arXiv:2601.01024v1 [cs.CV] 3 Jan 2026

ITSELF：注意力引导的细粒度对齐，用于视觉-语言检索

Tien-Huy Nguyen[1, 2, *] Huu-Loc Tran[1, 2*] Thanh Duc Ngo[1,2]
[1] 越南胡志明市信息技术大学[2] 越南国家大学，胡志明市，越南

摘要

视觉语言模型（*VLMs*）发展迅速，在基于文本的行人搜索（*TBPS*）任务中展现出巨大潜力。该任务需要捕捉图像与文本之间的细粒度关系，以区分不同个体。以往方法通过局部对齐应对这些挑战，但往往容易进行捷径学习并受到虚假相关性的影响，从而导致对齐失误。此外，注入先验知识可能会扭曲模态内结构。基于我们发现编码器的注意力在训练最早期便能从空间上精确地提取证据，并且为缓解这些问题，我们提出了***ITSELF***，一种用于隐式局部对齐的注意力引导框架。其核心是，带注意力库的引导表示（*GRAB*）将模型自身的注意力转换为由高显著性词元组成的注意力库，并在该库上应用局部目标，从而无需额外监督即可学习细粒度对应关系。为使选择过程可靠且避免冗余，我们提出了用于鲁棒选择的多层注意力（*MARS*），通过聚合各层注意力并执行具备多样性感知的前k选择；以及自适应令牌调度器（*ATS*），在训练过程中将保留预算从粗粒度逐步调度至细粒度，早期保留上下文，同时逐渐聚焦于具有判别性的细节。在三个广泛使用的*TBPS*基准数据集上进行的大量实验表明，最先进方法的性能以及强大的跨数据集泛化能力，证实了我们的方法在无需额外先验监督的情况下具有有效性和鲁棒性。我们的项目已公开发布于 *https://trhuuloc.github.io/itself*

1. 引言

基于文本的人物搜索（TBPS）旨在从大型图像库中找出与文本查询最匹配的人物 [15]。解决这一问题需要提取身份-

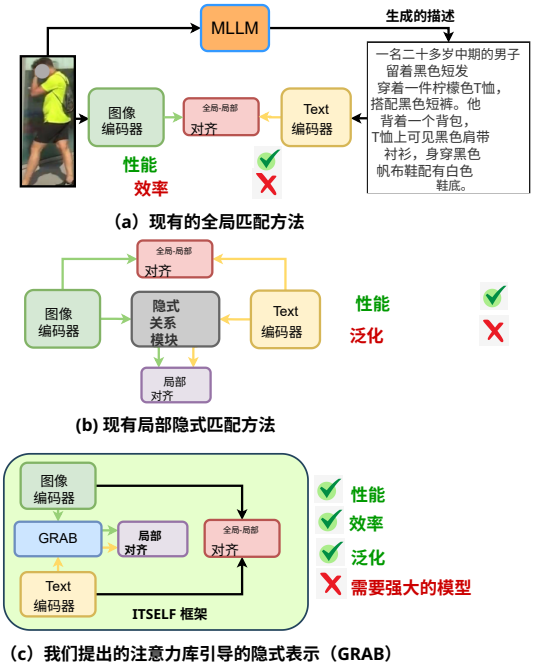


图 1. 基于文本的人物搜索范式的演进。(a) 全局匹配方法利用强大的多模态大语言模型合成额外的数据集。(b) 近期的局部隐式匹配方法对所有局部词元之间的关系进行隐式推理。(c) 我们的方法——采用 GRAB 的 ITSELF：一种由注意力引导的局部分支，通过隐式学习细粒度的判别性特征，实现更好的对齐。

利用图像和文本中的判别性线索，区分细微差异的个体。视觉语言模型（VLMs） [17, 22, 23, 36, 48], ， 尤其是 CLIP [26], 的最新进展，展现出应对这些细粒度挑战的巨大潜力。在此基础上，TBPS-CLIP [3] 率先将 CLIP 应用于基于文本的行人搜索（TBPS），随后出现的扩展方法 [13, 19, 21, 24, 43, 53] 进一步缩小了文本—图像鸿沟。然而，许多近期方法[20, 32, 39, 46, 52, 53] 依赖成本高昂的外部资源。例如，使用多模态大语言模型生成辅助数据（图1（a））虽然有效，却会增加计算和标注

*作者对本文贡献相同。

成本，并阻碍可扩展性和鲁棒性。

为说明这一局限，我们提出一个简单的问题：**TBPS模型如何在无需代价高昂的外部监督的情况下，自行捕捉细粒度的判别性细节？** 为尽量解决这一问题，近期研究致力于 隐式局部对齐 (**图 1(b)**，以探索判别性线索，而无需繁重的外部监督。嵌入空间对应关系方法 [9, 34, 38, 45]从隐式信号推断区域—短语匹配，但稀疏标签使这些对齐受到的约束较弱且不稳定。完全隐式的特征学习 [28, 31, 42] 优化局部损失，却忽视了特定文本—区域对的语义，因此无法保证对应关系的精确性。掩码建模式对齐 [11, 13]通过跨模态重建促进定位，但全局上下文捷径可能绕过真正的依赖关系。尽管研究路径各异，核心问题依然存在：局部性约束仍然太弱，导致监督信号分散、判别性区域选择能力减弱，并限制了稳健的局部特征学习。

注意力非常适合凸显细粒度线索并加强跨模态对应关系，但其在基于文本的行人搜索（TBPS）中的潜力仍未得到充分探索。我们通过一项简单的诊断实验来探究这一点：注意力引导的保留掩码（**图2**）。对于每张图像，我们计算图像编码器最后一层的注意力，保留排名前k个图像块，并屏蔽其余部分。随后，我们在 RSTP-Reid [51]上，针对多个保留比例，测量训练早期各轮中原始输入与对应输入之间的 R1 准确率差距。分析揭示了两个稳定的规律。首先，显著性很早便显现：到第 3 轮时，所有保留设置下的 R1 差距均低于 1 个百分点，表明保留的图像块已捕获几乎所有判别性证据。其次，注意力具有精确的空间定位能力：选中的图像块始终与语义明确的身体部位及携带物体相对应，为判别性区域与短语之间的对应关系提供了可靠的定位线索。

基于这些发现，我们提出*ITSELF*，一种用于基于文本的行人搜索（TBPS）中隐式局部对齐的注意力引导框架。ITSELF 优化全局文本—图像对齐损失，并结合注意力库引导表征（GRAB）加以增强。GRAB 是一种注意力驱动的局部分支，通过选择判别性标记来构建注意力库利用模型自身的注意力从两种模态中提取信息，无需额外监督（**图1(c)**）。与以往将局部对齐完全隐式处理、不给短语—区域对应关系施加直接约束，因而容易引发捷径学习和虚假相关性并导致错位的方法不同，我们的方法注入一种由注意力导出的局部性先验，将学习聚焦于信息量最大的区域，并在不同模态间一致地对齐这些区域，从而抑制噪声并强化全局对齐。从本质上说，我们的关键创新是一种由注意力驱动的隐式局部性机制，它将内部显著性图

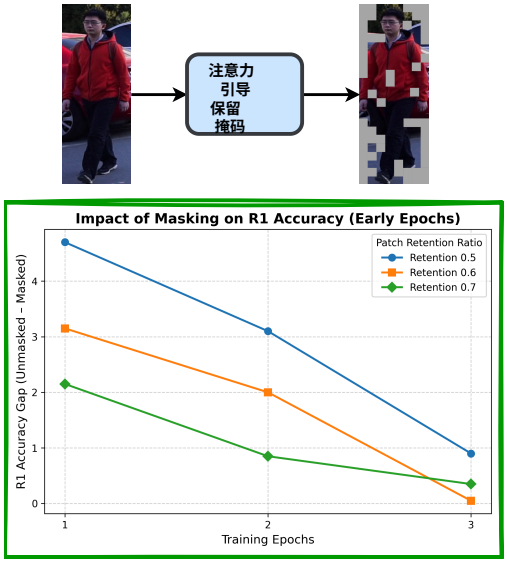


图2. 在 RSTP数据集上训练初期的不同掩码保留率下，未掩码图像与掩码图像之间的 Rank-1 准确率差距。

成为细粒度对齐的可靠锚点。

基于相关分析 [1, 4, 37] 表明，不同注意力头专注于互补线索，并在不同深度呈现出各异的模式。我们在*GRAB*中引入多层注意力鲁棒选择（*MARS*），以避免反复从单张注意力图中选取相同的主导词元。MARS 聚合不同层的注意力，并在不同模态间进行兼顾多样性的前k选择，确保互补覆盖。选出的词元填入 *GRAB* 的注意力库，其中的局部目标可强化模态间结构和模态内结构。

如（**图 2**）所示，未掩码输入与经注意力保留的输入之间的 R1差距在训练初期迅速缩小。受此启发，我们进一步引入自适应令牌调度器 （ATS），在训练初期维持较大的保留预算，以避免丢弃重要线索及由此导致的训练不稳定性，随后逐步降低预算，使模型聚焦于高置信度的细粒度词元。该调度策略可减少冗余和假阴性，并稳定局部学习。最后，按照近期的做法，我们采用 *CLIP* [26]作为骨干网络，使 ITSELF 能够迁移预训练知识，同时继续在基于文本的行人搜索（TBPS）任务上学习跨模态的隐式局部对应关系。总而言之，我们的主要贡献如下：

- **ITSELF 框架：** 一种新颖的注意力引导的隐式局部对齐框架*ITSELF*，借助*GRAB* 利用编码器注意力挖掘细粒度线索，并在无需额外监督的情况下强化全局对齐。
- **鲁棒选择与调度：**我们提出了*MARS*，该方法融合跨层注意力并执行

具有多样性感知的前-k 选择；以及*ATS*，该方法在训练过程中将保留预算从粗粒度逐步退火至细粒度，以稳定学习并防止过早的信息损失。

- **有力的实证结果：** 大量实验表明，我们的方法在 3 个广泛使用的基于文本的行人搜索基准上达到了最先进水平，并提升了跨数据集泛化能力，证实了该方法的有效性和鲁棒性。

2. 相关工作

2.1. 基于文本的行人搜索（TBPS）

过去几年，计算机视觉领域对基于文本的行人搜索表现出了浓厚兴趣 [2, 3, 13, 24, 25]。随着 [14, 26],等视觉-语言预训练技术的兴起，基于文本的行人搜索研究日益采用大规模预训练，以获得更强的跨模态表征。近期的一类工作通过引入辅助信号来提升基于文本的行人搜索性能。例如，一些方法利用人体解析或姿态估计 [40, 41] 来突出语义区域，另一些方法则采用外部行人重识别数据集[6, 29, 33, 40, 52, 53] 以更好地适应行人领域。这些策略能够改善细粒度对齐，但也会带来额外的训练成本、标注依赖或领域偏差。相比之下，我们的方法无需依赖外部数据集或工具，即可自主提取并对齐两种模态中的细粒度局部特征，从而有效弥合基于文本的行人搜索中的粒度与信息差距。

2.2. 基于文本的行人搜索的局部对齐

以往研究通过显式或隐式方法增强细粒度对齐。显式方法利用外部线索，例如人体解析网络 [40, 41]或大规模预训练 [35, 46]。然而，它们对大量外部监督、额外标注和计算资源的依赖，往往限制了其泛化能力。相比之下，隐式方法通过网络直接学习局部对应关系，无需外部数据。虽然这消除了对标注的依赖，但由于文本描述与特定图像区域之间的关系未得到显式约束 [13, 31, 44]，这些方法通常存在语义定位能力较弱的问题。因此，学习到的表征是否真正捕捉了细粒度跨模态细节，仍不确定。我们的方法建立在隐式的无标注方法之上，但引入了一个关键区别。我们不使用外部模型，也不学习语义定位能力较弱的特征，而是直接利用 CLIP 内部固有的注意力图。通过挖掘多个层中的细粒度线索，并选择信息量最大的区域，我们的方法生成了更具判别性的表征。这种轻量级设计无需额外监督或预训练，即可改善局部对齐。

3. 方法

本节概述我们提出的框架 ITSELF，见第3.1节。随后，我们将在第3.2节详细介绍核心机制 GRAB，该机制融合了 MARS 和 ATS。最后，第3.3 节将介绍整体流程的训练策略和推理过程。

3.1. ITSELF 框架

我们的框架由三个主要组件构成图 3：(a) 图像编码器 f_v，用于将图像编码为嵌入；(b) 文本编码器 f_t，用于根据图像描述生成文本嵌入；以及 (c) GRAB（利用注意力库进行引导式表征），它利用模型自身的注意力构建一个由高显著性词元组成的注意力库。随后在该库上应用局部目标，使模型无需额外监督即可学习细粒度对应关系。遵循先前的研究 [13, 24, 25], ，我们采用 CLIP ViT-B/16 作为视觉和文本模态的骨干网络。

图像编码器：给定输入图像 $I_i \in V$，我们将其划分为 $N = H \times W/P^2$ 个大小为 P 的不重叠图像块，将其展平并投影到 D 维空间，然后在序列前添加一个可学习的 [CLS] 词元及位置嵌入。该序列输入 Transformer 编码器，得到视觉嵌入 $\mathcal{V}_i = f_v(I_i) = \{v^i_{\text{global}}, v^i_{local}\} \in \mathbb{R}^{(1+N)\times D}$，其中 $v^i_{global} = v^i_{\text{cls}}$ 是全局嵌入，$v^i_{local} = \{v^i_j\}_{j=1}^N$ 是图像块嵌入。

文本编码器：对于文本，我们采用 CLIP 的 Transformer 编码器。给定图像描述 $T_i \in T$，先使用字节对编码进行分词，并在前后分别添加 [SOS]/[EOS]词元。序列经过嵌入后输入 Transformer，生成 $\mathcal{T}_i = f_t(T_i) = \{t^i_{\text{global}}, t^i_{\text{local}}\} \in \mathbb{R}^{(L+2)\times D}$，其中$t^i_{\text{global}} = t^i_e$（来自 [EOS]）是全局嵌入，而$t^i_{\text{local}} = \{t^i_j\}_{j=1}^L$ 是词元级嵌入。[SOS]词元 t^i_s 会被保留，但不参与使用。

3.2. 带注意力库的引导表示

3.2.1. 多层鲁棒选择注意力

为学习稳健的隐式局部表示，我们设计了 GRAB，它保留了一组多样且具有高度判别性的词元。在我们的研究发现基础上图2，具有持续较高注意力值的词元，即使在最早的训练轮次中，也能编码核心身份线索。然而，仅基于单个固定层选择词元存在固有的次优性，因为不同的 Transformer 层捕获不同类型的信息：浅层强调低层纹理，中间层捕获更广泛的上下文，而深层编码可能会抑制细粒度细节的语义抽象。为克服这一局限，我们引入**用于稳健选择的多层注意力（MARS）**，以聚合注意力信息

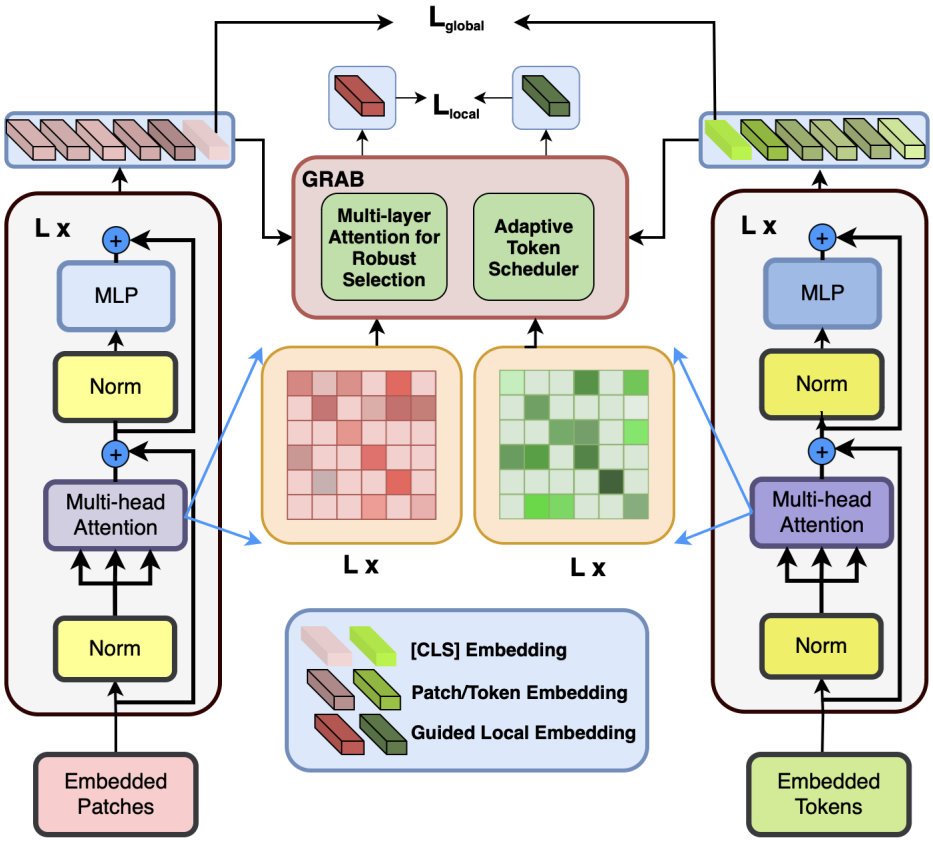


图 3. **我们提出的 ITSELF 概述**（一种注意力引导的隐式局部对齐框架）。该架构采用双流编码器，分别处理图像（左）和文本（右）。其核心是 GRAB（带注意力库的引导表示）模块，旨在学习细粒度的判别性线索。GRAB 由两个关键组件构成：MARS（多层鲁棒选择注意力），通过融合各层的注意力来选择信息丰富的图像块/词元；以及 ATS（自适应令牌调度器），在训练过程中将词元选择从粗略逐步调整为精细。模型采用双损失策略进行优化：局部损失 L_{local} 对齐引导式局部表示，全局损失 L_{global} 匹配最终的整体嵌入。这使 ITSELF 无需额外监督，也不会增加推理时开销，即可强化全局文本-图像对齐。

跨多个层，从而对图像块重要性提供更加稳定可靠的估计。形式化地，给定来自所选层的注意力图 $\mathbf{A}^{(\ell)} \in \mathbb{R}^{N\times N}$ $\ell \in \mathcal{L}$，我们通过移除注意力权重中最低的 δ_ℓ 比例进行去噪，从而滤除非信息性连接。随后，将剪枝后的注意力图与单位矩阵$\mathbf{I}$ 结合，以保留自依赖关系，并进行归一化：

$$\hat{\mathbf{A}}^{(\ell)} = \mathrm{Norm}\left(\frac{\mathrm{Discard}\left(\mathbf{A}^{(\ell)}, \delta_\ell\right) + \mathbf{I}}{2}\right), \qquad (1)$$

其中，Discard$(\cdot,\ \delta_\ell)$ 将最低的 δ_ℓ 比例元素置零，同时保留其余元素。

随后，通过顺序组合计算选定层之间的聚合注意力：

$$\mathbf{R} = \prod_{\ell\in\mathcal{L}} \hat{\mathbf{A}}_b^{(\ell)}, \quad \mathbf{R} \in \mathbb{R}^{N\times N}. \qquad (2)$$

基于聚合图 $\mathbf{R}$，我们通过选择信息量最大的词元获得局部特征$\mathbf{f}_{\mathrm{loc}i}$、$\mathbf{f}_t^{\mathrm{loc}}$：

$$\mathbf{f}_i^{\mathrm{loc}},\ \mathbf{f}_t^{\mathrm{loc}} = \mathrm{TopK}\left(\mathbf{R}_i, \mathbf{R}_t, v_{local}^i, t_{local}^i\right), \qquad (3)$$

其中$\mathbf{R}_i$ 和$\mathbf{R}_t$ 分别表示图像和文本模态的聚合注意力图。

通过根据$\mathbf{R}$对多样性感知的 top-k 个令牌进行排序和选择，MARS 隐式地捕获细粒度特征、增强判别信号、抑制背景噪声、强化全局嵌入，并生成更稳健、身份感知的特征空间。

3.2.2. 自适应令牌调度器

为进一步强化局部特征学习，我们引入自适应令牌调度器 （ATS）。与固定的词元预算不同，固定预算可能会保留过多背景信息，或过早丢弃有用细节；ATS逐渐降低

算法 1 GRAB：隐式局部对齐

要求： 图块/词元 特征 v^i_{local}, v^t_{local}; 注意力图 $\{\mathbf{A}^{(\ell)}\}$; 适配器(·)、GPO(·)；损失 $\mathcal{L}_{\text{local}}(\cdot,\ \cdot)$
确保： 局部对齐损失 $\mathcal{L}$
1: **聚合注意力（MARS）：** $\mathbf{R}_i$, $\mathbf{R}_t \leftarrow$ 计算得出 根据公式（2）计算。
2: **词元预算（ATS）：** $k \leftarrow$ 根据公式（4）计算。
3: **选择信息性词元：** $(\mathbf{f}_{\text{loc}i},\ \mathbf{f}_t^{\text{loc}}) \leftarrow$ 计算得出 根据公式（3），使用$\mathbf{R}_i$, $\mathbf{R}_t$ 和 k计算。
4: **局部适配（残差）：** $(\mathbf{f}_{\text{loc}i},\ \mathbf{f}_t^{\text{loc}}) \leftarrow$ 适配器$(\mathbf{f}_{\text{loc}i},\ \mathbf{f}_t^{\text{loc}}) + (\mathbf{f}_i^{loc},\ \mathbf{f}_t^{loc})$。
5: **池化：** $\mathbf{v} \leftarrow \text{GPO}(\mathbf{f}_i^{\text{loc}}),\quad \boldsymbol{\tau} \leftarrow \text{GPO}(\mathbf{f}_t^{\text{loc}})$。
6: **损失：** $\mathcal{L} \leftarrow \mathcal{L}_{\text{local}}(\mathbf{v}\ \boldsymbol{\tau})$。
7: **返回** $\mathcal{L}$。

训练过程中所选词元的数量。这种设计遵循由粗到细的范式：在训练早期，保留较大比例的词元，以避免丢失关键的身份线索；在后期，选择则逐渐聚焦于判别性较强的图像块。形式上，在步骤 t保留的词元数量定义为：

$$k_t = \begin{cases} \left\lfloor N\,\rho_{\text{start}} \left(\dfrac{\rho_{\text{end}}}{\rho_{\text{start}}}\right)^{\frac{t}{T}} \right\rfloor, & \text{if } t \le T, \\ \lfloor N\,\rho_{\text{end}} \rfloor, & \text{if } t > T, \end{cases} \tag{4}$$

其中， N 是词元数量， ρ_{start} 和 ρ_{end} 分别是初始和最终保留比例， t 是当前训练步数， T 是调度长度。这种逐渐缩小的方式可缓解早期信息损失并稳定训练，同时最终突出细粒度信号，以补充全局嵌入。

3.2.3. 隐式局部学习对齐

算法描述。为了用来自两种模态的注意力库填充最具判别性的线索，我们根据 MARS 返回的聚合注意力 $\mathbf{R}$ 为每个图像块/词元评分。在训练初期，为了稳定优化并避免遗漏显著区域，我们使用 ATS 安排逐步衰减的词元预算 k，使其随步数递减。每一步中，我们选取k 个在$\mathbf{R}$ 中注意力最高的图像块/词元，并将其加入注意力库。为了弥合模态偏移并细化局部特征，同时保留原始信号，选中的词元会经过一个轻量级的适配器（MLP），并通过残差连接进行处理。随后，我们应用GPO [5]以获得引导式局部嵌入$\mathbf{v}\ \boldsymbol{\tau}$，并计算局部对齐损失$\mathcal{L}_{\text{local}}$。完整流程总结于**算法1**。

3.3. 训练与推理

训练。为了优化全局外观和细粒度判别性词元，我们的方法结合了三元组对齐损失（TAL） [25] 和跨模态身份损失（CID） [53]。通过将这些专门的损失分别应用于全局和局部嵌入，模型能够学习更全面、更稳健的表征，以匹配文本描述与行人图像。

给定一个图像-文本对（I, T），损失定义如下：

$$\mathcal{L} = \mathcal{L}_{tal} + \mathcal{L}_{cid} \tag{5}$$

我们将此损失应用于全局嵌入（v_{global}, t_{global}）和引导式局部表示（$\mathbf{v}\ \tau$），得到

$$\mathcal{L}_{global} = \mathcal{L}(v_{global}, t_{global}), \mathcal{L}_{local} = \mathcal{L}(\mathbf{v}, \tau) \tag{6}$$

整体训练目标是两者之和：

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \mathcal{L}_{local} \tag{7}$$

推理。在推理过程中，最终的图文对相似度通过结合全局相似度和局部相似度计算得出。具体计算公式如下：

$$\mathcal{S} = \lambda_S \times \mathcal{S}_{global} + (1 - \lambda_S) \times \mathcal{S}_{local} \tag{8}$$

其中， $\mathcal{S}_{global}$ 和 $\mathcal{S}_{local}$ 分别表示全局相似度和局部相似度， λ_S 为权重因子。

4. 实验

4.1. 实验设置

数据集。 我们在广泛用于文本行人检索任务的公开数据集上评估所提出方法的有效性，包括 CUHK-PEDES [15], ICFG-PEDES [7], 和 RST-PReid [50]。关于这些数据集的更多详情，请参阅补充材料。

评估指标。 在评估中，我们在全部三个数据集上采用广泛使用的 Rank-k 准确率（k = 1、5、10）和平均精度均值（mAP）指标。

4.2. 定量结果

与最先进方法的比较。我们展示了在三个广泛使用的基准数据集上与最先进水平方法的比较结果**表 1**。我们的方法表现突出，在所有指标上均优于基于 CLIP 的竞争方法。在以 CLIP 为骨干网络的方法中，我们在所有数据集上都取得了新的最先进水平，**R@1** 较此前最强方法分别提升 **+1.01%**, **+1.55%**，以及 **+1.95%**；此外，我们在每个基准测试上都取得了最佳**平均精度均值**，其中在**RSTP-Reid** （**+2.17%** 平均精度均值）上的提升最大。这些优势在更高排名下依然保持，**R@5/R@10** 均达到最佳或并列最佳，表明检索性能提升广泛，而非仅限于单一指标。

方法	会议场所	图像编码器	文本编码器	CUHK-PEDES				ICFG-PEDES				RSTP-Reid			
				R@1	R@5	R@10	平均精度均值	R@1	R@5	R@10	平均精度均值	R@1	R@5	R@10	mAP
采用 ALBEF 骨干网络:															
RaSa [2]	IJCAI'23	Swin-B	BERT	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29	65.20	84.05	89.85	50.14
MARS [8]	TOMM'25	Swin-B	BERT	77.62	90.63	94.27	71.71	67.60	81.47	85.79	44.93	67.55	86.65	91.35	52.92
通过使用外部工具或 ReID领域预训练:															
UniPT [30]	ICCV'23	ViT-B/16	BERT	68.50	84.67	90.38	-	60.09	76.19	82.46	-	51.85	74.85	82.85	-
CFAM（B/16）[53]	CVPR'24	CLIP-ViT	CLIP-Xformer	72.87	88.61	92.87	64.92	62.17	79.57	85.32	36.34	59.40	81.35	88.50	46.04
SAP-SAM [40]	MM'24	CLIP-ViT	CLIP-Xformer	75.05	89.93	93.73	-	63.97	80.84	86.17	-	62.85	82.65	89.85	-
CFAM(L/14) [53]	CVPR'24	CLIP-ViT	CLIP-Xformer	75.60	90.53	94.36	67.27	65.38	81.17	86.35	39.42	62.45	83.55	91.10	49.50
DP [33]	AAAI'24	CLIP-ViT	CLIP-Xformer	75.66	90.59	94.07	66.58	65.61	81.73	86.95	39.14	62.48	83.77	89.93	48.86
APTM [46]	MM 2023	Swin-B	BERT	76.53	90.04	94.15	66.91	68.51	82.99	87.56	41.22	67.50	85.70	91.45	52.56
AUL [16]	AAAI'24	Swin-B	BERT	77.23	90.43	94.41	-	69.16	83.32	88.37	-	71.65	87.55	92.05	-
CAMeL [47]	TIFS'25	SG-Former	BERT	77.24	91.80	95.16	68.32	68.70	83.11	88.32	41.58	68.50	87.40	92.70	53.61
采用 CLIP 骨干网络:															
CLIP (ViT-B/16) [26]	ICML'21	CLIP-ViT	CLIP-Xformer	65.73	86.39	92.01	61.97	56.23	74.29	81.62	30.85	56.67	78.09	86.62	42.85
CFine [44]	TIP'23	CLIP-ViT	BERT	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
CSKT [18]	ICASSP'24	CLIP-ViT	CLIP-Xformer	69.70	86.92	91.80	62.74	58.90	77.31	83.56	33.87	57.75	81.30	88.35	46.43
DM-Adapter [19]	AAAI'25	CLIP-ViT	CLIP-Xformer	72.17	88.74	92.85	64.33	62.64	79.53	85.32	36.50	60.00	82.10	87.90	47.37
IRRA [13]	CVPR'23	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
TBPSCLIP [3]	AAAI'24	CLIP-ViT	CLIP-Xformer	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83	61.95	83.55	88.75	48.26
BiLMa [10]	ICCV'23	CLIP-ViT	CLIP-Xformer	74.03	89.59	93.62	66.57	63.83	80.15	85.74	38.26	61.20	81.50	88.80	48.51
MUM [49]	AAAI'24	CLIP-ViT	CLIP-Xformer	74.25	89.83	93.58	66.15	65.62	80.54	85.83	38.78	63.40	83.30	90.30	49.28
CRUE [12]	TOMM'25	CLIP-ViT	CLIP-Xformer	74.91	90.11	93.90	68.93	64.88	80.90	86.36	41.82	63.15	82.80	90.20	50.20
PLOT [24]	ECCV'24	CLIP-ViT	CLIP-Xformer	75.28	90.42	94.12	-	65.76	81.39	86.73	-	61.80	82.85	89.45	-
RDE [25]	CVPR'24	CLIP-ViT	CLIP-Xformer	75.94	90.14	94.12	67.56	67.68	82.47	87.36	40.06	65.35	83.95	89.90	50.88
Ours	**WACV'26**	CLIP-ViT	CLIP-Xformer	**76.95**	**90.64**	**94.36**	**69.38**	**69.23**	**82.84**	**87.62**	**43.80**	**67.30**	**85.60**	**90.50**	**53.05**

表1. 三个数据集上基于文本的人物搜索方法的性能。

关键的是，我们通过挖掘多层注意力（MARS）来引导隐式局部对齐，无需 依赖行人重识别领域的预训练。尽管设置精简，我们仍超越了利用额外资源或更大骨干网络的方法（例如 CFAM(L/14)、DP、UniPT）。值得注意的是，仅使用 CLIP，我们就在所有方法中取得了 **R@1ICFG-PEDES** 上的最佳成绩。

域泛化。 我们进一步评估了模型的跨域鲁棒性：在一个源数据集上进行训练，然后直接在另一个未见过的目标数据集上测试，且不进行微调。我们使用 CUHK-PEDES（C）、ICFG-PEDES（I）和 RSTPReid（R）构建了六种迁移设置（例如，C→I、I→R）。现有的局部隐式匹配方法（如 IRRA）通常能够取得较强的域内性能，但泛化能力较差，主要原因在于其关系模块过拟合于数据集特定模式，例如服装风格或标注偏差。因此，它们学习到的对应关系难以有效跨域迁移。相比之下，我们的方法通过将注意力库引导与平衡的局部–全局对齐相结合，避免了此类过拟合，从而获得语义上更加稳定的表征。如 表2 所示，我们的方法在六种迁移设置中均取得了最佳结果。例如，在 C→I 场景中，我们获得了 50.58% 的 R@1 和 27.32% 的平均精度均值，分别超过 RDE **2.40%** 和 **2.32%**。同样，在具有挑战性的 I→R 迁移中，我们将此前最佳的 R@1 提高了超过 2%。这些结果表明，与以往的局部隐式方法不同，我们的模型同时实现了较高的性能和强大的域泛化能力——这对于具有多样且不断变化的数据分布的真实世界 TBPS 场景至关重要。

4.3. 定性结果

前五名检索示例。 图 4 定性展示了我们的方法在RSTPReid基准测试上优于RDE。对于一名身穿黑色夹克和蓝色羽绒服的男子这一查询，我们的方法检索到全部五个正确匹配，而RDE仅找到三个。对于一名身穿橙色外套、摆出手插口袋姿势的男子这一查询，我们的方法在排名靠前的结果中正确检索到三张图像，而RDE仅找到两张，并将第二个真正例排在第四位。对于一名身穿带白色图案的黑色夹克的男孩，我们的方法再次表现出色，前三名检索结果中有三张正确图像，而RDE仅在R3位置找到一个正确匹配。这些示例凸显了我们的框架在将细粒度文本描述与相关视觉细节对齐方面的能力得到增强。

注意力比较。 图 5展示了 Grad-CAM [27]对 RDE [25]与我们的方法进行比较的可视化结果。我们的方法通过关注查询特定元素（例如服装和配饰），在最大限度减少溢出的同时实现了更清晰的定位，从而有效地在多人场景中分离出目标行人。相比之下，RDE 产生的热点区域较为分散，且往往与任务无关，导致文本—图像关联碎片化。这些结果凸显了我们框架更强的属性级保真度以及更优的

	方法	R1	R5	R10	mAP
C→I	IRRA [13]	42.41	62.11	69.62	21.77
	CLIP [26]	43.04	-	-	22.45
	RDE [25]	48.18	66.30	73.70	25.00
	Ours	**50.58**	**67.81**	**74.68**	**27.32**
I→C	IRRA [13]	33.48	56.29	66.33	31.56
	CLIP [26]	33.90	-	-	31.65
	RDE [25]	38.11	59.24	68.44	34.16
	Ours	**41.05**	**63.30**	**72.08**	**37.79**
I→R	IRRA [13]	45.30	69.25	78.80	36.82
	CLIP [26]	47.45	-	-	36.83
	RDE [25]	49.25	72.10	80.20	38.46
	Ours	**51.30**	**73.30**	**80.40**	**41.07**
R→I	IRRA [13]	32.30	49.67	57.80	20.54
	CLIP [26]	33.58	-	-	19.58
	RDE [25]	42.17	58.32	65.49	26.37
	Ours	**43.32**	**59.17**	**66.16**	**27.57**
C→R	CLIP [26]	52.55	-	-	39.97
	IRRA [13]	53.25	77.15	85.35	39.63
	RDE [25]	54.90	77.50	86.50	41.27
	Ours	**58.05**	**79.30**	**86.85**	**43.72**
R→C	IRRA [13]	32.80	55.26	65.81	30.29
	CLIP [26]	35.25	-	-	32.35
	RDE [25]	36.94	58.22	67.58	33.65
	Ours	**40.32**	**61.70**	**71.04**	**36.65**

表 2. 与最先进方法的比较（领域泛化）。其中，“C”表示 CUHK-PEDES，“I”表示 ICFG-PEDES，“R”表示 RSTPReid。

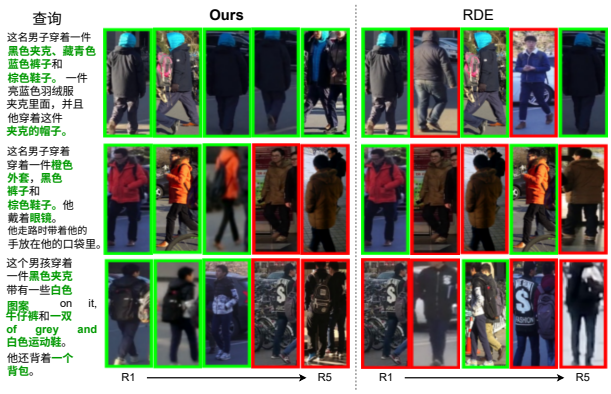


图 4. 在 RSTPReid 基准测试上进行文本到图像检索的定性结果，对比我们的方法与 RDE [25]。检索到的图像按相似度从高到低由左至右排列。正确匹配以绿色框标出，错误匹配则以红色框标出。绿色高亮的文本表示我们的方法有效捕捉到的描述性细节。

降低了跨身份混淆，这对现实世界中的行人搜索至关重要。

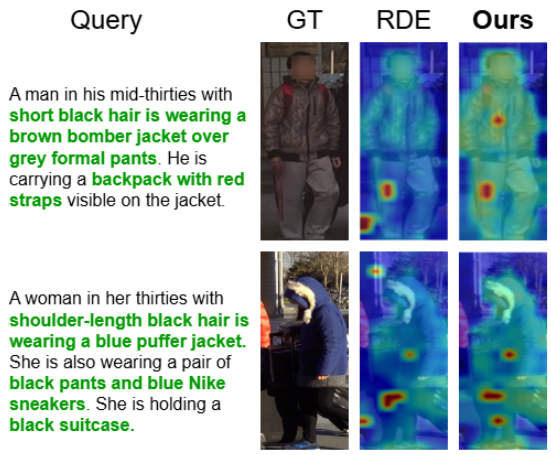


图 5. 对比 RDE [25] 与我们的方法通过 Grad-CAM 生成的注意力图。[27]

Top-K词元选择分析。 图12中的比较 表明，我们提出的 Top-K词元选择比基线更有效。基线生成的注意力图较为分散，且与查询不一致；而我们的方法生成了聚焦且语义相关的注意力图。该方法能够成功定位与关键词对应的图像区域，展现出将文本描述准确映射到正确视觉上下文中的更强能力，从而实现更准确的检索。

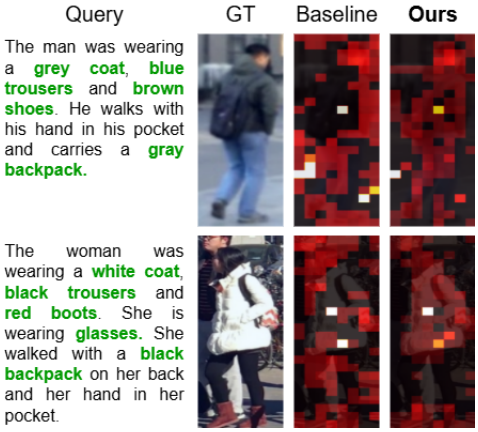


图 6. 基线方法与我们的方法在文本到图像检索中选取的 Top-K图像块对比。我们的方法能够突出显示与查询描述语义相关性更高的区域（例如服装颜色和配饰）。

4.4. 消融研究

各组件的有效性：为了全面评估我们所提

方法	GRAB			CUHK-PEDES				ICFG-PEDES				RSTPReid			
	SL	MARS	ATS	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
基线	✗	✗	✗	74.66	89.70	93.57	67.63	66.06	81.12	86.10	41.02	61.30	80.85	87.65	49.12
+SL	✓	✗	✗	76.49	90.20	94.35	58.97	67.93	82.36	87.12	42.38	65.00	85.45	90.15	52.26
+MARS	✗	✓	✗	76.90	90.46	94.35	**69.71**	69.07	82.72	87.51	43.67	66.95	85.15	90.40	**53.05**
+监督学习+ATS	✓	✗	✓	76.66	90.58	94.22	69.18	68.76	82.74	87.45	43.35	65.95	**85.60**	90.10	52.71
Ours	✗	✓	✓	**76.95**	**90.64**	**94.36**	69.38	**69.23**	**82.84**	**87.62**	**43.80**	**67.30**	**85.60**	**90.50**	**53.05**
Δ	-	-	-	+2.29	+0.94	+0.79	+1.75	+3.17	+1.72	+1.52	+2.78	+6.00	+4.75	+2.85	+3.93

表3. ITSELF 在三个数据集上各个组件的消融研究。

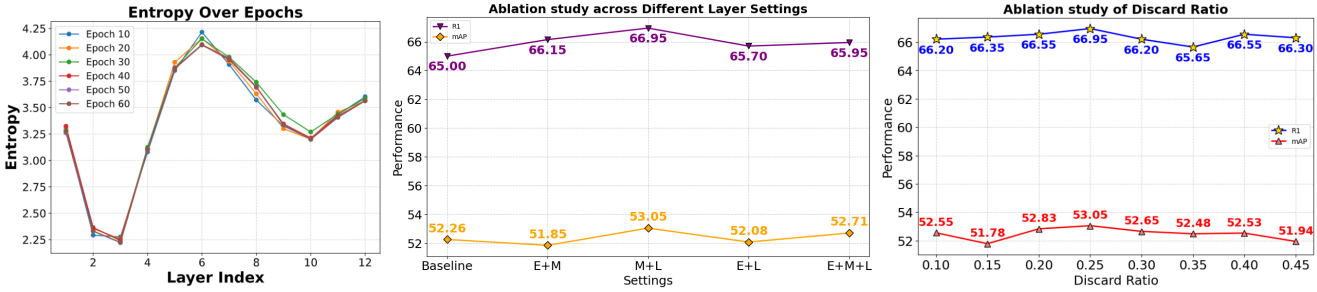


图7。（左）训练轮次中每个 CLIP 层的注意力熵。（中）为 MARS 选择的 CLIP 层消融实验（E：早期，M：中期，L：后期）。（右）MARS 丢弃比例的敏感性，在 R1 和平均精度均值上进行评估。

ITSELF 框架中各组件的贡献，我们在三个公开数据集上进行了系统的实证分析。详细实验结果见表 3。**MARS模块：**为评估 MARS 的有效性，我们首先将其与固定单层策略（SL）进行比较。根据注意力分布熵（图 7 （左）），我们选择第 3 层作为 SL，因为它的熵最低，表明置信度更高。虽然 SL 和 MARS 的表现都优于基线，但 MARS 在所有数据集上均持续取得更优结果，R1 分别提升了+2.24%、 +3.01% 和 +5.65%，充分体现了其相较于依赖固定 SL 的明显优势。**ATS模块：**在仅使用 SL 的基线上加入 ATS，R1 已能持续提升。在完整模型中结合 MARS 后，我们整体上取得了最佳性能，在三个数据集上相较基线分别有显著提升（+2.29%、+3.17%、 +6.00%）。这些结果证实，ATS 不仅能防止训练早期判别性线索的损失，还有助于实现更稳定的优化。

MARS 中层选择与丢弃比例分析：从图7中间的图可以看出，MARS 中的多层配置始终优于基线，其中中间层+后期层（*M*+*L*）组合在所有层类型组合（E、M、L 及其两两组合）中取得了最佳 R1 和平均精度均值。左图解释了原因：早期层注意力（尤其是第 3 层）的熵最低，这意味着注意力高度集中在少数低层级词元（边缘/纹理或背景）上，语义依据较弱，因此在融合时可能引入噪声。相比之下，中间层的熵更高，能够捕捉更广泛的上下文和关系；而后期层则会将注意力重新聚焦于显著的语义区域。因此，融合“中+后期层”可兼顾上下文覆盖与判别性聚焦，优于任何包含早期层注意力的方案。最后，右图显示，丢弃比例为 0.25 时性能最佳，这表明在融合过程中移除少量低注意力词元，可以滤除噪声并强化判别性线索。

5. 结论

本文提出了*ITSELF*，这是一种用于 TBPS隐式局部对齐 的新型注意力引导框架，可将 CLIP 的多层注意力转化为注意力库，无需额外监督，也不会增加推理时开销。在此基础上，*GRAB*通过隐式局部目标发掘并优化细粒度对应关系，以补充全局损失。为实现 GRAB，*MARS* 自适应地聚合并排序各层注意力，从而选出最具判别性的图像块/词元。同时，*ATS*从粗到细地安排词元保留预算，以减轻早期信息损失并稳定训练。在三个广泛使用的 TBPS 基准数据集上进行的大量实验表明，与基于 CLIP 的方法相比，我们的方法在所有指标上均达到**最先进方法**的性能，并提升了跨数据集泛化能力，证实了该方法的有效性、鲁棒性和实用性。

致谢

本研究得到越南国家大学胡志明市信息技术大学科学研究支持基金的资助。

补充材料

A. 实验细节

数据集。 我们在三个广泛使用的文本到图像行人检索基准测试上进行实验。

1. CUHK-PEDES [15]包含 40,206 张行人图像及与之配对的 80,412 条文本描述，涵盖 13,003 个身份；其中训练集、验证集和测试集分别包含 11,003、1,000 和 1,000 个身份。2. ICFG-PEDES [7]由 4,102 个个体身份对应的 54,522 组图文对组成，其中 34,674 组用于训练，19,848 组用于测试。3. RSTPReid [50] 包含 20,505 张图像，涉及 4,101 个个体身份；每个身份有 5 张图像，且每张图像均配有对应的两条人工标注文本描述。

实现细节 为与先前工作进行公平比较，我们使用预训练的 CLIP-ViT/B-16 [26] 模型初始化模态特定编码器，该模型版本与 IRRA [13]所用版本相同。为增加数据多样性，我们对图像采用随机水平翻转、随机裁剪和随机擦除，并对文本词元进行随机掩码、替换和移除。输入图像调整为 384 ×128，文本最大长度设为 77 个词元。我们使用 Adam优化器训练模型 60 个周期，初始学习率为 1×10^{-5}，并采用余弦学习率调度器。批次大小为 256，温度参数 τ 设为 0.015。MARS 的超参数设为中间层和后层，丢弃比例设为 0.25。归一化策略采用 L1归一化。对于 ATS，p start 和 p end 的值分别设为 0.65 和 0.5。我们将 t small 值设为基线达到最佳结果时所对应的周期。

B. 更多定量结果

为了进一步验证我们的方法，我们进行了大量定量实验和消融研究。我们分析了自适应令牌调度器（ATS）的影响，并在**表4** 中展示了逐步应用能够取得最佳效果。我们还在**表5**中评估了我们的方法在不同 CLIP骨干网络上的泛化能力，证实其性能始终优于基线。最后，我们在**表6**中详细比较了 MARS选择策略与几种基于启发式的方法，结果证实了我们所提方法的优越性。

设置	R@1	R@5	R@10	mAP
基线	65.00	85.45	90.15	52.26
轮次（ATS）	65.25	83.85	89.95	51.39
步数（ATS）	**65.95**	**85.70**	90.10	**52.71**

表 4. 自适应令牌调度器（ATS）效果的消融研究。基线不使用调度器，而 ATS 则应用于轮次级或步数级。结果表明，步数级调度性能最佳，与基线相比，R@1、R@5 和平均精度均值均有所提升。

设置	R@1	R@5	R@10	mAP
使用 ViT/B-16 CLIP 作为骨干网络				
基线	61.30	80.85	87.65	49.12
Ours	**67.30**	**85.60**	**90.50**	**53.05**
使用 ViT/B-32 CLIP 作为骨干网络				
基线	59.65	79.35	86.35	47.40
Ours	**64.50**	**84.10**	**90.40**	**50.28**

表5. 不同 CLIP骨干网络上的消融研究。无论应用于 ViT/B-16 还是轻量级 ViT/B-32 骨干网络，我们的方法都能持续提升性能，优于基线。值得注意的是，即使使用规模较小的 ViT/B-32，我们的方法在 R@1 和平均精度均值方面也取得了明显提升，证明其适用于不同模型容量。

设置	R@1	R@5	R@10	mAP
A	60.25	79.70	88.10	48.27
B	66.25	84.45	90.20	52.29
C	60.95	81.40	87.60	48.99
D	65.65	84.00	89.95	51.58
MARS	**66.95**	**85.15**	**90.40**	**53.05**

表 6. 基于注意力统计选择 Top-K 图像块的不同策略的消融研究。我们将方法（MARS）与四种基线策略进行比较：选择平均注意力最低的图像块（A）、平均注意力最高的图像块（B）、注意力标准差最小的图像块（C）以及注意力标准差最大的图像块（D）。结果清楚地表明，我们的 MARS 方法在所有评估指标上均优于其他基线方法。MARS 的表现最佳，R@1 达到 66.95%，平均精度均值达到 53.05%。这凸显了与仅依据注意力分数的平均值或标准差等简单启发式策略相比，我们的选择策略更加有效。

C. 更多定性结果

更多检索结果。**图8**、**图9**、**图10和图11**展示了 CUHK-PEDES、ICFG-PEDES 和 RSTPReid 基准测试中的更多定性比较

。这些示例一致表明，即使在涉及细粒度属性、小型配饰和视觉上相似的干扰项等具有挑战性的情况下，我们的方法也能检索到视觉和语义上准确的匹配结果。与基线相比，RDE [25], 以及 IRRA 等其他强劲方法 [13]和 TBPSCLIP [3], 相比，我们的方法在捕捉服装纹理、颜色搭配以及随身物品（如背包、手提包或手机）等细微线索方面展现出更强的鲁棒性。值得注意的是，我们的模型在领域偏移下依然可靠，能够应对从拥挤的街景到低光照图像等多种场景。这些结果进一步验证了我们的框架在生成更具判别性和泛化能力的文本-图像对齐方面的有效性。

更多注意力图可视化。图12 展示了我们的方法与 RDE 在 RSTPReid基准上的注意力图定性比较结果。结果表明，我们的模型能够始终关注文本查询中描述的、更具判别性且语义相关的区域，例如特定的服装颜色、配饰和携带物品（如背包、包或自行车）。相比之下，RDE 往往产生分散或错位的注意力，无法捕捉细粒度线索。这些可视化结果进一步凸显了我们的方法利用文本引导定位有意义的视觉区域的有效性，从而实现更准确的基于文本的人物检索。

参考文献

[1] Samira Abnar 和 Willem Zuidema。量化 Transformer 中的注意力流。载于《第58届计算语言学协会年会论文集》，第4190–4197页，线上，2020年。计算语言学协会。2[2] Yang Bai、Min Cao、Daming Gao、Ziqiang Cao、Chen Chen、Zhenfeng Fan、Liqiang Nie 和 Min Zhang。RaSa：用于基于文本的人物搜索的关系与敏感性感知表征学习。arXiv 预印本 arXiv:2305.13653，2023年。3，6[3] Min Cao、Yang Bai、Ziyin Zeng、Mang Ye 和 Min Zhang。用于基于文本的人物搜索的 CLIP 实证研究。载于《AAAI 人工智能会议论文集》，第465–473页，2024年。1，3，6，10[4] Hila Chefer、Shir Gur 和 Lior Wolf。超越注意力可视化的 Transformer 可解释性，2021年。2[5] Jiacheng Chen、Hexiang Hu、Hao Wu、Yuning Jiang 和 Changhu Wang。学习视觉语义嵌入的最佳池化策略，2021年。5[6] Haiwen Diao、Bo Wan、Ying Zhang、Xu Jia、Huchuan Lu 和 Long Chen。UniPT：面向迁移学习的通用并行调优，实现高效的参数与内存利用。arXiv 预印本 arXiv:2308.14316，2023年。3[7] Zefeng Ding、Changxing Ding、Zhiyin Shao 和 Dacheng Tao。用于文本到图像部件感知行人重识别的语义自对齐网络。arXiv 预印本 arXiv:2107.12666，2021年。5，9

[8] Alex Ergasti、Tomaso Fontanini、Claudio Ferrari、Massimo Bertozzi 和 Andrea Prati。《Mars：在基于文本的人物搜索中更加关注视觉属性》。arXiv预印本 arXiv:2407.04287，2024 年。6[9] Ammarah Farooq 、Muhammad Awais、Josef Kittler 和 Syed Safwan Khalid。《Axm-net：用于行人重识别的隐式跨模态特征对齐》。发表于 AAAI人工智能会议论文集，第4477–4485页，2022年。2 – 。[10] Takuro Fujii 和 Shuhei Tarashima。《BiLMa：用于基于文本的行人重识别的双向局部匹配》。发表于 IEEE/CVF国际计算机视觉会议论文集，第2786–2790页，2023年。6[11] Takuro Fujii 和 Shuhei Tarashima。《BiLMa：用于基于文本的行人重识别的双向局部匹配》，2023年。2[12] Shijuan Huang、Zongyi Li、Hefei Ling 和 Jianbo Li。《用于基于文本的人物搜索的跨模态关系与不确定性探索》。ACM多媒体计算、通信与应用汇刊，2025年。6[13] Ding Jiang 和 Mang Ye。《用于文本到图像人物检索的跨模态隐式关系推理与对齐》。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第2787–2797页，2023年。1, 2, 3, 6, 7, 9, 10[14] Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。《先对齐再融合：通过动量蒸馏进行视觉与语言表征学习》。神经信息处理系统进展，34：9694–9705，2021年。3[15] Shuang Li、Tong Xiao、Hongsheng Li、Bolei Zhou、Dayu Yue 和 Xiaogang Wang。《利用自然语言描述进行人物搜索》。发表于 IEEE计算机视觉与模式识别会议论文集，第1970–1979页，2017年。1, 5, 9 – 。[16] Shenshen Li、Chen He、Xing Xu、Fumin Shen、Yang Yang 和 Heng Tao Shen。《用于基于文本的人物检索的自适应不确定性学习》。发表于 AAAI人工智能会议论文集，第3172–3180页，2024年。6[17] Zongxia Li、Xiyang Wu、Hongyang Du、Fuxiao Liu、Huy Nghiem 和 Guangyao Shi。《大型视觉语言模型前沿综述：对齐、基准测试、评估与挑战》，2025年。1[18] Yating Liu、Yaowei Li、Zimo Liu、Wenming Yang、Yaowei Wang 和 Qingmin Liao。《基于CLIP的协同知识迁移，用于基于文本的人物检索》。发表于 ICASSP 2024—2024年 IEEE国际声学、语音与信号处理会议（ICASSP），第7935–7939页。IEEE，2024年。6[19] Yating Liu、Zimo Liu、Xiangyuan Lan、Wenming Yang、Yaowei Li 和 Qingmin Liao。《DM-Adapter：用于基于文本的人物检索的领域感知适配器混合》。发表于 AAAI人工智能会议论文集，第5703–5711页，2025年。1, 6[20] Tien-Huy Nguyen、Quang-Khai Tran 和 Anh-Tuan Quang-Hoang。《通过自集成提升视觉推理的泛化能力》，2024年。1

[21] Tien-Huy Nguyen、Huu-Loc Tran、Huu-Phong Phan-Nguyen 和 Quang-Vinh Dinh。《混合、统一且迭代：一种用于基于文本的人物异常检索的新型框架》，2025年。1
[22] Tinh-Anh Nguyen-Nhu、Triet Dao Hoang Minh、Dat To-Thanh、Phuc Le-Gia、Tuan Vo-Lan 和 Tien-Huy Nguyen。《Ster-vlm：增强参考视觉语言模型的时空方法》，2025年。1[23]Tinh-Anh Nguyen-Nhu、Huu-Loc Tran、Nguyen-Khang Le、Minh-Nhat Nguyen、Tien-Huy Nguyen、Hoang-Long Nguyen-Huu、Huu-Phong Phan-Nguyen、Huy-Thach Pham、Quan Nguyen、Hoang M. Le 和 Quang-Vinh Dinh。《一种轻量级时刻检索系统：结合全局重排序与稳健的自适应双向时间搜索》，2025年。1[24] Jicheol Park、Dongwon Kim、Boseung Jeong 和 Suha Kwak。《PLOT：通过部件槽注意力发现对应部件的基于文本的人物搜索》。发表于 欧洲计算机视觉会议，第474–490页。施普林格，2024年。1, 3, 6[25] Yang Qin、Yingke Chen、Dezhong Peng、Xi Peng、Joey Tianyi Zhou 和 Peng Hu。《用于文本到图像行人重识别的噪声对应学习》。发表于 IEEE/CVF计算机视觉与模式识别会议论文集，第27197–27206页，2024年。3, 5, 6, 7, 10[26]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《通过自然语言监督学习可迁移视觉模型》。发表于 国际机器学习会议，第8748–8763页。PmLR，2021年。1, 2, 3, 6, 7, 9[27] Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。《Grad-CAM：通过基于梯度的定位实现深度网络的视觉解释》。发表于 IEEE国际计算机视觉会议论文集，第618–626页，2017年。6, 7 – 。[28] Zhiyin Shao、Xinyu Zhang、Meng Fang、Zhifeng Lin、Jian Wang 和 Changxing Ding。《学习用于文本到图像行人重识别的粒度统一表示》，2022年。2[29] Zhiyin Shao、Xinyu Zhang、Changxing Ding、Jian Wang 和 Jingdong Wang。《利用伪文本进行统一预训练，以实现文本到图像行人重识别》。发表于 IEEE/CVF国际计算机视觉会议论文集，第11174–11184页，2023年。3[30] Zhiyin Shao、Xinyu Zhang、Changxing Ding、Jian Wang 和 Jingdong Wang。《利用伪文本进行统一预训练，以实现文本到图像行人重识别》，2023年。6[31] Xiujun Shu、Wei Wen、Haoqian Wu、Keyu Chen、Yiran Song、Ruizhi Qiao、Bo Ren 和 Xiao Wang。《看得更细，看得更多：用于基于文本的人物检索的隐式模态对齐》，2022年。2, 3[32] Zifan Song、Guosheng Hu 和 Cairong Zhao。《Diverse Person：定制专属的基于文本的人物搜索数据集》。AAAI人工智能会议论文集，38(5):4943–4951，2024年。1[33] Zifan Song、Guosheng Hu 和 Cairong Zhao。《Diverse Person：定制专属的基于文本的人物搜索数据集搜索》。在《AAAI人工智能会议论文集》中，页码 4943–4951，2024年。3, 6[34] Wei Suo、Mengyang Sun、Kai Niu、Yiqi Gao、Peng Wang、Yanning Zhang 和 Qi Wu。用于基于文本的人物搜索的简单而稳健的相关性过滤方法。载于《计算机视觉——ECCV 2022》，页码 726–742，瑞士沙姆，2022年。施普林格·自然瑞士。2[35] Wentan Tan、Changxing Ding、Jiayu Jiang、Fei Wang、Yib- ing Zhan 和 Dapeng Tao。借助多模态大语言模型的力量实现可迁移的文本到图像人物重识别。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 17127–17137，2024年。3[36] Huu-Loc Tran、Tinh-Anh Nguyen-Nhu、Huu-Phong Phan- Nguyen、Tien-Huy Nguyen、Nhat-Minh Nguyen-Dich、Anh Dao、Huy-Duc Do、Quan Nguyen、Hoang M. Le 和 Quang- Vinh Dinh。迈向高效且稳健的时刻检索系统：多粒度模型与时间重排序的统一框架，2025年。1[37] Elena Voita、David Talbot、Fedor Moiseev、Rico Sennrich 和 Ivan Titov。分析多头自注意力：少数专门化注意力头承担了繁重工作，其余则可剪枝。载于《计算语言学协会第57届年会论文集》，页码 5797–5808，意大利佛罗伦萨，2019年。计算语言学协会。2[38] Chengji Wang、Zhiming Luo、Yaojin Lin 和 Shaozi Li。通过多粒度嵌入学习实现基于文本的人物搜索。载于《第三十届国际人工智能联合会议论文集》，IJCAI-21，页码 1068–1074。国际人工智能联合会议组织，2021 2。主会场。[39] Yihao Wang、Meng Yang 和 Rui Cao。借助迁移人物SAM实现基于文本的人物检索的细粒度语义对齐。载于《第32届 ACM 多媒体国际会议论文集》，页码 5432–5441，美国纽约，2024年。计算机协会。1[40] Yihao Wang、Meng Yang 和 Rui Cao。借助迁移人物SAM实现基于文本的人物检索的细粒度语义对齐。载于《第32届 ACM 多媒体国际会议论文集》，页码 5432–5441，2024年。3, 6[41] Zhe Wang、Zhiyuan Fang、Jun Wang 和 Yezhou Yang。Vitaa：通过自然语言实现人物搜索中的视觉文本属性对齐。载于《欧洲计算机视觉会议》，页码 402–420。施普林格，2020年。3[42] Zijie Wang、Aichun Zhu、Jingyi Xue、Daihong Jiang、Chao Liu、Yifeng Li 和 Fangqiang Hu。Sum：用于基于文本的人物检索的序列化更新与匹配。《知识型系统》，248：108891，2022年。2[43] Shuanglin Yan、Neng Dong、Liyan Zhang 和 Jinhui Tang。CLIP驱动的细粒度文本图像人物重识别，2022年。1[44] Shuanglin Yan、Neng Dong、Liyan Zhang 和 Jinhui Tang。CLIP驱动的细粒度文本图像人物重识别。《IEEE图像处理汇刊》，32 6046，2023 6032 3, 6 : – 。[45] Shuanglin Yan、Hao Tang、Liyan Zhang 和 Jinhui Tang。面向基于文本的人物搜索的图像特定信息抑制与隐式局部对齐，2023年。2

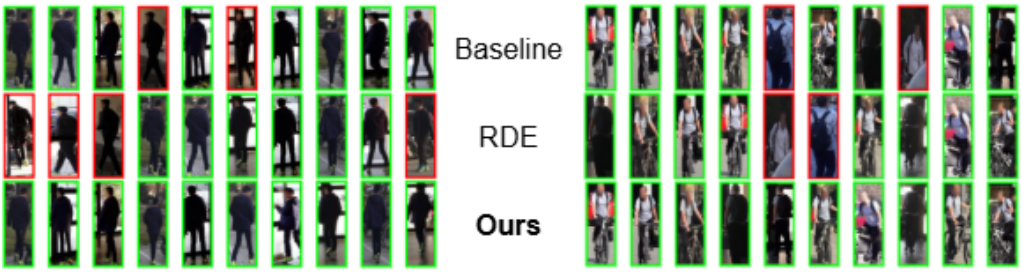
[46] Shuyu Yang、Yinan Zhou、Zhedong Zheng、Yaxiong Wang、Li Zhu 和 Yujiao Wu。迈向统一的基于文本的人物检索：大规模多属性与语言搜索基准。载于《第31届 ACM 多媒体国际会议论文集》，页码 4492–4501，2023年。1, 3, 6[47] Hang Yu、Jiahao Wen 和 Zhedong Zheng。Camel：面向基于文本的人物检索的跨模态自适应元学习。《IEEE信息取证与安全汇刊》，2025年。6[48] Jingyi Zhang、Jiaxing Huang、Sheng Jin 和 Shijian Lu。面向视觉任务的视觉语言模型：综述，2024年。1[49] Zhiwei Zhao、Bin Liu、Yan Lu、Qi Chu 和 Nenghai Yu。统一多模态不确定性建模与语义对齐，实现文本到图像行人重识别。载于《AAAI人工智能会议论文集》，页码 7534–7542，2024年。6[50] Aichun Zhu、Zijie Wang、Yifeng Li、Xili Wan、Jing Jin、Tian Wang、Fangqiang Hu 和 Gang Hua。DSSL：面向基于文本的人物检索的深度背景与行人分离学习。载于《第29届 ACM 多媒体国际会议论文集》，页码 209–217，2021年。5, 9[51] Aichun Zhu、Zijie Wang、Yifeng Li、Xili Wan、Jing Jin、Tian Wang、Fangqiang Hu 和 Gang Hua。DSSL：面向基于文本的人物检索的深度背景与行人分离学习，2021年。2[52] Jialong Zuo、Jiahao Hong、Feng Zhang、Changqian Yu、Hanyu Zhou、Changxin Gao、Nong Sang 和 Jingdong Wang。PLIP：用于人物表征学习的语言图像预训练。《神经信息处理系统进展》，37：45666–45702，2024年。1, 3[53] Jialong Zuo、Hanyu Zhou、Ying Nie、Feng Zhang、Tianyu Guo、Nong Sang、Yunhe Wang 和 Changxin Gao。UFineBench：迈向超细粒度的基于文本的人物检索。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，页码 22010–22019，2024年。1, 3, 5, 6



图8。CUHK-PEDES 基准上的更多示例

A middle-aged man with short black hair is wearing a **black insulated jacket with a hood**. He is also wearing **black fitted pants and neon green sneakers with a black Nike logo** on the sides.

A man in his mid 40's having a **bald patch**. He is wearing an off **white t-shirt and black trousers**. He is also holding a **black backpack.**



A man in his fifties with medium-length with a receding hair is wearing a **black bomber jacket**. He is also wearing a pair of **green chinos pants and blue sneakers.**

A young woman is wearing a persian **blue down jacket with white black checkered print** at the bottom and a hood having a fawn furry lining. She is also wearing **slim-fit black pants with dark and light blue shaded running shoes having a black lining** at the bottom and soles in white light blue colour.



A man with black short hair is wearing a **blue puffer jacket with a hood and a pair of fitted black pants**. He is wearing **green running shoes with white details** and his hands are inside the pocket.

A boy with black medium length hair is wearing a **dark blue hoodie that has green sleeves**. He is also wearing **grey denim loose fitted jeans with dark blue running shoes**. He has a **black and orange school bag** that has Mickey Mouse on it strapped on the shoulder.

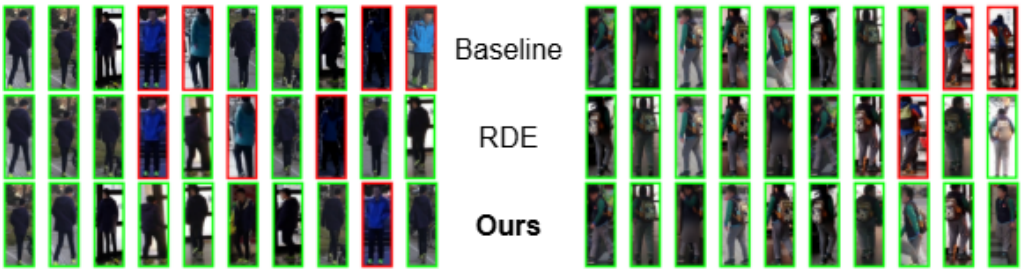


图9。ICFG-PEDES基准上的更多示例



图10。RSTPReid基准上的更多示例

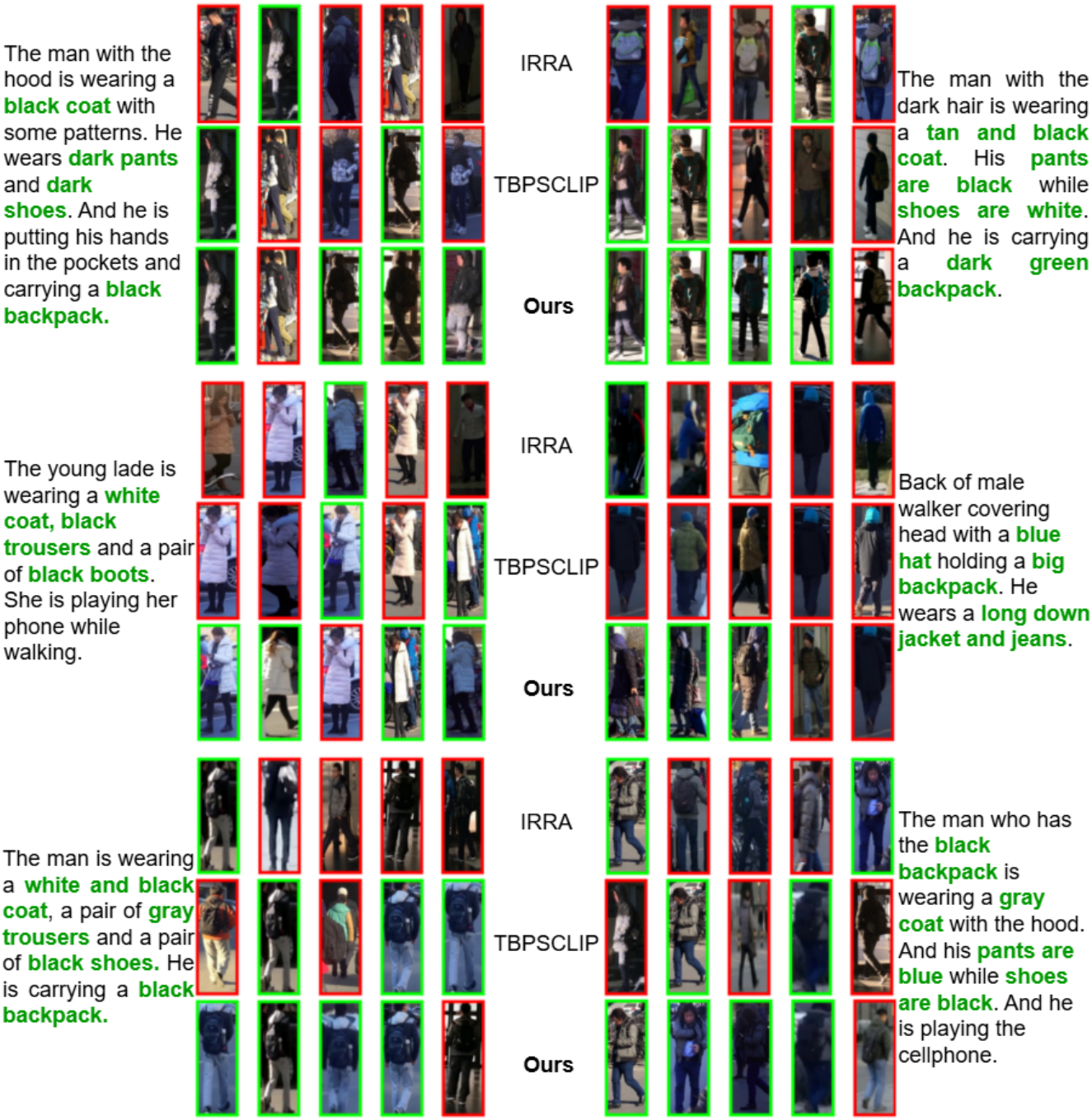


图 11. RSTPReid基准上与其他方法的更多示例对比

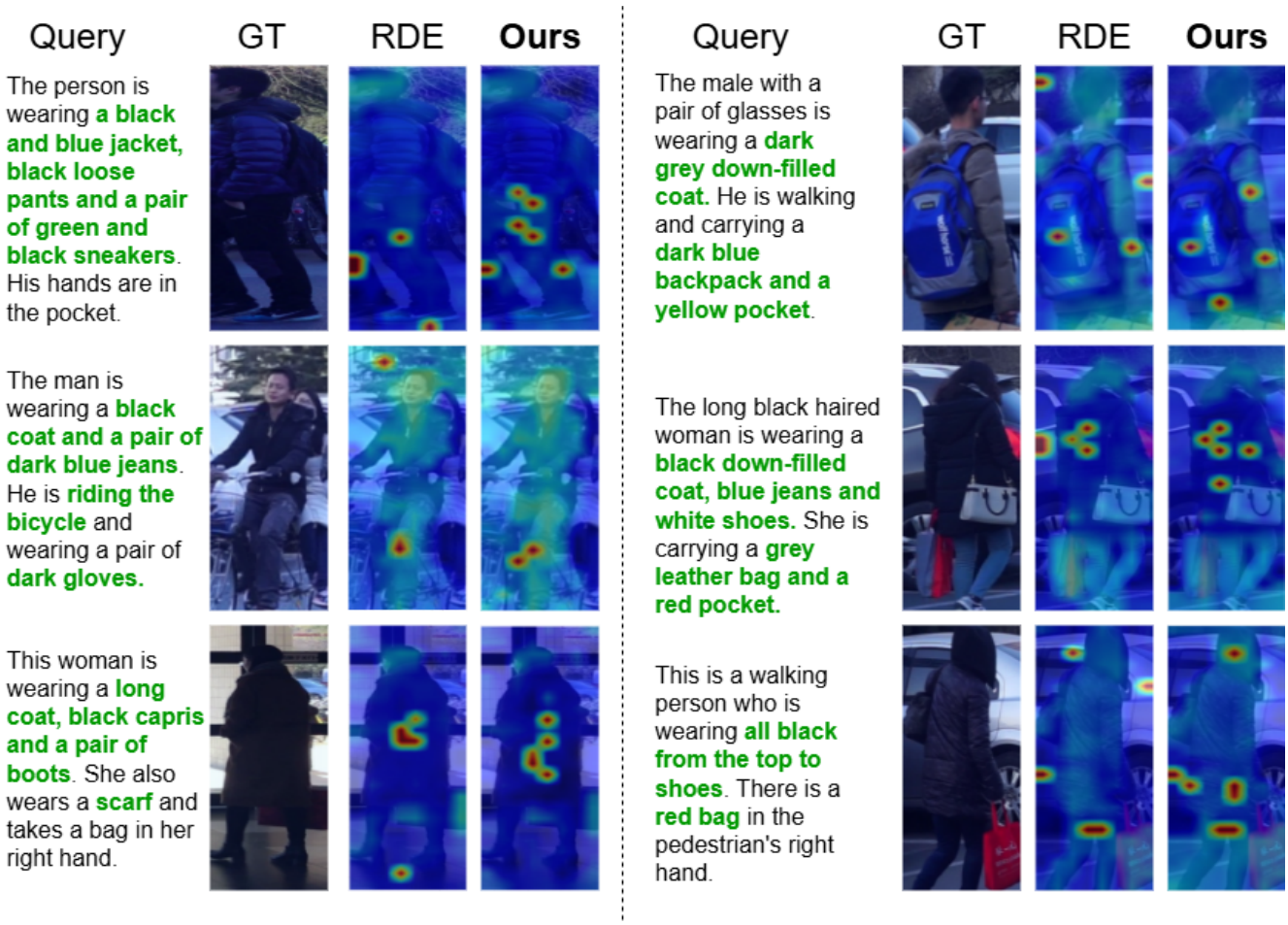


图12. 在RSTPReid基准上与RDE进行比较的更多注意力图示例