# CONFLUXCOUNT: FORMING AND ENUMERATING SEMANTIC CONFLUENCES FOR TEXT-GUIDED OBJECT COUNTING

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

## ABSTRACT

Text-guided object counting requires converting spatially distributed query evidence into one count per object, while retaining weak responses and avoiding duplicate counts. Existing density and detection readouts respectively preserve coverage or expose instances, but neither explicitly organizes distributed evidence into object-level counting units before enumeration. We introduce **ConfluxCount**, which instead counts *semantic confluences*: object-centered structures formed by a text-grounded semantic flow over a dense probe lattice. A crowding-adaptive determinantal point process (DPP) then enumerates these confluences, suppressing redundant probes while preserving nearby instances, and produces both counts and locations from a unified probabilistic readout. On CLOC-v1.1, the jointly trained model reduces MAE from 8.39 for Count Anything to 6.03. It also reaches 4.14 MAE on KubriCount with category text and 5.75 on MixCount with positive text in supporting evaluations.

## 1 INTRODUCTION

Object counting turns images into quantitative measurements for applications ranging from crowd analysis to microscopy and agriculture. Traditional counters target fixed categories such as people, vehicles, cells, and crops Zhang et al. (2016); Li et al. (2018); Xie et al. (2018); Cao et al. (2018); Wan et al. (2022); Tian et al. (2021); Sindagi et al. (2022); Wang et al. (2021), while exemplar-based methods let users specify the target with visual examples Ranjan et al. (2021); Shi et al. (2022); You et al. (2023); Đukić et al. (2023); Liu et al. (2022); Hsieh et al. (2017). Text-guided and multimodal counters let users describe the requested targets in natural language Jiang et al. (2023); Amini-Naieni et al. (2024); Lei et al. (2026); Lu et al. (2026). Vision–language grounding connects these queries to spatial evidence in an image Radford et al. (2021); Li et al. (2022); Liu et al. (2024); Wang et al. (2026). Reliable counting must then convert the evidence into one contribution per object, including small or occluded instances that may produce weak responses.

The central difficulty is assigning this distributed evidence to instances. Parts of one extended object can activate distant locations, while separate small objects can occupy adjacent locations with similar features. Spatial distance and response magnitude alone therefore cannot establish which responses belong together. A fixed confidence threshold can discard weak objects, and a fixed suppression radius can merge distinct neighbors. Figure 1 compares how counting representations handle this ambiguity. Density readouts integrate response maps, retaining coverage without explicit object identity Lempitsky & Zisserman (2010); Zhang et al. (2016); Li et al. (2018); detection and point readouts count retained candidates, so missed or duplicate candidates change the count Song et al. (2021); Carion et al. (2020); Nguyen et al. (2022). Hybrid counters retain both dense and sparse signals but reconcile them at the final readout Lei et al. (2026).

We address this gap with *semantic confluences*: object-level counting units that organize distributed query-conditioned evidence before enumeration. **ConfluxCount** forms these units by transporting a dense probe lattice toward object-centered attractors with a text-grounded semantic flow, and then applies a crowding-adaptive determinantal point process (DPP) to suppress residual redundancy while preserving nearby instances. This representation bridges dense semantic coverage and discrete object identity, enabling both counting and localization from a unified readout.



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.

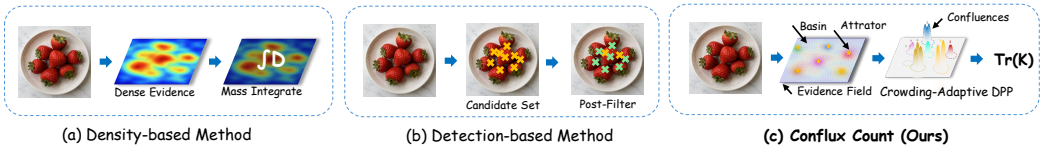



Figure 1: **Comparison of counting representations.** Density methods integrate continuous response mass, whereas detection methods count candidates retained by thresholding and NMS. ConfluxCount organizes dense probes into object-centered semantic confluences and counts distinct instances through a differentiable, crowding-adaptive DPP.

Our main contributions are summarized as follows:

- We introduce *semantic confluences* as an intermediate representation for object counting, organizing distributed query-conditioned evidence into object-centered units before enumeration.
- We propose **ConfluxCount**, which combines a text-grounded semantic flow for forming confluences with a crowding-adaptive DPP for their enumeration, producing both counts and instance locations through a unified probabilistic formulation.
- Experiments on CLOC-v1.1 show improved counting accuracy across all six domains, with supporting results on KubriCount and MixCount. Ablation studies show the benefit of forming confluences before counting and of adapting DPP repulsion to local crowding.

## 2 Related Work

**Text-guided object counting.** Fixed-category counting methods use density regression, point prediction, and detection Lempitsky & Zisserman (2010); Zhang et al. (2016); Li et al. (2018); Cao et al. (2018); Wan et al. (2022); Tian et al. (2021); Song et al. (2021); Chen et al. (2024); Liu et al. (2023); Liang et al. (2022); Lin et al. (2024); Liu et al. (2026b); Shu et al. (2024). Class-agnostic methods specify targets through visual exemplars or category-independent repetition cues Ranjan et al. (2021); Shi et al. (2022); You et al. (2023); Đukić et al. (2023); Liu et al. (2022); Hobley & Prisacariu (2023); Xu et al. (2023); Huang et al. (2024); Hsieh et al. (2017). Text-guided methods obtain target semantics from language–image representations, soft exemplars, and multimodal prompts Jiang et al. (2023); Paiss et al. (2023); Amini-Naieni et al. (2023); Kang et al. (2024); Shi et al. (2024); Pelhan et al. (2024a); Amini-Naieni et al. (2024); Pelhan et al. (2024b); Liu et al. (2025); Mondal et al. (2025); Qian et al. (2025); Zhu et al. (2024); Amini-Naieni & Zisserman (2026b;a); Lei et al. (2026), often building on vision–language and promptable foundation models Radford et al. (2021); Li et al. (2022); Liu et al. (2024); Cheng et al. (2024); Wang et al. (2025); Kirillov et al. (2023); Carion et al. (2026). Their grounding signals can be response maps, regions, or object queries. The number of these signals need not equal the number of physical objects, motivating a separate step that groups evidence before assigning cardinality.

**Counting representations and readouts.** Counting methods make different commitments about how and when visual evidence becomes discrete. Density-based approaches preserve image-wide evidence in a continuous response map and obtain the count by integration Lempitsky & Zisserman (2010); Zhang et al. (2016); Li et al. (2018); Wang et al. (2020); Wan et al. (2021; 2022), leaving object identity implicit. Detection- and point-based approaches instead represent objects explicitly, but their counts depend on candidate recall, matching, confidence thresholds, and duplicate suppression Song et al. (2021); Nguyen et al. (2022); Chen et al. (2024); Liu et al. (2023); Liang et al. (2022); Amini-Naieni et al. (2024). Recent hybrid methods retain both dense and instance-level evidence Lei et al. (2026), but still require the two forms to be reconciled when producing the final count. These readouts make different errors in crowded scenes: integration depends on calibrated response mass per object, whereas candidate counting depends on proposal recall and duplicate removal. ConfluxCount uses probe flow to organize the evidence before assigning cardinality with a DPP.

**Grouping and determinantal selection.** A complementary line of work groups local evidence around shared spatial modes, while determinantal point processes (DPPs) model diversity among jointly selected elements Kulesza & Taskar (2012); Azadi et al. (2017). In detection, DPPs have primarily been used to diversify or de-duplicate candidate predictions after those candidates have already been formed. ConfluxCount uses these ideas at a different stage of the counting process: a learned semantic flow first organizes dense probes into object-centered confluences, and a crowding-adaptive DPP then serves as the cardinality readout itself. This couples grouping and enumeration within the learned counting representation rather than treating de-duplication as post-processing.

# 3 METHOD

## 3.1 OVERVIEW

Given an image $I$ and a text query $c$, ConfluxCount predicts both the number and locations of the queried objects. As illustrated in Figure 2, the model separates counting into two stages: *confluence formation* and *confluence enumeration*. A text-grounded semantic flow guides a regular probe lattice toward object-centered attractors, causing probes supported by the same object to gather into a *semantic confluence*. A crowding-adaptive determinantal point process (DPP) then assigns cardinality to the resulting probe configuration while suppressing redundant evidence and preserving nearby

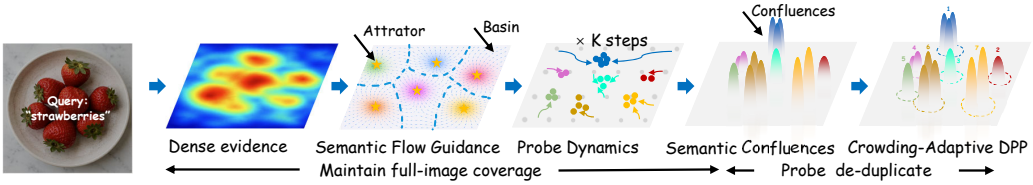
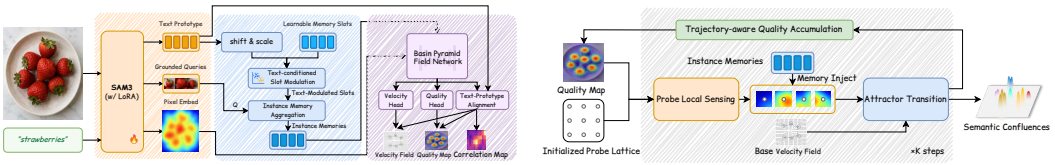



Figure 2: **Overview of ConfluxCount.** Text-grounded evidence induces a semantic flow that guides dense probes into object-centered confluences. A crowding-adaptive DPP then enumerates the confluences, yielding both object counts and locations.

Confluence formation and enumeration play complementary roles. The learned flow organizes distributed evidence into object-level structure, but finite-step dynamics may leave multiple probes around the same attractor. The DPP resolves this residual ambiguity jointly rather than treating individual probes as independent object predictions.

## 3.2 FROM TEXT-QUERY GROUNDING TO SEMANTIC FLOW




(a) Semantic conditioning and flow construction.

(b) Shared probe transitions toward confluences.

Figure 3: **Confluence formation.** Grounded queries are consolidated into instance memory, which conditions a dense velocity and quality field. Probes repeatedly sample this field and shared memory, updating their state, position, and support so that evidence from the same basin gathers around a common attractor.

Figure 3a illustrates how ConfluxCount converts text-grounded evidence into a query-conditioned semantic flow. A LoRA-adapted grounding model Hu et al. (2022) extracts $(P, \tau, Q) = \mathcal{G}(I, c)$, where $P$ contains dense pixel-level grounding features, $\tau$ represents the queried concept, and $Q$ contains grounded object queries. The dense features retain spatially distributed evidence, including weak responses from small, occluded, or low-contrast targets, while the grounded queries provide complementary object-level context.

Rather than treating grounded queries as candidate instances, we consolidate $Q$ into a fixed-length *instance memory*. Learnable slots $S_0$ are first modulated by the text prototype:

$$\widetilde{S} = S_0 \odot \big(1 + \gamma(\tau)\big) + \beta(\tau). \tag{1}$$

Cross-attention and feed-forward blocks then aggregate the grounded queries as $M = \mathrm{IMA}(\widetilde{S}, Q)$. The resulting memory summarizes query-relevant object-level evidence without imposing a one-query-per-object correspondence.

We next combine the dense features and instance memory through the Basin Pyramid Field Network, yielding a multiscale semantic field $F = \mathcal{B}(P, M)$. Fine-resolution features preserve spatial detail, intermediate features capture basin extent, and coarse features model interactions between neighboring instances. At the bottleneck, spatial tokens attend to $M$, injecting shared object-level context into the field construction.

From $F$, the network predicts a base velocity $v_0(x)$ and base quality $q_0(x)$ at each spatial location. The velocity controls how local evidence moves through the field, while the quality measures how strongly the location is supported by the text query. To make both query-selective, we compute local text alignment in a shared semantic space:

$$\kappa(x) = \langle \mathrm{norm}(W_f F(x)), \mathrm{norm}(W_\tau \tau) \rangle . \tag{2}$$

The alignment then modulates the two predictions as

$$v(x) = v_0(x) + C_v(\kappa(x)), \tag{3}$$

with the quality calibrated analogously as $q(x) = q_0(x) + b_q + \alpha_q \kappa(x)$. The resulting semantic flow therefore combines dense spatial evidence, object-level context, and text alignment to determine both where evidence should move and how strongly each trajectory is supported.

### 3.3 Probe Dynamics toward Semantic Confluences

As illustrated in Figure 3b, we convert the continuous semantic flow into object-level structure using a regular lattice of $N_\mathrm{p}$ probes. A probe is a tracer rather than an object query: it carries local evidence through the learned flow, and object identity emerges from the collective behavior of multiple probes.

Probe $i$ is initialized at position $p_i^0$ with semantic state $z_i^0 = \mathrm{Sample}(F, p_i^0)$ and support $s_i^0 = \mathrm{Sample}(q, p_i^0)$. We evolve all probes for $K$ parameter-shared transitions. At each step, a probe re-samples the semantic field at its current position, combines this local observation with the shared instance memory $M$ through cross-attention, and updates its semantic state. This repeated re-sampling couples appearance and motion: as a probe moves, its subsequent trajectory is conditioned on the evidence encountered at the new location.

The probe position is updated by combining the shared velocity field with a probe-specific residual:

$$p_i^{k+1} = \mathrm{clip}\big[p_i^k + \eta \tanh\big(v(p_i^k) + W_{\Delta p} z_i^{k+1}\big)\big]. \tag{4}$$

The shared velocity captures the object-level attraction geometry, while the residual adapts the motion to each probe's accumulated semantic evidence. In parallel, support is accumulated along the trajectory as $s_i^{k+1} = s_i^k + W_s z_i^{k+1}$, allowing the final score to reflect evidence gathered over the full path rather than at a single location.

After $K$ transitions, each probe is represented by $(p_i, s_i, a_i, g_i)$, denoting its converged position, accumulated support, appearance embedding, and local crowding descriptor, respectively. Probes supported by the same object tend to converge toward a common attractor, whereas probes associated with neighboring objects remain separated. We refer to these object-centered groups as *semantic confluences*.

To measure the many-to-one structure, we group final probes by spatial and semantic compatibility. Two probes are connected when their distance is below $\varepsilon_p$ and their appearance cosine similarity exceeds $\varepsilon_a$; the resulting connected components are denoted by $\mathcal{C} = \{C_m\}_{m=1}^{M}$, with $\chi(i)$ indicating the confluence containing probe $i$. These components define the confluences used in diagnostic analyses. During training and inference, the crowding-adaptive DPP assigns cardinality directly to the probe configuration.

### 3.4 CROWDING-ADAPTIVE DETERMINANTAL COUNTING

Semantic flow organizes distributed evidence into compact confluence clouds, but finite-step dynamics can leave multiple probes around the same object. Simply suppressing nearby probes is not sufficient: close probes may be redundant evidence for one object or valid evidence for several neighboring objects. We therefore use a crowding-adaptive determinantal point process (DPP) to enumerate the converged probe configuration jointly.

For each probe, the accumulated support is converted to positive evidence $e_i = \exp(s_i)$, while the crowding descriptor $g_i$ determines a gate $\pi_i = \epsilon + (1-\epsilon)\sigma(g_i)$. This gate divides the evidence between a *private* component, which can contribute independently, and a *shared* component, which competes with similar probes. Using the normalized appearance embedding $\bar{a}_i$ and spatial encoding $\phi(p_i)$, we define

$$U_i = \sqrt{e_i(1-\pi_i)}\big(\bar{a}_i \otimes \phi(p_i)\big), \qquad L = \mathrm{diag}(e \odot \pi) + UU^\top. \tag{5}$$

The diagonal term protects valid instances in crowded regions, whereas the shared low-rank term induces competition between probes that are similar in both appearance and position. The learned gate therefore adapts the amount of suppression to the local crowding pattern instead of applying a fixed rule everywhere. When several probes describe one object, their shared factors compete and reduce the sum of their marginal contributions. In crowded regions, stronger private evidence allows nearby probes to contribute separately when they support distinct objects.

Let $K = L(I+L)^{-1}$ denote the DPP marginal kernel. Its expected cardinality gives the predicted count:

$$\widehat{N} = \mathrm{tr}(K) = \sum\nolimits_i K_{ii}. \tag{6}$$

The diagonal marginal $K_{ii}$ measures how much probe $i$ contributes to the expected selected set. For analysis, the effective cardinality of a diagnostic confluence $C_m$ is simply $\kappa_m = \sum_{i \in C_m} K_{ii}$; ideally, an object confluence contributes approximately one unit, while background confluences contribute little.

This formulation makes enumeration part of the learned representation itself. Rather than thresholding probes or applying a fixed post-hoc suppression rule, ConfluxCount learns which nearby pieces of evidence should compete and which should coexist, while producing a differentiable count directly from the same DPP kernel.

Table 1: **Benchmark results on CLOC-v1.1.** We report full-test MAE, RMSE, and NAE, followed by per-domain MAE. Ins.Grd. indicates whether a method produces instance-grounded outputs. Best and second-best results are bold and underlined.

| Method | Venue | Ins.Grd. | Full Test | | | Single Domain MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MAE | RMSE | NAE | GS | RS | Histo. | Cell. | Agri. | Micro. |
| LOCA Đukić et al. (2023) | ICCV'23 | × | 68.09 | 179.34 | 14.70 | 53.45 | 81.41 | 65.86 | 124.36 | 473.81 | 39.47 |
| CLIP-Count Jiang et al. (2023) | ACMMM'23 | × | 30.97 | 119.65 | 4.48 | 23.28 | 25.82 | 57.98 | 121.17 | 298.45 | 22.40 |
| CounTX Amini-Naieni et al. (2023) | BMVC'23 | × | 31.79 | 123.17 | 4.74 | 22.79 | 30.07 | 64.64 | 112.22 | 301.10 | 27.41 |
| VLCounter Kang et al. (2024) | AAAI'24 | × | 30.24 | 123.09 | 3.06 | 22.85 | 22.46 | 55.14 | 125.83 | 306.38 | 37.72 |
| VA-Count Zhu et al. (2024) | ECCV'24 | × | 33.81 | 130.73 | 4.59 | 21.84 | 36.42 | 72.50 | 163.73 | 315.23 | 20.47 |
| CountGD Amini-Naieni et al. (2024) | NeurIPS'24 | ✓ | 56.62 | 1099.56 | 5.17 | 54.75 | 42.25 | 99.69 | 114.50 | 244.80 | 11.09 |
| YOLO-Count Zeng et al. (2025) | ICCV'25 | × | 33.28 | 133.66 | 1.08 | 23.95 | 27.70 | 72.50 | 168.37 | 309.30 | 24.33 |
| T2ICount Qian et al. (2025) | CVPR'25 | × | 49.70 | 156.97 | 8.83 | 32.85 | 62.48 | 73.06 | 151.30 | 296.37 | 203.51 |
| CountSE Liu et al. (2025) | ICCV'25 | × | 27.14 | 130.03 | 2.16 | 19.40 | 19.08 | 70.27 | 120.08 | 302.19 | 22.22 |
| CountGD++ Amini-Naieni & Zisserman (2026a) | CVPR'26 | ✓ | 22.01 | 122.75 | 0.95 | 12.08 | 10.69 | 55.83 | 222.25 | 321.99 | 43.80 |
| GLIP-L Li et al. (2022) | CVPR'22 | ✓ | 37.58 | 141.95 | 0.81 | 31.05 | 19.57 | 68.78 | 225.42 | 320.85 | 43.71 |
| YOLO-World-X Cheng et al. (2024) | CVPR'24 | ✓ | 35.06 | 140.19 | 0.76 | 28.02 | 17.81 | 67.93 | 226.07 | 321.13 | 42.49 |
| GrdDINO-SwinB Liu et al. (2024) | ECCV'24 | ✓ | 35.19 | 141.32 | 0.78 | 29.41 | 14.14 | 69.11 | 224.96 | 320.17 | 43.75 |
| YOLOE-v8L Wang et al. (2025) | ICCV'25 | ✓ | 40.21 | 142.47 | 0.99 | 33.94 | 21.78 | 69.99 | 228.28 | 321.97 | 44.80 |
| YOLOE-11L Wang et al. (2025) | ICCV'25 | ✓ | 40.24 | 142.46 | 1.00 | 33.98 | 21.79 | 69.95 | 228.24 | 321.88 | 44.81 |
| SAM3 Carion et al. (2026) | ICLR'26 | ✓ | 25.65 | 132.07 | 0.85 | 18.85 | 10.88 | 64.32 | 152.45 | 315.88 | 37.04 |
| LocateAnything-3B Wang et al. (2026) | ECCV'26 | ✓ | 29.82 | 137.82 | <u>0.64</u> | 25.03 | 11.87 | 62.55 | 173.45 | 296.75 | 16.96 |
| Count Anything Lei et al. (2026) | – | ✓ | <u>8.39</u> | <u>32.00</u> | 0.65 | <u>7.36</u> | <u>6.19</u> | <u>16.64</u> | <u>38.65</u> | <u>41.78</u> | <u>4.30</u> |
| **ConfluxCount (Ours)** | – | ✓ | **6.03** | **22.00** | **0.46** | **5.60** | **4.01** | **14.41** | **20.69** | **24.79** | **3.96** |

*Note:* Ins.Grd. denotes instance grounding capability. GS, RS, Histo., Cell., Agri., and Micro. denote General Scene, Remote Sensing, Histopathology, Cellular Microscopy, Agriculture, and Microbiology, respectively.

### 3.5 BASIN-CONSTRAINED LEARNING AND INFERENCE

ConfluxCount is trained from point annotations $\mathcal{X} = \{x_j\}_{j=1}^{N}$. Since a point specifies an object center but not the spatial extent of its attraction basin, we construct adaptive supervision from local object spacing. For annotation $x_j$, let $d_j$ denote its nearest-neighbor distance and define the basin radius as $r_j = \mathrm{clip}(\zeta d_j, r_{\min}, r_{\max})$. Each initial probe is assigned to its nearest annotation and treated as foreground only if it lies within the corresponding basin. This produces broader support around isolated objects and tighter basins in crowded regions.

We supervise confluence formation with complementary basin and flow objectives. The basin loss encourages probes originating inside target-supported regions to accumulate positive support while suppressing background probes, while the flow loss guides positive probes toward their assigned object centers. We additionally derive a soft crowding target from nearest-neighbor distance to supervise the crowding gate used by the DPP. Point supervision, however, specifies where objects occur but does not by itself enforce which text query they correspond to. We therefore add an in-batch prompt-contrastive objective on the grounded visual features. Let $\bar{f}_i(x)$ denote the $\ell_2$-normalized visual feature of image $i$ at location $x$, and $\bar{t}_k$ the normalized text prototype of prompt $k$. For each annotated object $x_j \in \mathcal{X}_i$, we optimize

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{\sum_i |\mathcal{X}_i|} \sum_i \sum_{x_j \in \mathcal{X}_i} \log \frac{\exp\big(\langle \bar{f}_i(x_j), \bar{t}_i \rangle / \tau\big)}{\sum_{k \in \mathcal{B} \setminus \mathcal{S}_i} \exp\big(\langle \bar{f}_i(x_j), \bar{t}_k \rangle / \tau\big)}, \tag{7}$$

where $\tau$ is a temperature and $\mathcal{S}_i = \{k \neq i : y_k = y_i\}$ masks other batch entries describing the same category, while retaining the matched prompt in the denominator. The remaining prompts provide semantic negatives, including fine-grained variants when present, encouraging query-specific grounding without separately sampled absent-category examples.

Counting further requires learning which probe explanations can coexist. Within each point's Voronoi cell, we select the nearest initial probe as its canonical representative. These representatives form $\mathcal{S}^\star$ for the DPP likelihood; Appendix A.3 specifies ties and empty cells:

$$\mathcal{L}_{\mathrm{dpp}} = \frac{\log\det(I + L) - \log\det L_{\mathcal{S}^\star}}{\sqrt{\max(1, |\mathcal{S}^\star|)}}. \tag{8}$$

A coverage term encourages the canonical probes to receive sufficient marginal probability, while a count loss calibrates the expected DPP cardinality $\mathrm{tr}(K)$ to the full ground-truth count. The complete objective is

$$\mathcal{L} = \sum\nolimits_{r \in \mathcal{R}} \lambda_r \mathcal{L}_r, \qquad \mathcal{R} = \{\mathrm{dpp}, \mathrm{basin}, \mathrm{flow}, \mathrm{crowd}, \mathrm{cov}, \mathrm{count}, \mathrm{con}\}. \tag{9}$$

The supervision therefore follows the structure of the model: basin and flow objectives form object-centered confluences, prompt contrast sharpens their semantic specificity, and set-level objectives determine their cardinality.

At inference, we report $\hat{N} = \mathrm{round}(\mathrm{tr}(K))$ and greedily select $\hat{N}$ locations from the same DPP kernel. Starting from an empty set, each step selects the probe with the largest conditional gain,

$$i^\star = \arg\max_{i \notin \mathcal{S}} \left[ L_{ii} - L_{\mathcal{S},i}^\top L_{\mathcal{S}}^{-1} L_{\mathcal{S},i} \right], \tag{10}$$

until $\hat{N}$ probes have been selected. This approximates the MAP set of the $\hat{N}$-conditioned DPP. Determinantal interactions reduce redundant selections, yielding counts and locations from one kernel without an additional detector or distance threshold.

## 4 EXPERIMENTS

We evaluate ConfluxCount on CLOC-v1.1, KubriCount, and MixCount. Unless otherwise stated, benchmark results use the model jointly trained on all three datasets. KubriCount inference uses only each annotation's category string as a text query, without exemplar boxes or negative prompts. Controlled ablations use a separate CLOC-only model matched to the Count Anything training setting Carion et al. (2026); Lei et al. (2026), and are compared against the corresponding full-model

baseline. ConfluxCount builds on a pretrained SAM3 image–text backbone with LoRA adaptation and evolves a fixed probe lattice for three shared transitions before the DPP readout. Appendix A gives the evaluation metrics, benchmark prompt interfaces, and training details for the CLOC-only ablation setting.

### 4.1 Comparison with Prior Work

Table 2: **Benchmark results on KubriCount.** We report MAE and RMSE overall and by difficulty. ConfluxCount uses category text only; published baselines use their native prompt interfaces Liu et al. (2026a). Best and second-best results are bold and underlined.

| Method | Overall | | L1 | | L2 | | L3 | | L4 | | L5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE | MAE | RMSE |
| ***Multimodal Large Language Models*** | | | | | | | | | | | | |
| Qwen3-VL-235B-A22B Bai et al. (2025) | 9.95 | 18.47 | 9.71 | 20.92 | 8.80 | 16.70 | 9.60 | 18.25 | 10.69 | 18.10 | 11.06 | 18.24 |
| InternVL3-78B Zhu et al. (2025) | 8.82 | 15.67 | 8.42 | 15.91 | 7.24 | 13.57 | 7.67 | 13.87 | 10.94 | 18.47 | 9.96 | 16.15 |
| LocateAnything-3B Wang et al. (2026) | 14.05 | 24.66 | 10.97 | 25.87 | 18.36 | 27.87 | 11.23 | 21.36 | 17.05 | 26.22 | 12.63 | 21.22 |
| Gemini-3-Flash Google DeepMind (2026) | 5.41 | 26.35 | 5.99 | 48.91 | 4.59 | 18.17 | 4.23 | 17.82 | **6.32** | <u>14.01</u> | 5.97 | 15.16 |
| Gemini-3-Pro Google DeepMind (2026) | 5.49 | 27.05 | 6.10 | 51.13 | 4.32 | 11.30 | 4.37 | 19.76 | <u>6.49</u> | 16.81 | 6.23 | 16.17 |
| GPT-5.2 OpenAI (2025) | 9.44 | 25.99 | 11.99 | 21.12 | 6.62 | 11.90 | 8.64 | 48.36 | 9.73 | 16.57 | 10.46 | 16.40 |
| Claude-4.5-Opus Anthropic (2025) | 8.87 | 16.53 | 10.67 | 21.25 | 7.81 | 14.20 | 8.17 | 15.53 | 8.83 | 15.66 | 8.93 | 15.17 |
| ***Counting Expert Models*** | | | | | | | | | | | | |
| FamNet Ranjan et al. (2021) | 21.17 | 37.49 | 21.55 | 30.72 | 26.85 | 50.57 | 20.09 | 35.49 | 16.98 | 27.83 | 19.77 | 37.21 |
| LOCA Đukić et al. (2023) | 16.69 | 29.07 | 9.80 | 15.60 | 24.15 | 39.30 | 17.31 | 32.93 | 16.78 | 26.87 | 14.74 | 23.99 |
| DAVE Pelhan et al. (2024a) | 18.04 | 31.51 | 12.18 | 19.68 | 22.22 | 39.34 | 19.71 | 39.18 | 17.95 | 26.94 | 17.86 | 27.20 |
| GeCo Pelhan et al. (2024b) | 10.82 | 18.63 | 8.23 | 14.98 | 13.33 | 22.97 | 10.19 | 17.89 | 12.49 | 19.97 | 9.58 | 15.68 |
| CounTR Liu et al. (2022) | 12.82 | 21.96 | 8.86 | 15.77 | 15.42 | 27.86 | 13.50 | 22.32 | 14.07 | 22.13 | 12.07 | 19.30 |
| CountGD Amini-Naieni et al. (2024) | 18.18 | 38.49 | 7.67 | 43.86 | 24.51 | 47.76 | 18.89 | 36.11 | 22.49 | 33.09 | 16.80 | 26.52 |
| CountGD++ Amini-Naieni & Zisserman (2026a) | 7.76 | 28.17 | 7.65 | 45.06 | 8.74 | 26.05 | 6.66 | 21.97 | 9.01 | 24.02 | 6.55 | 13.09 |
| Rex-Omni Jiang et al. (2026) | 14.66 | 42.78 | 6.89 | 37.86 | 23.34 | 51.88 | 12.01 | 42.24 | 18.29 | 42.98 | 11.87 | 36.10 |
| Count Anything Lei et al. (2026) | 11.01 | 19.84 | 4.59 | 13.70 | 14.32 | 23.98 | 8.40 | 14.72 | 18.75 | 28.00 | 9.01 | 14.27 |
| HieraCount Liu et al. (2026a) | <u>4.67</u> | <u>11.07</u> | **3.06** | <u>10.58</u> | <u>3.10</u> | <u>7.66</u> | <u>3.90</u> | <u>8.29</u> | 8.37 | 17.14 | <u>5.04</u> | <u>9.08</u> |
| **ConfluxCount (text only)** | **4.14** | **7.79** | <u>3.88</u> | **8.14** | **3.04** | **5.52** | **3.11** | **5.96** | 7.20 | **11.58** | **3.44** | **6.08** |

Table 3: **Benchmark results on MixCount.** The CountGD++ baseline was additionally trained on MixCount and uses positive text only Amini-Naieni & Zisserman (2026a); Dumery et al. (2026). Ins.Grd. denotes instance-grounded outputs. Best and second-best among listed rows are bold and underlined.

| Method | Ins.Grd. | Absolute Errors | | Relative Metrics | |
|---|---|---|---|---|---|
| | | MAE↓ | RMSE↓ | NAE↓ | Acc10%↑ |
| CountSE Liu et al. (2025) | × | 32.49 | 41.49 | 7.23 | 0.06 |
| CountGD Amini-Naieni et al. (2024) | ✓ | 33.00 | 42.25 | 5.64 | 0.05 |
| CountGD++ (trained on MixCount) | ✓ | <u>14.63</u> | <u>22.14</u> | 1.02 | 0.09 |
| LocateAnything-3B Wang et al. (2026) | ✓ | 19.20 | 25.53 | <u>0.90</u> | 0.11 |
| Count Anything Lei et al. (2026) | ✓ | 20.17 | 28.53 | 1.29 | <u>0.19</u> |
| **ConfluxCount (Ours)** | ✓ | **5.75** | **8.28** | **0.38** | **0.27** |

Tables 1–3 report results after joint training on all three datasets. On the primary CLOC-v1.1 benchmark, ConfluxCount reduces MAE from 8.39 for Count Anything to 6.03. On KubriCount, it obtains 4.14 MAE with category text alone versus HieraCount's 4.67 with text and exemplars Liu et al. (2026a); on MixCount, it obtains 5.75 MAE versus 14.63 for CountGD++ after additional MixCount training with positive text only Dumery et al. (2026). ConfluxCount leads all CLOC-v1.1 domains, all four listed MixCount metrics, and KubriCount RMSE at every level, though not KubriCount MAE at L1 or L4. The CLOC gain spans six visual domains. The MixCount comparison holds the positive-text interface fixed, while KubriCount baselines retain their published prompt resources, including exemplars (Appendix A.1). These are in-benchmark comparisons rather than transfer to an unseen dataset.

### 4.2 Do Semantic Confluences Emerge?

Strong counting accuracy alone does not establish that the model learns the intended object-level representation: the global count could be correct even if probe evidence were distributed incorrectly

across objects. We therefore examine whether the learned probe configuration exhibits the defining properties of semantic confluences.

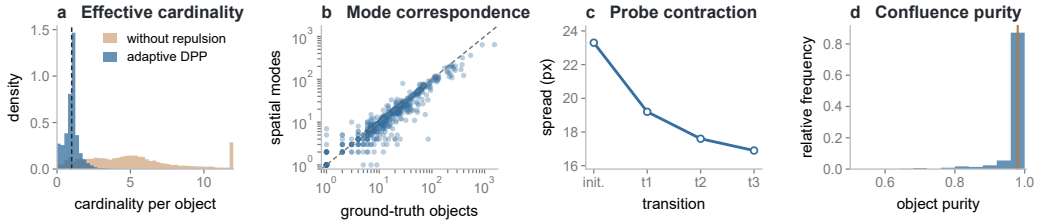



Figure 4: **Semantic confluences emerge as object-level counting units. (a)** DPP interaction concentrates effective cardinality near one per object. **(b)** The number of spatial modes closely tracks the ground-truth object count. **(c)** Probe spread decreases over successive transitions. **(d)** The resulting confluences are highly object-pure.

Figure 4 shows that the final representation is organized object by object rather than merely calibrated in aggregate. Each ground-truth basin initially contains a median of 36 probes, so individual probes cannot themselves serve as object predictions. Without determinantal repulsion, their inclusion mass accumulates to 5.5 units per object; the full DPP reduces this to a mean and median of 1.03, indicating that redundant evidence is consolidated into approximately one counting unit. This cardinality behavior is matched spatially: the inclusion field contains 0.94 modes per object with a mean object purity of 0.98, while within-object probe spread contracts by 28% over the three transitions. Together, these measurements show that probes converge into distinct, object-specific confluences and that the DPP assigns them approximately unit cardinality.

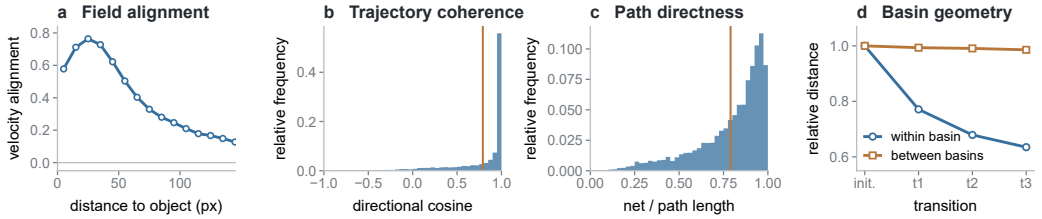



Figure 5: **Semantic flow selectively contracts object basins. (a–c)** Probe motion is object-directed, coherent, and direct. **(d)** Within-basin spread contracts while separation between neighboring basins is preserved.

Figure 5 measures how the probes move within and across object basins. Near each basin, velocity aligns most strongly with the assigned object; trajectories have a mean directional cosine and net-to-path ratio of 0.79. Within-basin spread falls to 0.63 of its initial value, while between-basin separation remains at 0.99. The probes contract toward their respective objects while neighboring basins remain separated.

### 4.3 Ablation Studies

All ablations use the CLOC-only setting matched to Count Anything Lei et al. (2026). We examine the counting readout, DPP components, training objectives, and point selection from the same kernel. Appendix B gives further results on hyperparameters, probe resolution, count-range behavior, robustness, and failure cases; Appendix C provides qualitative examples.

**How should confluence probes be counted?** Table 4 holds converged probes fixed, isolating enumeration from probe formation. The large errors from thresholding and spatial clustering indicate that probe strength and proximity alone do not resolve object identity. Density integration stays competitive, showing that useful counting evidence remains distributed across probes. The DPP reduces MAE from 6.55 for density integration to 6.12 by allocating inclusion mass jointly, accounting for redundancy among probes. Learned-NMS has lower NAE, while the DPP leads MAE and RMSE, placing its advantage in absolute count accuracy.

**Why is crowding-adaptive repulsion needed?** Without repulsion, probes contribute independently (Table 6); the error increase agrees with the excess inclusion mass per object in Figure 4.

Table 4: **Confluence readout ablation.** All methods use the same converged probes and differ only in the counting readout.

| Family | Readout | MAE↓ | RMSE↓ | NAE↓ |
|---|---|---|---|---|
| *Direct* | Density ($\int \lambda$) | 6.55 | 29.89 | 0.324 |
| | Threshold | 16.42 | 60.16 | 1.028 |
| *Rule* | NMS | 16.29 | 55.59 | 0.366 |
| | DBSCAN | 19.34 | 64.03 | 0.304 |
| | Cluster-count | 16.20 | 54.00 | 0.389 |
| | Mean-shift | 16.56 | 56.67 | 0.344 |
| *Learned* | Learned-NMS | 7.40 | 32.26 | **0.282** |
| | Diff-Cluster | 7.68 | 33.98 | 0.292 |
| *DPP* | **Conflux** | **6.12** | **27.23** | 0.350 |

Table 5: **Training objective ablation.** Removed losses are grouped by their roles in confluence formation, enumeration, and evidence allocation.

| Role | Removed | MAE↓ | RMSE↓ | NAE↓ |
|---|---|---|---|---|
| Set-level enum. | $\mathcal{L}_{\text{dpp}} + \mathcal{L}_{\text{count}}$ | 389.70 | 643.37 | 30.953 |
| | $\mathcal{L}_{\text{dpp}}$ | 7.34 | 28.65 | 0.456 |
| | $\mathcal{L}_{\text{count}}$ | 6.80 | 29.22 | 0.552 |
| Confluence formation | $\mathcal{L}_{\text{basin}} + \mathcal{L}_{\text{flow}}$ | 6.55 | 29.30 | 0.350 |
| | $\mathcal{L}_{\text{basin}}$ | 6.39 | 28.23 | 0.361 |
| | $\mathcal{L}_{\text{flow}}$ | 6.69 | 30.57 | 0.388 |
| Evidence allocation | $\mathcal{L}_{\text{crowd}} + \mathcal{L}_{\text{cov}}$ | 6.35 | 28.49 | 0.353 |
| | $\mathcal{L}_{\text{crowd}}$ | 6.58 | 28.85 | 0.394 |
| | $\mathcal{L}_{\text{cov}}$ | 6.40 | 28.14 | 0.361 |
| **Full** | ∅ | **6.12** | **27.23** | **0.350** |

Table 6: **Crowding-adaptive DPP ablation.** We vary similarity cues, repulsion, and the evidence gate.

| Similarity | | Readout | | Error↓ | | |
|---|---|---|---|---|---|---|
| App. | Pos. | Rep. | Gate | MAE | RMSE | NAE |
| ✓ | × | ✓ | Adapt. | 6.64 | 31.63 | 0.373 |
| × | ✓ | ✓ | Adapt. | 6.56 | 30.85 | 0.345 |
| N/A | N/A | × | $\pi$=1 | 18.59 | 121.98 | 0.760 |
| N/A | N/A | ✓ | $\pi$=0 | 6.89 | 27.99 | **0.314** |
| ✓ | ✓ | ✓ | Fixed | 8.74 | 31.42 | 0.410 |
| ✓ | ✓ | ✓ | Adapt. | **6.12** | **27.23** | 0.350 |

Table 7: **Localization readout comparison.** Both variants use the same DPP model and count; only point-set inference differs.

| Metric | Marginal NMS | Greedy fixed-$k$ DPP |
|---|---|---|
| F1↑ | **0.756** | 0.738 |
| LocErr(px)↓ | 5.5 | 5.5 |
| Recall@8↑ | **0.640** | 0.629 |
| GAME1↓ | **6.7** | 7.0 |
| LQ↑ | 0.699 | **0.703** |

With similarity cues fixed, adapting the gate reduces MAE from 8.74 to 6.12, supporting a local balance between suppressing duplicates and preserving distinct instances. The private diagonal lets nearby valid probes retain inclusion mass even when their shared appearance and position factors would otherwise make them compete. Appearance limits competition between unlike neighbors, while position limits competition between distant objects with similar features. Their combination improves absolute errors over either cue alone. Pure shared evidence gives lower NAE, but adaptive private evidence improves MAE and RMSE, revealing a trade-off in how strongly nearby probes compete.

**How do the training objectives shape the representation?** Removing both set-level losses in Table 5 causes the largest deterioration despite retaining geometric supervision. The likelihood constrains which canonical probes coexist, while count calibration constrains aggregate cardinality. Coverage can raise a canonical probe's marginal while leaving redundant probes active, explaining why local targets do not suffice to calibrate the aggregate. Removing flow is more harmful than removing basin supervision, consistent with its direct role in shaping trajectories. Crowding supervises the evidence split. Joint removal of either the formation or allocation pair can be less harmful than a single removal, so individual loss effects cannot be summed into independent contributions.

**Can the same kernel recover instance locations?** Table 7 fixes the model and predicted count, isolating which probes become output points. Greedy fixed-$k$ DPP matches NMS in localization error and slightly improves LQ, but gives lower F1 and recall and higher GAME1. Thus similar localization error can coexist with differences in instance recovery and spatial count allocation. Selection by conditional kernel gain avoids a separate suppression radius, while marginal-ranked NMS recovers more instances in this comparison.

## 5 Conclusion

ConfluxCount makes object-level organization an explicit step in text-guided counting. Semantic flow forms confluences from distributed query evidence, and a crowding-adaptive DPP controls which probe responses contribute together. The same kernel supports counting and localization. On CLOC-v1.1, the jointly trained model reduces MAE from 8.39 for Count Anything to 6.03, with KubriCount and MixCount providing supporting evaluations. CLOC-only analyses show how spatial contraction and redundancy control contribute to enumeration. These results support organizing dense evidence into counting units whose cardinality is learned jointly with their spatial structure.

### Reproducibility Statement

Section 3 specifies the semantic-flow model, probe dynamics, training objectives, and counting and localization procedures. Section 4 distinguishes the jointly trained benchmark model from the CLOC-only model used for controlled ablations. Appendix A defines the evaluation metrics and documents the CLOC-only data preprocessing, architecture, hyperparameters, and optimization schedule. Appendix B provides the additional sensitivity, robustness, count-range, and failure analyses.

### AI Use Statement

Generative AI tools were used for language editing and manuscript organization. The authors reviewed all AI-assisted changes and take responsibility for the final text, claims, and artifacts.

## References


Niki Amini-Naieni and Andrew Zisserman. CountGD++: Generalized prompting for open-world counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37725–37734, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Amini-Naieni_CountGD_Generalized_Prompting_for_Open-World_Counting_CVPR_2026_paper.html.

Niki Amini-Naieni and Andrew Zisserman. Open-World Object Counting in Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):2300–2308, 2026b. doi: 10.1609/aaai.v40i4.37214.

Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. In *British Machine Vision Conference*, 2023.

Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. CountGD: Multi-modal open-world counting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 48810–48837, 2024. doi: 10.52202/079017-1547.

Anthropic. Claude Opus 4.5 system card. System card, 2025. URL https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf.

Samaneh Azadi, Jiashi Feng, and Trevor Darrell. Learning detection with diverse proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7149–7157, 2017. doi: 10.1109/cvpr.2017.779.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *Eur. Conf. Comput. Vis.*, pp. 757–773, 2018. doi: 10.1007/978-3-030-01228-1_45.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020. doi: 10.1007/978-3-030-58452-8_13.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *Int. Conf. Learn. Represent.*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Ming-Hsuan Yang, and Sy-Yen Kuo. Improving Point-Based Crowd Counting and Localization Based on Auxiliary Point Guidance. In *Eur. Conf. Comput. Vis.*, pp. 428–444, 2024. doi: 10.1007/978-3-031-72691-0_24.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16901–16911, 2024. doi: 10.1109/cvpr52733.2024.01599.

Corentin Dumery, Niki Amini-Naieni, Shervin Naini, and Pascal Fua. The MixCount dataset: Bridging the data gap for open-vocabulary object counting. *arXiv preprint arXiv:2605.18063*, 2026. URL https://arxiv.org/abs/2605.18063.

Google DeepMind. Gemini 3 Pro model card. Model card, 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Michael Hobley and Victor Prisacariu. Learning To Count Anything: Reference-less class-agnostic counting with weak supervision. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2023.

Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Int. Conf. Comput. Vis.*, pp. 4145–4153, 2017. doi: 10.1109/iccv.2017.446.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2022.

Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, and Hongming Shan. Point, Segment and Count: A generalized framework for object counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17067–17076, 2024. doi: 10.1109/CVPR52733.2024.01615.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25472–25483, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jiang_Detect_Anything_via_Next_Point_Prediction_CVPR_2026_paper.html.

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. doi: 10.1145/3581783.3611789.

Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. VLCounter: Text-aware visual representation for zero-shot object counting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2714–2722, 2024. doi: 10.1609/aaai.v38i3.28050.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. doi: 10.1109/iccv51070.2023.00371.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044.

Mengqi Lei, Shuokun Cheng, Wei Bao, Shaoyi Du, Jun-Hai Yong, Siqi Li, and Yue Gao. Count anything. *arXiv preprint arXiv:2605.30846*, 2026. URL https://arxiv.org/abs/2605.30846.

Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, volume 23, pp. 1324–1332, 2010.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded Language-Image Pre-Training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10965–10975, 2022. doi: 10.1109/cvpr52688.2022.01069.

Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1091–1100, 2018. doi: 10.1109/cvpr.2018.00120.

Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *Eur. Conf. Comput. Vis.*, pp. 38–54, 2022. doi: 10.1007/978-3-031-19769-7_3.

Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. GramFormer: Learning crowd counting via graph-modulated transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3395–3403, 2024. doi: 10.1609/aaai.v38i4.28126.

Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. CounTR: Transformer-based generalised visual counting. In *British Machine Vision Conference*, 2022.

Chang Liu, Haoning Wu, and Weidi Xie. Count anything at any granularity. *arXiv preprint arXiv:2605.10887*, 2026a. URL https://arxiv.org/abs/2605.10887.

Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. Point-Query Quadtree for Crowd Counting, Localization, and More. In *Int. Conf. Comput. Vis.*, pp. 1676–1685, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3.

Shuai Liu, Peng Zhang, Shiwei Zhang, and Wei Ke. CountSE: Soft exemplar open-set object counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21536–21546, 2025. doi: 10.1109/ICCV51701.2025.02000.

Xinyan Liu, Guorong Li, Yuankai Qi, Zhenjun Han, Anton van den Hengel, Nicu Sebe, Ming-Hsuan Yang, and Qingming Huang. Consistency-aware anchor pyramid network for crowd localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(9):10192–10206, 2026b. doi: 10.1109/TPAMI.2024.3392013.

Lidong Lu, Guo Chen, Zhu Wei, Zhiqi Li, Yicheng Liu, and Tong Lu. AV-Reasoner: Improving and benchmarking clue-grounded audio-visual counting for MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33477–33487, 2026.

Anindya Mondal, Sauradip Nag, Xiatian Zhu, and Anjan Dutta. OmniCount: Multi-label object counting with semantic-geometric priors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19537–19545, 2025. doi: 10.1609/aaai.v39i18.34151.

Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *European Conference on Computer Vision*, pp. 348–365. Springer, 2022. doi: 10.1007/978-3-031-20044-1_20.

OpenAI. Update to the GPT-5 system card: GPT-5.2. System card, 2025. URL https://openai.com/index/gpt-5-system-card-update-gpt-5-2/.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023. doi: 10.1109/iccv51070.2023.00294.

Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. DAVE: A detect-and-verify paradigm for low-shot counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23293–23302, 2024a. doi: 10.1109/cvpr52733.2024.02198.

Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66260–66282, 2024b. doi: 10.52202/079017-2118.

Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau, Xiaopeng Hong, and Michael P. Pound. T2ICount: Enhancing cross-modal understanding for zero-shot counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 25336–25345, 2025. doi: 10.1109/cvpr52734.2025.02359.

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 *International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3394–3403, 2021. doi: 10.1109/cvpr46437.2021.00340.

Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9529–9538, 2022.

Zenglin Shi, Ying Sun, and Mengmi Zhang. Training-free object counting with prompts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 323–331, 2024. doi: 10.1109/wacv57701.2024.00039.

Weibo Shu, Jia Wan, and Antoni B Chan. Generalized characteristic function loss for crowd analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5):2882–2899, 2024. doi: 10.1109/TPAMI.2023.3336196.

Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2594–2609, 2022. doi: 10.1109/TPAMI.2020.3035969.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3365–3374, 2021. doi: 10.1109/iccv48922.2021.00335.

Ye Tian, Xiangxiang Chu, and Hongpeng Wang. CCTrans: Simplifying and improving crowd counting with transformer. *arXiv preprint arXiv:2109.14483*, 2021. URL `https://arxiv.org/abs/2109.14483`.

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. doi: 10.1109/iccv51070.2023.01730.

Jia Wan, Ziquan Liu, and Antoni B. Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1974–1983, 2021. doi: 10.1109/cvpr46437.2021.00201.

Jia Wan, Qingzhong Wang, and Antoni B Chan. Kernel-based density map generation for dense object counting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3):1357–1370, 2022. doi: 10.1109/TPAMI.2020.3022878.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *Int. Conf. Comput. Vis.*, pp. 24591–24602, 2025. doi: 10.1109/iccv51701.2025.02280.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1595–1607, 2020.

Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. NWPU-Crowd: A large-scale benchmark for crowd counting and localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(6):2141–2149, 2021. doi: 10.1109/TPAMI.2020.3013269.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. In *European Conference on Computer Vision*, pp. 336–357. Springer, 2026. doi: 10.1007/978-3-032-37718-0_20.

Weidi Xie, J. Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):283–292, 2018. doi: 10.1080/21681163.2016.1149104.

Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 15548–15557, 2023. doi: 10.1109/cvpr52729.2023.01492.

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. doi: 10.1109/wacv56688.2023.00625.

Guanning Zeng, Xiang Zhang, Zirui Wang, Haiyang Xu, Zeyuan Chen, Bingnan Li, and Zhuowen Tu. YOLO-Count: Differentiable object counting for text-to-image generation. In *Int. Conf. Comput. Vis.*, pp. 16765–16775, 2025. doi: 10.1109/iccv51701.2025.01557.

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. doi: 10.1109/cvpr.2016.70.

Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-Shot Object Counting With Good Exemplars. In *Eur. Conf. Comput. Vis.*, pp. 368–385, 2024. doi: 10.1007/978-3-031-72652-1_22.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL `https://arxiv.org/abs/2504.10479`.

# A Experimental Protocol and Reproducibility

## A.1 Evaluation Protocol

**Benchmark prompt interfaces.** For KubriCount, we pass each evaluated image and the annotation's `category` string directly to `model.count_image(images, prompts)`; for example, the query for one image may be "royal blue hat." This text is the sole test-time prompt at every level. We provide no exemplar box, positive or negative sample box, additional negative text, or Type-A/Type-B box coordinates at Level 4. The published KubriCount baselines in Table 2 retain the native prompt interfaces documented by the benchmark authors Liu et al. (2026a); HieraCount uses text and visual exemplars, while the benchmark's MLLM protocol additionally supplies box coordinates as text at Level 4. Their count errors and ours therefore share the benchmark metric but use different prompt resources.

The MixCount comparison uses positive-text-only prompting. The 14.63 MAE for the CountGD++ baseline in Table 3 comes from the MixCount paper's text-only row, using a checkpoint additionally trained on MixCount data Dumery et al. (2026). The same source reports 2.11 MAE when that model also receives negative text and positive and negative visual exemplars; this multimodal result uses a different input interface from ConfluxCount.

Following Count Anything and the official CLOC-v1.1 evaluation protocol, we report mean absolute error (MAE), root mean squared error (RMSE), and normalized absolute error (NAE) to evaluate image-level counting accuracy. These metrics measure whether the predicted total count is correct, while remaining agnostic to how the predicted instances are spatially distributed. We further introduce spatial normalized absolute error (SNAE) to evaluate whether the predicted count is assigned to the correct image regions.

**Image-level counting metrics.** Let $M$ denote the number of evaluated image–query pairs. For the $i$-th sample, let $N_i > 0$ be the ground-truth count and $\hat{N}_i$ be the predicted count. The three image-level metrics are defined as

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}\left|\hat{N}_i - N_i\right|, \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\hat{N}_i - N_i\right)^2}, \tag{12}$$

$$\mathrm{NAE} = \frac{1}{M}\sum_{i=1}^{M}\frac{\left|\hat{N}_i - N_i\right|}{N_i}. \tag{13}$$

MAE measures the average absolute counting deviation, whereas RMSE places greater emphasis on samples with large counting errors. NAE normalizes the error by the ground-truth count of each sample before averaging, preventing high-density images from dominating the evaluation. Lower values indicate better performance for all three metrics.

**Spatial normalized absolute error.** The image-level metrics above cannot distinguish two predictions with the same total count but substantially different spatial distributions. To measure spatial counting consistency, we introduce SNAE, which extends NAE to a hierarchy of spatial partitions.

At level $L$, each image is divided into a regular $2^L \times 2^L$ grid containing $4^L$ cells. Let $n_{i,c}^{(L)}$ and $\hat{n}_{i,c}^{(L)}$ denote the ground-truth and predicted counts in cell $c$, respectively. SNAE at level $L$ is defined as

$$\mathrm{SNAE}(L) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{c=1}^{4^L}\left|\hat{n}_{i,c}^{(L)} - n_{i,c}^{(L)}\right|. \tag{14}$$

For point-based predictions, $\hat{n}_{i,c}^{(L)}$ is obtained by counting the predicted points inside the cell. For density-based predictions, it is computed by integrating the predicted density over the corresponding region. This formulation therefore provides a unified spatial evaluation for density, detection, and point-process representations.

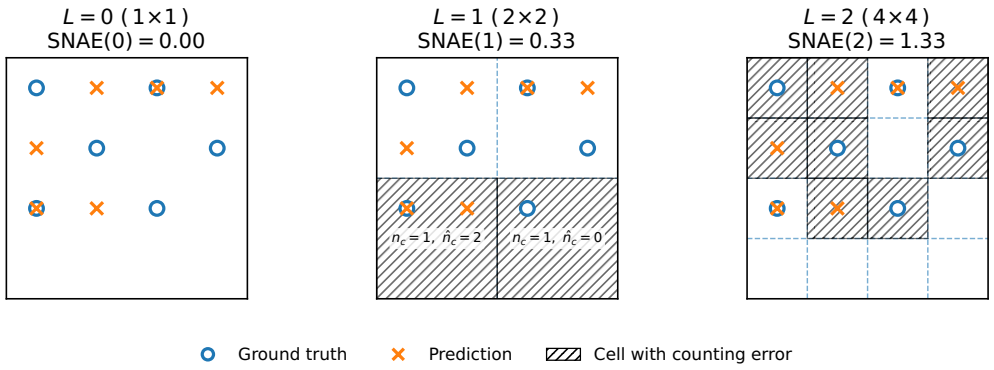



Figure 6: **Illustration of spatial normalized absolute error.** Ground-truth instances and predictions have the same global count, so $\mathrm{SNAE}(0) = \mathrm{NAE} = 0$. Finer partitions expose predictions assigned to incorrect spatial regions. Hatched cells contribute to the regional absolute error, increasing SNAE from $L = 1$ to $L = 2$.

SNAE is a spatial generalization of NAE. Increasing $L$ produces progressively finer partitions and imposes increasingly strict constraints on the spatial allocation of the predicted count. Because each coarse cell is partitioned into four child cells, the triangle inequality further gives

$$\mathrm{SNAE}(L+1) \geq \mathrm{SNAE}(L), \tag{15}$$

making the SNAE curve directly interpretable as counting error under increasing spatial resolution.

Unlike region-wise relative errors, SNAE normalizes the sum of regional errors by the total image count rather than by the count of each individual cell. It therefore remains well-defined for empty cells and penalizes false positive predictions in background regions without introducing division by zero. The subsequent macro-average assigns equal weight to each image, preserving the density-invariant property of NAE.

Figure 6 illustrates the behavior of SNAE. Although the example prediction has the correct image-level count and consequently zero error at $L = 0$, displaced predictions cross cell boundaries under finer partitions. These displacements create regional count deficits and surpluses, causing SNAE to increase with $L$. Therefore, low SNAE at fine partition levels indicates both accurate counting and more precise spatial localization.

### A.2 Training Data

This section describes the CLOC-only training setting used for the controlled ablation studies. The primary benchmark tables in the main paper instead use a model jointly trained on CLOC-v1.1, KubriCount, and MixCount.

We train ConfluxCount on the official training split of CLOC v1.1. The dataset exhibits a pronounced long-tailed count distribution: images containing hundreds or thousands of queried instances occur far less frequently than low-count scenes. Under uniform sampling, stochastic updates are therefore dominated by sparse images, providing limited supervision for activating and organizing large numbers of attractors in dense scenes.

To improve the representation of the dense tail, we adopt a soft power-law sampling strategy. Each training image $i$ is assigned the weight

$$w_i = \min\left(4,\ \left(1 + \frac{N_i}{300}\right)^{1/2}\right), \tag{16}$$

where $N_i$ denotes its ground-truth count. The weighting remains close to 1 for sparse images and gradually increases with the count, with a maximum sampling factor of 4. We implement this strategy using a distributed weighted sampler. This preserves the diversity of low-count scenes while exposing the model more frequently to images requiring hundreds or thousands of object-centered confluences.

All images are resized to $1008 \times 1008$ and normalized to $[-1, 1]$. The main recipe applies resizing and normalization only, as the intrinsic variation in object scale, appearance, and scene composition in CLOC v1.1 already provides substantial training diversity.

### A.3 Implementation Details

**Architecture.** ConfluxCount is built on a pretrained SAM3 image–text backbone. The image and text encoders are frozen, while rank-32 LoRA adapters are applied to the cross-attention layers. Grounded object queries and the text prototype are consolidated into 32 instance-memory tokens using two transformer blocks with eight attention heads. The semantic field is predicted on a $144 \times 144$ grid, and a stride-2 lattice initializes $72 \times 72 = 5{,}184$ probes. We use three shared probe transitions with a maximum displacement of $1.5$ feature-grid cells per step.

**DPP and losses.** The DPP factor combines 8-dimensional appearance and spatial embeddings, giving a rank-64 shared representation. The spatial bandwidth is $0.028$, the minimum private-evidence gate is $0.01$, and the quality-head bias is initialized to $-6$. We set $\lambda_{\text{dpp}} = \lambda_{\text{basin}} = \lambda_{\text{count}} = 1$, $\lambda_{\text{flow}} = \lambda_{\text{crowd}} = 0.5$, $\lambda_{\text{cov}} = 3$, and $\lambda_{\text{con}} = 0.2$.

**Canonical-probe routing.** In every training forward pass, each initial probe is assigned to its nearest ground-truth point by Euclidean distance. An equal-distance tie is resolved by the smaller point index. For each point with at least one assigned probe, we select the assigned probe nearest to that point in the initial lattice, resolving distance ties by the smaller global probe index. Flowed positions, quality scores, and bipartite matching do not enter this selection. If a point's Voronoi cell contains no probe, it has no canonical representative; we neither reuse another point's probe nor perform replacement matching. The resulting set of unique indices is used by both the DPP set likelihood and the marginal-coverage loss. All probes within target-supported basins still receive the basin, flow, and crowding supervision, and the count loss uses the full annotation count. We recompute the discrete assignment and selection for each forward pass without gradients; gradients from the DPP and coverage objectives flow through the selected probes' kernel values.

**Optimization.** We train for 60k iterations using AdamW with weight decay $10^{-4}$, a $500$-iteration linear warmup, and cosine decay to zero. The LoRA adapters, SAM decoder, and pixel decoder use learning rates of $2 \times 10^{-5}$, $1 \times 10^{-5}$, and $5 \times 10^{-6}$, respectively, while the newly introduced field, probe-dynamics, memory, and DPP modules use $2 \times 10^{-4}$.

## B Extended Ablations and Analyses

**Hyperparameter analysis.** Table 8 shows that ConfluxCount is generally stable across the tested step sizes, DPP bandwidths, and factor ranks. Transition depth has a clearer effect: moving from static probes to three transitions progressively improves MAE and NAE, while additional transitions provide no further benefit. The DPP rank exhibits a mild metric trade-off, with rank 16 favoring MAE/RMSE and rank 64 yielding the best NAE. We therefore use $T = 3$, $\delta = 1.5$, $\beta = 0.028$, and rank 64 as a balanced default rather than selecting hyperparameters separately for individual metrics.

Table 8: **Hyperparameter sensitivity.** We vary transition depth $T$, step size $\delta$, DPP bandwidth $\beta$, and factor rank while holding the remaining settings fixed. Default settings are bold.

| Group | Value | MAE↓ | RMSE↓ | NAE↓ |
|---|---|---|---|---|
| Transitions $T$ | 0 | 6.30 | 26.77 | 0.420 |
| | 1 | 6.27 | 26.47 | 0.388 |
| | **3** | **6.12** | 27.23 | **0.350** |
| | 5 | 6.22 | 29.15 | 0.359 |
| Step $\delta$ | 1.0 | 6.25 | 27.67 | 0.361 |
| | **1.5** | **6.12** | **27.23** | **0.350** |
| | 3.0 | 6.20 | 27.68 | 0.362 |
| Bandwidth $\beta$ | 0.014 | 6.22 | 28.17 | 0.362 |
| | **0.028** | **6.12** | **27.23** | **0.350** |
| | 0.056 | 6.27 | 27.69 | 0.365 |
| Rank | 16 | **6.00** | **26.80** | 0.370 |
| | **64** | 6.12 | 27.23 | **0.350** |
| | 256 | 6.26 | 27.65 | 0.366 |

**Probe-lattice resolution.** Table 9 reveals a clear accuracy–capacity trade-off. Reducing the lattice from 72×72 to 36×36 saves relatively little probe computation but substantially degrades counting accuracy. Figure 7 shows pronounced degradation in high-count scenes, indicating that insufficient probe capacity prevents nearby instances from being resolved. Increasing the lattice to 144×144 further improves all three metrics, but at a noticeably higher computational cost. We therefore use 72×72 as the default operating point, while the gains from the denser lattice also identify probe resolution as an important bottleneck in extremely dense scenes (Section B.4).

Table 9: **Probe-lattice resolution ablation.** We report counting accuracy and probe computation at three lattice resolutions.

| Stride | Lattice | MAE↓ | RMSE↓ | NAE↓ | Probe GFLOPs↓ |
|---|---|---|---|---|---|
| 1 | 144×144 | **5.86** | **25.70** | **0.260** | 230.9 |
| 2 | 72×72 | 6.12 | 27.23 | 0.350 | 158.7 |
| 4 | 36×36 | 10.57 | 70.26 | 0.380 | 140.6 |

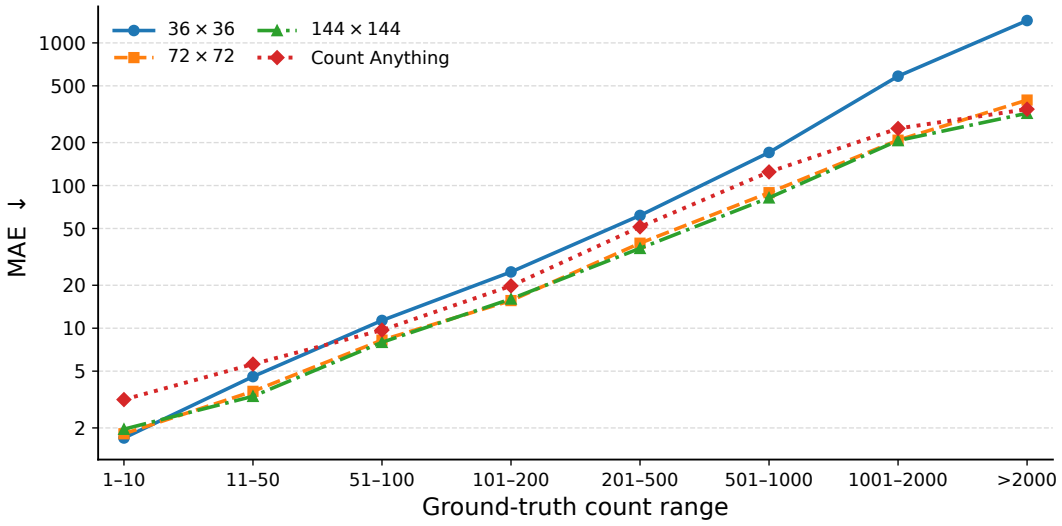
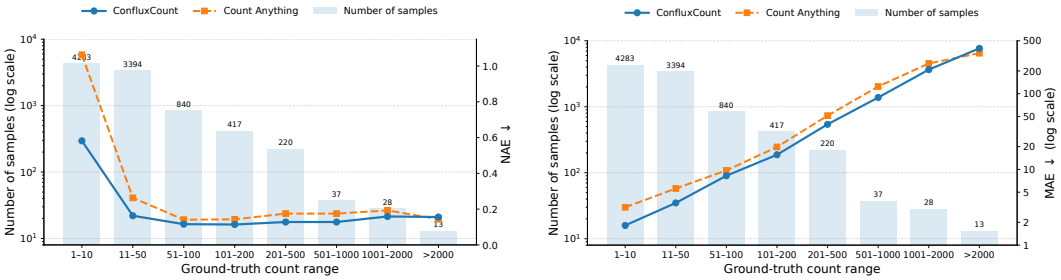



Figure 7: Counting error across count ranges for different probe resolutions.




(a) NAE across ground-truth count ranges.

(b) MAE across ground-truth count ranges.

Figure 8: **Counting performance across ground-truth count ranges on CLOC-v1.1.** The curves compare ConfluxCount and Count Anything over different count intervals, while the background bars indicate the number of test samples in each interval on a logarithmic scale. In terms of normalized absolute error, ConfluxCount consistently outperforms Count Anything for all ranges up to 2000, with the largest improvement in the sparse 1–10 regime. A similar trend is observed for MAE, where ConfluxCount remains better across all ranges up to 2000. Count Anything is marginally better only in the $> 2000$ regime, which contains very few test samples.

### B.1 Evaluation on CLOC-Test Sub-Benchmarks

Table 10: **Per-benchmark results on CLOC-v1.1.** We compare ConfluxCount and Count Anything using MAE, NAE, and SNAE at partition levels $L = 2$ and $L = 4$. Lower values are better, and the better result in each pair is bold.

| Benchmark | MAE↓ | | NAE↓ | | SNAE(2)↓ | | SNAE(4)↓ | |
|---|---|---|---|---|---|---|---|---|
| | ConfluxCount | CountAnything | ConfluxCount | CountAnything | ConfluxCount | CountAnything | ConfluxCount | CountAnything |
| *General scenes* | | | | | | | | |
| Objects365 | **2.30** | 3.32 | **0.324** | 0.515 | **0.404** | 0.630 | **0.642** | 0.809 |
| FSCD-LVIS | **10.51** | 13.30 | **0.173** | 0.238 | **0.279** | 0.330 | **0.499** | 0.551 |
| FSC-147 | **4.76** | 5.09 | 0.103 | **0.084** | **0.211** | 0.218 | **0.606** | 0.694 |
| Cityscapes | **2.11** | 2.62 | **0.708** | 0.791 | **0.835** | 0.924 | **1.087** | 1.094 |
| NWPU-CROWD | **8.76** | 9.07 | **0.090** | 0.105 | **0.170** | 0.172 | 0.371 | **0.348** |
| ShanghaiTech A | **72.19** | 73.24 | **0.131** | 0.150 | **0.178** | 0.189 | 0.344 | **0.335** |
| ShanghaiTech B | 13.63 | **10.57** | 0.070 | **0.061** | 0.110 | **0.106** | 0.260 | **0.246** |
| JHU-CROWD++ | **27.37** | 61.71 | **0.119** | 0.201 | **0.210** | 0.280 | **0.446** | 0.493 |
| *Remote sensing* | | | | | | | | |
| DIOR | **1.99** | 4.92 | **0.788** | 1.842 | **0.860** | 2.047 | **1.107** | 2.367 |
| VisDrone | **3.07** | 4.99 | **0.436** | 0.859 | **0.637** | 1.007 | **0.822** | 1.154 |
| NWPU-MOC | **10.31** | 13.95 | **0.299** | 0.497 | **0.468** | 0.638 | **0.719** | 0.873 |
| CARPK | 1.18 | **1.04** | 0.019 | **0.017** | 0.043 | **0.037** | 0.137 | **0.118** |
| UpCount | **5.97** | 8.24 | **0.096** | 0.148 | **0.185** | 0.251 | **0.343** | 0.398 |
| NWPU-VHR-10 | 2.38 | **2.09** | 1.328 | **1.154** | **1.194** | 1.257 | **1.380** | 1.507 |
| *Histopathology* | | | | | | | | |
| CoNIC | **3.93** | 16.29 | **0.493** | 1.747 | **0.884** | 2.079 | **1.116** | 2.350 |
| MoNuSAC | **15.24** | 19.00 | **0.745** | 2.464 | **1.068** | 2.618 | **1.286** | 2.902 |
| CellBinDB | **16.82** | 31.50 | **0.181** | 0.325 | 0.423 | **0.392** | 0.807 | **0.582** |
| EndoNuke | **10.91** | 14.77 | **0.090** | 0.113 | **0.151** | 0.183 | **0.419** | 0.465 |
| BCData | **26.14** | 27.24 | **0.099** | 0.115 | **0.149** | 0.160 | **0.311** | 0.322 |
| NuCLS | **2.95** | 9.90 | **0.125** | 0.434 | **0.301** | 0.562 | **0.652** | 1.025 |
| NuInsSeg | **7.86** | 13.48 | **0.177** | 0.238 | **0.360** | 0.419 | **0.607** | 0.619 |
| *Cell microscopy* | | | | | | | | |
| BriFiSeg | **2.06** | 4.38 | **0.082** | 0.179 | **0.201** | 0.334 | **0.490** | 0.685 |
| LIVECELL | **57.88** | 84.36 | **0.119** | 0.160 | **0.168** | 0.189 | 0.326 | **0.302** |
| VGG | **9.22** | 28.08 | **0.046** | 0.142 | **0.086** | 0.160 | **0.226** | 0.271 |
| *Agriculture* | | | | | | | | |
| GWHD | **4.00** | 4.51 | **0.100** | 0.115 | **0.161** | 0.197 | **0.367** | 0.409 |
| MTC | 4.81 | **4.33** | **0.215** | 0.290 | **0.298** | 0.372 | **0.484** | 0.564 |
| Soybean Pod | 274.77 | **257.90** | **0.138** | 0.150 | **0.207** | 0.217 | 0.398 | **0.375** |
| *Microorganisms* | | | | | | | | |
| AGAR | 2.31 | **2.24** | **0.100** | 0.103 | **0.115** | 0.128 | 0.212 | **0.175** |

Across the 28 CLOC-Test sub-benchmarks, ConfluxCount improves MAE on 22 and NAE on 24 (Table 10). The improvement is most consistent at moderate spatial resolution, where it leads on 25 datasets for SNAE(2). Large gains on aerial and biomedical test domains show that the advantage within CLOC-v1.1 is not confined to natural images; DIOR, CoNIC, NuCLS, and VGG are representative cases in which both global and spatial errors fall substantially.

The boundary becomes clearer at the finest partition: the lead narrows to 20 datasets for SNAE$(4)$, and Count Anything is better on several crowd, cell, and agriculture benchmarks. Because a small coordinate displacement can cross a fine cell boundary without changing the global count, this pattern separates robust cardinality gains from less uniform localization gains. ConfluxCount therefore improves spatial allocation broadly, but not universally at the finest scale.

### B.2 Accuracy across Count Ranges

Figure 8 separates absolute and relative error across count ranges. ConfluxCount is better in both metrics through $2000$ objects, with its largest relative gain in the sparse $1$–$10$ regime, where individual false positives matter most. Count Anything is slightly better beyond $2000$ objects in a bin of $13$ samples. We examine this dense-tail behavior and the default lattice's capacity in Section B.4.

### B.3 Robustness to Mismatched Queries

A reliable open-vocabulary counting model should suppress its output when the queried category is absent from the image. We therefore construct a mismatched image–query evaluation from the class-expanded CLOC-Test split. After removing duplicate image paths, the benchmark contains $7{,}860$ unique images. For each visual domain, we form a pool of categories that never appear in that domain and randomly sample one absent query for each image. The same sampled image–query pairs are used for all models, and every pair is assigned a ground-truth count of zero.

We evaluate mismatch robustness using the mean false count

$$\mathrm{MFC} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \hat{c}_i, \tag{17}$$

where $\hat{c}_i$ is the predicted count for the mismatched query. Since all ground-truth counts are zero, MFC directly measures the magnitude of spurious counting. We also report the query rejection rate

$$\mathrm{Rej} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \mathbb{I}[\hat{c}_i = 0]\,, \tag{18}$$

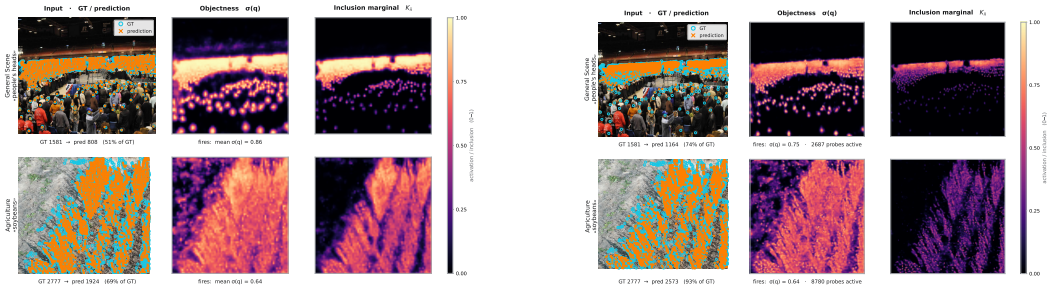
i.e., the fraction of mismatched queries for which the model predicts no instances. Lower MFC and higher Rej indicate stronger mismatch robustness.

Table 11: **Mismatched-query robustness.** Each image is paired with a category absent from its visual domain, and all models use the same image–query pairs. MFC denotes mean false count, and Rej denotes the fraction of zero-count predictions. Lower MFC and higher Rej are better.

| | **ConfluxCount** | | **CountAnything** | | **LocateAnything** | |
|---|---|---|---|---|---|---|
| **Domain** | **MFC↓** | **Rej↑** | **MFC↓** | **Rej↑** | **MFC↓** | **Rej↑** |
| All | **0.479** | **0.938** | 9.580 | 0.260 | 2.370 | 0.880 |
| General scenes | **0.552** | **0.932** | 8.750 | 0.260 | 2.120 | 0.880 |
| Remote sensing | **0.287** | **0.944** | 6.790 | 0.280 | 1.340 | 0.900 |
| Histopathology | **0.057** | **0.987** | 28.030 | 0.180 | 6.830 | 0.850 |
| Cellular microscopy | **0.669** | **0.987** | 38.840 | 0.230 | 16.800 | 0.830 |
| Agriculture | **0.051** | **0.993** | 13.730 | 0.340 | 0.910 | 0.970 |
| Microbiology | **0.073** | **0.986** | 16.720 | 0.110 | 4.540 | 0.860 |

Table 11 shows that ConfluxCount strongly suppresses spurious counting under absent queries. Overall, it reduces MFC to $0.479$ and rejects $93.8\%$ of mismatched queries, compared with $2.370/88.0\%$ for LocateAnything and $9.580/26.0\%$ for CountAnything. The improvement is consistent across all six domains, with MFC remaining below $0.7$ and rejection above $93\%$ in every case.

The gains are particularly pronounced in dense biomedical and microscopic domains, where unrelated structures can otherwise trigger large false counts. On histopathology, cellular microscopy, and microbiology, ConfluxCount keeps MFC between $0.057$ and $0.669$ with at least $98.6\%$ rejection, while both baselines produce substantially more false activations. This behavior is consistent with the semantic-specific supervision used during training: prompt-contrastive grounding separates the queried concept from competing categories, while empty-set supervision on mismatched




(a) Default $72\times72$ lattice.

(b) Resolution intervention at $144\times144$.

Figure 9: **Dense-scene capacity boundary.** Left: strong objectness is converted into too few inclusion units because the default lattice saturates and the DPP discounts unresolved neighbors. Right: a four-times denser lattice recovers substantial cardinality, but also increases redundancy and computation.

prompts suppresses probe evidence when no compatible object is present. As a result, absent queries produce little inclusion mass and the DPP cardinality naturally collapses toward zero.

### B.4 Failure Analysis in Extremely Dense Scenes

ConfluxCount systematically under-counts at the extreme dense tail. Figure 9a contrasts the post-dynamics objectness $\sigma(q_i)$ with the DPP marginal $K_{ii}$ for two representative failures. For active probes $\mathcal{A} = \{i \mid K_{ii} > 0.1\}$ in target regions, $\sigma(q_i) = q_i/(1+q_i)$ approximates the contribution before inter-probe repulsion, whereas $K_{ii}$ is the contribution after de-duplication and $\hat{N} = \sum_i K_{ii}$.

Both examples retain strong objectness over the target regions, locating the error after semantic activation. We quantify the suppression introduced by the DPP as

$$r_{\text{rep}} = 1 - \frac{\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} K_{ii}}{\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \sigma(q_i)}. \tag{19}$$

The discount is about $23\%$ in both cases. This is desirable when several probes describe one object, but excessive when distinct instances fall inside the effective repulsion bandwidth. A second ceiling follows from $K_{ii} \le 1$: a region cannot contribute more units than the probes that resolve it. The crowd activates only $1164$ probes for $1581$ heads and is therefore capacity-limited even without repulsion. The soybean field activates $3836$ probes for $2777$ objects, so its error is more directly associated with DPP discounting. Extreme density thus exposes two coupled limits: insufficient probe resolution and excessive repulsion among unresolved neighbors.

The resolution intervention in Figure 9b tests this diagnosis by increasing the lattice from $72\times72$ to $144\times144$. The finer lattice raises the active-probe counts to $2687$ and $8780$. Although the denser probe set also increases repulsion, predictions recover from $51\%$ to $74\%$ of ground truth in the crowd and from $69\%$ to $93\%$ in the soybean field. Resolution therefore recovers much of the missing cardinality, but at four times the probe count. Adaptive local refinement and density-dependent bandwidths are more promising remedies than uniformly increasing resolution.

# C Qualitative Analysis Across Domains

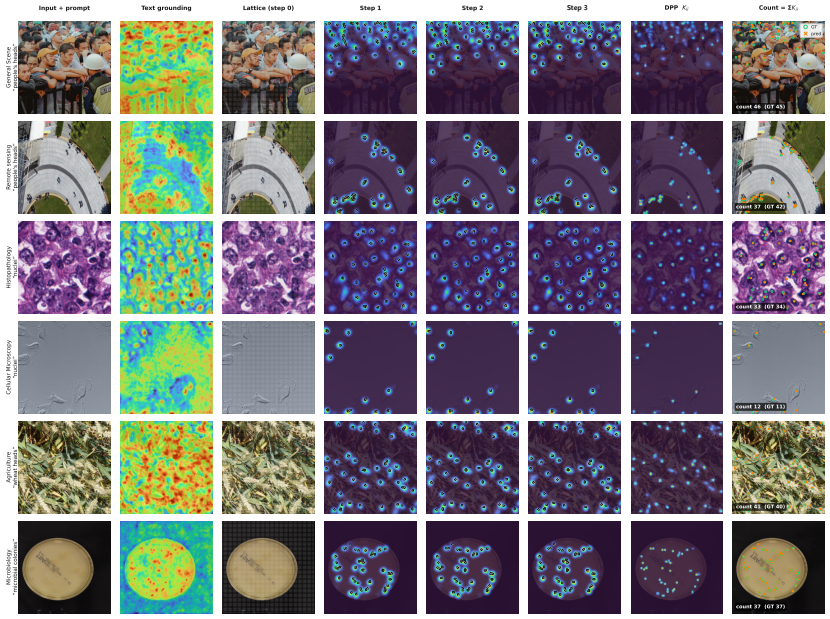



Figure 10: **Qualitative visualization of ConfluxCount across the six CLOC domains.** Rows span General Scenes, Remote Sensing, Histopathology, Cellular Microscopy, Agriculture, and Microbiology. Columns show the query and image, grounding response, initial lattice, three probe transitions, DPP marginals, and final locations. The same model and readout are used in every row.

Figure 10 traces inference across six visual domains. Grounding selects query-compatible regions; the lattice then contracts into compact modes as background support fades and same-object probes approach a shared center. The same progression appears in the cell, car, tissue, and crop examples despite their different appearances.

The DPP retains complementary probes and discounts redundant ones, producing expected counts and greedy-MAP locations from the same kernel. The final overlays broadly track the annotations, while crowded-row misses agree with the capacity boundary in Section B.4. The stages therefore realize a clear division of labor: select evidence, form confluences, and enumerate representatives.