Visual Textualization for Image Prompted Object Detection

Supplementary Material

6. Implementation details of VisTex-DINO and other comparison methods

We also implemented VisTex-OVLM on GroundingDINO-T [33], denoted as VisTex-DINO. The MSTB design mirrors that of VisTex-GLIP, employing two "fully connected (fc) + ReLU" layers. Based on the feature size and scale extracted by the pre-trained GroundingDINO-T's vision encoder, H is set to 100. Since the intermediate features of GroundingDINO have W values that vary with image size, we applied bilinear interpolation to also set W to 100. GroundingDINO's intermediate features are available at three scales, making $M = 3$. MSTB is applied to both the image backbone and the feature enhancer of GroundingDINO-T, with max pooling across stages used for non-parametric multi-stage fusion. All other training settings are consistent with those of VisTex-GLIP.

For the comparison methods, we selected recent top-performing approaches, adopting published results where available and reproducing results with recommended settings on PASCAL VOC and MSCOCO when necessary. Notably, to ensure a fair comparison, the weights used for OWL-ViT and OWL-ViT v2 are "CLIP ViT-L/14" and "CLIP B/16 ST+FT," respectively. These weights were chosen because they were not specifically trained to exclude categories that might appear in MSCOCO or LVIS, making their pre-training settings closer to those of the OVLM weights we used. We used class names as the text prompt for all methods that accept text prompts.

7. Performance on ODinW13 subsets

Following Sec. 4.3 of the main text, we provide detailed transfer results on the ODinW13 subsets [31] in Tab. 10. ODinW13 [29] is composed of 13 subsets from ODinW35, spanning specialized natural domains such as aquarium species, surgical instruments, and aerial imagery, among others. Although the categories in these 13 datasets may appear in the pre-training dataset of OVLM, their performance results still, to some extent, reflect the model's capability in real-world scenarios. The specifics of ODinW13 are outlined in Tab. 9. All methods, including ours, were trained on the MSCOCO base set, treating downstream task sets as novel sets and evaluating them with 2-shot support images. The results are presented in the table below. These results further validate VisTex-OVLM's superior transferability across diverse domains, demonstrating its robustness and adaptability in handling significant domain shifts.

8. Compatibility experiments on RegionCLIP and FIBER

We evaluated VisTex on RegionCLIP [60] under a one-shot setting using the Open-Vocabulary COCO and LVIS benchmark, where base and novel categories are disjoint (Tab. 6). The zero-shot (ZS) results were adopted from the original paper. In full fine-tuning (FF*), the model was fine-tuned with support images from both base and novel classes. In contrast, VisTex was trained only on base categories. Results show that VisTex still improves RegionCLIP's performance on novel categories. Furthermore, we tested VisTex on another object-level VLM, FIBER [5], under the same setting as Table 1 in the main text, and present the results in Tab. 7 to further demonstrate its generalization ability.

Table 6. Performance on RegionCLIP [60].

Method	COCO			LVIS	
	Novel AP50	Base AP50	All AP50	AP	APr
regionCLIP-ZS	39.3	61.6	55.7	32.3	22
regionCLIP-FF*	61.1	62.7	61.9	45.2	36.1
VisTex-regionCLIP	**65.3**	**68.2**	**67.4**	**47.8**	**40.3**

Table 7. Performance on FIBER [5] (mAP if not specified).

Method	LVIS MiniVal		Unseen medical datasets				
	AP	APr	MoNu	CCRCC	ConSeP	LIDC	Deeplesion
FIBER-ZS	35.8	29.5	0.3	0.6	1.3	0.1	0.3
FIBER-FF*	48.4	38.8	9.2	9.8	25.5	29.5	34.7
VisTex-FIBER	**49.6**	**41.6**	**10.5**	**11.4**	**26.8**	**32.9**	**36.4**

9. Computational Overhead and Preprocess Time

In Tab. 8, we report the computational overhead for processing one image using GLIP-L on RTX3090 with one support image, comparing it to MQ-Det and GLIP-FF. After an initial preprocessing step on the support image, textualized visual tokens are stored for reuse. Thus, the actual inference cost, aside from this initial preprocessing, remains identical to that of the original OVLM. For fair and direct comparison, the FLOPs and time corresponding to this one-time preprocessing step are highlighted in blue in the table.

10. More ablation and visualization results

10.1. Multi-scale textualizing block

We assessed the impact of multi-scale textualization and the parameter-sharing strategy (MSTB sharing), as shown in Tab. 11. With scales indexed by j (e.g., "0" represents the $\frac{H}{2^0} \cdot \frac{W}{2^0}$ scale), results indicate that using multi-scale features from the vision encoder better preserves OVLM's object-text alignment and boosts performance over single-layer features.

Table 8. Computational Overhead (Preprocessing costs: blue).

Method	FLOPs		#Param(Trainable)	Inference time
	Training	Inference		
MQ-Det	717.46G	243.25G	10.87M	0.553s
GLIP-FF*	653.15G	218.78G	397.59M	0.547s
VisTex-GLIP	702.11G	17.75G+218.78G	8.38M	0.031s+0.547s

Dataset	Objects of interest	Train	Val	Test
PascalVOC	Common objects (PascalVOC 2012)	13690	3422	\
AerialDrone	Boats, cars, etc. from drone images	52	15	7
Aquarium	Penguins, starfish, etc. in an aquarium	448	127	63
Rabbits	Cottontail rabbits	1980	19	10
EgoHands	Hands in ego-centric images	3840	480	480
Mushrooms	Two kinds of mushrooms	41	5	5
Packages	Delivery packages	19	4	3
Raccoon	Raccoon	150	29	17
Shellfish	Shrimp, lobster, and crab	406	116	58
Vehicles	Car, bus, motorcycle, truck, and ambulance	878	250	126
Pistols	Pistol	2377	297	297
Pothole	Potholes on the road	465	133	67
Thermal	Dogs and people in thermal images	142	41	20

Table 9. The objects of interest for each subset and the image number of each split in ODinW13.

MSTB sharing further reduces the required convolutional weights, enhancing textualization effectiveness. MSTB sharing creates synergy during training, easing the learning process for mapping features of various scales into the same text feature space and slightly improving performance,. MSTB sharing saves 5.15M parameters and improves nAP by 1.2% compared to non-sharing across scales.

10.2. Multi-stage fusion

Multi-stage fusion (MSF) merges features from multiple encoder stages into a single textualized visual token. The original GLIP has 8 stages. To maintain GLIP's object-text alignment, we map each stage's visual features into the BERT-derived text feature space using MSTB. As shown in Tab. 12, using stages $1 \rightarrow 8$ yields the best performance, while stages 1 and $1 \rightarrow 4$ perform slightly lower. However, stages $5 \rightarrow 8$ or just stage 8 result in significant drops, likely due to that as the neural network progresses, the features become more high-level and abstract, making the mapping learning more challenging. Continuously incorporating information from different stages starting from the lower layers helps reduce the difficulty of learning the mapping.

10.3. Ablation on shot fusion mode

When integrating multiple shot features, there are various fusion modes available. Tab. 13 presents the performance of different fusion approaches. Experimental results indicate that concatenation yields the best results, as it maximally preserves the information from all shots, thereby preventing information loss. Support samples provide detailed semantic guidance for query prediction, and concatenation allows the model to maintain the unique information of each shot while utilizing data from multiple support samples. This method helps GLIP better grasp the support sample distribution for query prediction, enhancing performance.

10.4. Ablation on MSF

MSF's innovation is its efficient use of OVLM's multi-stage object-text alignment without extra parameters. Tab. 12 in the main text shows multi-stage fusion's effectiveness. We conducted an MSF ablation study in Sec. 10.4 using different common fusion methods. Max pooling outperforms other non-parametric fusion methods. This is because max pooling can highlight the most informative features across stages while reducing the negative impact of redundant noise.

10.5. Ablation on image prompt engineering

This ablation study (Tab. 15) followed the same experimental setup outlined in the main text: (1) Conducted on VisTex-OVLM in a 2-shot setting on MSCOCO using GLIP-L; (2) mAP was reported for both base and novel classes, with all other settings kept optimal. Several methods for preprocessing and inputting image prompts were tested, following CLIPSeg [36] settings unless specified. Experimental results show that the "BG blur" technique performs best. It highlights the target object while preserving some background, unlike "crop" and "baseline." Additionally, it avoids overlaying original image pixels, preventing information loss.

10.6. More output visualization

We provide 10-shot output examples for VisTex-GLIP on PASCAL VOC in Fig. 5. The settings are corresponded to Tab. 2 in the main text.

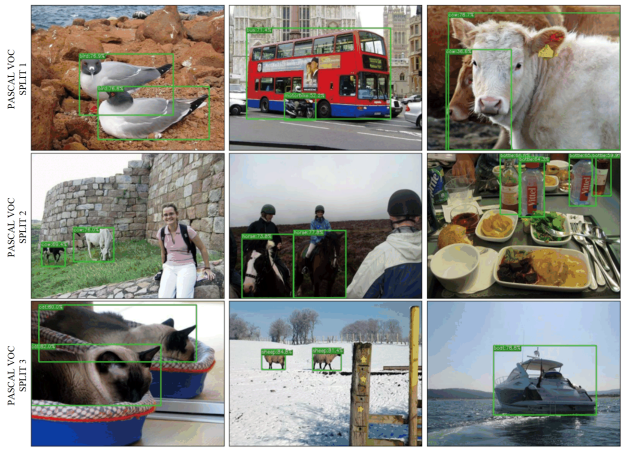

Figure 5. Visualization of VisTex-GLIP's 10-shot object detection results on PASCAL VOC. For simplicity, only detections of novel-class objects are illustrated. The settings are corresponded to Tab. 2 in the main text.

Method	AerialDrone	Vehicles	Aquarium	Mushrooms	Raccoon	Packages	Pothole	Shellfish	Rabbits	Pistols	Egohands	Pascalvoc	Thermal	Mean
Meta-DETR	16.5	32.5	18.2	36.1	34.2	31.4	22.8	28.2	36.6	34.1	35.2	30.1	37.1	30.2
DiGeo	22.6	37.3	25.1	49.2	41.4	44.2	26.7	39.8	47.0	38.7	41.6	42.2	39.7	38.1
DeFRCN	23.5	36.4	24.8	43.8	44.1	44.9	25.6	40.1	45.3	39.0	43.5	50.9	35.4	38.3
MFD	24.5	38.7	29.4	46.2	45.1	50.1	28.4	38.4	48.4	43.5	46.3	50.5	42.2	40.9
MQ-Det	21.2	36.5	30.0	45.3	41.1	41.6	28.3	37.5	46.9	44.1	39.7	48.4	40.3	38.5
VisTex-GLIP	**27.4**	**41.2**	**33.3**	**50.7**	**46.7**	**55.2**	**30.1**	**46.0**	**49.6**	**51.1**	**55.9**	**51.2**	**47.8**	**45.1**

Table 10. Detailed 2-shot transfer results on ODinW13 subsets. The best values are highlighted in **bold**.

MSTB	Scale (j)	Δ #Par(M)	2-shot		
			AP	bAP	nAP
w/o sharing	**0**	-8.34	47.1	47.0	47.1
	0+1	-7.36	48.8	**48.7**	49.2
	0+1+2	-3.68	48.6	47.9	48.8
	0+1+2+3	1.23	48.9	47.0	49.5
	0+1+2+3+4	5.15	49.2	48.5	50.6
sharing	**0+1+2+3+4**	0.00	**50.3**	48.6	**51.8**

Table 11. Ablation on multi-scale textualization and MSTB sharing. Δ#Par(M) represents the parameter amount offset relative to the optimal configuration. Best results are marked in **bold**.

Stages	Δ#Par(M)	2-shot		
		AP	bAP	nAP
1	**-30.67**	48.5	47.1	50.7
1→4	**-16.93**	48.8	47.7	49.3
1→8	**0.00**	**50.3**	**48.6**	**51.8**
5→8	**-16.93**	23.2	30.7	13.6
8	**-30.67**	20.5	28.9	13.9

Table 12. Ablation on mapping and fusing different numbers and sequences of stages. "i→j" indicates the stages used. Δ#Par(M) represents the parameter amount offset relative to the optimal configuration. Best results are marked in **bold**.

Shot Fusion Mode	2-shot		
	AP	bAP	nAP
element-wise addition	41.6	41	41.9
max	48.5	48.4	48.6
average	47.9	45.2	48.1
concat	**50.3**	**48.6**	**51.8**

Table 13. Ablation on shot fusion mode. Best results are marked in **bold**.

Stage Fusion Mode	2-shot		
	AP	bAP	nAP
element-wise Addition	36.7	37.4	31.6
max	**50.3**	**48.6**	**51.8**
average	46.8	48.3	41.4
concat	39.4	39.3	39.4

Table 14. Ablation on MSF. Best results are marked in **bold**.

Engineering Method	2-shot
	AP
baseline	16.3
BG blur	**49.6**
dye object red in grays image	19.5
add red object outline	28.3
crop	44.2
crop large context	30.8

Table 15. Ablation on image prompt engineering methods. "Baseline" indicates directly inputting the original image. "crop" means cropping out the target region based on ground truth bounding box while "crop large context" enlarges ground truth bounding box by $k = 10$ pixels. "BG blur" technique applies a shadow (intensity of 0.1) and Gaussian noise (kernel size of 15 and a standard deviation of 3) to the background area. Best results are marked in **bold**.

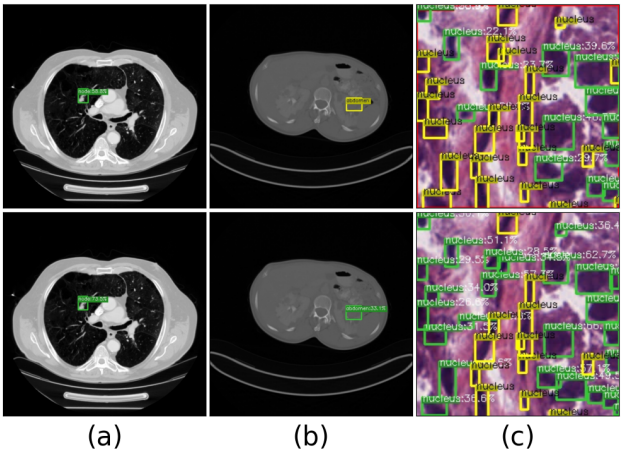

Figure 6. Output visualizations for medical datasets. The first row illustrates the inference results when models trained solely on the MSCOCO base set were directly applied to the medical datasets. The second row demonstrates results after briefly fine-tuning the MSTB in a 2-shot setting on the medical tasks. (a) LIDC, (b) Deepleision, (c) MoNu. Green, red, and yellow boxes denote true positives, false positives, and false negatives.

11. Output visualization for real-world downstream tasks

Fig. 8 and Fig. 6 present the output visualizations for real-world downstream tasks, including ODinW13 subsets and medical datasets (MoNu, LIDC, and Deeplesion).

The visualizations in Fig. 8 adhere to the settings described in Sec. 4.3 of the main text, where all methods, including ours, were trained on the MSCOCO base set. Downstream task sets were treated as novel sets, and evaluations were conducted using 2-shot support images, directly showcasing inference results on novel classes.

Fig. 6 focuses on datasets with larger domain gaps, specifically medical datasets. The first row illustrates the inference

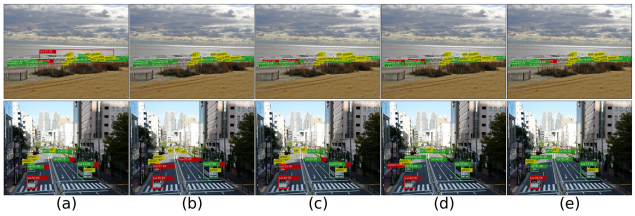

Figure 7. Failure cases. (a) GLIP-ZS, (b) GLIP-FF, (d) GLIP-MaPLe, (e) MQ-Det, (f) VisTex-GLIP. Green, red, and yellow: true positives, false positives, and false negatives.

results when models trained solely on the MSCOCO base set were directly applied to the medical datasets. The second row demonstrates results after briefly fine-tuning the MSTB in a 2-shot setting on the medical tasks.

We also provide some failure cases of our method on natural images in Fig. 7. As shown, these failures primarily occur in scenarios with dense or small objects, which is similar to the challenges observed in medical images in Fig. 6. We attribute this limitation to the weak representation of small objects in the pre-trained OVLM and the inherent difficulty of fitting novel category distributions with limited support samples. Addressing these issues will be a focus of future research.

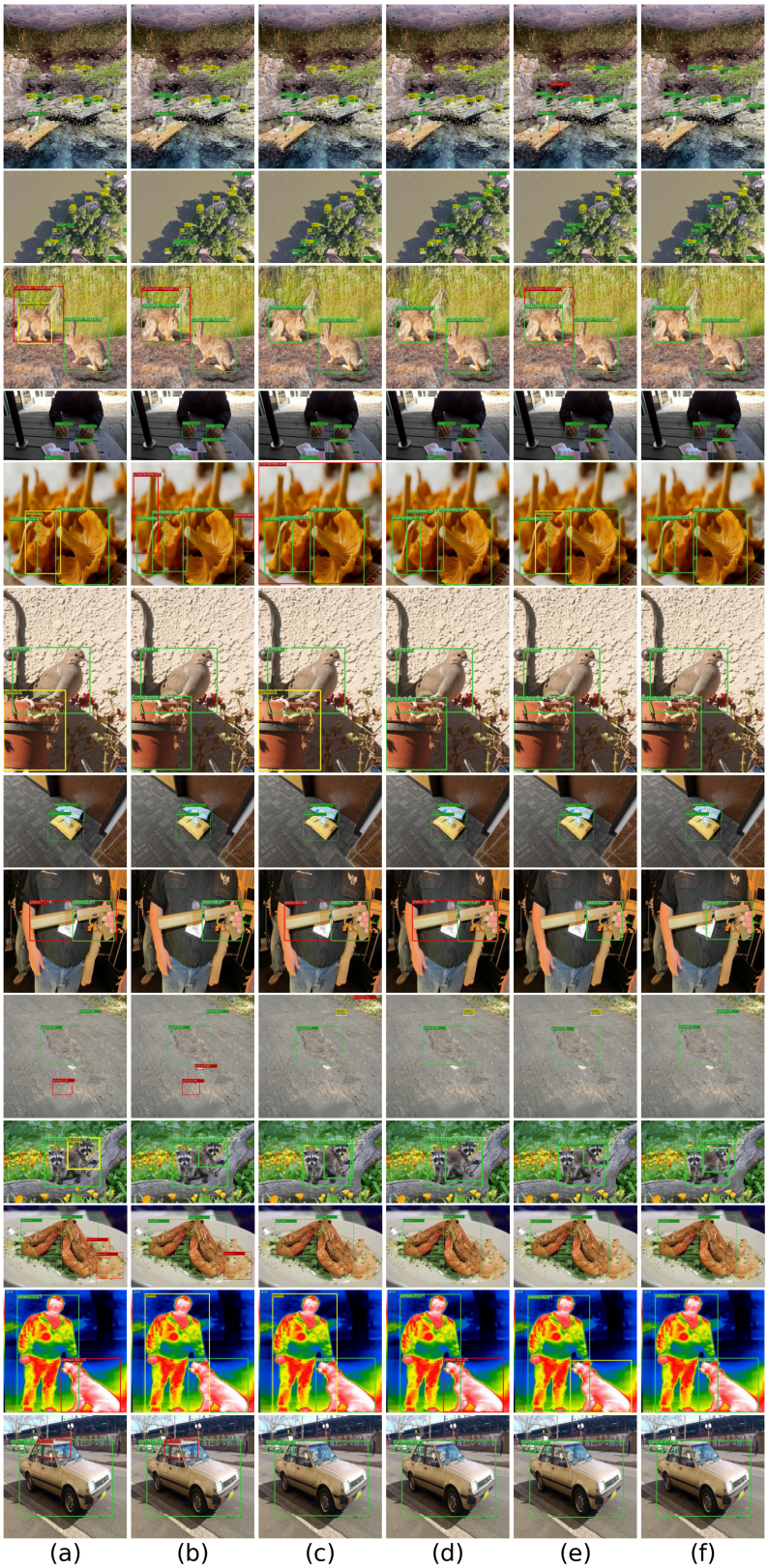

Figure 8. Comparison output visualizations on ODinW13. From top row to the bottom: Aquarium, AerialDrone, Rabbits, EgoHands, Mushrooms, PascalVOC, Packages, Pistols, Pothole, Raccoon, Shellfish, Thermal, Vehicles. (a) Meta-DETR, (b) DiGeo, (c) DeFRCN, (d) MFD, (e) MQ-Det, (f) VisTex-GLIP. Green, red, and yellow boxes denote true positives, false positives, and false negatives.