PET-DINO: Unifying Visual Cues into Grounding DINO with Prompt-Enriched Training

Supplementary Material

1. More Implementation Details

1.1. Detailed Illustration of DMD

Here, we further provide a detailed illustration of Dynamic Memory-Driven (DMD) Prompting, as shown in Figure 1. At iteration t, n categories are sampled from the Visual Cues Bank. For each sampled category, the corresponding stored visual prompts are aggregated to generate a memory-driven prompt. These memory-driven prompts, together with the current prompts and intra-batch prompts, are used in parallel to guide the model for detection, enabling parallel optimization, enhancing alignment with application scenarios, and strengthening open-category recognition. After participating in the guidance process, the current prompts and intra-batch prompts are stored by category to update the Visual Cues Bank, thereby enriching the visual cues for iteration t+1 and significantly reducing computational overhead. To ensure that stored visual prompts remain compatible with the evolving network, the Visual Cues Bank keeps up to M prompt embeddings per category during training, discarding the oldest when new ones are added.

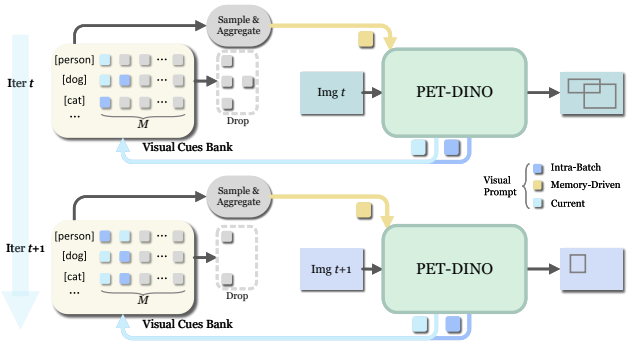


Figure 1. Dynamic Memory-Driven Prompting Diagram. During each iteration, the Visual Cues Bank updates its stored prompts with visual prompts used by PET-DINO, while PET-DINO utilizes the enriched prompts from the bank to improve training.

1.2. More Training Details

Both the matching cost computation and the final loss calculation incorporate classification losses, box L1 losses, and GIoU losses. For classification, we employ a contrastive loss between predicted objects and prompt embeddings to align the prompt representations with the implicit information in the image. Specifically, we compute the dot product between each output query and the prompt embeddings to obtain logits for each category and then apply focal loss to each logit. For Hungarian matching, we assign weights of 2.0, 5.0, and 2.0 to the classification, L1, and GIoU costs, respectively. The corresponding loss weights are 1.0, 5.0, and 2.0 in the final loss calculation.

We use automatic mixed precision for training. For image augmentation, we adopt standard techniques used in DETR-like methods, including multi-scale training and random flipping. Following DINO, we employ contrastive denoising training (CDN) to stabilize training and accelerate convergence.

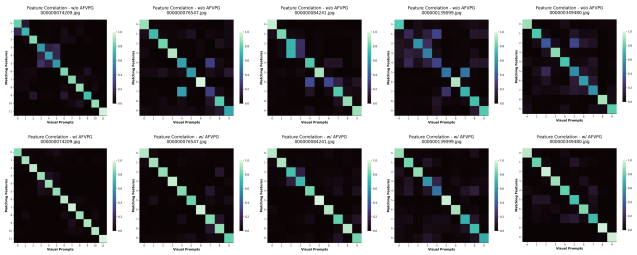


Figure 2. Feature correlation analysis between visual prompts and instance-level image features showing the impact of AFVPG.

2. Visualization Analysis of Proposed Modules

2.1. Feature correlation analysis of AFVPG

To visually demonstrate the effectiveness of AFVPG, we provide a feature correlation analysis as shown in Figure 2. We compute the cosine similarity, followed by Softmax normalization, between all candidate prompts and the image feature most similar to each visual prompt.

As illustrated, AFVPG (bottom) markedly enhances diagonal responses and suppresses off-diagonal interference, effectively mitigating the cross-category ambiguity observed in the model without AFVPG (top). This intrinsic alignment facilitates precise query selection, yielding substantial performance gains of 4.8 AP on visual-I via this simple yet effective design.

2.2. t-SNE visualization of IBP and DMD

To reveal the underlying reasons for the performance gains in Table 6, we conducted a t-SNE visualization on the visual prompt features of 10 randomly selected categories from the COCO dataset.

As shown in Figure 3, without IBP and DMD (left), the visual prompt features are scattered without clear boundaries between categories. In contrast, incorporating IBP

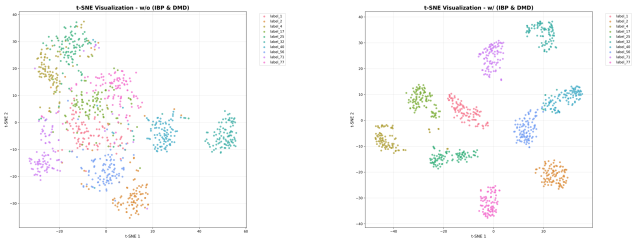


Figure 3. t-SNE visualization of visual prompt features showing the impact of IBP and DMD.

and DMD (right) leads to tighter intra-class aggregation and clearer inter-class separation. This highly discriminative semantic space ensures robust category differentiation in complex multi-category scenarios, serving as the primary driver of performance boost.

3. Visualizations

In this section, we comprehensively showcase the capabilities of PET-DINO across diverse scenarios by leveraging various types of prompts.

We evaluated zero-shot detection performance in dense object scenarios using PET-DINO in the interactive visual prompt detection mode. As shown in Figure 4, PET-DINO demonstrates excellent performance in single-category scenes. Furthermore, as illustrated in Figure 5, PET-DINO maintains strong performance in multi-category scenarios, accurately distinguishing between different categories. These results demonstrate that our prompt generation and training strategies enable the model to perform well in complex and dense detection scenarios, underscoring the strong potential of PET-DINO for real-world applications like automatic annotation and object counting.

As illustrated in Figure 6, PET-DINO can identify objects of the same category in different images based on a target selected from an exemplar image, paving the way for broader and more general application scenarios.

In Figure 7, we present zero-shot results based on pre-extracted generic visual prompts. The method achieves strong performance in various scenarios, demonstrating the effectiveness of our training strategies aligned with practical usage. It is especially valuable for adapting to novel scenarios without additional retraining, and can further be applied to cases where textual representations are difficult to align with the targets.

In Figure 8, we present zero-shot text prompt-based detection results. Our model maintains robust performance and demonstrates effectiveness across various scenarios, reflecting the harmony of our inheritance strategy.

In conclusion, empowered by our strategies, PET-DINO effectively adapts to diverse application scenarios, exhibiting high accuracy and robust generalization.

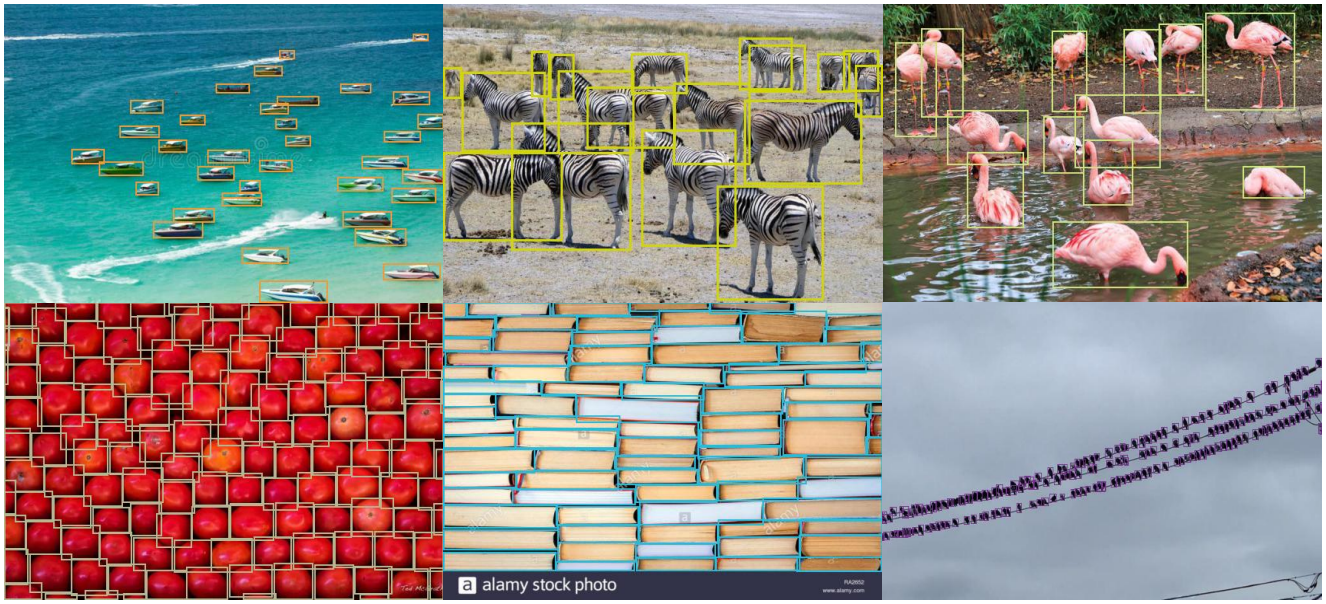


Figure 4. Zero-shot detection visualizations of **PET-DINO** on **interactive visual** prompt-based detection in **single**-category **dense** object scenarios.

Figure 5. Zero-shot detection visualizations of **PET-DINO** on **interactive visual** prompt-based detection in **multi**-category **dense** object scenarios.

Figure 6. Zero-shot detection visualizations of **PET-DINO** on **cross-image exemplar visual** prompt-based detection. Exemplars are shown above, and prediction outputs are shown below.

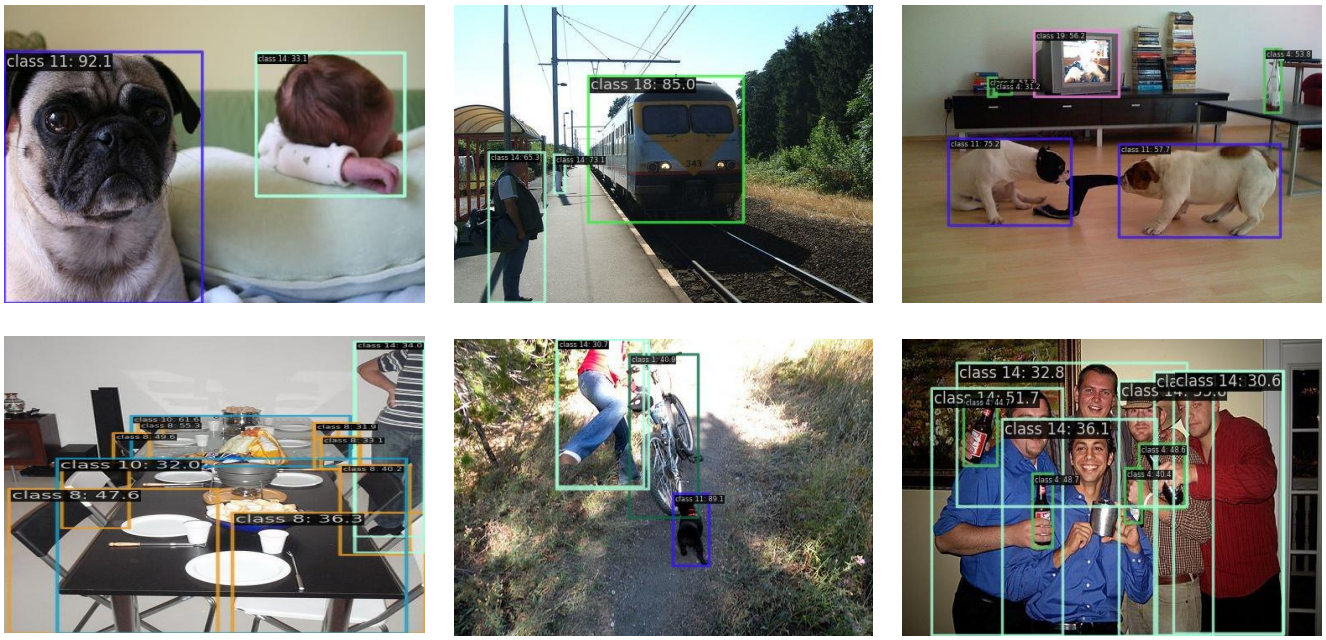
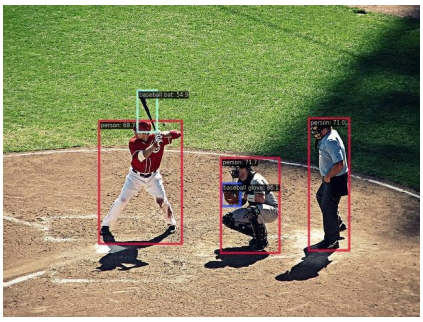
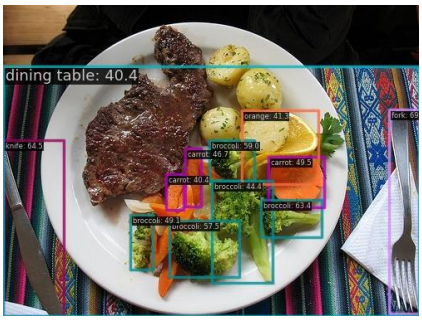


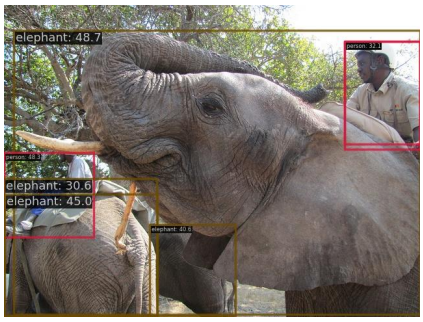
Figure 7. Zero-shot detection visualizations of **PET-DINO** with **class-level generic visual** prompts. The visual prompt embeddings are pre-extracted from the training set.



person, baseball glove, baseball bat, ···



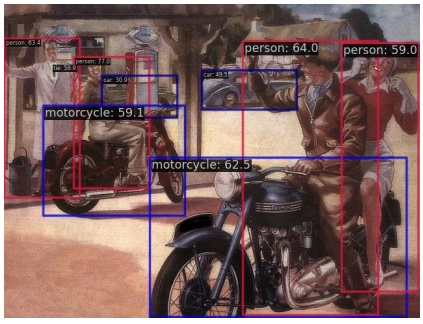
dinning table, fork, knife, carrot, broccoli, orange, ···



person, elephant, ···

person, tie, knife, cake, ···

person, motorcycle, car, truck, backpack, bottle, ···



person, motorcycle, car, ···

Figure 8. Zero-shot detection visualizations of **PET-DINO** with **text** prompts. The category names from the dataset are utilized as textual inputs for prompt generation.