# Beyond Semantic Alignment: Towards Dense Instance Perception in Vision-Language Models for Open-World Counting

**Anonymous authors**
Paper under double-blind review

## Abstract

Vision-Language Models (VLMs) achieve strong semantic understanding through large-scale image-text alignment, yet their ability to distinguish and enumerate nearby visual instances degrades substantially as object density increases. We investigate this semantic-to-instance perception gap through a unified language-guided counting benchmark built from four existing counting datasets and a controlled evaluation of representative visual encoders. Our analysis reveals that stronger semantic representations benefit counting mainly in sparse scenes, while their advantages diminish under increasing crowding. To alleviate this limitation, we propose **DIP-CountVLM**, a **D**ense **I**nstance **P**erception **Count**ing **VLM** for open-world scenarios. DIP-CountVLM combines **Instance-Aware Feature Fusion (IAFF)**, which adaptively aggregates hierarchical visual features to recover instance-sensitive cues, with **Dense Instance Token Reconstruction (DITR)**, which leverages image guidance to reconstruct spatially denser visual tokens. Extensive experiments show that DIP-CountVLM improves overall counting performance over a SigLIP2-based VLM baseline, with the most consistent gains in low- and medium-density regimes. Further analyses demonstrate that hierarchical feature fusion is most effective when coupled with dense token reconstruction, while extremely dense scenes remain challenging. These findings provide a systematic analysis of the visual representation bottlenecks underlying language-guided counting and a practical approach for alleviating them.

## 1 Introduction

Vision-language models (VLMs) Alayrac et al. (2022); Liu et al. (2023); Bai et al. (2023) have fundamentally reshaped visual understanding through large-scale image-text pretraining, bridging visual and linguistic spaces and enabling open-world perception beyond predefined categories. Unlike conventional object counting methods restricted to fixed object classes Lempitsky & Zisserman (2010); Bartunov & Vetrov (2018); Đukić et al. (2023), open-world object counting Amini-Naieni et al. (2024) aims to identify and enumerate arbitrary concepts specified by natural-language queries. Despite their strong semantic capabilities, however, current VLMs still struggle to distinguish and count nearby visual instances as object density increases. This limitation arises from semantic abstraction and coarse visual tokenization, which respectively weaken instance-specific details and spatial discrimination, leading to a *semantic-to-instance perception gap* in crowded scenes.

This observation raises a fundamental question: *why does strong semantic alignment not naturally translate into reliable dense instance perception?* We investigate this question from the perspective of visual representations. Existing visual foundation models, including DINOv2 Darcet et al. (2024), DINOv3 Siméoni et al. (2025), and SAM 2 Ravi et al. (2025), are trained with objectives that prioritize semantic abstraction, invariance, or region-level segmentation rather than explicit separation of neighboring instances. DINOv2 and DINOv3 learn semantically invariant representations through self-supervised objectives, while SAM2 focuses on prompt-driven mask prediction and accurate object boundaries. As illustrated



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

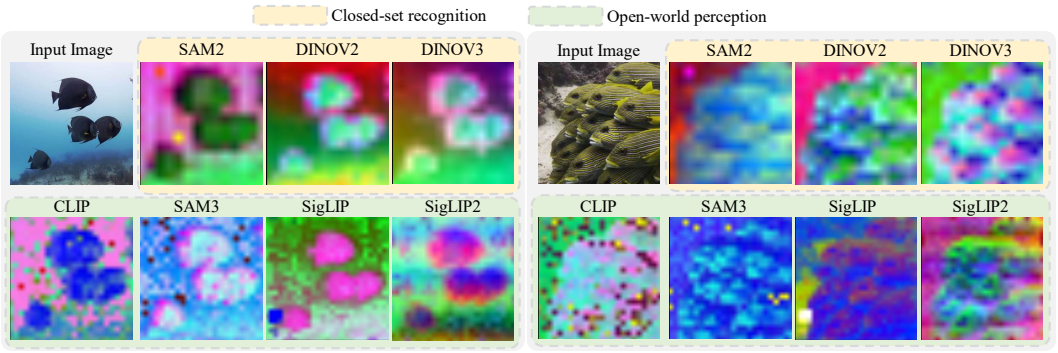



Figure 1: Comparison of VLMs with different visual backbones. VLM pretrained models (i.e. SigLIP2) are better for text-guided task due to their strong semantic alignment, but still struggle with instance perception in dense scenes.

in Figure 1, these representations remain informative in sparse scenes, but their responses become progressively entangled as density increases, making closely packed instances difficult to distinguish. This tendency is further inherited by vision-language models, where visual features are optimized to align with semantic concepts, potentially suppressing fine-grained differences between nearby instances.

Recent vision-language encoders further strengthen semantic alignment and dense visual representation. CLIP Radford et al. (2021) learns transferable visual representations through large-scale image-text contrastive learning, while SigLIP Zhai et al. (2023) replaces batch-dependent contrastive normalization with a sigmoid-based image-text matching objective. SigLIP2 Alabdulmohsin et al. (2023) further incorporates self-supervised objectives to improve localization and dense visual features while maintaining strong open-vocabulary semantics. Similarly, SAM 3 Carion et al. (2025) extends open-world visual localization through language-guided segmentation. Nevertheless, these advances are still primarily driven by semantic recognition and region-level grounding rather than explicit instance discrimination. Consequently, stronger semantic representations do not necessarily lead to proportional improvements in counting under increasing object density, exposing a semantic-to-instance perception gap in current VLMs.

To systematically study this gap, we establish a unified language-guided counting benchmark from four existing datasets and conduct controlled comparisons across representative visual encoders. Our analysis shows that stronger vision-language representations mainly benefit sparse scenes, while their advantages diminish as object density increases. Motivated by these observations, we propose **DIP-CountVLM**, a **D**ense **I**nstance **P**erception **Count**ing **VLM** that moves beyond semantic alignment toward finer-grained instance-aware visual representations. Instead of relying solely on the final-layer visual tokens, DIP-CountVLM introduces two complementary components. First, **Instance-Aware Feature Fusion (IAFF)** adaptively aggregates hierarchical visual representations to recover complementary instance-sensitive cues that may be weakened by high-level semantic abstraction. Second, **Dense Instance Token Reconstruction (DITR)** leverages image guidance to reconstruct spatially denser visual tokens, alleviating the spatial bottleneck introduced by patch-based tokenization. IAFF and DITR therefore address two complementary aspects of the semantic-to-instance gap: hierarchical feature preservation and spatial representation granularity. Extensive experiments show that DIP-CountVLM improves overall counting performance over the SigLIP2-based baseline, particularly in low- and medium-density regimes. Further analysis highlights the complementarity of hierarchical feature fusion and dense token reconstruction, while extremely dense scenes remain challenging. Our contributions are summarized as follows:

- We provide the first systematic analysis of how visual encoder design affects language-guided counting across density regimes, uncovering a *semantic-to-instance perception gap*: stronger vision-language representations improve counting in sparse scenes, but their advantages diminish as object density increases.

- We establish **DenseCountBench**, a unified benchmark that reorganizes four counting datasets under a common language-guided open-world protocol and density-aware evaluation, enabling controlled analysis of VLM counting from sparse to highly crowded scenes.
- We propose **DIP-CountVLM**, which moves beyond semantic alignment by jointly enhancing instance-sensitive features and spatial granularity with **IAFF** and **DITR**. It improves overall HitRate@100 from **49.1/7.9/2.5/11.4** to **56.7/8.3/3.8/11.5** across four benchmarks.

## 2 Related Work

**Object Counting Beyond Closed-Set Recognition.** Unlike traditional object counting methods Zhang et al. (2016); Deng et al. (2024) limited to predefined categories, open-world counting allows users to specify arbitrary target concepts. Early exemplar-based methods You et al. (2023) require visual references, while recent language-guided approaches Jiang et al. (2023); Yao et al. (2025) provide more flexible textual interfaces but suffer from insufficient fine-grained visual grounding. Despite strong semantic generalization, VLM-based methods such as COUNTGD Amini-Naieni et al. (2024) remain constrained by limited instance-level perception, since VLMs are primarily designed for global image-text alignment rather than dense object localization.

**Vision-Language Models and Semantic Alignment.** Recent advances in VLMs Radford et al. (2021) have transformed open-world visual understanding by aligning visual representations with language semantics. Models such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021) achieve strong zero-shot generalization through large-scale image-text pretraining, while BLIP Li et al. (2022a), BLIP-2 Li et al. (2023), and LLaVA Liu et al. (2023) further enhance multimodal reasoning by integrating vision encoders with large language models. Nevertheless, these models are mainly optimized for global semantic alignment and lack dense instance-level perception. Recent region-level methods, including GLIP Li et al. (2022b), OWL-ViT Minderer et al. (2022), and Grounding DINO Liu et al. (2024b), improve spatial grounding but remain primarily designed for detection rather than fine-grained instance understanding.

## 3 Method

Unlike existing VLMs Jiang et al. (2023); Yao et al. (2025) that mainly emphasize global semantic alignment, we propose DIP-CountVLM, a Dense Instance Perception VLM for open-world object counting. As shown in Figure 2, it enhances dense instance perception with two components: **Instance-Aware Feature Fusion (IAFF)**, which aggregates hierarchical Transformer features to capture instance-sensitive representations, and **Dense Instance Token Reconstruction (DITR)**, which reconstructs dense visual tokens for accurate language-guided counting in complex scenes.

### 3.1 DIP-CountVLM Overview

Given an image $I$ and a language query $Q$, open-world counting aims to predict the number of instances corresponding to the queried concept: $(I, Q) \rightarrow \hat{N}_Q$, where $\hat{N}_Q$ denotes the predicted count of the queried concept. Unlike conventional counting methods Zhang et al. (2016); Deng et al. (2024) restricted to predefined categories, open-world counting requires models to generalize toward arbitrary concepts while maintaining instance discrimination.

Our proposed DIP-CountVLM is built Qwen3-8B Yang et al. (2025) with a SigLIP2 Alabdulmohsin et al. (2023) vision encoder, which extracts hierarchical visual representations from multiple Transformer layers: $F = \{F^1, F^2, \cdots, F^L\}$, where $L$ is the number of layers and $F^l$ represents the visual tokens extracted from the $l$-th layer. DIP-CountVLM first applies **IAFF** to aggregate complementary instance cues across hierarchical representations: $F_I = \mathrm{IAFF}(F)$, where $F_I$ means the instance-enhanced visual representation.

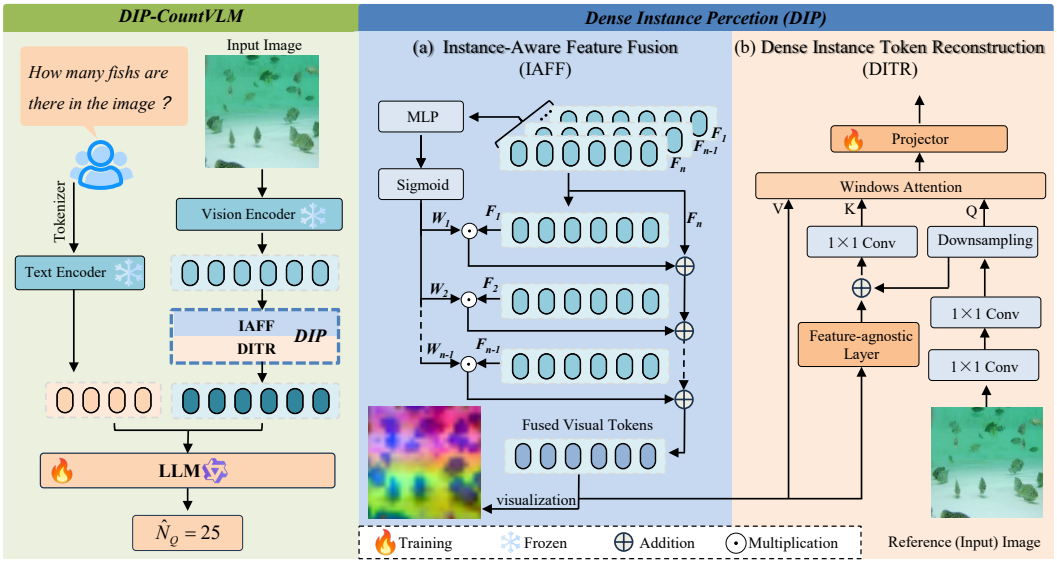



Figure 2: Architecture overview of **DIP-CountVLM**, consisting of two key components: **Instance-Aware Feature Fusion (IAFF)** and **Dense Instance Token Reconstruction (DITR)**. IAFF enhances instance-aware representations by aggregating hierarchical visual features, while DITR reconstructs dense visual tokens for language-guided counting.

To overcome the spatial information bottleneck caused by patch tokenization, **DITR** reconstructs dense spatial-aware visual representations by leveraging both the instance-enhanced features and the original image:

$$F_D = \mathrm{DITR}(I, F_I), \tag{1}$$

where $F_D$ denotes the reconstructed dense visual tokens. The dense representation provides enhanced spatial granularity and instance-level evidence for language-guided counting.

The reconstructed visual tokens are projected into the language embedding space through a visual projection layer $W_p$ where $Z_v = W_p F_D$. Given the projected visual tokens $Z_v$ and language query $Q$, the proposed DIP-CountVLM performs multimodal reasoning and predicts the target count: $\hat{N}_Q = P_\theta(Z_v, Q)$, where $P_\theta(\cdot)$ denotes the vision-language model employed in this work, namely Qwen3-8B Yang et al. (2025).

### 3.2 Instance-Aware Feature Fusion (IAFF)

Vision-language models are trained with image-text alignment objectives, where the final-layer visual representations emphasize high-level semantic abstraction. Although such semantic representations are effective for recognizing image-level concepts, dense counting requires distinguishing instances with similar appearances and precise spatial distributions.

We observe that instance-related information is distributed across different Transformer layers rather than concentrated in the final representation. Specifically, shallow layers preserve local appearance patterns and spatial structures, while deeper layers provide stronger semantic discrimination. Therefore, we propose **Instance-Aware Feature Fusion (IAFF)** to dynamically integrate hierarchical visual representations and recover instance-sensitive features for open-world counting. Different Transformer layers encode complementary visual information with different levels of semantic abstraction. Early layers retain detailed local structures that help separate adjacent instances, while deeper layers provide semantic representations for identifying queried concepts. Instead of relying solely on the final-layer features, IAFF learns adaptive fusion weights to selectively aggregate useful instance cues from different semantic levels.

**Dynamic Instance-aware Fusion.** To adaptively integrate hierarchical instance cues, IAFF introduces a lightweight routing network that predicts token-wise layer contribution weights. The final-layer representation is adopted as the semantic reference: $F_{\text{ref}} = F^L$.

For each spatial token $t$, the routing network estimates the importance of intermediate layers: $\alpha(t) = \sigma\left(W_2\delta(W_1 F_{\text{ref}}(t) + b_1) + b_2\right)$, where $\delta(\cdot)$ denotes the GELU activation function and $\sigma(\cdot)$ denotes the sigmoid function. The routing vector is formulated as: $\alpha(t) = [\alpha_1(t), \alpha_2(t), \cdots, \alpha_{L-1}(t)]$, where $\alpha_l(t)$ indicates the contribution of the $l$-th intermediate Transformer layer for token $t$. The final instance-aware representation is obtained through adaptive hierarchical aggregation:

$$F_{\text{IAFF}}(t) = F^L(t) + \sum_{l=1}^{L-1} \alpha_l(t) F^l(t). \tag{2}$$

To stabilize optimization, the bias term of the last routing layer is initialized as $b_2 = -5$, which produces small initial routing weights after sigmoid activation. Therefore, IAFF starts from the original semantic representation and gradually incorporates complementary hierarchical instance cues during training. Unlike fixed feature aggregation strategies, IAFF performs token-wise dynamic fusion, allowing different spatial regions to selectively exploit instance information from appropriate semantic levels. By combining local structural details with high-level semantic representations, IAFF improves dense instance perception and provides more reliable visual features for language-guided object counting.

### 3.3 Dense Instance Token Reconstruction (DITR)

Although IAFF enhances visual representations by aggregating hierarchical instance cues, the spatial granularity of visual tokens remains fundamentally constrained by the patch-based vision encoder. During patch embedding, local regions are compressed into individual tokens, where multiple neighboring instances may be aggregated into a single representation. Such spatial compression inevitably introduces instance ambiguity, especially in dense scenes containing small, overlapping, or visually similar objects.

To alleviate this spatial information bottleneck, we propose **Dense Instance Token Reconstruction (DITR)**, which reconstructs dense instance-aware visual representations from sparse semantic tokens. Different from conventional feature interpolation that only propagates existing semantic information, DITR performs image-guided reconstruction by exploiting the complementary information between instance-enhanced visual tokens and high-frequency image structures. Specifically, IAFF provides instance-level semantic constraints, while the input image preserves fine-grained appearance details and spatial boundaries. By jointly leveraging these two sources, DITR recovers dense visual tokens with enhanced instance sensitivity for open-world counting.

**Instance-aware Token Enhancement.** Given the instance-enhanced representation from IAFF: $F_I \in \mathbb{R}^{B\times N\times C}$, we reshape it into a spatial feature map $F_I^{2D} \in \mathbb{R}^{B\times C\times H\times W}$. The dense reconstruction is formulated as:

$$F_D = \text{DITR}(I, F_I^{2D}), \tag{3}$$

where $I$ denotes the input image and $F_D$ represents the reconstructed dense feature map. Different from interpolation-based upsampling, DITR performs image-guided reconstruction by jointly exploiting high-frequency image details and instance-aware semantic cues. The image guidance preserves local structures and object boundaries, while the instance-enhanced features constrain reconstruction toward counting-relevant regions. Consequently, DITR recovers spatially discriminative representations while maintaining semantic consistency with the original vision-language features. The reconstructed feature map is converted into dense visual tokens $F_D = \{f_1^D, f_2^D, \cdots, f_{N'}^D\}$, where $N' > N$ denotes the increased token number. These dense tokens provide finer spatial granularity and richer instance evidence, improving object localization and counting, especially for small and crowded instances.

**Dense Instance Representation for Counting.** The reconstructed dense features are projected into the language embedding space: $Z_v = W_p F_{\text{D}}$, where $W_p$ denotes the visual-language projection layer. The projected dense visual tokens are then integrated with the language query for multimodal reasoning. By recovering spatial details lost during patch tokenization, DITR bridges the gap between semantic alignment and instance-level perception, enabling VLMs to achieve more accurate open-world object counting while preserving their inherent open-world recognition capability.

# 4 Experiments

## 4.1 Experimental Settings

**Datasets.** Existing object counting datasets mainly focus on category-specific counting and lack benchmarks for language-guided open-world object counting. To address this gap, we construct an benchmark by organizing four counting datasets, including DeepFish Saleh et al. (2020) (460 samples), IOCFish5K Sun et al. (2023) (2000 samples), UnderwaterFish Wang et al. (2026) (2240 samples), and UNICBench Rong et al. (2026) (5508 samples). Following standard counting protocols, we categorize evaluation samples into three density levels according to the number of target instances: low-density (L, $\leq 10$), medium-density (M, 11–50), and high-density (H, $> 50$). To evaluate the dense counting capability of VLMs, all methods are trained on general-purpose vision-language instruction data following the LLaVA-1.5 Liu et al. (2024a) datasets and evaluated on the proposed benchmark.

**Evaluation Metrics.** Following previous counting evaluation protocols Rong et al. (2026), we adopt Hit Rate to measure counting accuracy under a given tolerance. For a relative tolerance $\tau$, the hit indicator for the $i$-th sample is defined as $\mathbb{I}_i^{(\tau)} = 1$ if $\frac{|\hat{y}_i - y_i|}{\max(1, y_i)} \leq \tau$, and 0 otherwise, where $\hat{y}_i$ and $y_i$ denote the predicted and ground-truth counts, respectively. The Hit Rate is then computed as: $\text{HitRate@}(1-\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_i^{(\tau)}$, where $N$ denotes the number of test samples. A prediction is considered correct if its relative error falls within the specified tolerance range. In our experiments, we report **HitRate@100%**, corresponding to $\tau = 0$, which requires the predicted count to exactly match the ground-truth annotation.

**Implementation Details.** Our proposed benchmarks and method are implemented in PyTorch, following the framework of Prismatic VLMs Karamcheti et al. (2024). We adopt the controlled single-stage training protocol with the same hyperparameters: batch size of 128, learning rate of 2e-5, AdamW optimizer, cosine learning rate scheduler with a warmup ratio of 0.03, max gradient norm of 1.0, and weight decay of 0.1. All models are trained using FP32 precision on 8 NVIDIA H20-3e 141GB GPUs. The vision encoder is kept frozen throughout training, while the vision-language projector and the language model are jointly optimized. We train two variants: Qwen3-8B Yang et al. (2025) and Llama-2-7B Touvron et al. (2023), each for a single stage without the separate projector alignment stage.

## 4.2 Analysis of Visual Encoders and Benchmark

Table 7 systematically evaluates how different visual encoders affect dense object counting in VLMs. Although stronger encoders yield better representations, their advantages diminish in dense scenes. With Qwen3-8B Yang et al. (2025), SigLIP2 performs best among individual encoders, achieving 49.1% and 11.4% HitRate on DeepFish and UNICBench, respectively; similar trends are observed with LLaMA2-7B Touvron et al. (2023), where SigLIP2 reaches 46.3% and 10.2%. Figure 5 further confirms that encoder differences are more evident in sparse scenes but become marginal as density increases. These results suggest that stronger visual representations alone cannot resolve the spatial ambiguity caused by coarse tokenization, highlighting the need for instance perception in dense counting. Since Qwen3-8B with SigLIP2 achieves the strongest overall performance, we adopt it as our baseline. Above 50 instances all VLMs collapse (best 1.0%, in Table 7); DIP-CountVLM alleviates this, giving the best high-density on UnderwaterFish (1.4% vs. 0.6%; Appendix C.1).

We further evaluate DIP-CountVLM on the same benchmarks using the same visual encoder setting. Compared with the strongest baseline VLM (Qwen3-8B with SigLIP2), DIP-CountVLM consistently achieves better performance across all benchmarks. Specifically, DIP-CountVLM improves the overall HitRate from 37.5% to 56.7% on DeepFish, from 7.9% to 8.3% on IOCFish5K, from 2.5% to 3.8% on UnderwaterFish, and from 11.4% to 11.5% on UNICBench. Moreover, substantial improvements are observed in low-density and medium-density scenarios, e.g., achieving 57.0% HitRate on low-density samples of DeepFish, compared with 49.3% from the SigLIP2-based baseline. These results demonstrate that DIP-CountVLM effectively enhances instance-level perception beyond existing VLM representations, enabling more accurate dense object counting.

Table 1: Comparison of different visual encoders and VLMs on four open-world counting benchmarks. Encoders are evaluated under the same VLM backbone, while the highlighted row shows the improvement of DIP-CountVLM over existing VLMs with the same visual encoder. Results are reported under low-density (L, $\leq$ 10), medium-density (M, 11–50), and high-density (H, $>$ 50) settings, as well as overall performance (higher is better).

| VisEnc | DeepFish | | | IOCfish5k | | | | UnderwaterFish | | | | UNICBench | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_{<10}$ | $M_{10-50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All |
| LLama2-7B-DinoV2 | 42.3 | 33.3 | 42.1 | 24.4 | 1.9 | 0.1 | 5.5 | 2.5 | 3.9 | 1.4 | 2.8 | 12.6 | 7.2 | 0.3 | 6.5 |
| LLama2-7B-DinoV3 | 46.5 | 33.3 | 46.3 | 32.1 | **2.1** | 0.0 | 7.0 | 2.9 | 4.5 | 1.4 | 3.2 | 12.1 | 7.5 | 0.5 | 6.6 |
| LLama2-7B-SAM2 | 30.6 | 0.0 | 30.2 | 17.3 | 0.0 | 0.0 | 3.5 | 0.0 | 0.0 | 0.0 | 0.0 | 6.3 | 4.7 | 0.2 | 3.8 |
| LLama2-7B-SAM3 | 27.3 | **50.0** | 27.6 | 15.6 | 1.7 | 0.0 | 3.6 | 2.9 | 2.0 | 1.4 | 1.8 | 15.1 | 10.0 | 0.5 | 8.5 |
| LLama2-7B-CLIP | 39.4 | 16.7 | 3 9.1 | 28.1 | 0.8 | 0.0 | 5.9 | 2.9 | 3.2 | 1.3 | 2.5 | 14.5 | 7.5 | 0.3 | 7.1 |
| LLama2-7B-SigLIP | 35.9 | 16.7 | 35.6 | 29.4 | 0.4 | 0.0 | 6.0 | **8.4** | 2.1 | 1.0 | 2.3 | 22.8 | 9.6 | 0.5 | 10.0 |
| LLama2-7B-SigLIP2 | 46.7 | 16.7 | 46.3 | 36.3 | 1.3 | 0.1 | 7.7 | 6.7 | 2.0 | 1.4 | 2.2 | 25.4 | 8.8 | 0.4 | 10.2 |
| Qwen3-8B-DinoV2 | 43.4 | 16.7 | 43.0 | 28.9 | 0.8 | 0.0 | 6.0 | 0.8 | 2.7 | 1.4 | 2.0 | 13.2 | 7.0 | 0.2 | 6.5 |
| Qwen3-8B-DinoV3 | 44.7 | 33.3 | 44.5 | 29.4 | **2.1** | 0.1 | 6.5 | 3.4 | **5.3** | 1.4 | 3.6 | 15.2 | 7.0 | 0.3 | 7.0 |
| Qwen3-8B-SAM2 | 26.2 | 0.0 | 25.8 | 17.3 | 0.4 | 0.0 | 3.6 | 0.0 | 0.0 | 0.0 | 0.0 | 11.4 | 5.9 | 0.1 | 5.5 |
| Qwen3-8B-SAM3 | 39.2 | 33.3 | 39.1 | 26.9 | 1.3 | 0.0 | 5.8 | 5.0 | 3.7 | 1.4 | 2.9 | 19.0 | 8.9 | 0.3 | 8.8 |
| Qwen3-8B-CLIP | 38.5 | 16.7 | 38.2 | 25.2 | 1.5 | 0.0 | 5.5 | 4.2 | 4.8 | 1.4 | 3.4 | 15.2 | 9.3 | 0.3 | 8.2 |
| Qwen3-8B-SigLIP | 42.3 | 33.3 | 42.1 | 29.4 | 0.8 | 0.0 | 6.1 | **8.4** | 3.5 | 1.0 | 3.0 | 25.4 | 9.5 | 0.4 | 10.5 |
| Qwen3-8B-SigLIP2 | 49.3 | 33.3 | 49.1 | 37.5 | 1.1 | 0.1 | 7.9 | 4.6 | 3.0 | 1.4 | 2.5 | 26.7 | **10.8** | 0.2 | 11.4 |
| DIP-CountVLM | **57.0** | 33.3 | **56.7** | **39.0** | 1.3 | 0.1 | **8.3** | 7.1 | 5.0 | 1.4 | **3.8** | **28.0** | 10.3 | 0.3 | **11.5** |

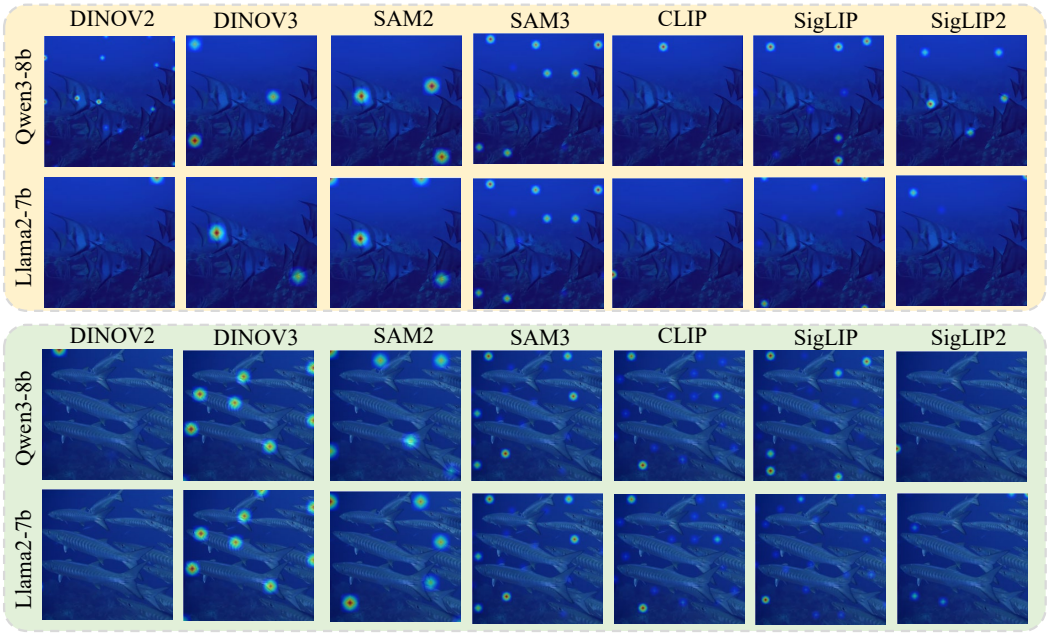



Figure 3: Comparison of attention maps generated by VLMs equipped with different visual encoders, including (a) DINOv2, (b) DINOv3, (c) SAM2, (d) SAM3, (e) CLIP, (f) SigLIP, and (g) SigLIP2. Results with Qwen3-8B and Llama2-7B on two counting scenes. Pred and GT denote the predicted and ground-truth object counts, respectively. Differences among visual encoders are more evident in sparse scenes but less so in dense scenes.

### 4.3 Analysis of Our DIP-CountVLM

Table 2 evaluates the contribution of IAFF and DITR in DIP-CountVLM. Starting from the Qwen3-8B Yang et al. (2025) with SigLIP2 Alabdulmohsin et al. (2023) baseline, we progressively introduce IAFF and DITR to analyze their individual effects.

**Effect of IAFF.** Compared with Table 2 (a) and (b), introducing IAFF alone brings limited and inconsistent improvements under different density conditions. Specifically, IAFF provides marginal gains in some high-density scenarios, improving the performance from 10.8% to 11.2% on DeepFish and from 0.2% to 0.4% on UNICBench. However, such improvements do not generalize across density variations, as the performance degrades in low-

Table 2: Ablation study of IAFF and DITR in DIP-CountVLM. The baseline model is built upon Qwen3-8B with SigLIP2 as the visual encoder. Starting from this baseline, we progressively introduce IAFF and DITR to analyze their contributions to open-world dense object counting. Results are reported on four benchmarks under different density levels, including low-density (L, $\leq 10$), medium-density (M, 11–50), and high-density (H, $> 50$).

| Exp. | DIP-CountVLM | | DeepFish | | | IOCfish5k | | | | UnderwaterFish | | | | UNICBench | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IAFF | DITR | $L_{<10}$ | $M_{10-50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All |
| (a) | ✗ | ✗ | 49.3 | 33.3 | 49.1 | 37.5 | 1.1 | 0.1 | 7.9 | 4.6 | 3.0 | 1.4 | 2.5 | 26.7 | 10.8 | 0.2 | 11.4 |
| (b) | ✓ | ✗ | 48.9 | 33.3 | 48.7 | 34.6 | 0.9 | 0.0 | 7.2 | 5.5 | 3.0 | 1.4 | 2.6 | 24.8 | 11.2 | 0.4 | 11.3 |
| (c) | ✗ | ✓ | 54.4 | 33.3 | 54.1 | 36.8 | 1.1 | 0.0 | 7.7 | 10.1 | 3.3 | 0.0 | 2.8 | 25.9 | 11.9 | 0.3 | 11.8 |
| (d) | ✓ | ✓ | **57.0** | **33.3** | **56.7** | **39.0** | **1.3** | **0.1** | **8.3** | **7.1** | **5.0** | **1.4** | **3.8** | **28.0** | **10.3** | **0.3** | **11.5** |

Table 3: Compared with fixed feature fusion, dynamic fusion adaptively integrates informative features and preserves instance-level details, especially in crowded scenes.

| IAFF | Dynamic Fusion | | | | Fixed Fusion | | | |
|---|---|---|---|---|---|---|---|---|
| | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All |
| **DeepFish** | 57.0 (+3.7) | 33.3 | - | 56.7 (+3.7) | 53.3 | 33.3 | - | 53.0 |
| **IOCfish5k** | 39.0 (+2.2) | 1.3 (-0.2) | 0.1 | 8.3 (+0.4) | 36.8 | 1.5 | 0.1 | 7.9 |
| **UnderwaterFish** | 7.1 (-1.3) | 5.0 (-0.3) | 1.4 | 3.8 (-0.4) | 8.4 | 5.3 | 1.4 | 4.2 |
| **UNICBench** | 28.0 (+2.8) | 10.3 (-1.0) | 0.3 (+0.2) | 11.5 (+0.2) | 25.2 | 11.3 | 0.1 | 11.3 |

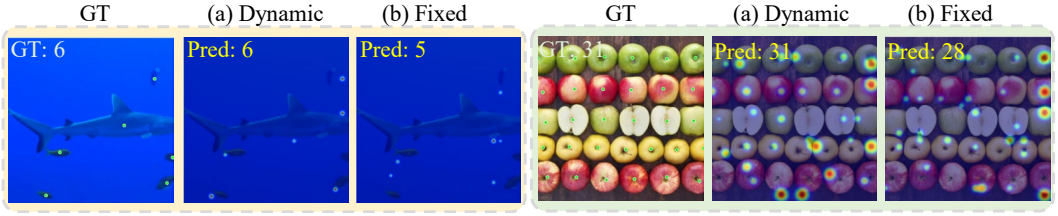



Figure 4: Comparison of fixed and dynamic feature fusion across sparse and dense scenarios.

density cases by -0.4% , -2.9% , and -1.9% on DeepFish, UnderwaterFish, and UNICBench, respectively. These results suggest that while hierarchical feature aggregation can enhance instance-related semantic representation, it alone cannot robustly improve counting performance across diverse object distributions.

**Effect of DITR.** Compared with the baseline, row (c) in Table 2 shows that DITR consistently improves counting performance across different density levels. It increases the overall HitRate from 49.1% to 54.1% on DeepFish and from 2.5% to 2.8% on UnderwaterFish, with particularly notable gains in low-density cases, improving from 49.3% to 54.4% on DeepFish and from 4.6% to 10.1% on UnderwaterFish. These results demonstrate that DITR alleviates the spatial bottleneck of patch-based tokenization by reconstructing fine-grained instance-aware tokens, reducing token collisions and improving instance for counting.

**Effect of Our DIP-CountVLM.** Combining IAFF and DITR achieves the best overall performance, demonstrating their complementary roles in dense object counting. Compared with DITR alone, incorporating IAFF further improves the overall HitRate from 54.1% to 56.7% on DeepFish and from 7.7% to 8.3% on IOCFish5K. This improvement suggests that IAFF provides instance-aware semantic guidance that benefits dense token reconstruction, while DITR complements IAFF by recovering the fine-grained spatial information that remains ambiguous after feature fusion. Together, IAFF and DITR jointly enhance semantic instance awareness and spatial instance discrimination, leading to more accurate counting.

### 4.4 Ablation Studies

We conduct ablation studies to validate our design choices, including the effectiveness of dynamic instance-aware routing, different DITR training strategies, and DITR integration locations within the vision-language architecture.

**Dynamic Fusion vs. Fixed Fusion.** Table 3 compares dynamic instance-aware fusion with fixed feature fusion across different object densities, where fixed fusion assigns a con-

stant weight of 0.5 to each feature level. Dynamic fusion consistently outperforms fixed fusion, improving HitRate on DeepFish (53.0%→56.7%), IOCfish5k (7.9%→8.3%), and UNICBench (11.37%→11.5%). Figure 4 further highlights its advantage in dense scenes, preserving 31 instances versus 28 for fixed fusion under the same GT of 31. These results demonstrate that adaptive feature aggregation better preserves discriminative instance-level information for open-world dense object counting.

Table 4: Compared with freezing DITR, jointly optimizing the module with the counting objective enables task-aware token refinement and better preserves instance-level details.

| DITR | Freezing | | | | Training | | | |
|---|---|---|---|---|---|---|---|---|
| | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All |
| **DeepFish** | 52.9 | 33.3 | - | 52.6 | 54.4 (+1.5) | 33.3 | - | 54.1 (+1.5) |
| **IOCfish5k** | 36.8 | 1.3 | 0.0 | 7.8 | 36.8 | 1.1 (-0.2) | 0.0 | 7.7 (-0.1) |
| **UnderwaterFish** | 8.4 | 3.8 | 0.0 | 2.8 | 10.1 (+1.7) | 3.3 (-0.5) | 0.0 | 3.8 (+1.0) |
| **UNICBench** | 25.8 | 10.7 | 0.3 | 11.2 | 25.9 (+0.1) | 11.9 (+1.2) | 0.3 | 11.8 (+0.6) |

**Effect of Freezing and Training DITR.** Table 4 compares freezing DITR after visual encoding with jointly optimizing it using the counting objective. Jointly training DITR consistently improves overall performance on DeepFish (52.6→54.1) and UNICBench (11.2→11.8), particularly in medium-density scenarios, demonstrating that task-aware optimization enables more effective dense token refinement and better preserves instance-sensitive details. However, the marginal gains on IOCfish5k and UnderwaterFish suggest that extreme density and severe instance ambiguity remain challenging even with adaptive token reconstruction. Overall, these results validate the importance of jointly training DITR for dense counting.

Table 5: Deep analysis of Frozen DITR in DIP-CountVLM. 'F' and 'T' denote frozen and trainable modules, respectively.

| | DITR | IAFF | $L_{<10}$ | $M_{10-20}$ | $M_{20-50}$ | $H_{50-100}$ | $H_{>100}$ | All |
|---|---|---|---|---|---|---|---|---|
| **DeepFish** | F | F | 51.5 | 33.3 | - | - | - | 51.3 |
| **DeepFish** | F | T | 51.8 (+0.3) | 33.3 | - | - | - | 51.5 (+0.2) |
| **DeepFish** | T | T | 57.0 (+5.5) | 33.3 | - | - | - | 56.7(+5.2) |
| **IOCfish5k** | F | F | 39.8 | 3.2 | 0.0 | 0.0 | 0.0 | 8.3 |
| **IOCfish5k** | F | T | 39.3 (-0.5) | 4.8 (+1.6) | 0.0 | 0.0 | 0.1 (+0.1) | 8.4 (+0.1) |
| **IOCfish5k** | T | T | 39.0 (-0.8) | 3.7(+0.5) | 0.0 | 0.0 | 0.1 (+0.1) | 8.3 |
| **UnderwaterFish** | F | F | 8.0 | 14.3 | 0.0 | 0.0 | 6.7 | 3.7 |
| **UnderwaterFish** | F | T | 7.6 (-0.4) | 14.8 (+0.5) | 0.0 | 0.0 | 8.0 (+1.3) | 3.9 (+0.2) |
| **UnderwaterFish** | T | T | 7.1 (-0.9) | 14.8 (+0.5) | 0.0 | 0.0 | 8.0 (+1.3) | 3.8 (+0.1) |
| **UNICBench** | F | F | 26.8 | 20.4 | 2.5 | 0.0 | 0.4 | 11.9 |
| **UNICBench** | F | T | 26.7 (-0.1) | 21.1 (+0.7) | 2.4 (-0.1) | 0.0 | 0.4 | 12.0 (+0.1) |
| **UNICBench** | T | T | 28.0 (+1.2) | 18.3 (-1.1) | 2.0 (-0.5) | 0.0 | 0.6 (+0.2) | 11.5 (-0.5) |

**Deep Analysis of Frozen DITRM.** Table 5 further analyzes the impact of training DITR across different density levels. With DITR frozen, adding IAFF yields only marginal gains, whereas jointly training DITR enables clear improvements, especially in low-density cases. For example, training DITR improves low-density performance on DeepFish and UNICBench from 51.5% to 57.0% and 26.8% to 28.0%, respectively, while also improving medium- and high-density performance on UnderwaterFish from 14.3% to 14.8% and 6.7% to 8.0%, respectively. These results validate the necessity of training DITR, which learns spatial representations and improves instance-level perception across diverse density scenarios.

## 5 Conclusion

In this work, we investigate the capability of VLM visual encoders for dense open-world object counting. We establish a benchmark to evaluate their performance across different densities, revealing that stronger visual representations alone are insufficient for dense instance perception. To address this limitation, we propose DIP-CountVLM, which combines Instance-Aware Feature Fusion (IAFF) for instance-sensitive feature enhancement and Dense Instance Token Reconstruction (DITR) for dense visual token reconstruction, enabling fine-grained instance perception for language-guided open-world object counting.

### AI use statement

In this work, we used generative AI tools for drafting parts of the research paper, specifically for grammar checking and polishing the manuscript text. We have not used generative AI tools for generating synthetic datasets, developing theoretical models, formulating or proving mathematical claims, proposing or refining hypotheses, designing research methodology, implementing methods, or interpreting results; these tasks are not applicable to this work. Additionally, we used generative AI tools for debugging software code and improving the readability of the manuscript. We have reviewed all AI-assisted work. Specifically, we manually verified all AI-suggested grammar corrections and code fixes for correctness and consistency with the intended meaning, and we confirmed that no AI-generated content was used for scientific claims or experimental results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### Reproducibility statement

We describe the artifacts that support the reproducibility of this work. All four benchmarks used in our experiments are publicly available, and we introduce no new dataset: DeepFish Saleh et al. (2020), IOCFish5K Sun et al. (2023), UnderwaterFish Wang et al. (2026), and UNICBench Rong et al. (2026). The construction of our evaluation protocol, including how each benchmark is split into low-, medium-, and high-density subsets, is described in Section 4.1 and Appendix B, and the definition of the reported HitRate@100 metric is given in Section 4.1. The complete specification of our proposed method is provided in Section 3, where the Instance-Aware Feature Fusion (IAFF) and Dense Instance Token Reconstruction (DITR) modules are formulated, and the full inference pipeline is additionally summarized as pseudocode with explicit equations in Appendix A. All training and evaluation settings are reported in Section 4.1, including the optimizer, learning-rate schedule, batch size, numerical precision, hardware, and the modules that are kept frozen or jointly optimized, together with the instruction-tuning data on which all compared models are trained. The same protocol is applied to our method and to all compared visual encoders and vision-language models, whose quantitative results are reported in Sections 4.2–4.4 and Appendix C, where we also provide the analysis of IAFF layer selection.

## References

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. In *Advances in Neural Information Processing Systems*, pp. 16406–16425, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2024.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence, 2026.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang,

Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025b.

Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 670–678. PMLR, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, François Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, pp. 2632–2652, 2024.

Lijia Deng, Qinghua Zhou, Shuihua Wang, Juan Manuel Górriz, and Yudong Zhang. Deep learning in crowd counting: A survey. *CAAI Transactions on Intelligence Technology*, 9 (5):1043–1077, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916, 2021.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4535–4545, 2023.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in Neural Information Processing Systems*, 23, 2010.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-VL: Towards real-world vision-language understanding, 2024.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Meta AI Blog, 2025. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed: 2026-09-26.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs, 2025.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pp. 728–755, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, pp. 28085–28128, 2025.

Chenggang Rong, Tao Han, Zhiyuan Zhao, Yaowu Fan, Jia Wan, Song Guo, Yuan Yuan, and Junyu Gao. Unicbench: Unified counting benchmark for MLLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23730–23740, June 2026.

Alzayat Saleh, Issam H. Laradji, Dmitry A. Konovalov, Michael Bradley, David Vazquez, and Marcus Sheaves. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Scientific Reports*, 10(1):14671, 2020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025.

Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13791–13801, June 2023.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu-hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar

Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

Nikola Đukić, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18872–18881, 2023.

Jinfeng Wang, Jinze Lv, Zhipeng Cheng, and Qiong Huang. Fcnet: Accurate counting model of fish in complex underwater environments. *Computers and Electronics in Agriculture*, 240:111172, 2026. ISSN 0168-1699.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai

Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Ziyu Yao, Xuxin Cheng, Zhiqi Huang, and Lei Li. Countllm: towards generalizable repetitive action counting via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19143–19153, 2025.

Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6315–6324, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 589–597, 2016.

# A Algorithm Details

**Overview.** Algorithm 1 summarizes the inference pipeline of DIP-CountVLM for open-world object counting. Given an input image $I$ and a language query $Q$, DIP-CountVLM first extracts hierarchical visual representations from the vision encoder. Then, the proposed Instance-Aware Adaptive Feature Fusion (IAFF) module adaptively aggregates multi-level visual features to recover instance-sensitive information. Subsequently, the Dense Instance Token Reconstruction (DITR) module enhances the spatial resolution of visual representations and generates dense instance-aware visual tokens. Finally, the enhanced visual tokens are projected into the language embedding space and processed by the VLM decoder to predict the target object count.

## A.1 Hierarchical Visual Representation Extraction

Given an input image $I$, the vision encoder $\Phi_v$ extracts hierarchical visual representations:

$$F = \{F^1, F^2, \cdots, F^L\} = \Phi_v(I), \tag{4}$$

where $F^l$ denotes the visual representation from the $l$-th layer. The shallow layers preserve fine-grained spatial structures, while deeper layers contain stronger semantic abstractions. However, existing VLMs mainly rely on the final-layer representation $F^L$, which is optimized for vision-language semantic alignment and may discard fine-grained instance-level information required for dense object counting.

## A.2 Instance-Aware Adaptive Feature Fusion

To bridge the semantic-instance gap, we propose the Instance-Aware Adaptive Feature Fusion (IAFF) module. The deepest feature representation is used as the semantic reference:

**Algorithm 1** DIP-CountVLM for Open-World Object Count

**Require:** Image $I$, language query $Q$, vision encoder $\Phi_v$, IAFF module $\mathcal{A}$, DITR module $\mathcal{D}$, projection layer $W_p$, VLM decoder $P_\theta$

**Ensure:** Predicted object count $\hat{N}_Q$

Extract hierarchical visual representations from the vision encoder:
  $F = \{F^1, F^2, \cdots, F^L\} = \Phi_v(I)$
Initialize semantic reference feature:
  $F_{\text{ref}} = F^L$
**for** each visual token $t$ **do**
  Predict layer-wise routing weights:
    $\alpha(t) = \sigma(W_2\delta(W_1 F_{\text{ref}}(t) + b_1) + b_2)$
  Fuse hierarchical instance cues:
    $F_{\text{IAFF}}(t) = F^L(t) + \sum_{l=1}^{L-1} \alpha_l(t) F^l(t)$
**end for**
Reshape fused tokens into spatial feature maps:
  $F^{2D}_{\text{IAFF}} \leftarrow F_{\text{IAFF}}$
Reconstruct dense instance-aware visual features:
  $F_D = \mathcal{D}(I, F^{2D}_{\text{IAFF}})$
Convert reconstructed feature maps into dense visual tokens:
  $F_D = \{f_1^D, f_2^D, \cdots, f_{N'}^D\}$
Project dense visual tokens into language embedding space:
  $Z_v = W_p F_D$
Perform vision-language reasoning:
  $\hat{N}_Q = P_\theta(Z_v, Q)$
**return** $\hat{N}_Q$

$$F_{\text{ref}} = F^L. \tag{5}$$

For each visual token $t$, IAFF predicts token-specific routing weights:

$$\alpha(t) = \sigma\left(W_2\delta(W_1 F_{\text{ref}}(t) + b_1) + b_2\right), \tag{6}$$

where $\delta(\cdot)$ denotes the nonlinear activation function and $\sigma(\cdot)$ represents the sigmoid function. The routing weights dynamically determine the contribution of different hierarchical features. The fused feature representation is formulated as:

$$F_{\text{IAFF}}(t) = F^L(t) + \sum_{l=1}^{L-1} \alpha_l(t) F^l(t). \tag{7}$$

Unlike conventional feature fusion with fixed weights, IAFF performs token-wise adaptive aggregation, enabling each spatial location to selectively integrate semantic and detailed visual information according to its instance perception requirement.

### A.3 Dense Instance Token Reconstruction

After hierarchical feature fusion, the enhanced tokens are reshaped into spatial feature maps:

$$F^{2D}_{\text{IAFF}} \leftarrow F_{\text{IAFF}}. \tag{8}$$

Since the original visual tokens extracted by VLM encoders usually have limited spatial resolution, we introduce the Dense Instance Token Reconstruction (DITR) module to recover dense spatial representations:

$$F_D = \mathcal{D}(I, F^{2D}_{\text{IAFF}}), \tag{9}$$

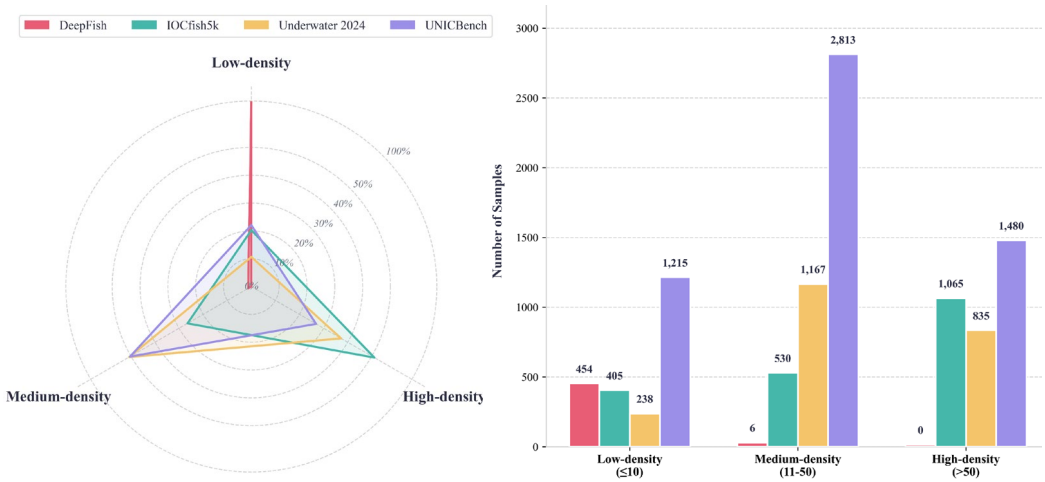



Figure 5: Sample density distribution comparison across four datasets. The left radar chart illustrates the relative proportional distribution of samples across three density categories, while the right grouped bar chart presents the absolute sample counts. Density levels are categorized as low-density (L, $\leq$ 10), medium-density (M, 11–50), and high-density (H, $>$ 50).

where $\mathcal{D}(\cdot)$ denotes the proposed dense reconstruction module. By leveraging both image-level spatial information and enhanced visual features, DITR reconstructs high-resolution instance-aware representations for dense object counting.

The reconstructed feature map is further converted into dense visual tokens:

$$F_D = \{f_1^D, f_2^D, \cdots, f_{N'}^D\}, \tag{10}$$

where $N'$ denotes the number of generated dense visual tokens.

### A.4 Vision-Language Counting Reasoning

The reconstructed dense visual tokens are projected into the language embedding space:

$$Z_v = W_p F_D. \tag{11}$$

The projected visual tokens are then combined with the language query $Q$ and processed by the VLM decoder:

$$\hat{N}_Q = P_\theta(Z_v, Q), \tag{12}$$

where $P_\theta$ represents the VLM reasoning module. The decoder performs language-conditioned reasoning and outputs the predicted object count $\hat{N}_Q$.

**Summary.** DIP-CountVLM enhances dense instance perception before vision-language reasoning. IAFF alleviates the semantic-instance gap by adaptively integrating hierarchical visual features, while DITR improves spatial granularity by reconstructing dense instance-aware tokens. Together, these components transform semantic-oriented VLM representations into counting-oriented representations, enabling accurate open-world object counting.

## B Dataset Density Distribution

Our benchmark is constructed from four publicly available datasets: DeepFish Saleh et al. (2020), IOCfish5k Sun et al. (2023), UnderwaterFish Wang et al. (2026), and UNICBench Rong et al. (2026). Dense object counting requires models to distinguish individual instances

Table 6: Analysis of feature layer selection in IAFF. By adaptively integrating hierarchical SigLIP2 representations, IAFF effectively exploits complementary spatial and semantic information, leading to improved dense object counting performance.

| Fusion Layers | DeepFish | | | IOCfish5k | | | | UnderwaterFish | | | | UNICBench | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_{<10}$ | $M_{10-50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All |
| {1, 4, 6, 10, 16, 24, 27} | 53.1 | 33.3 | 52.8 | 37.8 | 1.1 | 0.1 | 8.0 | **8.8** | 4.0 | 1.4 | 3.5 | 24.0 | 11.1 | 0.2 | 11.0 |
| {3, 6, 10, 14, 27} | 47.6 | 16.7 | 47.1 | 34.6 | 0.9 | 0.1 | 7.3 | 4.6 | 3.5 | 1.4 | 2.8 | 24.9 | 10.8 | 0.1 | 11.0 |
| {4, 6, 10, 16, 27} | 45.4 | 33.3 | 45.2 | 35.1 | 1.1 | 0.1 | 7.4 | 5.0 | 3.3 | 1.3 | 2.7 | 24.9 | **12.1** | 0.3 | 11.7 |
| All layers | **57.0** | **33.3** | **56.7** | **39.0** | **1.3** | **0.1** | **8.3** | 7.1 | **5.0** | **1.4** | **3.8** | **28.0** | 10.3 | **0.3** | **11.5** |

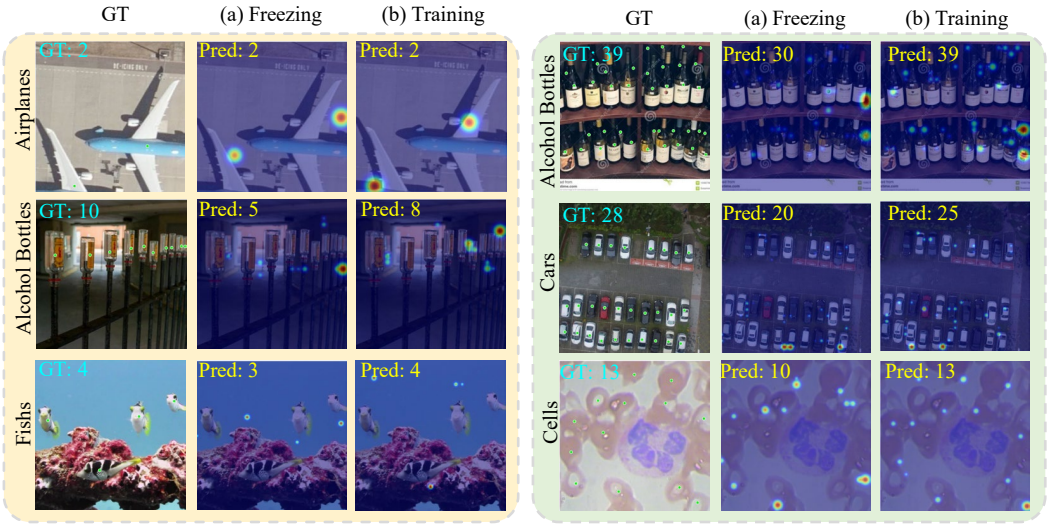



Figure 6: Visualization of estimated count predictions and density maps. Column (a) and (b) present predictions under Freezing DITR and Jointly Training DITR, respectively. Joint optimization enables the model to effectively separate overlapping/dense instances (e.g., Alcohol Bottles with GT: 39 and Cells with GT: 13), yielding precise localized responses and confirming the necessity of task-aware token refinement for dense counting.

under varying levels of object density. To evaluate this capability, we construct a benchmark containing diverse density patterns. As shown in Fig. 1, all samples are grouped into three categories based on instance numbers: low-density ($\leq$10), medium-density (11–50), and high-density (>50).

The benchmark presents a large density variation across datasets. While low-density samples focus on basic object recognition, medium- and high-density samples introduce more challenging conditions where semantic understanding alone is insufficient. These crowded scenarios require models to preserve fine-grained spatial information and instance-level visual representations for accurate counting.

The combined dataset includes diverse density distributions and imaging conditions, covering both sparse and highly crowded scenes. Such diversity allows a thorough evaluation of whether vision-language models can overcome the semantic-instance gap and acquire dense instance perception ability for open-world object counting.

## C Additional Experiments

In this section, we provide additional experiments to further validate the effectiveness and robustness of our proposed DIP-CountVLM. First, we investigate the impact of different SigLIP2 Tschannen et al. (2025) feature layer selections on IAFF, demonstrating that aggregating multi-level visual features consistently achieves superior counting performance and verifies the necessity of hierarchical feature integration. Second, we present qualitative

comparisons of DITR optimization strategies, showing that jointly optimizing DITR with the counting objective enables more accurate dense instance representations and alleviates the loss of fine-grained spatial information. Finally, we provide visual evidence of DIP-CountVLM on diverse open-world object counting scenarios, where the generated density maps and predictions further illustrate its capability to perceive and count instances under varying object categories and densities.

Table 7: Comparison of existing VLMs and DIP-CountVLM on four open-world counting benchmarks. The highlighted row shows the results of DIP-CountVLM. Results are reported under low-density (L, $\leq 10$), medium-density (M, 11–50), and high-density (H, $> 50$) settings, as well as overall performance (higher is better).

| Method | DeepFish | | | IOCfish5k | | | | UnderwaterFish | | | | UNICBench | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $L_{<10}$ | $M_{10-50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All | $L_{<10}$ | $M_{10-50}$ | $H_{>50}$ | All |
| DeepSeek-VL | 45.8 | 0.0 | 45.2 | 34.6 | 0.6 | 0.0 | 7.1 | 3.4 | 0.8 | 0.0 | 0.7 | 29.2 | 11.0 | 0.7 | 12.2 |
| DeepSeek-VL2 | 56.4 | 0.0 | 55.6 | 37.8 | 0.0 | 0.0 | 7.6 | 3.8 | 0.1 | 0.1 | 0.4 | 41.0 | 9.6 | 0.5 | 14.0 |
| Gemma4-E4B | 32.4 | 0.0 | 31.9 | 26.2 | 1.1 | 0.0 | 5.6 | 5.9 | 0.9 | 0.4 | 1.2 | 26.9 | 7.7 | 0.5 | 9.9 |
| InternVL3.5-8B | 43.4 | 0.0 | 42.8 | 29.6 | 0.4 | 0.0 | 6.1 | 4.2 | 0.0 | 0.0 | 0.4 | 42.8 | 14.8 | **1.0** | 17.2 |
| LLaVA-OneVision-2 | 55.5 | 16.7 | 55.0 | 35.6 | 1.9 | 0.1 | 7.7 | 6.7 | 0.2 | 0.2 | 0.8 | 42.1 | 13.3 | 0.5 | 16.2 |
| Llama4-Scout-17B | 37.9 | 0.0 | 37.4 | 37.5 | **4.3** | 0.0 | **8.7** | **13.9** | 1.5 | 0.0 | 2.2 | 39.9 | 11.6 | **1.0** | 15.0 |
| Phi4-multimodal | 35.7 | 0.0 | 35.2 | 28.6 | 1.5 | **0.3** | 6.3 | 2.1 | 0.3 | 0.0 | 0.3 | 30.8 | 8.2 | 0.9 | 11.2 |
| Qwen2.5-VL-7B | 45.8 | 0.0 | 45.2 | 30.9 | 0.2 | 0.1 | 6.3 | 2.1 | 0.1 | 0.0 | 0.2 | 41.6 | 13.9 | 0.5 | 16.3 |
| Qwen3-VL-8B | 55.3 | 0.0 | 54.5 | 34.1 | 1.7 | 0.0 | 7.3 | 7.6 | 0.3 | 0.6 | 1.2 | **45.1** | **16.1** | 0.5 | **18.3** |
| DIP-CountVLM | **57.0** | **33.3** | **56.7** | **39.0** | 1.3 | 0.1 | 8.3 | 7.1 | **5.0** | **1.4** | **3.8** | 28.0 | 10.3 | 0.3 | 11.5 |

## C.1 Comparison with General-Purpose VLMs

Table 7 compares DIP-CountVLM with nine open-source vision-language models on the four benchmarks, using the same evaluation protocol and HitRate@100% metric described in Section 4.1 of the original paper. The compared models span a wide range of scales and architectures, including DeepSeek-VL Lu et al. (2024) and DeepSeek-VL2 Wu et al. (2024), Gemma4-E4B Team et al. (2026), InternVL3.5-8B Wang et al. (2025), LLaVA-OneVision-2 An et al. (2026), Llama4-Scout-17B Meta AI (2025), Phi4-multimodal Microsoft et al. (2025), Qwen2.5-VL-7B Bai et al. (2025b), and Qwen3-VL-8B Bai et al. (2025a).

The most consistent pattern is the rapid degradation of all general-purpose VLMs as object density increases. On UNICBench, Qwen3-VL-8B attains 45.1% in low-density scenes but only 16.1% in medium-density and 0.5% in high-density scenes; Llama4-Scout-17B, the strongest model on IOCfish5k, similarly drops from 37.5% to 4.3% and 0.0% across the three density levels. The trend is even more pronounced on DeepFish, where nearly every compared VLM falls to 0.0% in the medium-density split, and on UnderwaterFish, where no baseline exceeds 0.6% in the high-density split. These models can therefore recognize the queried category in sparse scenes but fail once instances become crowded, indicating that the bottleneck lies in instance-level perception rather than in semantic understanding, which is the central premise of this work.

Increasing model capacity does not remove this limitation: despite the wide range of scales covered by the compared models, none exceeds 1.0% on the high-density split of UNICBench. This mirrors the behavior observed for visual encoders in Section 4.2 of the original paper and supports our claim that stronger semantic representations alone cannot recover the spatial evidence required for dense counting.

DIP-CountVLM benefits from its instance-aware design in precisely these regimes. It obtains the best overall HitRate on DeepFish (56.7%) and UnderwaterFish (3.8%), together with the best medium- and high-density results on UnderwaterFish (5.0% and 1.4%), where the relative margin over all baselines is the largest. On UNICBench, however, general-purpose backbones remain clearly better, with Qwen3-VL-8B reaching 45.1% in the low-density split and 18.3% overall; on IOCfish5k, DIP-CountVLM obtains the best low-density result, while Llama4-Scout-17B remains marginally better overall (8.7% vs. 8.3%). These results suggest that semantic grounding and instance perception are complementary: scaling the language backbone primarily strengthens open-vocabulary recognition, whereas IAFF and DITR address the spatial ambiguity that scaling alone leaves unresolved, making the two directions promising to combine.

### C.2 Effect of IAFF across Different Feature Levels

Table 6 studies the influence of different SigLIP2 layer selections on IAFF. The results demonstrate that exploiting all-layer features consistently achieves the best overall performance, validating the importance of comprehensive multi-level feature aggregation. Specifically, compared with the best-performing partial layer selection, all-layer fusion improves the overall HitRate from 45.2% to 56.7% on DeepFish, from 7.3% to 8.3% on IOCfish5k, and from 11.0% to 11.5% on UNICBench. On UnderwaterFish, it also achieves the highest overall performance of 3.8%, outperforming the selected layer combinations.

Furthermore, the improvements are consistently observed across different density levels. For example, on DeepFish, all-layer fusion achieves the best results in both low-density and medium-density cases, improving the low-density performance from 45.4% to 57.0% compared with the best partial layer setting. On IOCfish5k, it increases the low-density HitRate from 34.6% to 39.0%, while maintaining improvements in medium-density cases. Similarly, on UNICBench, all-layer fusion obtains the highest low-density performance of 28.0%, compared with 24.0% from partial layer fusion.

These results indicate that SigLIP2 Tschannen et al. (2025) layers provide complementary information for dense counting. Lower layers preserve spatial details beneficial for instance localization, whereas higher layers provide stronger semantic representations for object recognition. By incorporating all layers, IAFF can adaptively exploit complementary hierarchical features and select informative representations according to different object densities, leading to more accurate instance perception and counting performance.

### C.3 Qualitative Analysis of DITR Optimization

To intuitively analyze the effect of task-aware optimization, Figure 6 visualizes the predicted density maps under two training strategies: (a) freezing DITR and (b) jointly training DITR without IAFF.

As shown in Figure 6, freezing DITR often produces incomplete and diffuse density responses, especially in crowded regions where multiple instances are closely distributed. Although the frozen model can roughly localize object regions, it lacks sufficient instance-sensitive spatial information to distinguish neighboring objects, leading to missing responses and inaccurate density estimation. In contrast, jointly training DITR generates sharper and more concentrated density maps that better match the ground-truth object distributions. For example, in medium-to-high density scenarios such as Alcohol Bottles (GT: 39) and Cells (GT: 13), the trained DITR successfully separates individual instances and recovers more complete density peaks, achieving predictions closer to the ground truth (Pred: 39 and Pred: 13). These results demonstrate that task-aware optimization enables DITR to adapt its visual token reconstruction toward counting-specific spatial granularity, rather than relying on the original semantic-oriented representation.

Figure 8 further investigates the combined effect of IAFF and DITR training. Compared with the frozen DITR setting, the integration of IAFF provides richer multi-level visual cues, while jointly optimizing DITR further enhances the instance-awareness of the refined tokens. In sparse scenes, such as Airplanes and Fish, the frozen model with IAFF can identify the overall object regions, but its density responses remain scattered and incomplete, resulting in unstable counting results (e.g., Fish: GT 4, Pred 3). After jointly training DITR, the predicted responses become more localized around individual instances, improving the alignment between density peaks and object centers.

The improvements become more significant in dense scenarios with severe object overlap. For Alcohol Bottles (GT: 39) and Cells (GT: 13), the IAFF-enhanced model with frozen DITR still struggles to separate adjacent instances, producing blurred activations and under-counting (Pred: 30 and Pred: 10). After jointly training DITR, the model generates sharper and better-separated density responses, accurately recovering densely packed instances (Pred: 39 and Pred: 13). Similar improvements can be observed in Cars (GT: 28), where the combination of IAFF and trained DITR provides more complete coverage of small and crowded objects.

Overall, these qualitative comparisons reveal the complementary roles of IAFF and DITR. IAFF enriches visual representations by adaptively aggregating hierarchical features, while DITR further transforms these features into dense instance-aware visual tokens through task-aware optimization. Together, they narrow the gap between semantic visual representations and counting-oriented representations, enabling more precise and robust open-world object counting across diverse density levels.

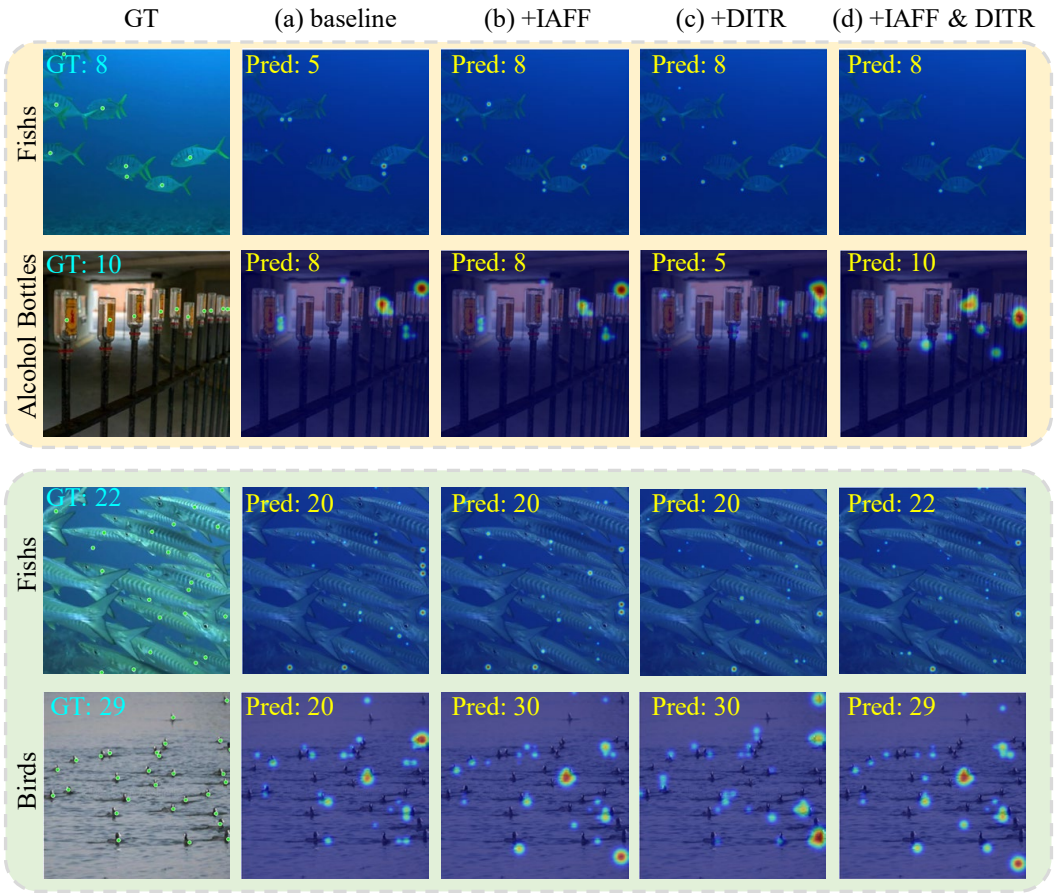



Figure 7: Qualitative visualization of counting results with different components. Each row presents the input image, ground truth count, and predicted counts from the (a) baseline, (b) IAFF, (c) DITR, and (d) IAFF+DITR (Our DIP-CountVLM). The first two rows show sparse cases, while the last two rows show dense cases. The results demonstrate that IAFF and DITR provide complementary improvements by enhancing semantic awareness and spatial instance perception, respectively.

### C.4 Visual Evidence of Our DIP-CountVLM

To further understand the effects of different components, we provide qualitative comparisons corresponding to the ablation results in Table 2 of the original paper. As shown in Fig. 7, the first two rows represent sparse object counting scenarios, where objects are relatively separated, while the last two rows correspond to dense object counting scenarios with severe object overlap and higher instance ambiguity.

**Effect of IAFF.** Compared with the baseline, introducing IAFF alone brings limited but interpretable improvements. As shown in Table 2 (a) and (b) of the original paper, IAFF improves performance in some cases, particularly for high-density scenarios, but the gains are inconsistent across different object distributions. The visualization results provide further evidence of this behavior. In sparse scenarios (the first two rows), both the

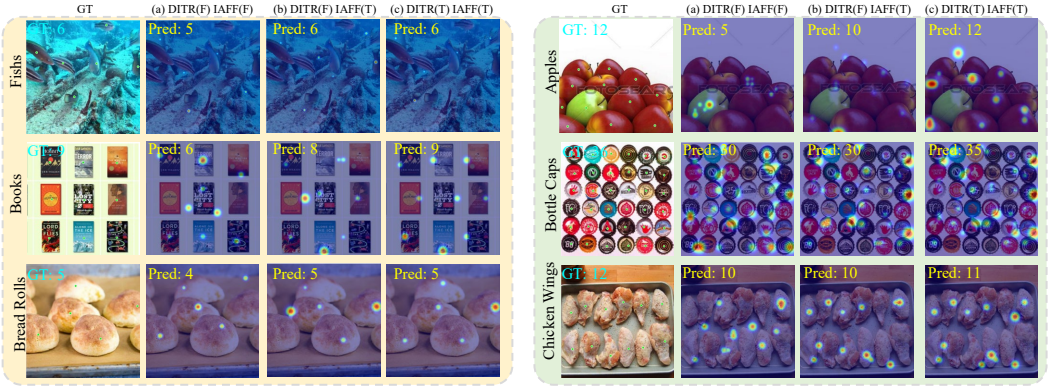



Figure 8: Qualitative visualization of the impact of DITR optimization with IAFF. Compared with the frozen DITR setting, jointly trained DITR generates more accurate density responses and better separates densely packed instances, validating the effectiveness of task-aware dense token refinement.

baseline and IAFF achieve comparable predictions, as individual objects can already be distinguished from coarse visual representations. IAFF improves hierarchical feature aggregation and strengthens instance-related semantic representation, helping maintain stable object responses. However, in dense scenarios (the last two rows), IAFF alone remains insufficient. Although enhanced feature fusion provides richer semantic information, it cannot fully resolve the spatial ambiguity caused by object overlap, resulting in missing or merged instances. These observations indicate that IAFF mainly improves semantic instance awareness but cannot overcome the spatial limitations of patch-based visual tokens.

**Effect of DITR.** Compared with the baseline, DITR consistently improves counting performance across different density levels, as demonstrated in Table 2 (a) and (c) of the original paper. The visualization results further demonstrate that DITR effectively recovers missing instances and improves counting completeness. In sparse cases, DITR reconstructs finer spatial details and preserves more object-level information, leading to accurate predictions (e.g., improving the predictions from 5 to 8 for fish). More importantly, in dense scenarios, where patch-level tokenization suffers from severe token collisions and limited spatial granularity, DITR provides clearer instance separation and reduces counting errors. For example, the baseline under-counts dense fish and bird instances (20 vs. 22 and 20 vs. 29), while DITR reconstructs more discriminative dense visual tokens and achieves predictions closer to the ground truth. These qualitative results are consistent with the quantitative improvements in Table 2 of the original paper, verifying that DITR effectively alleviates the spatial bottleneck of VLM representations by enhancing fine-grained instance perception.

**Effect of Combining IAFF and DITR.** Combining IAFF and DITR achieves the best overall performance, demonstrating their complementary roles in open-world object counting. As reported in Table 2 of the original paper, integrating IAFF with DITR consistently improves counting performance compared with using either component alone. The visualization results further explain this improvement. While DITR enhances spatial instance discrimination by reconstructing fine-grained visual tokens, IAFF provides instance-aware semantic guidance to better identify target objects during dense token enhancement. In sparse scenarios, their combination improves localization completeness and achieves more accurate counting by jointly leveraging semantic and spatial information. In dense scenarios, IAFF helps distinguish target objects and suppress irrelevant responses, while DITR preserves fine-grained spatial structures and reduces instance confusion, enabling the model to recover heavily overlapped objects. For example, the combined model achieves accurate predictions for dense fish and bird images (22 and 29 instances), whereas individual components still suffer from either semantic ambiguity or spatial information loss. These qualitative results are consistent with the quantitative improvements reported in Table 2

of the original paper, confirming that IAFF and DITR jointly bridge the semantic-instance gap for robust open-world object counting.

Overall, the visualization results are consistent with the quantitative ablation studies and demonstrate that IAFF and DITR address complementary limitations of VLM-based counting. IAFF enhances instance-aware semantic representation, while DITR improves fine-grained spatial instance perception. Their joint optimization effectively bridges the semantic-instance gap and enables robust counting across both sparse and dense object distributions.