dimensions before argmax. Component comparisons use the same dimensions and fixed class lists. Appendix E separates results under the author evaluation protocol.

D Qualitative examples and error analysis

Figure 4 reports the full-image IoU change for every GT-present part, including false positives. We exclude previously illustrated examples and require at least 500 unseen-part GT pixels, three annotated parts, positive gains for all parts, and three parts with at least 100 GT pixels each gaining 15 points or more. Eleven images qualify. We rank them by the number of parts gaining 20 points, then 15 points, then mean part-IoU gain, breaking ties by image ID. We first select one image per parent category, then fill to eight in rank order, with at most three per category. Figure 4 shows the highest-ranked tench, box turtle and jeep; Fig. 7 shows the other five, without repetition.

Across all 286 images, the per-image score improves on 139, declines on 144 and ties on three. This score averages IoU over each image's fixed GT-present fine classes, retaining zero-IoU entries and full-image false positives for those classes. Its mean rises from 46.96 to 51.55; the paired median is −0.02 points, with quartiles −2.62 and +7.85. Dataset mIoU instead pools pixels before averaging a fixed class list. Including all 147 fine classes in every image mean does not change the 139/144/3 signs, because GT-absent classes have zero IoU in both predictions. Figure 6 shows a bald-eagle regression.

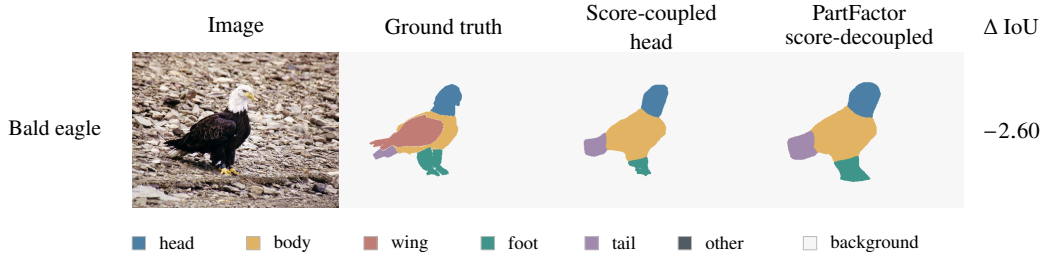


Figure 6: **Part-identity errors remain after foreground recovery.** This bald-eagle image is the median regression among eligible unseen-part examples. The heads follow Eqs. 3 and 4. Per-image fine IoU falls from 38.05% to 35.46%; incorrect wing and body regions remain in the output.

Part-to-background error in Table 4 divides rejected GT-part pixels by all GT pixels in the relevant seen or unseen group. For the score-coupled, uniform-weight score-decoupled and balanced score-decoupled heads, unseen error is 39.97%, 16.76% and 10.94%; seen error is 8.05%, 6.40% and 3.19%. These pixel-weighted rates describe one error type and are not class-averaged mIoU or mUSI.

E Published results under the original protocols

Table 5 lists published results under each method's original training and evaluation protocol. Table 1 uses the shared fixed-class, full-image evaluation protocol.

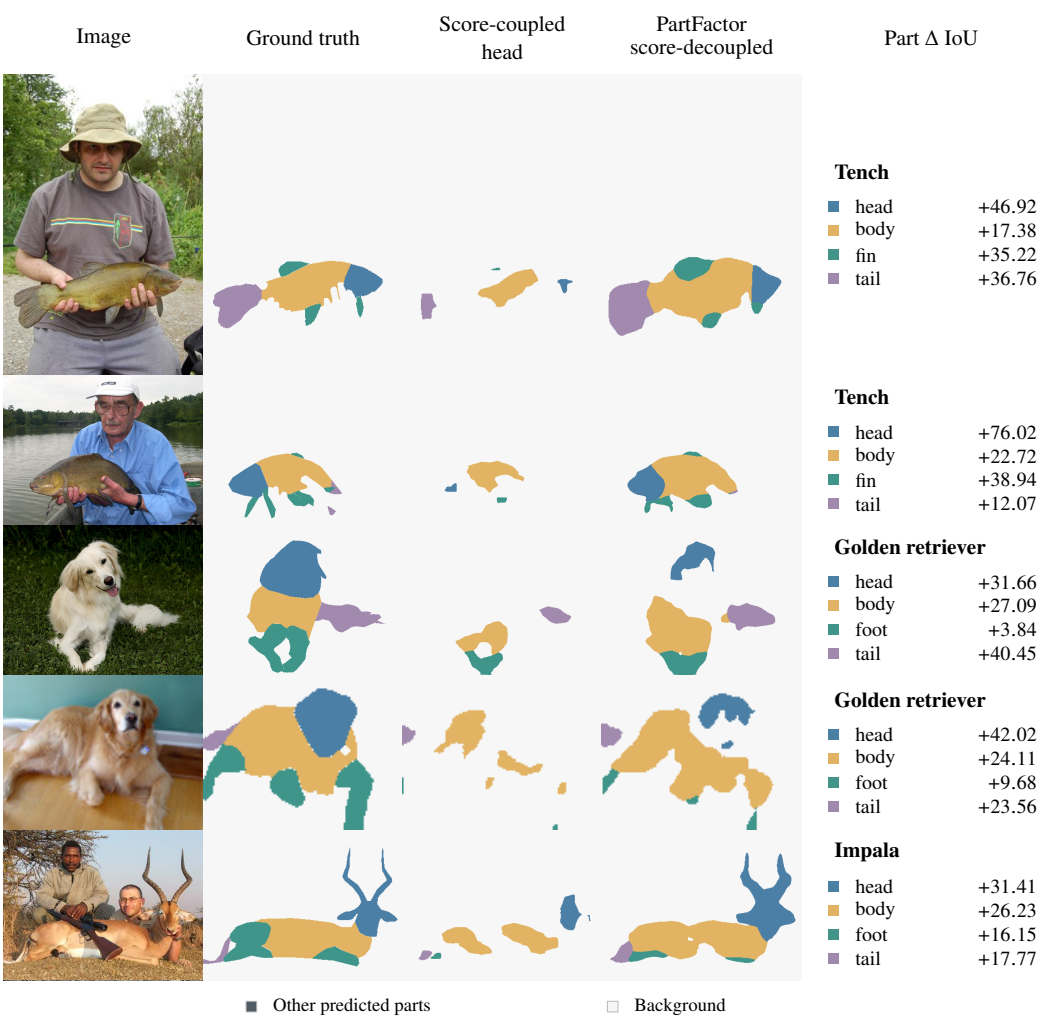


Figure 7: **Five additional multipart recovery examples.** These complete images are the remaining examples from the eight-image selection described above. None repeats Fig. 4, and all annotated-part gains are shown.

Table 5: **Published results for reference (%).** PI and OOD values are from Tables 5–6 of Li et al. (2026); PBAPS is Li et al. (2025). These results use the original training and evaluation protocols, not the main-text full-image protocol. A dash indicates an unreported result.

Method	PI Seen↑	PI Unseen↑	PI hIoU↑	OOD Unseen↑
CLIPSeg	32.39	12.27	17.80	6.69
CAT-Seg	33.58	23.04	27.33	19.83
PartCLIPSeg	38.82	19.47	25.94	10.33
PBAPS	44.87	41.40	43.06	–
PartCATSeg	57.33	53.07	55.12	40.17
HOPS	**60.81**	**57.46**	**59.08**	**45.08**

F Area analysis and additional error counts

For image i and GT-present class c, let G_{ic}, P_{ic} and T_{ic} count GT, predicted and correctly predicted pixels over the full image. Bin assignment uses only G_{ic}. We compute

$$E_{\mathrm{bg}}(B) = \frac{\sum_{(i,c)\in B} \#\{y_i = c, \widehat{y}_i = \mathrm{bg}\}}{\sum_{(i,c)\in B} G_{ic}}, \tag{8}$$

$$\mathrm{IoU}_{ic} = \frac{T_{ic}}{G_{ic} + P_{ic} - T_{ic}}, \qquad \overline{\Delta\mathrm{IoU}}(B) = \frac{1}{|B|} \sum_{(i,c)\in B} \left(\mathrm{IoU}_{ic}^{\mathrm{after}} - \mathrm{IoU}_{ic}^{\mathrm{before}}\right). \tag{9}$$

Zero-IoU regions remain in the average. A bin's net-correct-pixel share is its summed $T^{\mathrm{after}} - T^{\mathrm{before}}$ divided by the corresponding sum for the full seen or unseen group. All 540 seen and 334 unseen regions from the 286 images are included.

Table 6 bins by relative area, $G_{ic}/(H_i W_i)$. The smallest unseen bin loses 0.48 IoU points on average; the largest gains 18.35. The concentration of gains in larger regions therefore also holds when area is measured relative to image size.

False positives without a GT area bin. Predictions of a fine class absent from an image have no GT part area. Such false positives increase from 617,268 to 712,161 pixels; 281,441 and 310,933, respectively, lie on GT background. They remain in the dataset confusion matrices and the main fixed-class metrics. Among all 46,504,650 GT-background pixels, part predictions increase from 2,263,139 (4.87%) to 3,351,232 (7.21%).

Foreground transitions. Of the GT-part pixels assigned to background by the reference, 1,954,094 become correctly labeled parts and 309,992 become wrong parts. Conversely, 28,778 formerly correct pixels become background. Thus, the recovered foreground contains both correct labels and previously hidden part-identity errors.

G Resource accounting

Parameter counts. On PI, the part decoder and selected foreground head contain 2.72M task-trainable parameters, of which 22,177 belong to the PI foreground head. Frozen visual encoders are still required for inference. With the fixed vocabulary encoded, the complete inference model contains 846.32M parameters.

Inference and operation counts. PartFactor and HOPS are timed sequentially on the same NVIDIA RTX PRO 6000 Blackwell GPU with the weights used in Table 1, four CPU threads and batch size one. Both return Pred-All labels for the same 286 images in the same order. Timing runs from RGB loading to original-resolution labels, including the first call; model initialization and vocabulary encoding are excluded. PartFactor averages 47.28 ms per image and HOPS 214.54 ms, a 4.54× speedup. Additional complete passes give 43.75–43.92 ms for PartFactor and 210.31–213.70 ms for HOPS.

PartFactor reuses encoded text and executes its independent image encoders concurrently. HOPS retains its object detection, region processing and part prediction, while omitting the unused Oracle-Obj output and per-image memory-pool clearing. FLOPs include convolutions, matrix multiplications and attention in every visual encoder and task module, counting two operations per multiply–accumulate. Vocabulary encoding, resizing, elementwise operations and I/O are excluded.

Training cost. The measured PI training procedure, including both training stages and feature extraction, takes 11.51 minutes and peaks at 6.65 GiB of GPU memory. The part decoder runs for 1,250 steps at batch size 64; the foreground head runs for 20 epochs over 775 images, one image per step. Table 8 separates the stages. Timing includes model initialization and saving, but not foundation-model pretraining.

Table 6: **Foreground correction by relative part area.** A region is one GT part class in one image. Pixel shares use the corresponding seen or unseen group. Background error is pixel-weighted; ΔIoU averages paired region changes equally and includes full-image false positives. Before/after refers to score-coupled/score-decoupled prediction; values are percentages or percentage points.

Area (%)	Regions	Pixels (%)	BG error↓		ΔIoU↑
			Before	After	
Seen parts					
[0.00, 0.50)	51	0.28	37.05	**14.87**	-0.79
[0.50, 2.00)	110	2.75	22.83	**4.28**	+0.00
[2.00, 10.00)	256	29.79	10.27	**3.24**	-1.11
[10.00, 100.00]	123	67.17	6.34	**3.07**	+0.83
Unseen parts					
[0.00, 0.50)	43	0.29	59.23	**12.02**	-0.48
[0.50, 2.00)	63	2.18	49.22	**7.36**	+12.55
[2.00, 10.00)	135	21.20	36.89	**8.40**	+11.70
[10.00, 100.00]	93	76.33	40.50	**11.75**	+18.35

Table 7: **Foreground correction by absolute part area.** Each region is one GT part class in one image. Pixel shares are within the seen or unseen group. Background error is pixel-weighted, whereas ΔIoU is the equal-region mean change with full-image false positives included. Before/after refers to score-coupled/score-decoupled prediction; values are percentages or percentage points.

Area (10^3 pixels)	Regions	Pixels (%)	BG error↓		ΔIoU↑
			Before	After	
Seen parts					
[0.00, 1.00)	67	0.34	28.65	**8.83**	-0.22
[1.00, 5.00)	153	4.38	17.44	**3.57**	-1.26
[5.00, 20.00)	201	22.97	9.94	**2.74**	-0.37
[20.00, ∞)	119	72.31	6.78	**3.28**	+0.51
Unseen parts					
[0.00, 1.00)	46	0.31	58.25	**13.25**	+0.40
[1.00, 5.00)	97	4.35	43.92	**7.87**	+12.41
[5.00, 20.00)	115	20.85	38.38	**9.61**	+12.44
[20.00, ∞)	76	74.49	40.11	**11.48**	+18.47

Table 8: **PI task-training time by stage.** Measurements use one NVIDIA RTX PRO 6000 Blackwell GPU. Feature extraction covers 775 training and 200 validation images. Times exclude foundation-model pretraining.

Stage	Time (s)↓
Part-decoder training	454.95
Frozen feature extraction	114.51
Foreground-head training	119.83
Initialization	1.37
Total	690.67

H IMAGE-LEVEL EFFECTS AND POOLED CLASS IOU

This appendix reports image-level effects for the original single-run comparisons in Appendix A. The later three-seed OOD experiment is reported separately in Table 19. Let T_{ic}, P_{ic} and G_{ic} be the correctly predicted, predicted and GT pixel counts for class c in image i. For the fixed unseen class

list $\mathcal{U}$, benchmark mIoU is

$$\text{mIoU}_{\mathcal{U}} = \frac{1}{|\mathcal{U}|} \sum_{c \in \mathcal{U}} \frac{\sum_i T_{ic}}{\sum_i (G_{ic} + P_{ic} - T_{ic})}. \tag{10}$$

For the image-level diagnostic, we first average $T_{ic}/(G_{ic} + P_{ic} - T_{ic})$ over $C_i = \{c \neq \text{bg} : G_{ic} > 0\}$ within each image, then weight images equally. The unseen-only diagnostic uses $C_i \cap \mathcal{U}$ and omits images where that set is empty. All zero-IoU entries in these sets are retained, and all P_{ic} include predictions on the full image. The order of aggregation differs: dataset mIoU pools class counts first, whereas the image diagnostic averages IoUs computed within each image. For PI, $[139(13.83) - 144(4.23)]/286 \approx 4.59$ points explains the mean image gain despite slightly more declines. The unseen-only image mean rises from 36.33 to 49.12, whereas pooled unseen mIoU rises from 35.59% to 45.84%.

On OOD, mean image IoU is 23.47/28.14/54.74 for the part-training reference/scalar/spatial outputs. Relative to the part-training reference, the spatial head improves 2,661 images (90.08%), worsens 188 (6.36%), and ties on 105 (3.55%). Improved images gain 35.87 points on average; worsened images lose 16.35. Relative to the scalar, there are 2,743 improvements, 80 declines, and 131 ties. All 2,954 images contain evaluated unseen parts, so the all-part and unseen-only image diagnostics coincide here.

I Attribution on the PartFactor decoder

We train a new task decoder from public encoder initialization, then freeze it for every foreground configuration. Heads share 775 training images and 15,500 updates. Within each seed, initial head parameters and image order are identical across configurations; the three seeds vary these two factors. All configurations use the prespecified final epoch.

With visual inputs, changing the background from $b_0 + h$ to $A + h$ adds 11.00 unseen IoU points under uniform weights and 6.67 under balanced weights. At fixed $A + h$, removing scores adds 4.47 and 1.33 points, respectively. At fixed $b_0 + h$, it instead loses 3.16 and 3.65 points. The direction of the score-input effect therefore changes with the background parameterization.

Among the 111 images containing unseen parts, removing scores at fixed $A + h$ under uniform weights improves 80/78/83 images and worsens 28/30/25 across the three seeds, with three ties in each. Mean image-IoU gains are 4.65, 6.84 and 7.13 points. Only foreground training varies across these seeds; the part decoder is shared.

J Model construction and experimental procedure

J.1 From visual features to part probabilities

PartFactor learns an object-conditioned part decoder and a foreground head. For the decoder, DINOv3-L layers 6, 12, 18 and 24 each pass through a 1×1 projection to 256 channels. Learned softmax weights combine the spatially aligned maps. A residual convolution and local attention with eight heads over 5×5 neighborhoods produce S. Dense CLIP image–text matching supplies object masks and part evidence, while SigLIP 2 and an ImageNet-trained ResNet-50 supply the image-level object prior.

Each of the four selected objects has a soft mask that defines an object-relative coordinate frame. Mask and coordinate encodings augment S, and part text embeddings initialize spatial assignments. Each of three updates pools features with the current soft assignments, applies self-attention among the part representations and updates their spatial assignments. The final logits combine CLIP matching, text-matched structural evidence, object log priors and conditional part log probabilities. These components are trained before foreground adaptation.

The foreground head projects S and local RGB features to 32 and 16 channels. After spatial alignment, the concatenated features pass through a 3×3 convolution with 32 output channels, GELU and a scalar 1×1 convolution. On PI, the RGB input has three channels and the head contains 22,177 parameters. The score-coupled head additionally receives $[A - M, M - b_0, M, b_0]$. The score-decoupled head predicts a_ψ from visual features and sets $b = A + a_\psi$.

Table 9: **Distribution of per-image changes.** Image IoU averages GT-present part classes, retaining zero IoU and full-image false positives. The unseen-only row restricts the class set and omits images without unseen GT. Δ is after minus before, in percentage points. These image means are not dataset mIoU.

Comparison / image population	Improved↑	Declined↓	Tied	Mean Δ↑	Median Δ↑
PI: coupled → decoupled (all)	139	144	3	+4.59	−0.02
unseen only	78	30	3	+12.79	+7.92
OOD: base → spatial	2661	188	105	+31.27	+35.44
OOD: scalar → spatial	2743	80	131	+26.60	+25.30

Table 10: **Complete PI factorial results.** All heads share one independently trained decoder and are evaluated on all 286 original-resolution images, with fixed sets of 92 seen and 55 unseen classes retaining zero IoU. Unseen mIoU reports the mean and sample standard deviation over three head seeds; other metrics are means. Values are percentages. $P \to B$ and $B \to P$ are part-to-background and background-to-part pixel error rates. The pre-foreground decoder is a common reference; bold compares the four adapted configurations within each weighting.

Background	Inputs	Seen↑	Unseen↑	hIoU↑	$P \to B$↓	$B \to P$↓
Before foreground training		17.20	18.96	18.03	0.00	100.00
Uniform pixel weights						
$b_0 + h$	Visual + scores	50.00	36.42 ±2.30	42.12	19.08	4.60
$b_0 + h$	Visual	**50.67**	33.26 ±2.09	40.13	21.61	**4.45**
$A + h$	Visual + scores	50.42	39.79 ±2.37	44.46	17.02	4.85
$A + h$	Visual	50.25	**44.26** ±1.29	**47.06**	**11.85**	4.84
Balanced pixel weights						
$b_0 + h$	Visual + scores	**50.57**	41.79 ±1.46	45.76	13.36	5.93
$b_0 + h$	Visual	50.52	38.15 ±1.50	43.45	16.85	**5.86**
$A + h$	Visual + scores	50.41	43.49 ±1.47	46.66	10.56	6.56
$A + h$	Visual	50.52	**44.82** ±0.34	**47.49**	**7.38**	6.55

J.2 What the foreground loss optimizes

Let $z = 1$ for a GT part pixel and $z = 0$ for GT background. For a known part label y, categorical cross-entropy decomposes as

$$-\log p_y = -\log s - \log q_y; \qquad -\log p_{\text{bg}} = -\log(1 - s). \tag{11}$$

With q fixed, the term involving q_y has no foreground-head gradient. Under identical pixel weights on the grids in Appendix C, the trainable term is therefore

$$\ell_{\text{fg}}(u) = -z(u) \log s(u) - (1 - z(u)) \log(1 - s(u)). \tag{12}$$

Uniform weighting averages over valid pixels within each image. Balanced weighting averages the foreground and background group means, using the groups present. Both use the same part annotations and exclude padding and unknown labels. The last convolution is zero-initialized: the additive head starts at $b = b_0$, whereas the reparameterized head starts at $s = 1/2$.

J.3 Training stages

We train the part decoder before the foreground head. Once the decoder is frozen, source features and part logits are stored and reused for foreground updates. PI uses 775 foreground-training images and 200 foreground-validation images, with disjoint training and validation object lists. The factorial heads use Adam for 15,500 updates with the settings in Appendix C. The original OOD scalar and spatial comparison uses 1,024 training images and 20,480 updates. Its spatial head receives visual features and score summaries, so this comparison measures their combined contribution relative to a scalar. Appendix C describes the earlier nine-stage OOD decoder training and the 256-image foreground-validation set.

The PI factorial experiment starts from an independently trained decoder: 1,250 updates at batch size 64, using pretrained encoders and randomly initialized task parameters. It shares the foreground-training and validation images with the original PI experiment, but not the decoder in Table 3a. Every head in the factorial comparison uses this same new decoder.

J.4 Separating score inputs, background parameterization and weighting

We cross background parameterization ($b_0 + h$ or $A + h$) with the presence or absence of four score inputs. The four configurations are repeated under uniform and balanced pixel weights over three seeds. Within a seed, all eight heads share initial parameters, training images, image order and update count. Visual-only configurations zero the four score channels without changing convolution dimensions. Each factorial head has 23,329 parameters; the final visual head omits these channels and has 22,177. Every configuration is evaluated at the fixed final epoch.

For the source-trained PartCATSeg model, we use the same four parameterization/input combinations with balanced weights and three seeds (Table 18).

J.5 Numerical evaluation

PartFactor crops and bilinearly resizes the complete categorical posterior before argmax, retaining all candidate parts and background. PartCATSeg forms its posterior at original resolution (Table 18). Metrics aggregate per-image confusion counts. Across foreground-head seeds, attribution experiments reuse the same decoder predictions and enable deterministic numerical computation.

K Foreground diagnostics and additional comparisons

K.1 Background construction and supervision

In the independently trained PI attribution decoder, b_0 is a scalar 1×1 convolution of structural features. Part training uses annotated part pixels and ignores the surrounding unannotated canvas. Background remains in the categorical denominator on part pixels and therefore receives gradients, but has no positive examples in this phase. The OOD part decoder is also trained without positive background supervision; its final boundary refinement uses valid part labels. Tables label these intermediate outputs as *part-training references*. Foreground training then supplies both part and background labels. In particular, the PI comparison of 35.59% and 45.84% uses two trained foreground heads.

K.2 OOD controls and external training

The seven OOD heads share a frozen decoder, 1,024 source images and 20,480 updates, with the optimizer in Appendix C. All three seeds use the fixed final epoch. Visual factorial heads contain 24,049 parameters and use the same convolution dimensions; visual-only configurations zero the four score channels. The compact score head has 32 hidden channels. Widening it to 632 channels gives a parameter-matched head with 24,017 parameters and the same convolution–GELU–output topology. This matches parameter count while retaining score-only input. OOD uses 48-channel local RGB features, accounting for the parameter difference from PI. Table 2 lists the configurations, and Table 13 gives complete error metrics.

OOD CAT-Seg and PartCATSeg use their released configurations, batch size 4, AdamW and learning rate 2×10^{-4} for 20,000 updates. HOPS uses its released modules, batch size 4, a decoder learning rate of 10^{-4} and CLIP-attention/DINO learning rates of 10^{-6} for 20,000 updates. These models start from the public CAT-Seg base initialization and use the final checkpoint. The source population contains 16,517 annotated images after six known exact target duplicates are excluded. PI PartCATSeg also completes 20,000 updates; its validation-selected model gives 43.50/35.69/39.21% seen/unseen/hIoU. Published values remain separate in Appendix E.

The three-seed OOD comparisons use Pred-All with deterministic numerical computation. Table 2 reports their means and standard deviations; Appendices A and H report the earlier single-run comparison. Every official target image is retained, including duplicates within that population. Source foreground-training, foreground-validation and target sets have no exact file-content overlap.

K.3 Correct-part rejection and paired uncertainty

Table 14 reports correct-part rejection D from Eq. 6, correct unseen-part recall and background FPR separately. PartFactor uses the decoder grid, where q is identical between configurations. Both D and recall divide by all GT unseen-part pixels; recall counts final predictions with the correct unseen label. Background FPR divides by all GT-background pixels.

At original resolution, we scan $\log \max_c p_c - \log p_{\text{bg}}$; its fixed-grid counterpart is $\max_c f_c - b$. Thresholds cover $[-64, 64]$ in steps of 0.001, including both endpoint decisions. For each FPR budget, we take the highest attained correct recall without interpolating between thresholds. Lines show the mean over three seeds and shading their range. These curves diagnose the trade-off; model selection uses the prescribed training endpoint or source-validation rule.

For the PI interaction, we draw 2,000 paired image-bootstrap samples. Each sample uses identical image multiplicities for all configurations and seeds, rebuilds class confusion counts and averages the three seed-specific unseen mIoUs. Table 16 reports percentile intervals conditional on the trained heads. The original single-run PI interval instead uses the 10,000 draws in Appendix B.

K.4 HOPS foreground adaptation and calibration

HOPS heads use its own 384-channel visual features and pooled RGB, without another encoder. Visual and visual-and-score heads contain 27,425 parameters; the score-only head contains 1,217. All use $A + h$, uniform weights, the same 1,024 source images, 20,480 updates and three seeds. We then select a separate additive background offset for each head and for original HOPS on 256 source foreground-validation images. Selection maximizes fixed seen-class mIoU over $[-64, 64]$ in increments of 0.125. Ties prefer the offset nearest zero, then the smaller offset. The validation images are excluded from head training. Selected offsets are applied unchanged to all 2,954 target images.

Table 17 reports both uncalibrated and calibrated outputs. At its default operating point, the score-only head rejects nearly every part; source calibration raises unseen mIoU from 0.02% to 32.57%. The calibrated visual-only head reaches 32.32% but has 35.13% background FPR. The best unseen mIoU, 32.65%, comes from a global bias on original HOPS. Unseen-class IoU alone does not capture the differences in background behavior.

K.5 Source-validation bias calibration on PartFactor OOD

Calibration keeps all seven configurations at their existing final training weights. For each head and seed, we select a background offset on 256 source foreground-validation images from $[-64, 64]$ at increments of 0.125. The objective is mean IoU over the fixed 410 seen classes, retaining zeros. Ties prefer the offset nearest zero, then the smaller value, as for HOPS. Source scoring uses the decoder grid and valid labels of foreground validation.

Each selected offset is fixed before evaluation on all 2,954 target images. We set $b' = b + \delta$, form the complete 483-class posterior, crop padding and bilinearly interpolate to original resolution. Table 2 reports both operating points, Table 11 gives the calibrated error metrics, and Table 12 lists every selected offset.

Table 11: **Source-calibrated PartFactor OOD results.** Values are percentages, reported as means and sample standard deviations over three head seeds. FPR counts background predicted as a part; FNR counts parts predicted as background.

Head	Unseen↑	mUSI↓	mOSI↓	FPR↓	FNR↓
Scalar	33.47 ±0.01	40.15 ±0.10	53.79 ±0.09	28.17 ±0.15	22.00 ±0.13
Scores (compact)	36.47 ±0.71	39.10 ±0.50	45.12 ±1.19	18.73 ±0.98	26.64 ±1.38
Scores (parameter-matched)	36.92 ±0.39	40.65 ±0.66	41.82 ±1.30	15.49 ±1.50	28.95 ±1.72
$b_0 + h$ (visual + scores)	56.59 ±0.28	28.98 ±0.45	**10.73** ±0.24	**2.56** ±0.08	8.98 ±0.66
$b_0 + h$ (visual)	55.86 ±0.46	28.30 ±0.75	12.31 ±0.54	3.10 ±0.20	8.32 ±0.18
$A + h$ (visual + scores)	**57.08** ±0.45	**26.87** ±0.30	11.86 ±1.03	3.01 ±0.40	**7.42** ±1.72
$A + h$ (visual)	56.90 ±0.72	27.52 ±0.64	11.43 ±0.71	2.84 ±0.25	7.84 ±0.48

Table 12: **Source-selected background offsets.** Columns correspond to seeds 20260906, 20260907 and 20260908. Three decimals preserve the 0.125 search increment.

Head	Seed 1	Seed 2	Seed 3
Scalar	-1.875	-1.875	-1.875
Scores (compact)	-1.125	-1.625	-0.750
Scores (parameter-matched)	-1.500	0.375	-0.625
$b_0 + h$ (visual + scores)	-1.125	-0.750	-0.500
$b_0 + h$ (visual)	-2.000	-2.000	-1.875
$A + h$ (visual + scores)	-0.250	-1.125	-1.125
$A + h$ (visual)	-0.625	-0.625	-0.875

K.6 IMAGE-LEVEL EFFECTS

Table 19 reports per-image changes for the three-seed visual-and-score head relative to the scalar.

Table 13: **Complete OOD results before bias calibration.** Percentages, reported as means and sample standard deviations over three head seeds. FPR is background predicted as a part; FNR is a part predicted as background.

Head	Unseen↑	mUSI↓	mOSI↓	FPR↓	FNR↓
Scalar	24.80 ±0.07	67.83 ±0.12	35.73 ±0.07	6.09 ±0.05	59.46 ±0.17
Scores (compact)	30.82 ±4.24	57.89 ±7.33	33.95 ±3.65	6.99 ±3.10	52.58 ±12.04
$b_0 + h$ (visual + scores)	55.42 ±0.79	33.09 ±1.38	8.25 ±1.10	1.76 ±0.35	12.88 ±2.91
$b_0 + h$ (visual)	51.74 ±0.84	39.18 ±1.42	**6.73** ±0.60	**1.28** ±0.17	18.65 ±2.29
$A + h$ (visual + scores)	55.99 ±0.71	31.62 ±2.88	8.76 ±0.77	1.96 ±0.21	11.38 ±0.48
$A + h$ (visual)	**56.11** ±0.48	**31.40** ±1.08	8.85 ±0.48	1.98 ±0.15	**11.11** ±0.32
Scores (parameter-matched)	32.13 ±6.07	52.31 ±16.94	36.83 ±9.72	11.32 ±9.18	42.85 ±21.84

Table 14: **Correct identities retained and rejected on PI.** Percentages averaged over three seeds. D and correct recall divide by all GT unseen-part pixels on the decoder grid; FPR divides by GT-background pixels.

Configuration	$D \downarrow$	Recall↑	FPR↓
PartFactor / PI (uniform)			
$b_0 + h$ (visual + scores)	34.69	49.30	5.00
$b_0 + h$ (visual)	41.93	42.06	**4.79**
$A + h$ (visual + scores)	29.80	54.19	5.41
$A + h$ (visual)	**17.73**	**66.26**	5.23

Table 15: **Paired PI transitions after removing score inputs.** Rates are percentages. Correct recovery and loss use all GT unseen-part pixels as denominator; new and removed false positives use GT background. Each row describes a separate matched pair.

Pair	Recovered↑	Lost↓	New FP↓	Removed FP↑
$b_0 + h$ (uniform)	0.27	7.51	0.22	0.43
$A + h$ (uniform)	12.23	0.16	0.68	0.86
$b_0 + h$ (balanced)	0.73	9.42	0.40	0.51
$A + h$ (balanced)	8.31	0.24	0.82	0.93

Table 16: **PI score-input effects and parameterization interaction.** Effects are unseen-mIoU percentage points. Input effect is visual minus visual-and-score; interaction is the $A + h$ effect minus the $b_0 + h$ effect. Intervals use paired image bootstraps with trained heads fixed.

Contrast	Effect	95% interval
$b_0 + h$ (uniform)	-3.16	[-3.86, -2.68]
$A + h$ (uniform)	+4.47	[+3.56, +5.53]
Interaction (uniform)	+7.63	[+6.50, +9.17]
$b_0 + h$ (balanced)	-3.65	[-4.52, -2.71]
$A + h$ (balanced)	+1.33	[+0.40, +2.23]
Interaction (balanced)	+4.98	[+3.41, +6.49]

Table 17: **HOPS OOD with source-selected background calibration.** IoU columns are unseen mIoU before and after calibration. Spatial heads report mean and sample standard deviation over three seeds; FPR and background IoU describe calibrated outputs. All values are percentages.

Head	Before↑	After↑	FPR↓	BG IoU↑
Original HOPS	32.32	**32.65**	**1.46**	95.09
Scores	0.02 ±0.02	32.57 ±0.00	1.47 ±0.01	**95.87** ±0.02
Visual	12.17 ±3.26	32.32 ±0.00	35.13 ±12.07	64.85 ±12.05
Visual + scores	12.21 ±2.37	32.43 ±0.10	12.71 ±12.61	86.88 ±12.28

Table 18: **Foreground adaptation on source-trained PartCATSeg.** PI results with balanced weights; adapted heads report mean and sample standard deviation over three seeds. All values are percentages.

Head	Seen↑	Unseen↑	hIoU↑
Original	43.50	35.69	39.21
$b_0 + h$ (visual + scores)	45.37 ±0.51	35.94 ±0.70	40.10 ±0.58
$b_0 + h$ (visual)	44.96 ±0.59	**36.04** ±0.59	40.01 ±0.60
$A + h$ (visual + scores)	**45.67** ±0.49	35.92 ±0.17	**40.21** ±0.20
$A + h$ (visual)	45.40 ±0.46	35.15 ±1.06	39.62 ±0.83

Table 19: **Per-image OOD effects of the visual-and-score head versus the scalar.** Image IoU averages GT-present part classes and includes full-image false positives. Changes are percentage points; counts and image means are distinct from dataset mIoU.

Seed	Improved	Declined	Tied	Mean Δ	Median Δ
20260906	2751	75	128	27.05	26.20
20260907	2806	33	115	28.87	27.50
20260908	2810	27	117	28.92	27.53

Table 20: **HOPS correct unseen-part recall at matched background FPR.** All 2,954 OOD images are evaluated at original resolution. Each column gives the highest attained recall (%) without exceeding its FPR budget. Adapted heads report means and sample standard deviations over three seeds.

Foreground head	Background FPR budget (%)				
	0.10	0.50	1.00	2.00	5.00
Original HOPS	**25.17**	39.27	43.79	**44.94**	**44.94**
Scores	23.83 ±1.11	40.22 ±0.32	**43.86** ±0.02	44.88 ±0.03	**44.94** ±0.00
Visual	23.94 ±1.66	39.66 ±0.56	42.24 ±0.29	43.64 ±0.09	44.62 ±0.03
Visual + scores	24.82 ±1.21	**40.64** ±0.33	43.15 ±0.32	44.29 ±0.23	44.86 ±0.05

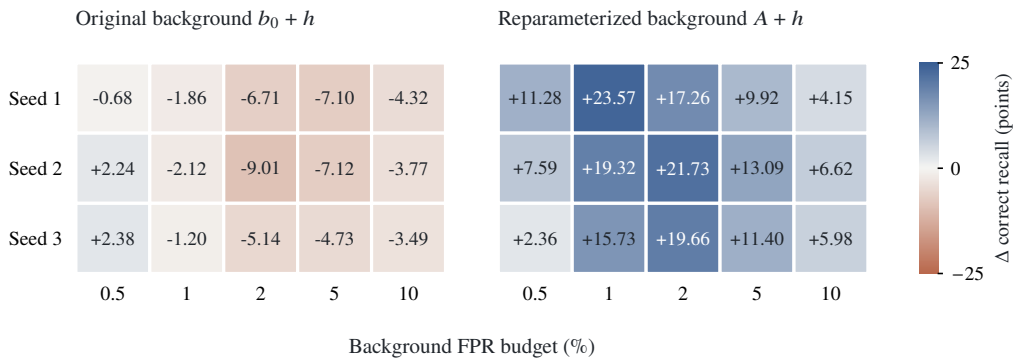


Figure 8: **Recall changes for every PI seed and error budget.** Each cell is visual minus visual-and-score correct unseen-part recall, in percentage points, using uniform weights and all 286 images. Rows correspond to seeds 20260906, 20260907 and 20260908. Both panels share a symmetric color scale. Recall is the highest attained value within each background FPR budget.