In this supplementary material, we first present the performance of our approach for distillation on SOTA detectors and other dataset. Next, we conduct more ablation studies for our approach to validate its effectiveness.

1. Distillation on SOTA Detectors

In the main paper, we have explored knowledge distillation on three baseline detectors. In this section, we perform KD on DINO [2] which is a stronger variant of DND-DETR. Results in Table 1 show that the student model’s performance is consistently improved with our proposed DETRDistill method. Furthermore, such a KD design can also be applicable when the number of queries in teacher and student is inconsistent, as shown in bottom block of Table 1.

| Setting | Backbone | AP50 | mAP |
|---------|----------|------|-----|
| Teacher | R-101    | 81.00| 80.99|
| Student | R-18     | 75.50| 75.49|
| Ours    |          | 50.1 (+1.5) | 53.4 (+1.8) |

Table 1. Results of our DETRDistill on DINO detector.

2. Distillation on Other Dataset

We also provide experiment on the PASCAL VOC dataset [1] and the performance in Table 2 shows that our DETRDistill gains 3.50 mAP and 2.36 mAP over the baselines, which indicate that DETRDistill is applicable across different datasets.

| Setting | Backbone | AP50 | mAP |
|---------|----------|------|-----|
| Teacher | R-101    | 81.00| 80.99|
| Student | R-50     | 79.30| 79.25|
| Ours    |          | 81.60 (+2.30) | 81.61 (+2.36) |

Table 2. Results on PASCAL VOC dataset. Train: trainval set of VOC2007 & VOC2012; Eval: VOC2007 test.

3. More Ablation Studies

Apart from the ablation studies presented in the main paper, we further provide more for our proposed approach.
Table 3. Comparison of the number of model parameters and computation cost on various detectors. Proportion of KD Cost is defined as KD Compute Cost / (Basic Compute Cost of the Teacher + Basic Compute Cost of the Student + KD Compute Cost).

| Teachers | AdaMixer | Deformable DETR | Conditional DETR |
|----------|----------|-----------------|------------------|
|           | Tea-R101 | Stu-R50        | Teacher-R101     | Stu-R50        | Tea-R101 | Stu-R50 |
| Model Params number (M) | 153.56 | 134.57 | 58.78 | 39.84 | 62.13 | 43.19 |
| Basic Compute Cost (GFLOPs) | 178.95 | 102.88 | 287.34 | 192.26 | 171.4 | 95.32 |
| KD Compute Cost (GFLOPs) | 24.49 | 51.10 | 18.35 | 6.43 |
| Proportion of KD Cost | 7.99 % | 9.63 % | 6.43 % | 6.43 % |

accuracy which shows the importance of incorporating the teacher’s assignment in the Query-prior Assignment Distillation module.

| Teacher Init | Teacher Group | Teacher Assigned |
|--------------|---------------|------------------|
| 43.6         | 42.3 (+0.0)   | **42.9 (+0.6)**  |

Table 4. Comparison of different variants for Query-prior Assignment Distillation module.

Analysis to the computation consumption in training We are interested to know the computational cost of our proposed distillation modules at the training phase. Since various detectors have different model architectures, different Flops are required and we report detector-specific computation consumption in Table. 3. It is clear that our proposed KD module only takes a small proportion of computation consumption in the whole model optimization which verifies the efficiency of our proposed approach.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[2] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.