RELATIONAL STRUCTURE AS A PRIOR FOR WEAKLY-SUPERVISED PART SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study weakly-supervised part semantic segmentation using only image-level object categories and the semantic parts present for each object, without part-level spatial supervision. In this setting, the main difficulty is to establish correspondence between fine-grained part concepts and visual regions when several visually similar parts may coexist within the same object. We observe that although their locations are unknown, part concepts are not independent: pretrained part-text representations already encode relations among them. We therefore use this relational structure as a prior for part-level visual-semantic alignment. Our method identifies part-specific visual support through relative competition among the parts present in the same object, forms image-specific visual prototypes, and adapts part representations with a shared orthogonal transformation that preserves their relational geometry. Experiments on PP-116 and ADE-234 show clear improvements under this weak supervision setting, supporting the use of pretrained relational structure for learning fine-grained part correspondence without part-level spatial annotations.

1 INTRODUCTION

Part semantic segmentation aims to assign a semantic part label to each valid pixel in an image. Existing methods usually rely on dense part masks or weaker spatial annotations, such as points and bounding boxes, to establish correspondence between part concepts and visual regions (Chen et al., 2014; He et al., 2022; Wu et al., 2024). We study a weaker setting where each training image provides only the object categories present in the image and, for each object category, the set of semantic parts present in that image, without part-level spatial annotations. Such supervision is practical for large image collections organized by part taxonomies, product catalogs with component lists, and adaptation to new domains or new part vocabularies where dense part masks are difficult to obtain. Part taxonomies are already defined in common part segmentation benchmarks (He et al., 2022; Wei et al., 2023), while image-level part-presence labels require no annotation of precise part locations or boundaries. The main problem is to recover dense part-to-region correspondence from labels that indicate which parts are present but provide no information about their locations.

Recent studies on representation alignment suggest that independently trained models can retain compatible pairwise structure across modalities even when their feature coordinates differ (Huh et al., 2024; Schnaus et al., 2025). Such observations make relational structure a useful constraint for connecting frozen visual and semantic representations. Part-level alignment introduces an additional difficulty because several related parts often appear within the same object and may respond to overlapping or visually similar regions. A single region can therefore have high similarity to several part concepts at the same time. Reliable part correspondence requires resolving the competition among candidate parts within each object.

We use this competition to identify part-specific visual support. For a given part, a region is more informative when its similarity to that part is higher than its similarity to the other parts present in the same object. Relative scores therefore provide a direct way to select part-specific regions and construct image-specific visual prototypes from image-level labels. These prototypes are generated from the current part representations and may contain incorrect assignments. Updating part representations toward such targets also changes the representation space used for subsequent target selection, allowing early assignment errors to accumulate during training. This leads to two

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.

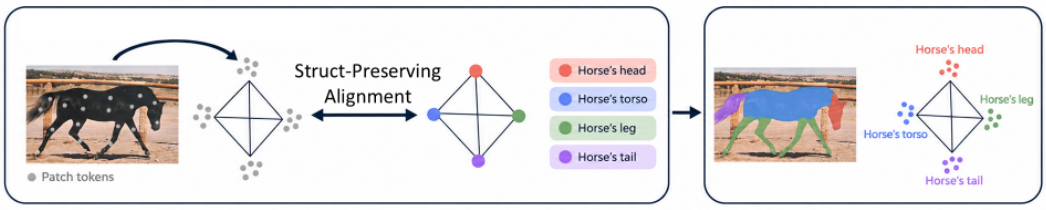


Figure 1: Conceptual illustration of our motivation. Without part-level spatial supervision, part locations are unknown, while pretrained part-text representations retain relations among semantic part concepts. We use this relational structure as a prior for learning part-level visual–semantic correspondence and preserve it during adaptation.

problems: locating part-specific visual targets and updating part representations without losing the semantic relations already encoded among part concepts. As illustrated in Fig. 1, we use the relational geometry of pretrained part representations as a prior during adaptation and preserve their pairwise relations while fitting the induced visual targets.

Our method contains two components. First, within the predicted object foreground, each region is scored relative to the other parts present in the object, and regions with high relative scores are pooled into an image-specific relative visual prototype. Second, all part representations are adapted using one shared orthogonal transformation. Orthogonality preserves inner products and vector norms, and therefore keeps the pairwise cosine relations among part representations unchanged during adaptation. We implement the method in the cross-modal representation space of Talk2DINO (Barsellotti et al., 2025), while keeping its visual encoder, text encoder, and pretrained object-level projector frozen. The relative score determines which regions are selected for each part, while the orthogonal mapping restricts how the part representation space changes when learning from these self-induced targets.

We evaluate our method on PP-116 and ADE20K-Part-234 (Chen et al., 2014; Wei et al., 2023; Zhou et al., 2017), using the same training data and supervision for all compared methods. Compared with Talk2DINO-Part, our method improves mIoU from 26.85 to 32.60 on PP-116 and from 5.33 to 6.38 on ADE-234. Ablation results show that relative part competition improves part-level correspondence, while structure-preserving adaptation gives a further improvement over unconstrained linear adaptation. We also observe that unconstrained linear adaptation increasingly follows its self-generated prototypes during training while its agreement with the ground-truth part structure decreases, whereas orthogonal adaptation remains more stable. To our knowledge, this is the first study of part semantic segmentation in which the target training set provides only image-level object and part-presence labels, without any part-level spatial supervision. Our contributions are threefold: **(i)** we introduce this weakly supervised setting for part semantic segmentation and identify intra-object part competition and changes to part relations when learning from self-induced visual targets as two key difficulties; **(ii)** we propose a simple method that combines relative part competition with exact preservation of part-level relational geometry; and **(iii)** experiments on two part segmentation benchmarks show the effectiveness of the proposed method and clarify the roles of the two components.

2 Related Work

2.1 Part Segmentation

Part semantic segmentation is commonly learned with spatial part annotations. Fully supervised methods such as Compositor and KPS use pixel-level part masks to learn fine-grained part representations and object–part relations (He et al., 2023; Gou et al., 2025). WPS-SAM reduces the annotation requirement to points or bounding boxes, while these annotations still provide explicit spatial cues for locating object parts (Wu et al., 2024). Open-vocabulary part segmentation studies a different supervision and generalization problem. Methods such as VLPart, OV-PARTS, PartCLIPSeg, and PartCATSeg learn from annotated base categories and extend part recognition to unseen objects

or part vocabularies (Sun et al., 2023; Wei et al., 2023; Choi et al., 2024; 2025). DM-PCL removes pixel-level part masks in a weakly-supervised few-shot setting, where textual part labels are provided for support images and the learned part information is transferred to query images (Han et al., 2025). We study target-dataset training where each training image provides only the object categories present in the image and, for each object category, the set of semantic parts present in that image. No points, bounding boxes, part masks, or support–query episodes are used. The resulting problem is to learn dense semantic part correspondence directly from image-level part-presence supervision.

2.2 Dense Visual–Semantic Alignment

Image–text supervision has been widely used to transfer image-level semantics to dense prediction. SegCLIP aggregates image patches into semantic regions, while TCL directly learns region–text alignment from image–text pairs (Luo et al., 2023; Cha et al., 2023). Another line of work connects language semantics with spatially informative visual representations. SimZSS aligns a text encoder with a frozen spatially-aware visual encoder, DINOv2 Meets Text learns image- and pixel-level vision–language alignment with frozen DINOv2 features, and Talk2DINO maps text representations into the DINOv2 visual space for open-vocabulary segmentation (Stegmüller et al., 2025; Jose et al., 2025; Barsellotti et al., 2025). Our method uses the object-level cross-modal space of Talk2DINO as its starting point.

These methods mainly establish dense correspondence for object-level or open-vocabulary semantic concepts. Part segmentation has a different ambiguity because several related concepts coexist inside the same object and can activate overlapping or visually similar regions. High region–text similarity alone may therefore assign the same region to several candidate parts. Our work focuses on this intra-object ambiguity and learns part-level correspondence from the relative responses among the parts known to be present.

2.3 Relational Structure in Representation Alignment

Relational structure provides a way to compare independently learned representation spaces without requiring their coordinates to coincide. The Platonic Representation Hypothesis reports that stronger models trained on different modalities can exhibit increasingly compatible relations among data points (Huh et al., 2024). Schnaus et al. (Schnaus et al., 2025) further show that pairwise relations within vision and language spaces can be used to recover cross-modal correspondence without parallel data. Geometry preservation has also been studied in embedding alignment. For example, orthogonal mappings preserve the internal geometry of monolingual word embeddings while aligning independently learned spaces (Artetxe et al., 2016).

These studies show that pairwise relational structure can provide useful information for correspondence between independently learned representation spaces. Our setting contains an additional dependency: the visual target of each part is selected using the current part representation, and the selected target is then used to update that representation. Incorrect assignments can therefore affect subsequent target selection. We preserve the pairwise cosine geometry of pretrained part representations during this adaptation process, while image-specific part prototypes are induced from the current visual features. Relational geometry thus serves as a fixed structural prior for learning from self-induced part targets.

3 Method

We consider part semantic segmentation with image-level semantic supervision. For each training image, the object categories present in the image are known, together with the semantic parts that appear for each object category. No part mask, point, bounding box, or instance location is used for part-level training. Two problems need to be addressed under this supervision. First, several related parts within the same object may respond strongly to the same or nearby visual regions, so absolute part–region similarity alone is insufficient for locating each part. Second, the visual targets used for adaptation are generated from the current part representations, and errors in these targets may affect later updates. We use relative part competition to locate part-specific visual targets and a shared orthogonal transformation to update part representations while keeping their pairwise cosine relations

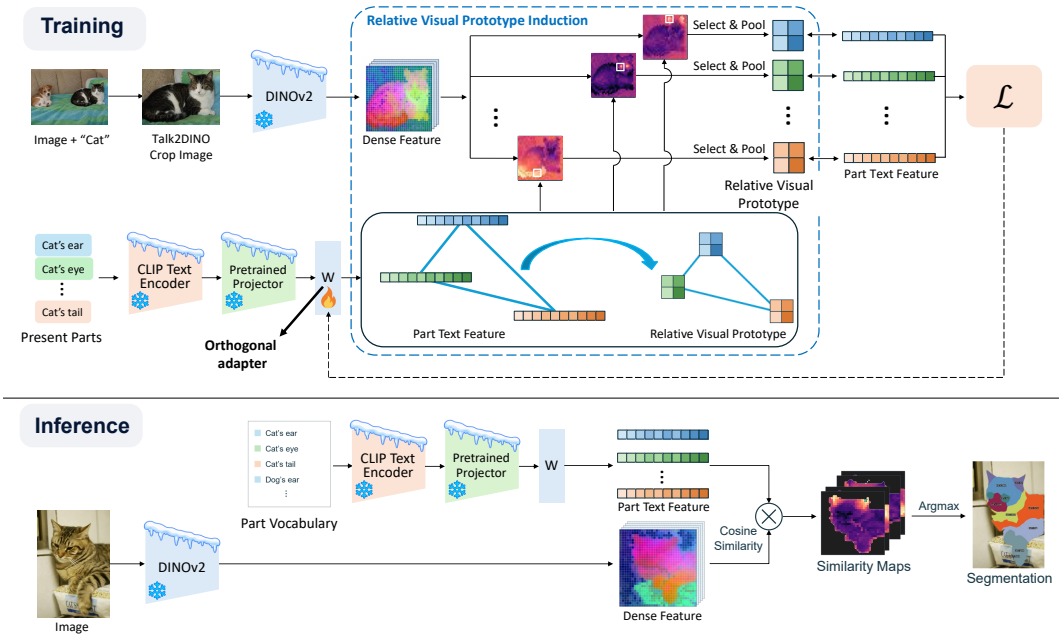


Figure 2: Overview of our method. During training, frozen Talk2DINO predicts the foreground of each image-present object category and frozen DINOv2 extracts dense visual features. Relative scores among the parts present in the object select part-specific visual support, which is pooled into image-specific relative visual prototypes. A shared orthogonal transformation adapts part-text representations toward these prototypes while preserving their pairwise cosine relations. At inference, the learned transformation is applied to the full part vocabulary and dense predictions are obtained by cosine similarity with DINOv2 features.

unchanged. The method is implemented in the frozen cross-modal space of Talk2DINO (Barsellotti et al., 2025).

3.1 Problem Formulation and Relational Prior

Let I denote a training image and $\{T_j\}_{j=1}^{M}$ the complete semantic part vocabulary. For an object category o present in I, we denote by

$$\mathcal{C}_I(o) \subseteq \{1, \dots, M\}$$

the set of semantic parts of o that are present in this image. The set $\mathcal{C}_I(o)$ is image-specific and can contain only a subset of the complete part taxonomy of o. The supervision therefore specifies which parts occur in the image, while their locations are unknown.

Let E_t be the frozen text encoder and P_{obj} the frozen text-to-visual projector pretrained by Talk2DINO. For each semantic part T_j, we define its representation in the DINOv2 visual space as

$$t_j = \text{Norm}\left(P_{\text{obj}}(E_t(T_j))\right), \qquad j = 1, \dots, M, \tag{1}$$

where $\text{Norm}(\cdot)$ denotes ℓ_2 normalization. Since P_{obj} is pretrained with image–caption supervision and mainly provides object-level semantic alignment, these representations are not expected to be accurately aligned with individual part regions. We use their pairwise cosine relations as a structural prior during part-level adaptation:

$$G_{ij} = \cos(t_i, t_j). \tag{2}$$

Here $G \in \mathbb{R}^{M \times M}$ describes the relational geometry of the full part vocabulary. Our goal is to improve part-level visual correspondence while keeping these relations fixed.

3.2 Relative Visual Prototype Induction

The image-level labels do not specify where a part is located. For a training image I and an object category o, we first use frozen Talk2DINO to predict the foreground of o. We denote the foreground DINOv2 patch set by Ω, and the normalized feature of patch $p \in \Omega$ by x_p. The predicted foreground only restricts the region in which part targets are searched; it does not determine the correspondence between individual parts and patches. All spatial information in Ω comes from the frozen object prediction rather than ground-truth part annotations. For simplicity, we write $\mathcal{C} = \mathcal{C}_I(o)$ below.

We introduce a shared transformation $W \in \mathbb{R}^{D_v \times D_v}$. For each part T_j present in the current image, its current representation is

$$\hat{t}_j = \mathrm{Norm}(t_j W), \qquad j \in \mathcal{C}. \tag{3}$$

The similarity between part T_j and foreground patch p is

$$S_{j,p} = \hat{t}_j^\top x_p. \tag{4}$$

A large $S_{j,p}$ does not necessarily mean that patch p corresponds to part j, because another part in the same object may obtain a similar or higher score on the same patch. We therefore define

$$R_{j,p} = S_{j,p} - \max_{\substack{k \in \mathcal{C} \\ k \neq j}} S_{k,p}, \tag{5}$$

when more than one part is present. For $|\mathcal{C}| = 1$, we set $R_{j,p} = S_{j,p}$.

The value $R_{j,p}$ is the similarity margin of part j over its strongest competing part in the same object. In particular, $R_{j,p} > 0$ means that part j has the highest score among all parts in $\mathcal{C}$ at patch p. We use this relative score to select the visual support of each part.

For part T_j, the patch with the largest relative score is selected as

$$a_j = \arg\max_{p \in \Omega} R_{j,p}. \tag{6}$$

We always include a_j, and further select at most $L - 1$ non-anchor patches with positive relative scores:

$$\mathcal{P}_j = \{a_j\} \cup \mathrm{Top}_{L-1}\left\{p \in \Omega \setminus \{a_j\} \,|\, R_{j,p} > 0\right\}. \tag{7}$$

The anchor ensures that each present part has at least one selected patch. Additional patches are used only when part j scores higher than every other present part. If no such additional patch exists, $\mathcal{P}_j$ contains only the anchor. The upper bound L limits the number of patches included in one prototype. Different parts are allowed to select the same patch, and no one-to-one matching is imposed.

The visual prototype of part T_j is obtained by averaging the selected patch features:

$$r_j = \mathrm{Norm}\left(\frac{1}{|\mathcal{P}_j|} \sum_{p \in \mathcal{P}_j} x_p\right). \tag{8}$$

We use r_j as the visual target for T_j. Since $\mathcal{P}_j$ is selected using the current $\hat{t}_j$, the prototype also depends on the current W. Updating W therefore changes the patches selected in later iterations. During each update, the prototypes are treated as fixed targets.

3.3 Structure-Preserving Part Adaptation

The visual prototypes above are generated by the current model and can contain incorrect part assignments. An unconstrained transformation can fit these prototypes by changing both the positions of individual part representations and their relations to other parts. Such changes also affect the relative scores used to construct later prototypes. We therefore restrict the shared transformation W to be orthogonal:

$$W^\top W = I. \tag{9}$$

The same W is applied to all M semantic parts. Since an orthogonal transformation preserves inner products and vector norms,

$$\cos(t_i W, t_j W) = \cos(t_i, t_j) = G_{ij}, \qquad \forall i, j \in \{1, \ldots, M\}. \tag{10}$$

Thus, W can rotate or reflect all part representations jointly while their pairwise cosine similarities remain unchanged. The relative score in Eq. 5 is computed within the parts present in the current object, whereas the constraint in Eq. 10 applies to the full part vocabulary.

We optimize W by aligning each adapted part representation with its corresponding visual prototype:

$$\mathcal{L} = \mathbb{E}_{(I,o)} \left[\frac{1}{|\mathcal{C}_I(o)|} \sum_{j \in \mathcal{C}_I(o)} \left(1 - \cos(\hat{t}_j, r_j)\right) \right]. \tag{11}$$

For each optimization step, the prototypes are first recomputed using the current W and are then held fixed when W is updated. Only W is optimized; the text encoder, visual encoder, and pretrained Talk2DINO projector remain frozen.

3.4 Inference

After training, the learned transformation W can be folded into the frozen text-to-visual projector. For every part in the full vocabulary, the adapted representation is

$$\hat{t}_j = \mathrm{Norm}\left(P_{\mathrm{obj}}(E_t(T_j))W\right), \qquad j = 1, \ldots, M. \tag{12}$$

Given a test image, the frozen DINOv2 encoder produces dense visual features. Let $x_{h,w}$ denote the normalized feature at spatial location (h, w). The predicted part label is

$$\hat{Y}[h, w] = \arg \max_{j=1,\ldots,M} \cos(x_{h,w}, \hat{t}_j). \tag{13}$$

The main evaluation therefore predicts directly over the full part vocabulary and does not require object labels or ground-truth spatial information at inference time.

4 Experiments

4.1 Experimental Setup

Datasets and supervision. We evaluate our method on two part semantic segmentation benchmarks, PASCAL-Part-116 (PP-116) (Chen et al., 2014; Wei et al., 2023) and ADE20K-Part-234 (ADE-234) (Zhou et al., 2017; Wei et al., 2023). PP-116 contains 116 semantic part categories, with 8,431 training images and 850 validation images. ADE-234 contains 234 semantic part categories associated with 44 object categories, with 7,347 training images and 1,016 validation images. We evaluate on the complete part vocabulary of each benchmark. For each training image, the supervision consists of the semantic object categories present in the image and, for each object category o, the corresponding set of present parts $\mathcal{C}_I(o)$. For PP-116, these labels are derived from the training object and part annotations; for ADE-234, the present part IDs are grouped according to the fixed part-to-object taxonomy. After these non-spatial labels are obtained, spatial annotation information is not used for object localization or part adaptation. The exact label conversion is described in Appendix A.1. Unless otherwise specified, ablations and analyses are conducted on PP-116.

Implementation details. We use DINOv2 ViT-B/14 with registers (Oquab et al., 2024; Dosovitskiy et al., 2021; Darcet et al., 2024) as the visual encoder and CLIP ViT-B/16 (Radford et al., 2021) as the text encoder. The text-to-visual projector is initialized from the Talk2DINO model pretrained on MS-COCO Captions 2014 (Chen et al., 2015; Barsellotti et al., 2025). DINOv2, CLIP, and the Talk2DINO projector remain frozen; only the shared transformation $W \in \mathbb{R}^{768\times768}$ is optimized. We initialize W as the identity matrix and train for 30 epochs with Adam (Kingma & Ba, 2015), a learning rate of 1×10^{-5}, a batch size of 16, and zero weight decay. After each optimizer step, if $W = U\Sigma V^\top$ is its singular value decomposition, we set $W \leftarrow UV^\top$ to maintain orthogonality. Each relative visual prototype contains at most $L = 8$ patches by default. Each valid image–object-category pair forms one training sample. For foreground prediction, frozen Talk2DINO compares the image-present object categories together with a background channel; we use a background threshold of 0.55 and apply PAMR (Araslanov & Roth, 2020). The resulting object crop is resized to 448×448 before extracting DINOv2 patch features. Additional training and foreground-generation details are given in Appendix A.3.

Table 1: Comparison on PP-116 and ADE-234. The "-Part" baselines are adapted to the target benchmark using the same training images, image-level object categories, and per-object part-presence labels.

Method	Visual Encoder	Text Encoder	PP-116		ADE-234	
			mIoU	mACC	mIoU	mACC
SimZSS-Part	DINOv2-B/14-Reg	Transformer-12L	10.83	23.83	6.20	16.01
SegCLIP-Part	CLIP ViT-B/16	CLIP Transformer	10.41	19.60	4.25	11.50
TCL-Part	CLIP ViT-B/16	CLIP Transformer	15.36	27.43	3.62	8.60
Talk2DINO-Part	DINOv2-B/14-Reg	CLIP-B/16	26.85	56.76	5.33	**20.06**
Ours	**DINOv2-B/14-Reg**	**CLIP-B/16**	**32.60**	**60.67**	**6.38**	15.82

Evaluation protocol. We report mIoU and mACC on both benchmarks. At inference, DINOv2 extracts dense features from the full image and all 116 or 234 part categories compete directly according to Eq. 13. No object labels, object masks, or part-presence labels are provided at test time, and PAMR is disabled for the final part prediction. Background and unlabeled pixels are ignored following the benchmark annotation conventions and are not treated as an additional semantic part category. All compared methods use the same class mapping, valid-pixel definition, and metric computation.

4.2 Comparison with Existing Methods

We compare with SimZSS (Stegmüller et al., 2025), SegCLIP (Luo et al., 2023), TCL (Cha et al., 2023), and Talk2DINO (Barsellotti et al., 2025). We denote their target-dataset adaptations as SimZSS-Part, SegCLIP-Part, TCL-Part, and Talk2DINO-Part, respectively. Each method starts from its model trained with MS-COCO Captions 2014 and is then adapted for 30 epochs on the target benchmark using the same training images and the same image-level object and part-presence supervision. No part-level spatial supervision is used during target-dataset adaptation.

On PP-116, our method improves Talk2DINO-Part from 26.85 to 32.60 mIoU and from 56.76 to 60.67 mACC. The gains are obtained under the same target training images and weak supervision, showing that part-specific target construction and structure-preserving adaptation improve the transfer from object-level alignment to part-level correspondence.

On ADE-234, our method obtains 6.38 mIoU, compared with 6.20 for SimZSS-Part and 5.33 for Talk2DINO-Part. Its mACC is 15.82, lower than the 20.06 of Talk2DINO-Part. The gain on this benchmark is therefore reflected mainly in spatial overlap, while class-wise accuracy remains limited.

4.3 Component Ablation

We isolate the two main design choices on PP-116. Talk2DINO (Direct) evaluates the COCO-Captions-pretrained projector on PP-116 without target-dataset adaptation. Absolute + Linear and Absolute + Orthogonal construct visual targets from absolute text–patch similarity and use a linear or orthogonal mapping, respectively. Relative + Linear replaces absolute similarity with the relative score in Eq. 5, while Relative + Orthogonal combines the relative score with the proposed orthogonal mapping.

Relative scoring. With the mapping fixed to a linear transformation, replacing absolute similarity with the relative score raises mIoU from 12.38 to 29.88 and mACC from 28.49 to 57.56. Absolute target selection also performs below Talk2DINO (Direct), showing that target-dataset adaptation can be harmful when visually similar parts repeatedly select the same foreground regions. Comparing each patch with the strongest competing part gives substantially better part-specific targets.

Structure-preserving adaptation. With relative target construction fixed, the orthogonal mapping improves mIoU from 29.88 to 32.60 and mACC from 57.56 to 60.67. The same direction is

Table 2: Component ablation on PP-116. Talk2DINO (Direct) denotes direct evaluation of the COCO-Captions-pretrained projector without target-dataset adaptation.

Variant	Target Score	Mapping	mIoU	mACC
Talk2DINO (Direct)	–	None	21.70	39.62
Absolute + Linear	Absolute	Linear	12.38	28.49
Absolute + Orthogonal	Absolute	Orthogonal	16.16	35.61
Relative + Linear	Relative	Linear	29.88	57.56
Relative + Orthogonal	**Relative**	**Orthogonal**	**32.60**	**60.67**

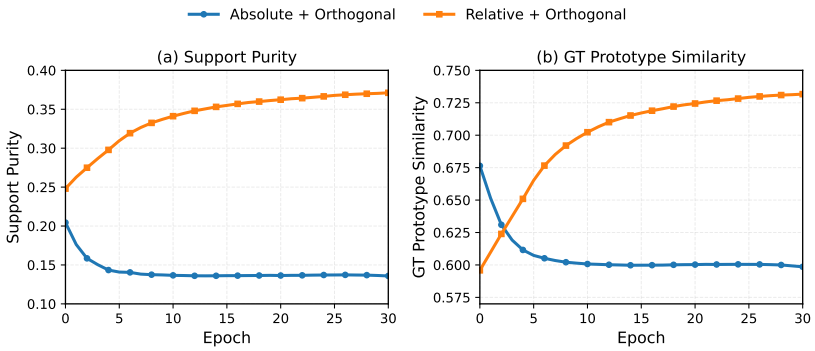


Figure 3: Post-hoc analysis of induced visual prototypes on PP-116. We compare Absolute + Orthogonal and Relative + Orthogonal. Support Purity is the fraction of selected support patches assigned to the target part by the GT patch labels, and GT Prototype Similarity is the cosine similarity between the induced prototype and the corresponding image-specific GT visual prototype.

observed with absolute target construction, where orthogonal adaptation improves mIoU from 12.38 to 16.16. Preserving the pairwise cosine relations among part representations therefore provides an additional gain after the visual target construction is fixed.

4.4 Analysis

Ground-truth spatial annotations in this subsection are used only for post-hoc analysis. They do not enter prototype construction, parameter optimization, or inference. Exact metric construction and aggregation are given in Appendix B.

Relative visual prototype quality. We first compare Absolute + Orthogonal and Relative + Orthogonal with the same orthogonal mapping and $L = 8$. Support Purity is the fraction of selected support patches whose ground-truth patch label matches the corresponding part, reported as a class-macro average. GT Prototype Similarity is the cosine similarity between an induced image-specific prototype and the visual prototype formed from the ground-truth patches of the same part in that crop.

As shown in Fig. 3(a), Support Purity decreases from 0.2044 to 0.1359 for Absolute + Orthogonal, while it increases from 0.2481 to 0.3711 for Relative + Orthogonal. The same difference appears at the prototype level. GT Prototype Similarity drops from 0.6764 to 0.5986 with absolute scoring, whereas relative scoring increases it from 0.5958 to 0.7317. Relative scoring therefore selects cleaner part-specific support and produces visual prototypes that become closer to the corresponding GT visual prototypes during training.

Structure preservation during adaptation. We next compare Relative + Linear and Relative + Orthogonal, keeping the relative target construction and $L = 8$ fixed. For this analysis, one GT visual prototype is constructed for each part by pooling DINOv2 patch features over its ground-truth regions in the PP-116 training set. For each parent-object group, we form the pairwise cosine relations among adapted part-text representations, induced visual prototypes, and GT visual proto-

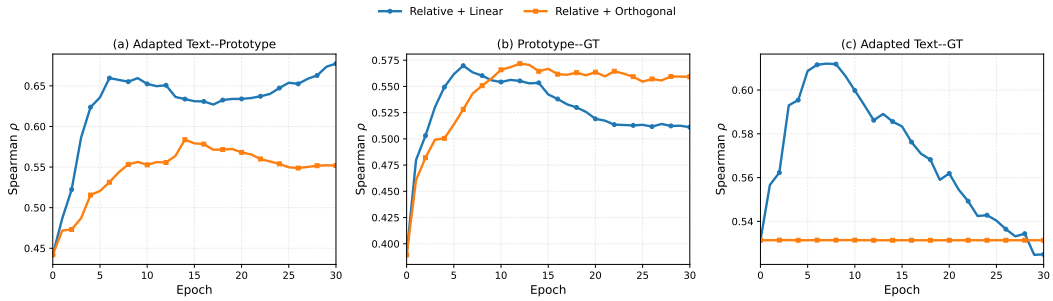


Figure 4: Relational structure during adaptation on PP-116. For Relative + Linear and Relative + Orthogonal, we report Spearman correlations between pairwise cosine-relation vectors of (a) adapted part text and induced visual prototypes, (b) induced and global GT visual prototypes, and (c) adapted part text and global GT visual prototypes, macro-averaged over valid parent-object groups. GT prototypes are used only for this post-hoc analysis.

types, and compute Spearman correlation between these relation vectors. The reported values are macro-averaged over parent-object groups with at least three valid parts.

Figure 4(a) shows that the linear mapping fits the relational structure of its own induced prototypes more strongly: Adapted Text–Prototype correlation increases from 0.4419 to 0.6774, compared with 0.5519 for the orthogonal mapping at epoch 30. This stronger agreement with self-induced targets does not persist when the targets are compared with the GT visual structure. In Fig. 4(b), Prototype–GT correlation for Relative + Linear peaks at 0.5697 at epoch 6 and then falls to 0.5111, while Relative + Orthogonal reaches 0.5717 at epoch 12 and ends at 0.5592.

Figure 4(c) gives the corresponding change on the text side. Relative + Orthogonal keeps Adapted Text–GT correlation at approximately 0.5314, as expected from the exact preservation of pairwise cosine relations under $W^\top W = I$. Relative + Linear initially increases the correlation to 0.6121 at epoch 7, but it later decreases to 0.5249. The linear mapping can therefore fit the current self-induced targets by changing the pairwise part relations, whereas the orthogonal mapping keeps these relations fixed while adapting to the visual targets. Sensitivity to the prototype support size and qualitative examples are reported in Appendix C.

5 Conclusion

We study part semantic segmentation without part-level spatial supervision, using only image-level object categories and per-object part-presence labels during target-dataset training. Our method exploits the relations already encoded among part representations to recover missing part-level visual correspondence: local competition is used to identify part-specific visual targets, while their relational structure is preserved during adaptation. Results on PP-116 and ADE-234 show consistent improvements under this weak supervision setting. More importantly, our findings suggest that when visual targets must be inferred from weak supervision and are therefore imperfect, preserving reliable pretrained relations can provide a useful prior for learning fine-grained part correspondence.

AI Use Statement

Generative AI tools were used only to assist with language polishing and improving the clarity and readability of the manuscript. All scientific ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of this work.

REFERENCES

Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Araslanov_Single-Stage_Semantic_Segmentation_From_Image_Labels_CVPR_2020_paper.html.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, 2016. doi: 10.18653/v1/D16-1250. URL https://aclanthology.org/D16-1250/.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22035, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Barsellotti_Talking_to_DINO_Bridging_Self-Supervised_Vision_Backbones_with_Language_for_ICCV_2025_paper.html.

Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11165–11174, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cha_Learning_To_Generate_Text-Grounded_Mask_for_Open-World_Semantic_Segmentation_From_CVPR_2023_paper.html.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Chen_Detect_What_You_2014_CVPR_paper.html.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. URL https://arxiv.org/abs/1504.00325.

Jiho Choi, Seonho Lee, Seungho Lee, Minhyun Lee, and Hyunjung Shim. Understanding multi-granularity for open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4358. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f7f47a73d631c0410cbc2748a8015241-Abstract-Conference.html.

Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, and Hyunjung Shim. Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9782–9793, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Choi_Fine-Grained_Image-Text_Correspondence_with_Cost_Aggregation_for_Open-Vocabulary_Part_Segmentation_CVPR_2025_paper.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Xuejian Gou, Fang Liu, Licheng Jiao, Shuo Li, Lingling Li, Hao Wang, Xu Liu, Puhua Chen, and Wenping Ma. Knowledge-guided part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5490–5500, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Gou_Knowledge-Guided_Part_Segmentation_ICCV_2025_paper.html.

Mengya Han, Yong Luo, Han Hu, Zengmao Wang, Lefei Zhang, Bo Du, Ling-Yu Duan, and Dacheng Tao. DM-PCL: Text-driven dual-modal prototype consistency learning for weakly-supervised few-shot part segmentation. *International Journal of Computer Vision*, 133(11):7553–7569, 2025. doi: 10.1007/s11263-025-02545-w.

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*, 2022. URL https://arxiv.org/abs/2112.00933.

Ju He, Jieneng Chen, Ming-Xian Lin, Qihang Yu, and Alan L. Yuille. Compositor: Bottom-up clustering and compositing for robust part and object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11259–11268, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/He_Compositor_Bottom-Up_Clustering_and_Compositing_for_Robust_Part_and_Object_CVPR_2023_paper.html.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jose_DINOv2_Meets_Text_A_Unified_Framework_for_Image-_and_Pixel-Level_CVPR_2025_paper.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23033–23044, 2023. URL https://proceedings.mlr.press/v202/luo23a.html.

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 Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24983–24992, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Schnaus_Its_a_Blind_Match_Towards_Vision-Language_Correspondence_without_Parallel_Data_CVPR_2025_paper.html.

Thomas Stegmüller, Tim Lebailly, Nikola ukić, Behzad Bozorgtabar, Tinne Tuytelaars, and Jean-Philippe Thiran. A simple framework for open-vocabulary zero-shot segmentation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QzPKSUUcud.

Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15453–15465, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Sun_Going_Denser_with_Open-Vocabulary_Part_Segmentation_ICCV_2023_paper.html.

Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. OV-PARTS: Towards open-vocabulary part segmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/dde53059fdb0f45e1e9ad9c66997d662-Abstract-Datasets_and_Benchmarks.html.

Xin-Jian Wu, Ruisong Zhang, Jie Qin, Shijie Ma, and Cheng-Lin Liu. WPS-SAM: Towards weakly-supervised part segmentation with foundation models. In *European Conference on Computer Vision*, pp. 314–333, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06063.pdf.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Scene_Parsing_Through_CVPR_2017_paper.html.

A Additional Experimental Details

A.1 Weak-label construction

PP-116. The dense training annotations are used only to derive the non-spatial supervision defined in Sec. 4. For each image, the semantic object mask gives the set of object categories present in the image. For an object category o, we collect the valid semantic part IDs that occur inside the semantic region of o and use this set as $\mathcal{C}_I(o)$. Only the object category and the resulting part-presence set are retained for subsequent training. Multiple instances of the same semantic object category are treated as one semantic category and yield one image–object-category training sample. No ground-truth object or part location is used after this conversion.

ADE-234. The ADE-234 training part mask contains semantic part IDs from the 234-part taxonomy. We first collect the valid part IDs present in an image and then map each part to its parent object category using the fixed benchmark part-to-object taxonomy. Parts with the same parent category form $\mathcal{C}_I(o)$. Thus, no ground-truth object mask is required for ADE-234. The part mask is used only for this presence conversion and is not used by the subsequent localization or adaptation procedure.

A.2 Object localization and feature extraction

During training, the frozen Talk2DINO model predicts object foregrounds using only the object categories known to be present in the image. These categories compete with a constant background channel, and the final semantic object mask is obtained by channel-wise argmax. We use the

Talk2DINO sliding-window setting with input scale $(2048, 448)$, crop size 448, stride 224, background threshold 0.55, and $\lambda_{\text{bg}} = 0.2$. PAMR is applied to the object score maps before the hard object prediction.

For every predicted object mask with non-zero area, we take its bounding box, convert it to a square box with an expansion ratio of 1.2, and resize the RGB crop to 448×448. DINOv2 ViT-B/14-Reg then produces a 32×32 grid of 768-dimensional patch features. The predicted object mask is mapped to the same patch grid using occupancy-based rasterization, so a non-empty foreground region is not removed solely by grid discretization. If Talk2DINO predicts an empty mask for a labeled object category, that image–object-category pair is omitted from part-level adaptation; no ground-truth spatial fallback is used.

A.3 Text features and optimization

We use the `sub_imagenet_template` prompt set for the part names. CLIP ViT-B/16 embeddings are averaged over prompts before the frozen Talk2DINO projector, following the Talk2DINO text path. The projected representations are ℓ_2-normalized before similarity computation. Table 3 summarizes the main optimization settings.

Table 3: Training settings used for the proposed method on PP-116 and ADE-234.

Setting	Value
Visual encoder	DINOv2 ViT-B/14-Reg
Text encoder	CLIP ViT-B/16
Projector initialization	Talk2DINO, COCO Captions 2014
Trainable parameter	shared $W \in \mathbb{R}^{768\times768}$
Initialization of W	identity
Optimizer	Adam
Learning rate	1×10^{-5}
Weight decay	0
Batch size	16
Epochs	30
Random seed	123
Maximum support size	$L = 8$
Object crop size	448×448
Orthogonality enforcement	SVD retraction after each update

One training sample corresponds to one valid image–object-category pair (I, o). Within a sample, the loss is first averaged over the parts in $\mathcal{C}_I(o)$; samples in a minibatch then receive equal weight. At every optimization step, the current W is used to recompute the relative scores, selected support, and image-specific prototypes. These prototypes are held fixed for the gradient update. After the Adam step, $W = U\Sigma V^\top$ is projected back to the orthogonal group by $W \leftarrow UV^\top$.

A.4 Inference and evaluation

After training, W is folded into the last affine layer of the Talk2DINO projector, so inference uses the same dense prediction path as the base model with the adapted part representations. Both benchmarks use full-image sliding inference with crop size 448 and stride 224 at input scale $(2048, 448)$. All part categories are evaluated jointly, and PAMR is disabled. PP-116 uses labels 0–115 as valid part classes and 255 as ignore. ADE-234 uses labels 0–233 as valid part classes and 65535 as ignore. No additional background part channel is introduced, and no label shifting is applied.

A.5 Target-dataset adaptation of baselines

The "-Part" baselines in Table 1 start from the corresponding models trained with MS-COCO Captions 2014 and are further adapted for 30 epochs on the target benchmark. All methods use the same target training split and the same image-level object and per-object part-presence supervision. Their original visual/text encoders and method-specific adaptation mechanisms are retained; only the target semantic vocabulary and supervision are changed to the part benchmark. No target part masks, boxes, or points are supplied during this stage.

B Analysis Protocols

Ground-truth spatial annotations are used in this section only to measure the quality of the learned support and relational structure. They do not enter prototype construction, parameter optimization, or test-time prediction.

B.1 Relative visual prototype quality

For each PP-116 training sample (I, o), the ground-truth part mask is cropped using the same predicted-object crop box used to construct the training sample. The cropped mask is resized directly to the 32×32 DINOv2 patch grid with nearest-neighbor interpolation, giving a single ground-truth part ID for every patch position. This conversion is performed only for the analysis in Fig. 3.

Let $\mathcal{P}_j$ be the selected support of part j and let y_p^{GT} be the ground-truth part ID assigned to patch p. Support Purity for a valid image–part instance is

$$\mathrm{Purity}_{I,o,j} = \frac{1}{|\mathcal{P}_j|} \sum_{p \in \mathcal{P}_j} \mathbf{1}[y_p^{\mathrm{GT}} = j]. \tag{14}$$

For the same instance, the GT visual prototype is formed from all DINOv2 patch features assigned to part j in that crop,

$$g_{I,o,j} = \mathrm{Norm}\left(\frac{1}{|\mathcal{G}_j|} \sum_{p \in \mathcal{G}_j} x_p \right), \qquad \mathcal{G}_j = \{p : y_p^{\mathrm{GT}} = j\}, \tag{15}$$

and GT Prototype Similarity is $\cos(r_j, g_{I,o,j})$. Instances for which $\mathcal{G}_j$ becomes empty after patch-grid discretization are excluded from this analysis. For both metrics, we first average valid image–part instances within each semantic part category and then report the macro average over valid part categories. These class-macro quantities are the curves shown in Fig. 3.

B.2 Relational structure analysis

The relational analysis in Fig. 4 uses one global GT visual prototype for each PP-116 part. Every training image is resized to 448×448, and its GT part mask is mapped to the 32×32 DINOv2 patch grid with nearest-neighbor interpolation. All DINOv2 patch tokens assigned to the same semantic part are pooled across the complete PP-116 training split and the resulting mean is ℓ_2-normalized.

The induced prototype bank is constructed separately at every epoch. For each valid image–object-category sample, we recompute the image-specific relative visual prototype using the corresponding checkpoint. Image-specific prototypes belonging to the same semantic part are averaged and normalized, yielding one induced prototype per part. The Linear and Orthogonal variants therefore use the same set of part instances and differ only in the learned mapping.

For a parent-object group containing K valid parts, we form the vector of all $K(K-1)/2$ pairwise cosine similarities from the upper triangle of its relation matrix. Spearman correlation is computed between the relation vectors of (i) adapted part text and induced prototypes, (ii) induced prototypes and GT prototypes, and (iii) adapted part text and GT prototypes. A parent-object group is retained only when at least three parts have both an induced prototype and a GT prototype and all compared correlations are defined. This gives 13 valid parent-object groups on PP-116. The value reported at each epoch is the unweighted mean of the per-group Spearman correlations.

C Additional Results

C.1 Sensitivity to prototype support size

Figure 5 varies the maximum number of patches L used by each relative visual prototype on PP-116. Performance improves rapidly when moving beyond the anchor-only setting and changes only slightly once L reaches the mid-range values. We use $L = 8$ in all main experiments; it is already close to the saturation region and does not rely on a narrow optimum.

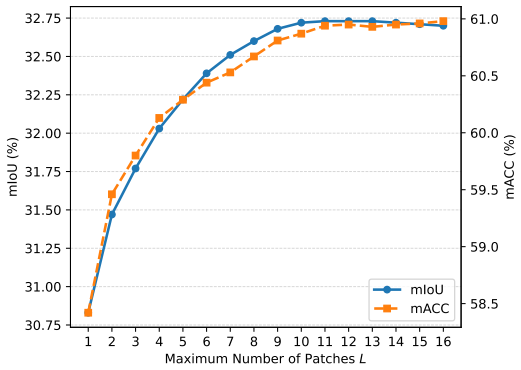


Figure 5: Sensitivity of mIoU and mACC to the maximum prototype support size L on PP-116. $L = 1$ uses only the anchor patch; larger values admit additional positive-margin support patches.

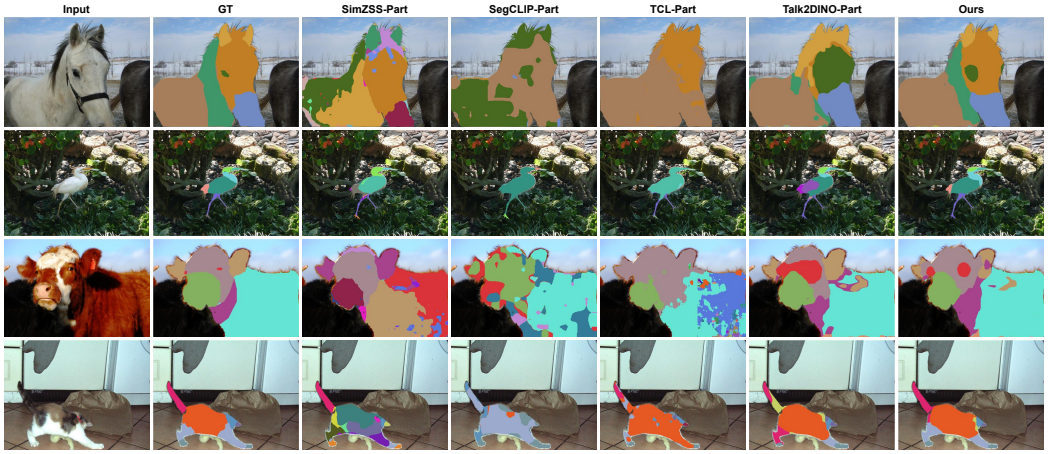


Figure 6: Qualitative comparison on PP-116. From left to right: input image, ground truth, Talk2DINO-Part, and our method.

C.2 Qualitative results

Figure 6 compares Talk2DINO-Part with our method on PP-116. The examples show the same trend as the quantitative results: the proposed method reduces confusion among parts belonging to the same object and gives more coherent part regions. On the horse example, it reduces the large erroneous eye/head response produced by Talk2DINO-Part; on the bird and cat examples, the torso and tail predictions are more consistent with the reference annotation. The cow example remains imperfect around the head and torso, indicating that errors in fine-grained localization are not completely removed.