WHAT YOU SHOW IS NOT WHAT YOU GET: RETHINKING DEMONSTRATION FOLLOWING IN VISUAL IN-CONTEXT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual in-context segmentation uses visual demonstrations to guide predictions for a new image. We identify background color mapping deviation, where non-target regions sharing a background label receive different output colors, and part-to-parent expansion, where predictions extend beyond the requested part into its parent object. These deviations can occur even when the relevant regions remain distinguishable in internal features. Controlled changes to demonstration colors and requested targets support an account in which demonstrations guide role and color assignments over familiar semantic regions, influenced by learned semantic and color preferences. Building on these findings, we introduce Semantic Contrast Prompt Selection (SCPS). Its semantic contrast encoding assigns separate colors to the requested part and the rest of its object while keeping the remaining scene black. SCPS combines existing retrieval scores with quality estimates from labeled development images to select demonstrations and their binary or semantic contrast encodings. Averaged across seven retrieval methods on three benchmarks, SCPS improves target IoU by 0.054 and reduces leakage into the rest of the object by 22.7% relative to the original binary selectors. Our findings offer a new perspective on understanding and improving demonstration following in visual in-context segmentation.

1 INTRODUCTION

In-context learning (ICL) enables models to perform tasks from demonstrations without parameter updates (Min et al., 2022). Visual in-context learning (VICL) extends this paradigm to vision through paired input and output images that guide predictions for a query image (Bar et al., 2022; Wang et al., 2023a). Research on language ICL distinguishes task recognition, which elicits familiar task knowledge, from task learning, which infers mappings unseen during pretraining (Pan et al., 2023). In VICL models supporting heterogeneous tasks, activations associated with specific tasks can guide predictions without the original demonstrations (Hojel et al., 2025), while unfamiliar demonstrated encodings can elicit familiar task outputs or encodings used during training (Khatri et al., 2026). These findings suggest that visual demonstrations can elicit learned behavior without ensuring adherence to the demonstrated mapping. In visual in-context segmentation, training strategies such as randomization of label colors encourage reliance on demonstrations within a single task family (Wang et al., 2023b). Whether such training leads to faithful demonstration following remains an open question.

A segmentation demonstration specifies both target regions and output encodings. We study adherence to these specifications through RGB predictions and identify two deviations, illustrated in Figure 1. First, although requested targets generally receive the demonstrated foreground color, non-target regions sharing the same background label can receive different output colors. Some receive the demonstrated non-black background color, while others remain black. We refer to this selective behavior as *background color mapping deviation*. Second, a prediction can extend beyond the demonstrated part into its enclosing object, which we refer to as the parent object. We call this *part-to-parent expansion*. This occurs even though the demonstration mask explicitly labels the rest of the parent object as background.

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.

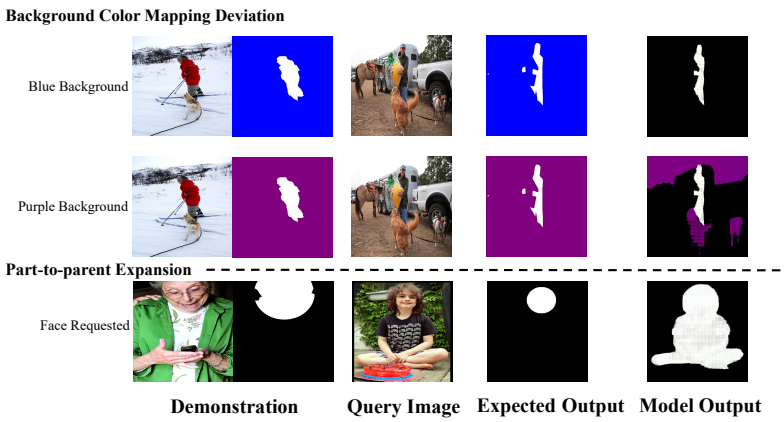


Figure 1: **Deviations from demonstration following.** Expected outputs apply the demonstrated colors to the annotated query regions. Top: non-target regions sharing the same background label receive different output colors. Bottom: a demonstration marking only the face yields a prediction extending to the whole person.

To understand background color mapping deviation, we vary demonstration colors and requested targets while examining predictions and internal representations. Non-target regions respond differently depending on their semantic identity, demonstrated role, and color, yet remain distinguishable from the remaining scene in late features. These findings support a mechanistic account in which the model recognizes familiar semantic regions and uses demonstrations to assign their roles and colors. Learned semantic and color preferences influence how these assignments appear in the output.

Our findings on selective background coloring also help explain part-to-parent expansion. Linear probes trained separately for each task on disjoint queries distinguish the requested part from the *parent-only region*, defined as the parent object excluding that part, in most analyzed expansion cases, although both receive the foreground color. A shared probe further shows that late features retain information about which part the demonstration requests. In face segmentation, changing the demonstrated background color selectively recolors much of the body while leaving the remaining scene predominantly black. Together, these findings connect part-to-parent expansion to how demonstration cues affect the output colors of internally distinguishable regions.

The generally reliable coloring of requested targets, together with the response of parent-only regions to background color changes, motivates explicitly marking the parent-only region with its own color. We call this a *semantic contrast encoding*, which assigns distinct colors to the target, parent-only region, and remaining scene. We keep the remaining scene black, consistent with its tendency to remain black in our experiments, and separately color the parent-only region to cue its exclusion from the target. Although this encoding reduces parent leakage, applying it uniformly can impair target segmentation. We therefore introduce Semantic Contrast Prompt Selection (SCPS), which combines existing selection scores with target quality estimates computed on development data to jointly select a demonstration and its binary or semantic contrast encoding.

Aggregated across three part segmentation benchmarks (Gong et al., 2017; He et al., 2022; Wei et al., 2023), SCPS improves target intersection-over-union (IoU) and Strict Target Success (STS) while reducing Parent Leakage (PL) relative to the original selectors for all seven evaluated retrieval methods. PL measures the fraction of the parent-only region incorrectly predicted as foreground, while STS requires accurate target segmentation and low leakage. Averaged across the seven methods, SCPS increases target IoU by 0.054, reduces PL by 22.7% relative to the corresponding baselines, and increases STS by 6.37 percentage points.

Our contributions are as follows:

- We identify background color mapping deviation and part-to-parent expansion in visual in-context segmentation. We propose a mechanistic account linking familiar semantic re-

gions, role and color assignments guided by demonstrations, and learned semantic and color preferences.

- We show that semantic regions and the target roles specified by demonstrations can remain readable in late features even when predictions violate the demonstration. Controlled changes to colors and requested targets, together with diagnostic feature interventions, connect these internal distinctions to output behavior.
- We introduce Semantic Contrast Prompt Selection (SCPS), which combines semantic contrast encoding with target quality calibration to jointly select demonstrations and their binary or semantic contrast encodings. Aggregated across three benchmarks, SCPS improves target IoU and STS while reducing PL for each of seven retrieval methods.

2 Related Work

Visual In-Context Learning. VICL encompasses frameworks supporting heterogeneous tasks through image inpainting (Bar et al., 2022), unified image prediction (Wang et al., 2023a), autoregressive modeling (Bai et al., 2024; Hao et al., 2024), and fusion of multiple prompts (Liao et al., 2026), alongside approaches specializing in low-level vision (Liu et al., 2024b; Chen et al., 2024; Rajagopalan & Patel, 2025) or segmentation. Within segmentation, SegGPT randomizes label colors to encourage reliance on demonstrations (Wang et al., 2023b), while SINE predicts multiple masks for different segmentation tasks to address task ambiguity (Liu et al., 2024a).[1] We focus on SegGPT, a representative model trained on diverse segmentation data, including semantic, instance, and part segmentation (Wang et al., 2023b), and integrated into automatic annotation tools such as Autodistill (Autodistill Team, 2023). Its broad training coverage and randomized label colors make it a relevant testbed for studying demonstration following, while its RGB outputs enable joint analysis of region selection and color mapping.

Demonstration Selection. Methods for selecting demonstrations for each query include retrieval based on similarity in frozen representations, such as CLIP and DINOv2 features (Radford et al., 2021; Oquab et al., 2023), exemplified by UnsupPR (Zhang et al., 2023), and learned retrievers such as SupPR (Zhang et al., 2023), LaPR (Luo et al., 2026), Partial2Global, and RH-Partial2Global (Xu et al., 2024; Wu et al., 2025). SCPS augments existing demonstration selectors by jointly selecting demonstrations and their binary or semantic contrast encodings, combining existing selection scores with target quality estimates computed on development data.

3 Problem Setup

3.1 Segmentation Demonstrations and Color Encodings

A demonstration pairs an image x_d with an RGB segmentation map y_d. Given a query image x_q, a frozen model predicts $\hat{y}_q = F_\theta(x_q; x_d, y_d)$. For the two-color demonstrations used in our analysis, the requested target receives foreground color c_f, and all other pixels receive background color c_b. The query prediction is expected to follow the same mapping. The standard binary demonstrations use white foreground and black background. Regions sharing the background label can include non-target objects and the remaining scene. We examine them separately. For a requested part T within its parent object P, the parent-only region is $P \setminus T$, and the remaining scene lies outside P.

For these color comparisons, we assign each predicted pixel to the nearest demonstrated foreground or background color using Euclidean distance in RGB space. When the background color is non-black, we also include black as a candidate color. Color adherence is the fraction of pixels in an annotated query region assigned to its demonstrated color.

3.2 Controlled Experiments and Feature Probes

We keep the demonstration image, target mask, and query fixed while changing the foreground or background color in the demonstration. We also compare requesting different objects while

[1] SINE reports part segmentation results for a variant trained with additional part annotations, but the corresponding checkpoint was not available in its official repository at the time of our experiments.

preserving each object's assigned color. For example, requesting the person assigns gray to the person and magenta to all other pixels; requesting the dog assigns magenta to the dog and gray to all other pixels.

We extract query features from SegGPT's final transformer block during the same forward pass as the RGB prediction. For the background-color experiment, we also extract features from the first transformer block to compare early and late representations. Linear probes are classifiers trained to distinguish annotated regions using these features. We train them on one set of queries and evaluate them on held-out queries using the area under the receiver operating characteristic curve (AUROC). They distinguish non-target semantic regions from the remaining scene, or requested parts from parent-only regions. In the color comparisons, we train a separate probe for each demonstration color setting while keeping the annotated region definitions unchanged. This lets us examine whether those regions remain distinguishable internally as their output colors change. Appendix A.1 provides the detailed analysis protocols and statistical procedures.

4 Understanding Deviations from Demonstration Following

We first examine how demonstration colors and requested targets affect background coloring. We then ask whether the affected regions remain distinguishable internally and connect these findings to part-to-parent expansion.

4.1 Background Color Mapping Deviation

Figure 1 shows that regions sharing the background label can receive different output colors. To examine this behavior across queries, we fix one demonstration marking the person as the target and evaluate 15 held-out COCO query images containing a person and a dog (Lin et al., 2014) under eight background colors, keeping the foreground white. As shown in Figure 2, purple produces the highest mean adherence to the demonstrated background color among dog pixels, at 6.3%. Some pixels in the remaining scene, which also contains other semantic regions, receive the demonstrated background color. Most pixels in both regions remain black on average. Thus, the dog and the remaining scene respond differently to the demonstrated background color despite sharing the same background label.

We next examine how changing the requested object affects coloring while keeping each object's assigned color unchanged. In one fixed demonstration and query pair containing a person and a dog, we assign gray to the person and magenta to the dog. The dog receives magenta on 0.0% of its pixels as a non-target but 90.0% as the target, while the person remains predominantly gray in both cases. We extend this comparison to 30 queries, three demonstrations, and four fixed color pairs. After averaging across demonstrations and color pairs within each query, the mean increase in adherence is 21.1 percentage points for the dog, whereas the person shows a mean decrease of 2.5 percentage points. Appendix A.2.3 provides the protocol and confidence intervals.

Thus, requesting a region can change whether it receives its assigned color, but the direction and magnitude differ between the person and dog. These observations suggest that background coloring depends jointly on semantic identity, demonstrated role, and color encoding.

4.2 Semantic Distinctions Behind Color Changes

We next ask whether a region that remains black also becomes indistinguishable from the scene in the model's features. For each background color, we train separate linear probes on early features from Block 0 and late features from Block 23, using the same 15 training queries and 15 held-out queries. Figure 2 shows that early features already distinguish the dog from the scene, while late features yield higher AUROC across all eight colors, with mean values ranging from 0.974 to 0.981. Together, the two panels show that strong linear separability in late features persists despite variation in adherence to the demonstrated background color.

The fixed person and dog example provides complementary evidence. When the non-target dog receives no magenta pixels, a probe trained on other queries still distinguishes it from the scene with AUROC 0.997. Additional experiments examine how demonstration content affects background

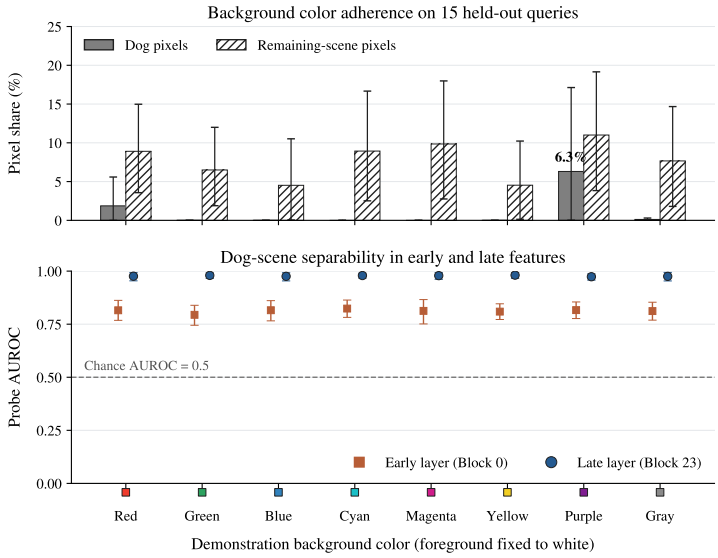


Figure 2: **Output colors and feature probes on the same 15 held-out queries.** Top: mean per-query fractions of dog and remaining-scene pixels receiving the demonstrated background color, with the foreground fixed to white. Bottom: dog-versus-scene probe AUROC for early layer and late layer features. Swatches show demonstration background colors; the dashed line marks chance AUROC (0.5). Error bars indicate 95% bootstrap confidence intervals over queries.

coloring and whether features distinguish target regions from non-target objects and the scene. These results are provided in Appendix A.2.

These results suggest that background coloring operates selectively over familiar semantic regions rather than uniformly over all non-target pixels. Together with the changes observed when different objects are requested, they support an account involving three learned behaviors: recognizing familiar semantic regions, interpreting demonstrations as cues for assigning target or non-target roles and their corresponding colors, and modulating these assignments through learned semantic and color preferences.

4.3 Part-to-Parent Expansion

Our findings on selective background coloring also help explain part-to-parent expansion. We first examine whether the requested part remains distinguishable from the parent-only region when both receive foreground predictions. Across three part segmentation benchmarks (Gong et al., 2017; He et al., 2022; Wei et al., 2023), we analyze 1,085 expansion cases from 38 tasks using binary demonstrations selected by a fixed retrieval procedure. We identify these cases by target recall ≥ 0.5 and PL ≥ 0.25, where recall measures the fraction of target pixels predicted as foreground and leakage measures the corresponding fraction of parent-only pixels. Linear probes trained separately for each task on other queries distinguish the requested part from the parent-only region with AUROC averaging 0.926 across expansion cases, and 94.6% of cases reach AUROC ≥ 0.75.

We further examine whether late features encode which part the demonstration requests. Holding the query and demonstration images fixed, we switch the requested part and train a shared linear probe across part identities and request conditions, with cross-fitting grouped by source query image. Across 13 request pairs and 200 source query images, a region's probe-predicted target probability increases by 0.184 when requested rather than unrequested (95% confidence interval: [0.168, 0.201]). Among expansion requests spanning 73 unique source query images, the mean increase is 0.302 (95% CI: $[0.271, 0.332]$). Thus, information about the demonstrated target role remains readable in late features even when the final prediction expands to the parent. Diagnostic interventions that use ground-truth masks to suppress features in the parent-only region reduce expansion and improve target IoU, whereas matched random controls yield no comparable improvement.

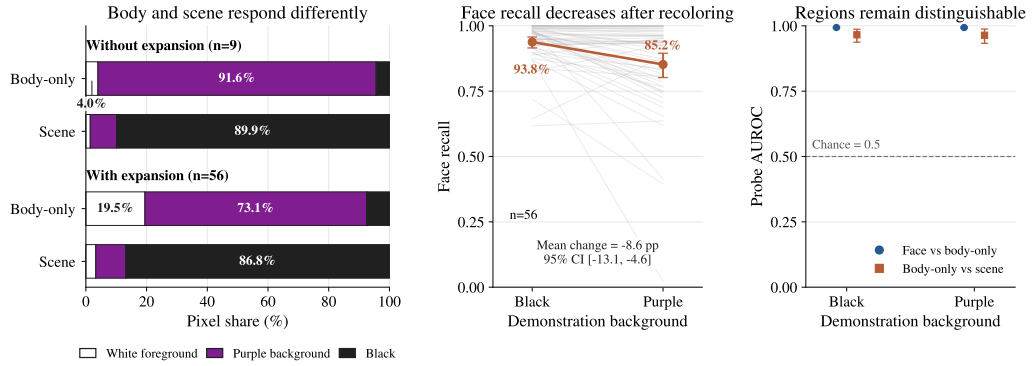


Figure 3: **Effects of changing all non-face demonstration pixels from black to purple.** Left: mean per-query output-color shares, grouped by the original binary predictions. Middle: paired face recall for the 56 expansion cases. Gray lines represent individual queries, and the colored line connects means. Right: face–body-only and body-only–scene probe AUROC under both backgrounds, with chance marked at 0.5. Error bars show 95% bootstrap confidence intervals over queries.

To investigate how background coloring affects expansion, we use 100 COCO query images with face regions defined from facial keypoints. Here, the body-only region is the person excluding the face. We keep the demonstrated face white and change every demonstration pixel outside the face from black to purple, assigning the same background color to the body-only region and the remaining scene. We group queries using only their original binary predictions. There are 56 expansion cases satisfying the criteria above and nine cases without expansion, defined by target recall ≥ 0.5 and PL ≤ 0.1. The remaining 35 have intermediate leakage or low recall.

Figure 3 compares output colors, face recall, and internal distinctions under black and purple demonstration backgrounds. After changing the background to purple, body-only pixels follow the demonstrated background color more strongly in cases without expansion than in expansion cases (91.6% versus 73.1%). Expansion cases retain more white body pixels (19.5% versus 4.0%). The scene remains predominantly black, with black pixel shares of 89.9% in cases without expansion and 86.8% in expansion cases. The demonstrated background color therefore appears selectively on the body rather than uniformly across all regions assigned that color.

This change also affects the requested target. In the expansion group, face recall decreases from 0.938 to 0.852, a reduction of 8.6 percentage points. Nevertheless, the regions remain distinguishable internally. Mean AUROC for distinguishing the face from the body-only region remains 0.993–0.994 under black and purple demonstration backgrounds, while that for distinguishing the body-only region from the scene remains 0.964–0.966.

The part segmentation results provide further support for this account. The requested part can remain internally distinguishable from the parent-only region even when both are predicted as foreground, and changing the demonstration background color can alter their output colors. The selective and incomplete recoloring of the body is consistent with learned semantic and color preferences influencing whether those assignments are expressed. These findings motivate explicitly marking the parent-only region in the demonstration while accounting for encoding effects on target segmentation. Appendix A.3 provides the detailed protocols, intervention controls, and additional results.

5 Semantic Contrast Prompt Selection (SCPS)

Our experiments show that requested targets generally follow foreground colors once predicted as foreground, with additional palette comparisons provided in Appendix A.2.2. Parent-only regions respond selectively to background colors, but changing the encoding can also reduce target recall. SCPS therefore combines explicit parent-only coloring with target quality calibration to select demonstrations and their encodings.

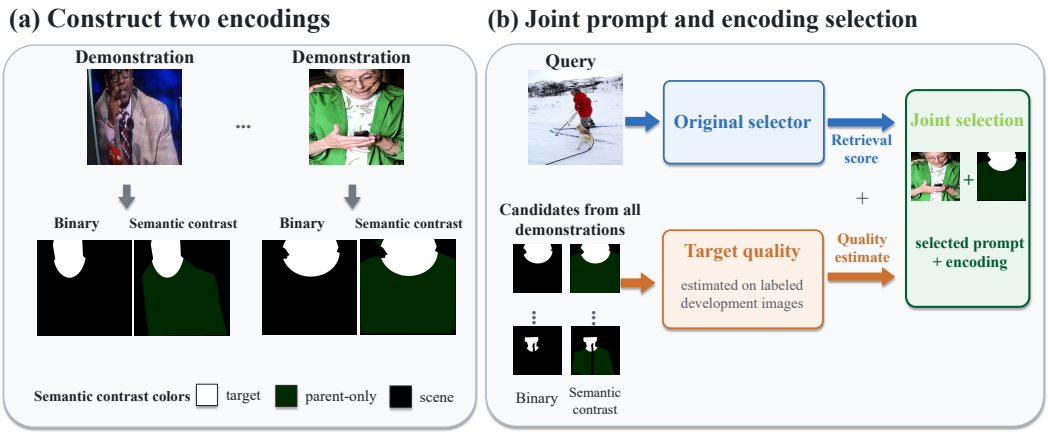


Figure 4: **Overview of Semantic Contrast Prompt Selection (SCPS).** (a) Each demonstration has binary and semantic contrast encodings. Semantic contrast assigns white to the target, dark green to the parent-only region, and black to the remaining scene. (b) SCPS combines normalized retrieval scores with weighted target quality estimates from labeled development images to jointly select a demonstration and its encoding.

5.1 Encoding the Target, Parent-Only Region, and Scene

Binary demonstrations assign the same background color to the parent-only region and the remaining scene. For a demonstrated target T_d within a parent object P_d, semantic contrast assigns distinct colors to T_d, $P_d \setminus T_d$, and the remaining scene, explicitly marking the parent-only region for exclusion from the target. Assigning a non-black background color to all non-target pixels may also cue the coloring of other semantic regions. We therefore color only the parent-only region and keep the remaining scene black. Constructing this encoding requires target and parent annotations for the demonstration.

Figure 4(a) illustrates the two encodings for each candidate demonstration. Both preserve the demonstration image and requested target extent. We use white for the target, dark green with normalized RGB value $(0, 0.16, 0)$ for the parent-only region, and black for the remaining scene. We select the parent-only color on the development data based on segmentation performance and fix it throughout evaluation. Only pixels assigned to white form the target mask.

5.2 Selecting Demonstrations and Encodings

As illustrated in Figure 4(b), SCPS combines retrieval scores with development estimates of target quality to select a demonstration and its encoding. We retain both binary and semantic contrast candidates. Each demonstration's original retrieval score is shared between its two encodings, while target quality is estimated separately for each version.

For each task t, we evaluate every candidate demonstration p under both encodings on the same labeled development images. We store the mean target IoU as $\widehat{Q}_{\text{target}}(t, p, m)$, where m denotes the encoding.

Given a query x_q, let $s_{\text{base}}(x_q, p)$ be the original selector's score for demonstration p, with larger values indicating stronger preference. We apply min–max normalization across the candidate demonstrations for each query to obtain scores in $[0, 1]$, assigning zero when all scores are equal. SCPS scores each demonstration and encoding by

$$s_{\text{SCPS}}(x_q, p, m) = \widetilde{s}_{\text{base}}(x_q, p) + \alpha\, \widehat{Q}_{\text{target}}(t, p, m), \tag{1}$$

where $\widetilde{s}_{\text{base}}$ denotes the normalized retrieval score and α controls the contribution of estimated target quality. We select α on development data separately for each retrieval method. We refer to this adjustment of the retrieval score as target quality calibration.

Table 1: Original binary selectors and SCPS. Top: averages over 47 equally weighted tasks, with a mean across methods. Bottom: averages over tasks within each benchmark and all methods. IoU denotes target IoU, and all metrics are fractions.

Method / benchmark	Original			SCPS		
	IoU ↑	PL ↓	STS ↑	IoU ↑	PL ↓	STS ↑
Appearance	.2133	.1522	.1451	.3037	.1153	.2466
CLIP (Radford et al., 2021)	.2331	.1596	.1663	.3059	.1110	.2514
DINOv2 (Oquab et al., 2023)	.2540	.1512	.1868	.3140	.1112	.2579
UnsupPR (Zhang et al., 2023)	.2457	.1486	.1819	.3104	.1093	.2559
SupPR (Zhang et al., 2023)	.2309	.1512	.1660	.2590	.1271	.2055
LaPR (Luo et al., 2026)	.2295	.1458	.1702	.2745	.1250	.2203
RH-Partial2Global† (Wu et al., 2025)	.2821	.1495	.2174	.2993	.1185	.2419
Mean	.2412	.1512	.1762	**.2952**	**.1168**	**.2399**
LIP	.4026	.0113	.4418	**.5199**	**.0095**	**.5880**
PartImageNet-OOD	.2229	.2500	.1312	**.2471**	**.1826**	**.1673**
Pascal-Part-116	.2294	.0823	.1709	**.3004**	**.0735**	**.2459**

†Our implementation follows the reliability filtering strategy of RH-Partial2Global before ranking demonstrations by query similarity.

SCPS selects the pair (p^*, m^*) with the highest score. For a fixed task and demonstration, the quality estimates determine the preferred encoding, while retrieval and quality jointly determine which demonstration is selected for the query. Using the stored estimates, SCPS requires one segmentation forward pass per new query and no test-query annotations.

6 Evaluating SCPS

6.1 Experimental Setup

We evaluate SegGPT at 448×448 resolution on LIP (Gong et al., 2017), PartImageNet-OOD (He et al., 2022), and Pascal-Part-116 (Wei et al., 2023): 47 tasks and 3,313 task–query pairs from 2,336 source images. Each task provides eight demonstrations with target and parent annotations. For the seven retrieval methods in Table 1, we retain the original retrieval scores and compare binary selection with SCPS using both encodings of the same demonstrations. Appearance ranks demonstrations by cosine similarity between globally averaged, layer-normalized SegGPT patch embeddings, without training an additional retrieval model.

On PACO-LVIS development data (Ramanathan et al., 2023), we select the parent-only color by mean STS, breaking ties by higher target IoU and then lower PL, and select α for each retrieval method by target IoU. Both remain fixed at test time. For each task, we estimate target quality for every demonstration and encoding using three labeled development splits from its benchmark, each containing ten images, except Bicycle Seat, which has seven. We evaluate the same test queries under each split. Test source images are disjoint from development images and demonstrations.

RGB outputs are clipped to $[0, 1]$ and assigned to the nearest color center by squared Euclidean distance, using white and black for binary encoding and white, dark green $(0, 0.16, 0)$, and black for semantic contrast. White pixels form the target mask, without spatial post-processing. We report target IoU, Parent Leakage (PL), and Strict Target Success (STS). PL is the fraction of parent-only pixels predicted as target. STS requires target IoU ≥ 0.5, target precision ≥ 0.6, and PL ≤ 0.1, where precision measures the fraction of predicted target pixels belonging to the target. Benchmark results average development splits and queries within each task, then weight tasks equally.

6.2 Main Results

Table 1 shows gains in all three metrics for every retrieval method after aggregating benchmarks, and for every benchmark after averaging methods. Mean target IoU increases by 0.054, PL decreases

Table 2: Encoding and calibration ablations, averaged across seven retrieval methods. Calibration adds estimated target quality to retrieval scores. Bold marks the best result per metric.

Candidate encodings	Calibration	Target IoU ↑	PL ↓	STS ↑
Binary	No	.2412	.1512	.1762
Semantic contrast	No	.2214	**.0716**	.1754
Binary	Yes	.2875	.1451	.2243
Semantic contrast	Yes	.2837	.0815	.2350
Both (SCPS)	Yes	**.2952**	.1168	**.2399**

by 22.7% relative to the baselines, and STS increases by 6.37 percentage points. On LIP, baseline leakage is low, and gains mainly concern target IoU and STS.

6.3 Ablation Studies

Table 2 uses matched demonstration pools and evaluation queries. Without calibration, semantic contrast reduces PL from 0.1512 to 0.0716 but lowers target IoU, motivating quality-guided selection. Averaged across the seven retrieval methods, calibration improves target IoU and STS under both encodings, and retaining both improves these metrics over either calibrated encoding alone. Compared with calibrated binary selection, SCPS reduces PL from 0.1451 to 0.1168 and raises STS from 0.2243 to 0.2399, with higher target IoU. Paired task-bootstrap intervals support all three average gains, showing the benefits of combining semantic contrast with quality-guided selection.

SCPS retains its gains under a shared output decoder. When both candidate sets retain binary encodings and use matched calibration, semantic contrast yields lower mean PL and higher mean STS than uniform background recoloring. Appendix A.4 provides the settings, detailed results, and controls.

7 Discussion and Limitations

Our findings motivate designing visual in-context segmentation models around the learned behaviors that shape demonstration following. Beyond varying output colors, training could vary whether the same semantic region is requested or excluded, and whether the target is a part or its parent. Studying these changes together may help identify and reduce dependencies among semantic identity, region role, and color that interfere with the demonstrated mapping. Evaluating internal representations alongside final predictions can reveal dependencies that output accuracy alone may obscure.

Our analysis primarily concerns SegGPT, whose RGB outputs expose both region selection and color mapping. Appendix A.5 provides complementary evidence from SINE's features and mask predictions, although SINE's restricted output encodings preclude the same color comparisons. Evaluating training strategies informed by these findings remains future work.

8 Conclusion

We identify background color mapping deviation and part-to-parent expansion in visual in-context segmentation. Across controlled changes to demonstrations, semantic regions can remain internally distinguishable while receiving colors that depart from the demonstrated mapping. These findings support an account in which demonstrations guide role and color assignments over familiar semantic regions, influenced by learned semantic and color preferences. Building on this account, SCPS combines semantic contrast encoding with target quality calibration to select a demonstration and its binary or semantic contrast encoding. Aggregated across three benchmarks, SCPS improves target IoU and STS while reducing PL for all seven evaluated retrieval methods. Our results motivate studying internal representations together with predictions when designing models that follow visual demonstrations more reliably.

AI USE STATEMENT

The authors conducted the initial experiments and observations, iteratively formulated, tested, and refined the hypotheses, and designed and performed the main experiments supporting the proposed mechanistic account. Generative AI tools helped suggest supplementary validation experiments, write experimental and analysis scripts, and discuss interpretations of the results. We also used these tools to polish the writing and identify relevant literature, including recent related work we had missed. The authors reviewed the AI-assisted contributions and take responsibility for the final content, results, and claims.

REFERENCES

Autodistill Team. Seggpt base models documentation, 2023. URL https://docs.autodistill.com/base_models/seggpt/. Accessed: 2025-08-18.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L. Yuille, Trevor Darrell, Jitendra Malik, and Alexei A. Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22861–22872, June 2024.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25005–25017. Curran Associates, Inc., 2022. doi: 10.52202/068431-1813. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf.

Xiangyu Chen, Yihao Liu, Yuandong Pu, Wenlong Zhang, Jiantao Zhou, Yu Qiao, and Chao Dong. Learning a low-level vision generalist via visual task prompt. In *ACM Multimedia 2024*, 2024.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Zhiwei Hao, Jianyuan Guo, Chengcheng Wang, Yehui Tang, Han Wu, Han Hu, Kai Han, and Chang Xu. Data-efficient large vision models through sequential autoregression. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17572–17596. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hao24b.html.

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 Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 128–145, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20074-8.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 257–273, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72775-7.

Sunil Khatri, Steven Landgraf, Markus Ulrich, and Simon Reiß. Beyond model size: Probing the gaps in visual in-context learning by training a tiny model, 2026. URL https://arxiv.org/abs/2606.10905.

Wenwen Liao, Jianbo Yu, Yuansong Wang, Qingchao Jiang, and Xiaofeng Yang. Enhancing visual in-context learning by multi-faceted fusion, 2026. URL https://arxiv.org/abs/2601.10107.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 25095–25119. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0791. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2cc0b08447bf9668db268e6c86364a6e-Paper-Conference.pdf.

Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30873–30891. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/liu24k.html.

Tianci Luo, Haohao Pan, Jinpeng Wang, Niu Lian, Xinrui Chen, Bin Chen, Shu-Tao Xia, and Chun Yuan. Love me, love my label: Rethinking the role of labels in prompt retrieval for visual in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38774–38783, June 2026.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527/.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Sudarshan Rajagopalan and Vishal M. Patel. Awracle: All-weather image restoration using visual in-context learning. In *AAAI Conference on Artificial Intelligence*, 2025. URL https://api.semanticscholar.org/CorpusID:272367682.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7141–7151, June 2023.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6830–6839, June 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, October 2023b.

Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. Ov-parts: Towards open-vocabulary part segmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70094–70114. Curran Associates, Inc., 2023. doi: 10.52202/075280-3071. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dde53059fdb0f45e1e9ad9c66997d662-Paper-Datasets_and_Benchmarks.pdf.

Wenxiao Wu, Jing-Hao Xue, Chengming Xu, Chen Liu, Xinwei Sun, Changxin Gao, Nong Sang, and Yanwei Fu. Towards reliable and holistic visual in-context learning prompt selection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=EAbuXJPKQz.

Chengming Xu, Chen Liu, Yikai Wang, Yuan Yao, and Yanwei Fu. Towards global optimal visual in-context learning prompt selection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 74945–74965. Curran Associates, Inc., 2024. doi: 10.52202/079017-2384. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8900e600e4f55be0a10cee4a14d7c3de-Paper-Conference.pdf.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 17773–17794. Curran Associates, Inc., 2023. doi: 10.52202/075280-0780. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/398ae57ed4fda79d0781c65c926d667b-Paper-Conference.pdf.

A Appendix

A.1 Additional analysis details

We provide analysis protocols, additional evidence on background coloring and part-to-parent expansion, SCPS settings and controls, and a complementary analysis of SINE.

A.1.1 Color comparisons and region definitions

The background experiment uses one demonstration requesting the person and 30 COCO queries. The foreground stays white while the background takes one of eight colors: red, green, blue, cyan, magenta, yellow, purple, or gray. Their normalized RGB values are $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, $(0, 1, 1)$, $(1, 0, 1)$, $(1, 1, 0)$, $(0.5, 0, 0.5)$, and $(0.5, 0.5, 0.5)$, respectively. The demonstration image, target mask, and query remain fixed across colors.

Color adherence is the fraction of pixels in an annotated query region assigned to its demonstrated color. We assign predictions to the nearest demonstrated foreground color, background color, or black by RGB distance. In the person, dog, and horse analyses, the remaining scene excludes those annotated objects but may contain other semantic regions, such as sky or vegetation. In part segmentation, the parent-only region is the parent object excluding the requested part, and the remaining scene lies outside the parent.

A.1.2 Feature probes

Unless otherwise specified, we extract SegGPT query features from the final transformer block, Block 23, after the model's final normalization, during the same forward pass as the RGB prediction.

For the early-layer comparison in Figure 2, we additionally extract query features from the output of the first transformer block, Block 0. A token's region occupancy is the fraction of its image patch covered by that region.

For background coloring, dog and scene tokens require occupancies of at least 0.25 and 0.80, respectively. We train a separate linear probe for each layer and background color using the same 15 training queries and 15 evaluation queries. We compute AUROC separately for each evaluation query and report the mean across queries.

For part segmentation, target and parent-only tokens require occupancies of at least 0.25 and 0.50, respectively; tokens meeting both criteria are excluded. We normalize features to unit L2 norm and fit linear logistic probes separately for each task. Queries are divided into three fixed folds, and each query is evaluated with a probe trained on the other two folds. Probes use AdamW for 300 steps with learning rate 0.02 and weight decay 10^{-4}, using at most 12,000 training tokens per class and 1,024 tokens per class in each batch.

For face segmentation, occupancy thresholds are 0.25 for the face, 0.50 for the body-only region, and 0.80 for the scene. Tokens meeting multiple thresholds are excluded. We use five folds grouped by source image and evaluate each image with probes trained on other images. Separate probes are fitted under black and purple backgrounds to test whether the annotated distinctions remain readable under each encoding.

The shared part-role probe uses the paired-request protocol described in Appendix A.3.5.

A.1.3 Averaging and confidence intervals

We compute color shares within each query before averaging across queries. Paired bootstrap comparisons resample query identities together across conditions; the comparison of requested objects uses 10,000 resamples. For SCPS, we average queries and development splits within each task, then resample task means. Feature intervention intervals resample tasks while retaining all sampled cases within each task. For the shared part-role probe, we average repeated part-pair comparisons within each source query image before bootstrapping source query images. All confidence intervals are 95% intervals.

A.2 Additional evidence on background coloring

A.2.1 Background colors and feature separability

Across all 30 queries, purple gives the highest mean dog adherence at 17.3%. In six queries, more than half of the dog pixels receive purple; no query reaches this level under the other seven colors. Figure 2 in the main paper uses the 15 held-out queries evaluated by the probes, where purple adherence averages 6.3%. Table 3 shows that the dog remains distinguishable from the scene under every tested color.

Table 3: Dog-versus-scene probes using late features from Block 23, fitted separately for each demonstration background color. Mean AUROC and 95% bootstrap confidence intervals are computed over the same 15 held-out queries used in Figure 2.

Background	Mean AUROC	95% CI
Red	0.9760	$[0.9542, 0.9930]$
Green	0.9798	$[0.9657, 0.9917]$
Blue	0.9755	$[0.9541, 0.9924]$
Cyan	0.9792	$[0.9667, 0.9901]$
Magenta	0.9787	$[0.9614, 0.9929]$
Yellow	0.9810	$[0.9658, 0.9932]$
Purple	0.9736	$[0.9577, 0.9870]$
Gray	0.9753	$[0.9537, 0.9939]$

Unit RGB-direction projection. We additionally compare continuous RGB predictions under purple $(0.5, 0, 0.5)$ and magenta $(1, 0, 1)$. For query i and demonstration background color c, we

compute the unit RGB-direction projection

$$s_{ic} = \left(\frac{1}{|D_i|} \sum_{p \in D_i} \mathbf{y}_{ic}(p) \right)^{\top} \frac{\mathbf{b}_c}{\|\mathbf{b}_c\|_2}, \tag{2}$$

where D_i is the annotated dog region, $\mathbf{y}_{ic}(p)$ is the floating-point RGB output before saving the outputs as PNG images or applying nearest-color decoding, and $\mathbf{b}_c$ is the demonstration background color. Purple and magenta have the same unit RGB direction, $(1/\sqrt{2}, 0, 1/\sqrt{2})$, while differing in magnitude. Across the 30 queries, mean s_{ic} is 0.1175 under purple and 0.0100 under magenta. The mean paired purple-minus-magenta difference is 0.1075, with a query-bootstrap 95% confidence interval of $[0.0385, 0.1853]$. The color dependence is therefore present in continuous predictions and is not solely induced by nearest-color decoding.

A.2.2 Foreground colors and target recovery

We vary the foreground over red, green, blue, cyan, magenta, yellow, white, and gray under black or purple backgrounds, producing 480 outputs on the same 30 queries. To distinguish the requested color from other colors, we assign predictions to the nearest of the eight foreground colors, the demonstrated background color, and black.

Table 4 reports the fraction of annotated target pixels receiving the requested color and target IoU using decoding with the demonstrated colors and black. Gray exceeds red in color share by 7.1 percentage points under black backgrounds and 14.2 points under purple backgrounds, with paired intervals $[4.5, 10.4]$ and $[10.4, 18.9]$. Among target pixels predicted as foreground, agreement with the demonstrated foreground color ranges from 95.1% to 100%. Foreground predictions therefore generally follow the specified color, while the recovered target extent varies with the palette.

Table 4: Foreground color comparisons on 30 queries. Shares of the requested color use the expanded palette described in the text and are percentages of annotated target pixels. Target IoU uses decoding with the demonstrated colors and black.

	Black background		Purple background	
Foreground	Color share (%)	Target IoU	Color share (%)	Target IoU
Red	73.5	.663	66.4	.615
Green	76.6	.681	74.6	.672
Blue	78.0	.687	74.7	.670
Cyan	78.6	.696	76.9	.685
Magenta	79.4	.703	74.2	.665
Yellow	78.9	.700	76.4	.687
White	78.4	.700	75.2	.684
Gray	80.6	.701	80.6	.699

A.2.3 Changing the requested object across queries

We extend the person and dog comparison to 30 queries, three demonstrations, and four color pairs fixed before the experiment. Demonstration images are disjoint from the queries. The person and dog are assigned, respectively, gray and magenta, white and purple, cyan and red, or yellow and blue.

We request each object under every demonstration–query and color-pair combination, producing 720 outputs in total. The requested object receives its assigned color and all other pixels receive the other object's color. For example, requesting the person gives gray person and magenta background, while requesting the dog gives magenta dog and gray background. Each object's assigned RGB value stays fixed while its role and the scene's background color change. Predictions are decoded using the two demonstrated colors and black.

We average the 12 demonstration and color combinations within each query. Table 5 reports adherence under each role and the paired difference. Requesting the dog increases adherence to its assigned color, whereas requesting the person produces a smaller decrease.

The original fixed demonstration and query pair gives mean increases of 76.7 percentage points for the dog and 4.2 points for the person across 42 color pairs. The broader comparison confirms a positive mean effect for the dog while showing that the size and direction of the effect depend on the semantic region and image conditions.

Table 5: Color adherence across 30 queries. Adherence is the percentage of the annotated region receiving its specified color. Differences and their query bootstrap intervals are in percentage points.

Region	Requested	Not requested	Difference [95% confidence interval]
Person	73.0	75.5	-2.5 $[-3.9, -1.1]$
Dog	54.1	33.0	$+21.1$ $[15.2, 26.6]$

A.2.4 Information about target and non-target roles

A separate linear classifier predicts whether a token belongs to the target, a non-target object, or the scene. We train it on 80 queries with requests for both person and dog and evaluate it on 48 disjoint queries with a held-out demonstration. Balanced accuracy is 0.929. When the person becomes requested, its mean predicted target probability increases by 0.878; the corresponding increase for the dog is 0.820. Internal features thus contain information about the demonstrated role as well as semantic region identity.

A.2.5 Demonstration content and region coloring

We compare two natural demonstrations, one containing a horse and one without a horse, on the same 30 queries containing a person, dog, and horse. Both demonstrations request the person. The query horse receives the demonstrated background color on 89.9% of its pixels with the horse-containing demonstration and 0.2% with the other demonstration, a paired difference of 89.7 percentage points with interval $[84.8, 93.6]$.

Probes trained on 15 queries and evaluated on the other 15 distinguish the horse from the scene with mean AUROC 0.977 and 0.974 under the respective demonstrations. Transferring the first probe to the second condition gives AUROC 0.918; the reverse transfer gives 0.972. Output coloring changes while the horse remains internally distinguishable. Because the demonstrations also differ in other image content, this comparison measures the effect of changing the demonstration as a whole.

For the fixed person and dog example, we also separate the demonstrated dog from the scene. With gray person and magenta background, the query person is 94.9% gray and the dog receives no magenta pixels. Keeping the demonstrated person gray and dog magenta while changing the scene to black yields 94.6% gray person pixels, 89.2% magenta dog pixels, and 99.2% black scene pixels. Explicitly distinguishing the dog from the scene changes whether its assigned color appears in the prediction.

A.3 Additional evidence on part-to-parent expansion

A.3.1 Part separability across queries

The part analysis uses one selected binary prediction for each of the 3,313 task–query pairs. A ridge regressor trained on PACO-LVIS development data predicts target IoU from one patch-embedding appearance similarity and 12 Block-23 prompt–query compatibility features. The compatibility features use demonstration target and parent annotations to construct regional prototypes, without using query annotations. For each query, we select the highest-scoring demonstration from eight candidates and analyze its precomputed binary prediction. This selector fixes demonstration choices for the part analyses and is separate from the seven retrieval methods evaluated in Table 1.

Target recall ≥ 0.5 and PL ≥ 0.25 identify 1,085 expansion cases from 38 tasks. Probes trained on other queries achieve mean AUROC 0.926 across these cases, with 94.6% reaching AUROC ≥ 0.75. With equal task weights, mean AUROC is 0.916, and the mean fraction reaching AUROC ≥ 0.75 is 91.7%.

A.3.2 FACE MASKS AND PAIRED RECOLORING

Face masks are ellipses derived from five visible COCO facial keypoints and intersected with the annotated person mask. The body-only region is the person excluding the face. The demonstration uses image ID 191577 and person annotation ID 1723055. Its face stays white while all other pixels change from black to purple $(0.5, 0, 0.5)$.

Binary predictions define the groups: 56 expansion cases, nine cases without expansion, and 35 with intermediate leakage or low recall. Table 6 reports the recoloring results. Under purple backgrounds, body adherence is 18.6 percentage points higher in cases without expansion than in expansion cases, with interval $[11.6, 25.8]$. Face recall in the expansion group changes by -8.6 percentage points, with paired interval $[-13.1, -4.6]$.

Table 6: Recoloring all non-face demonstration pixels purple. Groups are defined by binary predictions. Color shares are percentages; face recall is a fraction.

Measure	Without expansion ($n = 9$)	With expansion ($n = 56$)
Body-only purple share	91.6	73.1
Body-only white share	4.0	19.5
Body-only black share	4.4	7.5
Scene purple share	8.7	9.9
Scene black share	89.9	86.8
Face recall, black background	.785	.938
Face recall, purple background	.834	.852

A.3.3 DIAGNOSTIC FEATURE INTERVENTIONS

We sample 200 of the 1,085 expansion cases in Appendix A.3.1, stratifying by task. The sample covers 38 tasks. All conditions keep the query, selected demonstration, and binary encoding fixed.

We modify query features entering the decoder embedding. Let T, R, and S denote token sets for the target, parent-only region, and remaining scene, with mean features μ_T, μ_R, and μ_S. Target enhancement and parent-only suppression update token feature z_i by

$$\begin{aligned} z_i' &= z_i + \lambda_T(\mu_T - \mu_R), \quad i \in T, \\ z_i' &= (1 - \lambda_R)z_i + \lambda_R \mu_S, \quad i \in R, \end{aligned}$$

where $\lambda_T = \lambda_R = 0.25$. The first adds one quarter of the difference between the target and parent-only mean features; the second moves each parent-only token one quarter of the way toward the scene mean. We evaluate them separately and together. Ground-truth masks define the token sets and the diagnostic updates.

Random direction controls use updates with matched norms at the original locations. The random location control applies both components to two disjoint sets sampled from the full query token grid, matching the target and parent-only token counts. It preserves the target update; the parent component follows $\mu_S - \mu_R$, scaled to the mean norm of the original parent-only updates. Sampled locations may overlap the annotated regions by chance.

Table 7 shows that parent-only suppression substantially reduces leakage and improves target IoU. Target enhancement alone changes little, and random controls yield no comparable improvement. Using 10,000 paired bootstrap resamples of the 38 tasks, the target IoU increase from parent-only suppression is 0.4008, with interval $[0.3710, 0.4297]$. The change in PL is -0.6556, with interval $[-0.7122, -0.5957]$. Target recall changes by -0.0560, with interval $[-0.0654, -0.0474]$. Parent-only features contribute to the erroneous foreground prediction, although suppressing them also removes some target pixels.

A.3.4 SEPARABILITY AND THE RESPONSE TO SEMANTIC CONTRAST

Among the 1,085 expansion cases, we compare binary and semantic contrast encodings of the same demonstration. PL decreases without reducing target IoU in 46.9% of cases with probe AUROC ≥ 0.75 and 21.2% of the other cases, using equal task weights. Across the 20 tasks containing both groups, the paired difference is 25.4 percentage points with interval $[9.9, 38.7]$. Greater part separability is associated with a more favorable response to semantic contrast.

Table 7: Feature interventions on 200 expansion cases. Values are averaged across cases. Ground-truth masks are used only for this diagnostic, not by SCPS at test time.

Condition	Target IoU	Target recall	PL
Unmodified prediction	.2054	.8553	.6916
Target enhancement	.2073	.8569	.6923
Parent-only suppression	.6063	.7993	.0360
Both interventions	.6062	.8040	.0372
Random target directions	.2052	.8537	.6916
Random parent-only directions	.2058	.8551	.6902
Both random controls	.2059	.8544	.6894
Random location control	.1939	.7166	.5811

A.3.5 Demonstration-conditioned part-role information

We examine whether late features encode the requested role of a part by pairing demonstrations that request different parts of the same parent. Within each pair, the query image and demonstration image remain fixed, while the demonstration target mask changes. Paired demonstrations must have parent-mask IoU of at least 0.90 and requested-part IoU of at most 0.10. The evaluation contains 13 request pairs, 305 paired query instances, and 200 unique source query images.

We fit one shared linear probe across part identities and both request conditions. Five-fold cross-fitting groups instances by dataset and source query image, keeping all request versions of an image in the same fold. Requested-part and alternate-part tokens require region occupancy of at least 0.25; mixed tokens are excluded. The probe achieves a mean AUROC of 0.7962 across queries and request conditions, with balanced accuracy of 0.6996.

For each region, we compare its probe-predicted target probability when requested and unrequested. We first average repeated part-pair comparisons within each source query image, then bootstrap source query images. The mean increase is 0.1841 (95% CI: $[0.1677, 0.2014]$). The increase is positive for 94.0% of source query images.

The expansion subset contains 83 pair-linked request comparisons, corresponding to 81 unique task–query requests from 73 unique source query images. Two task–query requests each participate in two different part pairs. We average repeated comparisons within each source query image before computing the mean change, the positive-change fraction, and the bootstrap confidence interval, using these 73 images as the statistical units. The mean increase is 0.3017 (95% CI: $[0.2707, 0.3318]$), with positive changes for 98.6% of source query images. These results show that information about the demonstrated target role remains readable across queries, including when the output expands to the parent.

A.4 SCPS settings and additional evaluation

A.4.1 Tasks and development data

Table 8 lists the four LIP tasks, 21 PartImageNet-OOD tasks, and 22 Pascal-Part-116 tasks. They contain 400, 1,915, and 998 task–query pairs, respectively, totaling 3,313 pairs from 2,336 source images.

Each task provides eight demonstrations with target and parent annotations. We select the palette and calibration weights on PACO-LVIS development data. For each evaluated task, target quality is estimated on three labeled development splits from its benchmark, with ten images per split except Bicycle Seat, which has seven. We evaluate the same test queries with each split's estimates. Test source images are disjoint from development and demonstration images.

A.4.2 Palette and selection weights

We select the parent-only color by mean STS across PACO-LVIS development tasks, breaking ties by higher target IoU and then lower PL. The resulting encoding uses white for the target, dark green $(0, 0.16, 0)$ for the parent-only region, and black for the scene. We select α from

Table 8: Tasks used in the final benchmark. Names retain the identifiers in the evaluation records.

Benchmark	Tasks
LIP	arms person; face person; legs person; shoes person
PartImageNet-OOD	Bicycle Seat; Biped Foot; Biped Hand; Biped Head; Biped Tail; Bird Foot; Bird Head; Bird Tail; Bird Wing; Bottle Mouth; Car Tier; Fish Fin; Fish Head; Fish Tail; Quadruped Foot; Quadruped Head; Quadruped Tail; Reptile Foot; Reptile Head; Reptile Tail; Snake Head
Pascal-Part-116	aeroplane:stern; aeroplane:wing; bird:head; bus:window; car:window; cat:ear; cat:head; cat:leg; cat:neck; cat:paw; cat:tail; cow:head; dog:ear; dog:head; dog:leg; dog:muzzle; horse:head; person:hair; person:head; person:leg; person:lower arm; person:upper arm

$\{0, 0.25, 0.5, 1, 2, 4\}$ by development target IoU for each retrieval method. Selected values are 4 for Appearance, CLIP, DINOv2, and UnsupPR; 0.5 for SupPR; and 1 for LaPR and RH-Partial2Global. All settings remain fixed during test evaluation.

Both encodings of a demonstration share its normalized retrieval score but have separate target quality estimates. For a fixed task and demonstration, these estimates determine the preferred encoding. The retrieval score and quality estimate together determine which demonstration is selected for a query. The score includes no PL penalty.

Without target quality calibration, the binary and semantic contrast versions of a demonstration receive the same retrieval score. Consistently resolving these ties in favor of binary or semantic contrast therefore recovers the corresponding uncalibrated single-encoding baseline in Table 2 of the main paper.

A.4.3 Output decoding

We clip predictions to $[0, 1]$ and assign pixels to color centers by squared Euclidean RGB distance. Binary decoding uses white and black, assigning exact ties to background. Semantic contrast decoding uses white, dark green $(0, 0.16, 0)$, and black, with exact ties resolved in that order. Only pixels assigned to white form the target mask. We apply no spatial postprocessing.

A.4.4 Using the same decoder for both encodings

We decode the same RGB predictions with either white and black or white, dark green, and black. The model, demonstrations, quality estimates, selection weights, and selected pairs remain fixed. This control covers all 3,313 task–query pairs, seven retrieval methods, and three development splits, using floating-point predictions before image quantization.

Table 9 shows that gains persist under either shared decoder. Under white/black decoding, SCPS increases target IoU by 0.0544 with interval $[0.0393, 0.0703]$, reduces PL by 0.0341 with interval $[0.0134, 0.0572]$, and increases STS by 6.42 percentage points with interval $[4.54, 8.41]$.

Changing only the decoder changes target membership for 0.154% of annotated target pixels and 0.062% of parent-only pixels under original selection, averaged equally across tasks. The corresponding SCPS shares are 0.187% and 0.061%. The gains are therefore not explained by adding the green decoding center.

Table 9: Original selection and SCPS under native or shared decoders. Native decoding uses two centers for binary prompts and three for semantic contrast prompts. All metrics are fractions; tasks are weighted equally.

Decoder	Selection	Target IoU	PL	STS
Native	Original	.2412	.1512	.1762
Native	SCPS	.2952	.1168	.2399
White / black	Original	.2412	.1512	.1762
White / black	SCPS	.2956	.1170	.2404
White / green / black	Original	.2407	.1505	.1761
White / green / black	SCPS	.2949	.1164	.2400

A.4.5 WHERE THE SEPARATE COLOR IS ASSIGNED

Table 10 compares coloring the demonstrated parent-only region with coloring a translated region of equal area outside the parent. We start from the 200 intervention cases in Appendix A.3.3. Six do not permit an exactly area-matched translated region and are excluded, leaving 194 cases across 37 tasks. The query, demonstration image, target, and palette stay fixed. This comparison changes the demonstration encoding without running SCPS selection.

Coloring the parent-only region reduces PL more than coloring the translated region, with paired difference -0.3095 and bootstrap interval over tasks $[-0.3671, -0.2519]$. It also reduces target recall. The effect depends on where the color is assigned but can extend to the requested target, supporting selection that considers target quality.

Table 10: Location of the colored region, averaged equally across 37 tasks comprising 194 cases with area-matched control regions. All metrics are fractions.

Encoding	Target IoU	Precision	Recall	PL
Binary	.2140	.2261	.8562	.6681
Correct parent-only region	.1896	.2853	.4728	.3061
Translated control region	.2085	.2266	.7827	.6157

A.4.6 COMPARISON WITH CALIBRATED BINARY SELECTION

We compare SCPS with binary selection using the same target quality calibration. After averaging queries and development splits within each task, we compute paired differences and resample the 47 tasks. Table 11 reports gains from adding semantic contrast candidates. Averaged across methods, target IoU increases by 0.0078, PL decreases by 0.0283, and STS increases by 1.56 percentage points. All three mean intervals exclude zero; individual target IoU intervals for UnsupPR, SupPR, and LaPR include zero.

Table 11: SCPS minus calibrated binary selection. Each entry gives a mean paired task difference and its 95% bootstrap interval over tasks. IoU and PL differences are absolute fractions; STS differences are percentage points. The mean row averages methods within each task before resampling tasks.

Retrieval method	Δ Target IoU	Δ PL	Δ STS (pp)
Appearance	$+0.0104$	-0.0289	$+1.69$
	$[0.0032, 0.0185]$	$[-0.0461, -0.0142]$	$[0.49, 3.18]$
CLIP	$+0.0094$	-0.0321	$+1.89$
	$[0.0012, 0.0175]$	$[-0.0515, -0.0148]$	$[0.90, 3.06]$
DINOv2	$+0.0086$	-0.0249	$+1.50$
	$[0.0024, 0.0154]$	$[-0.0400, -0.0115]$	$[0.67, 2.42]$
UnsupPR	$+0.0061$	-0.0345	$+1.28$
	$[-0.0016, 0.0145]$	$[-0.0534, -0.0178]$	$[0.32, 2.45]$
SupPR	$+0.0053$	-0.0244	$+1.61$
	$[-0.0018, 0.0126]$	$[-0.0394, -0.0115]$	$[0.67, 2.71]$
LaPR	$+0.0053$	-0.0260	$+1.28$
	$[-0.0005, 0.0117]$	$[-0.0398, -0.0139]$	$[0.52, 2.15]$
RH-Partial2Global	$+0.0092$	-0.0273	$+1.70$
	$[0.0012, 0.0177]$	$[-0.0485, -0.0079]$	$[0.73, 2.97]$
Mean	$+0.0078$	-0.0283	$+1.56$
	$[0.0015, 0.0146]$	$[-0.0436, -0.0146]$	$[0.74, 2.58]$

A.4.7 RESULTS FOR EACH BENCHMARK AND RETRIEVAL METHOD

Table 12 expands the main benchmark averages. Target IoU and STS improve for every listed method on each benchmark. PL decreases for most combinations, with small increases in several settings where baseline leakage is already low. Results average development splits and weight tasks equally within each benchmark.

Table 12: Results for each benchmark for every retrieval method, averaged equally across tasks within each benchmark. IoU denotes target IoU. All metrics are fractions.

	Original binary			SCPS		
Method	IoU	PL	STS	IoU	PL	STS
LIP						
Appearance	.3842	.0152	.4125	.5476	.0076	.6292
CLIP	.3594	.0160	.3850	.5321	.0117	.5892
DINOv2	.4072	.0077	.4500	.5501	.0091	.6242
UnsupPR	.4165	.0123	.4575	.5458	.0071	.6208
SupPR	.4090	.0100	.4550	.4822	.0094	.5450
LaPR	.3883	.0075	.4350	.4847	.0094	.5508
RH-Partial2Global	.4534	.0102	.4975	.4965	.0119	.5567
PartImageNet-OOD						
Appearance	.2117	.2405	.1259	.2530	.1845	.1729
CLIP	.2160	.2704	.1195	.2522	.1715	.1748
DINOv2	.2232	.2560	.1295	.2574	.1743	.1744
UnsupPR	.2295	.2441	.1382	.2588	.1707	.1794
SupPR	.2176	.2549	.1285	.2240	.2042	.1467
LaPR	.2121	.2427	.1220	.2329	.2018	.1471
RH-Partial2Global	.2501	.2411	.1547	.2512	.1709	.1761
Pascal-Part-116						
Appearance	.1838	.0929	.1148	.3078	.0688	.2475
CLIP	.2265	.0799	.1711	.3161	.0713	.2631
DINOv2	.2556	.0773	.1936	.3251	.0694	.2710
UnsupPR	.2301	.0821	.1734	.3168	.0692	.2627
SupPR	.2112	.0779	.1492	.2518	.0749	.1998
LaPR	.2172	.0785	.1681	.2759	.0729	.2300
RH-Partial2Global	.2815	.0873	.2263	.3093	.0879	.2475

A.4.8 Comparison with uniform background recoloring

We compare semantic contrast with a uniform-background encoding that assigns white to the target and dark green $(0, 0.16, 0)$ to every non-target demonstration pixel. Both policies retain binary candidates and use the same demonstration pools, query-dependent selector scores, development splits, and calibration procedure. The uniform-background outputs are decoded using white, dark green, and black; using only white and dark green produces identical aggregate metrics.

Table 13 compares the candidate encoding sets. Relative to uniform recoloring, semantic contrast changes PL by -0.0291 (95% CI: $[-0.0432, -0.0160]$), STS by $+0.97$ percentage points (95% CI: $[0.21, 1.83]$), and target IoU by -0.0020 (95% CI: $[-0.0071, 0.0031]$). Explicitly distinguishing the parent-only region therefore improves average leakage control and strict target success beyond uniform background recoloring.

Table 13: Comparison of candidate encoding sets with target quality calibration. Results weight tasks equally and average across seven retrieval methods. All metrics are fractions.

Candidate encodings	Target IoU ↑	PL ↓	STS ↑
Binary	.2875	.1451	.2243
Binary + uniform background	.2973	.1458	.2303
Binary + semantic contrast	.2952	**.1168**	**.2399**
Binary + uniform background + semantic contrast	**.2976**	.1223	.2398

A.5 Semantic distinctions in SINE

SINE provides a complementary setting with native segmentation masks. With demonstrations requesting the person, we train a linear probe to distinguish dog from scene features on 80 queries and evaluate it on 48 disjoint queries. Each evaluation query uses three demonstrations, whose results are averaged within the query. We also measure this distinction using the continuous output score and thresholded mask.

Table 14: Dog versus scene AUROC in SINE under person requests, averaged across 48 held-out queries. The feature row uses a linear probe; output rows use the continuous score or binary mask directly.

Features or output	Mean query AUROC
Features after interaction with the demonstration	.988
Continuous output score	.874
Thresholded binary mask	.521

Table 14 shows that the dog remains distinguishable in internal features, while the binary output largely merges it with the scene. Both regions are background under a person request, so this is consistent with the requested segmentation. Internal distinctions can therefore persist even when output labels do not expose them. SINE's restricted output encodings do not support the color comparisons we conduct with SegGPT.