SLOT COUNT AS AN ENSEMBLE AXIS FOR OBJECT-CENTRIC LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Object-centric models decompose a scene into a fixed number of slots, yet the right number is rarely known. Too few slots merge objects, too many fragment them, and conventional inference commits to a single decomposition. We ask whether the decompositions that one model produces at different slot counts can instead be combined into a better one. We introduce PACE (Pairwise Agreement across Cardinality Ensembles), a two-part framework that answers this question positively. Prefix training samples a slot count at every step and activates the matching prefix of an ordered bank of slot initializers, so that a single checkpoint yields decompositions at many cardinalities from one encoder pass. Horizontal-Cut Consensus, a parameter-free readout, converts each soft partition into slot-invariant pairwise co-membership, averages this evidence, and cuts the resulting hierarchy into any requested number of regions, without slot matching, labels, or additional training. On COCO, PASCAL VOC2012, and MOVi-C, PACE improves FG-ARI by 4.67, 3.67, and 5.26 points over the standard seven-slot prediction, and mBO by 1.17 and 4.57 points on COCO and VOC2012. No single slot count recovers this grouping gain, even when the best one is chosen with evaluation labels. Controlled experiments trace the benefit to complementary errors. Coarse and fine requests fail in opposite ways, repeated sampling at one slot count saturates quickly, and a weak standalone partition can contribute more than a strong one. These results establish requested slot count as an ensemble axis and decouple the granularity of the grouping evidence from that of the final segmentation.

1 INTRODUCTION

Humans can generalize from limited visual experience to novel situations, in part by organizing perception around objects and their relations (Spelke & Kinzler, 2007; Lake et al., 2017; Lake & Baroni, 2018). This structured view of the world provides a natural basis for compositional reasoning: representations of entities and interactions can be reused across novel combinations and configurations (Battaglia et al., 2018; Wiedemer et al., 2023). Endowing learned visual representations with a similar structure is therefore a promising route toward models that can reason over complex scenes and generalize beyond the configurations observed during training.

Motivated by this view, object-centric learning seeks to decompose visual scenes into structured representations that support independent tracking and reasoning (Locatello et al., 2020; Peters & Kriegeskorte, 2021). The decomposition of an input scene, however, need not be unique. Objects may be represented either as wholes or in terms of their constituent parts, which may themselves be treated as wholes or recursively decomposed into finer components. Human annotations for segmentation tasks reflect this non-uniqueness (Palmer, 1977; Navon, 1977; Martin et al., 2001), and modern segmentation systems often accommodate multiple valid interpretations at different granularities (Kirillov et al., 2023; Ravi et al., 2024; Li et al., 2023). Early object-centric approaches such as Slot Attention (Locatello et al., 2020), on the other hand, cannot directly account for this variability. Instead, they partition input visual features among a fixed number K of latent slots. This fixed-cardinality constraint can limit the structure captured by the resulting decomposition. Smaller values of K may capture coarser structure but merge distinct entities, while larger values of K may resolve finer structures but lead to object fragmentation. Consequently, no single choice of K is guaranteed to expose the full range of useful grouping structure present in a scene.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

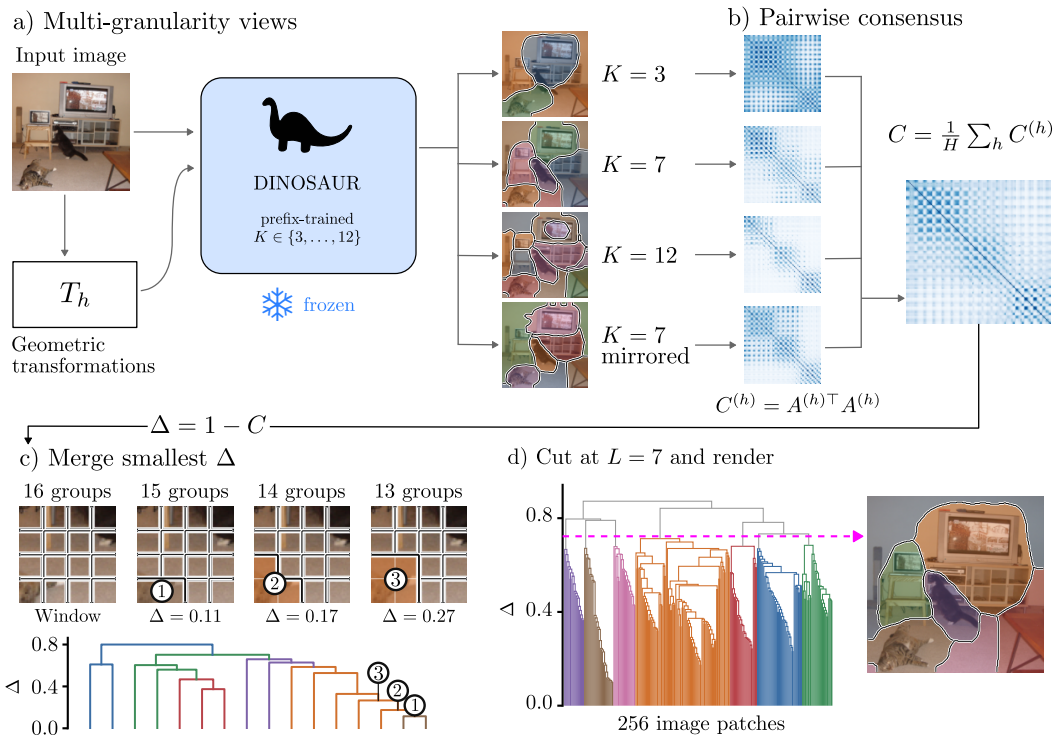

Figure 1: Overview of PACE. One prefix-trained model produces soft partitions at many cardinalities (a). The readout averages their slot-invariant patch co-membership (b), builds an average-linkage hierarchy (c), and cuts it at L regions to render the segmentation, without training or labels (d).

Recent work has sought to mitigate this limitation through adaptive-cardinality mechanisms (Fan et al., 2024; Liu et al., 2025; Chatzisavvas et al., 2026) and region-refinement methods (Tian et al., 2025; Lu & Guo, 2026), which respectively adapt the number of object representations or refine an initially obtained segmentation. Instead, we approach the problem from a different perspective, drawing inspiration from cluster-ensemble methods. These techniques represent partitions through pairwise co-membership, effectively removing dependence on arbitrary cluster identities and enabling evidence aggregation across predictions (Strehl & Ghosh, 2002; Fred & Jain, 2005). Specifically, we investigate whether complementary grouping information can be exploited across complete decompositions produced by the same object-centric model at different requested cardinalities K. The central research question addressed in this paper is therefore: *can we combine grouping evidence across several plausible decompositions of the same scene to improve its decomposition?*

To address this question we introduce **PACE** (**P**airwise **A**greement across **C**ardinality **E**nsembles), a two-part framework for multi-granularity object-centric segmentation. The first component, *prefix training*, dynamically samples the number of available slots K and activates the corresponding prefix of an ordered slot-initializer bank at each training step. This enables object-centric models to operate across multiple cardinalities with a shared encoder pass. The second component consists of a parameter-free *Horizontal-Cut Consensus* readout that converts each soft partition into slot-invariant pairwise patch co-membership and averages the resulting evidence. Average-linkage hierarchical clustering then produces a hierarchy that can be cut at a chosen output region count L, which allows to decouple the final segmentation granularity from the cardinalities of the constituent partitions. Figure 1 illustrates the complete pipeline. We make four primary contributions.

(i) **A training-free consensus readout for multi-granularity object decompositions.** We introduce Horizontal-Cut Consensus, which aggregates pairwise co-assignment evidence across predictions at different slot cardinalities and produces a single segmentation at any output cardinality. It requires no supervision, cross-prediction slot matching, or learned parameters.

(ii) **A cardinality-flexible training strategy for multi-granularity object parsing.** Prefix training enables a single model to produce meaningful decompositions across slot cardinalities while sharing a common image representation. It substantially outperforms applying fewer slots to a model trained only at a larger cardinality, and approaches the performance of separately trained models specialized to each slot count.

(iii) **An empirical validation on the advantage of object grouping over single-cardinality inference.** PACE improves FG-ARI over the standard single-K prediction by 4.67, 3.67, and 5.26 points on COCO, VOC2012, and MOVi-C, respectively. These gains are not recovered by any individual K, even when selected post hoc using ground-truth annotations. On COCO, it sets a new unsupervised state of the art in FG-ARI at a range of output region counts.

(iv) **An empirical characterization of why multi-granularity consensus works.** K induces complementary grouping errors: coarse decompositions merge objects, whereas fine decompositions increasingly split them. Consensus exploits it to improve scene segmentation.

2 Pairwise Agreement across Cardinality Ensembles (PACE)

2.1 Backbone and Prefix Training

Backbone. We build on the MLP-decoder variant of DINOSAUR (Seitzer et al., 2023), which applies Slot Attention (Locatello et al., 2020) to frozen vision-transformer features. Given an image $\boldsymbol{x}$, the encoder produces P patch features, which are mapped to the slot dimension by a trainable projection:

$$F = E(\boldsymbol{x}) \in \mathbb{R}^{P \times d_f}, \qquad G = \pi(F) \in \mathbb{R}^{P \times d}. \tag{1}$$

Slot Attention maps G to K latent slots. A broadcast MLP decoder then predicts, for each slot and patch, a feature proposal $\widehat{\boldsymbol{f}}_{sp}$ and mask logit a_{sp}. These define the soft patch partition $A \in [0,1]^{K \times P}$ and reconstructed features

$$A_{sp} = \mathrm{softmax}_s(a_{sp}), \qquad \widehat{\boldsymbol{f}}_p = \sum_{s=1}^{K} A_{sp} \widehat{\boldsymbol{f}}_{sp}, \qquad \mathcal{L}_{\mathrm{rec}} = \frac{1}{P d_f} \|\widehat{F} - F\|_F^2. \tag{2}$$

Whereas standard inference converts the decoder masks into a hard segmentation, our method retains the soft assignments A, which form the input to the consensus readout described below. We defer the standard Slot Attention recurrence to Appendix B.1.

Prefix training. Standard Slot Attention samples its slots from a shared learned Gaussian and typically trains at a fixed cardinality K (Locatello et al., 2020). To support multiple cardinalities within one checkpoint, we introduce an ordered bank of $K_{\max}$ position-specific Gaussian initializers. Each position s has a learned mean $\boldsymbol{\mu}_s$ and log-standard-deviation $\boldsymbol{\ell}_s$:

$$\boldsymbol{z}_s^{(0)} = \boldsymbol{\mu}_s + \exp(\boldsymbol{\ell}_s) \odot \boldsymbol{\epsilon}_s, \quad \boldsymbol{\epsilon}_s \sim \mathcal{N}(\mathbf{0}, I). \tag{3}$$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{\epsilon}_s$ is sampled during training. Requesting K slots passes Slot Attention the first K entries, $Z_{1:K}^{(0)} = (\boldsymbol{z}_1^{(0)}, \ldots, \boldsymbol{z}_K^{(0)})$ and leaves the rest unused.

We call $Z_{1:K}^{(0)}$ the active *prefix* and K the *cardinality* of that evaluation. All other parameters are shared across K. For the uniform schedule, we draw one cardinality for the whole batch from $\mathcal{K}_{\mathrm{train}}$ with probability $p(K) = 1/|\mathcal{K}_{\mathrm{train}}|$. Appendix D.3 ablates this choice against an endpoint-weighted schedule. Only the sampled prefix is evaluated and receives gradients, so a step has the cost of an ordinary Slot Attention step at that K. Over many steps the shared model minimizes

$$\mathcal{L}_{\mathrm{train}}(\theta) = \mathbb{E}_{\boldsymbol{x}} \left[\sum_{K \in \mathcal{K}_{\mathrm{train}}} p(K)\, \mathcal{L}_{\mathrm{rec}}^{(K)}(\boldsymbol{x}; \theta) \right]. \tag{4}$$

where $\mathcal{L}_{\mathrm{rec}}^{(K)}(\boldsymbol{x}; \theta)$ is the reconstruction loss under a prefix of length K. Position s is active only when the sampled cardinality is at least s, which gives exposure

$$\Pr(\text{position } s \text{ active}) = \sum_{K \in \mathcal{K}_{\mathrm{train}},\, K \geq s} p(K). \tag{5}$$

2.2 HORIZONTAL-CUT CONSENSUS

At inference, the frozen model produces a population $\mathcal{A}(\boldsymbol{x}) = \{A^{(h)}\}_{h=1}^{H}$ of soft partitions. Member h requests K_h slots and yields $A^{(h)} \in [0,1]^{K_h \times P}$. At inference, we call each evaluation of the frozen model a *view*, characterized by a requested cardinality K_h and an image transformation τ_h (identity or horizontal reflection and flip). Predictions from transformed inputs are spatially aligned to the original patch lattice before aggregation. The readout combines these partitions into a segmentation with a chosen output budget L, without additional training or segmentation labels.

Co-association. We represent each partition through pairwise patch co-membership:

$$C^{(h)} = (A^{(h)})^\top A^{(h)}, \quad C^{(h)}_{pq} = \sum_{s=1}^{K_h} A^{(h)}_{sp} A^{(h)}_{sq}. \tag{6}$$

The entry $C^{(h)}_{pq} \in [0,1]$ measures the evidence that patches p and q belong to the same slot. Since co-association sums over all slots, $C^{(h)}$ is order-agnostic. Every member produces a $P \times P$ matrix regardless of its cardinality, so no slot matching is required. This extends pairwise co-membership from cluster ensembles to soft object-centric partitions (Strehl & Ghosh, 2002; Fred & Jain, 2005; Punera & Ghosh, 2008). We average the evidence across members and convert it to dissimilarity:

$$C = \frac{1}{H} \sum_{h=1}^{H} C^{(h)}, \qquad D_{pq} = 1 - C_{pq} \quad (p \neq q). \tag{7}$$

Pairs have low dissimilarity when consistently assigned together, high when consistently kept apart.

Hierarchical clustering. We apply average-linkage agglomerative clustering to D (Murtagh & Contreras, 2012). Starting from one group per patch, the algorithm repeatedly merges the pair of groups with minimum mean dissimilarity:

$$\Delta(U,V) = \frac{1}{|U||V|} \sum_{p \in U} \sum_{q \in V} D_{pq}. \tag{8}$$

The resulting hierarchy starts with individual patches and progressively merges them into larger regions. Each merge records the dissimilarity between the groups being joined. Cutting the hierarchy at L groups assigns each patch to one of L regions, denoted by $\ell_L = \mathrm{Cut}_L(\mathcal{T})$, with region supports $R_r = \{p : \ell_L(p) = r\}$. The same hierarchy can be cut at different resolutions without recomputing the consensus. Consequently, the output cardinality L is independent of the member cardinalities $\{K_h\}_{h=1}^{H}$.

Dense rendering. The horizontal cut supplies discrete patch groups. To recover soft region evidence, we score each patch by its mean co-membership with each region:

$$B_{rp} = \frac{1}{|R_r|} \sum_{q \in R_r} C_{pq}, \quad M_{rp} = \frac{B_{rp}}{\max(\sum_j B_{jp}, \varepsilon)}. \tag{9}$$

Each row M_r is reshaped to the patch lattice and bilinearly upsampled to image resolution. The final dense segmentation is

$$\widehat{y}(u) = \arg \max_{r \in \{1,\dots,L\}} \mathrm{Up}(M_r)(u). \tag{10}$$

Thus, the hierarchy determines the region supports, while the averaged co-membership evidence determines the pixel-level assignments. The readout is invariant to permutations of the slot identities.

Proposition 1. *Let $H \geq 1$, $1 \leq L \leq P$, and let each $A^{(h)} \in [0,1]^{K_h \times P}$ have columns that sum to one on the same patch lattice. For arbitrary $K_h \times K_h$ permutation matrices Π_h, set $\widetilde{A}^{(h)} = \Pi_h A^{(h)}$. Let $\widetilde{C}$ be the consensus matrix computed from the $\widetilde{A}^{(h)}$. Then $\widetilde{C} = C$. With fixed clustering, region-ordering, and pixelwise tie rules, the dense outputs are identical.*

Proposition 1 shows that co-association does not depend on slot identities (proof in Appendix A.2).

Computational cost Constructing the co-association matrices requires $\mathcal{O}(P^2 \sum_{h=1}^{H} K_h)$ operations and $\mathcal{O}(P^2)$ working memory when members are accumulated sequentially. Average-linkage clustering adds pairwise computation over the P patches. For cardinality-only populations, all members reuse one encoder pass.

3 Experiments

3.1 Setup

Datasets. We evaluate PACE across three distinct visual settings. COCO 2017 provides cluttered natural scenes, with `train2017` for training and `val2017` for evaluation (Lin et al., 2014; COCO Consortium). VOC2012 offers a different category distribution and typically fewer annotated objects per image. We train on its 10 582 `trainaug` images and evaluate on the official 1449-image segmentation validation split (Everingham et al., 2010). MOVi-C provides synthetic multi-object scenes with dense instance annotations (Greff et al., 2022). We train a separate model per dataset without segmentation labels. MOVi-C frames are processed independently, without temporal cues.

Model and training. All primary comparisons within a dataset use the same prefix-trained checkpoint to isolate the effect of the readout from differences in training. Our DINOSAUR-derived model (Seitzer et al., 2023) uses frozen DINOv2 ViT-S/14 features (Oquab et al., 2024). A learned projection feeds three Slot Attention iterations and a broadcast MLP decoder. Each minibatch activates one prefix with $K \in \{3, 4, 5, 7, 9, 12\}$. COCO trains for 100 000 steps, and VOC2012 and MOVi-C for 40 000 steps. We train with three stochastic seeds, Adam, batch size 32, a peak learning rate of 2×10^{-4}, and cosine decay after a 5 % warm-up.

Evaluation protocol. We evaluate PACE with complementary metrics. FG-ARI measures foreground grouping, instance mBO overlap with individual objects, and matched mIoU the complete segmentation. Following DINOSAUR's seven-slot setting on COCO (Seitzer et al., 2023), we use the raw $K = 7$ prediction from the same checkpoint as our reference. For comparisons, we cut PACE hierarchies at $L = 7$ to match the reference output cardinality and retain this setting across COCO, VOC2012, and MOVi-C for consistency. We average three runs within each evaluation unit and compute 95 % BCa intervals by resampling images for COCO and VOC2012 and whole videos for MOVi-C. Appendix B defines the metrics, eligibility rules, and statistics.

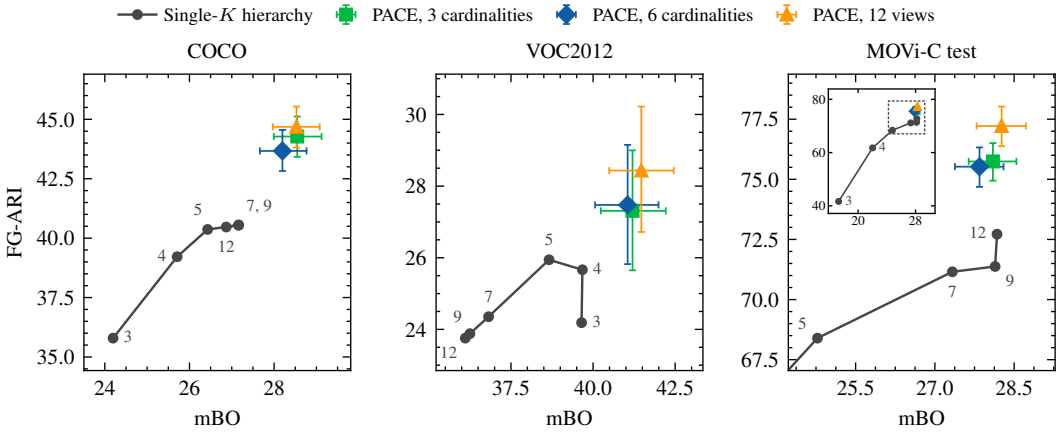

Figure 2: Single-K hierarchies (grey, labelled by K) and PACE at $L = 7$, from the same three models per dataset. PACE combines the 3 cardinalities $K = 3, 7, 12$, the 6 cardinalities $K = 3, 4, 5, 7, 9, 12$, or these 6 on the image and its horizontal mirror (12 views). Whiskers: 95% BCa intervals. The MOVi-C inset shows the full range; its dashed box marks the enlarged region.

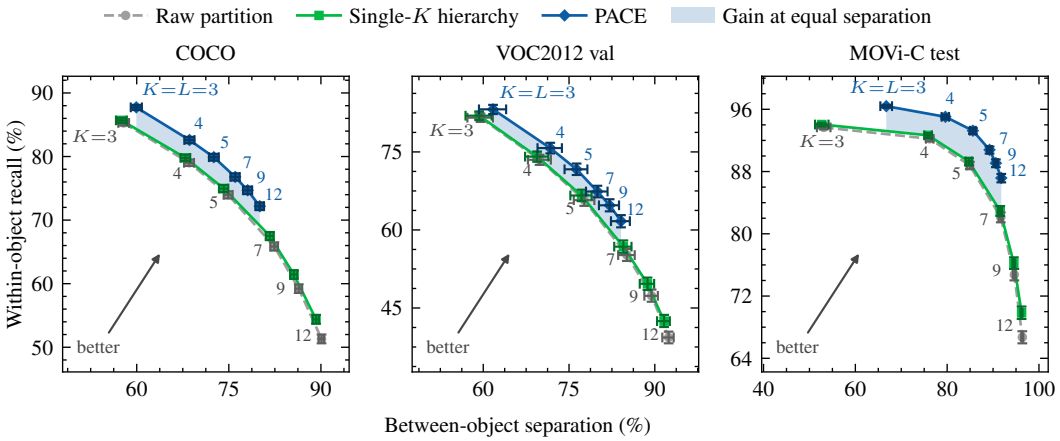

Figure 3: Between-object separation versus within-object recall (higher is better) as K increases, with hierarchies cut at $L = K$ and PACE combining all six cardinalities. Shading marks PACE's recall gain at equal separation. Three-seed means with 95% BCa intervals.

3.2 Results and Analysis

PACE extracts grouping information that is not available from any single cardinality in isolation. To quantify this advantage, we apply the same hierarchy and cut to each single partition with $K \in \{3, 4, 5, 7, 9, 12\}$, so that every output returns $L = 7$ regions from the same checkpoints (Figure 2). No single cardinality closes the gap. PACE exceeds every single-K hierarchy in FG-ARI on all three datasets, even the best one selected per metric with evaluation labels, and it also leads in mBO on COCO and VOC2012. The advantage is most pronounced in FG-ARI, which asks whether pairs of foreground pixels share a region as they do in the ground truth. This is precisely the pairwise relation that PACE averages, so a member that keeps an object intact can restore pairs that another member has split. mBO instead rewards the single region that best matches each object and therefore depends more on how finely the output is cut. MOVi-C makes this distinction visible, since its $K = 12$ hierarchy slightly exceeds six-cardinality PACE in overlap. Reflected views close this gap, and the analysis of output granularity below shows that it stems largely from the chosen output region count. Appendix S2 reports the corresponding comparison with the raw $K = 7$ prediction.

PACE mitigates the trade-off between splits and merges. FG-ARI conflates two kinds of error, splitting an object and merging distinct objects, and therefore conceals why no single K suffices. Figure 3 separates them into within-object pair recall and between-object pair separation. On every dataset, a larger request separates more pairs from different objects but also splits more pairs within the same object. Choosing K therefore only trades merged objects for split ones, so the single-K hierarchies trace one trade-off curve between the two errors, and a different K merely moves along it. This is the central motivation for combining cardinalities. PACE rises above this curve by drawing on both regimes at once. It retains more within-object pairs at matched separation, which indicates that its gain comes mainly from repairing fragmentation. Its separation exceeds that of single-K hierarchies at coarse cuts and falls below it at fine cuts. PACE therefore trades some additional merging for less fragmentation, consistent with its larger gains in FG-ARI than in mBO.

PACE derives its gains from complementary, scene-aligned grouping evidence rather than repeated sampling of the same decomposition regime. Figure 4 isolates this effect: repeated stochastic $K = 7$ inference saturates after only a few members, whereas PACE continues to improve as new cardinalities and reflected views enter the population. The explanation lies in what co-membership records. Because it is invariant to slot identity, a new initialization adds evidence only when it changes which patches are grouped together, and random restarts seldom do so. A different cardinality, by contrast, changes the grouping itself. Reflected views are informative for the same reason, as their masks are mapped back to the original coordinates and therefore perturb the decomposition while still describing the same scene. This scene correspondence proves essential. Averaging a partition with copies of itself leaves the output unchanged, whereas averaging partitions

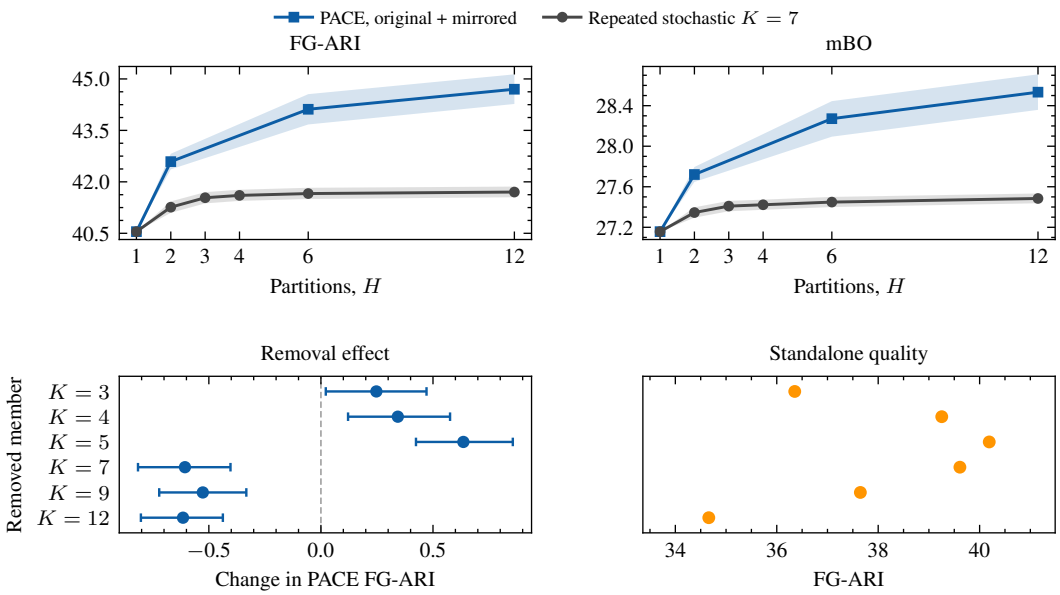

Figure 4: Population scaling and member contributions on COCO (three models). Top: scores as partitions H are added to one $K = 7$ hierarchy, with paired 95% BCa bands. Bottom: change in six-cardinality PACE FG-ARI when each member is removed (left), and that member's standalone FG-ARI (right).

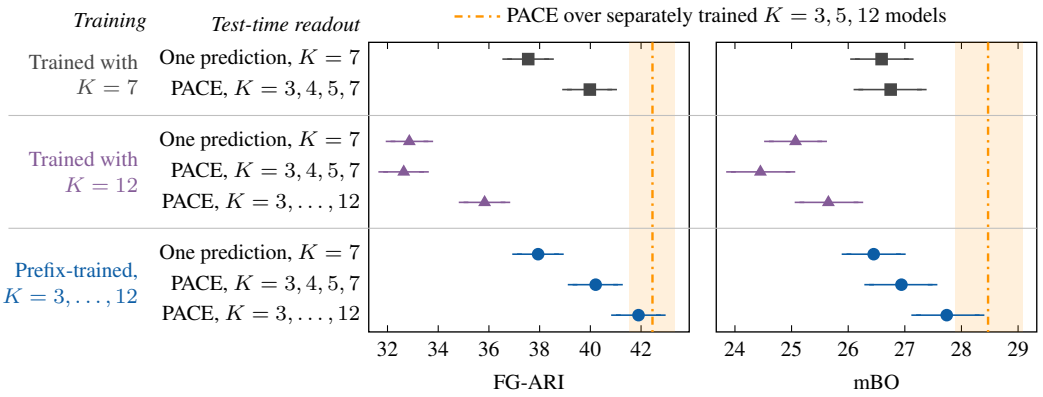

Figure 5: Attribution on COCO at $L = 7$ with 40,000-step models, one run each. A model runs with fewer slots than it was trained with by using the first K initializers of its bank (Section 2.1), and the model trained with $K = 7$ is not evaluated at $K = 9, 12$. Orange shows PACE over separately trained $K = 3, 5, 12$ models. Whiskers and band are 95% BCa intervals over 4,946 images.

of different images almost entirely destroys the grouping. The lower panels add a less intuitive observation. Removing $K = 5$, the strongest standalone partition, improves PACE, while removing the weak $K = 12$ partition harms it. A member is therefore valuable for the errors it does not share with the others rather than for its accuracy in isolation, which suggests that populations should be designed for complementarity. Appendices D.1 and D.2 extend this analysis to cardinality subsets and additional geometric transformations.

Prefix training approximates a multi-model ensemble Figure 5 compares prefix training with single-count training and with separate models per cardinality. PACE over the $K = 3, 4, 5, 7$ predictions of a conventional model trained with $K = 7$ already improves grouping, so the readout is a major source of the gain. Access to many slot counts is not sufficient, however. A model trained

with $K = 12$ can produce every deployed cardinality by keeping only its first K slot initializers, yet its partitions form an ineffective population. Its first slots were only ever trained alongside the rest and likely never learned to cover a whole scene alone. Useful members therefore appear to need training exposure to their own cardinality, which prefix training provides within a single checkpoint. Its benefit on cardinalities that the model trained with $K = 7$ already covers is small, and it mainly extends the range of useful cardinalities. The resulting population approaches separately trained models in foreground grouping, while those keep an advantage in overlap. In practice, one prefix-trained checkpoint and one encoder pass can thus stand in for a family of dedicated models.

Performance depends on output granularity. Because the hierarchy can be cut at any level, the output count becomes a deployment choice. Figure 6 traces the resulting trade-off and places it among published methods. Coarse cuts favour grouping and fine cuts favour overlap, since FG-ARI penalizes divided objects while mBO lets each object pick its best match from more regions. On COCO, every cut from $L = 4$ to $L = 12$ exceeds the FG-ARI of every cited unsupervised method by 0.1 to 1.0 points, a new state of the art within this comparison, although several methods achieve higher overlap. No cut reaches the strongest published overlap values, which belong to methods with transformer decoders, self-training, or feature-based refinement. On VOC2012, MetaSlot, DIAS, and the SSR-refined models lead on both metrics, while PACE exceeds the remaining methods in grouping. On MOVi-C, PACE comes close to the best framewise grouping but trails in overlap, and much of this gap is a matter of budget. Cutting the same hierarchy more finely raises overlap while leaving grouping essentially unchanged (Appendix Figure S3). Requested cardinality thus determines which grouping evidence is available, and the output count how finely it is returned.

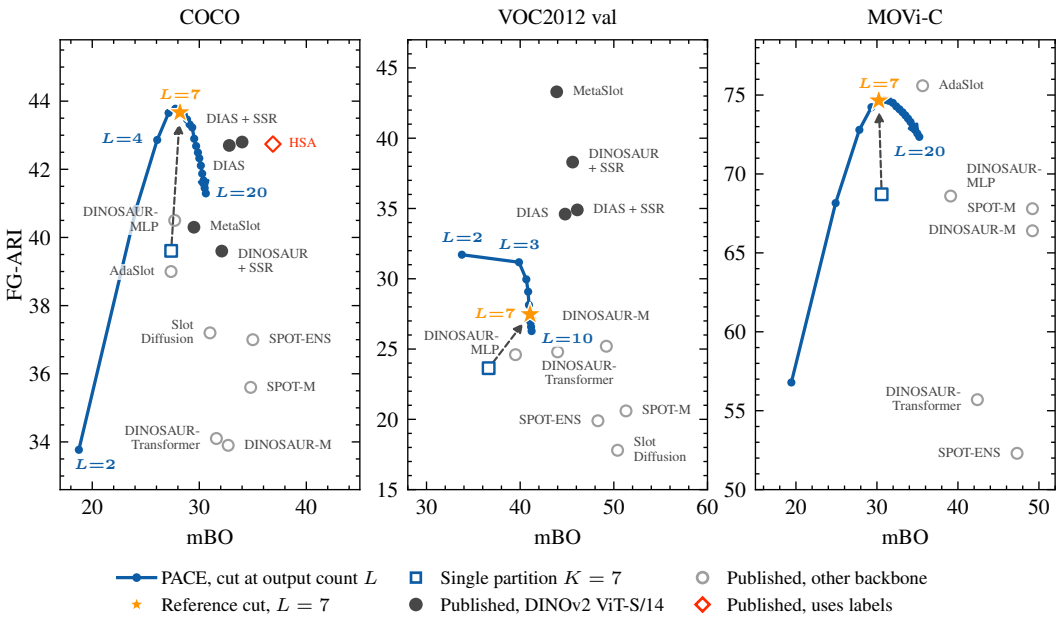

Figure 6: FG-ARI versus mBO for PACE across $L = 2, \ldots, 20$ (star: $L = 7$) and for published methods. MOVi-C uses the validation split. Filled circles: DINOv2 ViT-S/14; open circles: other backbones; red diamond: mask supervision.

4 Related Work

Object-centric learning. Slot Attention (Locatello et al., 2020) introduced an iterative attention mechanism that binds visual features to a set of latent slots and has become a foundation for subsequent object-centric models. Recent work has extended this framework to more complex natural images. DINOSAUR reconstructs frozen self-supervised ViT features (Seitzer et al., 2023), while SPOT introduces decoder-to-encoder self-training and autoregressive patch-order permutations (Kakogeorgiou et al., 2024). Other approaches combine features across multiple spatial

scales (Zhao et al., 2025), aggregate object representations across multiple transformer layers (Bock et al., 2026), incorporate additional semantic or instance-level guidance (Singh et al., 2025; Kori et al., 2024), or replaces the shared random initialization with separately optimized slot queries and bi-level optimization (Jia et al., 2023). *In contrast, we aggregate grouping information across multiple decompositions produced by the same model at different granularities.*

Adaptive cardinality and region consolidation. Slot cardinality affects grouping, with insufficient slots causing merges and excessive slots causing fragmentation (Zimmermann et al., 2023). AdaSlot selects active slots (Fan et al., 2024), MetaSlot removes redundant slots (Liu et al., 2025), and SlotCurri progressively increases capacity (Moon et al., 2026). *These methods seek to optimize the number of slots instance-wise, while we aggregate evidence across several slot cardinalities for each input scene.* Related methods consolidate regions through slot clustering, spectral merging, or redundancy removal (Aydemir et al., 2023; Tian et al., 2025; Chatzisavvas et al., 2026; Moon & Heo, 2026). *We instead combine complete scene decompositions for different values of K.*

Multiple predictions, refinement, and consensus. Masked Multi-Query Slot Attention combines multiple slot sets through Hungarian matching (Pramanik et al., 2024). *We instead combine different cardinalities through slot-invariant patch co-membership, which eliminates explicit correspondence.* SSR refines frozen object-centric predictions using transformer-derived affinities (Lu & Guo, 2026), while DenseCRF and spectral methods exploit image-space or feature-space evidence (Krähenbühl & Koltun, 2011; Melas-Kyriazi et al., 2022; Wang et al., 2022). *Our readout, on the other hand, derives its evidence solely from the model's predictions across granularities.* The underlying principle comes from cluster ensembles, which represent partitions through pairwise co-membership to remove arbitrary cluster identities (Strehl & Ghosh, 2002). Evidence accumulation and soft cluster ensembles aggregate hard or probabilistic assignments (Fred & Jain, 2005; Punera & Ghosh, 2008). *We extend this principle across slot counts and spatial coverage to obtain a fixed-budget consensus without additional training.*

5 Conclusion, Limitations, and Future Work

Conclusion. A slot model trained across cardinalities contains substantially more grouping information than is exposed by any single decomposition. PACE recovers this information by reconciling pairwise grouping evidence across requested slot counts. On COCO, VOC2012, and MOVi-C, consensus improves FG-ARI by 4.67, 3.67, and 5.26 points over the standard $K = 7$ prediction, respectively, without supervision, slot matching, or additional learned parameters. On COCO, every output count from $L = 4$ to $L = 12$ sets a new unsupervised state of the art in FG-ARI. These gains cannot be reproduced by selecting a single cardinality: coarse and fine decompositions exhibit complementary merge and split errors, repeated sampling of the same cardinality quickly saturates, and even weak standalone members can provide valuable evidence when their errors differ from those of the rest of the population. This perspective also separates the cardinality used to gather evidence from the cardinality of the final output. The requested slot counts determine which grouping hypotheses contribute to the consensus, while a single cut of the resulting hierarchy specifies the number of output regions at deployment. Prefix training makes this practical within a single checkpoint by exposing the model to multiple cardinalities during training. As a result, one model can provide useful decompositions across a range of slot counts, approaching the grouping quality of separately trained cardinality-specific models while sharing a common representation.

Limitations and future work. The output cardinality is currently selected externally; a label-free criterion based on hierarchy merge heights or agreement within the population could make this choice adaptive. Likewise, uniform averaging could be replaced by image-dependent weighting of consensus members. Errors shared across all decompositions cannot be corrected by consensus alone, and the pairwise co-association representation scales quadratically with the number of patches. These limitations suggest complementary directions for future work, including stronger decoders, boundary refinement, adaptive population weighting, and sparse approximations to pairwise grouping evidence.

Reproducibility Statement

The code is included as a separate archive in the supplementary material. It contains the readout, prefix training, the configuration of every model, one pipeline per group of experiments, the result records behind every table and figure, and the scripts that draw them. Section 2 and Appendix A specify the method, and Appendix B describes the models, metrics, and statistical procedures.

References

Görkay Aydemir, Weidi Xie, and Fatma Güney. Self-supervised object-centric learning for videos. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018. URL `https://arxiv.org/abs/1806.01261`.

Sebastian Bock, Leonie Schüßler, Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. MUFASA: A multi-layer framework for Slot Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27750–27760, 2026.

Christos Chatzisavvas, Panagiotis Rigas, George Ioannakis, Vassilis Katsouros, and Nikolaos Mitianoudis. When slots compete: Slot merging in object-centric learning. arXiv preprint arXiv:2603.11246, 2026.

COCO Consortium. COCO 2017 dataset. `https://cocodataset.org/`. Accessed Aug. 17, 2026.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, and Zheng Zhang. Adaptive Slot Attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23062–23071, 2024.

Ana L. N. Fred and Anil K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, 2005.

Klaus Greff, François Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Öztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3761, 2022.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Baoxiong Jia, Yu Liu, and Siyuan Huang. Improving object-centric learning with query optimization. In *International Conference on Learning Representations*, 2023.

Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzalos, and Nikos Komodakis. SPOT: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, pp. 4015–4026, 2023.

Avinash Kori, Francesco Locatello, Fabio De Sousa Ribeiro, Francesca Toni, and Ben Glocker. Grounded object-centric learning. In *International Conference on Learning Representations*, 2024.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011.

Joseph B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

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

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/lake18a.html`.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

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. arXiv preprint arXiv:2307.04767, 2023.

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 *European Conference on Computer Vision*, 2014.

Hongjia Liu, Rongzhen Zhao, Haohan Chen, and Joni Pajarinen. MetaSlot: Break through the fixed number of slots in object-centric learning. In *Advances in Neural Information Processing Systems*, 2025.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with Slot Attention. In *Advances in Neural Information Processing Systems*, 2020.

Xiaoqian Lu and Guangfu Guo. SSR: Similarity-shift refinement for training-free object-centric masks. arXiv preprint arXiv:2608.01103, 2026.

David R. Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 416–423, 2001.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

WonJun Moon and Jae-Pil Heo. Selective synergistic learning for video object-centric learning. arXiv preprint arXiv:2606.15527, 2026.

WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Reconstruction-guided slot curriculum: Addressing object over-fragmentation in video object-centric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25001–25010, 2026.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Stephen E. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4): 441–474, 1977.

Benjamin Peters and Nikolaus Kriegeskorte. Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 5:1127–1144, 2021.

Rishav Pramanik, José-Fabian Villa-Vásquez, and Marco Pedersoli. Masked multi-query Slot Attention for unsupervised object discovery. In *International Joint Conference on Neural Networks*, 2024.

Kunal Punera and Joydeep Ghosh. Consensus-based ensembles of soft clusterings. *Applied Artificial Intelligence*, 22(7–8):780–810, 2008.

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. arXiv preprint arXiv:2408.00714, 2024.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations*, 2023.

Roger N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2):125–140, 1962.

Krishnakant Singh, Simone Schaub-Meyer, and Stefan Roth. GLASS: Guided latent slot diffusion for object-centric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28673–28683, 2025.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1): 89–96, 2007. doi: 10.1111/j.1467-7687.2007.00569.x.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

Pinzhuo Tian, Shengjie Yang, Hang Yu, and Alex C. Kot. Pay attention to the foreground in object-centric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30281–30290, 2025.

Yangtao Wang, Xi Shen, Shell Hu, Yuan Yuan, James Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems*, volume 36, pp. 6941–6960, 2023.

Rongzhen Zhao, Vivienne Huiling Wang, Juho Kannala, and Joni Pajarinen. Multi-scale fusion for object representation. In *International Conference on Learning Representations*, 2025.

Roland S. Zimmermann, Sjoerd van Steenkiste, Mehdi S. M. Sajjadi, Thomas Kipf, and Klaus Greff. Sensitivity of slot-based object-centric models to their number of slots. arXiv preprint arXiv:2305.18890, 2023.

A Readout specification and mathematical properties

This supplement gives the inference specification, mathematical properties, experimental settings, and secondary analyses.

A.1 Complete inference procedure

Algorithm 1: Horizontal-Cut Consensus Readout

Input: Image $\boldsymbol{x}$, the frozen generator of Section 2.1, views $\{(K_h, \text{mirrored or not})\}_{h=1}^{H}$, and output budget L

Output: Dense segmentation $\widehat{y}$

for $h \leftarrow 1$ **to** H **do**
  Evaluate the frozen model at cardinality K_h on the view's image and flip the returned masks back if the view was mirrored, to obtain $A^{(h)}$;
  Form $C^{(h)} \leftarrow (A^{(h)})^\top A^{(h)}$;
Average $C \leftarrow 1/H \sum_h C^{(h)}$;
Form the off-diagonal dissimilarities $D_{pq} \leftarrow 1 - C_{pq}$ and build $\mathcal{T} \leftarrow \mathrm{HAC}_{\mathrm{avg}}(D)$;
Cut $\mathcal{T}$ so that exactly L groups $R_1, \ldots, R_L$ remain;
Compute $B_{rp} \leftarrow 1/|R_r| \sum_{q \in R_r} C_{pq}$ and normalize over r to obtain M;
Return $\widehat{y}(u) \leftarrow \arg\max_r \mathrm{Up}(M_r)(u)$;

A.2 Proof of slot-permutation invariance

Proof. The member-wise co-association matrix is $C^{(h)} = (A^{(h)})^\top A^{(h)}$. A permutation matrix Π_h rearranges slot indices, and its transpose reverses this rearrangement. Hence $\Pi_h^\top = \Pi_h^{-1}$, so $\Pi_h^\top \Pi_h = I_{K_h}$. It follows that

$$\begin{aligned}\widetilde{C}^{(h)} &= (\Pi_h A^{(h)})^\top (\Pi_h A^{(h)}) \\ &= (A^{(h)})^\top \underbrace{\Pi_h^\top \Pi_h}_{I_{K_h}} A^{(h)} = C^{(h)}.\end{aligned}$$

Thus, permuting the slots preserves all pairwise patch co-associations. Since this holds independently for every member,

$$\widetilde{C} = \frac{1}{H} \sum_{h=1}^{H} \widetilde{C}^{(h)} = C.$$

The clustering dissimilarities are therefore identical. □

B Experimental and statistical protocol

B.1 Background on Slot Attention

We use the multilayer-perceptron (MLP) decoder form of DINOSAUR, which applies Slot Attention to features from a pretrained vision model Locatello et al. (2020); Seitzer et al. (2023).

A frozen self-supervised vision transformer E divides an image $\boldsymbol{x}$ into P square patches and produces one feature vector for each patch. A trainable projection π then maps these features into the space used for grouping Seitzer et al. (2023); Oquab et al. (2024):

$$F = E(\boldsymbol{x}) \in \mathbb{R}^{P \times d_f}, \qquad G = \pi(F) \in \mathbb{R}^{P \times d}. \tag{S1}$$

Everything below is indexed by a patch $p \in \{1, \ldots, P\}$ on this lattice.

Slot Attention compresses the P patch vectors into K vectors, called *slots* Locatello et al. (2020). Each slot can summarize one candidate region, but no region or semantic category is assigned to it

in advance. Starting from Gaussian-initialized states $\{\boldsymbol{z}_s^{(0)}\}_{s=1}^K$, the following layer is repeated for $t = 0, \dots, T-1$:

$$\begin{aligned}
\boldsymbol{q}_s^{(t)} = W_q\,\mathrm{LN}(\boldsymbol{z}_s^{(t)}), \qquad \boldsymbol{k}_p = W_k\,\mathrm{LN}(\boldsymbol{g}_p), \qquad \boldsymbol{v}_p = W_v\,\mathrm{LN}(\boldsymbol{g}_p),\\
\alpha_{sp}^{(t)} = \mathrm{softmax}_s\left(\frac{(\boldsymbol{q}_s^{(t)})^\top \boldsymbol{k}_p}{\sqrt{d}}\right),\\
\bar{\alpha}_{sp}^{(t)} = \frac{\alpha_{sp}^{(t)}}{\sum_{r=1}^P \alpha_{sr}^{(t)} + \varepsilon},\\
\boldsymbol{u}_s^{(t)} = \sum_{p=1}^P \bar{\alpha}_{sp}^{(t)} \boldsymbol{v}_p,\\
\widetilde{\boldsymbol{z}}_s^{(t+1)} = \mathrm{GRU}(\boldsymbol{u}_s^{(t)}, \boldsymbol{z}_s^{(t)}),\\
\boldsymbol{z}_s^{(t+1)} = \widetilde{\boldsymbol{z}}_s^{(t+1)} + \mathrm{MLP}\Big(\mathrm{LN}(\widetilde{\boldsymbol{z}}_s^{(t+1)})\Big).
\end{aligned} \tag{S2}$$

The softmax is taken over slots, so $\sum_s \alpha_{sp}^{(t)} = 1$ for every patch. The second normalization forms a stable weighted mean over patches for each slot. In the GRU, $\boldsymbol{u}_s^{(t)}$ is the input and $\boldsymbol{z}_s^{(t)}$ is the recurrent state. After T iterations, the final slot set $Z^{(T)} = \{\boldsymbol{z}_s^{(T)}\}_{s=1}^K$ is passed to the decoder.

The final slots are passed to a broadcast MLP decoder D. A learned positional code tells the decoder which patch it is predicting. For every slot and patch, the decoder produces a feature proposal $\widehat{\boldsymbol{f}}_{sp}$ and an unnormalized mask score a_{sp}. Normalizing these scores over slots gives the decoder masks

$$A_{sp} = \frac{\exp(a_{sp})}{\sum_{j=1}^K \exp(a_{jp})}, \qquad \sum_{s=1}^K A_{sp} = 1, \tag{S3}$$

Training reconstructs the frozen features as the mask-weighted mixture

$$\widehat{\boldsymbol{f}}_p = \sum_{s=1}^K A_{sp} \widehat{\boldsymbol{f}}_{sp}, \qquad \mathcal{L}_{\mathrm{rec}} = \frac{1}{Pd_f}\big\|\widehat{F} - F\big\|_F^2, \tag{S4}$$

The probabilistic masks in A still require one slot assignment at every pixel. Standard practice upsamples each row A_s to image resolution and selects the largest value:

$$\widehat{y}_{\mathrm{raw}}(u) = \arg\max_{s \in \{1,\dots,K\}} \mathrm{Up}(A_s)(u). \tag{S5}$$

This operation discards uncertainty as a location split almost evenly between two slots is still assigned entirely to one. Our readout method keeps A and replaces only this final step.

B.2 Models of Additional Controls

Compute-matched model. Two resource-matched controls test whether one higher-resolution partition can use a comparable inference budget more effectively than consensus. The first matches runtime. It uses the primary frozen DINOv2 ViT-S/14 model with prefix training, a $K = 7$ prediction, and the same 100,000-step COCO schedule. Its 496×496 input reaches the encoder at 434×434, which produces a 31×31 patch lattice. The existing NVIDIA TITAN RTX protocol uses batch size 8, 10 warm-up iterations, and 40 measured iterations. Relative to the standard single-partition path, the higher-resolution model takes $3.10\times$ the runtime and six-cardinality consensus takes $2.98\times$. This comparison uses one training run per system.

Memory-matched model. The second control matches peak live GPU tensor allocation during batch-one online inference with model weights resident. Single-partition candidates vary encoder resolution and backbone and are screened without segmentation labels. Both the six-cardinality and twelve-view consensus budgets select the same DINOv2 ViT-S/14 model trained at fixed $K = 7$. Its 320-pixel input produces a 20×20 patch lattice, compared with the primary model's 16×16 lattice. The final RTX 3090 measurements give approximately 194.47 MiB for the single pass,

195.55 MiB for six-cardinality consensus, and 196.30 MiB for twelve-view consensus. The selected model follows the primary 100,000-step recipe for three seeds. Appendix B.4 details the resource selection and paired comparisons. The two controls use separate hardware and profiling protocols.

Attribution models. A separate group of models isolates the contribution of the training procedure. One is a conventional model trained at the fixed count $K = 7$ with no exposure to any other cardinality. Three more are trained at one count each, at $K = 3$, $K = 5$, and $K = 12$. These four models and the prefix-trained models they are compared against all use the primary architecture and a shortened 40,000-step COCO schedule, so the comparison holds the training budget fixed. Section 3.2 reports the comparison.

B.3 Segmentation metrics

Foreground adjusted Rand index. FG-ARI measures agreement between the predicted and the ground-truth grouping over foreground pixels alone. Let n_{ij} be the number of foreground pixels assigned to predicted region i and ground-truth object j, let $n_i^{\mathrm{pred}} = \sum_j n_{ij}$ and $n_j^{\mathrm{gt}} = \sum_i n_{ij}$ be its marginals, let n be the total foreground pixel count, and write $\mathrm{pairs}(r) = \binom{r}{2}$. With

$$N_{\mathrm{joint}} = \sum_{i,j} \mathrm{pairs}(n_{ij}), \quad N_{\mathrm{pred}} = \sum_i \mathrm{pairs}\left(n_i^{\mathrm{pred}}\right), \quad N_{\mathrm{gt}} = \sum_j \mathrm{pairs}\left(n_j^{\mathrm{gt}}\right), \tag{S6}$$

and the chance and ceiling terms $N_{\mathrm{expected}} = N_{\mathrm{pred}} N_{\mathrm{gt}} / \mathrm{pairs}(n)$ and $N_{\mathrm{maximum}} = (N_{\mathrm{pred}} + N_{\mathrm{gt}})/2$, the index is

$$\text{FG-ARI} = \frac{N_{\mathrm{joint}} - N_{\mathrm{expected}}}{N_{\mathrm{maximum}} - N_{\mathrm{expected}}}. \tag{S7}$$

N_{joint} counts foreground pixel pairs that both segmentations place together, and N_{expected} removes the agreement expected by chance. Because background is excluded, a region that extends past the true object boundary can still score well.

Instance mean best overlap. Let $G_1, \ldots, G_m$ be the ground-truth foreground instances and $P_1, \ldots, P_k$ the predicted regions. Each ground-truth instance independently selects the predicted region it overlaps best:

$$\mathrm{mBO} = \frac{1}{m} \sum_{j=1}^{m} \max_i \frac{|G_j \cap P_i|}{|G_j \cup P_i|}. \tag{S8}$$

The matching is independent per instance, so one predicted region may serve several instances. Background is not a target instance, and predicted pixels outside the matched instance enlarge the union and lower the score. Because every instance draws its own best match, a larger pool of candidate regions can raise mBO without any improvement in the decomposition, which is why comparisons across methods with different output budgets need care.

Mean intersection over union. mIoU scores the segmentation as a complete labelling, background included. Let $G_1, \ldots, G_m$ and $P_1, \ldots, P_k$ now include the background region on each side, and let $M_{ij} = |P_i \cap G_j| / |P_i \cup G_j|$. A one-to-one correspondence is obtained by maximum-weight bipartite matching, computed with the Hungarian algorithm Kuhn (1955),

$$\pi^\star = \arg\max_\pi \sum_{(i,j)\in\pi} M_{ij}, \qquad \mathrm{mIoU} = \frac{1}{m} \sum_{j=1}^{m} \widetilde{M}_j, \tag{S9}$$

where $\widetilde{M}_j = M_{ij}$ if G_j is matched to P_i and zero otherwise. Each predicted region may be used once, so unmatched ground-truth regions contribute zero and under-segmentation is penalised directly.

B.4 Aggregation, bootstrap, and multiplicity

Replicated quantities. The primary COCO, PASCAL VOC2012, and MOVi-C comparisons use three independently trained models per dataset. The main-model mechanism analyses also average those runs where reported. The training-cardinality schedule comparison uses three runs per schedule. Other single-checkpoint analyses are identified in their captions or tables.

Aggregation. A metric is computed independently for each image or video frame. For a contrast between two readouts we score both on the same input and take the paired difference. Results with three models first average that difference across models. Image datasets then retain one value per image. MOVi-C averages eligible frame values within each video before uncertainty is estimated. Single-seed transfer results retain the paired difference from that training run.

Interval construction. Image-dataset uncertainty uses a bootstrap over images. MOVi-C instead resamples whole videos because frames from one video are correlated. We draw 10,000 resamples of the applicable units, recompute the mean, and read a 95% interval from the resulting distribution. The interval is bias-corrected and accelerated, which adjusts the quantiles for skew and displacement of the point estimate Efron (1987). These intervals quantify sampling variation conditional on the evaluated checkpoints.

Multiplicity. Reporting many comparisons raises the chance that one crosses a fixed threshold by accident. COCO and PASCAL VOC2012 each use a pre-specified family of four comparisons. The family covers the six-member and twelve-member populations under FG-ARI and mBO. MOVi-C uses the same four contrasts within each split. Validation is the primary family and the test split is a separate secondary family. We apply Holm correction at $\alpha = 0.05$ within each family Holm (1979). Holm sorts the four p-values of a family from smallest to largest, compares the i-th against $\alpha/(m+1-i)$ with $m = 4$, and stops at the first comparison that fails. This holds the probability of any false rejection inside each family at 5%.

Memory-matched baseline selection. The memory control tests whether a higher-resolution single partition can use the same peak GPU-memory budget more effectively than consensus. We screen fixed-$K = 7$ models across input resolutions and DINOv2 encoder sizes using 64 COCO training images without segmentation labels. Selection uses peak live GPU tensor allocation during batch-one inference with model weights resident. Eligible candidates must use 90% to 100% of the corresponding consensus budget. Among candidates within $1.1\times$ the consensus latency, we select the one with the highest memory use. This procedure selects the DINOv2 ViT-S/14 model at 320-pixel input resolution for both consensus populations. Its peak allocation is 194.47 MiB, compared with 195.55 MiB for six-cardinality consensus and 196.30 MiB for twelve-view consensus. Profiling uses three fresh processes with 10 warm-up and 40 measured calls per process.

Resource-matched evaluation. All systems are evaluated on the same 256×256 COCO grid. The runtime-matched comparison uses one training run per system, while the memory-matched comparison averages three independently trained runs. For the memory control, paired differences are averaged across runs within each image. The pre-specified family contains four comparisons: six-cardinality and twelve-view consensus against the matched single partition under FG-ARI and instance mBO. Confidence intervals and multiplicity correction follow the bootstrap and Holm procedures described above.

B.5 Resource-matched controls

Resource-matched results. The gains from consensus could simply reflect its larger inference budget. To test this possibility, we compare consensus with denser single-partition models under matched runtime and peak GPU-memory budgets. Figure S1 asks whether spending the same resource budget on one stronger partition can reproduce the benefit of combining several decompositions.

Under matched runtime, the two approaches favour different aspects of segmentation. Consensus gives substantially stronger foreground grouping, whereas the denser single partition achieves higher instance overlap. Additional spatial resolution therefore improves object support, but does not recover the grouping information obtained from combining decompositions.

The same distinction becomes clearer under the memory-matched comparison. Increasing the resolution of a single partition improves on the standard-resolution baseline, yet its FG-ARI remains well below both consensus populations. In contrast, the mBO differences are small and their intervals include zero. The consensus advantage in foreground grouping therefore persists even when the competing single partition is allowed essentially the same peak GPU allocation.

This comparison also exposes a different cost profile. Because the consensus members are evaluated sequentially, weights can be reused and intermediate tensors released, keeping peak GPU allocation

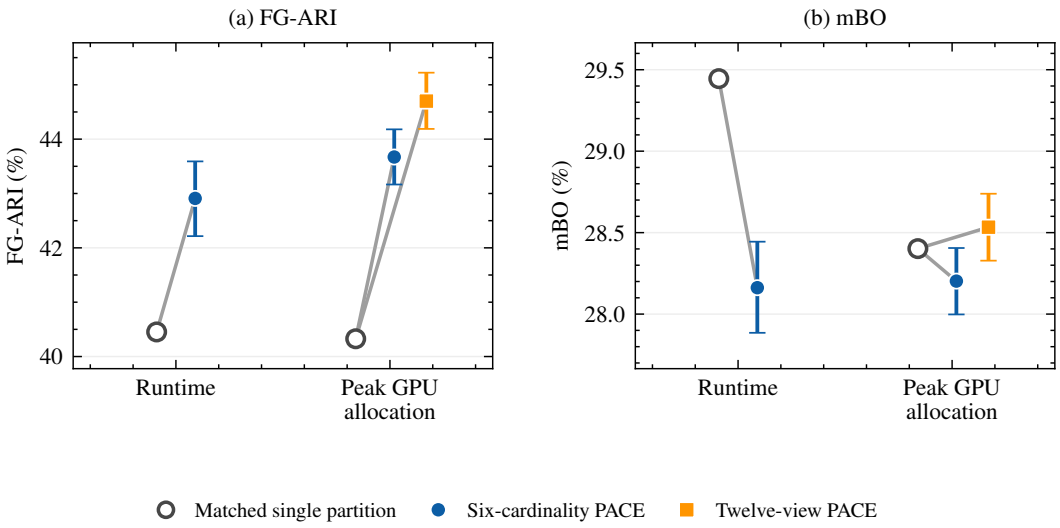

Figure S1: Resource-matched comparisons on COCO shown on the absolute metric scale. Panel (a) shows FG-ARI and panel (b) shows mBO. The horizontal categories distinguish matching by runtime from matching by peak GPU allocation. The runtime comparison includes six-cardinality PACE, while the peak-allocation comparison includes six-cardinality and twelve-view PACE. Hollow points show the matched single-partition controls, and coloured points show PACE. Lines connect each PACE result to its corresponding control. Whiskers show paired 95% BCa image-bootstrap intervals for the change relative to that control, translated onto the absolute metric scale.

close to that of a single prediction. The additional partitions instead increase latency. The comparison therefore supports a grouping benefit from complementary decompositions under matched resource budgets, while leaving runtime and memory as distinct deployment trade-offs.

Requested cardinality is especially economical because several partitions can reuse one encoded image. Spatial transformations require additional encoder passes. More broadly, the results reinforce the separation between the granularity used to generate candidate decompositions and the granularity ultimately returned by the readout.

C Output granularity and qualitative behavior

Requested cardinality K and readout budget L affect different stages.

C.1 Output-count sweep and single-cardinality hierarchies

At $L = 7$, six-cardinality PACE exceeds the strongest single-K hierarchy in FG-ARI by 3.11 [2.61, 3.61] points on COCO, 1.53 [0.92, 2.15] on VOC2012, and 2.81 [2.14, 3.48] on MOVi-C test. The corresponding mBO changes are 1.03 [0.85, 1.23], 1.38 [1.03, 1.73], and -0.33 [-0.49, -0.17]. These are paired 95% BCa intervals over images or whole videos. The best source K is chosen separately for each metric using evaluation labels, so the intervals condition on that dataset-level choice.

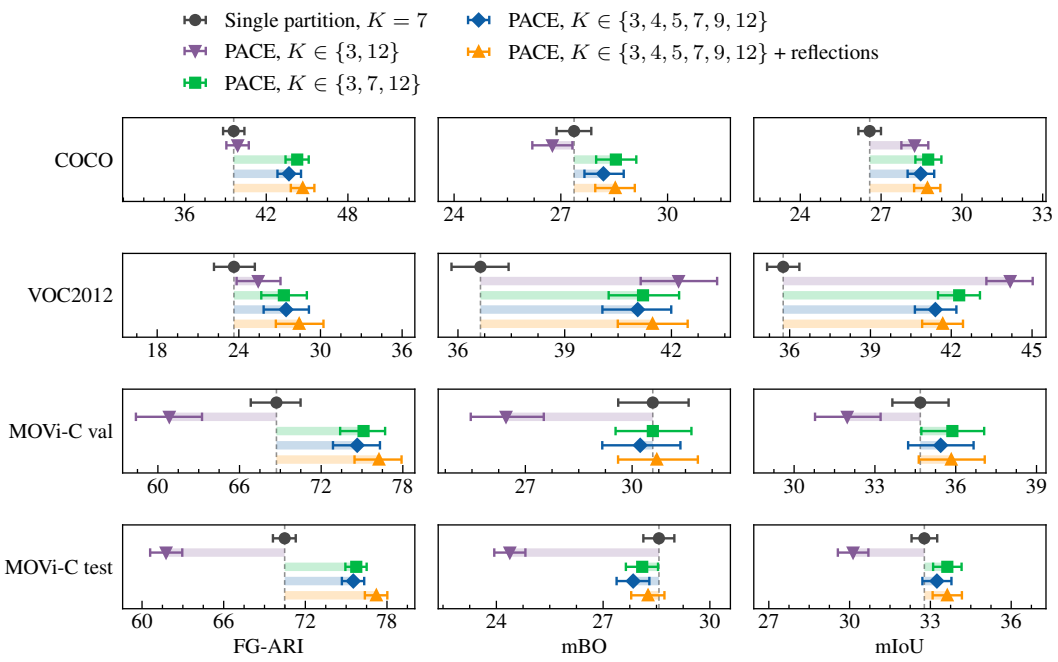

Figure S2: Full fixed-budget comparison on COCO, VOC2012, and MOVi-C validation and held-out-asset test. Raw $K = 7$ predictions are compared with PACE over the post hoc endpoint pair $\{3, 12\}$, three cardinalities $\{3, 7, 12\}$, six cardinalities $\{3, 4, 5, 7, 9, 12\}$, and those six cardinalities with horizontal reflection. All PACE outputs use $L = 7$. Markers show three-seed means with 95% BCa intervals over images or whole videos for MOVi-C. The dashed line marks the raw $K = 7$ mean in each panel.

All preceding results fix the output count at seven regions to ensure a fair comparison with the single-partition reference. Because this constraint is imposed by the evaluation rather than the hierarchy itself, which may be cut at any level, we next examine the sensitivity of performance to the chosen output granularity. Figure S3 varies the output count while holding each hierarchy fixed.

Consensus remains above the corresponding one-partition hierarchy in FG-ARI across the inspected range on COCO and VOC2012, and on MOVi-C from $L = 5$ onward. FG-ARI has an interior optimum because too few regions merge objects and too many fragment them. mBO improves as more regions become available because each ground-truth object can select its best overlapping prediction. The two metrics therefore prefer different cuts. This makes a single dataset-level L a metric-dependent compromise.

That reference uses $K = 7$. We also construct a separate hierarchy from each of the six requested cardinalities and cut it at every evaluated L. Figure S3 shows the highest mean score among these single-cardinality hierarchies at each output count. This oracle selects one source K per metric and count using the evaluation labels. On COCO and PASCAL VOC 2012, the mean score of six-cardinality consensus exceeds the oracle curve in both metrics across the inspected output counts. At $L = 7$, its grouping advantage remains positive against every individual source cardinality after Holm correction over the twelve metric-by-cardinality comparisons within each dataset.

On MOVi-C, consensus also exceeds every single-cardinality hierarchy in FG-ARI at $L = 7$, but its mBO is below the $K = 12$ hierarchy. Both directions survive the within-dataset Holm correction. Consensus does not exceed the best single-cardinality curve at every output count on this dataset.

The $K = 12$ source directly addresses whether a fine partition suffices. Applying HCC to that source and cutting at $L = 7$ improves its foreground grouping, but it remains below multi-cardinality consensus on COCO. These comparisons support a benefit from combining partitions beyond applying the same readout to one fine partition. The hierarchy clusters patches using soft masks, so its output can also reorganize the raw slot boundaries. On the held-out MOVi-C test split, finer hierarchy cuts

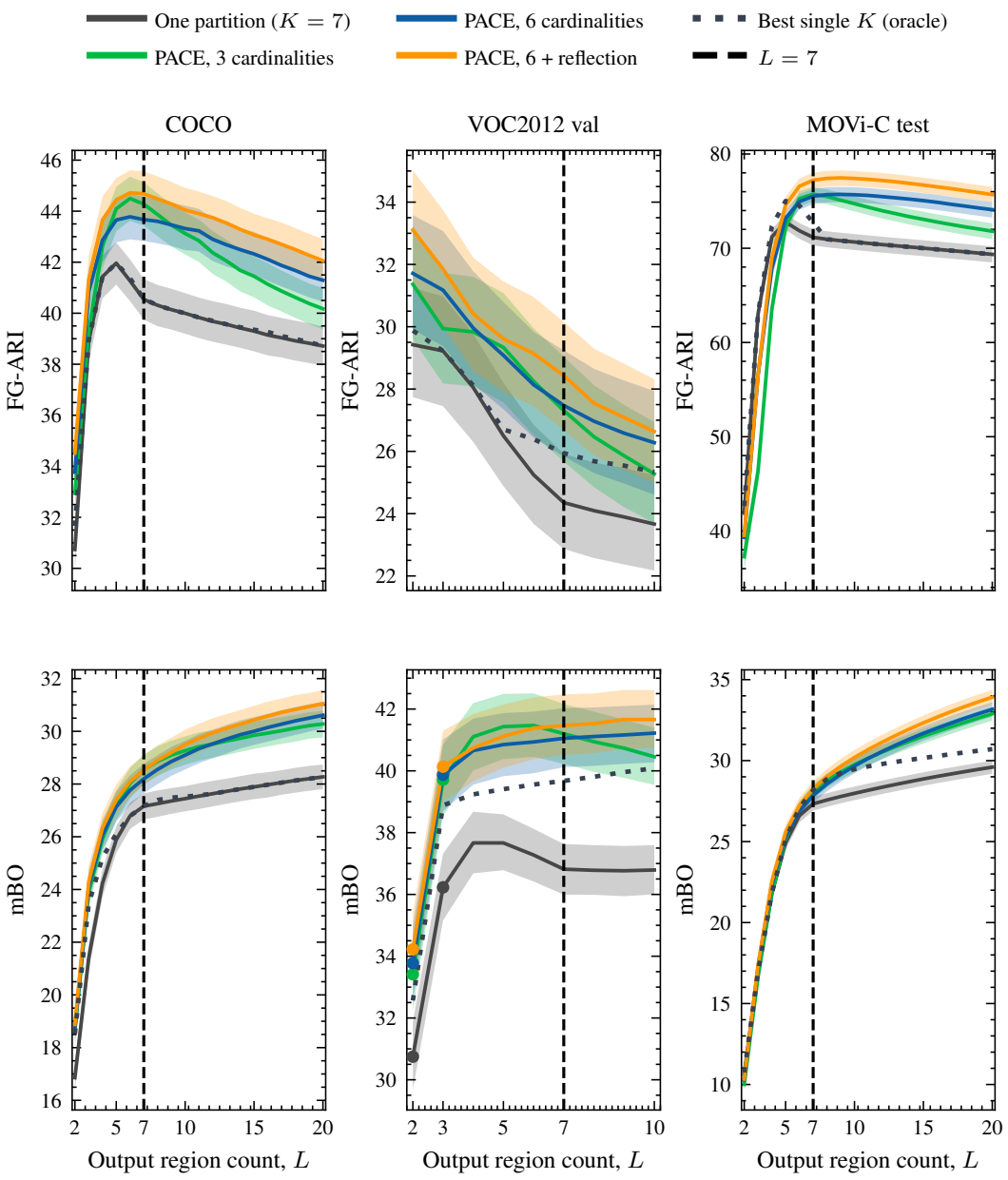

Figure S3: FG-ARI and mBO across output region count L on COCO and the MOVi-C test split for $L = 2, \ldots, 20$, and on VOC2012 validation for $L = 2, \ldots, 10$. Curves show the one-partition hierarchy, three-cardinality PACE, six-cardinality PACE, and six cardinalities with horizontally reflected predictions. The dotted grey curve shows the highest mean among six single-cardinality hierarchies at each output count. Its source K is selected separately for each metric and count using evaluation labels, so it is an oracle reference without a confidence band. Every dataset averages three independently trained models. Narrow shaded ribbons around the other curves show 95% BCa intervals over images for COCO and VOC and over videos for MOVi-C. The vertical dashed line marks the transferred $L = 7$ operating point.

improve object overlap while preserving foreground grouping, indicating that the overlap at $L = 7$ is partly limited by the output budget.

The requested cardinality K and the returned region count L therefore serve different roles. Several K values generate complementary intermediate evidence. The cut at L expresses the final output requirement.

Figure S4 scores each cardinality on its own.

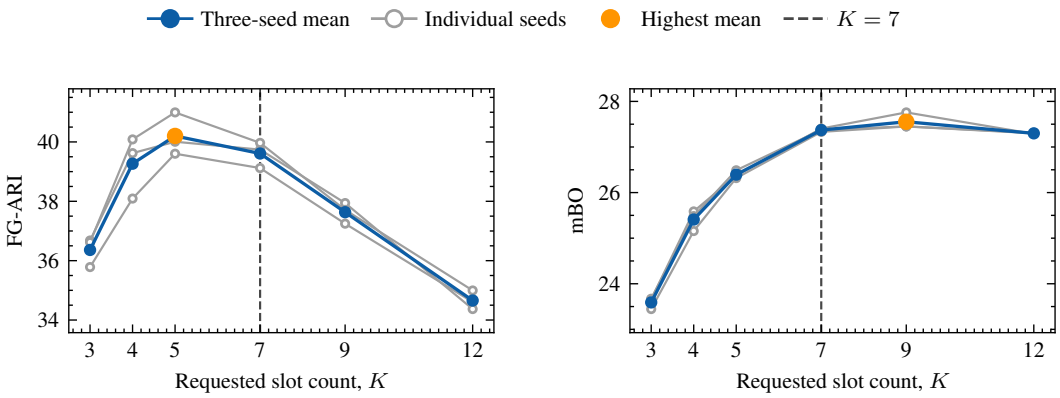

Figure S4: FG-ARI and mBO for single partitions at each requested slot count K on COCO, without consensus. The heavy line shows the mean over three independently trained 100,000-step models, with individual models as open markers. The dashed line marks the $K = 7$ baseline. Increasing K improves overlap but degrades foreground grouping.

FG-ARI favors lower cardinalities, while mBO benefits from finer decompositions. This reveals a trade-off between foreground grouping and object overlap and motivates the consensus readout, which aggregates complementary evidence across cardinalities and uses L to control the final granularity.

D Population composition and perturbations

D.1 Complementarity across cardinality subsets

The usefulness of a cardinality may depend on the other members of the population. Section 3.2 measures redundancy by removing one member from the complete population. The subset-averaged contribution instead evaluates its addition across every subset of the other cardinalities. Table S1 reports both quantities on the three 100,000-step COCO models together with standalone quality.

Table S1: Contribution of each requested cardinality on the three 100,000-step COCO models at $L = 7$, averaged within image across models. Leave-one-out removes each cardinality from the full six-member population, while subset-averaged effects measure its mean contribution across all subsets of the remaining cardinalities relative to $K = 7$. Standalone FG-ARI evaluates each partition individually. Values are percentage points.

Cardinality	LOO Δ FG-ARI [95% CI]	Subset contribution	Standalone FG-ARI
$K = 12$	$-0.62\,[-0.80, -0.44]$	$+0.89$	34.66
$K = 9$	$-0.53\,[-0.72, -0.33]$	$+0.82$	37.65
$K = 7$	$-0.61\,[-0.82, -0.40]$	reference	39.61
$K = 5$	$+0.64\,[+0.42, +0.86]$	$+0.14$	40.19
$K = 4$	$+0.34\,[+0.12, +0.58]$	$+0.36$	39.25
$K = 3$	$+0.25\,[+0.02, +0.47]$	$+0.29$	36.35

Both contribution measures invert the standalone-quality ranking. Cardinalities above the reference provide the strongest complementary evidence. $K = 12$ has the lowest standalone score but the largest subset-averaged contribution, while removing either $K = 12$ or $K = 9$ reduces FG-ARI. Lower cardinalities contribute less to the full population, despite $K = 5$ being the strongest standalone partition. All leave-one-out confidence intervals exclude zero.

Population composition thus matters more than size as carefully selected smaller subsets can outperform the complete set at lower cost, although reducing population size generally hurts performance on average. This highlights informed view selection as a promising route to more efficient consensus.

D.2 Diversity from additional input perturbations

Cardinality diversity provides complementary partitions of the same image. We next ask whether changing the input itself reveals additional grouping evidence, and whether this benefit depends on the choice of perturbation. We screen geometric and appearance transformations alongside spatial occlusions, aligning all predictions to the original patch lattice. Occluded regions are excluded from co-association, allowing the remaining visible patches to contribute evidence.

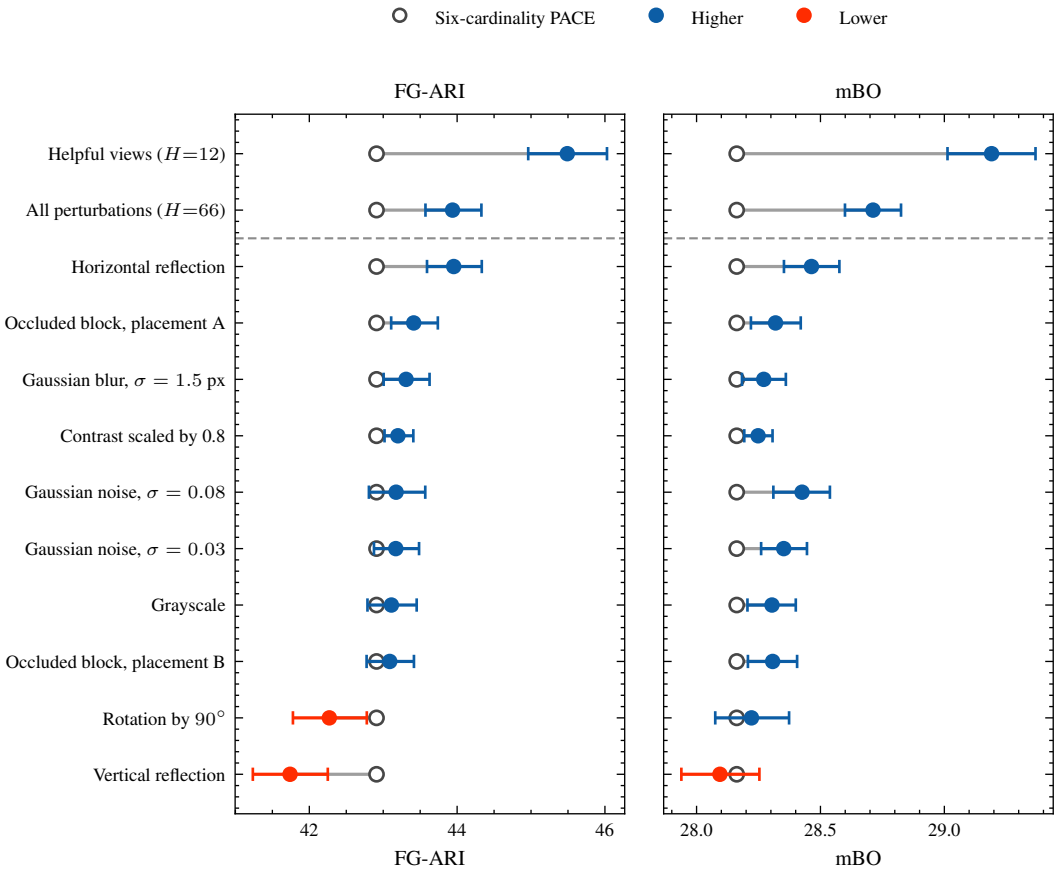

Figure S5: Effect of individual perturbations and combined view populations on COCO. Each individual perturbation is paired with the original-image cardinality population. The combined populations compare exhaustive aggregation with a compact selection of complementary views. Points show absolute scores, and whiskers show paired bootstrap confidence intervals relative to the cardinality-only reference.

Figure S5 reveals that the value of additional views depends strongly on how they alter the available grouping evidence. Occlusion and reduced contrast provide the clearest improvements, suggesting that partial visibility and changes in appearance can expose useful alternative partitions. Blur, noise, and grayscale conversion contribute less. In contrast, vertical reflection and rotation weaken foreground grouping, consistent with sensitivity to the upright orientation prevalent in the training images.

The combined populations reveal a further distinction between useful diversity and population size. An exhaustive combination of perturbations performs worse than a compact selection comprising the

original image, horizontal reflection, cropping, and occlusion. This selected population also requires substantially less inference. Thus, additional predictions are most valuable when they contribute complementary grouping evidence. Indiscriminate aggregation can instead dilute consensus, making population composition a more useful design consideration than member count alone.

Alignment and validity. Predictions from transformed images are aligned to the original patch lattice before aggregation. For horizontal reflection, we flip the predicted masks back to their original orientation. For crops and occlusions, a binary validity mask $v_{hp} \in \{0, 1\}$ indicates whether member h provides evidence for patch p. We average co-association only over members that cover both patches:

$$C_{pq} = \frac{\sum_h v_{hp} v_{hq} \sum_{s=1}^{K_h} A_{sp}^{(h)} A_{sq}^{(h)}}{\sum_h v_{hp} v_{hq}}, \tag{S10}$$

whenever the denominator is positive.

An uncovered pair is therefore excluded from that member's evidence. All evaluated populations include full-image members, so every patch pair has support. For complete images and reflections, all validity indicators equal one and the expression reduces to ordinary averaging.

D.3 Uniform versus endpoint-weighted training

The uniform schedule used in prefix-training is compared with an endpoint-weighted schedule on COCO. This alternative sets $p(3) = 0.15$, $p(12) = 0.25$, and $p(K) = 0.075$ for each interior count. It gives the endpoints extra exposure. The schedule was designed to offset the decline in row exposure toward $K_{\max}$. Three models were trained under each schedule for 40,000 steps on COCO. Their configurations differ only in the sampling distribution. The realized frequencies match the intended probabilities.

Table S2: Matched training-schedule comparison at $L = 7$, three independently trained 40,000-step COCO models per schedule, averaged over models. Disagreement is the label-free mean pairwise $1 - \mathrm{ARI}$ between population members. Changes are uniform minus endpoint weighted, with paired 95% BCa intervals over images.

Quantity	Endpoint weighted	Uniform	Change [95% CI]
FG-ARI	42.09	42.29	$+0.20\,[-0.09, +0.49]$
mBO	27.70	27.83	$+0.13\,[+0.02, +0.23]$
Population disagreement	0.4974	0.4920	$-0.0054\,[-0.0062, -0.0046]$

The schedule has little effect. FG-ARI is statistically indistinguishable between the two. Uniform sampling improves mBO by 0.13 points and lowers population disagreement slightly. All three effects are at most 0.20 points. The consensus effects in Section 3.2 are $+4.06$ FG-ARI and $+0.83$ mBO. Per-cardinality changes alternate in sign. The sampling schedule therefore has no material effect on the reported conclusions.

E Readout and weighting ablations

E.1 Ablation protocol

The consensus readout contains several design choices between the population of soft partitions and the final dense segmentation. We evaluate these choices through controlled ablations that alter one readout component at a time. The component ablations use the six-cardinality COCO population and, except for the output-count controls, fix the hierarchy cut at $L = 7$.

The reference readout follows Section 2.2. It forms soft patch co-association matrices, averages them uniformly, converts co-association to dissimilarity through $D = 1 - C$, constructs an average-linkage hierarchy, cuts the hierarchy at $L = 7$, and renders the resulting patch groups through consensus-based region membership. The ablations examine four stages of this construction: evidence formation, patch grouping, the dissimilarity transform, and partition rendering. A separate experiment tests whether the population members should receive unequal weights.

E.2 READOUT ALTERNATIVES

Let $A^{(h)}$, C, D, ℓ_L, and R_r follow the definitions in Section 2.2.

E.2.1 EVIDENCE CONSTRUCTION

The reference co-association retains the full soft slot assignments. The first ablation family tests whether this graded assignment information is important, whether performance depends on assignment sharpness, and whether uncertain patch assignments should contribute less to the consensus.

To vary assignment sharpness without changing the most likely slot, we apply a temperature τ before co-association:

$$A_{sp}^{(h,\tau)} = \frac{\left(A_{sp}^{(h)}\right)^{1/\tau}}{\sum_t \left(A_{tp}^{(h)}\right)^{1/\tau}}, \qquad C_{pq}^{(\tau)} = \frac{1}{H}\sum_{h=1}^{H}\sum_s A_{sp}^{(h,\tau)} A_{sq}^{(h,\tau)}. \tag{S11}$$

The reference corresponds to $\tau = 1$. Values below one sharpen the assignments, while values above one make them more diffuse. The **hard co-association control** replaces each assignment column by a one-hot vector at its largest entry. It therefore removes graded assignment uncertainty entirely.

We also test whether local assignment confidence provides a useful reliability signal. Define

$$u_{hp} = 1 + \frac{\sum_s A_{sp}^{(h)} \log A_{sp}^{(h)}}{\log K_{\max}}, \qquad C_{pq}^{\text{ent}} = \frac{\sum_h u_{hp} u_{hq} C_{pq}^{(h)}}{\sum_h u_{hp} u_{hq}}. \tag{S12}$$

This construction gives greater influence to a patch pair when both assignments have low entropy.

A further control isolates an implementation choice in the co-association matrix. The **direct-mask control** inserts the cached decoder masks directly into Eq. equation 6 without additional normalisation.

E.2.2 PATCH GROUPING

The co-association matrix specifies pairwise evidence, but a grouping rule is still required to convert this evidence into regions. The reference uses average-linkage agglomerative clustering. We first compare alternative linkage criteria within the same hierarchical construction, then compare the hierarchy with methods that partition the patches directly.

For groups U and V, average, single, and complete linkage are

$$\Delta_{\text{avg}}(U,V) = \frac{1}{|U||V|}\sum_{p\in U}\sum_{q\in V} D_{pq}, \tag{S13}$$

$$\Delta_{\text{single}}(U,V) = \min_{p\in U,\ q\in V} D_{pq}, \tag{S14}$$

$$\Delta_{\text{complete}}(U,V) = \max_{p\in U,\ q\in V} D_{pq}. \tag{S15}$$

Average linkage aggregates evidence over all cross-group patch pairs. **Single linkage** instead permits one close patch pair to determine a merge, while **complete linkage** requires all cross-group pairs to remain sufficiently close.

Weighted linkage differs from average linkage in how previous merges contribute. If $U = X \cup Y$, it updates the distance to another group V according to

$$\Delta_{\text{weighted}}(U,V) = \frac{1}{2}\left[\Delta(X,V) + \Delta(Y,V)\right]. \tag{S16}$$

The two child groups therefore receive equal influence regardless of their sizes.

The **direct partitioning controls** test whether the hierarchical construction itself contributes beyond the underlying consensus evidence. Define the concatenated patch representation

$$\boldsymbol{z}_p = \left[A_{:p}^{(1)}; \ldots; A_{:p}^{(H)}\right]. \tag{S17}$$

The k-**means control** partitions $\{\boldsymbol{z}_p\}_{p=1}^{P}$ into L Euclidean clusters. **Spectral clustering** instead forms

$$S = Q^{-1/2} C Q^{-1/2}, \qquad Q = \mathrm{diag}(C\mathbf{1}), \tag{S18}$$

then applies k-means to the row-normalised top L eigenvectors of S.

Anchor aggregation tests a different alternative to slot-invariant co-association. It explicitly aligns every member to the $K = 7$ partition through mask intersection over union. Matching is one-to-one for $K_h \leq 7$ and many-to-one for $K_h > 7$. Masks assigned to the same anchor are averaged to produce seven outputs. This control tests whether explicit slot alignment can replace the pairwise consensus construction.

E.2.3 Dissimilarity transform

Hierarchical clustering requires a dissimilarity between patches, whereas the consensus matrix provides a similarity $C_{pq} \in [0, 1]$. The reference uses the direct complement $D_{pq} = 1 - C_{pq}$, which preserves the co-association scale linearly. We therefore test two alternative transforms that retain the ordering induced by C but change the spacing between dissimilarities.

The **negative-log transform**

$$D_{pq}^{\log} = -\log(\max(C_{pq}, \varepsilon)). \tag{S19}$$

is close to $1 - C_{pq}$ when C_{pq} is near one, but assigns increasingly large distances to pairs with weak co-association. It therefore tests whether the hierarchy benefits from stronger separation of poorly supported patch pairs.

As a complementary control, the **rank transform** discards the magnitudes of the co-association values and retains only their global ordering, following the ordinal treatment of dissimilarities in non-metric multidimensional scaling Shepard (1962); Kruskal (1964):

$$D_{pq}^{\mathrm{rank}} = \frac{\mathrm{rank}_{i<j}\,(1 - C_{ij}; , 1 - C_{pq})}{\binom{P}{2} - 1}. \tag{S20}$$

The rank is computed over distinct patch pairs and starts at zero. This transform maps the ordered pairwise dissimilarities to approximately uniform spacing. It therefore tests whether the absolute differences between co-association values contain useful information beyond their relative ordering.

Together, these controls distinguish sensitivity to the particular linear choice $1 - C$ from sensitivity to the underlying ordering of the pairwise consensus evidence.

E.2.4 Partition rendering

The hierarchy defines region supports on the patch lattice, but evaluation requires a dense pixel-level segmentation. We therefore compare several ways of rendering the resulting partition.

The **reference renderer** retains the soft consensus evidence after the hierarchy cut. It computes the region evidence and normalised memberships from Eqs. equation 9 and equation 9, then applies bilinear upsampling followed by a pixelwise argmax.

The **leave-one-out renderer** tests whether a patch's own diagonal consensus entry contributes materially to its region assignment. For a patch p inside region R_r, the self-term is excluded:

$$B_{rp}^{\mathrm{ex}} = \frac{\sum_{\substack{q \in R_r \\ q \neq p}} C_{pq}}{|R_r| - \mathbf{1}[p \in R_r]}. \tag{S21}$$

A singleton region receives a zero self-score.

The **discrete renderer** tests whether the soft consensus evidence remains useful once the hierarchy has fixed the region supports. It replaces the soft memberships with the binary partition induced by the hierarchy cut:

$$M_{rp}^{\mathrm{cut}} = \mathbf{1}[\ell_L(p) = r]. \tag{S22}$$

These binary patch maps are then bilinearly upsampled.

The **nearest-patch renderer** additionally tests the effect of interpolation. It upsamples the discrete patch labels ℓ_L directly with nearest-neighbour interpolation, so each output pixel inherits the label of its nearest patch.

Finally, the **member-mask renderer** tests whether the hierarchy can provide only the region structure while the original slot masks provide the rendering evidence. For each region R_r, we first measure how strongly slot s of member h supports that region:

$$w_{sr}^{(h)} = \frac{1}{|R_r|} \sum_{p \in R_r} A_{sp}^{(h)}, \qquad \widehat{w}_{sr}^{(h)} = \frac{w_{sr}^{(h)}}{\sum_t w_{tr}^{(h)}}, \tag{S23}$$

and use these region-specific slot weights to reconstruct a soft region score:

$$B_{rp}^{\mathrm{mem}} = \frac{1}{H} \sum_{h=1}^{H} \sum_{s} \widehat{w}_{sr}^{(h)} A_{sp}^{(h)}. \tag{S24}$$

The resulting region scores are normalised across regions as in Eq. equation 9.

E.3 Readout component results

Table S3 reports the component ablations on one 100,000-step COCO model. Each comparison changes one readout choice relative to the reference construction.

Table S3: Single-component readout ablations on one 100,000-step COCO model over 4,946 images. The hierarchy is cut at $L = 7$. Within each stage, the best value in each metric column is shown in bold. The two metrics select different configurations for pairwise evidence and for the renderer.

Stage	Configuration	FG-ARI	mBO
Reference	selected readout	42.91	28.16
Pairwise evidence	hard co-association	41.16	27.96
	mask temperature 0.5	42.45	**28.38**
	mask temperature 2	42.06	27.30
	mask temperature 4	40.48	26.37
	direct decoder masks	**42.91**	28.16
	entropy-weighted pair evidence	42.88	28.22
Patch grouping	complete linkage	**41.19**	**27.34**
	weighted linkage	40.13	26.57
	single linkage	33.09	22.82
	k-means on concatenated masks	39.42	26.17
	spectral clustering	39.28	26.08
	anchor aggregation	38.91	27.12
Dissimilarity	negative logarithm	43.13	28.20
	rescaled rank	**43.17**	**28.27**
Renderer	member masks, Eq. equation S24	43.51	28.05
	leave-one-out membership, Eq. equation S21	**43.52**	27.98
	discrete cut, Eq. equation S22	42.60	**28.20**
	nearest-patch labels	41.85	28.08

The evidence ablations show that the readout benefits from graded slot assignments. Hard co-association reduces FG-ARI by 1.74 points, and every tested departure from the reference mask temperature also reduces foreground grouping. The direct-mask control reproduces the reference result, as expected from the normalisation in Eq. equation 2. Entropy weighting has negligible effect, which provides no evidence that assignment entropy is a useful local reliability signal.

The grouping ablations provide the strongest distinction between alternative constructions. Single linkage performs substantially worse because one close patch pair can connect otherwise dissimilar

groups. Complete linkage also reduces performance because one dissimilar pair can prevent a merge. Weighted linkage gives equal influence to child groups of unequal size and performs below average linkage. The three direct partitioning controls also reduce FG-ARI. These results support average-linkage aggregation of the co-association evidence and indicate that the hierarchy contributes beyond the pairwise representation alone.

The logarithmic and rank-based dissimilarities change FG-ARI by only 0.22 and 0.26 points. The construction is therefore relatively insensitive to the exact numerical transformation from co-association to dissimilarity.

The rendering variants have a smaller effect than changes to the grouping stage. Discrete rendering reduces FG-ARI, while nearest-neighbour patch assignment causes a further decline. This result indicates that soft region evidence provides an additional benefit after the hierarchical structure has been constructed. Leave-one-out membership yields a slight improvement in FG-ARI, which suggests that the gain does not depend on a patch's own self-evidence. Reconstruction from the original member masks produces a similar result and demonstrates that the hierarchy can also provide the region-level structure for an alternative soft rendering scheme.

Average linkage performs better than the tested grouping alternatives; this is a controlled empirical result rather than a universal claim about linkage rules. The member-mask and leave-one-out renderers improve FG-ARI but reduce mBO. Consequently, the default balances simplicity and the two metrics, and is not the best-performing configuration in every component sweep.

E.4 Member weighting

The reference consensus averages all population members uniformly. This choice makes no assumption about which requested cardinalities provide more reliable grouping evidence. We test whether simple unsupervised reliability signals can improve the consensus through unequal member weights.

A weighted consensus takes the form

$$C(w) = \frac{\sum_{h=1}^{H} w_h C^{(h)}}{\sum_{h=1}^{H} w_h}, \qquad w_h \geq 0. \tag{S25}$$

E.4.1 Weighting rules

We consider four weighting rules that assign member reliability from different observable properties.

The **cardinality rule** tests whether consensus improves when finer or coarser decompositions are systematically preferred. We index the cardinality-ordered members from coarse to fine by

$$c_h = \frac{2h - H - 1}{H - 1}. \tag{S26}$$

The corresponding weights are

$$w_h^{\mathrm{tilt}} \propto \exp(\alpha c_h), \tag{S27}$$

where positive α gives greater weight to finer members and negative α favours coarser members.

The **view-invariance rule** treats consistency under horizontal reflection as evidence that a member provides a stable decomposition. Let $\widetilde{C}^{(h)}$ denote the co-association matrix obtained after horizontal reflection. We measure the stability of member h as

$$t_h = 1 - \frac{1}{P^2} \sum_{p,q} \left| C_{pq}^{(h)} - \widetilde{C}_{pq}^{(h)} \right|, \tag{S28}$$

and assign weights

$$w_h^{\mathrm{view}} \propto \exp\left(\beta \left(t_h - \max_g t_g \right) \right). \tag{S29}$$

The parameter β controls how strongly the rule concentrates weight on the most stable members.

The **population-agreement rule** gives greater influence to members whose co-association structure is similar to that of the population as a whole. We define

$$a_h = \frac{1}{H} \sum_{g} \cos\left(C^{(h)}, C^{(g)} \right), \tag{S30}$$

with weights

$$w_h^{\text{agree}} \propto a_h. \tag{S31}$$

Finally, the **reconstruction rule** tests whether the generator's feature-reconstruction error provides a useful proxy for member quality. Let e_h denote the reconstruction error of member h, and define its normalised value as

$$\widehat{e}_h = \frac{e_h - \min_g e_g}{\mathrm{sd}(e)}. \tag{S32}$$

Members with lower reconstruction error then receive greater weight:

$$w_h^{\text{recon}} \propto \exp(-\widehat{e}_h). \tag{S33}$$

All weights are normalised through Eq. equation S25. We additionally evaluate a fixed grid of 385 weight vectors on the simplex

$$\left\{ w \geq 0 : \sum_h w_h = 1 \right\}. \tag{S34}$$

This free search uses evaluation performance to select a fixed image-independent weighting and therefore serves only as an empirical upper reference for this class of weighting rules.

E.4.2 WEIGHTING RESULTS

Table S4 reports the weighting ablation on one 40,000-step COCO model.

Table S4: Member-weighting ablation on one 40,000-step COCO model at $L = 7$ over 4,946 images. The best value in each metric column is shown in bold.

Weighting rule	FG-ARI	mBO
Uniform	42.00	27.76
Cardinality tilt, $\alpha = +0.5$	42.50	27.85
View invariance, $\beta = 8$	42.38	27.91
Population agreement	42.03	27.73
Reconstruction error	41.50	27.46
Free search, 385 candidates	**42.61**	**27.99**

The cardinality tilt is the strongest rule-based alternative and gives greater weight to finer members. Its 0.50-point FG-ARI improvement remains below the 0.61-point gain from the free search. Both halves of the cardinality range remain useful, consistent with the subset analysis in Appendix D.1. View invariance provides a smaller improvement. Population agreement closely matches uniform weighting, while reconstruction error does not provide a useful proxy for grouping quality.

The limited advantage of the free search shows that fixed unequal weights provide little additional headroom over the simple cardinality tilt. Uniform weighting therefore remains a simple label-free default. The larger per-image oracle gap indicates that further gains would require a more informative image-conditioned reliability signal.