CDICS: Delving Into Fine-Grained Attribute for In-Context Segmentation via Compositional Prompts and Phased Decoupling

Supplementary Material

A. Dataset Reconstruction

The data utilized in this work originates from two existing part-level benchmarks: PACO [3] and PartImageNet [1]. While PACO provides valuable part attribute annotations, its color attributes are relatively coarse-grained. To enhance the granularity and reliability of part color labels, we introduce an algorithmic approach to generate precise "dominant color" annotations for part instances (in Section A.1). A filtering strategy is employed to ensure that each label faithfully represents the primary color characteristics of the part's appearance (in Section A.2).

A.1 Dominant Color Extraction Pipeline

For each region of the segmented part, we designed a standardized pipeline for extracting the dominant color. First, all pixels within the part region are quantized to suppress noise. Subsequently, the pixels are converted from the RGB to the CIELAB color space, where k-means clustering ($k = 5$) is performed. This process groups the pixels into k discrete color clusters, effectively summarizing the color distribution within the region. A cluster is designated as the "dominant color" if its pixel proportion exceeds a threshold $\tau_{\text{dom}} = 0.70$. This threshold ensures that the selected color covers at least 70% of the region, effectively filtering out interference from secondary colors. Color labels are only provided for part instances meeting the dominant color ratio. This approach prevents color confusion and ensures every labeled color in the dataset is clear and accurate.

A.2 Construction of Task-Specific Datasets

To address the issue of colors that are numerically distinct yet perceptually similar, we incorporate the CIEDE2000 (ΔE_{00}) color difference formula [2]. If the color difference between any two colors is less than a similarity threshold, $\Delta E_{00} < \tau_{\text{sim}}$, they are considered to belong to the same color class. Based on the user study detailed below, we set $\tau_{\text{sim}} = 5$, which corresponds to the empirical threshold at which human observers perceive two colors as equivalent.

For the compositional in-context learning task, the algorithmically extracted dominant color of a part, $\mathcal{C}_{\text{part}}$, is compared with a reference color from the prompt, $\mathcal{C}_{\text{prompt}}$. If the perceptual similarity condition is met ($\Delta E_{00}(\mathcal{C}_{\text{part}}, \mathcal{C}_{\text{prompt}}) < \tau_{\text{sim}}$), the mask corresponding to the object containing that part is included in the ground-truth (GT) mask.

For providing color information to conventional in-context models, which cannot directly process a compositional prompt comprising a color-part-species triplet, we retrieve a reference image for each target image that best matches in terms of object category, part type, and dominant color.

For constructing data for referring segmentation models, we generate textual captions using the template: `The {category} with a {color} {part}` and `The {category} with a {part}`. The `{color}` name is sourced from a Wikipedia library of color name-RGB pairs, selecting the name that is closest to the dominant color of the target part. This ensures consistency between the color description in the prompt and the reference image.

A.3 User Studies and Hyperparameter Calibration

We recruited 21 participants to conduct two subjective evaluation studies to validate and determine the thresholds τ_{dom} and τ_{sim} mentioned above.

1. **Dominant Color Representativeness Study**: We presented participants with n part images and asked them to judge whether a given color (the candidate dominant color extracted by our algorithm) was representative of the part. The samples were divided into four bins based on the pixel proportion p of the candidate dominant color cluster: $p > 0.70$, $0.50 < p \leq 0.70$, $0.30 < p \leq 0.50$, and $p \leq 0.30$. The results showed that when $p > 0.70$, the agreement rate for the color being representative reached 75.5%, significantly higher than in the other bins.
2. **Perceptual Color Similarity Study**: We showed participants $2n$ pairs of dominant color patches extracted from our dataset and asked if the two colors in each pair looked "essentially the same." These pairs were categorized into five bins based on their ΔE_{00} value: $[0, 5)$, $[5, 10)$, $[10, 20)$, $[20, 40)$, and ≥ 40. The results indicated that for pairs with $\Delta E_{00} < 5$, the agreement rate for perceptual identity exceeded 75%. In contrast, when $\Delta E_{00} \geq 10$, the agreement rate dropped sharply to below 32%.

Based on the quantitative results from these two user studies, we finalized the dominance significance threshold at $\tau_{\text{dom}} = 0.70$ and the perceptual color similarity threshold at $\tau_{\text{sim}} = 5$.

B. Visualization of Compositional Multiple Prompt-driven In-context Segmentation

A notable feature of our decoupled architecture is its flexibility and extensibility. We found that, by iteratively applying the model's Conditional Constraint Stage (Stage 2), multiple constraints can be chained for certain samples. Figure 1 shows this process: the model first performs a general in-context segmentation (Stage 1), then generates a mask output based on the first pair of part-color-specific attributes (Stage 2), and subsequently re-applies Stage 2 by applying a second pair of part-color-specific attributes constraints to produce a mask output (Stage 2 again), achieving a multi-conditional constrained result. Compared to a single prompt, this multi-stage approach allows for segmentation under more complex and hierarchical conditions. This implies that our framework is applicable to real-world tasks requiring more refined and diversified targets.

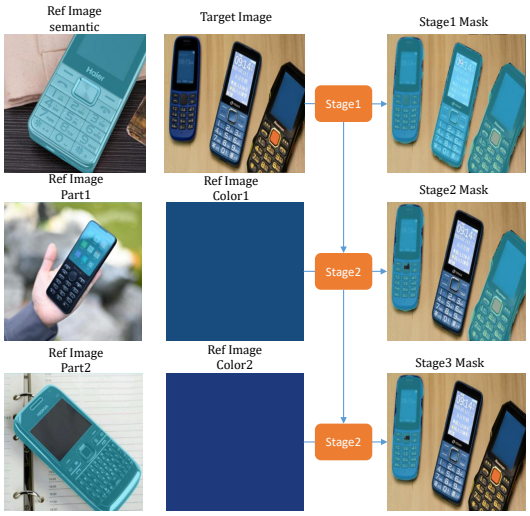

Figure 1. Visualizations of the iterative constraint application. Starting with a generic segmentation from Stage 1, the model iteratively applies the Conditional Constraint Stage (Stage 2) twice. It sequentially incorporates specific part-color attribute pairs (Part1-Color1, then Part2-Color2) to progressively refine the segmentation mask, demonstrating the model's capability for handling hierarchical multi-condition constraints.

C. Qualitative Analysis of Failure Cases

To comprehensively understand the applicable boundaries of the CDICS framework, we visualize two typical failure cases in Figure 2, which highlight the inherent challenges of fine-grained in-context segmentation.

(a) Limitation of Resolution and Scale (Figure 2(a)): When a target object is located in the far background and appears extremely small, it may be entirely missed by the model. Because its fine-grained parts (e.g., the ears of the distant dog) lack sufficient pixel-level detail, the model fails to extract reliable part features or align them spatially. Consequently, the model only successfully segments the prominent instance (the large dog), generating no mask for the distant tiny instance.

(b) Ambiguity from Coupled Similarities and Vague Prompts (Figure 2(b)): A fundamental challenge arises when a distractor in the target image shares highly similar color and part morphology with the prompt, while the semantic prompt itself lacks holistic context. In this case, the reference semantic prompt does not provide the distinct global structure of the entire object. Because the distractor (a pair of scissors) possesses blades with matching color and shape, this combination of strong part-color similarity and vague global semantic guidance misleads the model, resulting in the erroneous segmentation of the scissors.

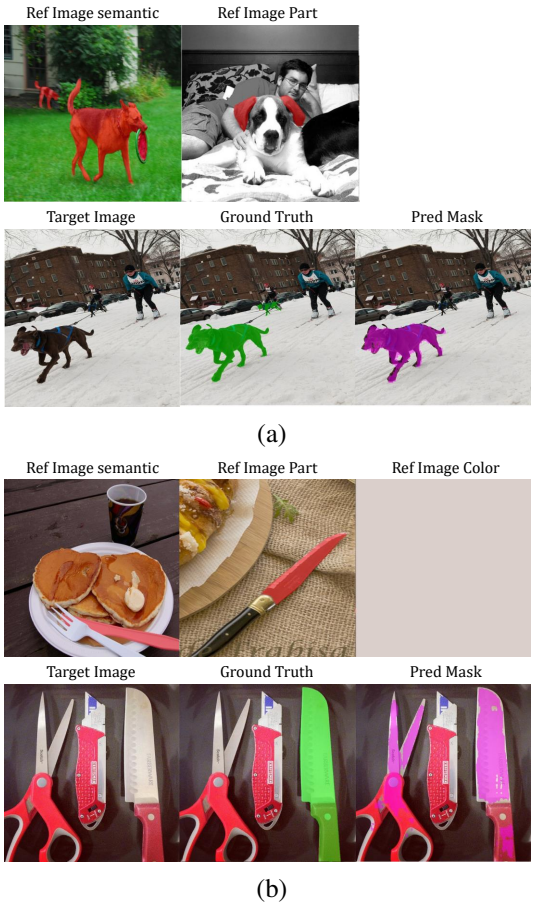

Figure 2. Qualitative failure cases. (a) The model segments only the prominent large dog, completely missing the small dog in the background. (b) The model incorrectly segments the scissor blades because they share highly similar color and morphological features with the reference image.

References

[1] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 1

[2] M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(5):340–350, 2001. 1

[3] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 1