Appendices

A. Further implementation details.

The set of strong augmentations passed to the student model includes color jittering, random solarization, grayscale augmentation, crop-resize, and random-flip, while the teacher only receives flip and crop-resize transfomrations with a cropping scale $75 - 100\%$ of the original image size. As with previous MGOD methods [2, 3, 18], the set M of motion masks is obtained via the mapping of RAFT optical flow [40] to instance masks, learned on the synthetic dataset FlyingThings3D [32] using the approach described in [7] . Also, the model trained on KITTI is initialized through training on TRI-PD (following [2, 3, 18]). Burn-in duration and the regularization strength α are jointly analyzed in section 5.3. They are set to 400 and 0.3 respectively on the photo-realistic/real-world datasets TRI-PD and KITTI. Regarding MOVI-E dataset, we report the results using a pretrained encoder (from DINOv2), following the most recent approaches. We used an input resolution of 256×256, adjusted to 266×266 for compatibility with the ViT-based encoder (with DINOv2 pre-training). In this setting, a shorter burn-in duration (60 epochs) was sufficient to achieve the reported performance. The distillation phase runs for $500, 100$ and 200 epochs on TRI-PD, KITTI and MOVI-E, respectively. Both during distillation and burn-in, the learning rate is set to 5.10^{-4}, with batch size 8 on TRI-PD and KITTI, and 32 on MOVI-E.

During the evaluation, the student model is applied to sequences of images of the same length as those used during training.

B. Beyond motion: assessing DIOD's compatibility with a different source of pseudo-labels.

In the main paper, we focused on the use of motion segments as a guidance signal (segments of moving objects, extracted from optical flow). We made this choice since motion information provides a clear definition of what an object is: we aim to localize objects capable of moving, as initiated in [2]. It is worth noting that object definition is lacking in most object discovery methods, and remains an open problem in the field.

In this section, our aim is to verify the compatibility of DIOD with sources of pseudo-labels other than motion. We conduct this experiment on the MOVi-E dataset and, instead of using motion segments, we generate other pseudo-labels to guide the model's attention. To do this, we use our best DIOD* model trained on the TRI-PD dataset for inference only on MOVi-E videos. Given the domain discrepancy between TRI-PD and MOVi, the object segments generated are noisy. To further ensure high signal sparsity, we keep only the three largest segments per frame, excluding the background. We run both the burn-in and distillation phases for 60 epochs with regularisation $\alpha = 0.3$. The results obtained (table .5) show that DIOD effectively handles the noise and sparsity of this new guidance signal, suggesting its compatibility with other pseudo-label sources. These results encourage further investigation into the use of distillation with more generic object discovery frameworks.

Method	Bao et al.	MoTok	SAVI++	PPMP	STEVE	Ours (after burn-in)	DIOD
Fg-ARI	51.6	<u>66.7</u>	41.3	63.1	54.1	57.7	**68.9**

Table .5. Object discovery performance of DIOD on MOVi-E dataset, using a different source of pseudo-labels.

C. Additional qualitative comparison of DIOD

In Figure 5, we provide an additional qualitative comparison on the test set of TRI-PD [2] to visualize the enhancements brought by DIOD . We display in different colors the attention of each slot activated during inference. In BMOD[18] and our approach, the background being modeled within one slot, the content of this specific slot is not displayed. We compare DIOD with state-of-the-art approaches in motion-guided object discovery [2, 18] (left), where the results show significant improvements: better distinction in DIOD between object/non-object regions (background modeling), improved separation of objects even when close and of the same semantic category, and the detection of more objects that are generally difficult-to-capture or/and static. Furthermore, we present in column 4 the result after the burn-in phase of DIOD. Here, one can observe a stronger attenuation of noise than in [18] due to the regularization across the entire image described in the main paper, section 3.2. However, this noise is still present at the end of the burn-in phase, and many objects are missed. These limitations are addressed in DIOD (after distillation) as indicated by column 5.

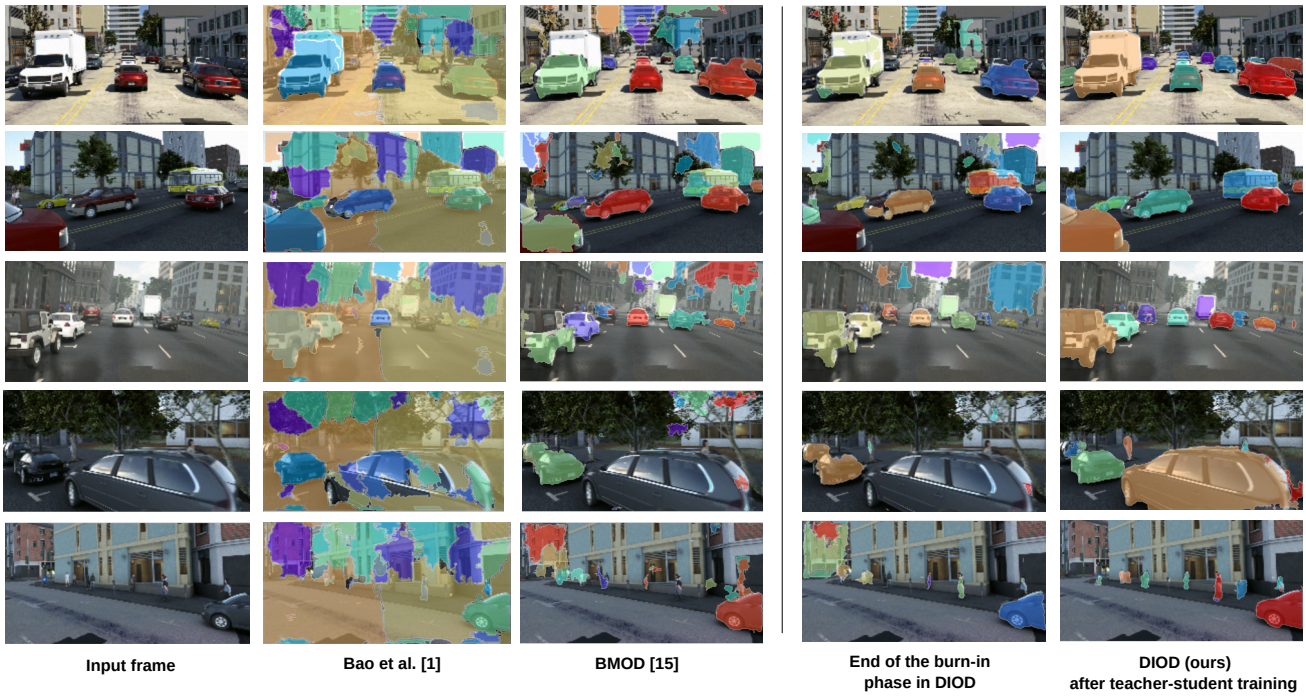

Figure 5. **Qualitative comparison on the test set of TRI-PD [2]**: comparing state-of-the-art methods [2, 18] (left) with our results following the burn-in phase (column 4) and after distillation (DIOD) in column 5.

D. Upper-bound performance of DIOD: use of GT motion masks

DIOD showed to handle different sources of noise contained in the motion guidance signal. This includes issues such as incomplete objects, missing labels, random background segments, and merging of nearby objects (refer to figure 6 for examples on this noise). In this section, we conduct an ablation study to assess DIOD's effectiveness in completing sparse labels. For this purpose, we use the ground truth (GT) masks of moving objects, which represent a **sparse** but **clean** signal. We conduct the experiment on TRI-PD dataset and compare DIOD with the latest motion-guided object discovery methods, all of which use the same GT masks. The improvement brought by DIOD (table .6) reflects its ability to complete sparse labels.

Method	Bao et al.[2]	MoTok[3]	BMOD([2])[18]	BMOD([3])[18]	DIOD
Fg-ARI	59.6	76.3	74.1	81.5	**87.5**

Table .6. Object discovery performance over foreground regions using the ground truth motion segments on TRI-PD dataset.

E. Illustrating the motion-guidance signal

In the main paper, we discussed the various types of noise encountered in the motion guidance signal (masks of moving objects). We present in figure 6 examples of this motion-supervision on several scenes from the training data. These examples illustrate the types of noise described in the main paper, providing a qualitative assessment of the noise level and the sparsity of this supervision: the absence of static objects and the omission of several moving objects. In the second column, we display the outputs of our method (the student model) on these same scenes at the end of training. These results particularly demonstrate the handling of noise in static regions and the completion of missing labels. It is important to note that this information in the student model is being continuously transferred to the teacher during training. This results in an improved quality of supervision provided by the teacher model, which justifies the favorable results obtained by DIOD.

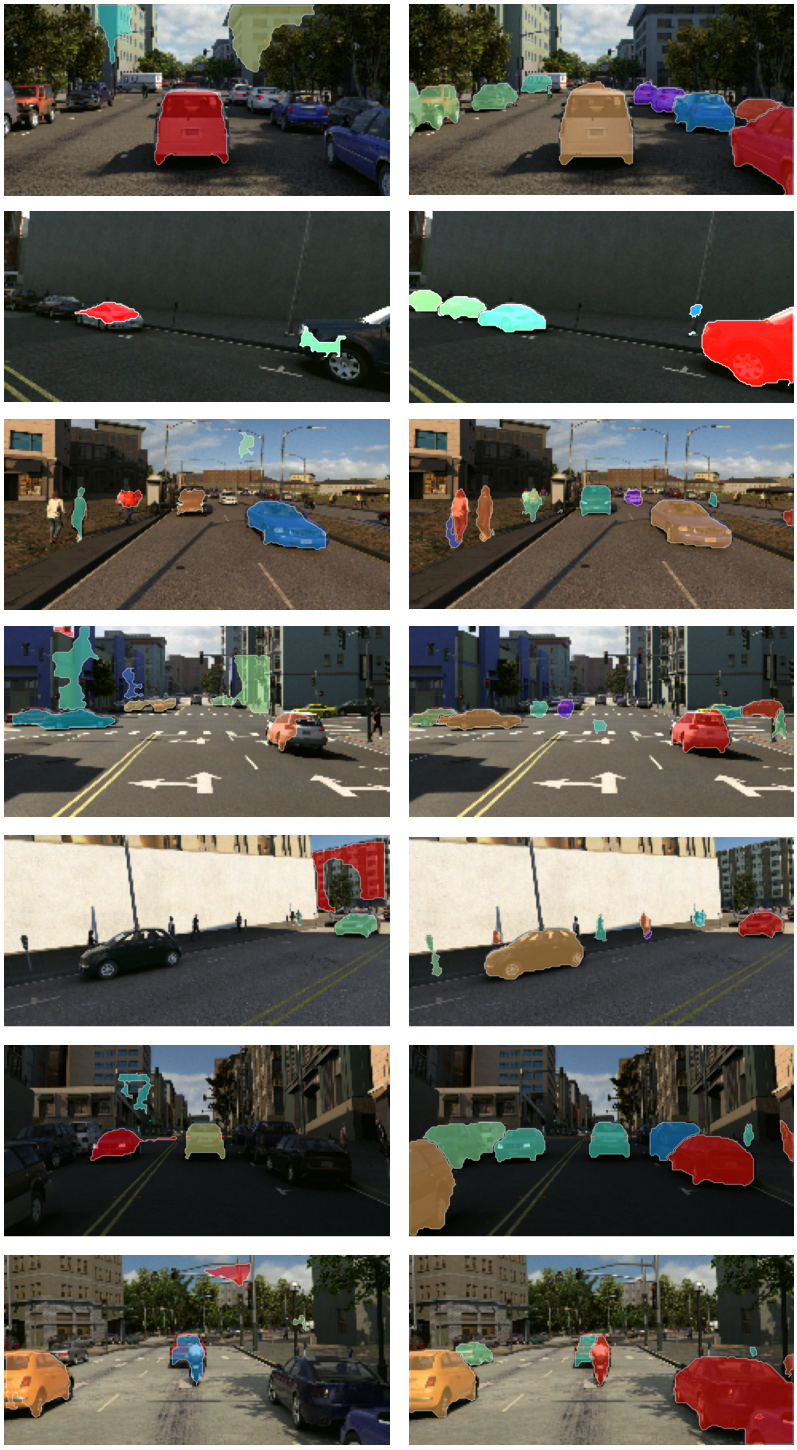

Figure 6. Illustration of the improvement achieved by DIOD as shown in column 2, leveraging the noisy and sparse motion masks (column 1) to direct the slots' attention. These example frames are from the training set of TRI-PD [2]

F. One-to-one vs. one-to-many

In the main paper, we conducted a quantitative comparison of two configuration choices in the design of DIOD as detailed in section 3.3.2. We illustrate in figure 7 the qualitative differences between them. In the one-to-one configuration, treating the content of a slot —which may capture multiple objects— as a single instance for supervision from the teacher introduces a semantic bias. This is evident in figure 7 (row 1), where the model heavily relies on semantic cues, resulting in imprecise localization of instances. This manifests as imprecise boundaries, merging of nearby objects belonging to the same semantic class, and even the introduction of noise by extending into semantically similar regions near the objects of interest (e.g., mistaking tree trunks or poles for the *person* class). These issues are addressed in the second configuration adopted in DIOD, where the model better represents individual instances (row 2).

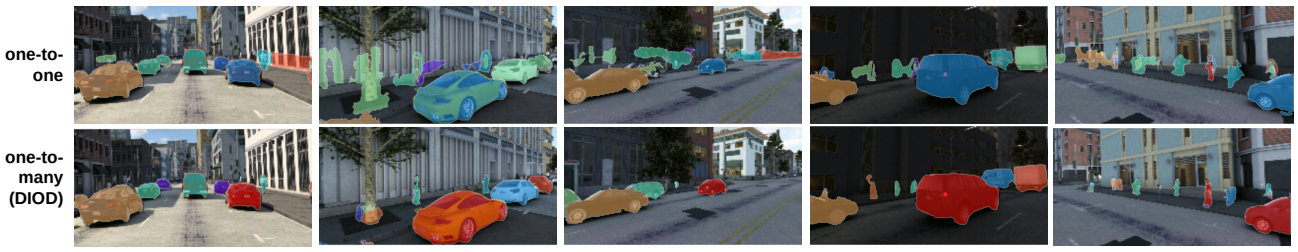

Figure 7. Qualitative comparison between *one-to-one* and *one-to-many* configurations in the connection of teacher and student models' attentions.