STEP-DETR: Advancing DETR-based Semi-Supervised Object Detection with Super Teacher and Pseudo-Label Guided Text Queries

Tahira Shehzadi[1,2,3], Khurram Azeem Hashmi[1,2,3], Shalini Sarode[1,2,3], Didier Stricker[1,2,3], Muhammad Zeshan Afzal[1,2,3]
[1]DFKI, [2]RPTU Kaiserslautern-Landau, [3]MindGarage-RPTU
tahira.shehzadi@dfki.de

In this supplementary document, we provide additional details of the proposed STEP-DETR. Section 1 outlines the implementation detail, including data pre-processing and training configurations. Section 2 presents extended experimental results and analyses of each module's contribution. Section 3 includes visualizations of pseudo-labels, queries, and predictions of proposed STEP-DETR.

1. Additional Details of Implementation.

The STEP-DETR approach is implemented using the MMDetection framework [1], with data pre-processing methods adapted from Soft-Teacher [3]. The model is trained on an A100 GPU with 80GB memory, requiring approximately two days to complete 120,000 iterations. In the COCO-Partial setup, the static teacher is trained on labeled data for 12 epochs; after this, the network is trained for 120,000 iterations, processing five images per iteration. The first 60,000 iterations utilize a one-to-many assignment strategy, followed by a one-to-one assignment for the remaining iterations. The labeled-to-unlabeled data ratio is set at 1:4. In the Pascal VOC setup; the training involves 40,000 iterations with a one-to-many assignment strategy, transitioning to one-to-one for the next 40,000 iterations, with the same labeled-to-unlabeled ratio (For every labeled image, there are four unlabeled ones. This standard ratio, as in Soft-Teacher and Semi-DETR, ensures fair comparison). Across all setups, a confidence threshold of 0.4 is used. The model employs the Adam optimizer with a learning rate of 0.001 and no learning rate decay to ensure a fair comparison with Semi-DETR [4].

Data Augmentation. We adopt the data augmentation strategy proposed in Semi-DETR [4], as detailed in Table 2. To generate pseudo-labels, weak augmentation is applied to the unlabeled data, ensuring reliable pseudo-label generation. During the student's training phase, both labeled and unlabeled data undergo strong augmentation, enhancing the student's robustness and generalization capabilities.

2. Additional Details of Modules.

In this section, we conduct additional experiments to evaluate the contribution of each module in the STEP-DETR framework.

Method	COCO-Partial		
	mAP	AP_{50}	AP_{75}
Exp-1	43.5	59.7	46.8
Exp-2	43.5	59.8	46.8
Exp-3	45.7	63.1	49.3

Table 1. Performance comparisons of different variants of Teacher. The mAP results for Exp-1, Exp-2, and Exp-3, highlighting the impact of incorporating the Static Teacher in the Super Teacher module. Performance on employing only the Dynamic Teacher (Exp-1), the Dynamic Teacher initially trained on limited labeled data and generating pseudo-labels for the student (Exp-2) and the full Super Teacher module with both Static and Dynamic Teachers (Exp-3).

Impact of Super Teacher: We perform additional experiments to assess the efficacy of our Super Teacher as follows:

1. Is the Static Teacher in the Super Teacher module crucial? Can we train the Dynamic Teacher independently and then have it update its parameters based on the student's learning, or is the Static Teacher necessary for optimal performance?
2. How does the absence of the Static Teacher affect training? How does Static Teacher affect the training process and overall performance?

To address these two questions, we conduct three experiments as follows:

Exp-1: In this experiment, we use only the Dynamic Teacher, which updates its parameters via the student EMA update. The setup for this experiment is illustrated in Figure 1(a). As shown in Table 1, the performance drops to 43.5 mAP without the Static Teacher. This result highlights the critical role of the Static Teacher in our Super Teacher

Augmentation	Labeled image training	Unlabeled image training	Pseudo-label generation
Scale Jitter	shortest edge $\in [480, 800]$	shortest edge $\in [480, 800]$	shortest edge $\in [480, 800]$
Solarize Jitter	$p = 0.25$, ratio$\in (0, 1)$	$p = 0.25$, ratio$\in (0, 1)$	-
Brightness	$p = 0.25$, ratio$\in (0, 1)$	$p = 0.25$, ratio$\in (0, 1)$	-
Contrast Jitter	$p = 0.25$, ratio$\in (0, 1)$	$p = 0.25$, ratio$\in (0, 1)$	-
Sharpness Jitter	$p = 0.25$, ratio$\in (0, 1)$	$p = 0.25$, ratio$\in (0, 1)$	-
Translation	-	$p = 0.3$, translation ratio$\in (0, 1)$	-
Rotate	-	$p = 0.3$, angle$\in (0, 30°)$	-
Shift	-	$p = 0.3$, angle$\in (0, 30°)$	-
Cutout	num$\in (1, 5)$, ratio$\in (0.05, 0.2)$	num$\in (1, 5)$, ratio$\in (0.05, 0.2)$	-

Table 2. Data augmentations used in our approach. p indicate the probability of choosing a certain type of augmentation.

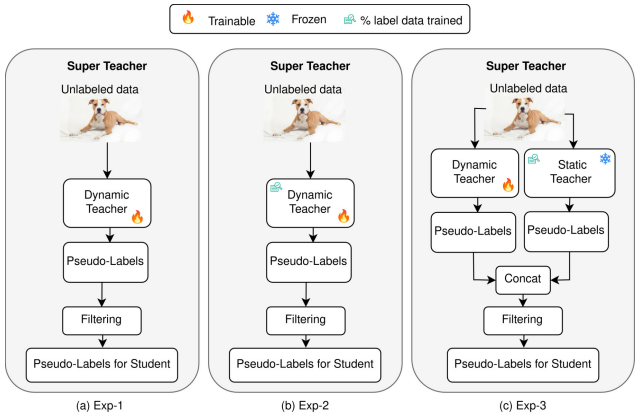

Figure 1. Experimental setups for the Super Teacher module and its components. **(a)** Exp-1: The Super Teacher generates pseudo-labels only through the Dynamic Teacher that is updated using the student's EMA updates, without the Static Teacher. **(b)** Exp-2: In Super Teacher, the Dynamic Teacher is first trained on the available labeled data (e.g., 10%) and then generates pseudo-labels for the student. It is updated using the student's EMA updates without the Static Teacher. **(c)** Exp-3: In Super Teacher, the Static Teacher is trained on the labeled dataset and provides stable pseudo-labels to the student, with its parameters frozen. The Dynamic Teacher continues to update based on the student's learning. The pseudo-labels from both the Static Teacher and the Dynamic Teacher are concatenated and fed to the student, resulting in improved performance.

module. The likely reason for the effectiveness of the Static Teacher is that the Dynamic Teacher updates according to the student's learning progress. At the early stages of training, the student results in poor-quality labels from Dynamic Teacher. It highlights the Static Teacher's stabilizing role in Dynamic Teacher's guidance.

Exp-2: In this experiment, the Dynamic Teacher is first trained on the available labeled data (e.g., 10%) and then generates pseudo-labels for the student. As shown in Table 1 and illustrated in Figure 1(b), this approach results in a performance of 43.5 mAP. Despite the initial training on labeled data, the Dynamic Teacher does not improve pseudo-label quality as the student updates it. The students' early inaccurate predictions influence the updates to the Dynamic Teacher. As the student makes low-quality predictions, these errors are passed to the Dynamic Teacher, hindering its ability to refine pseudo-labels effectively. It creates a feedback loop that limits performance and highlights the challenge of relying solely on the Dynamic Teacher without the Static Teacher.

Exp-3: The Static Teacher is first trained on the labeled dataset, and its parameters are frozen. The student and Dynamic Teacher continue updating during the learning process, but the Static Teacher's parameters remain frozen. This setup is illustrated in Figure 1(c). As shown in Table 1, this approach achieves a performance of 45.7 mAP. The role of the Static Teacher is to provide stable and consistent pseudo-labels based on its pre-learned knowledge, which helps mitigate the impact of noisy predictions from the student during the early stages of training. By offering reliable supervision, the Static Teacher ensures a more robust learning process than relying solely on the Dynamic Teacher, whose updates are affected by the student's noisy predictions.

Category-level results. Table 3 shows Category-level results of all categories for Static Teacher, Dynamic Teacher and Super Teacher.

Super Teacher role and benefit: It contains Static Teacher trained on label data and Dynamic Teacher that is updated with Student learning (Exp3). It ensures reliable guidance, preserves labeled data knowledge, and provides strong pseudo-labels from the first iteraiton.

Static Teacher need and updation: It is needed to generate consistent and reliable pseudo-labels from the beginning and is not updated (as shown in Figure 3) to prevent error propagation caused by student's early inaccurate predictions. Updating Static Teacher declines accuracy by 6.8%.

Pseudo-labels for Rare vs. Common Classes. In Figure 2, we demonstrate how the Dynamic and Super Teacher generate pseudo-labels for rare and common classes. In the Dynamic Teacher, common classes exhibit higher confidence than rare classes, as observed in Figure 2(a) at the start of training. As a result, queries for rare classes, which initially

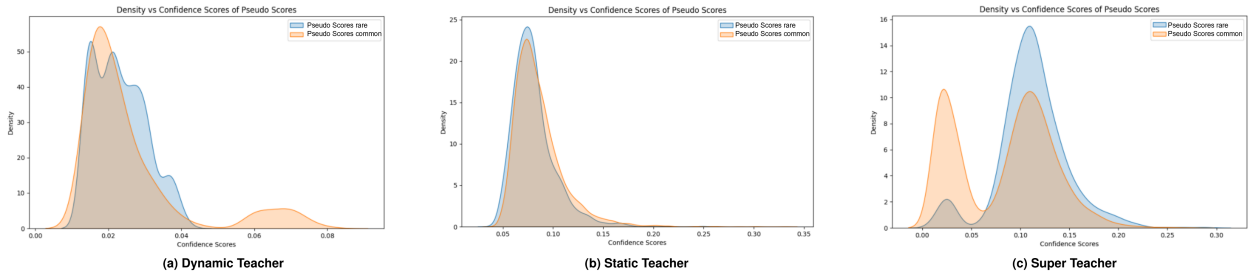

Figure 2. Illustration of rare and common class handling by three types of teachers: Dynamic Teacher, Static Teacher, and Super Teacher. (a) In Dynamic Teacher, common classes exhibit higher confidence, while rare classes are filtered out due to low-confidence pseudo-labels. This limits the model's ability to learn from rare classes effectively during early training. (b) Static Teacher provides high-confidence pseudo-labels for both rare and common classes, overcoming the limitations of the Dynamic Teacher but lacking adaptability. (c) Super Teacher combines the strengths of both Static and Dynamic Teachers, providing a balanced approach that ensures improved handling of both rare and common classes.

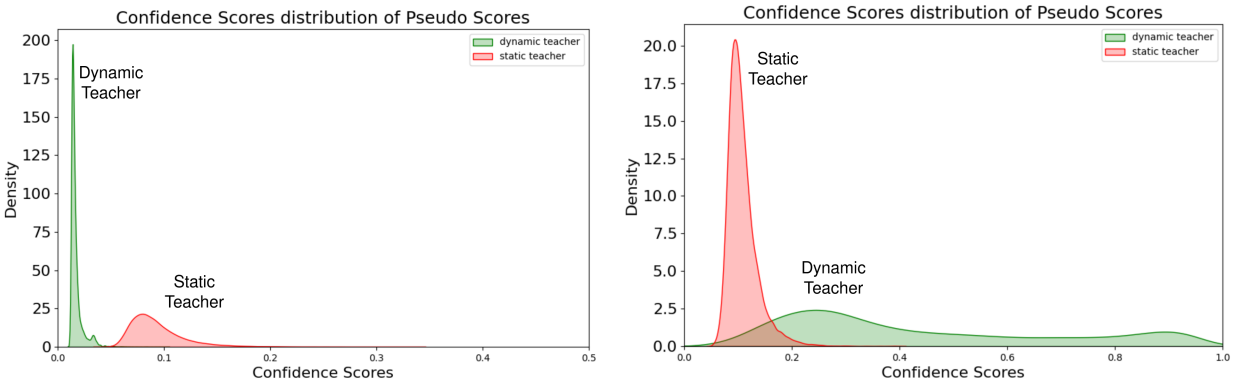

Figure 3. Illustration of the behavior of Static Teacher and Dynamic Teacher during student learning. (a) At the start of training (0k iterations), the Static Teacher provides fixed pseudo-labels that remain constant throughout the learning process. At the same time, the Dynamic Teacher begins updating its pseudo-labels based on the student's learning. (b) By the end of training (120k iterations), the Static Teacher retains its original pseudo-labels, offering consistent guidance. In contrast, the Dynamic Teacher has adapted its pseudo-labels to align with the student's learning progression. This comparison highlights the static and adaptive characteristics of the two teacher types during training.

have low confidence, are not generated since their pseudo-labels are filtered out. This filtering mechanism limits the super teacher-student ability to effectively learn from rare classes during the early stages of training. Figure 2(b) shows the pseudo-label scores for rare and common classes. In the Super Teacher, which combines both the Dynamic and Static Teachers, high-confidence pseudo-labels are provided for rare classes as well, as shown in Figure 2(c). After filtering, pseudo-labels for rare classes remain, enabling their queries to be generated. This approach improves students' learning ability from rare classes and enhances overall performance.

Static Teacher vs. Dynamic Teacher Behavior During Training. Figure 3 illustrates the behavior of the Static Teacher and Dynamic Teacher during the student's learning process. At the start of training, shown in Figure 3(a), the Static Teacher provides fixed pseudo-labels that remain the same throughout the training process, while the Dynamic Teacher begins updating its pseudo-labels based on the student's learning. By the end of the training, as shown in Figure 3(b), the Static Teacher retains its original pseudo-labels, offering consistent guidance, whereas the Dynamic Teacher has adapted its pseudo-labels over time to align with the student's learning and progression. This comparison highlights the static and adaptive characteristics of Static and Dynamic teacher throughout the training process.

Quantity of Pseudo-label Text Queries. Prior works [2, 4] exclude rare class queries by filtering out low-confidence rare-class instances. We incorporate them through text queries to enhance detection performance We analyze the quantity of pseudo-label text queries using a single batch example, as illustrated in Figure 4(a). Each image in the batch has a varying number of pseudo-labels from Super

Category	AP (Static)	AP (Dynamic)	AP (Super)	Category	AP (Static)	AP (Dynamic)	AP (Super)
person	0.481	0.572	0.584	bicycle	0.221	0.321	0.333
car	0.340	0.441	0.462	motorcycle	0.322	0.451	0.469
airplane	0.603	0.686	0.712	bus	0.585	0.674	0.694
train	0.589	0.676	0.693	truck	0.281	0.348	0.370
boat	0.188	0.282	0.293	traffic light	0.211	0.281	0.284
fire hydrant	0.600	0.688	0.715	stop sign	0.581	0.621	0.627
parking meter	0.386	0.482	0.514	bench	0.185	0.245	0.258
bird	0.312	0.396	0.415	cat	0.628	0.791	0.796
dog	0.564	0.659	0.673	horse	0.484	0.624	0.656
sheep	0.419	0.560	0.574	cow	0.472	0.610	0.611
elephant	0.605	0.715	0.724	bear	0.694	0.754	0.757
zebra	0.623	0.701	0.708	giraffe	0.615	0.713	0.716
backpack	0.089	0.135	0.137	umbrella	0.283	0.389	0.415
handbag	0.071	0.103	0.139	tie	0.260	0.363	0.370
suitcase	0.283	0.450	0.459	frisbee	0.560	0.683	0.688
skis	0.155	0.240	0.272	snowboard	0.239	0.345	0.433
sports ball	0.394	0.471	0.478	kite	0.351	0.460	0.486
baseball bat	0.208	0.286	0.343	baseball glove	0.282	0.387	0.415
skateboard	0.420	0.529	0.574	surfboard	0.264	0.390	0.440
tennis racket	0.397	0.487	0.526	bottle	0.281	0.385	0.407
wine glass	0.242	0.372	0.387	cup	0.325	0.424	0.433
fork	0.201	0.353	0.415	knife	0.073	0.179	0.214
spoon	0.069	0.133	0.199	bowl	0.304	0.424	0.460
banana	0.177	0.243	0.274	apple	0.132	0.161	0.182
sandwich	0.274	0.364	0.387	orange	0.265	0.285	0.288
broccoli	0.194	0.238	0.252	carrot	0.137	0.196	0.241
hot dog	0.214	0.383	0.436	pizza	0.472	0.566	0.580
donut	0.385	0.505	0.547	cake	0.230	0.366	0.396
chair	0.182	0.286	0.306	couch	0.334	0.442	0.474
potted plant	0.179	0.250	0.475	bed	0.387	0.482	0.513
dining table	0.220	0.284	0.314	toilet	0.518	0.622	0.625
tv	0.508	0.580	0.594	laptop	0.434	0.627	0.635
mouse	0.521	0.614	0.631	remote	0.193	0.314	0.360
keyboard	0.408	0.541	0.583	cell phone	0.281	0.355	0.389
microwave	0.443	0.576	0.591	oven	0.234	0.356	0.362
toaster	0.207	0.372	0.374	sink	0.290	0.387	0.396
refrigerator	0.428	0.599	0.615	book	0.096	0.136	0.148
clock	0.468	0.540	0.544	vase	0.281	0.382	0.419
scissors	0.056	0.284	0.375	teddy bear	0.349	0.479	0.539
hair drier	0.148	0.158	0.180	toothbrush	0.103	0.238	0.286

Table 3. Comparison of Average Precision (AP) for Static Teacher, Dynamic Teacher, and Super Teacher.

Teacher; the highest pseudo-label count across all images is used as a baseline to ensure uniformity across the batch. To maintain consistency, these pseudo-labels are repeated five times. For each repetition, corresponding bounding boxes and text embeddings are generated and fed as text queries to the decoder of super teacher-student.

Quantity of Denoising Text-guided Object Queries. It's hard to distinguish object-background from denoising. Adding noise to denoising queries typically shifts bounding box coordinates without altering class labels, making it harder for the classifier to distinguish the background. This issue worsens in semi-supervised learning, where denoising queries are generated from noisy pseudo-labels. For this, we employ Denoising Text-guided Object Queries. Unlike prior denoising approaches, we assign correct textual embeddings to foreground-positive queries, leveraging pseudo-labels from the Super Teacher as text queries, while assigning random embeddings to background-negative queries

Figure 4(b) illustrates the process of generating denoising text-guided object queries for images in a batch. To ensure

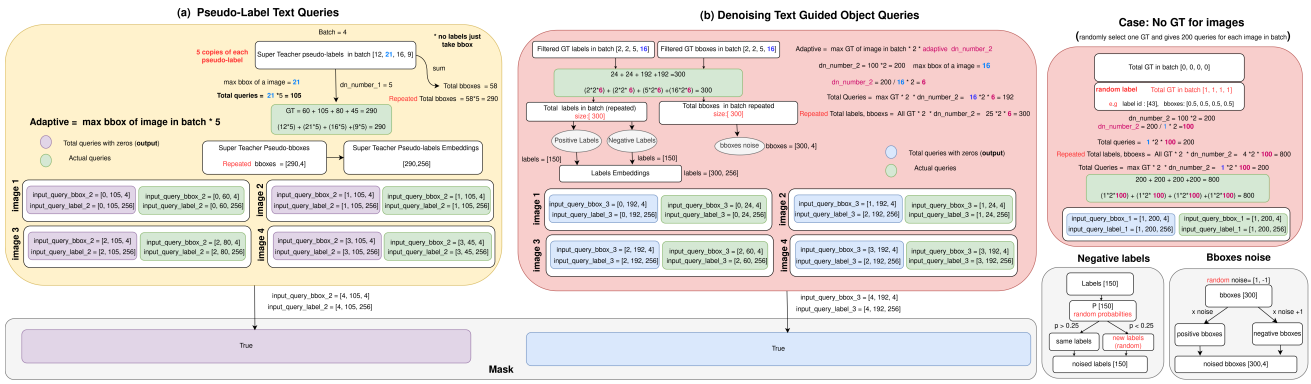

Figure 4. Running Example of Pseudo-Label Text Queries and Denoising Text-Guided Queries for a Batch. **(a)** Pseudo-label text queries take the maximum pseudo-label count in a batch from the Super Teacher as the baseline, repeating these labels multiple times for consistency. Corresponding bounding boxes and text embeddings are used as text queries. **(b)** Denoising text-guided object queries generates both positive and negative object queries, where negative queries contain more noise than positive queries. Negative query labels are randomly selected based on probabilities, while positive query labels are from the Super Teacher. For images without ground-truth, random labels and bounding boxes are assigned. Here, it can be observed that denoising text-guided queries are larger in quantity compared to pseudo-label text queries. For a more detailed and clearer view, please zoom in.

Figure 5. Qualitative Comparison of pseudo-labels at 60k iterations. (a) Ground-truth (b) Pseudo-labels in Semi-DETR (c) Pseudo-labels with both Dynamic and Static Teachers in STEP-DETR. Compared to Semi-DETR, our approach generates more accurate query proposals for each unlabeled image. Ground-truths are outlined in green, while positive query proposals are highlighted in blue. STEP-DETR outperforms Semi-DETR, as evidenced by the positive proposals from both Dynamic and Static teachers. Additionally, STEP-DETR introduces text queries based on positive proposals from the Super Teacher and refines the Denoising Text-Guided Object Queries module, significantly improving proposal quality and identification accuracy.

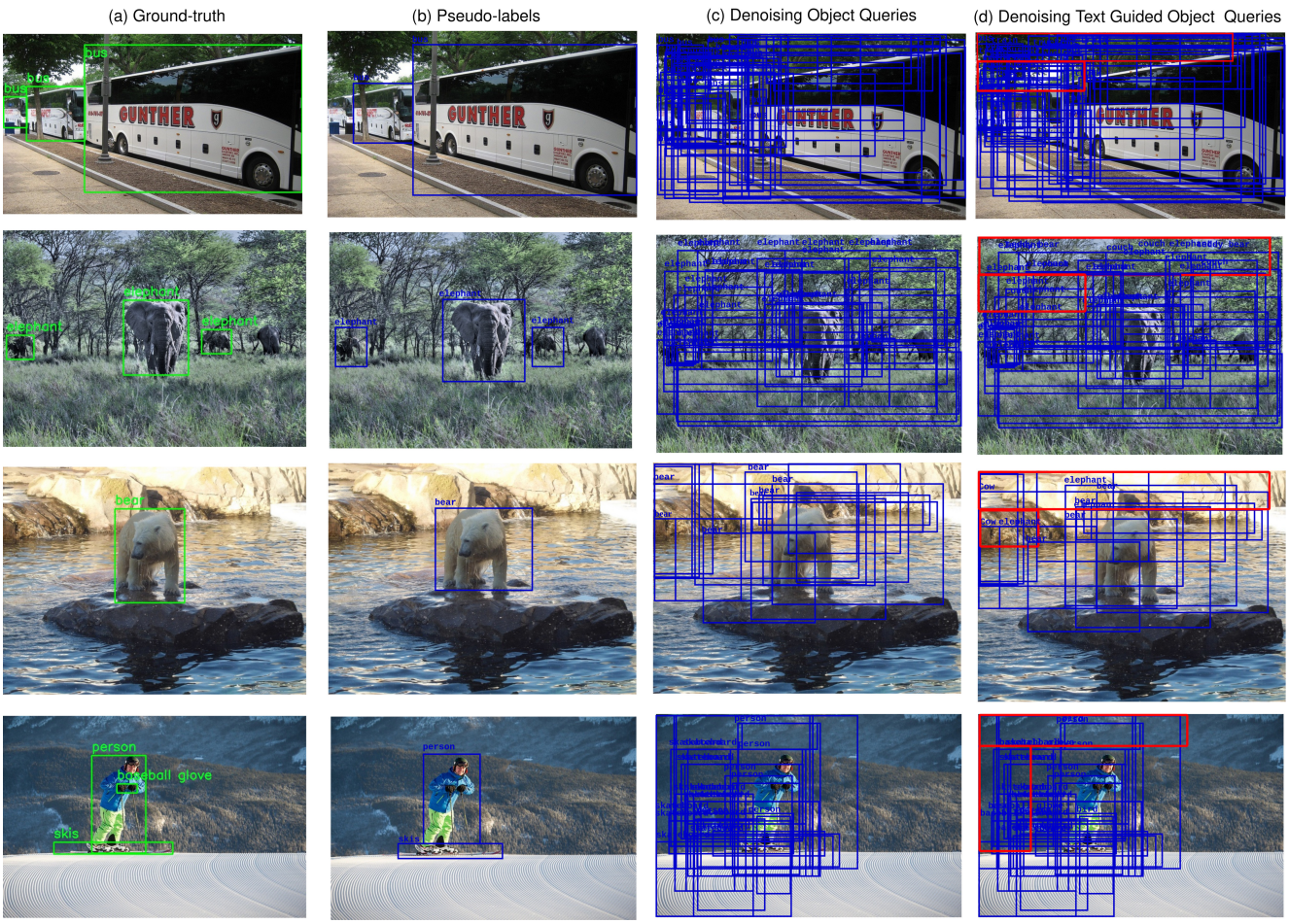

Figure 6. Qualitative Comparison of Standard Denoising Object Queries and Denoising Text-Guided Object Queries. (a) Ground-truth (b) Pseudo-labels of Super Teacher (c) Standard Denoising Object Queries (d) Denoising Text-Guided Object Queries. Text-guided queries generate more accurate proposals for each unlabeled image than standard denoising queries. To generate foreground query-box pairs, the correct text embedding from the Super Teacher is added to the positive object queries. Additionally, background query-box pairs are generated by randomly assigning labels from the total categories to the object queries of boxes indexed based on probability, as in red box regions. STEP-DETR introduces text queries based on the positive proposals from the Super Teacher and refines them through the Query Refinement module. This process filters queries based on pseudo-labels, improving proposal quality and enhancing performance.

robust and diverse query generation, the filtered pseudo-ground-truth per image is doubled and scaled by an adaptive DN factor. These queries are then categorized evenly into positive and negative groups. Bounding boxes with added noise are assigned to these queries, with negative queries incorporating higher noise levels compared to positive ones. Positive queries are derived from the Super Teacher, whereas negative queries receive labels randomly sampled from the class distribution, guided by randomly assigned probabilities. For images without pseudo ground-truth, random labels and bounding boxes are generated to maintain uniformity in query generation across the entire batch. This method introduces controlled noise and variability, enhancing the model's adaptability and ability to generalize. As shown in Figure 4, the denoising text-guided queries are larger in quantity compared to pseudo-label text queries.

Confidence score of rare vs common categories. We provide additional results of rare category comparison for supervised baseline DINO vs STEP-DETR in Table 4. With 10% label data, rare categories like "fire hydrant" has lower confidence Figure 2b and 60.0 AP. STEP-DETR improves both.

Approach		rare category:AP	confidence score
Supervised	(10%)	fire hydrant: 60.0	Figure 2b
Semi-supervised	(10%)	fire hydrant: 71.5	Figure 2c

Table 4. Additional results of rare category comparison for supervised vs STEP-DETR.

3. Additional Visual Analysis of Pseudo-labels, Queries and Output Predictions.

Visual Analysis of Pseudo-labels. Figure 5 presents a qualitative comparison of pseudo-labels at 60k iterations, showing pseudo-labels generated using only the Dynamic Teacher as in Semi-DETR and those generated using the Super Teacher in STEP-DETR. By leveraging superior pseudo-labels from the Super Teacher, STEP-DETR outperforms Semi-DETR, producing more accurate query proposals for each unlabeled image.

Visual Analysis of Denoising Queries. Figure 6 provides a comparison between Standard Denoising Object Queries and Denoising Text-Guided Object Queries. The text-guided queries outperform the standard ones in generating more accurate query proposals for each unlabeled image. This improvement allows for a clearer and more reliable differentiation between object and background queries, thereby enhancing the model's ability to distinguish foreground objects from the background. To further address redundancy and improve model efficiency, we refine the denoising text-guided object queries using pseudo-label text queries in Query Refinement module. This refinement removes irrelevant or redundant queries, particularly among background queries, as highlighted in the red-boxed regions in Figure 6. By filtering out these redundant queries, it enhances training efficiency and performance.

Visual Analysis of Output Predictions. Figure 7 provides qualitative comparisons between the baseline Semi-DETR and STEP-DETR. While Semi-DETR misses objects like the hotdog, STEP-DETR demonstrates notable improvements by detecting missed objects and providing more accurate bounding boxes, highlighting its superior performance in object detection tasks.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 1

[2] Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5850, 2024. 3

[3] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *CoRR*, abs/2106.09018, 2021. 1

[4] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23809–23818, 2023. 1, 3

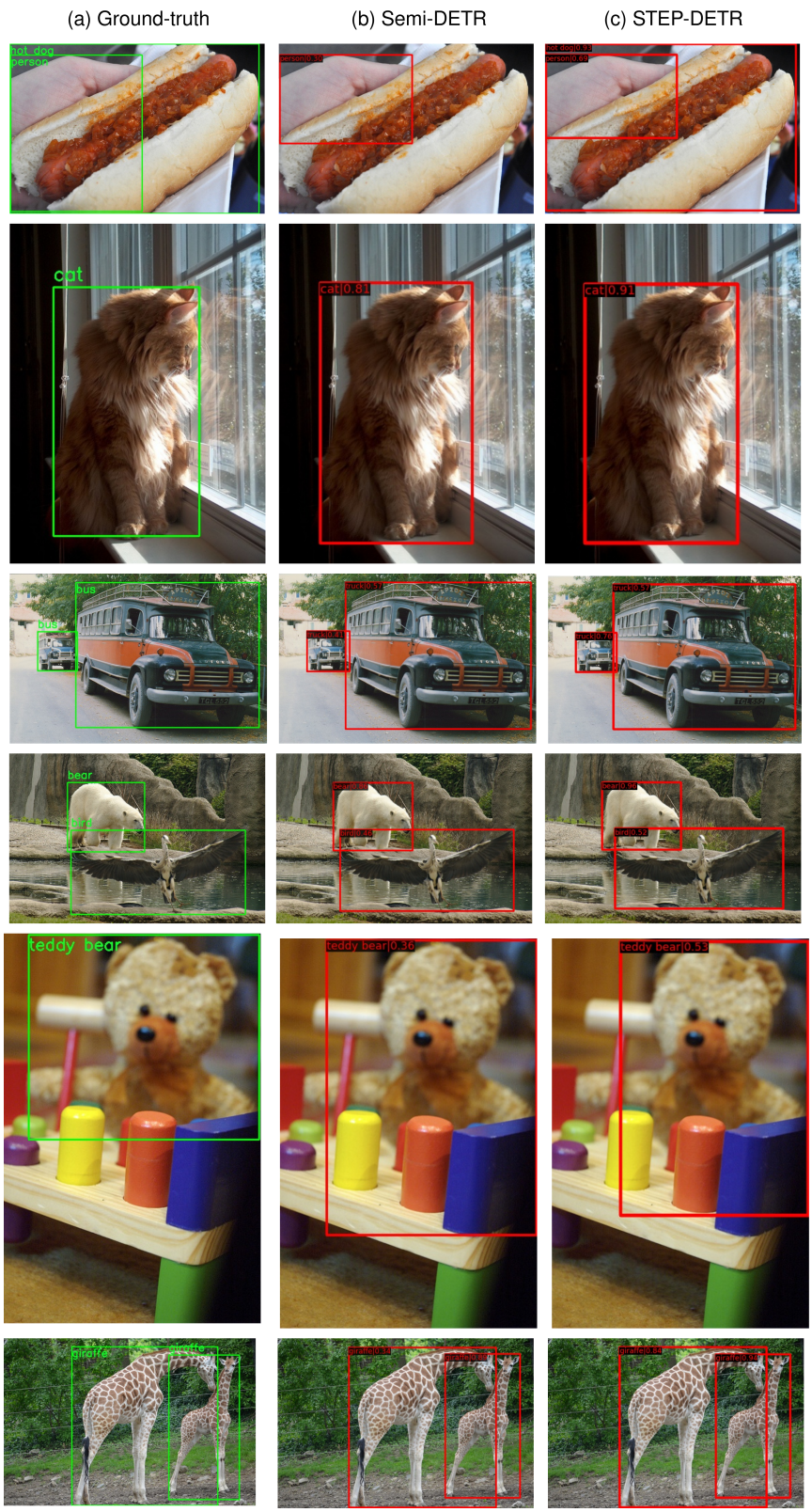

Figure 7. Qualitative comparisons between the baseline Semi-DETR and STEP-DETR on COCO validation data. (a) Ground-truth (b) Semi-DETR predictions. (c) STEP-DETR predictions. STEP-DETR demonstrates improved detection accuracy.