点击与计数：面向类别无关目标计数的多模态原型适应

Amani Aifi[1], **Wassim Bouachir**[1],**Ahmed Zgaren**[1] 和 **Nizar Bouguila**[2]Université TÉLUQ，蒙特利尔，魁北克，加拿大[2]康考迪亚大学，蒙特利尔，魁北克，加拿大 aifi.amani@univ.teluq.ca, {Wassim.Bouachir, Ahmed.Zgaren}@teluq.ca，nizar.bouguila@concordia.ca

arXiv:2611.28321v2 [cs.CV] 25 Nov 2026

摘要—我们提出一种多模态、类别无关的密度式目标计数框架，可在单一迭代细化流程中通过视觉示例、文本描述或两者共同指定感兴趣对象。现有计数方法通常分为两类：一类融合两种模态——手绘边界框和文本描述——来指定目标对象；另一类是单模态方法，仅依赖边界框示例或文本。近期有少数方法用一个点或一次点击取代边界框。我们利用冻结的视觉提示分割模型，将单次点击扩展为精确的对象掩码，并将其输入迭代原型细化过程；该过程还会在每个细化步骤融合文本描述。我们的贡献有两项：第一，提出一种由点击生成的示例提取模块，免除以往基于示例的计数方法对边界框的要求；第二，提出一种逐步文本融合机制，在整个细化过程中结合视觉和文本提示。结合文本引导后，我们的模型实现了最先进的计数精度，在所有指标上均优于以往基于点击的计数方法，且在平均绝对误差方面优于边界框示例基准架构。代码见 github.com/ameniifi/Click-and-Count。

索引词—**类别无关目标计数、少样本学习、基于点击的示例、原型适应、多模态学习、视觉语言模型**

I. 引言

目标计数是一项基础的计算机视觉任务，旨在估计图像中某一目标对象类别的实例数量。它在诸如人群分析、工业检测、农业监测和生物图像分析等应用中发挥着重要作用。传统的计数方法通常从标注数据集 [1], [2], 中学习类别特定表示，这限制了它们在没有额外

训练或标注。

为了解决这一局限，类别无关目标计数应运而生，成为一种更灵活的形式 [3], [4], ，其中目标对象类别通过额外指导在推理时指定。模型不再依赖固定的类别集合，而是必须根据用户提供的信息（例如视觉示例、语义描述或空间提示）识别并计数任意对象。然而，如何提供有效的指导仍然是一项关键挑战，因为对象表示的质量

© 2027 IEEE。允许将本材料用于个人用途。对于任何其他用途，无论是在当前还是未来的媒体中，包括为广告或促销目的重印/再版本材料、创作新的合集作品、转售或再分发至服务器或列表，或在其他作品中重复使用本作品中任何受版权保护的部分，均必须获得 IEEE 的许可。

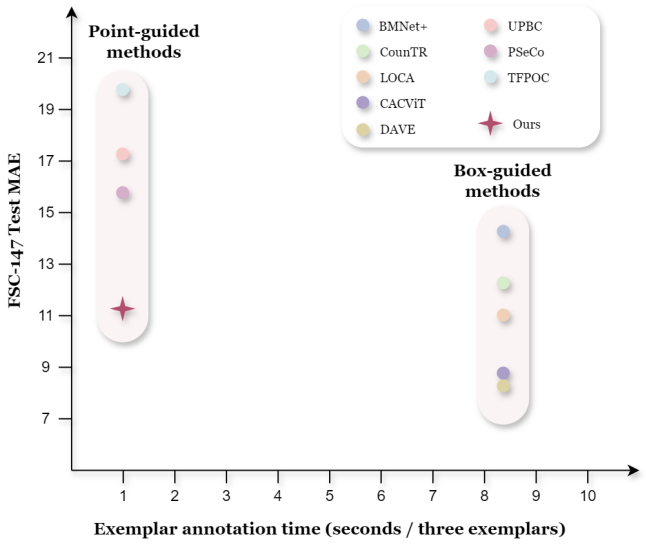


图1. 标注时间是通过对随机选取的样本中标注三个示例所需的时间取平均值估算得出的。此比较仅考虑视觉示例标注（点击和边界框），不包括文本提示。基于点击的标注大幅减少了标注工作量，同时保持了具有竞争力的计数性能。

会直接影响计数性能。

类别无关计数的一种广泛应用范式依赖于视觉示例 [5], [6],，用户通过提供目标对象的示例来引导模型。这些示例提供了宝贵的实例级信息，包括对象外观和空间特征。然而，获取此类示例通常需要精确定位对象，往往要手动绘制边界框。尽管边界框提供了丰富的空间监督，但创建多个紧密贴合的标注十分耗时，也限制了基于示例的计数系统的可扩展性。

如图1所示，点击标注提供了一种简单得多的替代方案：每个示例只需一个点来指示目标对象的位置。然而，点击所提供的信息远少于边界框，也无法直接描述完整的对象范围。因此，一个关键挑战是如何将稀疏的点交互转化为信息丰富的对象表示，同时保留

基于示例的计数的优势。

另一个挑战涉及语义信息的运用。文本描述 [7], [8] 无需手动进行视觉标注即可提供类别级知识，但缺少外观、尺度和空间结构等实例特定细节。因此，视觉信息与语义信息对目标物体的描述互为补充，却都不完整。如何有效结合这两种模态，仍是提升目标计数泛化能力的重要方向。

在这项工作中，我们提出 C&C（点击计数），一种多模态点击引导计数框架，通过极少的用户交互重新构建基于示例的目标计数。C&C 无需手动指定边界框，而是利用稀疏点击，并通过点击驱动的目标提取过程来恢复目标感知视觉表示。这些表示能够捕获空间与外观信息，从而根据轻量级用户标注迭代细化原型。

我们的主要贡献总结如下:

- 我们提出 C&C（点击计数），一种点击引导的类别无关计数框架，利用稀疏点标注而非手动定义的边界框实现基于示例的计数。
- 我们提出了一种由点击驱动的示例表示策略，将最少的用户交互转化为包含空间信息和外观信息的目标感知视觉原型。
- 我们引入了语义原型条件化作为一种可选的多模态扩展，使文本信息能够在迭代优化过程中补充视觉原型。
- 大量实验表明，C&C 在仅需大幅简化用户交互的同时，仍能实现具有竞争力的计数性能，并且能够有效泛化至未见目标类别。

用于复现我们结果的代码和数据已发布于 https://github.com/ameniifi/Click-and-Count。

II. 相关工作

计数方法分为基于检测的方法和基于密度的方法：前者通过边界框或掩码定位单个实例， [9], [10], 后者则对预测密度图进行积分以得出计数[1], [2]；后者对遮挡和拥挤场景更为稳健。类别无关计数将这一理念扩展到仅在推理时指定的类别。FSC-147 基准测试集[4] 为这一任务设立了标准协议，并推动了基于示例的方法的发展。这类方法将边界框示例与图像特征进行匹配，包括 GMN [3], 、FamNet[4], 、BMNet [5],、SPDCN [11], 、CounTR [12], 和 LOCA [6]。边界框示例也为 DAVE [13], 的小样本变体提供引导，该方法会验证候选检测结果，并利用已验证的检测结果细化密度图。在这些方法中，边界框都用于指定要计数的对象。在我们的基于密度的框架中，视觉示例仍发挥这一作用，而我们通过单次点击来构建这些示例。

近期研究开始采用边界框之外的提示来降低标注成本。文本引导方法（ZSOC [7],、CLIP-Count [8], 、GroundingREC [14]）利用视觉语言预训练，但缺少针对具体实例的外观线索。基于点的方法以点击取代边界框: PSeCo [15] 和 TFPOC[16]探索了点监督计数和基于分割的计数，而 UPBC [17]则将边界框、点和文本提示统一为共享表示。不同于这些方法，我们将用户提供的每次点击扩展为掩码派生形状和外观原型，并在每个迭代细化步骤中融合文本。

III. 方法

A. 问题定义

给定一幅输入图像

$$I \in \mathbb{R}^{3\times H\times W}, \tag{1}$$

其中， H 和 W 分别表示图像的高度和宽度，基于示例的目标计数旨在预测属于目标对象类别的所有对象实例的空间分布。

模型接收一组描述感兴趣对象的示例规格:

$$C = \{c_i\}_{i=1}^{N}, \qquad c_i = (x_i, y_i), \tag{2}$$

其中，每个 c_i 表示一次用户点击的空间坐标，该点击指示一个示例对象实例； $N = 3$表示示例的数量（每个示例对应一次点击），遵循FSC-147协议。

模型根据这些示例预测一张密度图:

$$R \in \mathbb{R}^{H\times W}, \tag{3}$$

其中，估计的对象计数通过以下方式获得:

$$\hat{N} = \sum_{x,y} R(x, y). \tag{4}$$

我们的框架是一种多模态迭代原型式目标计数方法，旨在最大限度减少示例标注，同时保留准确的对象表示。所提框架无需手动绘制边界框，而是采用半自动示例提取机制，将每个示例的一次用户点击与可提示视觉分割模型相结合，从而恢复完整的对象区域。随后，利用预测的对象掩码提取互补的形状和外观原型，以最少的人工干预提供丰富的视觉表示。

除了视觉信息外，该框架还通过预训练的视觉语言模型引入语义引导。在迭代细化之前及过程中，将语义特征注入视觉原型，生成多模态对象原型，共同编码外观、形状和类别级语义，从而实现稳健的类别无关目标计数。

B. 所提框架概述

图2概述了所提出的点击引导多模态计数框架。该框架旨在通过结合点击引导的视觉提取与语义引导，仅凭少量用户交互构建强大的对象原型。

所提框架包含四个主要阶段：
1) 视觉特征提取：将输入图像编码为特征表示，用于原型构建和对象定位。2) 点击引导的示例构建：将用户提供的点击输入可提示的视觉分割模型，以恢复完整的对象区域。利用生成的掩码提取互补的形状和外观原型。3) 多模态原型细化：利用从文本描述中提取的语义信息增强视觉原型。通过整合外观、图像上下文和语义线索，对生成的原型进行迭代细化。4) 原型匹配与密度预测：将细化后的对象原型与图像特征匹配，生成最终的密度图和对象计数。

C. 视觉特征提取

输入图像经预训练的卷积骨干网络处理后提取多尺度特征，再将这些特征调整至相同分辨率、拼接并投影到统一嵌入空间。随后，带有位置编码的 Transformer 编码器捕获图像区域之间的长程交互，生成编码后的视觉表示 $F_E \in \mathbb{R}^{d\times h\times w}$ （嵌入维度为 d，空间尺寸为$h \times w$），作为后续原型提取和细化阶段的视觉记忆。

D. 点击引导的示例表示

所提框架引入了一种基于点击的示例表示，将稀疏的用户交互转化为感知对象的视觉提示。

1) 对象掩码提取：对于每个示例点击 c_i，该框架使用基于提示的视觉分割模块，识别与点击位置对应的完整对象区域。

给定输入图像和一个用户交互点，分割模块会预测一个对象掩码：

$$M_i = \Phi(I, c_i), \tag{5}$$

其中，M_i 表示与第 i个示例对应的预测对象区域。

分割模块在训练期间保持冻结状态，仅用于从稀疏的用户交互中提取对象掩码。这些掩码提供对象级空间信息，随后用于构建示例表示。

2) 掩码引导的原型构建：提取的目标掩码提供对象级空间信息，用于构建示例表示。对于每个掩码 M_i，我们提取两种互补的表示：描述对象几何形状的形状原型，以及捕捉其视觉特征的外观原型。

a) 形状原型：M_i 的前景区域可得一个紧致的边界区域 $b_i = (x_1, y_1, x_2, y_2)$，其边界由前景位置的最小值和最大值确定，并据此计算物体宽度和高度 $w_i = x_2 - x_1$ 和 $h_i = y_2 - y_1$。这些几何描述符被转换为形状原型 $Q_i^S = \phi(w_i, h_i) \in \mathbb{R}^{s^2\times d}$。

b) 外观原型：将对象掩码调整为编码后的图像特征的空间分辨率。随后将掩码应用于特征图，以保留物体特定信息：

$$F_i^m = F_E \odot \tilde{M}_i. \tag{6}$$

使用相应的边界区域提取掩码对象区域：

$$P_i = \mathrm{RoIAlign}(F_i^m, \hat{b}_i), \tag{7}$$

同时，掩码覆盖率和原始未掩码特征可计算如下：

$$C_i = \mathrm{RoIAlign}(\tilde{M}_i, \hat{b}_i), \quad B_i = \mathrm{RoIAlign}(F_E, \hat{b}_i). \tag{8}$$

外观原型由掩码特征表示构建。由于掩码可能会降低部分覆盖空间位置处的特征响应，我们使用相应的覆盖率对掩码特征进行归一化：

$$Q_i^A(u,v) = \frac{P_i(u,v)}{\max(C_i(u,v), \tau)}, \tag{9}$$

其中，C_i 表示前景覆盖率，τ 是用于防止数值不稳定性的小常数。

对于覆盖率不足的位置（$C_i(u, v) \le \tau$），将掩码表示替换为原始区域特征，$Q_i^A(u, v) = B_i(u, v)$，以减轻掩码边界不准确带来的影响，同时保留以对象为中心的特征提取优势。由于掩码表示和未掩码表示均直接来自 F_E，因此无需额外投影，从而得到 $Q_i^A \in \mathbb{R}^{s^2\times d}$。

E. 语义原型条件化

尽管视觉原型提供了实例级信息，但仅凭外观线索可能不足以捕捉目标类别的语义身份。因此，我们使用冻结的文本编码器引入语义引导。

给定文本描述 T，文本编码器提取出语义表示：

$$z_T = E_{\mathrm{text}}(T), \tag{10}$$

该表示随后被投影到原型嵌入空间：

$$\tilde{z}_T = \mathrm{Proj}(z_T), \tag{11}$$

其中，Proj(·) 是先进行可学习线性投影、再进行层归一化的操作。

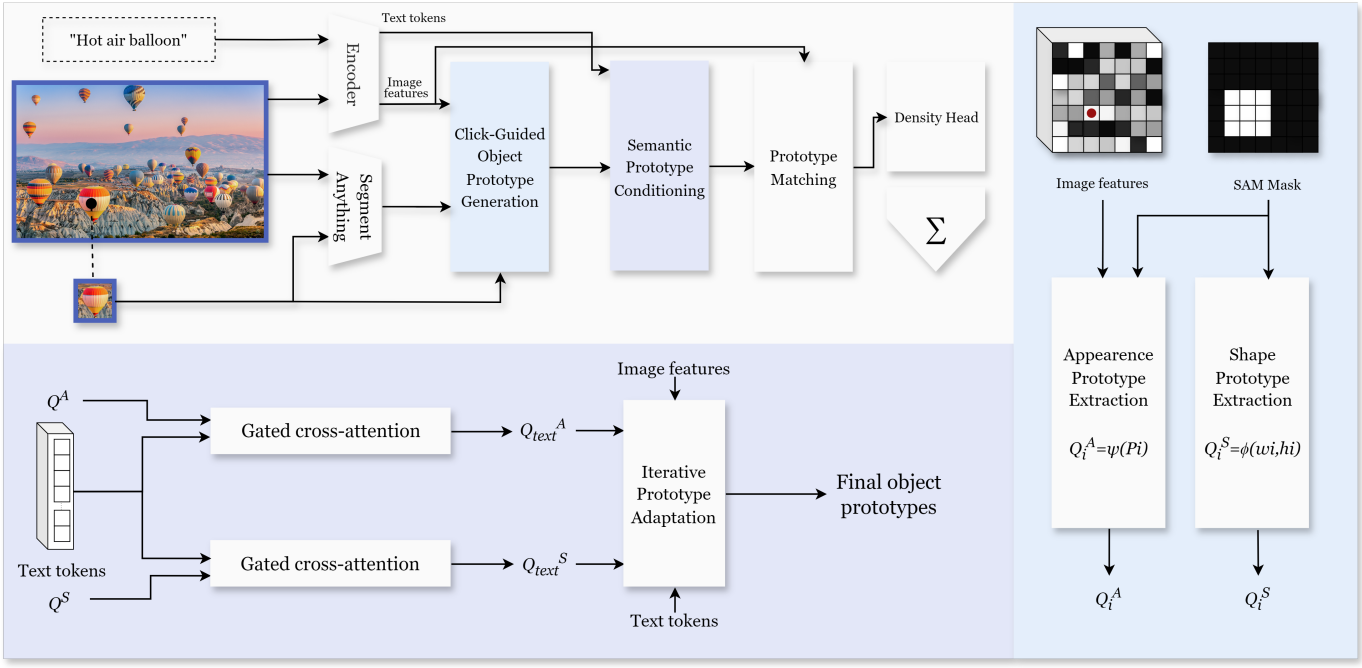


图 2. 所提出的点击引导的多模态目标计数框架概览。所提出的组件以彩色突出显示。点击引导的示例构建模块（蓝色）将用户点击转换为对象掩码，并提取形状和外观表示；而语义条件化模块（紫色）则将类别级信息注入原型中。细化后的多模态原型与图像特征进行匹配，以生成最终的密度图和对象计数。

为融入语义信息并保留原始原型结构，我们采用门控交叉注意力机制。对于输入原型 Q，语义条件操作定义为：

$$\mathcal{T}(Q, \tilde{z}_T) = Q + \tanh(g)\,\mathrm{MHA}(\mathrm{LN}(Q), \tilde{z}_T, \tilde{z}_T), \qquad (12)$$

其中，MHA(·) 表示多头交叉注意力，g 是初始化为零的可学习标量门控。这样的初始化使模型能够保留原始视觉原型表示，并逐步学习语义校正。

语义条件化分别应用于形状原型和外观原型：$Q^S_{\text{text}} = \mathcal{T}(Q^S, \tilde{z}_T)$ 和 $Q^A_{\text{text}} = \mathcal{T}(Q^A, \tilde{z}_T)$。

随后，在迭代原型细化过程中使用这些经过语义增强的原型。为了在视觉推理过程中保持语义一致性，在外观与图像上下文聚合之后引入了额外的门控文本交叉注意力操作：

$$Q_\ell^{(3)} = Q_\ell^{(2)} + \tanh(g_\ell)\,\mathrm{MHA}(\mathrm{LN}(Q_\ell^{(2)}), \tilde{z}_T^{(\ell)}, \tilde{z}_T^{(\ell)}). \qquad (13)$$

每个细化层都有一个独立的可学习门控 g_ℓ 和文本投影 $\tilde{z}_T^{(\ell)}$。当语义信息不可用时，残差文本分支会变为非激活状态，使细化过程仅使用视觉信息即可运行。

F. 原型匹配与密度预测

经过迭代细化后，最终原型 Q_L 用于在图像特征空间中识别目标物体。每个细化后的原型都会与编码后的图像特征通过逐通道相关操作进行匹配，从而生成示例特定响应图。

随后将得到的响应图进行聚合，并通过密度回归解码器处理，以生成最终的密度图 R，再根据公式（4）得到预测计数 N。

在整体框架中，原型匹配和密度回归阶段保持不变，使所提出的改进集中于点击引导的示例构建和语义原型条件化模块。

G. 优化目标

模型使用密度回归目标进行训练。给定真实密度图 R^*，主要训练目标是最小化预测密度分布与目标密度分布之间的差异：

$$\mathcal{L}_{\text{density}} = \frac{1}{N_o} \|R - R^*\|_2^2, \qquad (14)$$

其中，R 和 R^* 分别表示预测密度图和真实密度图，N_o 表示标注对象的数量。

为改进中间原型细化过程，我们对中间密度预测施加辅助监督。当语义描述可用时，额外的对齐目标会促使细化原型与提取的文本表示保持一致。

完整的训练目标定义如下：

$$\mathcal{L} = \mathcal{L}_{\text{density}} + \lambda_{aux}\mathcal{L}_{aux} + \lambda_{text}\mathcal{L}_{text}. \qquad (15)$$

仅当提供有效的文本信息时，语义对齐项才会启用。否则，模型仅使用密度估计目标进行优化。

IV. 实验

我们将点击引导计数框架与类别无关计数基线进行比较，评估以点击引导表示替换示例框的效果，并对视觉示例构建和语义原型条件化进行消融实验。

A. 数据集

FSC-147 [4]。我们使用 FSC-147 数据集作为类别无关目标计数的主要基准。该数据集包含 6,135 张图像，涵盖 147 个物体类别。遵循标准协议，这些类别被划分为 89 个训练类别、29 个验证类别和 29 个测试类别。每张图像最初都标注了三个用于定义计数参考物体的示例边界框。

在我们的框架中，每个示例框都被一个点击所替换，并将该点击用作 SAM 提示，以生成对象掩码。

基于 COCO 的计数基准。为了评估类别无关计数模型在类别偏移下的泛化能力，我们基于 COCO 2017 的实例标注构建了一个额外的计数基准。不同于标准的 COCO 评估协议（其中相同类别可能同时出现在训练图像和评估图像中），我们采用类别不相交设置，即评估类别完全不包含在训练数据中。

具体而言，我们从 COCO 的可用类别中随机选取 20 % 作为留出类别，并将其从训练划分中移除。其余类别用于训练，而验证和测试样本仅从留出类别中构建。我们还根据图像标识进一步划分留出样本，以确保评估针对的是未见类别。

为构建计数样本，我们仅考虑非拥挤物体实例，并保留包含至少三个同类别物体的图像。我们依据最小面积和边界框尺寸约束筛选对象实例，以去除极小或不可靠的标注。对于每个图像-类别对，我们选取三个对象实例作为示例，并将其空间位置转换为点提示，作为所提框架的引导。由此构建的基准用于评估模型能否将学到的对象表示适配到训练期间未见过的类别。

数据集统计信息汇总于表 II。

B. 实验设置

视觉编码器采用经 SwAV 预训练的 ResNet-50 主干网络；多尺度特征被投影到共同的嵌入空间，并通过带位置编码的 3 层 Transformer 编码器处理，以生成图像记忆。图像被调整为 512×512 ，特征缩减因子为 $\rho = 8$；原型精炼模块采用嵌入维度 $d = 256$、8 个注意力头、核大小 $s = 3$，并执行三个精炼步骤。目标掩码根据点击标注预先离线计算，以加快训练和推理。

模型使用 AdamW 进行优化（lr $= 10^{-4}$，权重衰减 10^{-4}），训练 200 个 epoch，每个 GPU 的批量大小为 4，共使用两块 NVIDIA GPU；在第 120 个 epoch 后降低学习率，并将梯度裁剪范数设为 0.1。辅助损失权重为 0.3；使用语义引导时，文本对齐损失权重为 0.1，并在 30 个 epoch 的预热阶段后启用。

C. 评估指标

遵循以往的类别无关计数工作，我们使用平均绝对误差（MAE）、均方根误差（RMSE）和归一化绝对误差（NAE）对模型进行评估：

$$MAE = \frac{1}{|D|}\sum_{i\in D}|\hat{N}_i - N_i|, \tag{16}$$

$$RMSE = \sqrt{\frac{1}{|D|}\sum_{i\in D}(\hat{N}_i - N_i)^2}, \tag{17}$$

$$NAE = \frac{1}{|D|}\sum_{i\in D}\frac{|\hat{N}_i - N_i|}{N_i}, \tag{18}$$

ˆ其中， N_i 表示通过对预测密度图进行积分得到的预测对象计数， N_i 表示真实对象计数。数据集大小表示为$|D|$。

D. 实验结果

FSC-147 上的结果。我们在 FSC-147 基准测试集上评估了该框架。该数据集包含多种物体类别，为类别无关目标计数提供了具有挑战性的场景。表 I 展示了在不同提示模态下与现有方法的比较，包括边界框、文本和基于点的引导。

结果表明，基于点击的目标表示可以有效替代人工标注的边界框：在显著降低标注负担的同时，其性能可与采用边界框监督的方法相媲美。通过文本条件引入语义信息后，性能进一步优于单独使用点引导的情况：点和文本提示的组合在所有比较方法中取得了最低的平均绝对误差，表明实例级视觉线索与类别级语义对于原型细化具有互补作用。然而，这种改进在不同指标上并不均衡：尽管点+文本取得了总体最低的平均绝对误差，但其均方根误差仍高于采用边界框监督的 LOCA 基线（测试集上为 68.69，对比 56.97），这表明，即使点击生成的原型能够提升平均情况下的准确性，它们在一部分具有挑战性的样本上仍更容易产生较大误差。

COCO衍生基准上的结果。在类别不相交的 COCO衍生基准上，我们的模型取得了 MAE

表 I 不同提示设置下 FSC-147 目标计数与最先进方法的比较。

提示	方法	验证集平均绝对误差	验证集均方根误差	测试集平均绝对误差	测试集均方根误差
box	FR [18]	45.45	112.53	41.64	141.04
	FSOD [19]	36.36	115.00	32.53	140.65
	C-DETR [20]	–	–	16.79	123.56
	SAM [21], [22]	31.20	100.83	27.97	131.24
	SAM-Free [16]	–	–	19.95	132.16
	LOCA [6]	10.24	32.56	10.79	56.97
	PSeCo [15]	15.31	68.34	13.05	112.86
text	ZSOC [7]	26.93	88.63	22.09	115.17
	CLIP-C [8]	18.79	61.18	17.78	106.62
	TFPOC [16]	–	–	24.79	137.15
	UPBC [17]	16.92	58.92	16.81	105.83
	GroundingDINO [23], [24]	12.07	57.02	11.52	97.37
点	TFPOC [16]	–	–	20.10	132.83
	UPBC [17]	17.16	59.38	15.86	103.27
	Ours	**11.57**	**35.69**	**12.31**	**89.40**
点 + 文本	**Ours**	**9.95**	**36.03**	**9.77**	**68.69**

表 II 用于评估的数据集统计信息。

数据集属性	FSC-147	源自 COCO
图像	6,135	66,823
类别	147	80
划分协议	类别不相交	类别不相交
每个样本的实例数	≥ 3	≥ 3
评估任务	类别无关计数	类别无关计数

表 III 基于点击（仅点）的 C&C 模型在 FSC-147 和所提出的类别不相交的 COCO 衍生计数基准上的性能，评估时不使用语义文本指导。

数据集	MAE	RMSE	NAE
FSC-147（点）	11.57 / 12.31	35.69 / 89.40	0.23 / 0.26
源自 COCO（点）	1.57 / 1.48	2.23 / 2.11	0.27 / 0.28

为 1.57/1.48、RMSE 为 2.23/2.11（验证集/测试集，表 III）。不同数据集划分上的结果高度一致，表明在类别偏移下原型构建稳定，并能将计数能力迁移至训练期间未见过的类别。

由于源自 COCO 的图像中每张图像的实例数少于 FSC-147，因此不同数据集之间的原始平均绝对误差/均方根误差无法直接比较；我们改为比较归一化绝对误差（NAE）。两个基准数据集的 NAE 值相近（验证集/测试集分别为 0.23/0.26 和 0.27/0.28），表明尽管计数分布不同，相对计数精度仍保持一致。

E. 消融实验

表 IV 在 FSC-147 上分别验证了提出的两个组件。将边界框示例替换为每个示例对应的一次点击

表 IV 示例类型与文本引导的消融：边界框示例（LOCA）、仅点击与点击+文本（我们的方法）。

方法	验证集平均绝对误差	验证集均方根误差	测试集平均绝对误差	测试集均方根误差
LOCA（边界框）	10.24	32.56	10.79	56.97
我们的方法（点）	11.57	35.69	12.31	89.40
我们的方法 + 文本	**9.95**	**36.03**	**9.77**	**68.69**

（我们的方法）在 +1.33/+1.52 平均绝对误差（验证集/测试集）下的代价，并使测试集均方根误差从 56.97 升至 89.40。加入文本（我们的方法 +文本）后，平均绝对误差降低（验证集从 11.57 降至 9.95，测试集从 12.31 降至 9.77），测试集均方根误差也降低（从 89.40 降至 68.69），使平均绝对误差低于 LOCA；然而，验证集均方根误差略有上升（从 35.69 升至 36.03），测试集均方根误差仍高于 LOCA。

五、结论

本文提出 C&C，一种点击引导的类别无关计数框架，可将稀疏点击转化为感知物体的示例，从而免除边界框标注。语义原型条件化进一步在细化过程中融入文本信息。在 FSC-147 和一个类别不相交的 COCO 衍生基准上的实验表明，该方法效果显著，并能泛化至未见类别。未来工作将探索更丰富的视觉语言表示，以及更具适应性的多模态原型学习。

致谢

本研究获得加拿大人工智能预防自杀研究主席项目（CRC-2023- 00036）的支持。

参考文献

[1] Y. Zhang、D. Zhou、S. Chen、S. Gao 和 Y. Ma，“通过多列卷积神经网络进行单图像人群计数”，载于 IEEE计算机视觉与模式识别会议（CVPR），2016年，第 589–597 页。 [2] Y. Li、X. Zhang 和 D. Chen，“CSRNet：用于理解高度拥挤场景的空洞卷积神经网络”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2018年，第 1091–1100 页。 [3] E. Lu、W. Xie 和 A. Zisserman，“类别无关计数”，载于 亚洲计算机视觉会议（ACCV），2018年，第 669–684 页。 [4] V. Ranjan、U. Sharma、T. Nguyen 和 M. Hoai，“学习计数一切”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2021年，第 3394–3403 页。 [5] M. Shi、H. Lu、C. Feng、C. Liu 和 Z. Cao，“表征、比较与学习：一种类别无关计数的相似性感知框架”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2022年，第 9529–9538 页。 [6] N. Đuki´c、A. Lukeži˘c、V. Zavrtanik 和 M. Kristan，“采用迭代原型适应的低样本目标计数网络”，载于 IEEE/CVF国际计算机视觉会议（ICCV），2023年，第 18 872– 18 881 页。[7] J. Xu、H. Le、V. Nguyen、V. Ranjan 和 D. Samaras，“零样本目标计数”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2023年，第 15 548–15 557 页。 [8] R. Jiang、L. Liu 和 C. Chen，“CLIP-Count：迈向文本引导的零样本目标计数”，载于 ACM国际多媒体会议（ACM MM），2023年，第 4535–4545 页。 [9] S. Ren、K. He、R. Girshick 和 J. Sun，“Faster R-CNN：借助区域提议网络实现实时目标检测”，载于 神经信息处理系统进展（NeurIPS），第 28 卷，2015年，第 91–99 页。[10] K. He、G. Gkioxari、P. Dollár 和 R. Girshick，“Mask R-CNN”，载于 IEEE国际计算机视觉会议（ICCV），2017年，第 2980–2988 页。[11] W. Lin、K. Yang、X. Ma、J. Gao、L. Liu、S. Liu、J. Hou、S. Yi 和 A. B. Chan，“用于示例引导的类别无关计数的尺度先验可变形卷积”，载于 英国机器视觉会议（BMVC），2022年，第 313 页。 [12] C. Liu、Y. Zhong、A. Zisserman 和 W. Xie，“CounTR：基于 Transformer 的通用视觉计数”，载于 英国机器视觉会议（BMVC），2022年。 [13] J. Pelhan、A. Lukeži˘c、V. Zavrtanik 和 M. Kristan，“DAVE——一种用于低样本计数的检测与验证范式”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2024年，第 23 293– 23 302 页。[14] S. Dai、J. Liu 和 N.-M. Cheung，“指代表达式计数”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2024年，第 16 985–16 995 页。 [15] Z. Huang、M. Dai、Y. Zhang、J. Zhang 和 H. Shan，“点选、分割与计数：目标计数通用框架”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2024年，第 17 067–17 076 页。 [16] Z. Shi、Y. Sun 和 M. Zhang，“利用提示进行免训练目标计数”，载于 IEEE/CVF冬季计算机视觉应用会议（WACV），2024年，第 323–331 页。 [17] W. Lin 和 A. B. Chan，“基于不动点方法的统一提示计数”，载于 AAAI人工智能会议论文集，第 38 卷，第 4 期，2024年，第 3468–3476 页。 [18] B. Kang、Z. Liu、X. Wang、F. Yu、J. Feng 和 T. Darrell，“通过特征重加权实现少样本目标检测”，载于 IEEE/CVF国际计算机视觉会议（ICCV），2019年，第 8420–8429 页。 [19] Q. Fan、W. Zhuo、C.-K. Tang 和 Y.-W. Tai，“结合注意力RPN和多关系检测器的少样本目标检测”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2020年，第 4013–4022 页。[20] T. Nguyen、C. Pham、K. Nguyen 和 M. Hoai，“少样本目标计数与检测”，载于 欧洲计算机视觉会议（ECCV），2022年，第 348–365 页。 [21] A. Kirillov、E. Mintun、N. Ravi、H. Mao、C. Rolland、L. Gustafson、T. Xiao、S. Whitehead、A. C. Berg、W.-Y. Lo、P. Dollár 和 R. Girshick，“分割一切”，载于 IEEE/CVF国际计算机视觉会议（ICCV），2023年，第 3992–4003 页。

[22] Z. Ma、X. Hong 和 Q. Shangguan，“SAM 能数任何东西吗？关于 SAM 计数的一项实证研究”，*arXiv* 预印本 *arXiv:2304 10817*，2023年。. [23] S. Liu、Z. Zeng、T. Ren、F. Li、H. Zhang、J. Yang、Q. Jiang、C. Li、J. Yang、H. Su、J. Zhu 和 L. Zhang，“Grounding DINO：将 DINO 与基础预训练相结合，实现开放集目标检测”，载于欧洲计算机视觉会议（*ECCV*），2024年，第 38–55 页。[24] S. Liu、P. Zhang、S. Zhang 和 W. Ke，“CountSE：软示例开放集目标计数”，载于*IEEE/CVF*国际计算机视觉会议（*ICCV*），2025年，第 21 536–21546 页。