Teaching VLMs to Localize Specific Objects from In-context Examples
Supplementary Material

Sivan Doveh*[2] Nimrod Shabtay*[1,3] Eli Schwartz[1] Hilde Kuehne[1,4] Raja Giryes[3] Rogerio Feris[5]
Leonid Karlinsky[5] James Glass[6] Assaf Arbelle[1] Shimon Ullman[2] M. Jehanzeb Mirza[6]
[1]IBM Research [2]Weizmann Institute of Science [3]Tel Aviv University,
[4]Tuebingen AI Center [5]MIT-IBM [6]MIT CSAIL

In the following, we provide additional experiments and further explanations that offer deeper insights and enhance the clarity of the main manuscript. In Section 1, we present the detailed results for Qwen2-VL-7B retention of generalization, and the LLaVA-OV model (with and without) fine-tuning, complementing the ablation results from the main text. In Section 2, we detail the methodology for converting pixel-level segmentation maps from datasets like PDM and PerSeg into bounding box annotations, ensuring compatibility with our application. In Section 3, we provide extensive qualitative visualizations across various datasets and few-shot settings, highlighting both successful localizations and challenges in instance-level discrimination. Comprehensive examples for 1-shot to 8-shot scenarios are included to illustrate the robustness and adaptability of our method. Finally, in Section 4, we include a prompt ablation study that explores how different prompt variants influence the model's localization performance, thereby highlighting best practices for designing prompts in inference.

1. Detailed Results

In the following, we first provide extended results of experiments performed in the main manuscript highlighting the retention of generalization abilities of our fine-tuned models, then provide results showing the generalization of fine-tuning using our method across network architectures.

1.1. Retention of generalization

We extend Table 4 of the main paper for the 7B model. In Table 1 we show the minimal performance degradation of IPLoc-7B compared to the base model on several tasks, namely GQA, Seed-Bench and POPE.

1.2. Retention of generalization across architectures

The detailed results obtained from base LLaVA-OV and the fine-tuned model (on our dataset) are presented in Table 2. We find that our IPLoc consistently improves the base model on all the few-shot splits we test on.

Dataset	Qwen2-VL-7B	IPLoc
GQA	62.34	61.11
SEED	72.06	71.16
POPE	88.35	88.04

Table 1. **Retention of Generalization.** Ablation highlighting the retention of generalization abilities of our fine-tuned model for the task of interest.

Dataset	Shots	LLaVA-OV	IPLoc-LLaVA-OV
PDM	1	11.10	12.29
	2	13.85	15.03
PerSeg	1	43.01	52.96
	2	40.08	57.90
	3	30.01	56.51
	4	14.03	18.11
ICL-LASOT	1	12.45	13.99
	2	15.66	16.88
	4	18.64	21.04
	8	7.62	7.80
Average		20.65	27.25

Table 2. **Generalization across different VLMs.** We report the detailed results of base and IPLoc-finetuned LLaVa-OV model.

These results provide insights regarding the generalization of our fine-tuning methodology across different vision language models (VLMs). Note that in the main manuscript (Table 2), we fine-tuned Qwen2-VL [2].

2. Segmentation Masks to Bounding Boxes

The PDM [1] and PerSeg [3] datasets provide images annotated with pixel-level segmentation maps, where each item is uniquely labeled. Since our method requires bounding box annotations, we directly utilize the segmentation

Model	Prompt	PDM		PerSeg				ICL-LASOT				Average
		1-shot	2-shot	1-shot	2-shot	3-shot	4-shot	1-shot	2-shot	4-shot	8-shot	
Qwen2-VL-72B	Org. Prompt (1)	33.65	32.37	0.81	0.90	0.29	1.26	4.82	0.40	1.63	0.76	7.69
Qwen2-VL-72B	Prompt 1 (2)	16.88	12.97	41.88	35.47	49.51	39.96	42.66	38.05	33.54	31.17	**34.21**
Qwen2-VL-72B	Prompt 2 (3)	19.96	13.11	44.75	28.19	43.43	37.36	43.14	40.04	31.29	22.81	32.41
Qwen2-VL-72B	Prompt 3 (4)	21.23	12.29	34.91	28.68	45.50	31.04	43.01	37.65	31.38	27.23	31.29
InternVL2-76B	Org. Prompt (1)	23.28	14.20	34.51	24.18	44.94	36.57	40.67	23.21	31.72	21.65	29.49
InternVL2-76B	Prompt 1 (2)	27.48	21.87	49.01	41.16	43.22	34.63	42.47	45.61	39.71	21.74	36.69
InternVL2-76B	Prompt 2 (3)	31.26	28.67	50.01	51.46	49.37	45.70	47.68	42.90	38.90	27.21	**41.32**
InternVL2-76B	Prompt 3 (4)	29.11	20.85	46.06	42.30	50.03	42.34	46.07	43.76	38.86	29.18	38.86

Table 3. **Prompt ablation for Qwen2-VL-72B and InternVL2-76B.** We employ multiple prompt strategies to gain the best performance (% IoU) out of the big LMMs (Qwen2-VL-72B, InternVL2-76B). Suprisingly, the default prompt which the models were originally trained with perform the worse, while more explicit instructions work better.

maps. Each segmentation map, containing unique object labels, is processed to extract bounding box coordinates (x_1, y_1, x_2, y_2) that enclose the objects. These bounding box annotations are derived from the segmentation maps without the need for additional pixel-level computation, as the mapping between labels and objects is predefined.

3. Comprehensive Visual Analysis Across Datasets

We present an extensive set of qualitative results demonstrating our method's performance across multiple benchmark datasets (PDM, PerSeg, ICL-LASOT) under various few-shot settings. Our visualization framework employs a consistent color-coding scheme where red bounding boxes denote ground truth annotations in the support frames, while blue bounding boxes indicate our model's predictions on query frames. For each example, we show the support shots (containing the target object with its localization) followed by the corresponding query image in the final column.

As shown in Figure 7, our method successfully localizes objects with just a single support frame across diverse scenarios. Further, in Figure 8, Figure 9 and Figure 10 we can see our method also performs well with more shots. We also highlight some challenges of instance-level discrimination in Figure 11, where semantically similar objects lead to incorrect localizations.

4. Prompt Ablation

4.1. Qwen2-VL-72B and InternVL2-76B Prompt Variants

Our investigation reveals a surprising relationship between instruction format and bounding box prediction accuracy in large multimodal models (Qwen2-VL-72B and InternVL2-76B). Contrary to expectations, the default instructions employed during model training consistently yielded poorer performance. The model ignores the instruction and outputs a caption (see the attached examples in Tables 5, 6, 7, 8, 9, 10, 11, and 12). In contrast, more explicit and verbally detailed instructions produced significantly better results, and the outputs are indeed bounding boxes. In Table 3 we present a comprehensive analysis of various instruction formats and their corresponding performance outcomes across both models. The best prompt over all datasets is reported in Table 2 of the main paper for each model.

We evaluated the models using four distinct prompting strategies, labeled Original prompt and prompt 1 through prompt 3:

Original prompt:
“`<ref>category</ref>`”

Figure 1. Qwen2-VL Original Prompt

Prompt 1:
“`Please provide the bounding box of the element {element}, return the bounding box in the following format:` $[x_0, y_0, x_1, y_1]$”

Figure 2. Qwen2-VL / InternVL2 Prompt 1

4.1.1. Qwen2-VL-72B Examples:

In Tables 5, 6, 7 and 8, we provide additional and detailed examples of the variants of the prompts we used throughout our experiments.

Prompt 2:
“Task: Locate the element in the image. Provide its bounding box coordinates in the format $[x_{min}, y_{min}, x_{max}, y_{max}]$”

Figure 3. Qwen2-VL / InternVL2 Prompt 2

Prompt 3:
“Please analyze this image and locate the exact element. Return the precise bounding box coordinates using this format: $[x_{min}, y_{min}, x_{max}, y_{max}]$ The coordinates should tightly bound only the element, nothing more. Take your time to carefully examine the image and provide the most accurate bounding box possible.”

Figure 4. Qwen2-VL / InternVL2 Prompt 3

“Please provide the bounding box of the element element”

Figure 5. GPT-4o Prompt 1

“Please provide the bounding box of the element element, return the bounding box coordinates the following format: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Do not output anything else besides the coordinate”

Figure 6. GPT-4o Prompt 2

4.1.2. InternVL2-76B Examples:

In Tables 9, 10, 11 and 12 we provide additional and detailed examples of the variants of the prompts we used throughout our experiments.

4.2. GPT-4o Prompt Variants

Following the successful experimental approaches implemented with Intern-VL2-76B and Qwen2-VL-72B, we employ the same instruction analysis for GPT-4o (gpt-4o-2024-08-06). We employ several prompt strategies to gain the best performance- Table 4 shows the full results over different prompts. Concretely, we utilized the following prompt configurations 5 and 6.

In addition, we use the best resolution that the model can work with by setting the ”detail: high” argument. The best prompt over all datasets is reported in Table 2 of the main paper for each model.

4.2.1. GPT-4o Examples:

In Tables 13 and 14 we provide additional and detailed examples of the variants of the prompts we used throughout our experiments.

References

[1] Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where’s waldo: Diffusion features for personalized segmentation and retrieval. *NeurIPS*, 2024. 1

[2] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[3] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 1

Model	Prompt	PDM		PerSeg				ICL-LASOT				Average
		1-shot	2-shot	1-shot	2-shot	3-shot	4-shot	1-shot	2-shot	4-shot	8-shot	
GPT-4o	Prompt 1 (5)	4.83	7.70	26.06	25.53	26.13	25.19	11.27	15.18	18.60	19.05	18.23
GPT-4o	Prompt 2 (6)	11.27	13.29	33.43	41.41	44.77	45.72	18.13	20.61	23.08	24.61	**27.63**

Table 4. **Instruction format on GPT-4o.** We employ multiple prompt strategies to gain the best performance (% IoU) out of the model.

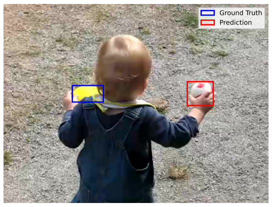
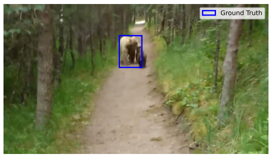
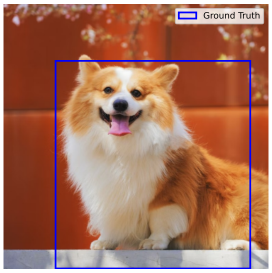
Prompt	Response	Visualization
'<ref>balloon</ref>', '[456.25, 497.91, 517.18, 577.08]', '<ref>balloon</ref>'	[444, 186, 506, 244]	
'<ref>bear</ref>', '[519.53, 36.1, 606.25, 273.61]', '<ref>bear</ref>', '[564.84, 320.83, 692.18, 577.77]', '<ref>bear</ref>'	The image shows a group of bears walking along a dirt path in a forested area. The path is surrounded by trees and greenery. The bears appear to be moving in a line, with one bear following another.	
'<ref>dog</ref>', '[219.23, 56.41, 998.71, 926.92]', '<ref>dog</ref>', '[120.35, 215.10, 998.71, 859.15]', '<ref>dog</ref>', '[18.92, 257.64, 998.54, 951.96]', '<ref>dog</ref>'	The image shows a fluffy dog with a white and orange coat, sitting and looking towards the camera. The dog has a happy expression with its tongue out. The background features a warm, orange hue with some blurred flowers, suggesting a pleasant, sunny environment.	

Table 5. **Qwen2-VL-72B with the original prompt examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

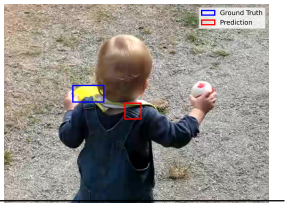
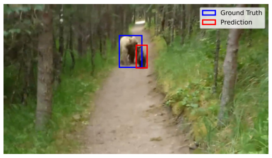
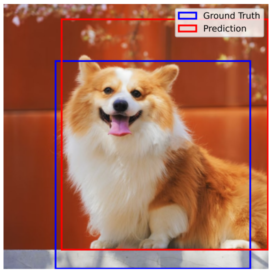
Prompt	Response	Visualization
’Please provide the bounding box of the element balloon, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[456.25, 497.916, 517.1875, 577.083]’, ’Please provide the bounding box of the element balloon, return the bounding box in the following format: [x0, y0, x1, y1]’	[292, 239, 331, 276]	
’Please provide the bounding box of the element bear, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[519.53, 36.11, 606.25, 273.61]’, ’Please provide the bounding box of the element bear, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[564.84, 320.833, 692.18, 577.77]’, ’Please provide the bounding box of the element bear, return the bounding box in the following format: [x0, y0, x1, y1]’	[641, 196, 694, 306]	
’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[219.23, 56.41, 998.71, 926.92]’, ’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[120.35, 215.10, 998.71, 859.15]’, ’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[18.92, 257.64, 998.54, 951.96]’, ’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’	[171, 44, 779, 723]	

Table 6. **Qwen2-VL-72B with prompt variant 1 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

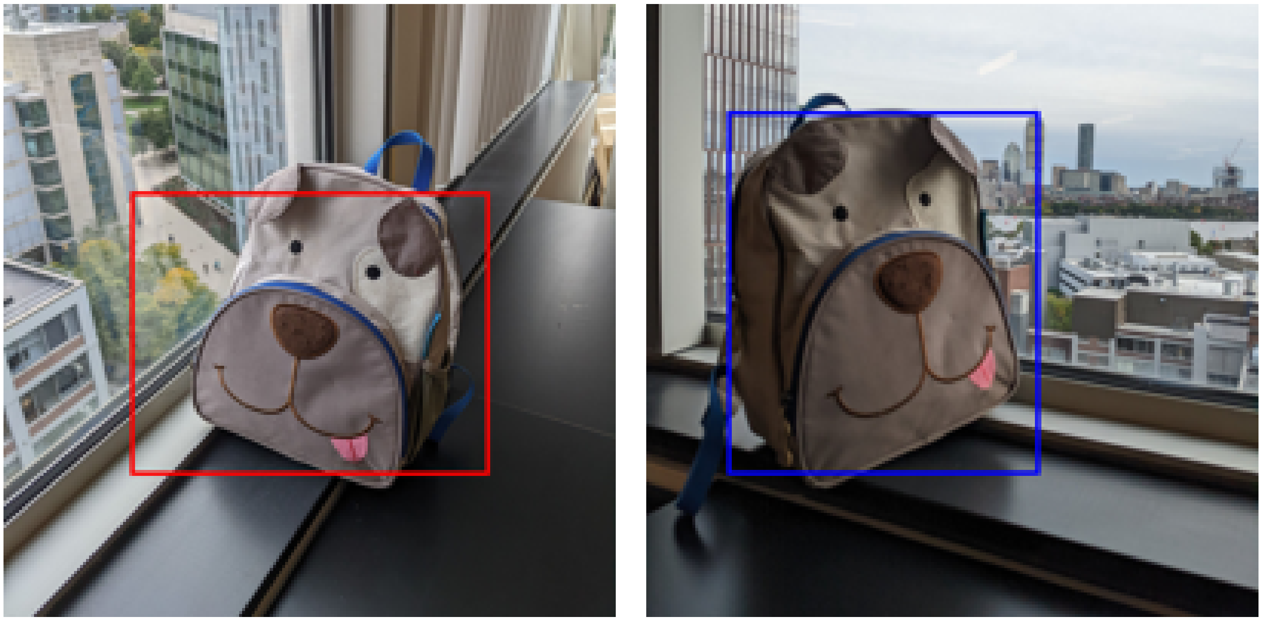


Figure 7. **One-shot Localization Results.** Representative examples of object localization using a single support frame in the PerSeg dataset.

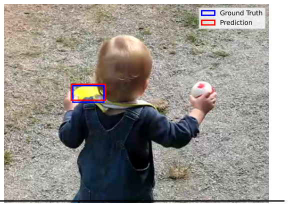
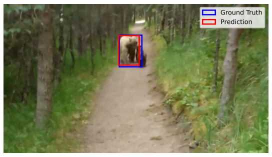
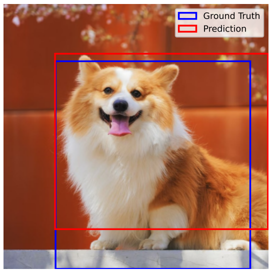
Prompt	Response	Visualization
’Task: Locate the balloon in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[456.25, 497.91, 517.18, 577.08]’, ’Task: Locate the balloon in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’	[199, 263, 222, 294]	
’Task: Locate the bear in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[519.53, 36.11, 606.25, 273.61]’, ’Task: Locate the bear in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[564.84, 320.8, 692.18, 577.77]’, ’Task: Locate the bear in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’	[199, 263, 222, 294]	
’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[219.23, 56.4, 998.71, 926.92]’, ’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[120.35, 215.10, 998.71, 859.15]’, ’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[18.92, 257.64, 998.54, 951.96]’, ’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’	[151, 144, 779, 662]	

Table 7. **Qwen2-VL-72B with prompt variant 2 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset.

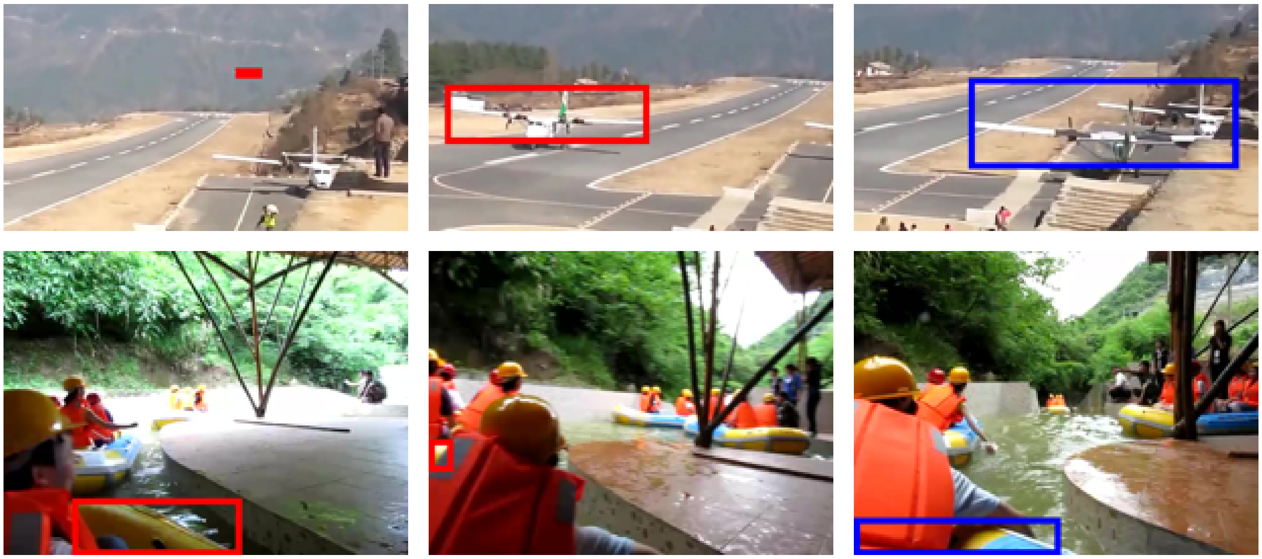


Figure 8. **Two-shot Localization Results.** Performance visualization with two support frames on the LaSOT and PDM datasets.

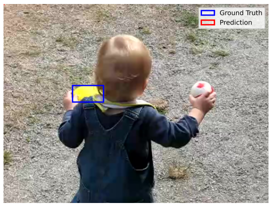
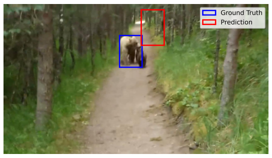
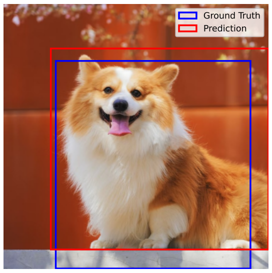
Prompt	Response	Visualization
’Please analyze this image and locate the exact balloon. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the balloon, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[456.25, 497.91, 517.18, 577.08]’, ’Please analyze this image and locate the exact balloon. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the balloon, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’	[0, 0, 0, 0]	
’Please analyze this image and locate the exact bear. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the bear, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[519.53, 36.11, 606.25, 273.61]’, ’Please analyze this image and locate the exact bear. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the bear, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[564.84, 320.83, 692.18, 577.77]’, ’Please analyze this image and locate the exact bear. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the bear, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’	[665, 25, 776, 197]	
’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[219.230, 56.41, 998.71, 926.92]’, ’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[120.35, 215.10, 998.71, 859.15]’, ’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[18.92, 257.64, 998.54, 951.96]’, ’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’	[138, 130, 780, 723]	

Table 8. **Qwen2-VL-72B with prompt variant 3 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset.

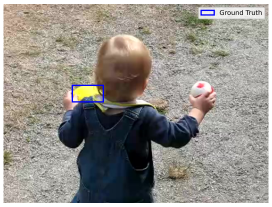
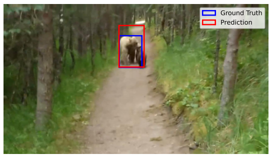
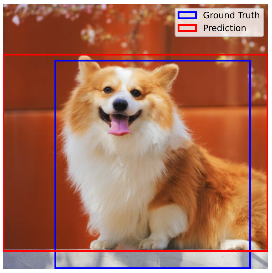
Prompt	**Response**	**Visualization**
“'<ref>balloon</ref>', '[456.25, 497.91, 517.18, 577.08]', '<ref>balloon</ref>'”	The second image shows a young child holding a yellow ball and a red and white ball, with one in each hand. The background is a gravel path with some green grass and a few scattered objects nearby. The child is wearing a blue shirt and blue overalls.	
“'<ref>bear</ref>', '[519.53, 36.1, 606.25, 273.61]', '<ref>bear</ref>', '[564.84, 320.83, 692.18, 577.77]', '<ref>bear</ref>'”	[558, 100, 679, 300]	
“'<ref>dog</ref>', '[219.23, 56.41, 998.71, 926.92]', '<ref>dog</ref>', '[120.35, 215.10, 998.71, 859.15]', '<ref>dog</ref>', '[18.92, 257.64, 998.54, 951.96]', '<ref>dog</ref>'”	[0, 149, 778, 728]	

Table 9. **InternVL2-76B with the original prompt examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

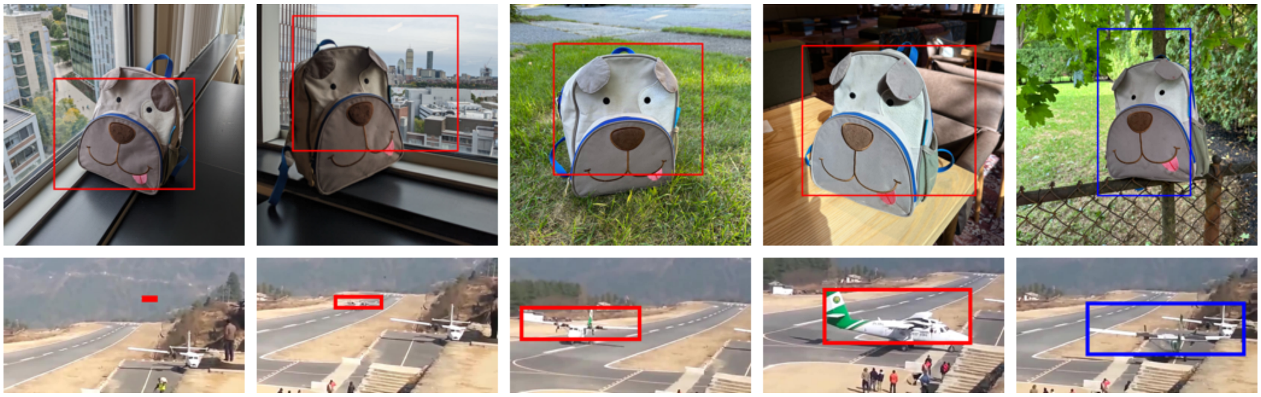


Figure 9. **Four-shot Localization Examples.** Visualization of our method's localization results when provided with four support frames on PerSeg and LaSOT datasets.

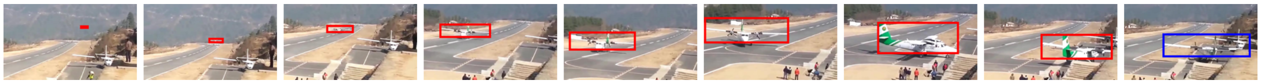


Figure 10. **Eight-shot Setting Results.** Demonstration of our method's localization behavior in the eight-shot setting on the LaSOT dataset.

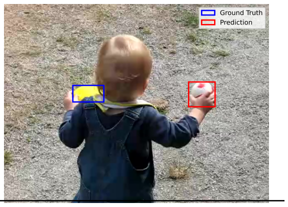
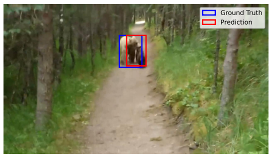
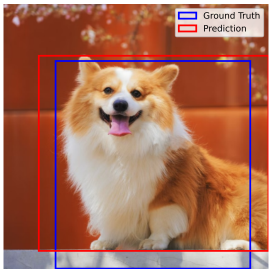
Prompt	**Response**	**Visualization**
“Please provide the bounding box of the element balloon, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[456.25, 497.916, 517.18, 577.08]’, ’Please provide the bounding box of the element balloon, return the bounding box in the following format: [x0, y0, x1, y1]”	[446, 187, 510, 247]	
“Please provide the bounding box of the element bear, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[519.53, 36.11, 606.25, 273.61]’, ’Please provide the bounding box of the element bear, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[564.84, 320.833, 692.18, 577.77]’, ’Please provide the bounding box of the element bear, return the bounding box in the following format: [x0, y0, x1, y1]”	[595, 149, 689, 296]	
‘’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[219.23, 56.41, 998.71, 926.92]’, ’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[120.35, 215.10, 998.71, 859.15]’, ’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]’, ’[18.92, 257.64, 998.54, 951.96]’, ’Please provide the bounding box of the element dog, return the bounding box in the following format: [x0, y0, x1, y1]”	[103, 151, 780, 726]	

Table 10. **InternVL2-76B with prompt variant 1 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

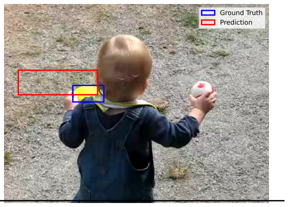
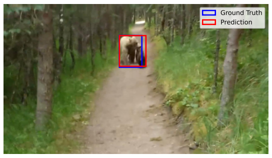
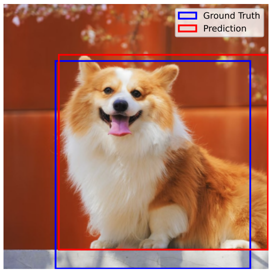
Prompt	Response	Visualization
“Task: Locate the balloon in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[456.25, 497.91, 517.18, 577.08]’, ’Task: Locate the balloon in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]”	[34, 159, 226, 218]	
“Task: Locate the bear in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[519.53, 36.11, 606.25, 273.61]’, ’Task: Locate the bear in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[564.84, 320.8, 692.18, 577.77]’, ’Task: Locate the bear in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]”	560, 150, 688, 298]	
“Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[219.23, 56.4, 998.71, 926.92]’, ’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[120.35, 215.10, 998.71, 859.15]’, ’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]’, ’[18.92, 257.64, 998.54, 951.96]’, ’Task: Locate the dog in the image. Provide its bounding box coordinates in the format [x_min, y_min, x_max, y_max]”	[162, 148, 778, 723]	

Table 11. **InternVL2-76B with prompt variant 2 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

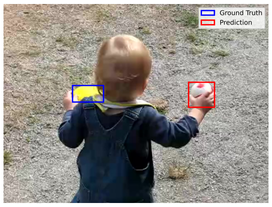
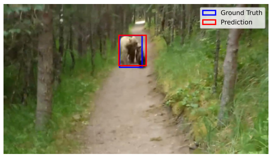
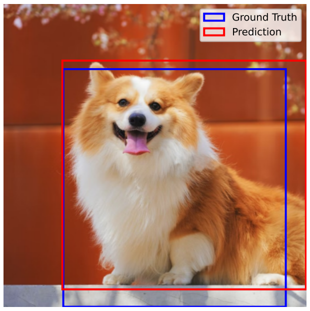
Prompt	Response	Visualization
“Please analyze this image and locate the exact balloon. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the balloon, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[456.25, 497.91, 517.18, 577.08]’, ’Please analyze this image and locate the exact balloon. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the balloon, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.”	446, 188, 509, 248]	
“Please analyze this image and locate the exact bear. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the bear, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[519.53, 36.11, 606.25, 273.61]’, ’Please analyze this image and locate the exact bear. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the bear, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[564.84, 320.83, 692.18, 577.77]’, ’Please analyze this image and locate the exact bear. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the bear, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.”	[556, 149, 688, 295]	
“Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[219.230, 56.41, 998.71, 926.92]’, ’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[120.35, 215.10, 998.71, 859.15]’, ’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.’, ’[18.92, 257.64, 998.54, 951.96]’, ’Please analyze this image and locate the exact dog. Return the precise bounding box coordinates using this format: [x_min, y_min, x_max, y_max] The coordinates should tightly bound only the dog, nothing more Take your time to carefully examine the image and provide the most accurate bounding box possible.”	[150, 144, 778, 734]	

Table 12. **InternVL2-76B with prompt variant 3 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset.

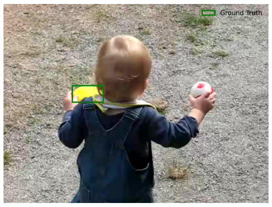
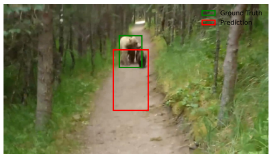
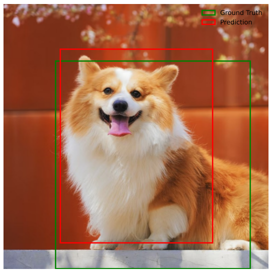
Prompt	Response	Visualization
“Please provide the bounding box of the element balloon, ’[456.25, 497.916, 517.18, 577.08]’, ’Please provide the bounding box of the element balloon”	The yellow balloon is not clearly visible. However, there is a yellow object in the child’s left hand. Would you like details on that instead?	
“Please provide the bounding box of the element bear, ’[519.53, 36.11, 606.25, 273.61]’, ’Please provide the bounding box of the element bear, ’[564.84, 320.833, 692.18, 577.77]’, ’Please provide the bounding box of the element bear”	[531, 218, 700, 510]	
“Please provide the bounding box of the element dog, ’[219.23, 56.41, 998.71, 926.92]’, ’Please provide the bounding box of the element dog, ’[120.35, 215.10, 998.71, 859.15]’, ’Please provide the bounding box of the element dog, ’[18.92, 257.64, 998.54, 951.96]’, ’Please provide the bounding box of the element dog”	[167, 132, 616, 703]	

Table 13. **GPT-4o with prompt variant 1 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

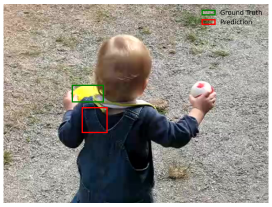
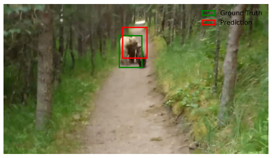
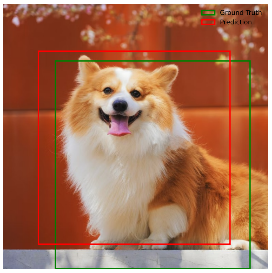
Prompt	**Response**	**Visualization**
“Please provide the bounding box of the element balloon, return the bounding box coordinates the following format: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Do not output anything else besides the coordinates”, ’[456.25, 497.916, 517.18, 577.08]’, ”Please provide the bounding box of the element balloon, return the bounding box coordinates the following format: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Do not output anything else besides the coordinates”	[190, 250, 250, 310]	
“Please provide the bounding box of the element bear, return the bounding box coordinates the following format: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Do not output anything else besides the coordinates”, ’[519.53, 36.11, 606.25, 273.61]’, ”Please provide the bounding box of the element bear, return the bounding box coordinates the following format: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Do not output anything else besides the coordinates”, ’[564.84, 320.833, 692.18, 577.77]’, ”Please provide the bounding box of the element bear, return the bounding box coordinates the following format: $[x_{min}, y_{min}, x_{max}, y_{max}]$. Do not output anything else besides the coordinates”	[576, 110, 694, 258]	
“Please provide the bounding box of the element dog, ’[219.23, 56.41, 998.71, 926.92]’, ’Please provide the bounding box of the element dog, ’[120.35, 215.10, 998.71, 859.15]’, ’Please provide the bounding box of the element dog, ’[18.92, 257.64, 998.54, 951.96]’, ’Please provide the bounding box of the element dog”	[167, 132, 616, 703]	

Table 14. **GPT-4o with prompt variant 2 examples.** We present examples from each dataset. Top row is 1-Shot from the PDM dataset, middle row is 2-shots from the ICL-LASOT test set and the bottom row is 3-shot from the PerSeg dataset

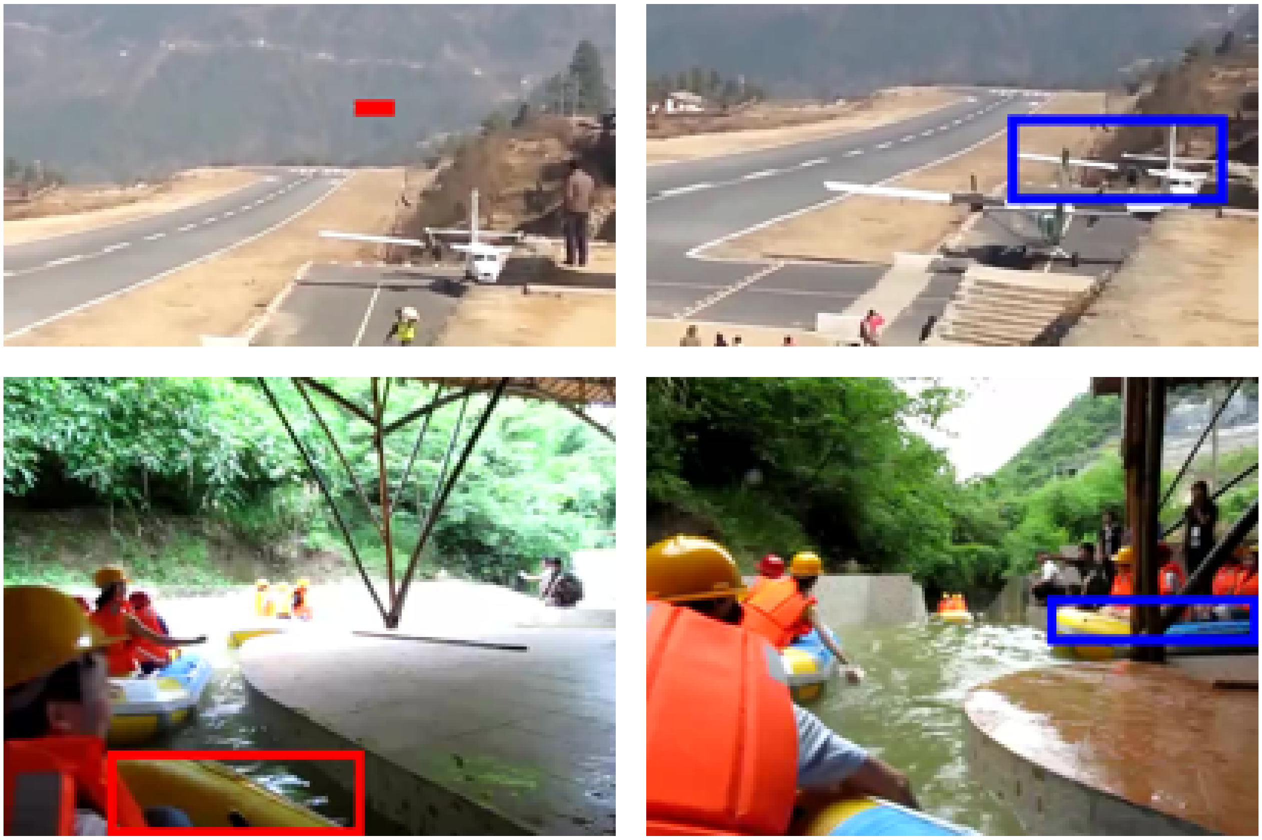


Figure 11. **Challenging Cases in One-shot Setting.** Examples where the model identifies semantically similar objects (incorrect airplane and boat) but fails to distinguish the specific target instance, highlighting the complexity of instance-level discrimination.