Supplementary Material:
Visual Modality Prompt for Adapting Vision-Language Object Detectors

Heitor R. Medeiros* Atif Belal Srikanth Muralidharan
Eric Granger Marco Pedersoli
LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
International Laboratory on Learning Systems (ILLS)

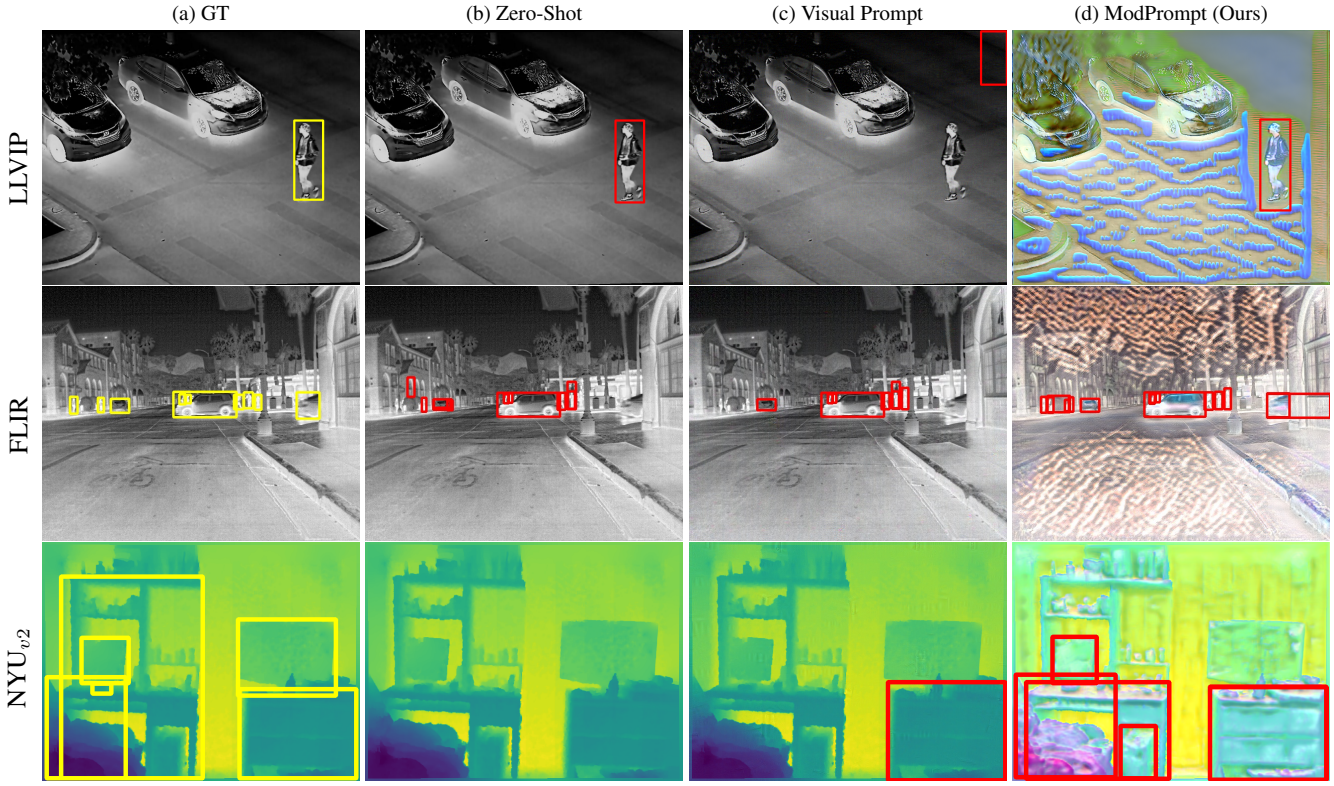

Figure 1. **Detections of different approaches across modalities**: LLVIP and FLIR datasets (infrared) and NYU$_{v2}$ (depth). Each column corresponds to a different approach: **(a) GT (Ground Truth):** Shows in yellow the ground-truth bounding boxes for objects. **(b) Zero-Shot:** Displays detections (in red) from a zero-shot model. This model misses several detections and predicts inaccurate boxes without specific tuning. **(c) Visual Prompt:** Illustrates detections with weight map visual prompt added to the image. It shows improvements over zero-shot, with more accurate detection, but still misses some objects. **(d) ModPrompt (Ours):** Detections from our proposed model. ModPrompt generates artifacts on the image to enhance objects and suppress the background, facilitating the detector.

In the supplementary material, we provide additional information to reproduce our work. The source code is available at https://github.com/heitorrapela/ModPrompt. The supplementary material is divided into the following sections: Section 1 with additional details regarding the implementation and architecture of the vision-language object detectors used in our work. Then in Section 2 we formally define text-prompt tuning in more detail. And in Section 3 we provide some additional results. Specifically, in Section 3.1 we provide additional main results on the FLIR-IR dataset. Then, we provide results with different backbones, YOLO-World-Large and Grounding DINO-

*Contact: heitor.rapela-medeiros.1@ens.etsmtl.ca

B in Section 3.2 on FLIR-IR and NYU$_{v2}$-DEPTH. Further, in Section 3.3 we provide the results for the FLIR-IR dataset with the learnable MPDR. In Section 3.4 we provide results of ablations using different visual prompt strategies. Then, in Section 3.5 we compare our method with state-of-the-art modality translation OD methods, and in Section 3.6 we show additional visualizations. Finally, in Section 4 we provide some limitations of our work and possible future directions.

1. Additional Details of Vision-Language ODs

For the YOLO-World, we use AdamW optimizer with a learning rate $2e^{-4}$, weight decay 0.05, and batch size 8. And for the Grounding-DINO, we use AdamW optimizer with a learning rate $1e^{-4}$, weight decay $1e^{-4}$, and batch size 8. For the main manuscript, we used YOLO-World Small and Grounding-DINO Tiny. For the experiments with text, we extract the embeddings and optimize them without the text encoder for efficient adaptation of the embedding space. Additionally, we provide results with bigger backbones to further corroborate our findings.

2. Formal definition of Text-Prompt Tuning

Following our definitions of visual prompts for object detection in the main manuscript, we define the text-prompt tuning using our notation in this section. YOLO-World follows the YOLOv8 loss from Jocher et al. [5], with a text contrastive head to provide the object-text similarities; for more details about YOLO-World loss check [1]. Here, we provide a generic definition, independent of the model. Thus, we define the text-prompt cost function ($\mathcal{C}_{\text{tp}}(\phi)$), with the following Equation:

$$\mathcal{C}_{\text{tp}}(\phi) = \frac{1}{|\mathcal{D}|} \sum_{(x_t, Y) \in \mathcal{D}} \mathcal{L}_{text}(g_\psi(x_t + h_\phi), Y), \quad (1)$$

where x_t is the input text, g_ψ is the text-encoder, h_ϕ is the additional prompt parametrized by ϕ. In the case of YOLO-World, $\mathcal{L}_{text}$ can be seen as label assignment of Feng et al. [2] to match the predictions with ground-truth annotations, with Binary Cross Entropy (BCE), and assign each positive prediction with a text index as the classification label.

3. Additional Results

3.1. Main Results on FLIR-IR data:

In Table 1 we compare the performance of our method against the baselines on the FLIR-IR dataset. It can be observed that our ModPrompt achieves the highest performance in terms of APs for YOLO-World, and for Grounding DINO, the AP$_{50}$ and AP results were the highest, while the AP$_{75}$ is equally good as the random prompt. The FLIR-IR dataset is a more challenging dataset composed of 3 classes, some small bounding boxes, and missing annotations, which make the problem more difficult when the input image is being changed only with detector feedback and without the availability of the RGB Groundtruth images for image-to-image translation during training. We observe that ModPrompt performs better when objects are well-defined in the image and when objects are not too small, otherwise, like all other input-level pixel strategies it faces challenges, especially on refined bounding-box localization, which can be seen with AP$_{75}$ and AP, whereas in AP$_{50}$ it always shows good results.

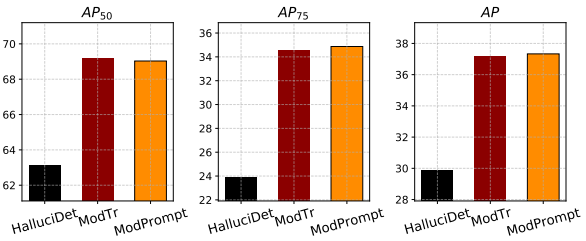

Figure 2. **Detection performance on FLIR-IR dataset of different modality translators for OD in terms of APs.**

3.2. Results with Different Detection Backbones:

In this section, we provide results for the YOLO-World-Large and Grounding DINO-Big models with different visual prompt strategies and our ModPrompt. In Table 2, we show that ModPrompt is better than all visual prompt methods for FLIR and NYU$_{v2}$.

3.3. MPDR with FLIR-IR data:

In Table 3, we provide additional results for FLIR-IR with our MPDR module. We emphasize that the knowledge preservation strategy improves performance in many cases. However, this dataset is too noisy, which compromises translation methods such as ModPrompt, resulting in degradation of performance in some cases.

3.4. Ablation Studies on Visual Prompts:

We evaluate different variations of the visual prompt adaptation methods. Specifically, we compare the performance when different input patch sizes are used; for instance, p$_s$ = 30 refers to a patch size of 30 pixels. In this study, we test multiple patch sizes for each of the visual prompt methods and report the performance in Table 4. We evaluate ModPrompt using two different translators with U-Net based backbones, MobileNet (MB) [4] and ResNet (RES) [3]. Here, we provide the additional results on the FLIR-IR dataset.

Dataset	Method	YOLO-World			Grounding DINO		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
FLIR - IR	Zero-Shot (ZS)	64.70 ± 0.00	32.10 ± 0.00	34.90 ± 0.00	64.30 ± 0.00	29.30 ± 0.00	32.90 ± 0.00
	Head Finetuning (HFT)	73.90 ± 0.14	36.47 ± 0.12	40.00 ± 0.00	67.27 ± 0.15	33.60 ± 0.26	36.20 ± 0.10
	Full Finetuning (FT)	80.47 ± 1.11	41.13 ± 0.41	44.17 ± 0.39	80.73 ± 0.06	42.17 ± 0.78	44.17 ± 0.25
	Visual Prompt (Fixed)	45.60 ± 0.08	21.90 ± 0.08	23.77 ± 0.05	64.27 ± 0.06	29.47 ± 0.06	32.83 ± 0.12
	Visual Prompt (Random)	43.27 ± 0.29	20.63 ± 0.17	22.47 ± 0.12	64.03 ± 0.08	**29.50 ± 0.00**	32.77 ± 0.15
	Visual Prompt (Padding)	45.63 ± 1.44	22.13 ± 1.08	23.87 ± 0.91	61.73 ± 0.23	27.60 ± 0.17	31.20 ± 0.20
	Visual Prompt (WM)	54.43 ± 0.78	26.37 ± 0.54	28.67 ± 0.40	54.73 ± 0.08	23.20 ± 0.10	27.20 ± 0.10
	Visual Prompt (WM_{v2})	52.43 ± 0.50	25.10 ± 0.22	27.50 ± 0.22	54.80 ± 0.10	23.13 ± 0.21	27.27 ± 0.06
	ModPrompt (Ours)	**69.03 ± 1.06**	**34.87 ± 0.40**	**37.33 ± 0.12**	**65.03 ± 0.05**	29.23 ± 0.55	**32.90 ± 0.26**

Table 1. **Detection performance (APs) for YOLO-World and Grounding DINO for the FLIR-IR dataset.** The different visual prompt adaptation techniques are compared with our ModPrompt, and the zero-shot (ZS), head finetuning (HFT), and full finetuning (FT) are also reported, where the full finetuning is the upper bound.

Detector	Method	FLIR-IR			NYU_{v2}-DEPTH		
		AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP
YOLO-World-Large	Zero-Shot (ZS)	71.60 ± 0.00	37.60 ± 0.00	39.30 ± 0.00	05.30 ± 0.00	03.70 ± 0.00	03.50 ± 0.00
	Head Finetuning (HFT)	82.27 ± 0.21	43.53 ± 0.12	45.93 ± 0.05	24.43 ± 0.17	14.63 ± 0.29	14.53 ± 0.17
	Full Finetuning (FT)	84.33 ± 0.45	44.00 ± 0.29	46.70 ± 0.22	54.13 ± 0.05	40.43 ± 0.09	37.33 ± 0.17
	Visual Prompt (Fixed)	71.83 ± 0.09	37.70 ± 0.00	39.40 ± 0.00	05.20 ± 0.00	03.60 ± 0.00	03.40 ± 0.00
	Visual Prompt (Random)	71.40 ± 0.14	37.40 ± 0.08	39.23 ± 0.05	04.67 ± 0.12	03.17 ± 0.09	02.97 ± 0.09
	Visual Prompt (Padding)	66.27 ± 0.12	33.83 ± 0.05	35.97 ± 0.05	02.87 ± 0.12	01.80 ± 0.08	01.80 ± 0.08
	Visual Prompt (WM)	71.70 ± 0.14	37.53 ± 0.25	39.43 ± 0.05	15.10 ± 0.45	09.83 ± 0.26	09.50 ± 0.24
	Visual Prompt (WM_{v2})	70.90 ± 0.08	36.73 ± 0.19	38.73 ± 0.05	14.53 ± 0.41	09.57 ± 0.33	09.13 ± 0.31
	ModPrompt (Ours)	**77.13 ± 0.29**	**41.37 ± 0.69**	**43.23 ± 0.24**	**39.27 ± 0.53**	**28.93 ± 0.17**	**26.73 ± 0.09**
Grounding DINO-Big	Zero-Shot (ZS)	64.80 ± 0.00	29.10 ± 0.00	32.90 ± 0.00	08.50 ± 0.00	05.70 ± 0.00	05.40 ± 0.00
	Head Finetuning (HFT)	68.40 ± 0.14	34.60 ± 0.28	36.95 ± 0.21	08.50 ± 0.00	06.10 ± 0.10	05.67 ± 0.06
	Full Finetuning (FT)	81.97 ± 0.25	34.60 ± 0.28	45.85 ± 0.07	53.93 ± 0.40	40.53 ± 0.15	37.50 ± 0.26
	Visual Prompt (Fixed)	64.87 ± 0.06	29.17 ± 0.06	33.00 ± 0.00	08.53 ± 0.06	05.73 ± 0.06	05.53 ± 0.06
	Visual Prompt (Random)	64.83 ± 0.06	29.13 ± 0.06	32.93 ± 0.06	08.53 ± 0.06	05.77 ± 0.15	05.47 ± 0.06
	Visual Prompt (Padding)	62.63 ± 0.06	27.27 ± 0.06	31.53 ± 0.06	07.93 ± 0.06	05.13 ± 0.06	04.83 ± 0.06
	Visual Prompt (WM)	56.77 ± 0.31	21.93 ± 0.55	27.00 ± 0.36	05.57 ± 0.06	03.23 ± 0.06	03.17 ± 0.06
	Visual Prompt (WM_{v2})	57.03 ± 0.40	22.20 ± 0.40	27.20 ± 0.30	05.73 ± 0.06	03.33 ± 0.06	03.33 ± 0.06
	ModPrompt (Ours)	**65.73 ± 0.15**	**30.07 ± 0.12**	**33.47 ± 0.12**	**25.30 ± 0.53**	**17.60 ± 0.20**	**16.57 ± 0.35**

Table 2. **Detection performance (APs) for YOLO-World-Large and Grounding DINO-B on FLIR-IR and NYU_{v2}-Depth datasets.** Each visual prompt adaptation strategy is compared with our ModPrompt.

3.5. Comparison with SOTA Modality Translation OD methods:

Our ModPrompt technique is compared with recent state-of-the-art modality translation methods for ODs: HalluciDet [7] and ModTr [6]. In Fig. 2, we observe that our results are better in all APs for the FLIR dataset.

3.6. Qualitative Results:

In this section, we provide more visual results for the methods compared, where the performance of each model can be shown by the bounding box predictions. For instance, in Figure 1, we can see that Zero-Shot (ZS) is performing well on the FLIR-IR dataset, apart from some false positives. However, the ZS doesn't perform well on the modality that is too different from the pre-training weight, such as depth (NYU_{v2} dataset). For the Visual Prompt (weight map), we have some false positives and missing detections (e.g., the person in the first row for LLVIP and a wrong bounding box). For ModPrompt, we see that we have some good overall detections because the encoder-decoder architecture tends to suppress a little bit of background, as we can see in the first row or second row, but the bounding boxes are not as precise as the right ones in the ZS (see

Detector	Method	FLIR-IR		
		AP_{50}	AP_{75}	AP
YOLO-World	**Fixed**	63.93 ± 0.26 (+0.63)	31.90 ± 0.08 (-0.60)	34.63 ± 0.05 (+0.33)
	Random	63.30 ± 0.14 (+0.43)	31.73 ± 0.09 (-0.44)	34.27 ± 0.12 (+0.17)
	Padding	59.23 ± 0.12 (+0.66)	29.10 ± 0.08 (-0.10)	31.50 ± 0.08 (+0.17)
	WeightMap	63.10 ± 0.28 (+0.33)	31.60 ± 0.29 (-0.10)	34.20 ± 0.08 (+0.33)
	ModPrompt	**72.73 ± 0.00 (-1.74)**	**37.70 ± 0.00 (-0.60)**	**40.37 ± 0.00 (-0.73)**
Grounding DINO	**Fixed**	66.53 ± 0.98 (+2.26)	**32.83 ± 2.50 (+3.36)**	**34.83 ± 0.90** (+2.00)
	Random	66.10 ± 1.08 (+2.07)	31.40 ± 1.31 (+1.90)	34.53 ± 0.90 (+1.76)
	Padding	63.33 ± 1.10 (+1.60)	29.73 ± 1.27 (+2.13)	32.93 ± 0.90 (+1.73)
	WeightMap	55.60 ± 0.92 (+0.87)	24.37 ± 0.65 (+1.17)	28.30 ± 0.66 (+1.10)
	ModPrompt	**67.80 ± 0.14 (+2.77)**	31.10 ± 0.31 (+1.87)	33.70 ± 0.09 (+0.80)

Table 3. **Detection performance (APs) for YOLO-World and Grounding DINO on FLIR-IR data.** Each visual prompt adaptation strategy is compared with the learnable MPDR (results in parentheses are the gain with the MPDR module), which is responsible for updating the new modality embeddings and not changing the original embedding knowledge.

Method	Variation	FLIR - IR		
		AP_{50}	AP_{75}	AP
Fixed	30	45.60 ± 0.08	21.90 ± 0.08	23.77 ± 0.05
	300	29.30 ± 0.37	13.50 ± 0.54	15.00 ± 0.37
Random	30	43.27 ± 0.29	20.63 ± 0.17	22.47 ± 0.12
	300	19.13 ± 0.33	09.00 ± 0.42	09.80 ± 0.33
Padding	30	45.63 ± 1.44	22.13 ± 1.08	23.87 ± 0.91
	200	00.53 ± 0.12	00.17 ± 0.17	00.27 ± 0.09
ModPrompt	MB	66.80 ± 0.29	**35.23 ± 0.38**	36.53 ± 0.12
	RES	**69.03 ± 1.06**	34.87 ± 0.40	**37.33 ± 0.12**

Table 4. **Detection performance (APs) for YOLO-World on FLIR-IR data.** We compared the main visual prompt strategies *fixed*, *random*, *padding*, and ModPrompt. The variations consist of the number of prompt pixels ($p_s = 30$, 200 or 300) and for ModPrompt, the MobileNet (MB) or ResNet (RES).

first-row comparison between ZS and ModPrompt). Additionally, in Figure 3, we provide a batch of 8 images from the test, and we can observe a similar trend for depth and IR modalities as discussed above. Surprisingly, in some cases of the FLIR dataset (challenging dataset with really hard small bounding boxes and some missing labels), our method tends to detect objects that are not labeled in the ground truth (for instance a small person in the first row and first column, behind the two cars on the left, which are not labeled, but our method get its right). This shows the effectiveness of our method and exemplifies the ability of visual language detectors to detect unseen objects.

4. Limitations and future works

Limitations: We believe our work still has some limitations, which we believe can be further improved in subsequent works. For instance, adaptation strategies still require target labels, which could be explored in other tasks, such as unsupervised or semi-supervised approaches. Additionally, we argue that our method is still not perfect for small objects, and it incorporates some noise, which can be further minimized by other loss constraints if we have access to additional source data (which we didn't during training). Some of the limitations were already discussed in the qualitative results, which can be summarized as difficulties with small objects and duplications of bounding box predictions.

Future works: Future works could improve the conditioning on both text and vision, and exploit more label-efficient adaptation strategies such as test-time adaptation or few-shot learning.

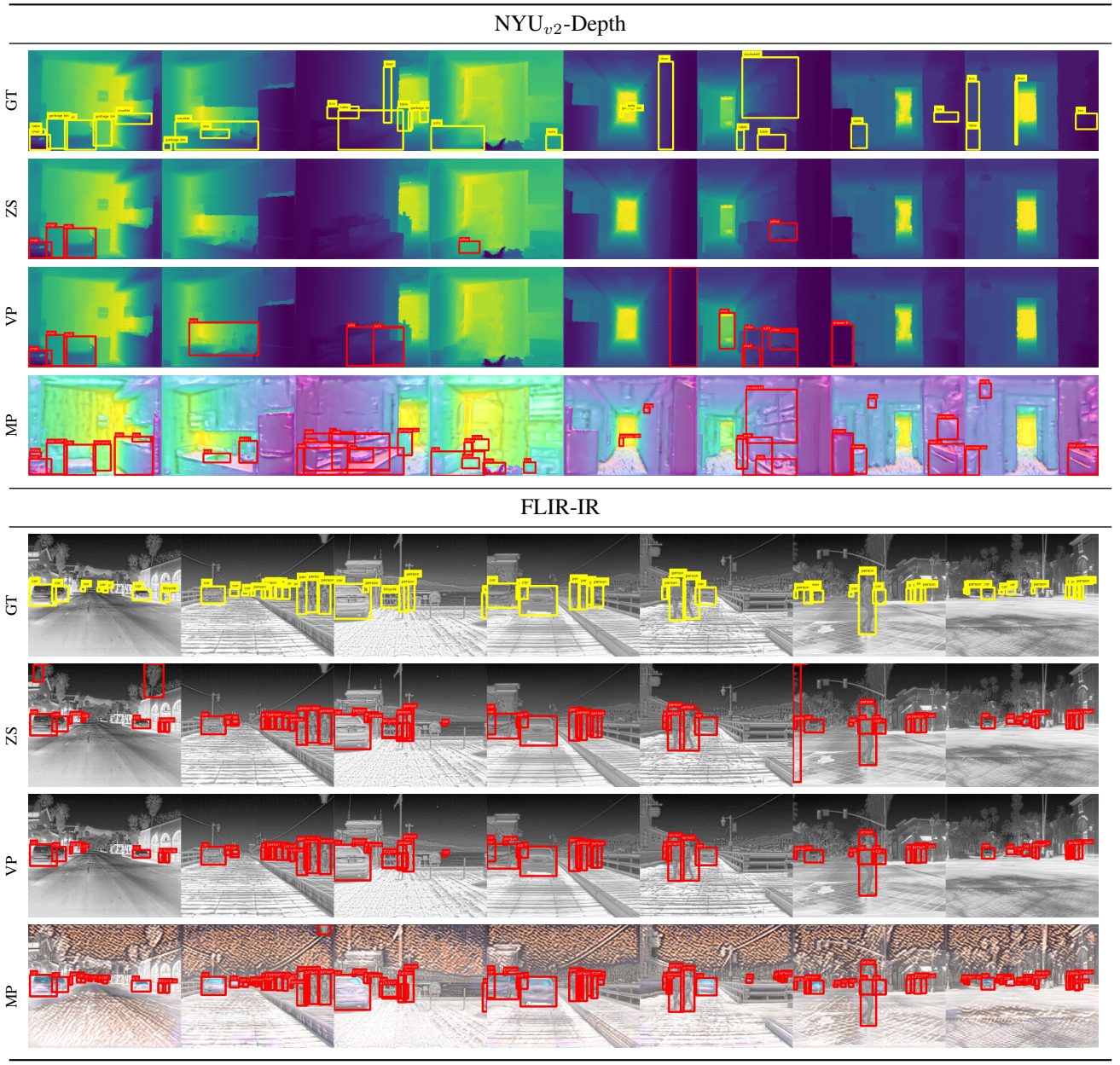

Figure 3. **Detections of different approaches across modalities for YOLO-World**: NYU$_{v2}$ (depth) and FLIR (infrared). Each row corresponds to a different approach: **GT (Ground Truth):** Shows in yellow the ground-truth bounding boxes for objects. **ZS (Zero-Shot):** Displays detections (in red) from a zero-shot model YOLO-World-s. **VP (Visual Prompt):** Illustrates detections with weight map visual prompt added to the image. **MP (ModPrompt):** Detections from our proposed model.

References

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2

[2] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3490–3499. IEEE Computer Society, 2021. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[5] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023. 2

[6] Heitor Rapela Medeiros, Masih Aminbeidokhti, Fidel Alejandro Guerrero Peña, David Latortue, Eric Granger, and Marco Pedersoli. Modality translation for object detection adaptation without forgetting prior knowledge. In *European Conference on Computer Vision*, pages 51–68. Springer, 2024. 3

[7] Heitor Rapela Medeiros, Fidel A Guerrero Pena, Masih Aminbeidokhti, Thomas Dubail, Eric Granger, and Marco Pedersoli. Hallucidet: Hallucinating rgb modality for person detection through privileged information. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1444–1453, 2024. 3