Method	Spatial Prior	Reference Coordinate	Target Prediction	Cross-Attn.	Reference Prior Update	Discriminative PE
DETR	No	No	$[cx, cy, w, h]$	Standard		✓
Deformable DETR	Implicit	4D	$[dcx, dcy, w, h]$	Deformable Points		✓
SMCA-DETR	Implicit	4D	$[\Delta cx, \Delta cy, w, h]$	Gaussian Points		
Conditional DETR	Implicit	2D	$[\Delta cx, \Delta cy, w, h]$	Conditional		
Anchor DETR	Explicit	2D	$[\Delta cx, \Delta cy, w, h]$	Standard		✓
DAB-DETR	Explicit	4D	$[\Delta cx, \Delta cy, \Delta w, \Delta h]$	Conditional	✓	
SAM-DETR-w/SMCA	Explicit	4D	$[\Delta cx, \Delta cy, \Delta w, \Delta h]$	Gaussian Points	✓	
SAP-DETR (Ours)	Explicit	2D+4D	$[\Delta x, \Delta y, \Delta \ell, \Delta t, \Delta r, \Delta b]$	Conditional Side	✓	✓

Table 7. Comparison of DETR-like models and our proposed SAP-DETR.

Appendix

A. Comparison of DETR Family

Tab. 7 detailedly compares various representative properties for the DETR family. DETR [2] follows the vanilla Transformer structure and leverages the learnable positional encodings to help Transformer distinguish paralleled input queries. However, such learnable positional encodings without any spatial prior help severely affect the convergency speed of the Transformer detector. To this end, the mainstream approaches make effort to introduce different spatial prior into DETR, which can be divided into implicit and explicit methods. Specifically, the former decouples reference coordinates from the learnable positional encodings, while the latter directly sets a 2D/4D coordinate for each query and maps such low-dimensional coordinate into a high dimension positional encoding via the sinusoidal PE [24].

From the perspective of the spatial prior indoctrination, a straightforward way for object query is to predict the offset between their reference and the target bounding boxes. For example, previous approaches [5, 16, 25] only regress the offset of center points, while the current approaches [13, 26] directly regress the 4D offset based on the reference coordinate. Another spatial prior indoctrination benefits from the redesign of the cross-attention mechanism. Deformable DETR [31], SMCA [5], and SAM-DETR [26] aggregate multiple extreme point regions from the content features by directly predicting the coordinates of these points from the object queries. Conditional DETR [16] and DAB-DETR [13] utilize a Gaussian-like positional cross-attention map to attend to distinct regions dynamically. Take a close insight at the Gaussian map, the region of box sides and center point are attended by different heads in the multi-head attention mechanism. From the perspective of the spatial prior update, the prevailing approaches [13, 26] apply a cascaded way to refine the box prediction as well as update the reference spatial prior. *However, all of these methods view center points as the reference spatial prior, eroding the discrimination of the positional encodings during performing the redundant prediction, thereby confusing the Transformer detector as well as leading to the slow model convergency.*

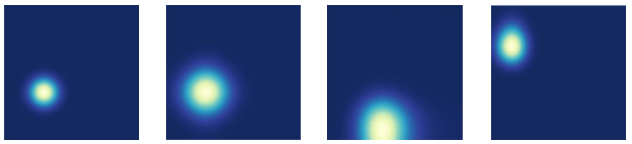
In our proposed SAP-DETR, such confusing reference spatial prior is replaced by the query-specific reference point. Specifically, each object query in SAP-DETR is assigned a non-overlapping fixed grid-region, which prompts queries to consider the grid area as a salient region to attend to image features and compensate for the over-smooth/inadequacy during center-based detection by localizing each side of the bounding box layer by layer. Considering the sparseness of the reference points, the movable strategy is proposed to enhance small/slender object detection. Therefore, there exists the 2D+4D reference spatial prior in the proposed SAP-DETR, and the final prediction is based on such a 6D reference coordinates ($[\Delta x, \Delta y, \Delta\, \Delta t, \Delta r, \Delta b]$). Taking an insight into the Conditional attention mechanism, we investigate that the highlight region is most relevant to four sides of bounding boxes, hence facilitating the final box localization. More intuitively, we devise the PECA to indicate the location of bounding box sides to object queries, where they should attend from context image features.

B. Temperature Consistency in PE

Following DETR, we also use the 2D sinusoidal function $\mathrm{PE}(x, y)$ as positional encoding. Given a position, the $\mathrm{PE}_{\mathrm{pos}}$ is calculated by

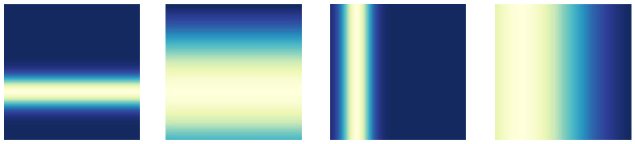
$$\mathrm{PE}^{T}_{\mathrm{pos}}(i) = \begin{cases} \sin(\mathrm{pos} \cdot \omega_t) & i = 2t \\ \cos(\mathrm{pos} \cdot \omega_t) & i = 2t+1, \end{cases} \qquad (6)$$
$$\omega_t = T^{-2t/d}, \quad t = 1, \cdots, d/2,$$

where T is an adjustable temperature and i is the channel index of the positional embedding. As shown in Fig. 5, the receptive field size of the positional attention map tends to become wider with increasing temperature [13]. Before the softmax operation, the positional query-to-key similarity $\mathbf{A}$ in the cross-attention mechanism is computed by a dot-product between query position $\mathrm{PE}^{T_q}_{\mathrm{pos}_q}$ and key position $\mathrm{PE}^{T_k}_{\mathrm{pos}_k}$. Clearly, the resulting positional similarity in Fig. 5(a) and (b) subjects to a Gaussian-like distribution. We fix the



(a) T_q=T_k=20 (b) T_q=T_k=10^4 (c) T_q=20, T_k=10^4 (d) T_q=10^4, T_k=20

Figure 5. Positional attention maps. Given two sequential PE of query-key pairs, we fix one PE of the query, reshape its sequential attention map for all PE of the key into original 2D image size.



(a) Head1 (b) Head3 (c) Head5 (d) Head7

Figure 6. Positional attention maps in each head.

$\mathrm{PE}_{\mathrm{pos}_q}^{T_q}$ and then the center of $\mathbf{A}$ is calculated by

$$\begin{aligned}\mathbf{A}_{\mathrm{PE}} &= \mathrm{PE}_{\mathrm{pos}_{\mathrm{q}}}^{T_{\mathrm{q}}} \cdot \mathrm{PE}_{\mathrm{pos}_{\mathrm{k}}}^{T_{\mathrm{k}}\top} \\ &= \sum_{t}^{d/2} \sin(\omega_t^{\mathrm{q}}\mathrm{pos}_{\mathrm{q}})\sin(\omega_t^{\mathrm{k}}\mathrm{pos}_{\mathrm{k}}) + \cos(\omega_t^{\mathrm{q}}\mathrm{pos}_{\mathrm{q}})\cos(\omega_t^{\mathrm{k}}\mathrm{pos}_{\mathrm{k}}) \\ &= \sum_{k}^{d/2} \cos(\omega_t^{\mathrm{q}}\mathrm{pos}_{\mathrm{q}} - \omega_t^{\mathrm{k}}\mathrm{pos}_{\mathrm{k}}), \quad \mathrm{pos}_{\mathrm{q}}, \mathrm{pos}_{\mathrm{k}} \in \mathfrak{X} = [0, \frac{\pi}{2}].\end{aligned} \tag{7}$$

By fixing the $\mathrm{pos}_{\mathrm{q}}$, the center $\mathrm{pos}_{\mathrm{k}}^{\mathrm{center}(k)}$ of $\mathbf{A}_{\mathrm{PE}}$ of each dimension $k \in \{1, \cdots, d/2\}$ is calculated by

$$\begin{aligned}\mathrm{pos}_{\mathrm{k}}^{\mathrm{center}(t)} &= \underset{\mathrm{pos}_{\mathrm{k}} \in \mathfrak{X}}{\mathrm{argmax}}(\cos(\omega_t^{\mathrm{q}}\mathrm{pos}_{\mathrm{q}} - \omega_t^{\mathrm{k}}\mathrm{pos}_{\mathrm{k}})) \\ &= \underset{\mathrm{pos}_{\mathrm{k}} \in \mathfrak{X}}{\mathrm{argmin}}(\omega_t^{\mathrm{q}}\mathrm{pos}_{\mathrm{q}} - \omega_t^{\mathrm{k}}\mathrm{pos}_{\mathrm{k}}) \\ &= (T_{\mathrm{k}}/T_{\mathrm{q}})^{2t/d}\mathrm{pos}_{\mathrm{q}}.\end{aligned} \tag{8}$$

Consequently, there exists an offset center for each channel of the positional attention map if $T_{\mathrm{k}} \neq T_{\mathrm{q}}$. Literally, each channel of the positional attention map can be viewed as a superposition by several horizontal and vertical line masks (see Fig. 6). So it is easy to illustrate the offset center and irregular width/height of the positional attention maps as shown in Fig. 5(c) and (d).

Without loss of generality, we eliminate the effect of conditional scaling transformation and fix the temperature of encoder's positional encoding to 20. As shown in Tab. 9, the reported results compare the different temperature settings based on PECA. Clearly, both point and box site positional encodings are benefit from a relative small consistent temperature, especially when concatenating with box side PE.

C. Scaling Transformation for PE

Revisiting Conditional Spatial Query Prediction. Given a set of content queries and their corresponding reference points, the conditional spatial query prediction adaptively maps the reference points into high-dimensional positional embeddings according to a spatial transformation generated by content queries. Let $\boldsymbol{r}^{\mathrm{q}} \in \mathbb{R}^{\mathrm{k}}$ denotes the 2D unnormalized reference point, $\mathbf{e} \in \mathbb{R}^d$ denotes the content query, and $\mathbf{T} \in \mathbb{R}^d$ indexes the scaling spatial transformation where d is the query dimension. Then the conditional spatial query prediction is calculated by

$$\mathbf{p}_{\mathrm{q}} = \mathbf{T} \cdot \mathrm{PE}(\mathrm{sigmoid}(\boldsymbol{r}^{\mathrm{q}})), \quad \mathbf{T} = \mathrm{FFN}(\mathbf{e}), \tag{9}$$

where FFN is a feed-forward network consisting of a linear layer, a ReLU activation, and a linear layer. PE is the sinusoidal positional encoding as illustrated in Eq. (6). In Conditional DETR [16], the unnormalized reference point is either a learnable 2D coordinate or generated by its corresponding content query.

Scaling Transformation in PECA. As introduced in Section Appendix B, the proposed PECA concatenates both point and box side PEs for conditional spatial cross-attention. Following the scaling transformation of Conditional DETR, we also conduct ablations on different ways of scaling transformation in PECA. The following settings are involved:

- Comparing the effectiveness of scaling transformation with and without box side PE concatenation.
- Comparing the effectiveness of scaling consistency in both point PE and box side PE, and then considering three types of ablation: no scaling, shared, and independent scaling transformation.
- Exploiting a learnable diagonal matrix to transform the positional encoding of the key-vector, which also can be shared between point PE and box side PE.

Tab. 8 summarizes the results of the ablation study on the 3-layer encoder-decoder Transformer neck. There exists a large gap between the performances of the model with no key-vector scaling transformation and counterparts with the transformation. We speculate that the scaling transformation of key-vector PE may cause the decoder confusion in extreme region localization, while the transformation on query-vector PE (point or box side) would facilitate it to focus on the spatial information within the content embeddings to the content image features. In addition, we observe that the main function of the point PE is to keep reference-specific for each query, and its effectiveness on box side attention will be weakened when concatenating the box side PE. Finally, we use a shared scaling transformation for both point and box side PEs. More visualization on PECA without the scaling transformation $\mathbf{T}$ please see our journal version.

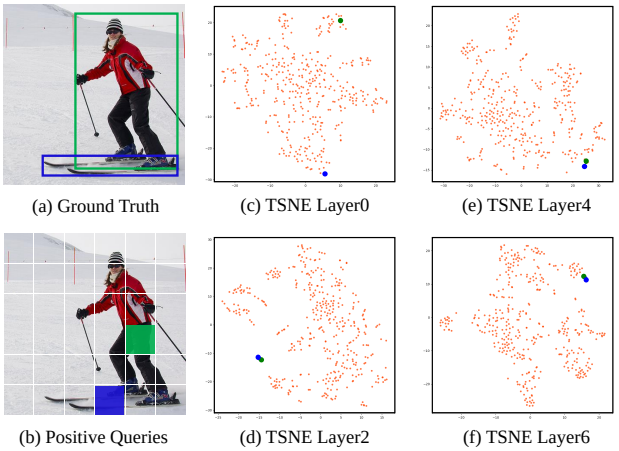
Concatenate Box Side PE	Scaling Transformation for PE			AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	$\mathbf{T}_k$	$\mathbf{T}_{qp}$	$\mathbf{T}_{qb}$						
✗	○	○	-	33.3	54.4	33.9	13.1	36.3	52.2
✗	●	○	-	32.3	53.4	32.7	12.8	35.4	50.4
✗	○	●	-	34.4	55.3	35.6	14.5	37.7	53.3
✗	●	●	-	32.6	53.7	33.1	12.2	35.5	51.1
✓	○	○	○	34.0	54.5	35.0	13.9	37.1	52.8
✓	●	○	○	33.2	54.0	34.0	13.3	36.4	51.7
✓	○	●	○	34.7	55.3	35.8	14.4	37.8	53.1
✓	○	○	●	35.1	55.1	36.7	14.9	38.2	53.5
✓	●	●	●	32.6	53.4	33.2	12.3	35.6	52.0
✓	○	◐	◑	35.2	55.1	36.6	15.7	38.5	**53.9**
✓	○	●	●	**35.2**	**55.4**	**36.8**	**15.8**	**38.5**	53.6

◐ and ◑ denote different independent scaling transformations.
○ and ● denote no scaling and shared scaling transformations, respectively.

Table 8. Ablation study on the scaling transformation of PE.

Concatenate Box Side PE	Temperature of PE			AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	T_k	T_{qp}	T_{qb}						
✗	20	1000	-	31.8	52.8	32.1	12.8	34.4	50.5
✗	1000	20	-	32.1	53.0	32.5	12.7	35.2	50.6
✗	20	20	-	32.2	53.2	32.7	12.7	35.0	51.0
✓	20	1000	1000	32.2	52.9	32.9	12.8	35.1	51.1
✓	1000	20	20	32.3	52.7	32.8	13.3	35.2	50.8
✓	20	20	20	**33.0**	**53.6**	**33.5**	**13.7**	**36.3**	**52.1**

Table 9. Ablation Study on the temperature consistency of PE.



(a) Ground Truth (c) TSNE Layer0 (e) TSNE Layer4

(b) Positive Queries (d) TSNE Layer2 (f) TSNE Layer6

Figure 7. Visualization of t-SNE. Both grids and slots in t-SNE represent object queries, where the green and blue color are the positive queries, corresponding to the same colored ground truth.

D. Independent Prediction Heads

Taking a close insight into the semantic representation of these object queries, we map each query output into a 2D distribution via t-SNE [23]. As shown in Fig. 7, each dot here represents an query output from the decoder layer. It can be seen that the instance objects (blue and green dots in Fig. 7(c)-(f)) whose location at the edge/corner of the distribution are easy to distinguish from the background queries. More precisely, the instance objects, except from the first decoder layer, are at a closer distance than the semantic-close queries. Inspired by this, we employ a dedicated classification head for the first decoder layer and a shared head for the others in the auxiliary training process.

Tab. 10 reports the ablation study on the 3-layer encoder-decoder decoder neck. As we can see, the detach operation generally boosts the detector performance by ∼0.3%AP, and the independent box prediction head is conducive to the Transformer detector for further improvements. Moreover, There exists a slight performance drop when using the independent classification prediction head.

E. Movable Reference Points

We evaluate two types of training strategies for reference points. As illustrated in Fig. 8(a), we tile the mesh-grid reference points for their initialization and set such coordinates as fixed/learnable parameters. By visualizing the learnable reference points in Fig. 8(b), their distribution are observed to be uniform within the image, similar to the learnable anchor points in Anchor DETR [25]. It indicates that the learnable reference coordinates would not be affected by properties of the target regression. We further hypothesize that there exists partial denominators between salient points and the center anchor points, to a certain extent.

As introduced in Sec. 3.1, the proposed movable reference points significantly facilitate detecting small and slender objects, which are omitted caused by the sparseness of the reference point distribution. The experiments in Sec. 4.2 demonstrate that the performance of small object detection is prompted after applying the movable strategy. Dialectically,

Detach	Indep. Prediction Head		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	$Head_{cls}$	$Head_{bbox}$						
✗			34.6	54.8	35.7	14.4	37.6	52.8
✗	✓		34.7	54.8	36.0	**16.2**	37.6	53.2
✗		✓	34.8	55.0	35.8	15.3	37.8	53.5
✗	✓	✓	34.7	54.8	36.1	14.5	38.1	52.7
✓			35.0	55.1	36.5	15.6	38.3	53.0
✓	✓		34.6	55.1	35.9	14.6	37.9	52.1
✓		✓	**35.2**	**55.4**	**36.8**	15.8	**38.5**	53.6
✓	✓	✓	35.0	55.2	36.3	14.7	38.2	**54.0**

Table 10. Ablation study on the independent prediction head.

we conduct another ablation on the number of queries to verify that such a vulnerability is attribute to the query sparsity. Fig. 9 describes the performance histogram of 3-layer detectors based on both 12-epoch and 36-epoch training schemes. Without the help of the movable component, the standard SAP-DETR relatively benefits more from the query number growth compared to the counterpart. Along with query number increase, the performance gap is reduced progressively (from 1.2 AP to 0.2 AP), which further verifies our sparsity analysis and the effectiveness of the movable strategy.

To further demonstrate the effectiveness of the movable strategy, the update processes of salient points are plotted in Fig. 10 and Fig. 11. Indeed, some small and slender objects can be localized well after moving the reference point within the objects. However, some queries whose reference points are located within the large objects behave an unstable matching result that the matched queries in the latter layers are inconsistent with the previous layers. Hence there exists a slight performance deterioration for large object detection after adding the movable reference points.

F. Training Details and More Configurations

Warm Up Training Strategy. In the early training process, the bipartite matching in Transformer detectors may appear to be fragile and instable, where the positive label are assigned to one false prediction. This phenomenon is also reported in DN-DETR [9]. Following the conventional training strategy, we conduct a warm-up step during the early training process. In our experiments, we set warm-up steps to 400 and 1000 iterations for 3-layer and 6-layer Encoder-Decoder Transformer detectors, respectively.

Detailed Configurations. We list the all configurations in Tab. 11. For each number of query in Appendix E, the batch size of 8 is applied in our 3-layer SAP-DETR.

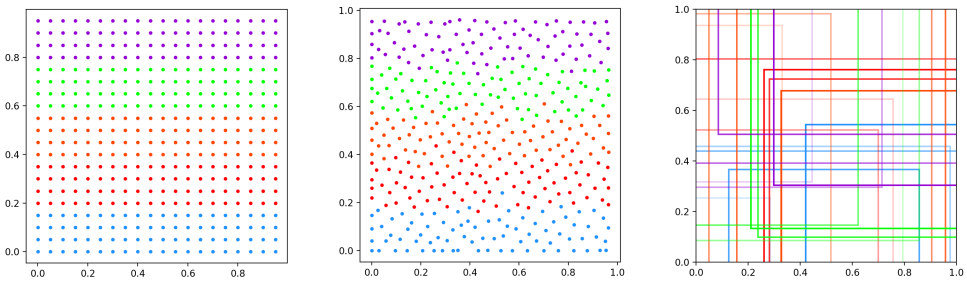
Item	Value	Item	Value
lr	1e-4	mask_loss	1
lr_backbone	1e-5	obj_loss	1
weight_decay	1e-4	class_loss	1
k_pe_temp	20	bbox_loss	5
q_point_pe_temp	20	giou_loss	2
q_bbox_pe_temp	20	obj_cost	2
enc_layers	3 / 6	class_cost	2
dec_layers	3 / 6	class_cost	2
dim_feedforward	2048	bbox_cost	5
hidden_dim	256	giou_cost	2
dropout	0.0	inner_cost	9999
nheads	8	focal_alpha	0.25
warm_up	1000	transformer_activation	relu
batch_size	4×4	num_queries	400

Table 11. All configurations of SAP-DETR

G. Visualization of Attention Maps

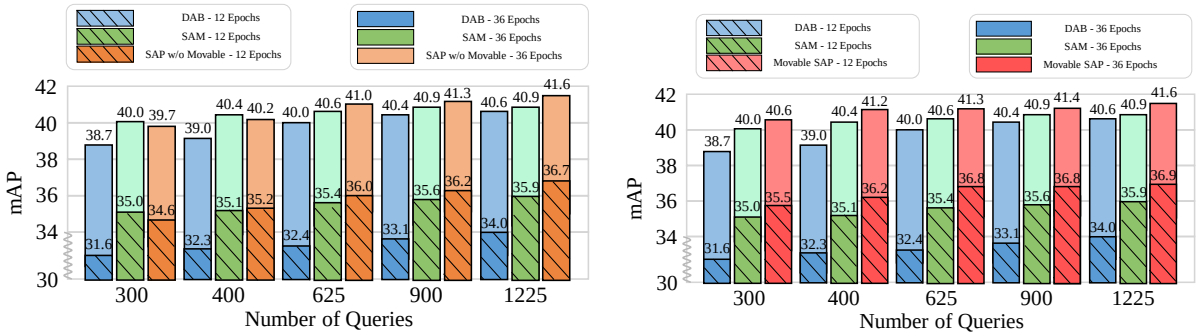
Visualization of Query-Specific Region. To understand how query-specific reference point affect on the object queries aggregation, we visualize the cross-attention map and the output bounding box for each query based on DAB-DETR and our proposed SAP-DETR in Fig. 12 to Fig. 15. Precisely, we visualize the query-specific region in various scenes. For example, the #785 validation image with sample background and sparse instance, the #71226 validation image with complex background and different scale objects, the #1000 validation image with sophisticated instance objects, and the #3255 validation image with sophisticated small instance objects. Compared with redundant prediction and wilderness attention region in DAB-DETR, each query of SAP-DETR only has a compact attention receptive field except for the positive instance query, which benefits from the query-specific reference point and PECA attention mechansim, hence resulting in a superior convergency speed.

Visualization of PECA. Fig. 16 visualizes both content and side attention generated by the proposed PECA. For each positive object query, we visualize each head attention map from the cross-attention mechanism. Then we compare them with the conditional spatial cross-attention. All models are based on ResNet-50 and 6-layer encoder-decoder structure under 50 training epochs. Intuitively, our content attention region mostly falls within the foreground content features, whereas a proportion of the head of Conditional DETR focus on the background. For the side attention, the attention maps of Conditional DETR are inaccurate, with several attention regions outside the bounding box. These inaccurate regions make it fail to locate the extremities efficiently and accurately. The visualization proves the effectiveness of PECA for extreme region attention and partial object detection.



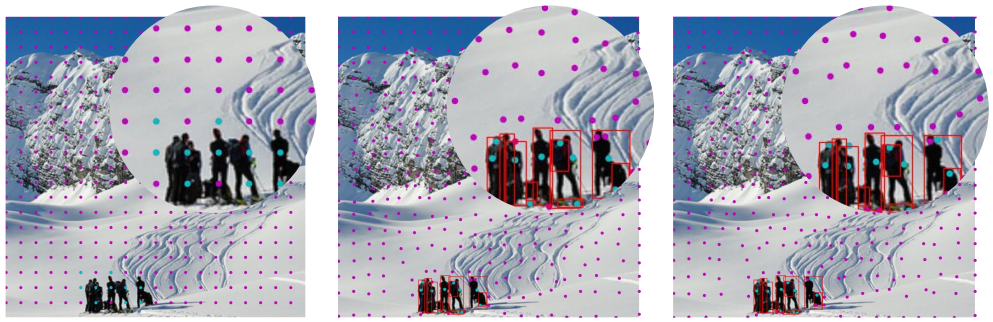
(a) Fixed Reference Point (b) Learnable Reference Point (c) Learnable Box Sides

Figure 8. Distribution.



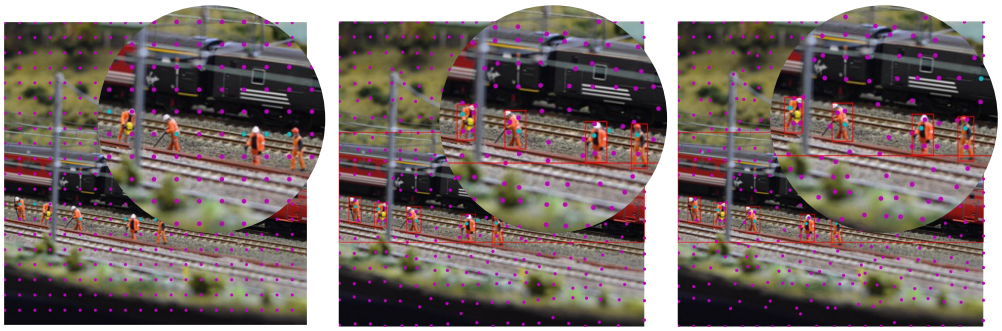
(a) w/o Movable Reference Point (b) Movable Reference Point

Figure 9. Comparison of performance and training losses curves between our purposed SAP-DETR and the current SOTA methods.



Initialized Reference Point Output Point from Layer #0 Output Point from Layer #5

Figure 10. Movable point update for COCO validation image #3255.



Initialized Reference Point Output Point from Layer #0 Output Point from Layer #5

Figure 11. Movable point update for COCO validation image #14473.

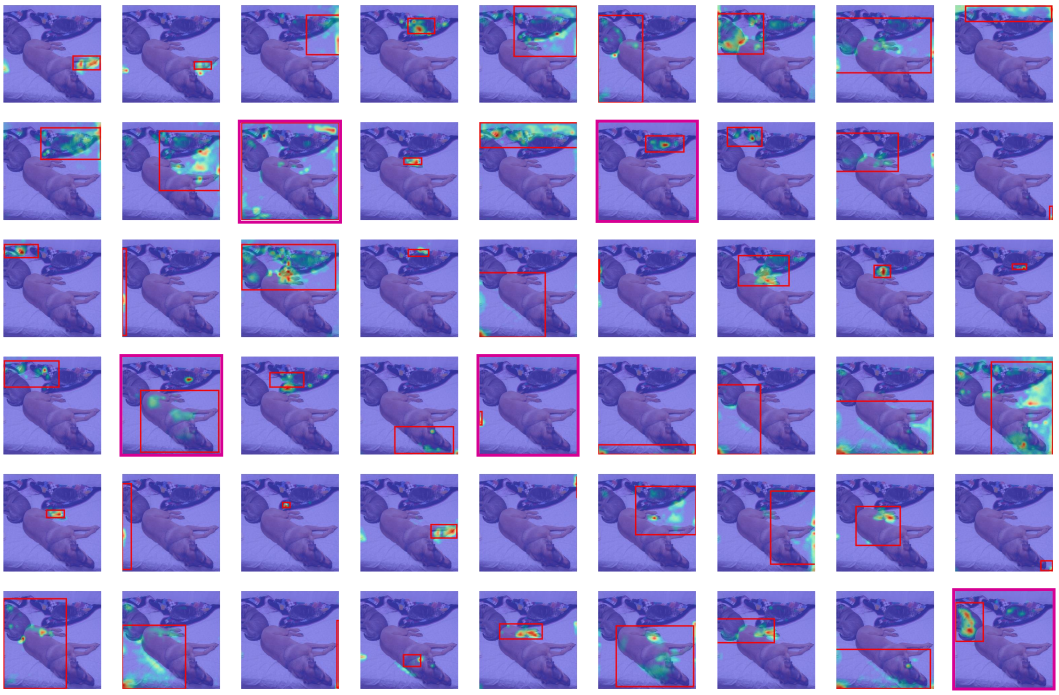
(a) Ground Truth of COCO validation image #785

(b) DAB-DETR for COCO validation image #785

(c) SAP-DETR for COCO validation image #785

Figure 12. Visualization of partial object queries in both SAP-DETR and DAB-DETR.

(a) Ground Truth of COCO validation image #71226

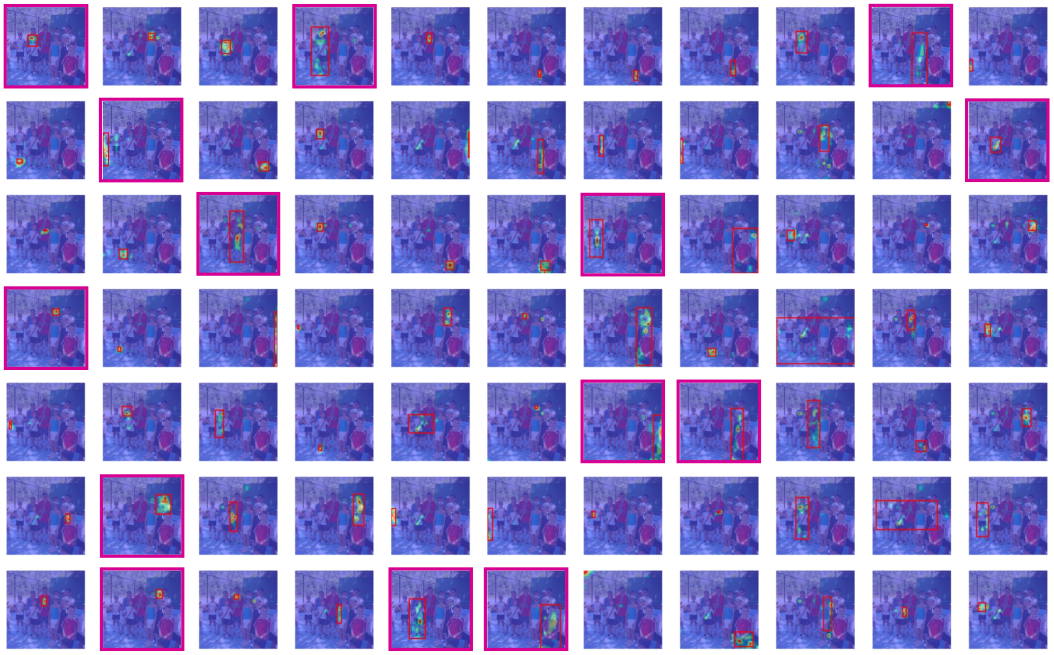


(b) DAB-DETR for COCO validation image #71226

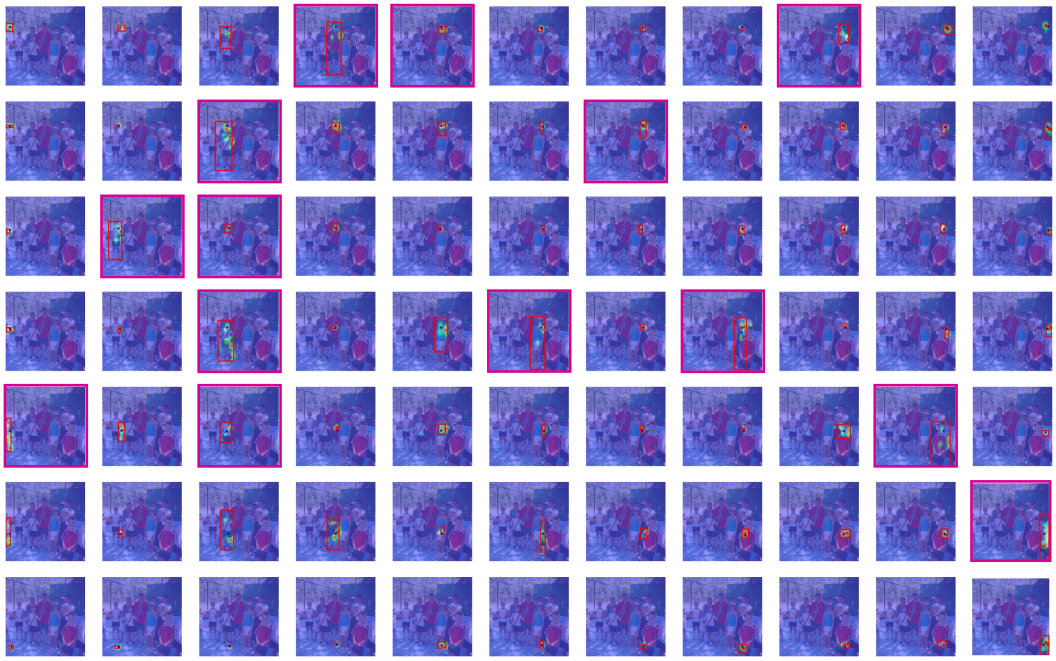
(c) SAP-DETR for COCO validation image #71226

Figure 13. Visualization of partial object queries in both SAP-DETR and DAB-DETR.

(a) Ground Truth of COCO validation image #1000



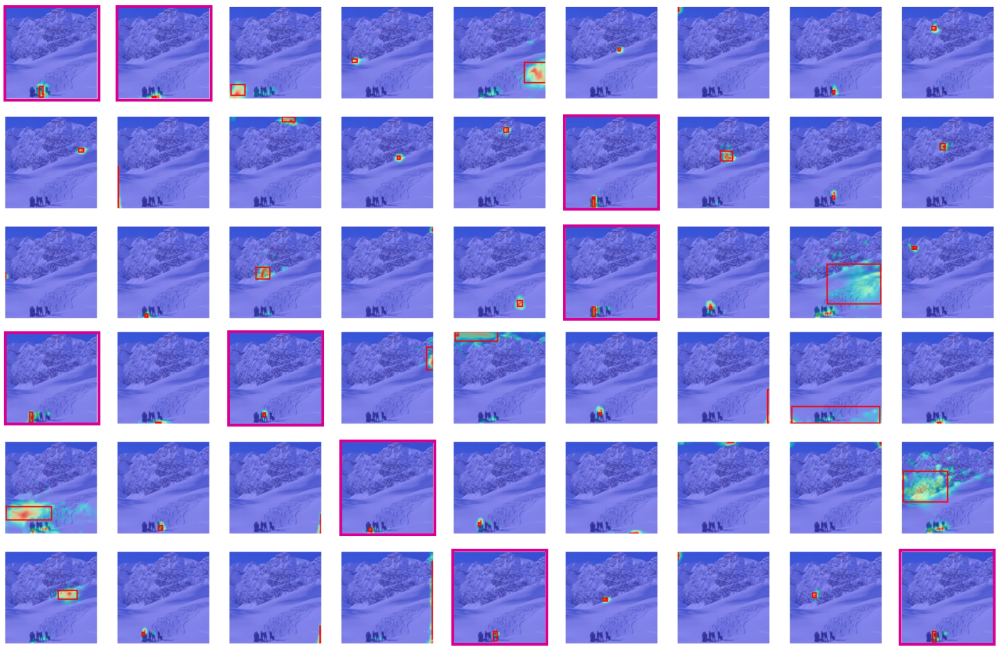
(b) DAB-DETR for COCO validation image #1000



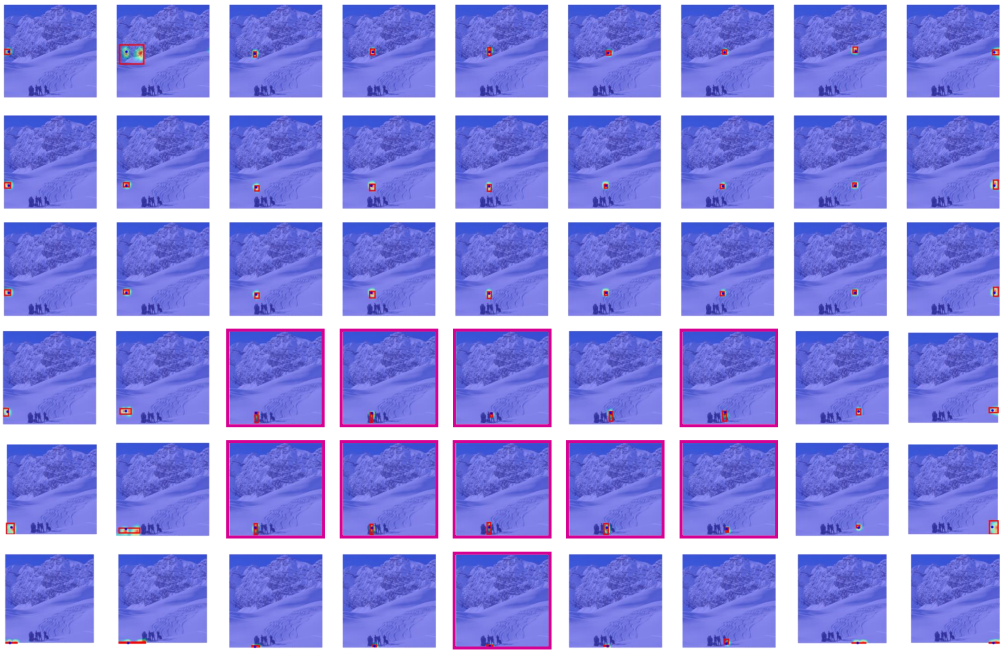
(c) SAP-DETR for COCO validation image #1000

Figure 14. Visualization of partial object queries in both SAP-DETR and DAB-DETR.

(a) Ground Truth of COCO validation image #3225

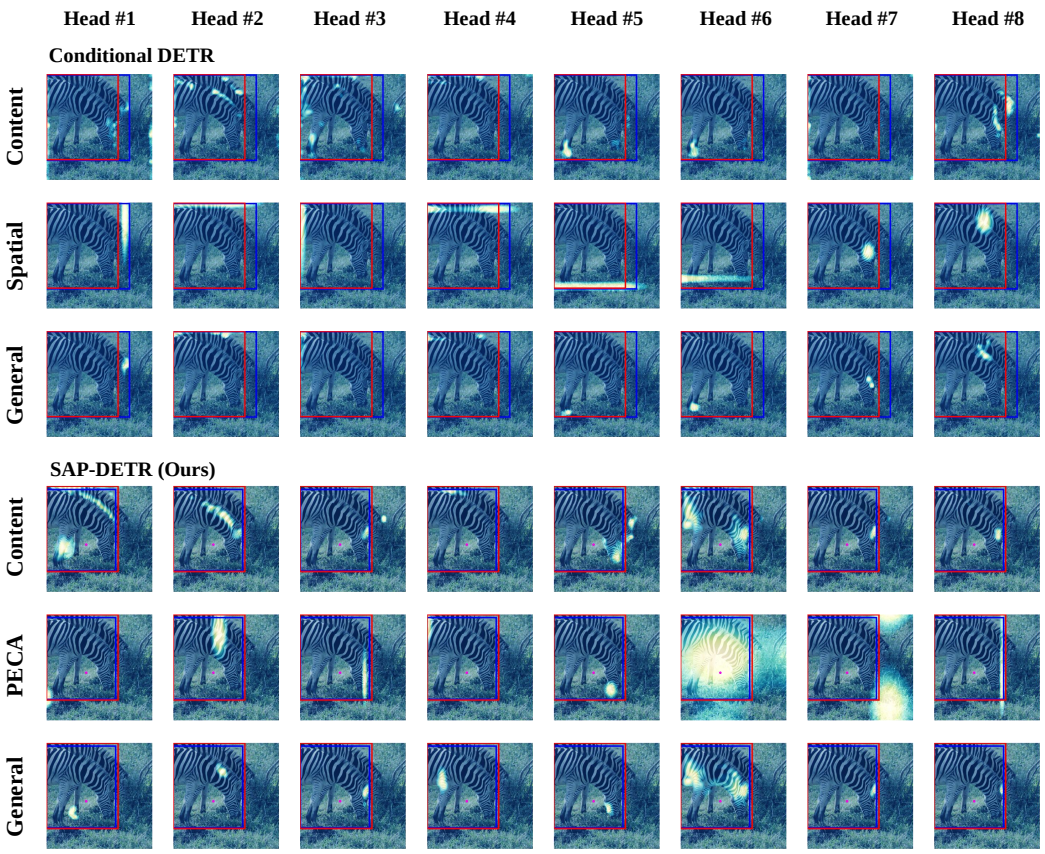


(b) DAB-DETR for COCO validation image #3255



(c) SAP-DETR for COCO validation image #3255

Figure 15. Visualization of partial object queries in both SAP-DETR and DAB-DETR.



(a) Comparison on COCO validation image #159311

(b) Comparison on COCO validation image #507975

Figure 16. Comparison of PECA between Conditional DETR and SAP-DETR.