ICCV HONOLULU HAWAII
OCT 19-23, 2025

# Visual Modality Prompt for Adapting Vision-Language Object Detectors

#10337

**Heitor R. Medeiros, Atif Belal, Srikanth Muralidharan, Eric Granger, Marco Pedersoli**
LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada

ILLS International Laboratory on Learning Systems

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Université du Québec ÉTS Engineering for Industry

LIVIA LABORATOIRE D'IMAGERIE, DE VISION ET D'INTELLIGENCE ARTIFICIELLE

DISTECH CONTROLS™

# Outline

1. Background and Motivation
2. Related Works
3. ModPrompt
4. Experiments
5. Conclusion

# 1. Background and Motivation

# 1.1 - Object Detectors

- Object detection is the task of locating objects on images.
  - Input: Image.
  - Output: class labels and bounding-box of detected objects.


**Figure 1.** Object detection task. Image taken from [1].

[1] Li et al. CS231n: Convolutional Neural Networks for Visual Recognition (Lecture 11). accessed 22 March 2022, http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf.

# 1.2 - Vision-Language Models (VLMs)

- A VLM is composed of an image encoder and a text encoder.

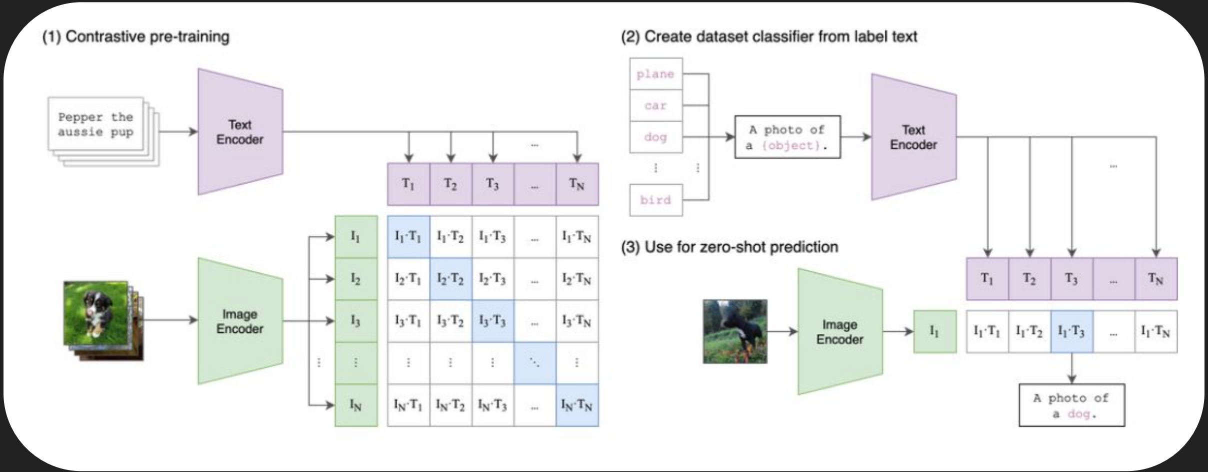



**Figure 2.** CLIP framework [2].

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

# 1.3 - What are VL-ODs?

- Vision-Language Object Detectors (VL-ODs):
  - They have a text encoder, a vision encoder, and a fusion head.
  - Advantages: open-vocabulary, zero-shot detection.

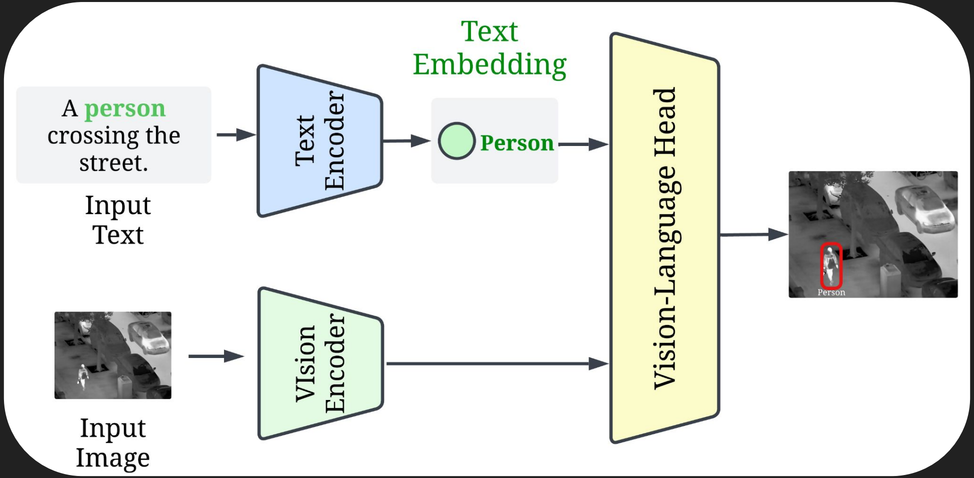



**Figure 3.** VL-OD illustration.

# 1.4 - Modality Adaptation

- Modality adaptation uses detection feedback for image translation.
- Adapts detectors for different input distributions.

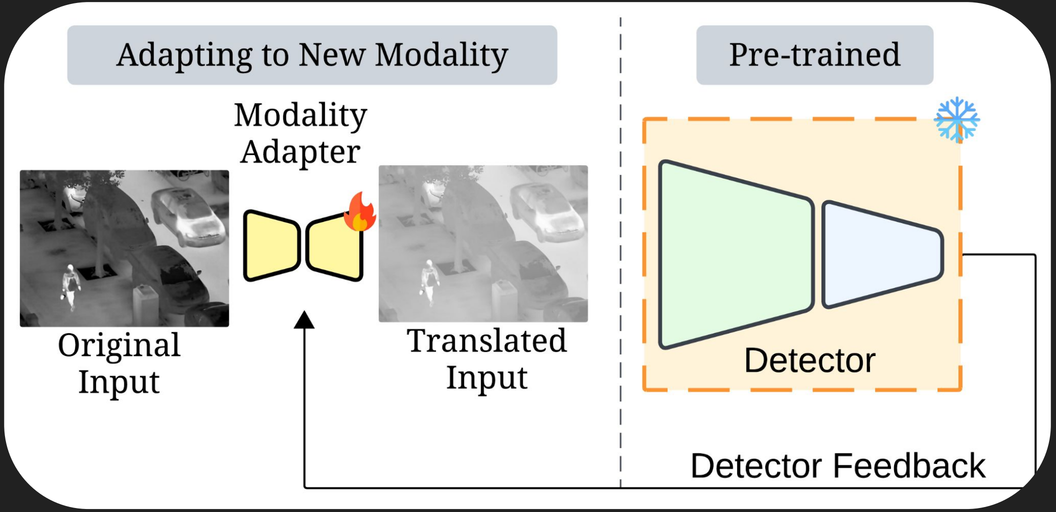



**Figure 4.** Modality Adaptation framework.

# 2. Related Works

# 2.1 - Vision-Language Object Detectors

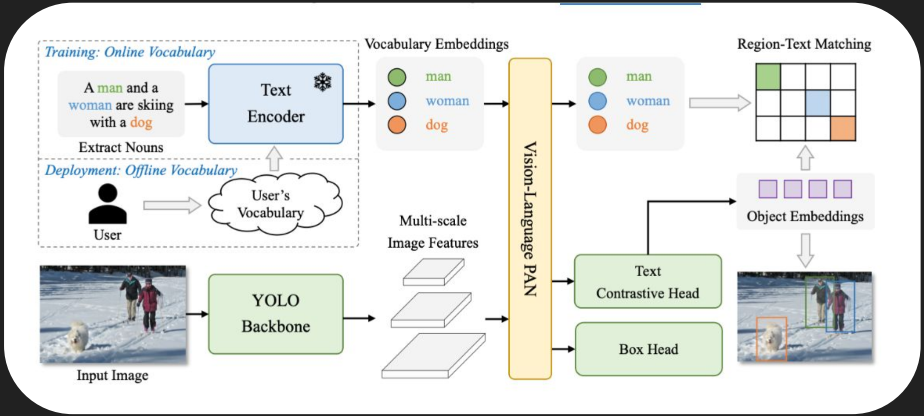



**Figure 5.** Yolo-World architecture [3].

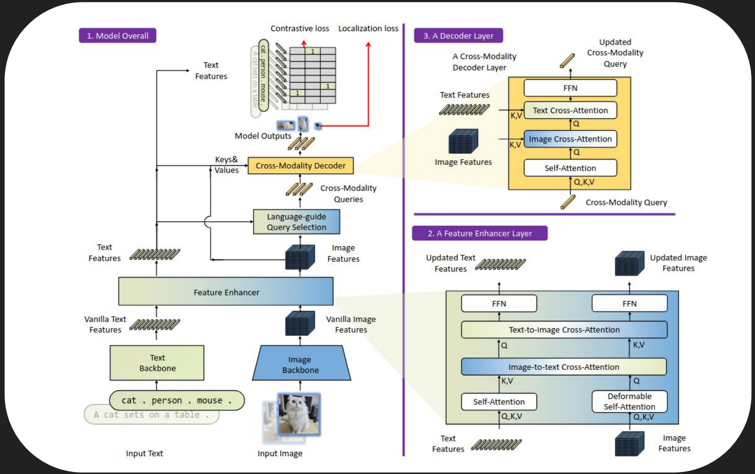



**Figure 6.** Grounding DINO architecture [4].

- YOLO-World is pre-trained on large-scale data. It re-parameterizes vocabulary embeddings as parameters into the model and achieve superior inference speed.
- Grounding DINO effectively fuse language and vision modalities, it proposes a tight fusion solution.


[3] Cheng, Tianheng, et al. "Yolo-world: Real-time open-vocabulary object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.
[4] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

## 2.2 - Modality Adaptation

- HalluciDet leverages RGB detector knowledge to guide IR-to-RGB translation for improved detection with a task-driven hallucination loss.
- ModTr adapts new-modality inputs (e.g. IR) via a small translator network so the original RGB-trained detector can be reused unchanged. It preserves the original detector's.

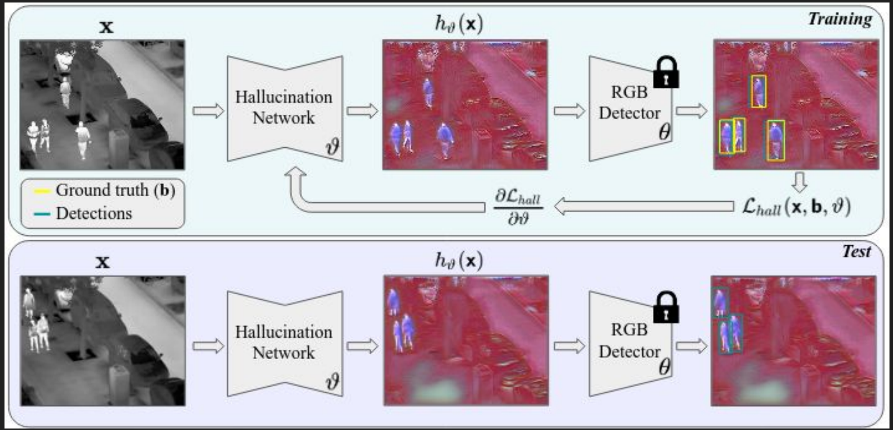



**Figure 7.** HalluciDet framework [5].

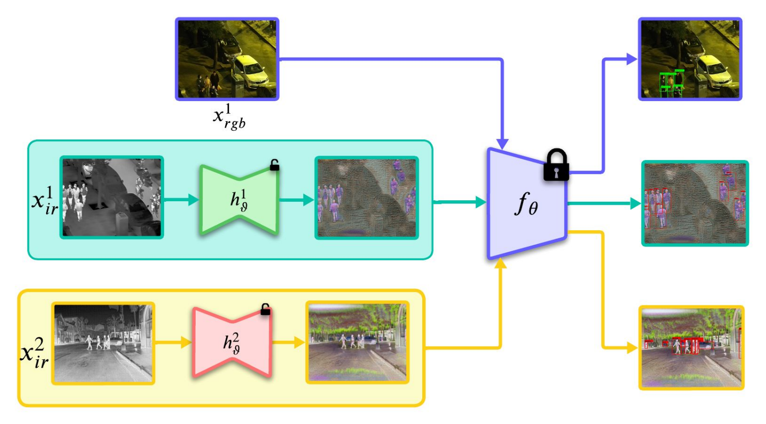



**Figure 8.** ModTr framework [6].


[5] Medeiros, Heitor Rapela, et al. "HalluciDet: hallucinating RGB modality for person detection through privileged information." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
[6] Medeiros, Heitor Rapela, et al. "Modality translation for object detection adaptation without forgetting prior knowledge." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

# 2.3 - Different strategies to adapt to new modalities

- Our work investigates how to efficiently adapt VL-ODs.
- VL-ODs suffer under modality shift:
    - IR, depth, event-based, LiDAR.

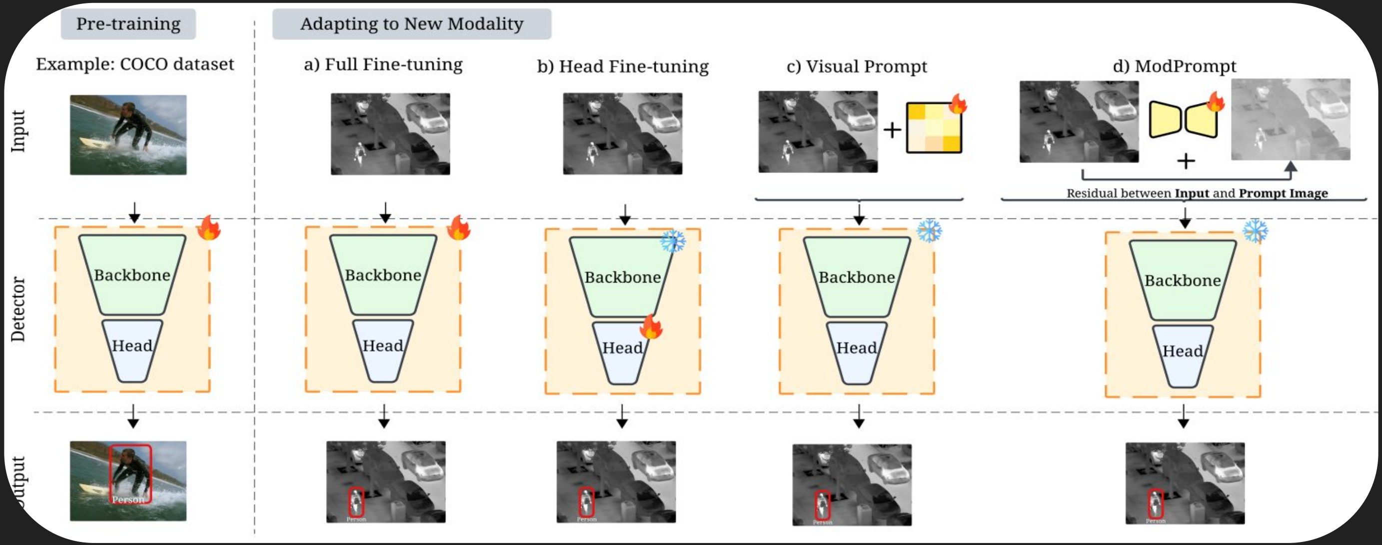



**Figure 9.** Different adaptation approaches for new modalities.

# 2.3 - Different strategies to adapt to new modalities

- Full fine-tuning VL-ODs is too costly.
- Previous methods for modality adaptation did not investigate VL-ODs, and visual prompt strategies focused on the downstream classification task.

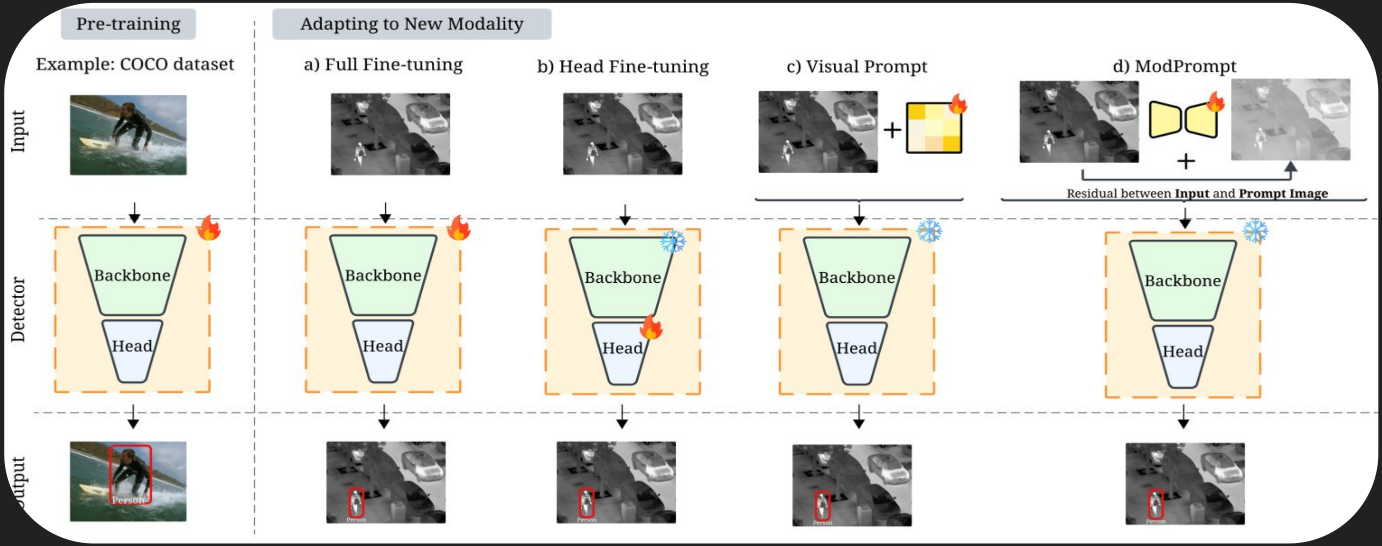



**Figure 9.** Different adaptation approaches for new modalities.

# 3. ModPrompt

# 3.1 - ModPrompt - Main Contributions

- ModPrompt translates inputs at the pixel level for better modality alignment while preserving encoder knowledge via a backbone-agnostic design.

- It overcomes the failure of traditional pixel-level prompts, yielding superior cross-modality detection.

- ModPrompt achieves near fine-tuning performance across diverse modalities while retaining zero-shot cability.

# 3.2 - ModPrompt

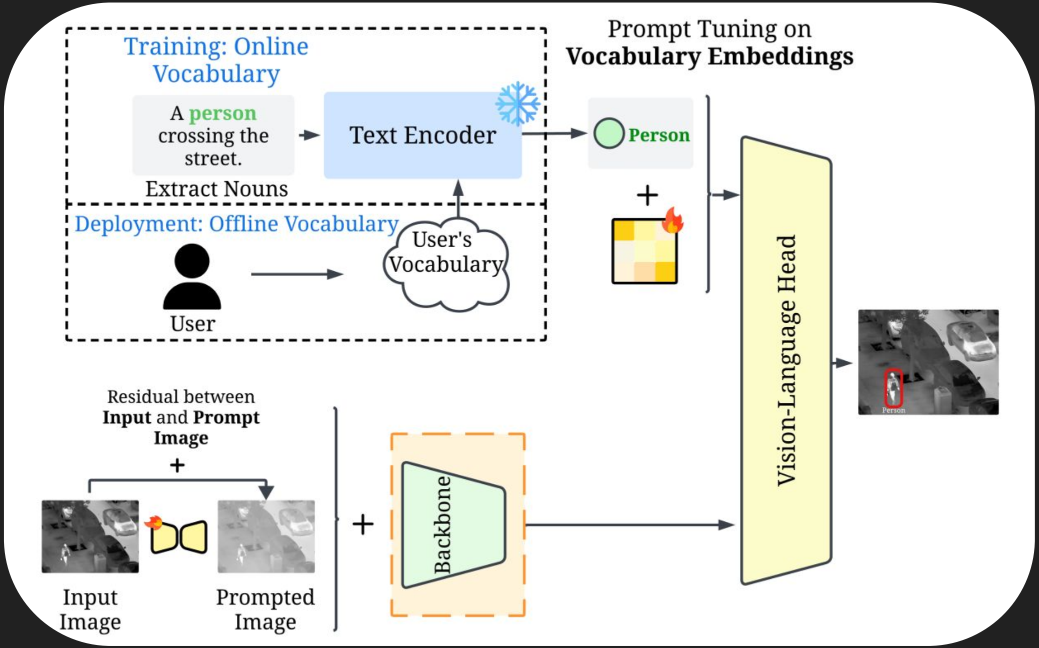



**Figure 10.** Our Proposed: **ModPrompt**.

- The detector output is used to calculate the ModPrompt loss and update its parameters.
- ModPrompt loss is defined as follows:

$$\mathcal{C}_{\text{mp}}(\vartheta) = \frac{1}{|\mathcal{D}|} \sum_{(x,Y)\in\mathcal{D}} \mathcal{L}_{det}(f_\theta(x + h_\vartheta(x)), Y)$$

# 4. Results

# 4.1 - Datasets and Evaluation

- **Datasets:** LLVIP [7], FLIR [8] and NYUv2 [9].

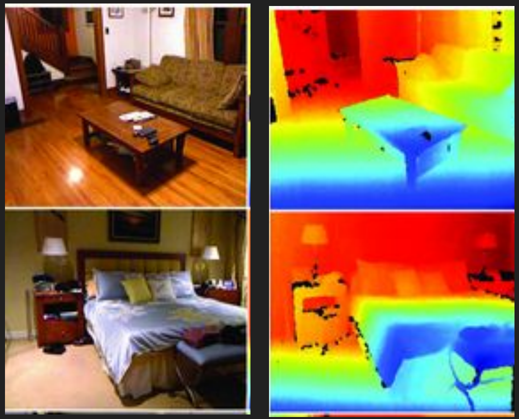
LLVIP

IR

FLIR

IR

NYUv2



Depth

- **Evaluation:** AP detection performance.

[7] Jia, Xinyu, et al. "LLVIP: A Visible-infrared Paired Dataset for Low-light Vision." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
[8] FLIR Thermal Dataset. Accessed: Jan. 23, 2025. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form
[9] Silberman, Nathan, et al. "Indoor segmentation and support inference from rgbd images." European conference on computer vision. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

# 4.2 - Comparison with Visual Prompt Strategies

| Dataset | Method | YOLO-World | | | Grounding DINO | | |
|---|---|---|---|---|---|---|---|
| | | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ | AP |
| LLVIP - IR | Zero-Shot (ZS) | 81.00 ± 0.00 | 57.80 ± 0.00 | 53.20 ± 0.00 | 85.50 ± 0.00 | 62.70 ± 0.00 | 56.50 ± 0.00 |
| | Head Finetuning (HFT) | 93.57 ± 0.05 | 73.83 ± 0.19 | 64.80 ± 0.08 | 87.53 ± 0.06 | 65.57 ± 0.23 | 58.10 ± 0.20 |
| | Full Finetuning (FT) | 97.43 ± 0.05 | 77.93 ± 0.21 | 67.73 ± 0.09 | 97.17 ± 0.31 | 79.93 ± 0.83 | 67.83 ± 0.96 |
| | Visual Prompt (Fixed) | 70.30 ± 7.89 | 45.67 ± 6.97 | 43.53 ± 5.79 | 83.83 ± 0.06 | 61.53 ± 0.23 | 55.13 ± 0.15 |
| | Visual Prompt (Random) | 60.13 ± 0.29 | 38.73 ± 0.17 | 36.87 ± 0.12 | 83.87 ± 0.06 | 61.37 ± 0.06 | 55.03 ± 0.06 |
| | Visual Prompt (Padding) | 79.87 ± 1.00 | 51.77 ± 0.90 | 49.30 ± 0.83 | 82.73 ± 0.31 | 60.00 ± 0.35 | 55.13 ± 0.15 |
| | Visual Prompt (WM) | 82.00 ± 1.59 | 53.90 ± 1.06 | 50.90 ± 0.94 | 69.57 ± 0.93 | 41.37 ± 1.27 | 40.77 ± 0.87 |
| | Visual Prompt ($WM_{v2}$) | 74.10 ± 0.43 | 46.47 ± 0.62 | 44.70 ± 0.22 | 69.87 ± 1.12 | 41.77 ± 1.30 | 41.13 ± 0.96 |
| | **ModPrompt (Ours)** | **92.80 ± 0.29** | **70.73 ± 1.02** | **62.87 ± 0.63** | **93.13 ± 0.15** | **67.17 ± 0.78** | **60.10 ± 0.50** |

**Table 1. Detection performance** (APs) for **YOLO-World** and **Grounding DINO** for **LLVIP-IR**.

# 4.3 - Comparison with SOTA Modality Adaptation

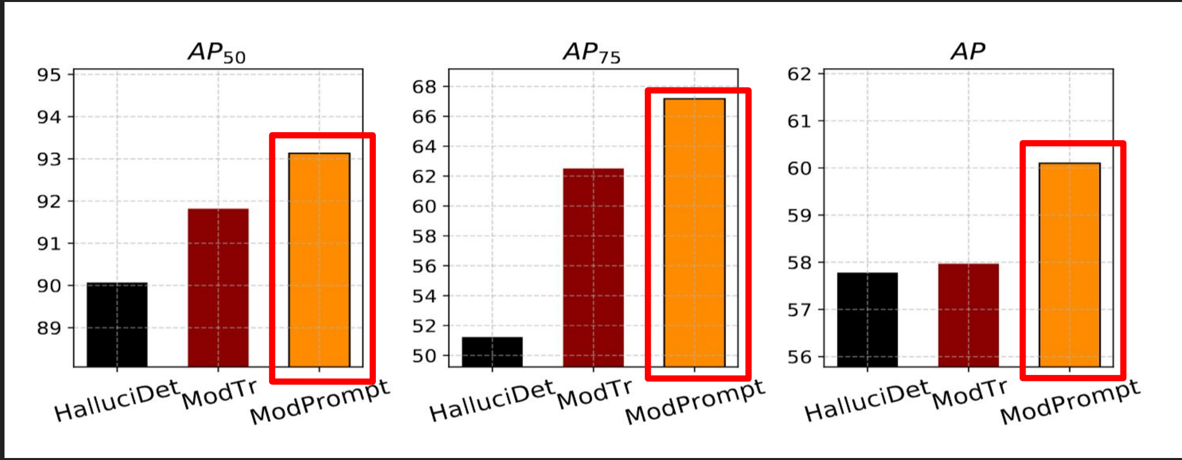



**Figure 11. Detection performance** on **LLVIP** for different **SOTA Modality Translation OD** methods.

- ModPrompt has **better localization** quality.
- **Improves** $AP_{50}$, $AP_{75}$, and AP over **HalluciDet** and **ModTr**.

# 4.4 - Knowledge Preservation

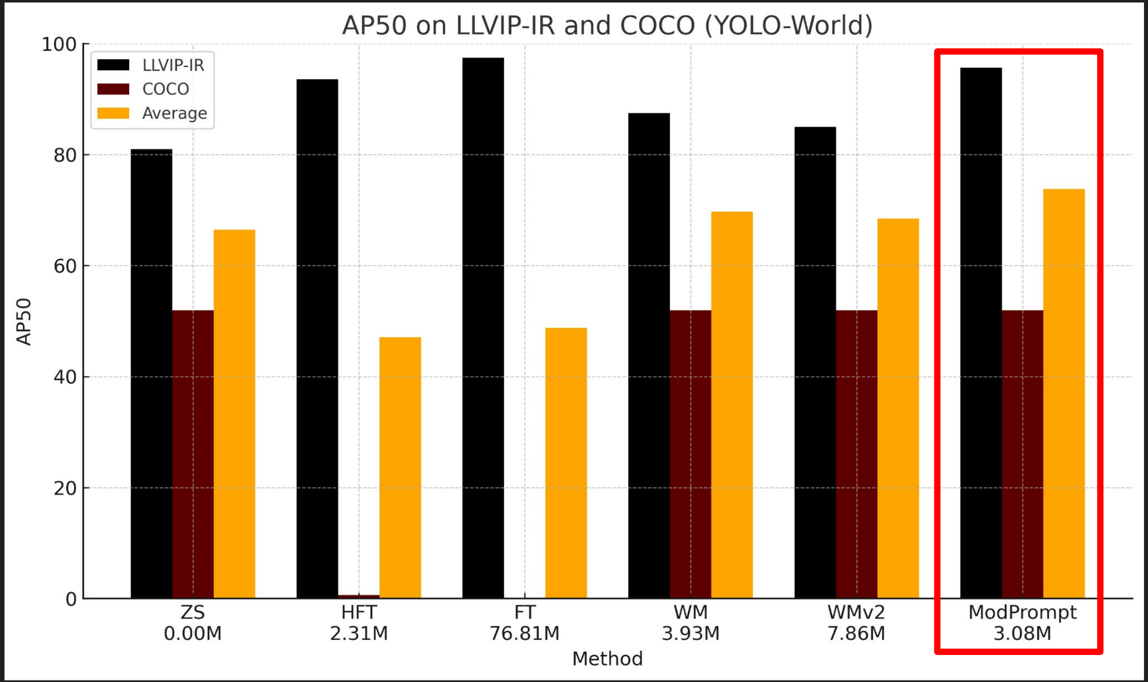



**Figure 12. AP50** of **YOLO-World** on **LLVIP-IR** and **COCO** data for **knowledge preservation**.

# 4.5 - Visual Prompt Ablation

| Method | Variation | LLVIP - IR | | |
|---|---|---|---|---|
| | | $AP_{50}$ | $AP_{75}$ | AP |
| **Fixed** | 30 | 61.60 ± 0.75 | 39.93 ± 0.52 | 37.97 ± 0.56 |
| | 300 | 70.30 ± 7.89 | 45.67 ± 6.97 | 43.53 ± 5.79 |
| **Random** | 30 | 60.13 ± 0.29 | 38.73 ± 0.17 | 36.87 ± 0.12 |
| | 300 | 56.27 ± 0.46 | 33.73 ± 0.62 | 33.13 ± 0.42 |
| **Padding** | 30 | 79.87 ± 1.00 | 51.77 ± 0.90 | 49.30 ± 0.83 |
| | 300 | 39.53 ± 2.36 | 15.90 ± 1.02 | 19.07 ± 1.18 |
| **ModPrompt** | MB | **92.80 ± 0.29** | **70.73 ± 1.02** | **62.87 ± 0.63** |
| | RES | 91.03 ± 0.12 | 68.40 ± 1.10 | 61.43 ± 0.58 |

**a)** LLVIP - IR

| Method | Variation | $NYU_{v2}$ - Depth | | |
|---|---|---|---|---|
| | | $AP_{50}$ | $AP_{75}$ | AP |
| **Fixed** | 30 | 04.67 ± 0.05 | 03.07 ± 0.05 | 02.90 ± 0.00 |
| | 300 | 03.43 ± 0.05 | 02.00 ± 0.08 | 02.10 ± 0.00 |
| **Random** | 30 | 04.23 ± 0.12 | 02.63 ± 0.05 | 02.53 ± 0.05 |
| | 300 | 01.53 ± 0.17 | 00.77 ± 0.12 | 00.87 ± 0.12 |
| **Padding** | 30 | 03.97 ± 0.05 | 02.50 ± 0.00 | 02.43 ± 0.05 |
| | 200 | 00.37 ± 0.12 | 00.10 ± 0.08 | 00.17 ± 0.05 |
| **ModPrompt** | MB | 35.37 ± 0.12 | 25.20 ± 0.24 | 23.27 ± 0.17 |
| | RES | **37.17 ± 0.57** | **27.50 ± 0.64** | **24.93 ± 0.50** |

**b)** NYUv2 - Depth

**Table 2.** Comparison of visual prompt strategies: fixed, random, padding, and ModPrompt. a) LLVIP and b) NYUv2 - Depth.

# 4.6 - Training and Test Speed

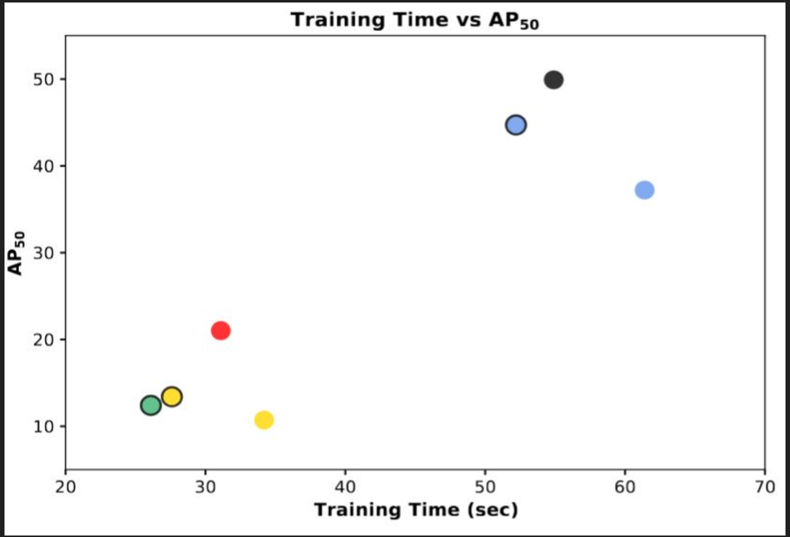



**Figure 13.** Training Time vs. Detection Performance.

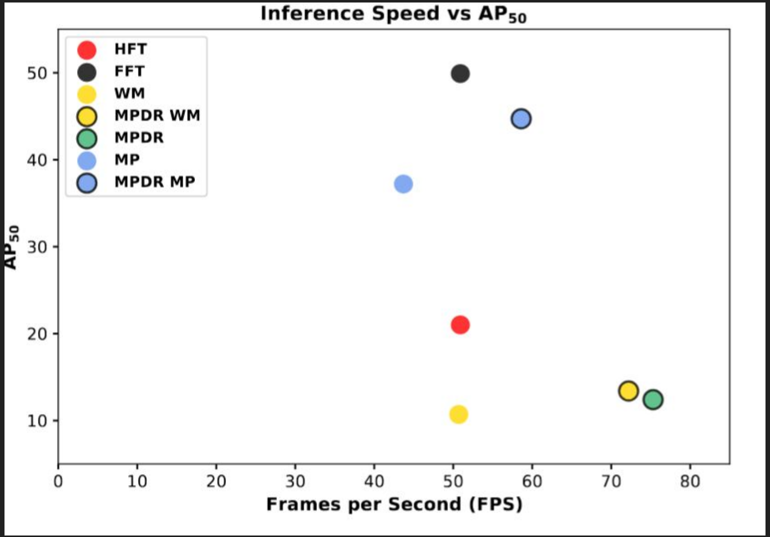



**Figure 14.** Inference Speed vs. Detection Performance.

# 4.7 - Qualitative Results

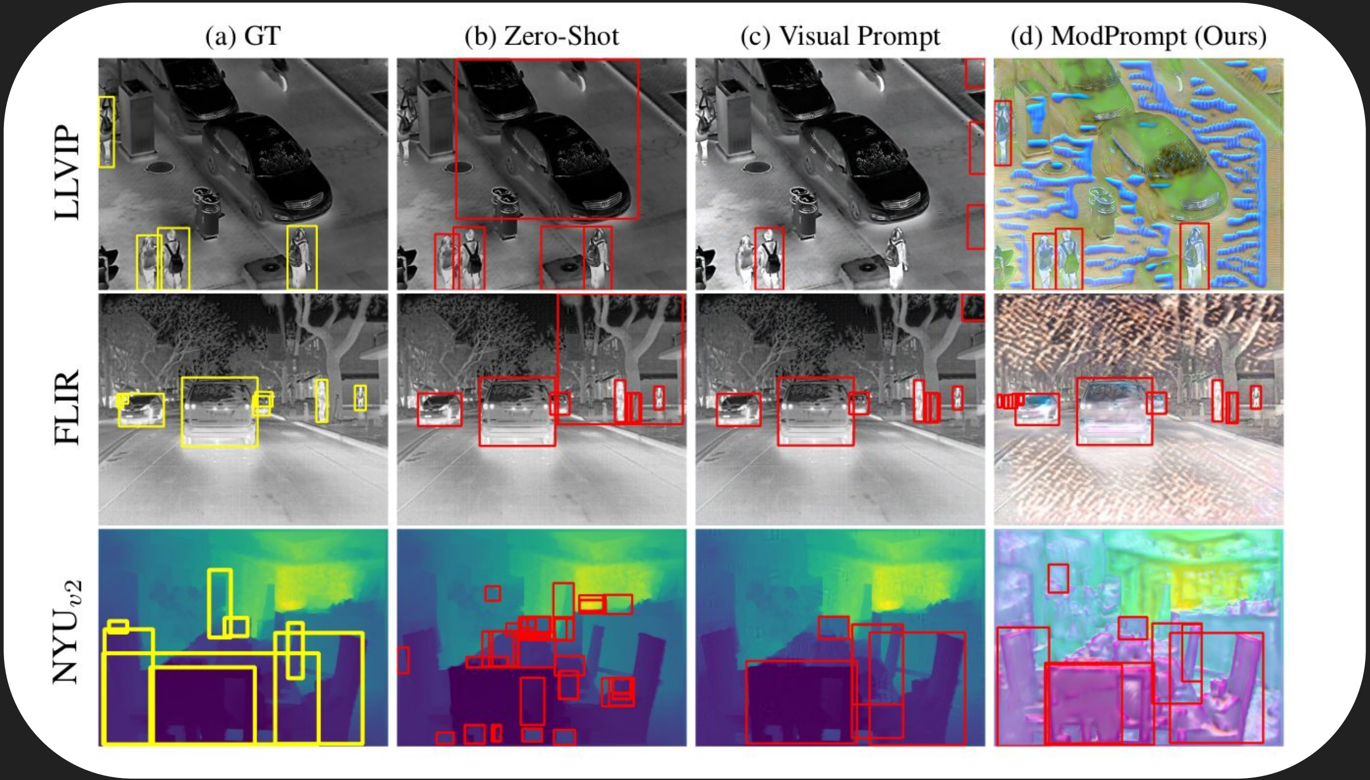



**Figure 15.** Detection over different methods.

# 5. Conclusion

# 5 - Conclusion

- We propose ModPrompt, a novel method that adapts VL-ODs across modalities with conditional visual prompts.
- ModPrompt preserves ZS, is efficient (<5% params), and is competitive with FT. Our residual tune text embeddings are toggleable at inference.
- Our technique outperformed competitors across different visual modalities, such as IR, Depth, Event-based, and LIDAR.

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Paper

Code

ILLS International Laboratory on Learning Systems

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Université du Québec ÉTS Engineering for Industry

LIVIA LABORATOIRE D'IMAGERIE, DE VISION ET D'INTELLIGENCE ARTIFICIELLE

DISTECH CONTROLS™

# Visual Modality Prompt for Adapting Vision-Language Object Detectors

#10337

**Heitor R. Medeiros, Atif Belal, Srikanth Muralidharan, Eric Granger, Marco Pedersoli**
LIVIA, ILLS, Dept. of Systems Engineering, ETS Montreal, Canada