SPOR: SEMANTICS-PRESERVING ORDINAL REFINEMENT FOR OPEN-VOCABULARY PANOPTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary panoptic segmentation requires recognizing categories unseen during training while accurately separating object instances and delineating coherent stuff regions. However, masks generated from appearance and semantic cues alone may erroneously merge semantically similar regions at different depth layers. We introduce SPOR (**S**emantics-**P**reserving **O**rdinal **R**efinement), a framework that uses within-image depth ordering to constrain spatial grouping in open-vocabulary panoptic segmentation. SPOR converts monocular depth estimates into normalized ordinal ranks, yielding a representation invariant to strictly increasing transformations of depth values. Rather than equating depth continuity with segment identity, it applies ordinal geometry selectively throughout mask formation. Ordinal affinities regulate local and cross-scale feature aggregation, while query-conditioned boundary routing uses reliable ordinal intervals to regulate cross-attention at uncertain boundaries while protecting confident mask cores. A training-only partition consistency objective combines ground-truth segment relations with ordinal gaps and cross-scale agreement to encourage within-segment coherence and cross-segment separation. Direct geometry-dependent modifications are confined to the mask-formation pathway, leaving pretrained semantic embeddings and the visual feature used for semantic pooling unchanged. Built on ODISE and fine-tuned on COCO Panoptic, SPOR achieves 55.5 PQ on COCO and 22.8 PQ in zero-shot transfer to ADE20K, while retaining broadly comparable performance across five open-vocabulary semantic segmentation benchmarks. These results suggest that ordinal scene structure can complement semantic recognition by constraining how pixels are grouped.

1 INTRODUCTION

Recognizing two objects as belonging to the same category does not reveal where one instance ends and the other begins. Likewise, regions with similar appearance may occupy different layers of a scene. Open-vocabulary panoptic segmentation must reconcile these distinctions: it requires recognizing categories beyond the training vocabulary while separating individual thing instances and delineating coherent stuff regions (Kirillov et al., 2019; Xu et al., 2023a). Accurate semantic recognition is essential, but does not by itself resolve the spatial organization required for panoptic prediction.

Pretrained vision–language and diffusion models have enabled substantial progress in open-vocabulary segmentation through dense image–text alignment and transferable mask representations (Li et al., 2022; Liang et al., 2023; Xu et al., 2023a; Yu et al., 2023). However, semantic compatibility alone does not determine which pixels should exchange information or share a mask. Feature aggregation may mix evidence across scene layers, while a decoder query may read incompatible regions when its predicted foreground crosses an object boundary. Masked attention restricts query access to predicted foreground regions (Cheng et al., 2022), but this support remains dependent on an evolving mask estimate. We therefore ask:

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.

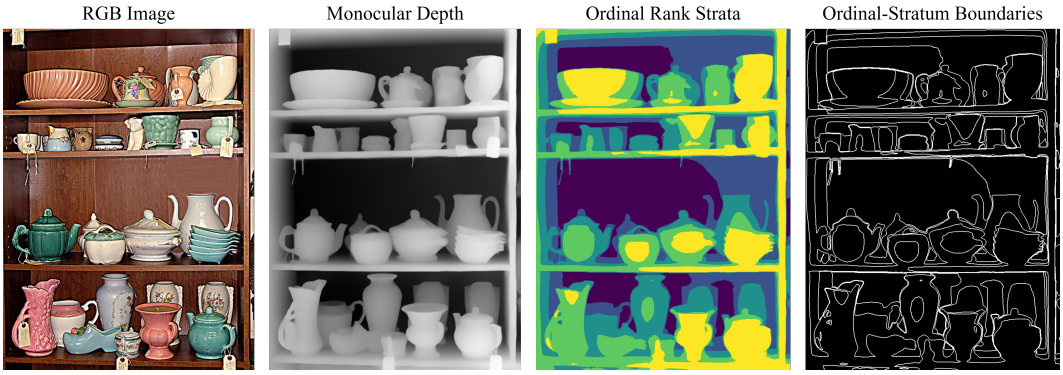

Figure 1: Construction of the ordinal-depth representation. From left to right: the input RGB image, its monocular pseudo-depth estimate, the tie-aware ordinal rank map quantized into five strata, and the boundaries induced by transitions between adjacent strata. The resulting representation preserves within-image depth ordering without relying on metric depth values.

How can relative scene depth constrain information propagation and pixel assignment in open-vocabulary panoptic segmentation, helping distinguish regions that are semantically similar but geometrically separated?

Depth provides structural cues complementary to visual appearance. Geometry-conditioned operators and RGB-D representation learning demonstrate its value for segmentation (Wang & Neumann, 2018; Yin et al., 2024; 2025). We focus on within-image depth ordering—the relative ranking of valid depth estimates—rather than absolute depth values. Specifically, as illustrated in Fig. 1, we convert a monocular depth map into a tie-aware normalized rank map over valid pixels. This representation preserves within-image depth order and is invariant to strictly increasing transformations of depth values. For visual clarity, the figure quantizes the continuous rank map into five ordered strata and delineates their boundaries; this discretization is used only for illustration, whereas the model operates on continuous ranks and their local variations. Nevertheless, ordinal continuity is not equivalent to segment identity: adjacent instances may have similar depths, while a single object may span a substantial depth range. Ordinal geometry should therefore serve as a selective constraint on mask formation, rather than a substitute for semantic or panoptic supervision.

We introduce *SPOR: Semantics-Preserving Ordinal Refinement*, which incorporates ordinal geometry into three complementary stages of mask formation. *Ordinal Depth Feature Aggregation* (ODFA) regulates local propagation through ordinal affinities and combines adjacent pyramid levels through depth-conditioned gates. Its decoupled mask-support branch refines mask prediction features while retaining the unmodified feature for semantic pooling. *Query-Conditioned Ordinal Boundary Routing* (QOBR) estimates a dominant ordinal interval from each query's preceding mask prediction and uses reliable intervals to regulate uncertain boundary reads in cross-attention. It protects high-confidence cores and retains baseline routing when query support is unreliable. Finally, *Reliable Ordinal Partition Consistency* (ROPC) regularizes query-induced pixel partitions using ground-truth segment relations weighted by ordinal gaps and cross-scale agreement. ROPC operates only during training and introduces no additional inference-time computation.

Together, these components connect relative scene structure to feature propagation, query access, and partition supervision. SPOR confines direct depth-dependent modifications to the mask-formation pathway, leaving text/category embeddings, CLIP features, and the semantic pooling feature unchanged. This architectural separation, which we term semantics preservation, allows geometry to refine spatial grouping without directly perturbing the pretrained semantic representations. Our contributions are:

- We introduce SPOR, a unified framework for incorporating ordinal scene geometry into open-vocabulary panoptic mask formation while leaving the semantic feature pathway unchanged.

- We develop ordinal affinity-guided feature aggregation and query-conditioned boundary routing to regulate information flow at complementary stages of mask prediction.
- We propose a training-only, reliability-weighted partition consistency objective that combines ground-truth segment relations with ordinal geometry.

2 Related Works

2.1 Open-Vocabulary Panoptic Segmentation

Panoptic segmentation assigns a semantic category to each pixel and a distinct identity to each thing instance (Kirillov et al., 2019). MaskFormer and Mask2Former established mask-based prediction as a unified formulation for semantic, instance, and panoptic segmentation (Cheng et al., 2021; 2022). Open-vocabulary panoptic segmentation extends this formulation to categories specified by text at inference time, while retaining the distinction between thing instances and stuff regions.

Early approaches transferred vision–language representations to panoptic prediction in different ways. MaskCLIP integrates mask tokens into a pretrained CLIP visual encoder (Ding et al., 2023), whereas OPSNet uses embedding modulation to exchange information between a segmentation network and CLIP (Chen et al., 2023). ODISE combines frozen text-to-image diffusion features with discriminative vision–language features for mask prediction and classification (Xu et al., 2023a). FC-CLIP instead shares a frozen convolutional CLIP backbone between mask generation and open-vocabulary classification (Yu et al., 2023). More recently, UMG-CLIP has strengthened image-, region-, and pixel-level vision–language alignment (Shi et al., 2024); EOV-Seg has pursued efficient single-stage prediction through vocabulary-aware feature selection and spatially aware embeddings (Niu et al., 2025); and SAM-CP has composed text–patch and patch–instance affinities to produce semantic and instance predictions (Chen et al., 2025). Building on this progress, our work investigates ordinal geometry as a complementary structural prior for feature aggregation, query routing, and mask-partition consistency.

2.2 Geometry-Guided Segmentation

Depth provides structural cues complementary to visual appearance. Early approaches exploit geometric depth encodings (Gupta et al., 2014), RGB–depth feature fusion (Hazirbas et al., 2017; Lee et al., 2017), or depth-conditioned convolution and pooling (Wang & Neumann, 2018). More recently, DFormer learns transferable RGB-D representations through joint pretraining and dedicated RGB-D blocks (Yin et al., 2024). DFormerv2 further uses depth and spatial distances as geometric priors to allocate self-attention weights (Yin et al., 2025). While these studies establish geometry-aware representations and aggregation for RGB-D segmentation, our work investigates within-image ordinal relationships for open-vocabulary panoptic segmentation, guiding feature propagation, query routing, and partition consistency without directly modifying the semantic feature pathway.

3 Method

3.1 Problem Definition

Let $\mathcal{C}_{\text{train}}$ and $\mathcal{C}_{\text{test}}$ denote the training and test category sets, respectively; the latter may include categories unseen during training. Given an image $I \in \mathbb{R}^{3\times H\times W}$, an aligned monocular depth map $d \in \mathbb{R}^{H\times W}$, and a validity map $V \in \{0,1\}^{H\times W}$, the model predicts non-overlapping panoptic masks and their category labels. Panoptic annotations are provided during training. At inference, the model receives the image, the depth observation, and candidate category names from $\mathcal{C}_{\text{test}}$. Geometry is modeled only at pixels marked valid by V and having finite depth values.

3.2 Method Overview

As shown in Fig. 2, SPOR constructs a normalized ordinal rank map R from the valid depth values in each image. After sorting the depths, tied values are assigned the average of their occupied rank positions, and the ranks are normalized according to the number of valid pixels. The three modules

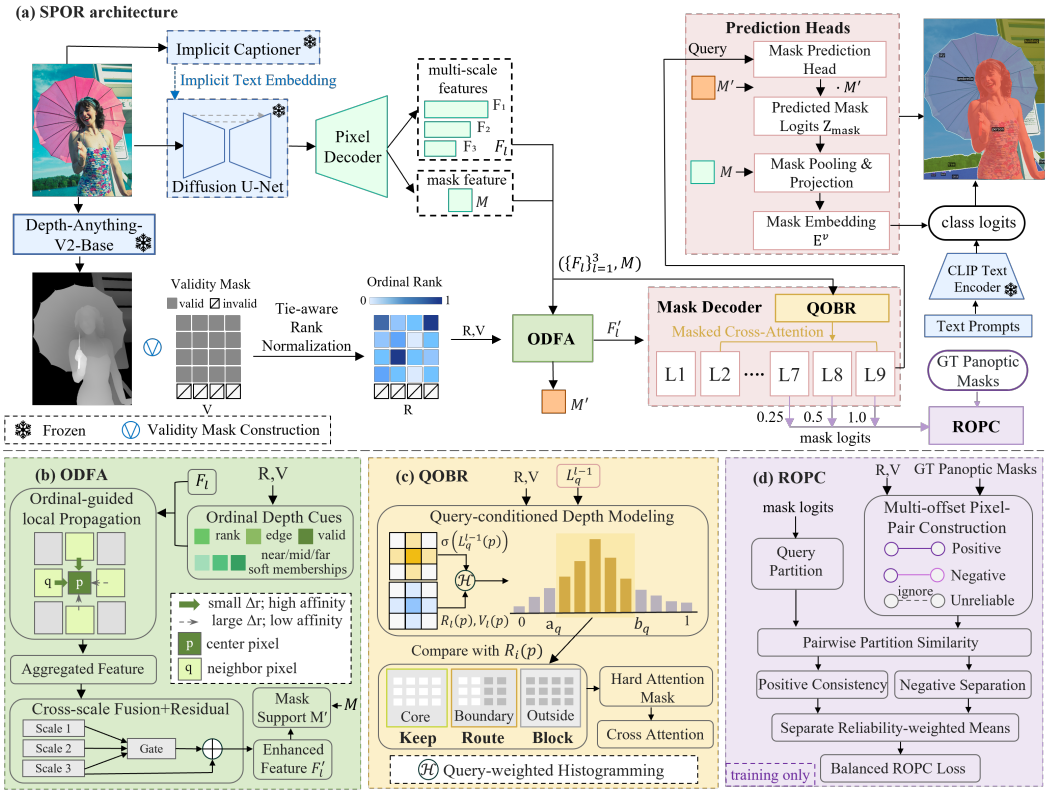

Figure 2: Overview of SPOR. (a) Overall pipeline: the RGB image and estimated monocular depth provide visual features and a shared ordinal representation for open-vocabulary panoptic segmentation. (b)–(d) ODFA, QOBR, and ROPC, respectively, introduce ordinal geometry at complementary stages of spatial mask formation.

use ordinal information defined in the same way, with validity-aware pooling at their respective resolutions. ODFA uses ordinal proximity to constrain local and cross-scale feature aggregation, and incorporates geometry into multi-scale features and mask-support features through residual updates. QOBR estimates a dominant ordinal interval for each query from the preceding decoder stage's mask logits. It then adjusts visibility in masked cross-attention, preserving the query core and routing uncertain boundary pixels. During training, ROPC combines panoptic annotations and ordinal ranks to construct reliability-weighted pixel pairs at multiple offsets, and imposes partition consistency and separation constraints on the final prediction and the last two auxiliary predictions.

3.3 Ordinal Depth Feature Aggregation

ODFA limits feature mixing across ordinal-depth boundaries. Let l index a feature-pyramid level and $F_l \in \mathbb{R}^{C\times h_l\times w_l}$ be its C-channel visual feature. Validity-aware pooling of the full-resolution rank map R and validity map V defined above yields the level-wise rank r_l and validity v_l. The depth cue D_l consists of r_l, local ordinal edges e_l, v_l, and three soft ordinal-layer memberships. We project F_l and D_l into visual and depth representations H_l and E_l, respectively.

Let p be a pixel and $u \in \mathcal{N}_4(p)$ an in-bounds pixel in its four-connected neighborhood. Their ordinal affinity is

$$a_l(p,u) = v_l(p)v_l(u)\exp\left(-\frac{|r_l(p)-r_l(u)|}{\tau}\right), \tag{1}$$

where $\tau > 0$ controls decay with ordinal distance. ODFA aggregates the projected feature as

$$\bar{H}_l(p) = \frac{\sum_{u\in\mathcal{N}_4(p)} a_l(p,u)H_l(u) + \varepsilon H_l(p)}{\sum_{u\in\mathcal{N}_4(p)} a_l(p,u) + \varepsilon}. \tag{2}$$

Here, $\varepsilon > 0$ keeps the denominator nonzero and retains the center feature when no valid neighbor is available. The same ordinal constraint is applied to spatially aligned features from adjacent levels. Let $\mathcal{S}_l$ contain level l and its available neighbors, and let $\bar{H}_{s\to l}$ denote a source feature from $s \in \mathcal{S}_l$ after alignment and aggregation at level l. Pixel-wise gates $g_{l,s}$ conditioned on H_l and E_l form $z_l(p) = \sum_{s\in\mathcal{S}_l} g_{l,s}(p)\bar{H}_{s\to l}(p)$. The visual feature is then updated by a bounded residual:

$$F_l' = F_l + \rho \tanh(P_l([H_l, z_l, E_l])) . \tag{3}$$

Brackets denote channel concatenation, P_l is a level-specific residual projection, and $\rho > 0$ bounds its magnitude. The output convolution of P_l is zero-initialized, giving $F_l' = F_l$ at initialization. A separate mask-support branch likewise refines the mask feature through a bounded residual. Semantic pooling retains the pre-ODFA visual feature as its feature input, although changes to the predicted masks may indirectly affect the pooled result.

3.4 Query-Conditioned Ordinal Boundary Routing

ODFA constrains feature aggregation, but a decoder query may still attend to pixels from ordinally incompatible regions. Starting from the second decoder layer, QOBR uses the preceding mask prediction to control pixel visibility. Let t index the current decoder layer, q a query, and $Z_q^{(t-1)}$ its preceding mask-logit map. Each operation aligns this map and the depth ranks to the required resolution. We denote the aligned rank and validity at pixel p by $r(p)$ and $v(p)$, and define $m_q(p) = \sigma(Z_q^{(t-1)}(p))$, where σ is the sigmoid function.

To estimate a query interval, QOBR constructs a 16-bin ordinal histogram at the highest-resolution feature scale using weights

$$w_q(p) = \left[\frac{m_q(p) - 0.20}{0.80}\right]_+^2 v(p), \tag{4}$$

where $[x]_+ = \max(x, 0)$. A connected interval $[a_q, b_q]$ is expanded from the dominant bin toward a target of 70% of the weighted foreground mass, with a maximum width of 0.35. Mask confidence, valid pixel support, depth coverage, and histogram concentration yield a reliability score $c_q \in [0, 1]$; unreliable queries retain the original attention routing.

At the current cross-attention resolution, let $\delta_q(p)$ be the distance from $r(p)$ to $[a_q, b_q]$, and let $\gamma_q = \mathrm{clip}(b_q - a_q + 0.05, 0.08, 0.35)$ be the transition width. Here, $\mathrm{clip}(x, a, b)$ restricts x to $[a, b]$. Ordinal compatibility is

$$B_q(p) = v(p)\,\mathrm{clip}\left(1 - \frac{\delta_q(p)}{\gamma_q}, -1.5, 1\right). \tag{5}$$

Invalid-depth pixels thus receive no ordinal correction. Pixels with preceding mask probability $m_q(p) \geq 0.70$ remain visible, whereas those with $m_q(p) < 0.20$ remain masked. For the remaining pixels, QOBR computes

$$\widetilde{Z}_q(p) = Z_q^{(t-1)}(p) + \eta\lambda_t c_q B_q(p), \qquad \eta = 0.50, \tag{6}$$

where λ_t is a layer-dependent routing strength with an activation warm-up. A candidate pixel is visible to the query only if $\sigma(\widetilde{Z}_q(p)) \geq 0.5$. The routing score constructs the cross-attention mask; it does not directly overwrite predicted mask or class logits.

3.5 Reliable Ordinal Partition Consistency

ROPC regularizes the predicted partition during training. Let j index a decoder stage, Q be the number of queries, and $M_i^{(j)}(p)$ be the mask logit of query i at pixel p. After appending a fixed zero logit for the void state, a softmax with temperature $T > 0$ gives

$$\Pi^{(j)}(p) = \mathrm{softmax}\left([M_1^{(j)}(p), \ldots, M_Q^{(j)}(p), 0]/T\right). \tag{7}$$

Here, $\Pi_i^{(j)}(p)$ is the soft assignment of pixel p to query i. The non-void same-query agreement between pixels p and u is

$$K_{pu}^{(j)} = \sum_{i=1}^{Q} \Pi_i^{(j)}(p)\Pi_i^{(j)}(u). \tag{8}$$

Let $G(p)$ be the ground-truth panoptic segment ID at pixel p. We sample pixel pairs at multiple spatial offsets with valid annotations and valid depth at both the current and a coarser scale. Pairs from the same segment form E^+; pairs from different segments form E^-. Let $\Delta_{pu} = |r(p) - r(u)|$ be the current-scale ordinal gap and Δ^c_{pu} its coarser-scale counterpart. Their agreement defines reliability $c_{pu} = \exp(-|\Delta_{pu} - \Delta^c_{pu}|/\sigma)$, where $\sigma > 0$ controls sensitivity. The pair weights are $w^+_{pu} = c_{pu} \exp(-\Delta_{pu}/\tau_+)$ for same-segment pairs and $w^-_{pu} = c_{pu}[1 - \exp(-\Delta_{pu}/\tau_-)]$ for different-segment pairs, where $\tau_+, \tau_- > 0$ control their responses to the ordinal gap.

Let $s \in \{+, -\}$ index pair type, with $\ell_+(K) = 1 - K$ and $\ell_-(K) = K$ penalizing separation of same-segment pixels and merging of different-segment pixels, respectively. Define $W_s = \max(\sum_{(p,u)\in E^s} w^s_{pu}, \varepsilon)$, where $\varepsilon > 0$ prevents a zero denominator. Both terms combine into one stage loss:

$$\mathcal{L}^{(j)}_{\mathrm{ROPC}} = \frac{1}{2} \sum_{s\in\{+,-\}} \frac{1}{W_s} \sum_{(p,u)\in E^s} w^s_{pu}\, \ell_s\Big(K^{(j)}_{pu}\Big). \tag{9}$$

At most 4096 pairs of each type are sampled per image and stage. The loss is applied to the final prediction and the last two auxiliary predictions with weights 1, 0.5, and 0.25, respectively. Its overall weight reaches 0.05 at iteration 500 and 0.1 at iteration 2000.

4 Experiments

4.1 Implementation Details

4.1.1 Datasets

We train SPOR on the COCO 2017 panoptic training set (Lin et al., 2014), which contains approximately 118K images and 133 panoptic categories, including 80 thing and 53 stuff categories. Each RGB image is paired with a monocular pseudo-depth map generated offline using Depth Anything V2-Base (Yang et al., 2024), which contains 97.5M parameters. The pseudo-depth values are used only to derive within-image ordinal relations and are never treated as metric depth supervision.

We report in-domain performance on COCO and evaluate cross-dataset open-vocabulary generalization on ADE20K (Zhou et al., 2017), PASCAL VOC (Everingham et al., 2010), and PASCAL Context (Mottaghi et al., 2014). ADE20K is evaluated under two label spaces: A-150, containing 150 common classes, and A-847, containing all 847 classes. PASCAL Context is evaluated using PC-59 with 59 common classes and PC-459 with all 459 classes. For PASCAL VOC, we adopt the PAS-21 setting, which contains 20 foreground classes and one background class.

4.1.2 Architecture

SPOR builds upon ODISE (Xu et al., 2023a) by introducing ordinal geometric reasoning while preserving its open-vocabulary semantic pathway. Given an input image, the visual backbone extracts multi-scale diffusion features at timestep $t = 0$, which are subsequently decoded by Mask2Former into 100 query masks. SPOR converts the image-aligned monocular pseudo-depth map into a within-image ordinal representation and uses it to guide feature aggregation, query routing, and mask-partition consistency. The ordinal representation is restricted to this structural prediction pathway and does not modify the CLIP embedding space or provide category-level supervision.

4.1.3 Training Strategy

We initialize SPOR from the official label-supervised ODISE checkpoint and fine-tune it on COCO 2017 for 6,000 iterations. Training is conducted on eight NVIDIA A100-SXM4-80GB GPUs with a total batch size of 64. We employ AdamW (Loshchilov & Hutter, 2019) with a base learning rate of 1×10^{-4} and a weight decay of 0.05.

Our training objective retains the original ODISE losses and additionally includes ROPC, which regularizes the query-induced pixel partition using reliable ordinal-depth relations. ROPC is applied to the final three decoder stages with relative weights of 0.25, 0.5, and 1.0, respectively.

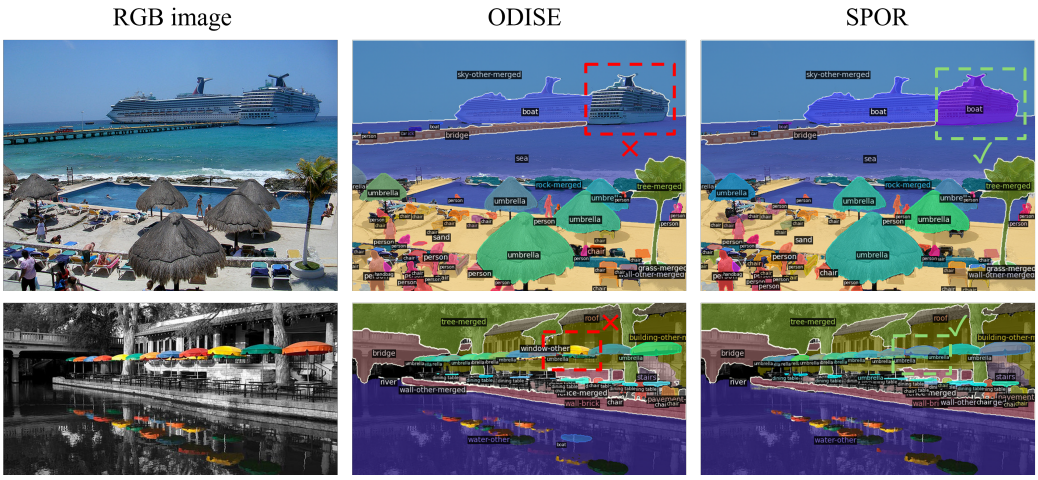

Figure 3: Qualitative comparison of ODISE and SPOR in complex scenes. The red dashed boxes highlight objects that ODISE fails to segment correctly, while the green dashed boxes show the corresponding segmentation results produced by SPOR.

4.2 Main Results

4.2.1 Open-Vocabulary Panoptic Segmentation

We follow the standard cross-dataset evaluation protocol by training SPOR on COCO Panoptic and directly evaluating it on ADE20K in a zero-shot manner. As shown in Table 1, SPOR achieves 55.53 PQ, 46.22 AP, and 65.33 mIoU on COCO, outperforming ODISE by 0.1, 0.2, and 0.1 points, respectively. The qualitative comparisons in Figure 3 further illustrate this advantage in complex scenes. ODISE tends to produce incomplete or incorrect predictions for large objects and structurally ambiguous regions, whereas SPOR generates more complete object masks and more coherent region partitions. Together, the quantitative and qualitative results suggest that ordinal geometry improves the spatial organization of diffusion features and the reliability of panoptic partitioning while preserving open-vocabulary semantic recognition.

On ADE20K, SPOR improves PQ by 0.2 points over ODISE, while AP remains comparable and mIoU decreases slightly. This performance profile is consistent with the structure-oriented design of SPOR: ordinal geometry operates on feature propagation, query routing, and mask partitioning without directly modifying open-vocabulary category alignment. Its primary benefit is therefore reflected in PQ, which jointly evaluates recognition and partition quality. MasQCLIP achieves higher PQ and mIoU on ADE20K, indicating that cross-dataset recognition remains strongly influenced by the underlying vision–language representation. In contrast, SPOR addresses the underexplored role of geometric structure in diffusion-based open-vocabulary segmentation, improving PQ on both COCO and ADE20K while largely retaining the original cross-dataset semantic transfer capability.

4.2.2 Open-Vocabulary Semantic Segmentation

Table 2 compares SPOR with existing methods on five open-vocabulary semantic segmentation benchmarks. SPOR achieves 11.1, 15.2, 29.6, 56.9, and 84.9 mIoU on A-847, PC-459, A-150, PC-59, and PAS-21, respectively. Compared with ODISE, SPOR improves PC-459 and PAS-21 by 0.7 and 0.3 points, respectively, while remaining competitive on the other three datasets. In particular, SPOR attains the best result on PAS-21, demonstrating that the introduction of ordinal structure preserves the recognition of common semantic categories. The improvements on both PC-459 and PAS-21, which differ substantially in label-space granularity, further suggest that the benefit of structural modeling is not restricted to a specific vocabulary scale.

SPOR incorporates ordinal geometry into feature propagation, query routing, and mask formation while retaining the original semantic recognition backbone. These experiments therefore examine whether the resulting structural enhancement can improve region-level reasoning without disrupting

Table 1: Open-vocabulary panoptic segmentation results on COCO and ADE20K. All values are reported in percentage (%). Higher is better. blue, purple, and orange cells mark the first-, second-, and third-highest reported values in each column, respectively.

Method	COCO			ADE20K		
	PQ	AP	mIoU	PQ	AP	mIoU
MaskCLIP (Ding et al., 2023)	–	–	–	15.1	6.0	23.7
FreeSeg (Qin et al., 2023)	–	–	–	16.3	6.5	24.6
ODISE (Xu et al., 2023a)	55.4	46.0	65.2	22.6	14.4	29.9
OPSNet (Chen et al., 2023)	57.9	–	–	19.0	–	–
MasQCLIP (Xu et al., 2023b)	48.5	–	62.0	23.3	–	30.4
EOV-Seg-M (Niu et al., 2025)	–	–	–	18.7	8.5	25.5
SPOR (Ours)	55.5	46.2	65.3	22.8	14.4	29.6

Table 2: Open-vocabulary semantic segmentation comparison on five datasets. All values are mIoU (%). "–" denotes an officially unreported result. blue, purple, and orange cells mark the first-, second-, and third-highest reported values in each column, respectively.

Method	A-847	PC-459	A-150	PC-59	PAS-21
SPNet (Xian et al., 2019)	–	–	–	24.3	18.3
GroupViT (Xu et al., 2022a)	4.3	4.9	10.6	25.9	50.7
OpenSeg (Ghiasi et al., 2022)	6.3	9.0	21.1	42.1	–
LSeg+ (Ghiasi et al., 2022)	3.8	7.8	18.0	46.5	–
SimBaseline (Xu et al., 2022b)	–	–	15.3	–	74.5
ZegFormer (Ding et al., 2022)	–	–	16.4	–	73.3
OVSeg (Liang et al., 2023)	9.0	12.4	29.6	55.7	–
DeOP (Han et al., 2023)	7.1	9.4	22.9	48.8	–
MaskCLIP (Ding et al., 2023)	8.2	10.0	23.7	45.9	–
ODISE (Xu et al., 2023a)	11.1	14.5	29.9	57.3	84.6
EOV-Seg-M (Niu et al., 2025)	7.8	12.2	25.5	51.8	–
SPOR (Ours)	11.1	15.2	29.6	56.9	84.9

open-vocabulary category alignment. The competitive results across all five benchmarks, together with the improvements on PC-459 and PAS-21, indicate that ordinal geometry provides complementary structural cues rather than overwriting the semantic knowledge encoded by the pretrained representation. Combined with the panoptic segmentation results, these findings show that SPOR improves the structural modeling of diffusion-based segmentation while retaining robust cross-dataset semantic transfer.

4.3 Ablation Study

Component-wise analysis. We conduct a cumulative ablation study by progressively introducing ODFA, QOBR, and ROPC. As shown in Table 3, compared with the variant without any of the three components, the complete model improves PQ/AP by 0.38/0.30 points on COCO and by 0.48/0.34 points on ADE20K, while mIoU remains largely unchanged. The consistent improvements in PQ and AP, in contrast to the minor variations in mIoU, indicate that the performance gains primarily arise from more accurate region grouping, instance discrimination, and mask prediction, rather than broad changes in category-level semantic recognition. The progressive ablation results further show that ODFA first improves the feature representations used for region prediction, QOBR subsequently enhances the correspondence between object queries and image regions, and ROPC further reduces inconsistencies among region predictions. The three components therefore complement one another throughout the region prediction process, resulting in more pronounced gains in PQ and AP.

The incremental gains introduced by the individual components are relatively moderate because all variants share the same pretrained vision–language backbone and segmentation architecture, which already provide strong semantic representations. The proposed components perform structured

Table 3: Component ablation of SPOR on COCO and ADE20K.

	Components			COCO			ADE20K		
Configuration	ODFA	QOBR	ROPC	PQ	AP	mIoU	PQ	AP	mIoU
SPOR w/o ODFA, QOBR, ROPC	–	–	–	55.15	45.92	**65.36**	22.29	14.01	**29.70**
SPOR w/o QOBR, ROPC	✓	–	–	55.32	46.03	65.36	<u>22.58</u>	14.12	<u>29.63</u>
SPOR w/o ROPC	✓	✓	–	<u>55.41</u>	<u>46.18</u>	65.31	**22.77**	<u>14.13</u>	29.61
SPOR	✓	✓	✓	**55.53**	**46.22**	<u>65.33</u>	**22.77**	**14.35**	<u>29.63</u>

'w/o' denotes 'without.' The best results are highlighted in bold, and the second-best results are underlined.

Table 4: Component ablation of SPOR on five open-vocabulary semantic segmentation datasets. All results are reported in mIoU (%).

	Components			Open-Vocabulary Semantic Segmentation				
Configuration	ODFA	QOBR	ROPC	A-847	PC-459	A-150	PC-59	PAS-21
SPOR w/o ODFA, QOBR, ROPC	–	–	–	<u>11.13</u>	14.09	**29.70**	56.75	<u>84.80</u>
SPOR w/o QOBR, ROPC	✓	–	–	**11.26**	14.68	<u>29.63</u>	56.82	84.62
SPOR w/o ROPC	✓	✓	–	11.00	<u>14.74</u>	29.61	**56.97**	84.75
SPOR	✓	✓	✓	11.06	**15.19**	<u>29.63</u>	<u>56.89</u>	**84.87**

'w/o' denotes 'without.' The best results are highlighted in bold, and the second-best results are underlined.

refinement over these established representations rather than relearning the underlying semantic knowledge. Therefore, despite the moderate absolute gains, the consistent improvements across multiple region-level metrics provide clear evidence of the effectiveness of the individual components.

As shown in Table 4, the most notable cross-dataset improvement is observed on PC-459, where the complete model outperforms the variant without any of the three components by 1.10 mIoU, while the results on the other four benchmarks remain broadly comparable. Given the differences among these benchmarks in label granularity, category distribution, and image domain, the concentrated gain on PC-459 suggests that the improvements in region representation and region–semantic correspondence are particularly effective under this evaluation setting. Meanwhile, the stable results on the remaining benchmarks indicate that this improvement is not achieved at the expense of performance in other evaluation scenarios. Overall, our method exhibits stable cross-dataset transfer across diverse open-vocabulary benchmarks.

5 Conclusion

We propose SPOR, a semantics-preserving ordinal refinement framework for open-vocabulary panoptic segmentation. SPOR uses within-image depth ordering to guide mask formation without directly modifying pretrained semantic embeddings or the visual features used for semantic pooling. Experiments across panoptic and semantic segmentation benchmarks show that SPOR improves the spatial partitioning of panoptic masks while retaining cross-dataset open-vocabulary semantic transfer. These findings indicate that relative depth can provide spatial constraints complementary to appearance and semantic cues in open-vocabulary panoptic segmentation.

AI use statement

In this work, we used generative AI tools to assist with searching for and summarizing relevant literature, translating and editing portions of the manuscript, and assessing the clarity of the method description. We reviewed all AI-assisted suggestions and checked the references, technical descriptions, and equations included in the paper. We take full responsibility for the final content, including AI-assisted text and claims.

REFERENCES

Pengfei Chen, Lingxi Xie, Xinyue Huo, Xuehui Yu, Xiaopeng Zhang, Yingfei Sun, Zhenjun Han, and Qi Tian. SAM-CP: Marrying SAM with composable prompts for versatile segmentation. In *International Conference on Learning Representations*, 2025.

Xi Chen, Shuang Li, Ser-Nam Lim, Antonio Torralba, and Hengshuang Zhao. Open-vocabulary panoptic segmentation with embedding modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1141–1150, October 2023.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11583–11592, 2022.

Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8090–8102. PMLR, 2023.

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

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision – ECCV 2022*, pp. 540–557, 2022.

Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, 2014.

Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1086–1096, 2023.

Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Computer Vision – ACCV 2016*, pp. 213–228. Springer, 2017.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Seungyong Lee, Seong-Jin Park, and Ki-Sang Hong. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Roozbeh Mottaghi, Xiyang Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

Hongwei Niu, Jie Hu, Jianghang Lin, Guannan Jiang, and Shengchuan Zhang. Eov-seg: Efficient open-vocabulary panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6254–6262, 2025. doi: 10.1609/AAAI.V39I6.32669.

Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Yuxi Ren, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, and Xingang Wang. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19446–19455, June 2023. doi: 10.1109/CVPR52729.2023.01863.

Bowen Shi, Peisen Zhao, Zichen Wang, Yuhang Zhang, Yaoming Wang, Jin Li, Wenrui Dai, Junni Zou, Hongkai Xiong, Qi Tian, and Xiaopeng Zhang. UMG-CLIP: A unified multi-granularity vision generalist for open-world understanding. In *European Conference on Computer Vision*, pp. 259–277, 2024.

Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D segmentation. In *Proceedings of the European Conference on Computer Vision*, 2018.

Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero- and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8256–8265, 2019.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022a.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision – ECCV 2022*, pp. 736–753, 2022b.

Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 887–898, October 2023b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

Bo-Wen Yin, Jiao-Long Cao, Ming-Ming Cheng, and Qibin Hou. DFormerv2: Geometry self-attention for RGBD semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19345–19355, 2025.

Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. DFormer: Rethinking RGBD representation learning for semantic segmentation. In *International Conference on Learning Representations*, 2024.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In *Advances in Neural Information Processing Systems*, 2023.

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.

A APPENDIX

A.1 CODE AVAILABILITY

The source code for SPOR is available at the following anonymous repository:

`https://anonymous.4open.science/r/SPOR-C4A9`