arXiv:2603.01666v1 [cs.CL] 2 Mar 2026

# 超越网格：利用布局信息的多向量检索与解析后的视觉文档表示

**Yibo Yan** [1 2 3] **Mingdong Ou** †[2] **Yi Cao** [2] **Xin Zou** [1 3] **Shuliang Liu** [1 3] **Jiahao Huo** [1 2] **Yu Huang** [1 2] **James Kwok** [3] **Xuming Hu** †[1 3]

## 摘要

要充分发挥视觉丰富文档的潜力，检索系统不仅要理解文本，还要理解复杂的版面布局；这是视觉文档检索（VDR）面临的核心挑战。现有的多向量架构虽然功能强大，却存在关键的存储瓶颈；当前的优化策略，如嵌入合并、剪枝或引入抽象令牌，都无法在不牺牲性能或忽略重要布局线索的情况下解决这一问题。为此，我们提出 `ColParse`这一新范式，利用文档解析模型生成一小组融入布局信息的子图像嵌入，再将其与全局页面级向量融合，从而构建紧凑且具备结构感知能力的多向量表示。大量实验表明，`ColParse` 可将存储需求降低逾95%，同时在众多基准测试和基准模型上显著提升性能。`ColParse`由此弥合了多向量检索的细粒度准确性与大规模部署实际需求之间的关键差距，为构建高效且可解释的多模态信息系统开辟了新路径。

## 1. 引言

视觉文档检索（VDR）旨在从大规模语料库中检索相关文档页面，已成为现代信息检索领域的基石（Mei 等人，2025年；Yan 等人，2026a）。与自然图像检索不同，视觉文档（如学术论文、金融报告和发票）的特点是文本内容、复杂版面布局和图形元素密集交织，如图1所示。为有效捕捉这些细粒度细节，该领域主要采用多向量检索架构（Faysse 等人，2024年；G¨unther 等人，2025年；Team，2025年）。这类模型将每个文档页面表示为一组图块级嵌入，并采用诸如`MaxSim`这样的后期交互机制来计算相关性（Khattab 和 Zaharia，2020年；Santhanam 等人，2022年）。这种范式擅长将特定查询短语与文档中的相应视觉或文本区域对齐；对于视觉文档检索所涉及的高精度信息检索任务而言，这一能力至关重要。

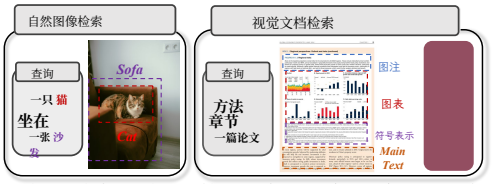



图*1.* 自然图像检索与视觉文档检索（VDR）的比较。

尽管性能更优，多向量视觉文档检索模型的广泛应用仍受一个关键瓶颈阻碍：存储开销高得令人难以承受（Jayaram 等人，2024年；Shrestha 等人，2024年；Liu 和 Mao，2023年）。为每个页面存储数百甚至数千个嵌入向量，使大规模部署在实践中面临巨大挑战。为解决这一问题，研究社区探索了多种优化策略，如图2所示。❶ 一类研究致力于合并补丁嵌入，例如 Light-ColPali（Ma 等，2025年）采用聚类技术聚合相似向量。然而，这种方法往往会削弱细粒度信息，导致性能不稳定。❷ 另一种方向是剪枝，DocPruner（Yan 等，2025年）等框架旨在丢弃冗余嵌入。但在激进压缩的情况下，这些方法难以维持性能。❸ 第三种范式以 MetaEmbed（Xiao 等，2025年）为代表，引入一组抽象的可学习令牌，以形成紧凑的多向量表示。虽然这一方法颇具创新性，但这些令牌缺乏对文档内在布局结构的明确依据，因此难以捕捉关键的布局特定语义。

[1]香港科技大学（广州） [2]阿里云计算 [3]香港科技大学。联系邮箱：Yibo Yan <yanyibo70@gmail.com>、Mingdong Ou <mingdong.omd@alibaba-inc.com>、Xuming Hu<xuminghu@hkust-gz.edu.cn>。

预印本。*2026*年*3*月*3*日。

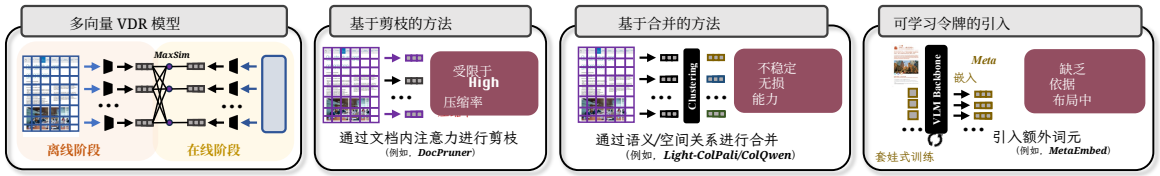



图2. 多向量视觉文档检索模型及其效率瓶颈的三种主要优化策略示意

一种多向量视觉文档检索模型，以及针对其效率瓶颈的三种主要优化策略。

为克服现有方法的局限，我们提出 ColParse，**一种构建多向量表示的新范式，其设计从根本上契合视觉文档的结构特性**。不同于在统一的图像块网格或抽象令牌上进行处理，ColParse 首先采用专门的文档解析模型，将每个文档页面智能地划分为少量 $k$ 具有语义意义且充分考虑布局的子图像（例如 表格、图表、段落），其中 $k$ 通常少于 10 个。随后，逐一使用标准单向量检索模型对这些 $k$ 子图像进行编码，得到 $k$ 局部向量。同时，对整个文档页面进行编码，以生成一个捕捉整体语境的全局向量。最后，我们将全局向量加权后逐元素加到每个 $k$ 局部向量上，从而融合这些表征。此过程为每个文档生成 $k$ 融合向量，同时整合细粒度的布局特定细节与页面级整体语境。

我们在 24 个不同的视觉文档检索数据集（Meng 等人，2025年）上进行了全面实验，以验证所提出框架的有效性和稳健性。 ColParse在应用于 10 种不同的主流单向量模型时，始终带来显著的性能提升，nDCG@5 平均提高超过 10 分。这体现了其作为免训练、即插即用模块的卓越灵活性。通过将视觉文档独特的结构属性与多向量检索的强大机制深度融合，ColParse 在检索性能与存储效率之间实现了新的权衡。我们的主要贡献如下：

❶ **一种新颖的多向量构建范式：**我们提出了视觉文档检索领域首个布局感知的多向量表示构建范式，通过利用文档解析，突破了传统多向量模型的存储效率瓶颈。

❷ **灵活且稳健的框架：**我们的方法被设计为一个免训练、即插即用的框架，在众多现有单向量模型上均展现出稳健且显著的性能提升，凸显了其通用性和易于采用的特点。

❸ **卓越性能与增强的可解释性：**ColParse 通过使检索结果能够追溯到特定的、已解析的布局组件，提供了固有的可解释性，从而显著提升了其实用性以及在真实世界工业应用中的潜力。

## 2. 相关工作

### 2.1. 视觉文档检索

视觉文档检索已成为理解视觉丰富文档的关键任务，它超越了往往会丢失关键布局信息的传统基于 OCR 的流程（Zhang 等人，2025b；Most 等人，2025年）。视觉语言模型（VLMs）的出现引入了端到端的单向量方法（例如 DSE（Ma 等人，2024a）、GME（Zhang 等人，2024b）和 UniSE（Liu 等人，2025b）），但这些方法往往难以捕捉密集文档所需的细粒度语义。多向量范式带来了重大飞跃，该范式由 ColPali（Faysse 等人，2024年）开创，它将页面表示为大量图块级嵌入，并采用晚期交互以实现更优匹配。近期工作试图在多个层面优化这一范式：在模型层面，探索如 Modern-VBERT（Teiletche 等人，2025年）这样的双向架构；在数据层面，通过先进的数据合成和困难负样本挖掘，如 Llama Nemoretriever Colembed（Xu 等人，2025年）等工作所示；在训练层面，则借助新的目标函数和多任务框架，如 jina-embeddings-v4（G¨unther 等人，2025年）。尽管性能出色，这些多向量模型仍带来了严重的存储瓶颈。

### 2.2. 多向量检索

多向量范式最初由 ColBERT（Khattab 和 Zaharia，2020年）在文本检索领域推广开来，它将文档表示为词元级嵌入集合，通过后期交互机制实现细粒度匹配（Qian 等人，2022年；Lee 等人，2023年）。随后，BGE-M3-Embedding（Chen 等人，2024年）和 Jina-ColBERT-v2（Jha 等人，2024年）等模型进一步改进了这一范式在文本领域的应用。ColPali（Faysse 等人，2024年）成功将这一范式拓展至多模态检索，将重点转向视觉文档；与自然图像相比，视觉文档本身更为复杂。尽管这些模型性能更优，但图块级嵌入带来的高昂存储成本，成为它们面临的关键效率瓶颈（Liu 和 Mao，2023年；Shrestha 等人，2024年；Park 等人，2025年）。当前的优化工作主要分为三类，但每类方法都存在固有缺陷。（i）如 DocPruner（Yan 等人，2025年）和 Prune-then-Merge（Yan 等人，2026b）所采用的冗余嵌入剪枝，在压缩力度较大时往往难以维持性能，

。（ii）**合并**相似嵌入，例如通过聚类实现的 Light-ColPali（Ma 等人，2025年），可能会稀释细粒度信息，导致性能不稳定。（iii）**引入抽象的可学习令牌**，由 MetaEmbed（Xiao 等人，2025年）和 CausalEmbed（Huo 等人，2026年）开创，能够创建紧凑表示，但这些表示缺乏对文档固有布局结构的明确依据。相比之下，`ColParse` 通过利用文档解析生成一组紧凑的、由布局提供信息的嵌入，从而解决这些局限。

### 2.3. 文档解析视觉语言模型

文档解析视觉语言模型已成为将视觉丰富的文档图像转换为 LaTeX 或 Markdown 等结构化格式的关键工具（Zhang 等人, 2024a;Ouyang 等人, 2025年; Zhang 等人, 2025c）。早期模型，如 Nougat（Blecher 等人, 2023年）和 Donut（Kim 等人,2022年），采用了端到端的序列到序列方法，但往往难以承受高分辨率输入的计算成本。为平衡准确率与效率，一种更新的多阶段范式日益受到关注。MinerU2.5（Niu 等人,2025年）等模型便是典型代表：它首先在下采样图像上进行高效的版面分析，再对裁剪区域进行有针对性的高分辨率识别。这种由粗到细的策略同样见于 Dolphin（Feng 等人, 2025年）和 MonkeyOCR（Zhang 等人, 2025a）等模型，有效缓解了端到端处理高分辨率图像的 $O(N^2)$ 复杂度。对于 `ColParse`，我们选择 MinerU2.5 作为文档解析器，因其具备最先进的准确率与效率。与其他文档解析模型的定量比较将在章节 4.2.3 中给出。

## 3. 方法

在本章节中，我们首先在多向量范式下对视觉文档检索任务进行形式化。随后，我们介绍`ColParse` 框架，并详细说明其生成紧凑且由布局提供信息的文档表示的多阶段流程。伪代码见附录 A。

### 3.1. 任务形式化

视觉文档检索的主要目标是，给定文本查询 $q$，从大规模语料库 $\mathcal{C} = \{d_1, d_2, \ldots, d_{|\mathcal{C}|}\}$ 中检索出相关文档页面的排序列表。在传统的多向量检索范式中，文档页面 $d$ 会先被渲染为图像，然后进行均匀划分划分为由 $N_p$ 图像块组成的网格，$\{p_j\}_{j=1}^{N_p}$。视觉语言模型作为编码器 $\Phi(\cdot)$，将每个图像块 $p_j$ 映射为 $D$ 维嵌入，从而得到一组大型的图像块级文档嵌入 $\mathbf{D}_{\text{grid}} = \{\mathbf{d}_j\}_{j=1}^{N_p}$，其中每个 $\mathbf{d}_j \in \mathbb{R}^D$。与此同时，同一编码器将文本查询 $q$ 映射为一组

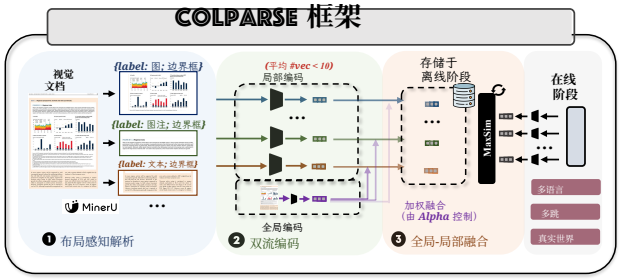



图 3. `ColParse` 框架的简化示意图。

由 $N_q$ 个词元级嵌入 $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_q}$组成，其中每个$\mathbf{q}_i \in \mathbb{R}^D$。随后，使用后期交互机制（通常为 MaxSim）计算查询与文档之间的相关性分数 $s(q, d)$，定义如下：

$$s(q,d) = \sum_{i=1}^{N_q} \max_{j=1}^{N_p} (\mathbf{q}_i^\top \mathbf{d}_j). \tag{1}$$

其中，假设向量经过 L2归一化。尽管这种基于网格的方法在细粒度匹配方面表现出色，但其存储成本高得令人难以承受：每个文档页面需要 $O(N_p \times D)$，因为 $N_p$ 可能达到数百甚至数千。

我们的工作旨在解决这一关键瓶颈。我们希望用高度紧凑、具备结构感知能力的多向量表示$\mathbf{D}_{\texttt{ColParse}}$取代庞大且版面无关的集合$\mathbf{D}^{\text{grid}}$，该表示仅包含 $k$ 个向量，其中$k \ll N_p$。这种新表示应能将存储需求大幅降低至 $O(k \times D)$，同时通过明确基于文档的语义版面来提升检索性能。

### 3.2. `ColParse` 框架

`ColParse` 是一种即插即用框架，通过“超越网格”彻底革新了多向量表示的构建方式。它不依赖均匀图块，而是利用结构理解生成紧凑且语义丰富的嵌入。如图3所示，我们的框架针对每个文档图像离线运行，包含三个阶段：$d \in \mathbb{R}^{H \times W \times 3}$（1）布局感知文档解析，（2）双流编码，以及（3）全局-局部融合。

#### 3.2.1. 布局感知文档解析

的基础步骤是将文档图像拆解为其组成语义组件。我们采用专用的现成文档解析模型，($\Psi_{\text{parse}}$)，它充当布局检测器。对于给定的文档图像 $d$，解析器会识别出 $k$ 个不同的布局区域，并输出相应的边界框和内容类型标签：$[\{b_j, c_j\}_{j=1}^{k}] = \Psi_{\text{parse}}(d)$。其中，$b_j = (x_{j1}, y_{j1}, x_{j2}, y_{j2})$ 是第 $j$ 个区域的边界框，用于定义其在原始图像中的坐标。 $c_j$

是来自预定义内容类型集合的类别标签$\mathcal{C}_{\text{types}} = \{$‘标题’，‘表格’，‘图’ $\ldots\}$，用于指示该区域的语义性质。利用这些边界框，我们裁剪原始图像 $d$，提取一组 $k$ 子图像 $\mathcal{S}_d$。子图像的数量 $k$由解析器根据文档的复杂程度动态确定，通常很少（例如，$k < 10$）。该过程可表述为 $\mathcal{S}_d = \{s_1, s_2, \ldots, s_k\}$，其中 $s_j = \text{Crop}(d, b_j)$。每个子图像 $s_j \in \mathbb{R}^{H_j \times W_j \times 3}$ 的尺寸各不相同，可灵活适应检测到的布局组件的尺寸。这种智能的内容感知分割构成了我们紧凑表示的基础。

#### 3.2.2. 双流编码

文档解析完成后，我们使用标准的单向量检索模型 $\Phi_{\text{enc}}(\cdot) : \mathbb{R}^{H' \times W' \times 3} \to \mathbb{R}^D$ 生成嵌入，该模型作为基础编码器。此阶段通过两条并行流运行，以同时捕捉局部的布局特定细节和全局的页面级上下文。

**局部编码**。 $k$ 中的每个大小可变的子图像 $s_j$ 均会调整尺寸，并独立输入编码器 $\Phi_{\text{enc}}(\cdot)$，以生成对应的 $D$ 维局部向量，$\mathbf{v}^{(j)}_{\text{局部}}$。此过程会产生一组 $k$ 局部嵌入，其中每个嵌入都表示一个独立的语义单元：$\mathbf{D}_{\text{局部}} = \{\mathbf{v}^{(j)}_{\text{局部}} \in \mathbb{R}^D \mid \mathbf{v}^{(j)}_{\text{局部}} = \Phi_{\text{enc}}(s_j), \forall s_j \in \mathcal{S}_d\}_{j=1}^{k}$.

**全局编码**。与此同时，将完整、未经裁剪的文档页面图像 $d$ 输入同一编码器$\Phi_{\text{enc}}(\cdot)$，生成一个单一的 $D$ 维全局向量，$\mathbf{v}_{\text{global}}$。该向量作为页面的整体摘要，捕捉全局上下文以及各布局组件之间的关系：$\mathbf{v}_{\text{global}} = \Phi_{\text{enc}}(d) \in \mathbb{R}^D$. 这种双流设计确保最终表示既能体现各个布局元素的特异性，也能利用整个页面的更广泛上下文。

#### 3.2.3. 用于最终表示的全局-局部融合

我们不采用简单求和，而是使用加权融合策略，从而能够灵活调节这两种关键信息流之间的平衡。对于每个 $k$ 局部向量，$\mathbf{v}^{(j)}_{\text{局部}}$，我们使用平衡因子$\alpha \in [0, 1]$，注入单个全局向量 $\mathbf{v}_{\text{global}}$所捕获的整体上下文。由此得到的融合向量 $\mathbf{d}^{(j)}_{\text{融合}} \in \mathbb{R}^D$，协同整合了细粒度细节与全局上下文。融合通过加权逐元素相加实现：$\mathbf{d}^{(j)}_{\text{融合}} = \alpha \cdot \mathbf{v}_{\text{global}} + (1 - \alpha) \cdot \mathbf{v}^{(j)}_{\text{局部}} \quad \forall j \in \{1, \ldots, k\}$. 这种加权机制使我们能够灵活调整表示的侧重点。

此融合操作将应用于所有 $k$ 局部向量，为文档生成最终的多向量表示 $d$，记作 $\mathbf{D}_{\texttt{ColParse}} = \{\mathbf{d}^{(j)}_{\text{融合}}\}_{j=1}^{k} \subset \mathbb{R}^{k \times D}$。随后，这组仅包含 $k$ 结构感知且经上下文增强的向量将被存储，以供在线检索，从而实现大幅削减存储空间的目标。

#### 3.2.4. 使用 ColParse进行后期交互评分

在在线检索阶段，使用紧凑表示$\mathbf{D}_{\texttt{ColParse}}$，即可高效计算查询 $q$ 与文档 $d$ 之间的相关性分数。首先，按照常规方式将查询编码为词元级嵌入$\mathbf{Q} = \{\mathbf{q}_i \in \mathbb{R}^D\}_{i=1}^{N_q}$。随后，对 $k$ 融合文档向量计算 MaxSim 分数：

$$s_{\texttt{ColParse}}(q, d) = \sum_{i=1}^{N_q} \max_{j=1}^{k} (\mathbf{q}_i^\top \mathbf{d}^{(j)}_{\text{fused}}). \tag{2}$$

将搜索对象从 $N_p$ 基于网格的向量替换为仅搜索 $k$ 布局感知向量后， ColParse 不仅大幅减少了存储占用，还使后期交互机制能够聚焦于文档中语义上最显著的部分。

### 3.3. 理论基础

我们从信息瓶颈（IB）的角度，为 ColParse 提供理论依据。我们证明，该框架的架构通过（1）解析以解耦源信息，以及（2）利用上下文侧信息对其进行细化，为解决视觉文档检索中难以处理的 IB 目标提供了一种有原则的替代方案。

#### 3.3.1. 作为信息瓶颈的视觉文档检索压缩问题

设 $D$ 为表示文档图像的随机变量， $Q$ 为表示查询的随机变量，它们来自未知分布 $P(Q)$。设 $R$ 为相关性变量，是 $D$ 和 $Q$ 的函数。目标是通过求解 IB 拉格朗日函数，学习一个压缩函数 $g$，将文档 $D$ 映射为紧凑表示 $Z = g(D)$：

$$\min_g \mathcal{L}(Z) = I(Z; D) - \beta \mathbb{E}_Q[I(Z; R(D, Q))]. \tag{3}$$

该目标旨在最小化 $Z$ 保留的关于源 $D$ 的信息（压缩），同时最大化其保留的关于相关性 $R$ 的信息（预测）。由于需要对未知查询分布取期望 $\mathbb{E}_Q$，这个问题在索引阶段难以求解。

#### 3.3.2. 通过解析实现信息解耦

ColParse的第一阶段，即解析，会变换输入空间。令 $\Psi_{\text{parse}}(D) = \{S_1, \ldots, S_k\}$ 为从文档 $D$导出的子图像（语义区域）集合。这些区域构成该文档核心语义内容的一个划分。根据互信息链式法则，原始文档中的信息可通过其

组成部分表示： $I(D;\ R) = I(S_1, S_2, \ldots, S_k;\ R)$。我们提出**语义集中公理**：对于任意给定的查询$Q$，相关性信号 $R$ 主要由单个主要语义区域 $S_{j^*} \in \{S_j\}$决定，其中$j^*$ 是最相关区域的索引。这意味着其余区域近似满足条件独立性：$I(S_{\neg j^*}; R|S_{j^*}) \approx 0$, 其中 $S_{\neg j^*} = \{S_j\}_{j \neq j^*}$. 。该公理导出如下近似： $I(D; R) \approx I(S_{j^*}; R)$. 这种分解将问题从压缩单一整体变量 $D$ 转变为创建一组表征，每个潜在主要通道 $S_j$各对应一个表征。这为IB问题提供了结构先验，从而说明应创建多向量集合 $\{g_j(S_j)\}_{j=1}^k$，而非单一向量 $g(D)$。

3.3.3. 通过协同融合进行上下文精炼

`ColParse` 生成两组中间表征：局部向量 $\{V_j = \Phi_{\text{enc}}(S_j)\}_{j=1}^k$ 和全局向量 $V_{\text{global}} = \Phi_{\text{enc}}(D)$。这些编码步骤中的每一步本身都是一个信息瓶颈，受数据处理不等式（DPI）约束： $I(V_j;\ R) \leq I(S_j;\ R) \quad \forall j \in \{1, \ldots, k\}$; $I(V_{\text{global}}; R) \leq I(D; R)$。 `ColParse` 的核心是融合步骤，该步骤创建最终表示集合 $\{Z_j = V_j + V_{\text{global}}\}_{j=1}^k$。为分析其优势，考虑局部—全局对（$V_j$, $V_{\text{global}}$）所包含的关于相关性 $R$的联合信息。根据链式法则：

$$I(V_j, V_{\text{global}}; R) = I(V_j; R) + \underbrace{I(V_{\text{global}}; R|V_j)}_{\textbf{Contextual Information Gain}}. \quad (4)$$

该增益项量化了在给定局部区域的情况下，全局上下文为相关性带来的新信息。当全局上下文有助于消除局部内容的歧义时，该项非零。融合函数 $f(V_j,\ V_{\text{global}}) = V_j + V_{\text{global}}$创建向量 $Z_j$。根据数据处理不等式，这种融合也构成一个瓶颈：

$$I(Z_j; R) = I(V_j + V_{\text{global}}; R) \leq I(V_j, V_{\text{global}}; R). \quad (5)$$

这种融合的目标是构造一个紧凑的向量 $Z_j$，使其比单独的局部向量 $V_j$ 包含更多信息，同时捕获相当一部分上下文增益。区域 $j$ 的信息净增量为：

$$\Delta I_j = I(Z_j; R) - I(V_j; R). \quad (6)$$

`ColParse` 的成功取决于这种融合是否有效，即，能否确保 $\Delta I_j > 0$。当融合操作能够成功编码上下文信息增益时，这一点便能成立。简单的向量加法是一种无需参数且计算高效的实现机制。因此，最终表示 $\mathbf{D}_{\texttt{ColParse}} = \{Z_j\}_{j=1}^k$ 是一组经上下文优化、彼此解耦的向量，能够更有效地保留与查询相关的信息，从而为视觉文档检索压缩问题提供更优的解决方案。

更多理论分析请参见附录 B。

# 4. 实验

## 4.1. 实验设置

基准测试与评估。为确保评估全面，我们在五个主流 VDR 基准测试套件上评估 `ColParse`的性能，这些套件共包含 24 个不同的数据集：ViDoRe-V1（Faysse 等人，2024年）、ViDoRe-V2（Mac´e 等人，2025年）、VisRAG（Yu 等人，2024年）、ViDoSeek（Wang 等人，2025年）和 MMLongBench（Ma 等人，2024b）。我们将框架应用于十种知名单向量检索模型，以验证其通用性和即插即用特性：VLM2Vec-V1-2B/7B（Jiang 等人，2024年）、VLM2Vec-V2-2B（Meng 等人，2025年）、LamRA-Ret（Liu 等人，2025a）、GME-2B/7B（Zhang 等人，2024b）、UniME-V2-2B/7B（Gu 等人，2025年）、B3-2B/7B（Thirukovalluru 等人，2025年）。基准测试和模型的详细信息分别见附录 C.1 和 C.2。按照 VDR 领域的惯例（Wasserman 等人，2025年；ILLUIN，2025年），我们采用 5 处归一化折损累计增益（nDCG@5）作为评估指标。

**基线方法**。为了展示 `ColParse` 的优越性，我们将其与涵盖五个不同类别的多种基线方法进行比较。❶ **基础方法：**此类别代表十种单向量模型在未经任何多向量适配时的原始性能。❷ **多图像：**该方法将从文档中解析出的所有子图像同时输入基准模型，利用其原生的多图像输入支持，生成一个代表性向量。❸ **布局分块：**此类别探索直接作用于基准模型最终层词元嵌入的策略。在布局解析结果的指导下，这些词元被分块并聚合，而解析出的边界框之外的词元则被丢弃。我们评估四种变体：类型聚类（通过语义聚类合并内容类型相同的词元）、类型均值（通过均值池化合并内容类型相同的词元）、子图像聚类（对来自同一子图像区域的词元进行聚类）以及子图像均值（对来自同一子图像区域的词元进行池化）。❹ **语义分块：**与布局引导的方法不同，该基线方法对最终层的全部词元进行层次聚类，以生成固定数量的向量（默认为 10 个），且不考虑文档的布局结构。❺ **单转多：**此类别通过分别编码每个解析出的子图像来模拟 `ColParse`，但采用不同于我们向量融合的评分机制。两种变体分别为加法评分和乘法评分：文档最终得分分别通过将各个查询-子图像相似度分数与全局图像-子图像相似度分数相加或相乘来计算。

**实现细节**。 为确保公平比较和可复现性，我们的整个评估流程均基于

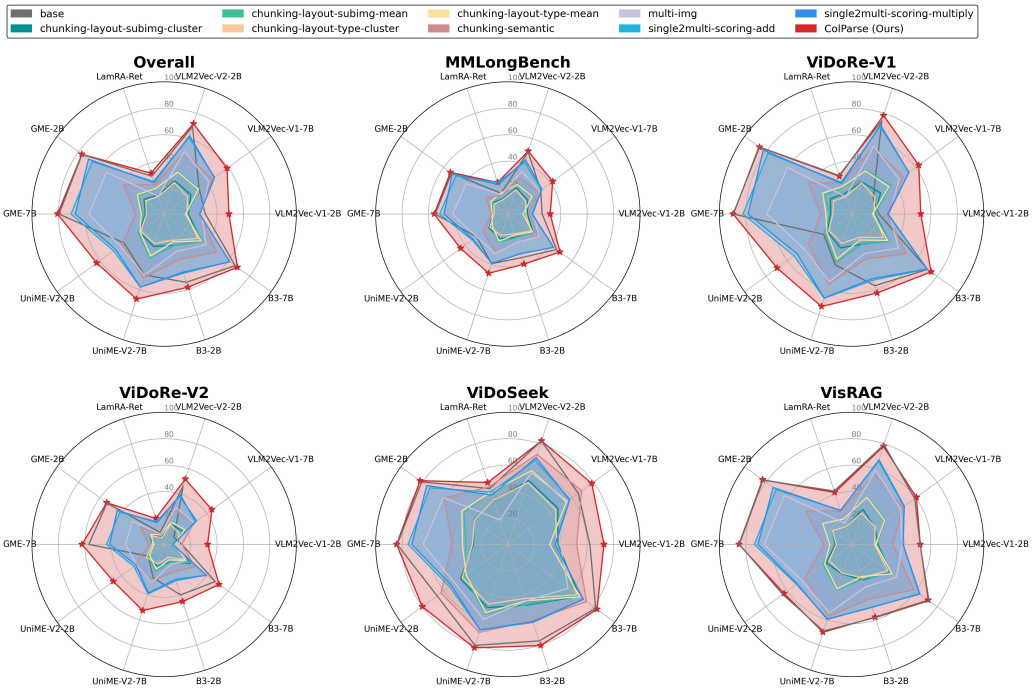



图 4. 在十种主流单向量多模态检索模型上，比较了 与各基线方法在五个视觉文档检索基准测试上的性能（以 nDCG@5 评估）。受篇幅限制，详细结果请参阅 表2 和 表3 。

MMEB 代码库[1]。我们在所有实验中均采用 MinerU2。5[2] (Niu 等人，2025年) 作为统一的文档解析模型，因其具备最先进的性能和高效性（MinerU2.5 的详细信息见附录 C.3）。对于ColParse，我们选择 $\alpha$ 取值范围为 0.1 至 0.9，间隔为 0.1，并为每个基准模型选取最优超参数，以报告最终结果（已在第 4.2.3 章中验证）。实验在配备 NVIDIA A100（80G）GPU 的集群上进行。完整代码库将在论文录用后公开。

## 4.2. 实验分析

4.2.1. 主要结果

**ColParse 在各种基准测试中，始终优于单向量基座模型和现有的多向量优化基线。**如表2和表3（附录C.4）所示，ColParse 在 ViDoRe-V1 基准测试中，VLM2Vec-V1-2B 的平均 nDCG@5 提升了 31.64 分，其 7B 版本则提升了 42.69 分，表现十分突出。图4进一步直观展示了这一优势，其中代表 ColParse 的红色包络线在全部十种主流模型中始终构成最外侧边界。这表明，融入布局信息的子图像表示能够捕捉关键的细粒度细节，而这些细节通常会在单一整体的全局嵌入中被“稀释”。

**该框架展现出卓越的通用性和稳健性，无需训练即可作为即插即用模块，适用于各种基于视觉语言模型的嵌入。**在包括 VLM2Vec、GME、UniME 和 B3 在内的十种不同模型中，ColParse无论模型的架构或参数规模如何，都能持续提升性能。值得注意的是，即使对于性能出色的 GME-7B 模型，ColParse仍能保持最先进的表现，而语义分块（c-sem）等其他优化策略则会导致其在 ViDoRe-V1 上的性能从 89.36 骤降至 23.21。这种普适性表明，布局感知是视觉文档理解中一种基础且与模型无关的增强方式，可在表示层面实现。

**布局感知分解在多向量构建方面，显著优于传统的词元级分块或聚类方法。**定量结果表明，cl-t-c（类型聚类）和 cl-s-m（子图像均值池化）等基线方法往往会导致性能显著下降；例如，使用词元级聚类时，VLM2Vec-V2-2B 的性能从 74.16 暴跌至 24.85。相比之下，ColParse 通过保留语义区域的视觉完整性，而非聚合抽象的词元嵌入，将同一模型的性能提升至 78.41。这一现象表明，在解析后的区域中保持原始的视觉语义对齐，优于对后期特征进行事后启发式聚合。

**ColParse 在处理需要多跳推理的复杂长篇文档时表现出更优的效果。**

[1]https://github.com/TIGER-AI-Lab/VLM2Vec[2] https://github.com/opendatalab/MinerU

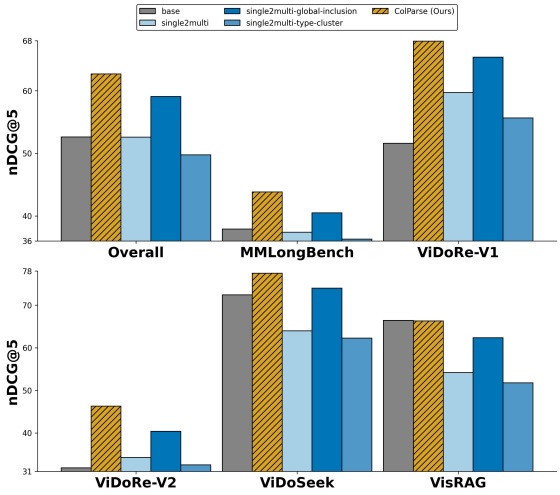



图5. ColParse 及其变体的变体研究。

在 MMLongBench 数据集上（表2）， ColParse 将 VLM2Vec-V1-2B 的平均 nDCG@5 从 25.93 提升至 32.07，将 UniME-V2-2B 从 29.31 提升至 44.21，优于所有其他压缩基线方法。这一性能飞跃在需要跨页信息定位的案例中尤为明显。我们推测，通过智能地将页面划分为关键布局组件（例如，表格、图表），ColParse 可以降低后期交互阶段的“信噪比”，使模型能够聚焦于语义上最显著的区域。

**全局-局部融合策略对于为孤立的语义区域提供必要的上下文锚定至关重要。** 与 单转多(s2m-add/mul) 基线方法的比较分析显示于 表3， ColParse在 VisRAG 和 ViDoSeek 等具有挑战性的任务中始终表现领先；使用 VLM2Vec-V1- 2B 时，在 VisRAG 上取得 51.96，而 s2m-add仅为 38.94。这一显著差距表明，仅靠局部子图像往往缺乏准确检索所需的整体上下文（例如，表格缺少其前文段落的上下文）。通过加权融合实现的协同作用使 ColParse 能够兼顾区域特异性和页面级语义。

#### 4.2.2. 变体研究

我们评估了三种具体变体：(i)**单转多**(s2m)变体作为布局分解表示的基线，它提取并编码子图像，将其表示为不含任何全局信息的独立向量。(ii) **单转多-类型聚类** (s2m-t-c)变体通过将同一语义类别中的所有子图像嵌入聚合为一个代表性类型级向量，在 s2m 的基础上进一步压缩表示。(iii) **单转多-全局纳入**(s2m-g-i)变体将原始整体页面级嵌入作为一个独立的全局向量，添加到 s2m 子图像集合中，以提供整页上下文。

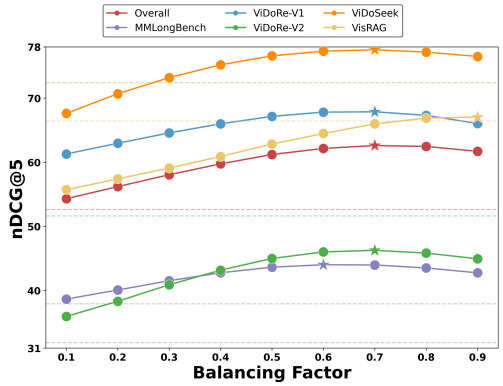



图 6. 不同 平衡因子下 ColParse 的平均性能比较。虚线表示基准结果；星形点表示表现最佳的平衡因子。模型级比较见 图 10 。

**协同的全局-局部融合在为具备布局感知能力的表征提供上下文方面，显著优于简单纳入全局向量。**如图5所示，ColParse 在所有十种基准模型和基准测试中始终优于s2m-g-i 变体。例如，在 ViDoRe-V1 上，VLM2Vec-V1-2B 的 nDCG@5 比 s2m-g-i高出 2.44 分。图示结果进一步展示了这一优势：在图9（附录C.5）的雷达图中， ColParse （红色包络）始终覆盖最大面积；尤其是在 VisRAG 等密集型任务中，其在 7B 模型上的表现领先s2m-g-i 逾 2 分。这表明，逐元素融合比简单纳入更深层地发挥条件化机制的作用，使局部特征能够被文档的全局语义环境从根本上“重新加权”。

更多分析请参见附录C.5 ，受篇幅限制，此处不再赘述。

#### 4.2.3. 超参数分析

平衡因子的影响。我们研究了 ColParse 对平衡因子 $\alpha$ 的敏感性，如图 6 所示（详见附录 C.6.1）。首先，在 $\alpha \in [0.1, 0.9]$ 的整个取值范围内，检索性能都十分稳健，并且所有测试模型的表现始终优于单向量基线。例如，对于 VLM2Vec-V1-2B，即使 $\alpha$ 低至 0.1，总体得分仍达到 35.97，显著高于基线结果 31.18。其次，性能通常呈凸轨迹，最佳平衡点一般偏向全局上下文，以有效地为孤立布局组件提供依据。从定量结果来看，大多数模型在 $\alpha \in [0.6, 0.8]$ 范围内达到最佳性能。最后，适度的协同融合至关重要，因为局部特异性过强或全局主导性过强都会导致次优结果。对于 B3-7B，性能从 $\alpha = 0.1$ 时的 62.09 稳步提升，在 $\alpha = 0.7$ 时达到峰值 68.51；随后，随着表示过度受全局向量主导，性能在 $\alpha = 0.9$ 时略有下降。

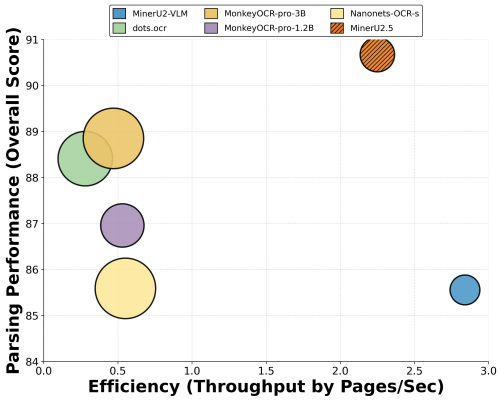



图7。MinerU2.5 与同类模型的对比。纵轴表示 OmniDocBench 的总体得分（Ouyang 等人，2025年），横轴表示端到端吞吐量（页/秒），气泡大小表示参数规模。

*表 1.* 效率分析： `ColParse` 在表现最佳的模型 GME-7B 及其多向量对应模型上的效率分析（采用原始设置）。**性能** 表示 MMEB-visdoc 的总体得分，**存储** 指每篇文档存储的向量数量，**延迟** 表示每篇文档的平均编码时间。* 表示多向量模型采用对齐配置进行训练。橙色 箭头表示更优，蓝色 箭头表示更差。

| 模型 | 性能 | 存储 | 延迟 |
|---|---|---|---|
| GME-7B | 79.50 | 1.00 | 0.30 |
| +ColParse | 80.61↑1.11 | 5.90↑4.9 | 0.81↑0.51 |
| ColQwen* | 80.02↑0.52 | 768.00↑767 | 0.41↑0.11 |

**文档解析模型的影响**。 我们评估了各种文档解析模型的能力，并得出结论：如 图7所示（详见附录C.6.2）。首先，与现有专用视觉语言模型相比，MinerU2.5 在所有语义和结构维度上都展现出更高的解析准确率。从定量结果来看，它在 OmniDocBench（Ouyang 等人，2025年）上取得了业界领先的总体得分 90.67，显著优于最佳基线模型 MonkeyOCR-pro-3B（88.85），并且在文本（0.047）和阅读顺序（0.044）识别方面均保持最低的错误率。其次，MinerU2.5 实现了工业级吞吐量，确保了其在大规模 `ColParse` 文档语料库中的实用性。例如，其端到端处理速度达到 2.25 页/秒， 4× 快于 Nanonets-OCR-s（0.55 页/秒）等高参数量模型。因此，我们选择 MinerU2.5 作为统一解析器，因为它处于最优的帕累托前沿，在保持出色吞吐量的同时，提供了最高的解析质量。

#### 4.2.4. 效率分析

为确保公平比较，我们使用 GME-7B（性能最佳的单向量模型）和 ColQwen（架构与训练配置相匹配的多向量基线）评估 `ColParse` 。我们得出结论，我们的范式能够实现更优的检索性能，同时大幅削减传统多向量模型固有的存储开销。具体而言，如表 1所示， `ColParse` 在 MMEB-visdoc 上的总体得分为 80.61，优于多向量模型 ColQwen（80.02），而每个页面仅需 5.9 个向量（各数据集的详细记录见附录C.6.3），与 ColQwen 的 768 个向量相比，存储开销削减幅度超过 99%。此外，尽管 `ColParse` 会略微增加编码延迟，但仍是非常适合真实世界部署的实用方案。虽然每份文档的延迟升至 0.81 秒，但仍显著低于传统基于 OCR 的流程平均 7 秒的延迟，而且可通过并行处理优化离线解析阶段，从而降低索引开销。

#### 4.2.5. 案例研究




图 *8.* 一个典型案例的示意图。

我们在图 8 中开展了一项案例研究，以展示 `ColParse` 除卓越的检索性能之外的可解释性。在处理需要特定细节的查询时，例如碳排放减少百分比，`ColParse`不仅能检索到正确的文档页面，还能精确定位得分最高的解析子向量（例如，某个特定文本块）的 `MaxSim` 得分。这种布局感知粒度支持细粒度的信息回溯，使系统能够直接向用户展示具体的证据区域。在财务审计或法律审查等实际工业场景中，这种可说明性极具价值，因为核实信息来源的能力与检索准确率同样重要。

## 5. 结论

在本文中，我们提出了 `ColParse`，这是一种新颖的布局感知范式，旨在克服多向量视觉文档检索中的关键存储效率瓶颈。我们的框架将文档解析器生成的布局感知子图像嵌入与整体全局向量融合，从而生成紧凑的多向量表示。大量实验表明， `ColParse` 在不同基础模型和数据集上始终提升检索性能，同时实现了超过 95% 的存储压缩。最终，`ColParse` 为这一领域开辟了新方向，确立了深入理解文档结构对于构建实用多模态信息系统至关重要。

## 影响声明

**伦理考量**。我们认为，我们提出的ColParse 框架不会引发新的伦理问题。该框架旨在以原则性且注重资源效率的方式，提升视觉文档检索系统的效率和性能。通过利用现有文档解析技术，生成更紧凑且更易于解读的表示，我们的方法推动了负责任的人工智能发展，并遵循信息检索研究中既定的伦理准则，同时不依赖敏感或专有数据。

**社会影响**。 ColParse 通过从存储密集型的基于网格的表示转向轻量级的布局感知方法，为多模态信息检索引入了新范式。这从根本上解决了细粒度检索准确率与大规模部署的实际存储成本之间的关键矛盾。通过将存储需求降低 95% 以上，同时提升性能，ColParse大幅降低了部署最先进视觉文档理解系统的门槛。这有望推动高级信息检索在多个领域普及，包括学术研究、企业知识管理和数字档案。此外，其固有的可解释性能够将检索结果关联到具体的文档组成部分，从而提高透明度，增强用户对人工智能驱动的信息系统的信任。

## 参考文献


Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. Nougat：面向学术文档的神经光学理解。*arXiv* 预印本 *arXiv:2308.13418*，2023年。

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., 和 Liu, Z. M3-embedding：通过自知识蒸馏实现多语言、多功能、多粒度文本嵌入。载于计算语言学协会 *ACL 2024* 论文集，第 2318–2335 页，2024年。

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., 和Colombo, P. Colpali：利用视觉语言模型实现高效文档检索。*arXiv* 预印本 *arXiv:2407.01449*，2024年。

Feng, H., Wei, S., Fei, X., Shi, W., Han, Y., Liao, L., Lu, J., Wu, B., Liu, Q., Lin, C., 等。Dolphin：通过异构锚点提示进行文档图像解析。*arXiv* 预印本 *arXiv:2505.14059*，2025年。

Gu, T., Yang, K., Zhang, K., An, X., Feng, Z., Zhang, Y., Cai, W., Deng, J. 和 Bing, L. UniME-V2：以多模态大语言模型作为评判者进行通用多模态嵌入学习，2025年。网址 https://arxiv.org/abs/2510.13515。

G¨unther, M., Sturua, S., Akram, M. K., Mohr, I., Ungureanu, A., Wang, B., Eslami, S., Martens, S., Werk, M., Wang, N. 等人。jina-embeddings-v4：面向多模态多语言检索的通用嵌入。*arXiv*预印本 *arXiv:2506.18902*，2025年。Huo, J., Huang, Y., Yan, Y., Pan, Y., Cao, Y., Ou, M., Yu, P. S. 和 Hu, X. CausalEmbed：面向视觉文档嵌入的潜在空间自回归多向量生成。*arXiv*预印本 *arXiv:2601.21262*，2026年。

ILLUIN. ViDoRe V3：面向企业应用场景的全面检索评估，2025年11月。网址 https://huggingface.co/blog/QuentinJG/introducing-vidore-v3。

Jayaram, R.、Dhulipala, L.、Hadian, M.、Lee, J. D. 和 Mirrokni, V. Muvera：通过固定维度编码实现多向量检索。神经信息处理系统进展，37:101042–101073，2024年。

Jha, R.、Wang, B.、G¨unther, M.、Mastrapas, G.、Sturua, S.、Mohr, I.、Koukounas, A.、Akram, M. K.、Wang, N. 和 Xiao, H. Jina-ColBERT-v2：通用型多语言晚期交互检索器。*arXiv* 预印本 *arXiv:2408.16672*，2024年。

Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., and Chen, W. VLM2Vec：面向大规模多模态嵌入任务的视觉语言模型训练。*arXiv* 预印本 *arXiv:2410.05160*，2024年。

Khattab, O. and Zaharia, M. ColBERT：通过基于 BERT 的上下文晚期交互，高效且有效地进行段落搜索。发表于第 *43* 届 *ACM SIGIR* 信息检索研究与发展国际会议论文集，第 39–48 页，2020年。

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. 无 OCR 的文档理解 Transformer。发表于欧洲计算机视觉会议，第 498–517 页。施普林格，2022年。

Lee, J., Dai, Z., Duddu, S. M. K., Lei, T., Naim, I., Chang, M.-W., and Zhao, V. 重新思考词元检索在多向量检索中的作用。神经信息处理系统进展，36:15384–15405，2023年。

Liu, Q. 和 Mao, J. 理解多向量稠密检索模型。载于第 *32* 届 *ACM* 国际信息与知识管理会议论文集，第 4110–4114 页，2023年。

Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., 和 Xie, W. Lamra：将大型多模态模型作为您的高级检索助手。发表于计算机视觉与模式识别会议论文集，第 4015–4025 页，2025a。

Liu, Z., Liu, Z., Liang, Z., Zhou, J., Xiao, S., Gao, C., Zhang, C. J., 和 Lian, D. 任何信息都只值一张截图：以可视化信息检索统一搜索。发表于计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 19238–19261 页，2025b。

Ma, X., Lin, S.-C., Li, M., Chen, W., 和 Lin, J. 通过文档截图嵌入统一多模态检索。载于 Al-Onaizan, Y., Bansal, M., 和 Chen, Y.-N.（编），*2024*年自然语言处理实证方法会议论文集，第 6492–6505 页，美国佛罗里达州迈阿密，2024年11月。计算语言学协会。doi: 10.18653/v1/2024.emnlp-main.373.

URL https://aclanthology.org/2024.emnlp-main.373/。

Ma, Y., Zang, Y., Chen, L., Chen, M., Jiao, Y., Li, X., Lu, X., Liu, Z., Ma, Y., Dong, X., 等。Mmlongbench-doc：利用可视化对长上下文文档理解进行基准测试。神经信息处理系统进展，37:95963–96010，2024b。Ma, Y., Li, J., Zang, Y., Wu, X., Dong, X., Zhang, P., Cao, Y., Duan, H., Wang, J., Cao, Y., 等。迈向存储高效的视觉文档检索：减少图块级嵌入的实证研究。*arXiv*预印本 *arXiv:2506.04997*，2025。Mac´e, Q., Loison, A., 和 Faysse, M. ViDoRe基准v2：提升视觉检索的标准。*arXiv*预印本 *arXiv:2505.17166*，2025。Mei, L., Mo, S., Yang, Z., 和 Chen, C. 多模态检索增强生成综述。*arXiv*预印本 *arXiv:2504.08748*，2025。Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Chen, Z., Xu, R., Xiong, C., 等。VLM2Vec-v2：推进视频、图像和视觉文档的多模态嵌入。*arXiv*预印本 *arXiv:2507.04590*，2025。Most, A., Winjum, J., Bhattarai, M., Jones, S., Ranasinghe, N. R., Biswas, A., 和 O'Malley, D. 迷失在光学字符识别的转换中？基于视觉的方法实现稳健的文档检索。载于*2025*年*ACM*文档工程研讨会论文集，第 1–10 页，2025年。Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., 等。MinerU2.5：一种用于高效高分辨率文档解析的解耦式视觉语言模型。*arXiv*预印本 *arXiv:2509.22186*，2025年。Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., 等。OmniDocBench：对多样化 PDF 文档解析进行基准测试

附有全面注释。在计算机视觉与模式识别会议论文集，第 24838–24848 页，2025年。Park, C., Jeong, S., Kim, M., Lim, K. 和 Lee, Y.-H. SCV：采用序列压缩向量的轻量且高效的多向量检索。发表于第 *31* 届计算语言学协会国际会议：产业分会论文集，第 760–770 页，2025年。Qian, Y., Lee, J., Duddu, S. M. K., Dai, Z., Brahma, S., Naim, I., Lei, T. 和 Zhao, V. Y. 多向量检索作为稀疏对齐。*arXiv* 预印本 *arXiv:2211.01267*，2022年。Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C. 和 Zaharia, M. ColBERTv2：通过轻量级后期交互实现高效检索。发表于*2022*年北美计算语言学协会分会年会：人类语言技术论文集，第 3715–3734 页，2022年。Shrestha, S., Reddy, N. 和 Li, Z. ESPN：内存高效的多向量信息检索。发表于*2024*年 *ACM SIGPLAN* 内存管理国际研讨会论文集，第 95–107 页，2024年。Team, N. Nomic Embed Multimodal：用于视觉文档检索的交错文本、图像和截图，2025年。网址：https://nomic.ai/blog/posts/nomic-embed-multimodal。

Teiletche, P., Mac´e, Q., Conti, M., Loison, A., Viaud, G., Colombo, P. 和 Faysse, M. ModernVBERT：迈向更小型的视觉文档检索器。*arXiv* 预印本 *arXiv:2510.01149*，2025年。Thirukovalluru, R., Meng, R., Liu, Y., Su, M., Nie, P., Yavuz, S., Zhou, Y., Chen, W., Dhingra, B. 等。通过智能批次挖掘打破对比学习的批次壁垒（B3）。*arXiv* 预印本 *arXiv:2505.11293*，2025年。Tishby, N., Pereira, F. C. 和 Bialek, W. 信息瓶颈方法。*arXiv* 预印本 *physics/0004057*，2000年。Wang, Q., Ding, R., Chen, Z., Wu, W., Wang, S., Xie, P. 和 Zhao, F. ViDoRAG：通过动态迭代推理智能体实现视觉文档检索增强生成。*arXiv* 预印本 *arXiv:2502.18017*，2025年。Wasserman, N., Pony, R., Naparstek, O., Goldfarb, A. R., Schwartz, E., Barzelay, U. 和 Karlinsky, L. Real-MM-RAG：真实世界多模态检索基准测试。*arXiv* 预印本 *arXiv:2502.12342*，2025年。Xiao, Z., Ma, Q., Gu, M., Chen, C.-c. J., Chen, X., Ordonez, V. 和 Mohan, V. MetaEmbed：扩展多模态

在测试时采用灵活的后期交互进行检索。*arXiv* 预印本 *arXiv:2509.18095*，2025年。

Xu, M., Moreira, G., Ak, R., Osmulski, R., Babakhin, Y., Yu, Z., Schifferer, B., and Oldridge, E. Llama Nemoretriever Colembed：性能领先的文本-图像检索模型。*arXiv* 预印本 *arXiv:2507.05513*，2025年。

Yan, Y., Xu, G., Zou, X., Liu, S., Kwok, J., and Hu, X. DocPruner：一种通过自适应块级嵌入剪枝实现多向量视觉文档检索的存储高效框架。*arXiv* 预印本 *arXiv:2509.23883*，2025年。

Yan, Y., Huo, J., Feng, G., Ou, M., Cao, Y., Zou, X., Liu, S., Lyu, Y., Huang, Y., Li, J., 等。解锁多模态文档智能：从当前成就到视觉文档检索的未来前沿。*arXiv* 预印本 *arXiv:2602.19961*，2026a。

Yan, Y., Ou, M., Cao, Y., Zou, X., Huo, J., Liu, S., Kwok, J., and Hu, X. 雕塑向量空间：通过先剪枝后合并框架实现高效的多向量视觉文档检索。*arXiv* 预印本 *arXiv:2602.19549*，2026b。Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., et al. VisRAG：面向多模态文档的基于视觉的检索增强生成。*arXiv* 预印本 *arXiv:2410.10594*，2024年。Zhang, J., Liu, Y., Wu, Z., Pang, G., Ye, Z., Zhong, Y., Ma, J., Wei, T., Xu, H., Chen, W., et al. MonkeyOCR v1.5 技术报告：解锁复杂模式下稳健的文档解析。*arXiv* 预印本 *arXiv:2511.10390*，2025a。Zhang, J., Zhang, Q., Wang, B., Ouyang, L., Wen, Z., Li, Y., Chow, K.-H., He, C., and Zhang, W. OCR 阻碍检索增强生成：评估 OCR 对检索增强生成的级联影响。载于 *IEEE/CVF*国际计算机视觉会议论文集，第 17443–17453 页，2025b。Zhang, Q., Wang, B., Huang, V. S.-J., Zhang, J., Wang, Z., Liang, H., He, C., and Zhang, W. 揭秘文档解析：结构化信息抽取的技术、挑战与前景。*arXiv* 预印本 *arXiv:2410.21169*，2024a。Zhang, Q., Zhang, J., Ren, Z., Ouyang, L., Wen, Z., Niu, J., Qu, Y., Wang, B., Chow, K.-H., He, C., et al. Docr-inspector：利用视觉语言模型对文档解析进行细粒度自动评估。*arXiv* 预印本 *arXiv:2512.10619*，2025c。

Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., 和 Zhang, M. GME：通过多模态大语言模型改进通用多模态检索。*arXiv*预印本 *arXiv:2412.16855*，2024b。

## A. 算法工作流程

我们将所提出的ColParse 框架的完整工作流程形式化为两个不同的算法。**算法1** 详细说明了离线索引过程，其中ColParse 通过其顺序三阶段流程生成一组高度紧凑的文档嵌入。随后，**算法2** 展示了在线检索阶段，其中最终相关性分数经由MaxSim 运算，使用这组压缩嵌入高效计算得出。

**算法 1** ColParse的离线索引过程

**输入：** 文档图像 $d \in \mathbb{R}^{H\times W\times 3}$;
文档解析器模型 $\Psi_{\text{parse}}$;
单向量编码器 $\Phi_{\text{enc}}$： $\mathbb{R}^{H'\times W'\times 3} \to \mathbb{R}^{D}$
**输出：** A 紧凑 多向量表示 $\mathbf{D}_{\texttt{ColParse}} \subset \mathbb{R}^{k\times D}$

/* 阶段 1：布局感知文档解析 */
$[\{b_j, c_j\}]_{j=1}^{k} \leftarrow \Psi_{\text{parse}}(d)$; // 获取 $k$ 边界框和内容类型
$\mathcal{S}_d \leftarrow \emptyset$ **对于** $j \leftarrow 1$ **到** $k$ **执行**
  $s_j \leftarrow$ 裁剪$(d,\ b_j)$; // 使用边界框 $b_j$裁剪文档图像 $d$
  $\mathcal{S}_d \leftarrow \mathcal{S}_d \cup \{s_j\}$
**end**
/* Stage 2: Dual-Stream Encoding */
$\mathbf{D}_{\text{local}} \leftarrow$ **对于每个子图像** $s_j \in \mathcal{S}_d$ **执行**
  $\mathbf{v}^{(j)}_{\text{局部}} \leftarrow \Phi_{\text{enc}}(s_j)$; // 编码局部区域
  $\mathbf{D}_{\text{local}} \leftarrow \mathbf{D}_{\text{local}} \cup \{\mathbf{v}^{(j)}_{\text{局部}}\}$
**end**
$\mathbf{v}_{\text{global}} \leftarrow \Phi_{\text{enc}}(d)$ ; // 编码整页以获取全局上下文
/* Stage 3: Global-Local Fusion */
$\mathbf{D}_{\texttt{ColParse}} \leftarrow \emptyset$ **对每个局部向量** $\mathbf{v}^{(j)}_{\text{局部}} \in \mathbf{D}_{local}$ **执行**
  $\mathbf{d}^{(j)}_{\text{融合}} \leftarrow \mathbf{v}^{(j)}_{\text{局部}} + \mathbf{v}_{\text{global}}$ ; // 通过逐元素加法融合相加
  $\mathbf{D}_{\texttt{ColParse}} \leftarrow \mathbf{D}_{\texttt{ColParse}} \cup \{\mathbf{d}^{(j)}_{\text{融合}}\}$
**end**
**返回** $\mathbf{D}_{\texttt{ColParse}}$

## B. 更多理论分析

本章节对第 3 章节中介绍的概念进行了详细的理论阐述，并依据ColParse 信息论的基本原理构建框架。

### B.1. 信息论预备知识

我们首先定义分析中使用的核心概念。

**定义 B.1**（互信息）。两个随机变量 $X$ 和 $Y$之间的互信息 $I(X;Y)$

算法2 在线检索流程 ColParse输入：一个文本查询 $q$; 一个预先计算的紧凑表示 $\mathbf{D}_{\texttt{ColParse}} = \{\mathbf{d}^{(j)}_{\text{融合}}\}_{j=1}^{k}$; 一个编码器 $\Phi_{\text{enc}}$输出：相关性分数 $s_{\texttt{ColParse}}(q,d)$ /* 步骤 1：编码查询 */ $\mathbf{Q} \leftarrow \Phi_{\text{enc}}(q)$; // 将 $q$ 编码为 $N_q$ 词元向量 $\{\mathbf{q}_i\}$/* 步骤 2：后期交互评分（MaxSim） */ $score \leftarrow 0$ 对于每个查询向量 $\mathbf{q}_i \in \mathbf{Q}$，执行 $max\ sim \leftarrow -\infty$ 对于每个融合文档向量 $\mathbf{d}^{(j)}_{\text{融合}} \in \mathbf{D}_{\texttt{ColParse}}$，执行 $sim \leftarrow \mathbf{q}_i^{\top}\mathbf{d}^{(j)}_{\text{融合}}$; // 假设向量已进行 L2归一化$max\ sim \leftarrow \max(max\ sim, sim)$ end $score \leftarrow score + max\ sim$ ; // 汇总最大相似度 end

**返回** $score$

衡量它们之间的相互依赖程度。其定义如下：

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}. \tag{7}$$

其中，$p(x,\ y)$ 是联合概率分布，$p(x)$ 和 $p(y)$ 是边缘分布。$I(X;Y) = 0$ 当且仅当 $X$ 和 $Y$ 相互独立时成立。

**定义 B.2**（条件互信息）。条件互信息 $I(X;Y|Z)$ 衡量在已知第三个变量 $Z$ 的情况下，$X$ 和 $Y$之间的互信息：

$$I(X;Y|Z) = \mathbb{E}_{z\sim p(z)}[I(X;Y|Z=z)]. \tag{8}$$

**定理 B.3**（互信息链式法则）。对于一组随机变量 $\{X_1,\ldots,X_n\}$ 和另一个变量 $Y$，链式法则表明：

$$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\ldots,X_{i-1}). \tag{9}$$

该法则是分解复杂系统信息内容的基础。

**定理 B.4**（数据处理不等式（DPI））。对于任意随机变量的马尔可夫链 $X \to Y \to Z$，其中在给定 $Y$ 的条件下 $Z$ 条件独立于 $X$，下列不等式成立：

$$I(X;Z) \le I(X;Y)\ and\ I(X;Z) \le I(Y;Z). \tag{10}$$

该定理形式化了这样一种观点：后处理（从 $Y$到 $Z$ 的步骤）无法增加关于原始来源 $X$ 的信息。

### B.2. 视觉文档检索中的信息瓶颈（IB）原理

如第 3.3 节所述，视觉文档检索的压缩任务可以表述为一个 IB 问题（Tishby 等人，2000）。目标是找到文档 $Z$ 的一种压缩表示 $D$，使其最大化关于相关性变量 $R$ 的信息，同时最小化关于来源 $D$ 本身的信息。

**难处理性证明。** IB 拉格朗日函数（主文中的式3 ）要求对所有可能查询的分布 $P$（$Q$）取期望。

$$\mathcal{L}(Z) = I(Z;D) - \beta \int_{q\in\mathcal{Q}} P(q) I(Z;R(D,q))dq. \quad (11)$$

由于在文档索引时， $P(Q)$未知且可能是无限的，因此无法直接优化这一目标。因此，实际方法必须依赖有原则的近似或替代目标来实现这一理想目标。`ColParse` 提供了这样一种替代方案。

### B.3. 结构解耦的合理性

`ColParse`的解析阶段 $\Psi_{\text{parse}}(D) = \{S_1, \ldots, S_k\}$，由语义集中公理提供依据。现在，我们给出更形式化的论证。

**公理 B.5** (语义集中)。 对于给定的查询$Q = q$，存在一个主要语义区域 $S_{j^*} \in \{S_j\}$，其中包含了判断相关性所需的几乎全部信息。其余区域 $S_{\neg j^*} = \{S_j\}_{j\neq j^*}$提供的额外信息微乎其微。

$$I(S_{\neg j^*}; R | S_{j^*}, Q = q) \approx 0. \quad (12)$$

论证。 该公理是关于用户查询和文档本质的一项经验性假设。对于查询“2023年第三季度的收入是多少？”，答案几乎肯定完全包含在某个财务表格中。一旦确定了正确的表格，其他区域（例如摘要、方法图）中的信息在条件意义上便无关紧要。

□

**推论 B.6** (分解表示的信息等价性)。 在语义集中公理下，整个文档与相关性变量之间的互信息，近似等于任一单独语义区域所包含的最大信息量。

$$I(D;R) \approx \max_{j\in\{1,\ldots,k\}} I(S_j;R). \quad (13)$$

证明。 根据链式法则， $I(D; R) = I(S_1, \ldots, S_k; R)$。对于特定查询 $q$，令 $j^*$ 为主要区域的索引。我们有：

$$I(D;R|Q=q) = I(S_{j^*};R|Q=q) + I(S_{\neg j^*};R|S_{j^*},Q=q). \quad (14)$$

应用公理B.5，第二项消失：$I(D;R|Q=q) \approx I(S_{j^*};R|Q=q)$。对 $P(Q)$ 取期望，并利用性质 $\mathbb{E}[\max(X_i)] \geq \max(\mathbb{E}[X_i])$，我们得到如下近似：总信息可由最佳可能信道中的信息很好地表示，从而为多向量方法提供了依据。 □

### B.4. 协同融合的合理性

融合阶段将局部向量 $\{V_j = \Phi_{\text{enc}}(S_j)\}$与全局向量 $V_{\text{global}} = \Phi_{\text{enc}}(D)$结合，生成最终表示 $\{Z_j = V_j + V_{\text{global}}\}$。

**定义 B.7**（上下文信息增益）。 对于区域 $j$，上下文信息增益是指在已知局部信息 $V_j$ 的情况下，全局上下文$V_{\text{global}}$提供的有关相关性 $R$ 的额外信息。

$$G_j^{\text{context}} \triangleq I(V_{\text{global}}; R | V_j). \quad (15)$$

**定理 B.8**（融合表示中的信息）。融合向量 $Z_j = V_j + V_{global}$ 中包含的信息不超过其组成部分的联合信息。

$$I(Z_j;R) \leq I(V_j, V_{global}; R). \quad (16)$$

证明。 融合向量 $Z_j$ 是$V_j$ 和 $V_{\text{global}}$的确定性函数。这构成马尔可夫链 $(V_j, V_{\text{global}}) \to Z_j \to R$。将数据处理不等式（定理B.4）应用于该链，即可直接得出结果。 □

**推论 B.9** (信息改善条件)。当且仅当融合函数成功捕获了非零部分的上下文信息增益时，融合步骤才有益（即， $Z_j$ 比单独的 $V_j$ 包含更多信息）。

$$\Delta I_j = I(Z_j;R) - I(V_j;R) > 0 \iff I(Z_j;R|V_j) > 0. \quad (17)$$

证明。 根据链式法则， $I(Z_j, V_j; R) = I(V_j; R) + I(Z_j; R|V_j)$。由于 $Z_j$ 是 $V_j$ 和 $V_{\text{global}}$的函数，已知 $V_j$ 并不能使 $Z_j$ 完全确定。项$I(Z_j; R|V_j)$表示：即使 $V_j$保持不变， $Z_j$ 中的变化（由 $V_{\text{global}}$引起）仍能提供关于 $R$的信息。要实现正的净改进 $\Delta I_j > 0$，就必须使该条件项为正；这反过来意味着融合过程必须编码了一部分上下文增益 $G_j^{\text{context}}$。向量加法 $V_j + V_{\text{global}}$ 是实现这一目的的一种简单而有效的函数，因为它会在点积评分过程中与查询向量发生非线性交互：$\mathbf{q}^\top(\mathbf{v}_j + \mathbf{v}^{\text{global}})$，从而使模型能够利用局部和全局信号。 □

# C. 更多实验分析

## C.1. 基准测试详情

为确保对我们框架进行全面且稳健的评估，我们以五个主流的视觉文档检索基准测试套件为实验基础；这些套件均集成于 MMEB 的 visdoc 章节中（Meng 等人，2025年）。以下基准测试共同涵盖了多种文档类型、复杂程度各异的查询以及不同的检索场景，从多个维度展现了模型性能。

► **ViDoRe-V1（Faysse 等人，2024年）[3]：** 作为页面级视觉文档检索的基础基准测试，ViDoRe-V1 是首批系统评估系统处理视觉丰富型文档能力的基准之一。它结合了重新利用的学术视觉问答数据集与实用的特定主题任务，凸显了传统纯文本检索系统在处理包含复杂布局、表格和图表的文档时固有的不足。

► **ViDoRe-V2（Mac´e 等人，2025年）[4]：** 作为 ViDoRe-V1 的后续版本，ViDoRe-V2 旨在通过引入更具挑战性、更贴近现实的检索场景来提高基准要求，以应对原始版本上观察到的性能饱和问题。其核心贡献包括采用通过合成与人在环混合流程生成的长篇、跨文档和多语言查询，从而减少抽取偏差，并更准确地反映真实世界中的用户交互。

► **VisRAG（Yu 等人，2024年）[5]：** VisRAG 基准测试旨在通过汇总并精炼多个现有视觉问答数据集，专门评估基于视觉的检索增强生成流程。其主要贡献在于将涵盖科学图表、图表、信息图和演示幻灯片在内的广泛文档类型统一到单一评估框架下，并结合关键的筛选流程，以去除依赖上下文的问题，确保其适用于开放检索任务。

► **ViDoSeek（Wang 等人，2025年）[6]：** ViDoSeek 是一项新颖的基准测试，旨在评估面向富含视觉信息的文档、需要复杂推理的端到端检索增强生成系统。它的主要贡献在于提供了一个大型文档语料库，其中每个查询都对应唯一答案，因此能够在大规模场景下，更真实且严格地评估检索及后续推理阶段。

► **MMLongBench（Ma 等人，2024b）[7]：** MMLongBench 专为评估大型视觉语言模型的长上下文、多模态理解能力而设计。它的突出特点是采用篇幅较长的文档（平均 47.5 页），并包含相当比例的跨页问题，这些问题需要多跳推理；同时还设置了无法回答的问题，以检测模型幻觉，从而严格检验模型从海量上下文中定位并综合信息的能力。

## C.2. 模型详情

我们从近期文献中选取十个具有代表性的单向量多模态检索模型，作为实验的基础模型。这些模型基于多种架构和预训练范式构建，为评估我们所提出框架的通用性与有效性提供了全面的测试平台。

► **VLM2Vec-V1-2B/7B（Jiang 等人，2024年）[8]：** 作为通用多模态嵌入领域的先驱性工作，VLM2Vec 引入了一种对比训练框架，使任意视觉语言模型都能适用于广泛的任务。其核心贡献是将多种多模态任务（例如，分类、视觉问答、检索）重新表述为统一的指令遵循式排序问题，使模型能够为图像和文本学习通用嵌入。

► **VLM2Vec-V2-2B（Meng 等人，2025年）[9]：** 该模型在前代基础上扩展了多模态嵌入的适用范围，将视频和视觉文档纳入其中，同时涵盖图像和文本。其主要贡献是引入了更全面的基准测试和统一的训练策略，使单个模型能够有效学习静态、时序和结构化视觉数据格式的表征。

► **LamRA-Ret-7B（Liu 等人，2025a）[10]：** LamRA 探索将生成式大型多模态模型重新用于检索任务，并在单一指令跟随框架下统一多种检索场景。其关键创新在于采用两阶段训练策略：首先在纯语言任务上对模型进行预训练，然后进行多模态指令微调，逐步使生成式模型适应检索。

► **GME-2B/7B（Zhang 等人，2024b）[11]：** 通用多模态嵌入器（GME）框架专注于

[3] https://huggingface.co/collections/vidore/vidore-benchmark
[4] https://huggingface.co/collections/vidore/vidore-benchmark-v2
[5] https://huggingface.co/collections/ openbmb/visrag
[6] https://huggingface.co/datasets/Qiuchen-Wang/ViDoSeek
[7] https://huggingface.co/datasets/yubo2333/MMLongBench-Doc
[8] https://huggingface.co/TIGER-Lab/ VLM2Vec-Full
[9] https://huggingface.co/VLM2Vec/ VLM2Vec-V2.0
[10] https://huggingface.co/code-kunkun/ LamRA-Ret
[11] https://huggingface.co/collections/Alibaba-NLP/gme-models

通过利用更多样化的训练数据组合来提升通用多模态检索能力，其中包括单模态、跨模态和融合模态示例。其得分贡献在于提出了一种新颖的数据合成流程，用于创建大规模、高质量的融合模态数据，从而显著增强模型处理复杂查询和检索视觉文档的能力。

► **UniME-V2-2B/7B（Gu 等人，2025年）[12]：** UniME-V2 通过利用多模态大语言模型作为“评审”，为查询-候选项对生成软语义匹配得分，从而增强表示学习。这种多模态大语言模型评审机制有助于更有效地挖掘困难负例，使嵌入模型能够学习更细粒度的语义差异，显著提升其判别能力。

► **B3-2B/7B（Thirukovalluru 等人，2025年）[13]：** 打破批次壁垒（ B3 ）提出了一种新颖的批次构建策略，用于对比学习，通过精心构建富含困难负例的高质量批次来提升训练效果。它不采用随机采样，而是利用教师模型和基于图的社区检测，将相互构成挑战的样本归为一组，从而提高训练效率，即使批次规模显著缩小，也能达到最先进的性能。

### C.3. MinerU2.5 详情

为解决直接处理高分辨率文档图像所带来的巨大计算开销（$O(N^2)$ 复杂度）与信息损失之间的权衡问题，MinerU2.5 创新性地采用了一种由粗到细的两阶段解耦策略：

1. **第一阶段：全局布局分析。** 在此阶段，模型首先将输入文档图像缩放为中等分辨率的缩略图（例如，1036×1036 像素）。随后，模型在该缩略图上快速执行全局布局分析，以较低的计算成本识别所有结构元素（例如段落、表格、公式和图表）及其位置。

2. **第二阶段：局部内容识别。** 根据第一阶段检测到的布局信息，模型从**原始高分辨率**图像中精确裁剪出相应的语义区域。随后，模型在这些原生分辨率的裁剪图像块上并行执行精细内容识别（例如，文本光学字符识别、表格结构化、公式转录）。这种方法既能保持较高的识别准确率，又可避免对整幅高分辨率图像进行冗余计算。

**算法3**详细介绍布局引导的图像切分 `ColParse`中使用的流程。

**算法 3** 布局引导的图像切分 for `ColParse`

**输入：** 一张文档图像 $d \in \mathbb{R}^{H\times W\times 3}$；
一个布局检测器模型 $\Psi_{\text{split}}$（例如 DocLayoutY-OLO）；
最小面积比阈值 $\tau$（默认值为 0.01）；
子图像数量上限 $N_{\max}$（默认值为 20）；
网格回退参数： $R_{\text{grid}}$， $C_{\text{grid}}$
**输出：** 裁剪后的子图像列表 $\mathcal{S}_d$；
内容类型标签列表 $\mathcal{C}_d$（可选）

/* 步骤 1：语义布局检测 */
总面积 $\leftarrow H \times W$ $\mathcal{B} \leftarrow \emptyset, \mathcal{C}_d \leftarrow \emptyset$
**如果** $\Psi_{split}$ 可用**则**
  $\mathcal{R} \leftarrow \Psi_{\text{split}}$.预测($d$)； // 返回 {边界框的列表, 类别、得分}
  **如果** $\mathcal{R}$ 不为空 **则**
    **对于**每个区域 $r \in \mathcal{R}$**执行**
      $b \leftarrow (x_1, y_1, x_2, y_2)$ 来自 $r$.多边形 $c \leftarrow$ 映射类别 ID($r$.类别 ID) $\mathcal{B} \leftarrow \mathcal{B} \cup \{(b,\ c,\ \text{centerY}(b),\ \text{centerX}(b))\}$
    **end**
  **end**
**end**

/* 步骤 2：回退与排序机制 */
**如果** $\mathcal{B}$ 为空**则**
  $\mathcal{B} \leftarrow$ 基于网格的分割($H, W, R_{\text{grid}}, C_{\text{grid}}$)； // 回退到网格
**else**
  $\mathcal{B} \leftarrow$ 按阅读顺序排序($\mathcal{B}$)； // 按垂直方向带，然后按水平方向
**end**

/* 步骤 3：筛选、裁剪和输出 */
$\mathcal{S}_d \leftarrow \emptyset$，计数 $\leftarrow 0$ **对于**每个$(b,\ c) \in \mathcal{B}$**执行**
  **如果**计数 $\geq N_{max}$ **then**
    **中断**
  **end**
  区域 $\leftarrow$ 宽度($b$) ×高度($b$) **如果** 区域/总面积 $\geq \tau$ **then**
    $s \leftarrow$ 裁剪($d, b$)； // 从原始图像中提取区域图像
    $\mathcal{S}_d \leftarrow \mathcal{S}_d \cup \{s\}$ $\mathcal{C}_d \leftarrow \mathcal{C}_d \cup \{c\}$ 计数 $\leftarrow$ 计数$+1$
  **end**
**end**
**返回** $\mathcal{S}_d, \mathcal{C}_d$

### C.4. 主要结果

请参阅表2和表3，了解以下所有结果： `ColParse` 以及五个基准测试中的基线方法。

[12] https://huggingface.co/collections/TianchengGu/unime-v2 [13] https://huggingface.co/raghavlite

表2。 在 MMLongBench 和 ViDoRe-V1 基准测试上的性能比较。对于每个模型区块，我们将每列中表现最佳的优化方法加粗（基础结果除外）。优化方法的平均分及其相对于基准模型的相对增益见下（↑/↓）。

| 方法 | MMLongBench | | ViDoRe-V1 | |
|---|---|---|---|---|
| | 文档 | Avg. | arXiv | Avg. |
| VLM2Vec-V1-2B | 25.62 26.23 | 25.93 | 17.80 13.98 39.41 9.18 36.32 10.56 16.39 15.96 23.56 24.11 | 20.73 |
| s2m-add | 21.54 | 18.31↓7.62 | 35.07 6.62 36.51 33.61 | 27.77↑7.04 |
| s2m-mul | 22.07 15.10 | 18.59↓7.34 | 34.91 16.12 52.61 6.57 36.65 10.34 23.90 30.89 28.88 33.60 | 27.45↑6.72 |
| cl-t-c | 16.57 | | 13.97 3.97 21.62 8.73 23.20 26.72 | 17.01↓3.72 |
| cl-t-m | 14.35 8.67 | 11.51↓14.42 | 16.47 2.98 26.97 16.27 18.48 8.59 14.40 23.03 18.71 13.10 | 15.90↓4.83 |
| cl-s-c | 18.29 | | 18.83 4.87 22.80 20.85 | 17.92↓2.81 |
| cl-s-m | 15.45 9.06 | 12.26↓13.67 | 15.73 2.54 27.46 15.68 18.99 9.76 13.52 24.73 22.59 20.72 | 17.17↓3.56 |
| c-sem | 18.76 | 16.32↓9.61 | 29.33 5.87 35.70 30.44 | 26.88↑6.15 |
| 多图像 | 23.61 15.85 | 19.73↓6.20 | 37.78 13.96 54.20 11.35 40.50 9.10 20.36 32.15 36.75 32.04 | **28.82**↑8.09 |
| ColParse | **34.31** | | **47.66** | |
| VLM2Vec-V1-7B | 23.85 37.63 | 30.74 | 28.07 17.93 44.47 2.06 16.78 5.86 17.93 25.04 28.90 14.59 | 20.16 |
| s2m-add | 34.57 | | 50.30 66.38 | |
| s2m-mul | 35.29 28.49 | 31.89↑1.15 | 50.46 26.34 67.28 36.45 64.13 23.84 69.43 61.94 68.35 66.99 | 53.52↑33.36 |
| cl-t-c | 18.75 | | 17.10 7.71 26.41 26.65 | 21.63↑1.47 |
| cl-t-m | 25.34 20.46 | 22.90↓7.84 | 28.06 7.00 47.43 29.48 45.92 19.42 27.66 47.56 44.96 52.94 | 35.04↑14.88 |
| cl-s-c | 22.25 | | 21.80 8.11 30.92 32.58 | 27.12↑6.96 |
| cl-s-m | 26.31 20.38 | 23.35↓7.39 | 28.58 8.85 49.12 32.67 46.07 20.85 31.96 50.86 50.85 49.99 | 36.98↑16.82 |
| c-sem | 31.36 | | 45.77 66.76 | |
| 多图像 | 33.77 25.60 | 29.69↓1.05 | 49.40 19.55 62.09 28.19 66.34 17.19 41.89 51.44 60.84 48.89 | 44.58↑24.42 |
| ColParse | **43.34** | | **60.47** | |
| VLM2Vec-V2-2B | 48.55 50.34 | 49.45 | 78.98 38.51 82.21 64.57 87.64 44.68 85.06 82.99 89.89 87.08 | 74.16 |
| s2m-add | 43.33 | | 66.60 80.36 | 69.59↓4.57 |
| s2m-mul | 45.72 40.66 | 43.19↓6.26 | 68.04 39.80 75.45 58.87 69.90 54.31 90.93 84.53 84.56 82.40 | 70.88↓3.28 |
| cl-t-c | 20.08 | | 28.47 6.46 29.32 22.54 | |
| cl-t-m | 25.04 21.94 | 23.49↓25.96 | 44.82 7.28 42.98 31.23 38.96 21.19 26.38 47.14 45.34 37.44 | 34.28↓39.88 |
| cl-s-c | 23.25 | | 29.34 8.76 27.77 27.64 | |
| cl-s-m | 26.08 22.30 | 24.19↓25.26 | 44.95 7.16 45.42 19.48 39.29 25.44 31.57 47.18 48.84 40.13 | 34.95↓39.21 |
| c-sem | 29.94 | | 61.30 70.10 | |
| 多图像 | 38.69 29.00 | 33.85↓15.60 | 65.39 25.80 70.54 27.92 71.56 33.79 55.93 62.93 72.27 53.14 | 53.93↓20.23 |
| ColParse | **49.49** | | **80.17** | **78.41**↑4.25 |
| LamRA-Ret | 19.78 13.24 | 16.51 | 29.31 19.56 63.00 15.83 51.44 7.70 21.10 29.81 37.18 31.95 | 30.69 |
| s2m-add | **32.18** | | 9.80 14.37 24.81 | 25.36↓5.33 |
| s2m-mul | 30.52 16.37 | 23.45↑6.94 | 9.71 14.98 45.61 17.37 27.91 17.20 20.45 28.32 38.60 23.01 | 24.32↓6.37 |
| cl-t-c | 14.88 8.09 | 11.49↓5.02 | 5.57 2.22 19.76 6.59 14.23 20.52 | 13.89↓16.80 |
| cl-t-m | 20.43 13.28 | 16.86↑0.35 | 7.47 3.85 30.91 17.54 13.41 12.37 27.06 35.63 39.20 31.88 | 21.93↓8.76 |
| cl-s-c | 15.32 9.05 | 12.19↓4.32 | 6.78 2.62 20.53 9.11 17.05 22.25 | 16.31↓14.38 |
| cl-s-m | 19.84 12.91 | 16.38↓0.13 | 7.75 4.82 32.16 15.61 14.02 10.23 24.81 34.66 37.31 27.20 | 20.86↓9.83 |
| c-sem | 19.62 | 15.16↓1.35 | 13.51 5.91 34.43 9.13 32.30 **37.01** | 25.04↓5.65 |
| 多图像 | 23.32 7.71 | 15.52↓0.99 | 6.90 6.10 21.26 8.00 22.26 13.60 13.91 15.91 20.30 9.67 | 13.79↓16.90 |
| ColParse | 30.74 | | **17.27** 26.29 | **30.10**↓0.59 |
| GME-2B | 52.07 53.14 | 52.61 | 82.59 56.46 88.97 89.72 93.20 70.33 98.49 92.15 98.15 95.65 | 86.57 |
| s2m-add | 50.92 | | 71.85 | 78.12↓8.45 |
| s2m-mul | **53.82**50.06 | 51.94↓0.67 | 76.87 47.23 85.49 81.53 84.77 **74.63**95.72 **93.07** 93.85 92.18 | 82.53↓4.04 |
| cl-t-c | 15.54 | | 14.96 4.70 25.32 19.34 | 16.85↓69.72 |
| cl-t-m | 17.95 13.61 | 15.78↓36.83 | 30.20 3.58 33.07 30.13 31.16 16.11 22.27 33.99 36.32 33.28 | 27.01↓59.56 |
| cl-s-c | 16.03 | | 15.90 5.88 22.23 22.00 | 19.84↓66.73 |

下页继续

表2 –接上页

| 方法 | MMLongBench | | | ViDoRe-V1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 文档 | | Avg. | arXiv | | | | | | | | | | Avg. |
| cl-s-m | 16.39 | 12.39 | 14.39↓38.22 | 31.79 | 4.01 | 30.55 | 19.79 | 31.61 | 14.18 | 24.92 | 30.34 | 34.24 | 28.50 | 24.99↓61.58 |
| c-sem | 23.78 | | | 45.19 | | | | | | | | | 53.58 | |
| 多图像 | 45.10 | 32.87 | 38.99↓13.62 | 67.63 | 27.89 | 74.78 | 47.02 | 78.34 | 43.20 | 59.99 | 63.76 | 73.68 | 63.19 | 59.95↓26.62 |
| ColParse | 53.06 | | | **82.39** | | | | | | | | | | **86.11**↓0.46 |
| GME-7B | 54.01 | 55.80 | 54.91 | 87.59 | 56.05 | 91.96 | 94.25 | 93.72 | 76.26 | 99.63 | 95.45 | 99.63 | 99.06 | 89.36 |
| s2m-add | 53.57 | | | 75.91 | | | | | | | | | 92.17 | 82.13↓7.23 |
| s2m-mul | 50.73 | 45.95 | 48.34↓6.57 | 72.29 | 43.38 | 83.23 | 72.85 | 78.28 | 69.64 | 92.92 | 90.23 | 90.05 | 90.17 | 78.30↓11.06 |
| cl-t-c | 12.53 | 8.39 | 10.46↓44.45 | 7.32 | 5.16 | 17.36 | | | | | | | 13.97 | 14.21↓75.15 |
| cl-t-m | 14.64 | 9.03 | 11.84↓43.07 | 6.78 | 3.80 | 26.36 | 24.67 | 17.82 | 15.52 | 19.12 | 26.65 | 24.74 | 28.60 | 19.41↓69.95 |
| cl-s-c | 13.08 | 8.93 | 11.01↓43.90 | 7.25 | 3.85 | 16.83 | | | | | | | 21.64 | 15.77↓73.59 |
| cl-s-m | 13.62 | 7.93 | 10.78↓44.13 | 6.88 | 4.73 | 22.77 | 18.40 | 18.43 | 10.43 | 16.80 | 21.03 | 26.40 | 20.86 | 16.67↓72.69 |
| c-sem | 15.42 | 9.22 | 12.32↓42.59 | 15.02 | 6.13 | 27.52 | | | | | | | 26.13 | 23.21↓66.15 |
| 多图像 | 47.50 | 36.01 | 41.76↓13.15 | 72.48 | 33.71 | 78.84 | 45.50 | 84.88 | 45.65 | 64.59 | 68.82 | 75.90 | 69.50 | 63.99↓25.37 |
| ColParse | **54.96** | | | **87.35** | | | | | | | | | | **89.56**↑0.20 |
| UniME-V2-2B | 18.52 | 40.10 | 29.31 | 36.52 | 12.43 | 42.41 | 14.09 | 51.11 | 7.39 | 20.23 | 32.96 | 24.21 | 19.25 | 26.06 |
| s2m-add | 36.66 | | | 50.78 | | | | | | | | | | |
| s2m-mul | 36.06 | 28.97 | 32.52↑3.21 | 50.76 | 23.93 | 58.23 | 32.85 | 54.90 | 34.94 | 63.01 | 64.67 | 66.31 | 61.07 | 51.07↑25.01 |
| cl-t-c | 19.45 | | | 19.03 | 7.04 | 25.79 | | | | | | | 26.67 | 21.39↓4.67 |
| cl-t-m | 16.70 | 12.63 | 14.67↓14.64 | 21.88 | 4.30 | 33.79 | 26.88 | 20.47 | 9.65 | 15.83 | 36.40 | 34.18 | 27.10 | 23.05↓3.01 |
| cl-s-c | 19.23 | | | 19.53 | 6.89 | 23.75 | | | | | | | 27.73 | 22.61↓3.45 |
| cl-s-m | 17.77 | 13.19 | 15.48↓13.83 | 23.67 | 5.33 | 36.60 | 32.03 | 20.17 | 11.46 | 19.46 | 37.16 | 36.87 | 26.76 | 24.95↓1.11 |
| c-sem | 25.11 | | 23.00↓6.31 | 38.24 | | | | | | | | | 52.80 | |
| 多图像 | 33.99 | 24.39 | 29.19↓0.12 | 50.79 | 19.08 | 60.61 | 30.15 | 58.91 | 24.36 | 43.08 | 46.76 | 61.06 | 50.79 | 44.56↑18.50 |
| ColParse | **44.22** | | | **62.39** | | | | | | | | | | |
| UniME-V2-7B | 33.19 | 45.72 | 39.46 | 63.23 | 24.91 | 65.25 | 11.16 | 41.54 | 14.18 | 41.89 | 40.56 | 57.44 | 42.78 | 40.29 |
| s2m-add | 40.28 | | | 55.17 | | | | | | | | | 89.16 | |
| s2m-mul | 40.57 | 38.14 | 39.36↓0.10 | 55.29 | 33.72 | 70.38 | 56.55 | 65.82 | 44.48 | 84.10 | 81.55 | 86.00 | **89.89** | 66.78↑26.49 |
| cl-t-c | 20.22 | | | 24.63 | 5.79 | 29.62 | | | | | | | 28.61 | 23.34↓16.95 |
| cl-t-m | 23.20 | 19.90 | 21.55↓17.91 | 34.15 | 5.21 | 49.47 | 34.87 | 32.02 | 18.91 | 35.36 | 47.97 | 43.55 | 56.22 | 35.77↓4.52 |
| cl-s-c | 20.91 | | | 23.98 | 8.28 | 28.30 | | | | | | | 43.46 | 26.91↓13.38 |
| cl-s-m | 24.81 | 19.71 | 22.26↓17.20 | 35.09 | 6.60 | 49.60 | 36.77 | 33.75 | 20.75 | 37.45 | 48.31 | 55.85 | 60.28 | 38.45↓1.84 |
| c-sem | 31.71 | | | 59.37 | | | | | | | | | 75.85 | |
| 多图像 | 34.79 | 28.05 | 31.42↓8.04 | 53.99 | 23.06 | 67.47 | 40.81 | 70.58 | 30.36 | 52.38 | 65.49 | 72.19 | 62.65 | 53.90↑13.61 |
| ColParse | **45.90** | | | **64.78** | | | | | | | | | | |
| B3-2B | 37.10 | 32.07 | 34.59 | 57.00 | 29.38 | 68.09 | 48.31 | 71.55 | 18.09 | 74.13 | 64.64 | 75.44 | 63.13 | 56.98 |
| s2m-add | 35.86 | | | 47.93 | | | | | | | | | 65.45 | 52.60↓4.38 |
| s2m-mul | 35.66 | 26.97 | 31.32↓3.27 | 47.89 | 24.47 | 64.82 | 35.42 | 49.59 | 32.43 | 63.95 | 63.09 | 66.79 | 64.47 | 51.29↓5.69 |
| cl-t-c | 17.88 | | | 19.64 | 4.53 | 32.73 | | | 9.10 | 12.68 | | | 18.64 | 18.58↓38.40 |
| cl-t-m | 18.36 | 14.95 | 16.66↓17.93 | 18.38 | 5.34 | 33.44 | 13.13 | 19.91 | 12.42 | 14.05 | 32.81 | 30.84 | 26.31 | 20.66↓36.32 |
| cl-s-c | 20.70 | | | 16.92 | 5.67 | 34.77 | | | | | | | 36.75 | 24.07↓32.91 |
| cl-s-m | 20.41 | 17.63 | 19.02↓15.57 | 19.74 | 4.72 | 35.90 | 17.36 | 20.74 | 15.77 | 19.76 | 38.43 | 36.44 | 33.89 | 24.28↓32.70 |
| c-sem | 23.80 | | | 34.98 | | | | | | | | | 46.73 | 35.58↓21.40 |
| 多图像 | 28.93 | 16.01 | 22.47↓12.12 | 41.03 | 13.38 | 46.20 | 20.46 | 46.89 | 10.26 | 31.74 | 38.83 | 32.44 | 32.80 | 31.40↓25.58 |
| ColParse | **42.06** | | | **56.47** | | | | | | | | | | **62.74**↑5.76 |
| B3-7B | 46.09 | 45.10 | 45.60 | 68.95 | 43.38 | 79.86 | 66.56 | 84.12 | 37.06 | 81.01 | 81.25 | 88.57 | 81.30 | 71.21 |
| s2m-add | 44.95 | | | 59.11 | | | | | | | | | 86.34 | 70.79↓0.42 |
| s2m-mul | 45.11 | 40.61 | 42.86↓2.74 | 59.50 | 38.63 | 75.74 | 69.25 | 71.18 | **51.72** | 87.36 | 82.07 | 85.17 | 86.23 | 70.69↓0.52 |
| cl-t-c | 23.96 | | | 25.53 | 8.05 | 47.76 | | | | | | | 33.61 | |

续见下页

表2 –接上一页

| 方法 | MMLongBench | | ViDoRe-V1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 文档 | Avg. | arXiv | | | | | | | | | Avg. |
| cl-t-m | 24.79 | 21.29 | 23.04$_{\downarrow 22.56}$ | 31.72 | 9.21 | 52.63 | 18.92 | 28.84 | 22.25 | 31.48 | 47.83 | 51.71 | 43.89 | 33.85$_{\downarrow 37.36}$ |
| cl-s-c | 25.66 | | | 24.40 | | | | | | | | | 52.12 | |
| cl-s-m | 25.52 | 20.72 | 23.12$_{\downarrow 22.48}$ | 24.78 | 11.31 | 49.68 | 26.51 | 32.71 | 26.20 | 37.63 | 46.29 | 51.53 | 44.47 | 35.11$_{\downarrow 36.10}$ |
| c-sem | 29.88 | | | 56.08 | | | | | | | | | 70.35 | 51.13$_{\downarrow 20.08}$ |
| 多图像 | 35.37 | 22.45 | 28.91$_{\downarrow 16.69}$ | 52.60 | 22.38 | 56.80 | 27.96 | 65.24 | 16.05 | 53.57 | 47.96 | 43.69 | 51.39 | 43.76$_{\downarrow 27.45}$ |
| ColParse | **49.11** | | | **67.68** | | | | | | | | | | **74.41**$_{\uparrow 3.20}$ |

表3. 在 ViDoRe-V2、ViDoSeek 和 VisRAG 基准测试上的性能比较。对于每个模型区块，我们将各列中表现最佳的优化方法加粗（基础结果除外）。优化方法的平均分及相对增益（↑/↓）均以基准模型为参照。

| 方法 | ViDoRe-V2 | | | | | ViDoSeek | | | VisRAG | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bio-L | Eco-R | ESG-H | ESG-M | Avg. | Doc | Page | Avg. | arXiv | 图表 | InfoV | MP-Doc | Plot | 幻灯片 | Avg. |
| VLM2Vec-V1-2B | 6.88 | 14.15 | 12.25 | 20.54 | 13.46 | 56.40 | 67.73 | 62.07 | 41.68 | 58.21 | 70.79 | 42.74 | 23.83 | 74.07 | 51.89 |
| s2m-add | 10.40 | 8.59 | 4.68 | 3.46 | 6.78↓6.68 | 40.78 | 31.23 | 36.01↓26.06 | 28.94 | 44.79 | 59.76 | 30.86 | 9.67 | 59.59 | 38.94↓12.95 |
| s2m-mul | 10.00 | 8.40 | 5.09 | 3.51 | 6.75↓6.71 | 40.63 | 31.00 | 35.82↓26.25 | 28.85 | 44.64 | 60.48 | 30.55 | 9.83 | 59.52 | 38.98↓12.91 |
| cl-t-c | 14.86 | 14.49 | 4.14 | 4.48 | 9.49↓3.97 | 42.17 | 31.65 | 36.91↓25.16 | 13.56 | 18.52 | 25.66 | 13.11 | 1.41 | 31.91 | 17.36↓34.53 |
| cl-t-m | 13.34 | 4.60 | 8.78 | 5.15 | 7.97↓5.49 | 36.71 | 26.80 | 31.76↓30.31 | 8.29 | 18.27 | 37.74 | 11.85 | 1.13 | 32.46 | 18.29↓33.60 |
| cl-s-c | 16.38 | 10.03 | 4.93 | 4.64 | 9.00↓4.46 | 43.89 | 31.95 | 37.92↓24.15 | 13.03 | 27.01 | 28.61 | 18.82 | 1.54 | 31.31 | 20.05↓31.84 |
| cl-s-m | 13.46 | 6.50 | 6.89 | 3.74 | 7.65↓5.81 | 37.20 | 26.86 | 32.03↓30.04 | 8.21 | 22.04 | 37.37 | 12.65 | 1.15 | 33.31 | 19.12↓32.77 |
| c-sem | 20.32 | 11.01 | 11.86 | 10.44 | 13.41↓0.05 | 56.19 | 47.96 | 52.08↓9.99 | 16.53 | 44.39 | 43.88 | 24.77 | 3.86 | 54.32 | 31.29↓20.60 |
| 多图像 | 15.59 | 14.90 | 5.70 | 5.95 | 10.54↓2.92 | 45.24 | 34.26 | 39.75↓22.32 | 30.20 | 41.53 | 60.40 | 29.53 | 8.44 | 60.10 | 38.37↓13.52 |
| **ColParse** | **30.33** | **29.55** | **33.21** | **38.33** | **32.86**↑19.40 | **75.23** | **70.19** | **72.71**↑10.64 | **38.18** | **60.09** | **69.44** | **48.29** | **18.83** | **76.95** | **51.96**↑0.07 |
| VLM2Vec-V1-7B | 4.93 | 13.74 | 6.82 | 11.27 | 9.19 | 54.26 | 77.39 | 65.83 | 52.58 | 69.83 | 71.43 | 52.86 | 34.24 | 73.22 | 59.03 |
| s2m-add | 29.49 | 38.26 | 31.73 | 22.80 | 30.57↑21.38 | 62.50 | 53.64 | 58.07↓7.76 | 45.56 | 48.98 | 68.39 | 49.01 | 11.01 | 70.61 | 48.93↓10.10 |
| s2m-mul | 28.79 | 37.40 | 29.66 | 21.43 | 29.32↑20.13 | 62.34 | 52.73 | 57.54↓8.29 | 45.29 | 50.21 | 69.10 | 49.54 | 11.02 | 70.82 | 49.33↓9.70 |
| cl-t-c | 17.52 | 18.22 | 11.68 | 8.82 | 14.06↑4.87 | 49.02 | 40.25 | 44.64↓21.19 | 15.12 | 21.70 | 36.61 | 19.07 | 2.22 | 37.89 | 22.10↓36.93 |
| cl-t-m | 22.08 | 14.23 | 24.74 | 15.61 | 19.17↑9.98 | 58.38 | 51.13 | 54.76↓11.07 | 18.75 | 30.49 | 58.40 | 22.24 | 3.90 | 52.32 | 31.02↓28.01 |
| cl-s-c | 19.74 | 22.03 | 14.36 | 12.74 | 17.22↑8.03 | 51.57 | 41.97 | 46.77↓19.06 | 15.24 | 17.35 | 36.59 | 24.24 | 1.79 | 35.59 | 21.80↓37.23 |
| cl-s-m | 23.08 | 16.08 | 20.53 | 14.47 | 18.54↑9.35 | 58.59 | 50.84 | 54.72↓11.11 | 18.76 | 31.65 | 58.91 | 24.90 | 3.10 | 52.63 | 31.66↓27.37 |
| c-sem | 29.88 | 31.92 | 30.24 | 20.34 | 28.10↑18.91 | 72.03 | 66.00 | 69.02↑3.19 | 39.77 | 57.94 | 63.94 | 41.41 | 14.39 | 69.66 | 47.85↓11.18 |
| 多图像 | 30.91 | 39.93 | 23.80 | 17.07 | 27.93↑18.74 | 62.03 | 49.71 | 55.87↓9.96 | 45.32 | 44.84 | 64.95 | 40.71 | 10.55 | 67.95 | 45.72↓13.31 |
| **ColParse** | **42.63** | **42.89** | **50.55** | **42.86** | **44.73**↑35.54 | **78.34** | **78.61** | **78.48**↑12.65 | **54.43** | **70.30** | **69.27** | **58.49** | **33.46** | **77.98** | **60.66**↑1.63 |
| VLM2Vec-V2-2B | 44.45 | 45.77 | 48.77 | 46.98 | 46.49 | 80.88 | 83.68 | 82.28 | 77.38 | 82.30 | 86.27 | 71.60 | 66.96 | 92.04 | 79.43 |
| s2m-add | 41.02 | 49.68 | 41.29 | 20.26 | 38.06↓8.43 | 69.01 | 64.83 | 66.92↓15.36 | 62.37 | 73.06 | 78.41 | 75.52 | 22.67 | 85.73 | 66.29↓13.14 |
| s2m-mul | 42.43 | 50.96 | 42.90 | 21.91 | 39.55↓6.94 | 71.66 | 66.41 | 69.04↓13.24 | 63.30 | 73.50 | 79.84 | 76.41 | 23.36 | 87.29 | 67.28↓12.15 |
| cl-t-c | 19.84 | 21.50 | 12.92 | 9.75 | 16.00↓30.49 | 48.48 | 49.99 | 49.24↓33.04 | 25.86 | 22.73 | 34.54 | 19.89 | 10.59 | 36.62 | 25.04↓54.39 |
| cl-t-m | 25.10 | 18.29 | 15.42 | 12.37 | 17.80↓28.69 | 57.69 | 59.96 | 58.83↓23.45 | 37.64 | 42.56 | 56.47 | 24.76 | 11.54 | 51.93 | 37.48↓41.95 |
| cl-s-c | 23.22 | 18.36 | 10.47 | 12.71 | 16.19↓30.30 | 49.73 | 51.97 | 50.85↓31.43 | 28.13 | 27.93 | 34.99 | 25.08 | 10.77 | 40.66 | 27.93↓51.50 |
| cl-s-m | 27.68 | 18.68 | 17.65 | 9.93 | 18.49↓28.00 | 57.31 | 61.50 | 59.41↓22.87 | 37.55 | 43.07 | 55.65 | 26.56 | 11.74 | 52.27 | 37.81↓41.62 |
| c-sem | 36.74 | 31.11 | 20.49 | 21.25 | 27.40↓19.09 | 71.91 | 71.47 | 71.69↓10.59 | 53.17 | 61.34 | 67.94 | 53.96 | 23.78 | 74.81 | 55.83↓23.60 |
| 多图像 | 30.01 | 45.94 | 23.58 | 17.56 | 29.27↓17.22 | 63.17 | 50.95 | 57.06↓25.22 | 62.83 | 61.38 | 75.92 | 54.34 | 24.62 | 75.43 | 59.09↓20.34 |
| **ColParse** | **50.06** | **53.76** | **57.41** | **46.40** | **51.91**↓5.42 | **80.94** | **83.87** | **82.41**↑0.13 | **77.18** | **78.05** | **84.37** | **78.07** | **58.74** | **91.95** | **78.06**↓1.37 |
| LamRA-Ret | 10.75 | 9.65 | 6.32 | 11.18 | 9.48 | 60.17 | 28.81 | 44.49 | 11.17 | 63.50 | 59.78 | 33.57 | 29.42 | 57.59 | 42.51 |
| s2m-add | 12.36 | **21.34** | 18.02 | 21.30 | 18.26↑8.78 | 55.39 | 27.91 | 41.65↓2.84 | 2.75 | 33.31 | 44.51 | 28.61 | 4.86 | 49.08 | 27.19↓15.32 |
| s2m-mul | 9.91 | 17.36 | 21.00 | 20.88 | 17.29↑7.81 | 52.53 | 25.95 | 39.24↓5.25 | 2.75 | 33.53 | 43.10 | 28.28 | 5.19 | 49.10 | 26.99↓15.52 |
| cl-t-c | 6.69 | 7.44 | 5.85 | 1.05 | 5.26↓4.22 | 39.52 | 30.63 | 35.08↓9.41 | 4.26 | 19.69 | 21.94 | 9.90 | 1.53 | 26.54 | 13.98↓28.53 |
| cl-t-m | 7.75 | 10.84 | 8.20 | 3.97 | 7.69↓1.79 | 44.87 | 38.69 | 41.78↓2.71 | 5.46 | 22.85 | 38.15 | 17.45 | 2.13 | 35.03 | 20.18↓22.33 |
| cl-s-c | 7.57 | 6.57 | 6.90 | 1.97 | 5.75↓3.73 | 38.41 | 30.98 | 34.70↓9.79 | 5.19 | 17.67 | 21.50 | 11.02 | 1.52 | 25.81 | 13.79↓28.72 |
| cl-s-m | 8.55 | 10.98 | 7.70 | 4.07 | 7.83↓1.65 | 45.50 | 38.86 | 42.18↓2.31 | 5.45 | 23.56 | 37.71 | 17.17 | 2.24 | 35.58 | 20.29↓22.22 |
| c-sem | 8.67 | 9.41 | 8.69 | 4.43 | 7.80↓1.68 | 53.12 | 38.51 | 45.82↑1.33 | **9.07** | 40.44 | 41.90 | 18.39 | 3.92 | 44.45 | 26.36↓16.15 |
| 多图像 | 3.36 | 10.48 | 13.43 | 6.24 | 8.38↓1.10 | 28.18 | 10.96 | 19.57↓24.92 | 1.22 | 23.90 | 35.05 | 16.83 | 5.71 | 18.51 | 16.87↓25.64 |
| **ColParse** | **15.81** | 17.65 | **23.75** | **25.48** | **20.67**↑11.19 | **58.55** | **39.99** | **49.27**↑4.78 | 5.91 | **60.76** | **54.91** | **38.77** | **25.27** | **62.34** | **41.33**↓1.18 |
| GME-2B | 54.25 | 50.65 | 59.44 | 49.15 | 53.37 | 81.44 | 79.62 | 80.53 | 81.37 | 81.70 | 91.31 | 85.03 | 63.81 | 93.60 | 82.80 |
| s2m-add | 48.33 | 46.46 | 40.98 | 30.02 | 41.45↓11.92 | 82.20 | 63.50 | 72.85↓7.68 | 70.43 | 77.10 | 85.04 | 80.94 | 20.83 | 89.96 | 70.72↓12.08 |
| s2m-mul | 51.15 | 49.41 | 46.69 | 28.65 | 43.98↓9.39 | 83.64 | 67.99 | 75.82↓4.71 | 74.25 | 77.91 | 88.07 | 84.75 | 22.40 | 92.99 | 73.40↓9.40 |
| cl-t-c | 17.04 | 18.20 | 2.88 | 5.79 | 10.98↓42.39 | 34.72 | 28.03 | 31.38↓49.15 | 14.40 | 18.17 | 30.07 | 14.03 | 4.30 | 26.48 | 17.91↓64.89 |
| cl-t-m | 18.60 | 12.72 | 12.39 | 7.72 | 12.86↓40.51 | 45.17 | 40.84 | 43.01↓37.52 | 25.41 | 31.73 | 44.53 | 16.56 | 2.77 | 40.18 | 26.86↓55.94 |
| cl-s-c | 17.47 | 19.49 | 10.88 | 5.99 | 13.46↓39.91 | 34.42 | 29.29 | 31.86↓48.67 | 15.10 | 14.91 | 29.78 | 15.49 | 3.92 | 28.86 | 18.01↓64.79 |
| cl-s-m | 21.66 | 11.82 | 14.37 | 5.63 | 13.37↓40.00 | 42.98 | 38.99 | 40.99↓39.54 | 24.76 | 30.32 | 45.48 | 17.41 | 2.60 | 39.35 | 26.65↓56.15 |
| c-sem | 24.51 | 22.71 | 22.72 | 18.57 | 22.13↓31.24 | 65.58 | 53.29 | 59.44↓21.09 | 38.80 | 50.88 | 52.09 | 35.27 | 18.64 | 63.25 | 43.16↓39.64 |
| 多图像 | 32.00 | 41.67 | 25.86 | 19.83 | 29.84↓23.53 | 73.00 | 47.58 | 60.29↓20.24 | 67.87 | 67.77 | 81.98 | 58.54 | 22.85 | 77.89 | 62.82↓19.98 |
| **ColParse** | **55.14** | **52.08** | **56.24** | **51.36** | **53.71**↑0.34 | **83.93** | **80.16** | **82.05**↑1.52 | **81.65** | **83.87** | **91.06** | **85.89** | **63.17** | **93.76** | **83.23**↑0.43 |
| GME-7B | 53.66 | 54.34 | 65.38 | 54.32 | 56.93 | 83.21 | 84.18 | 83.70 | 87.20 | 82.32 | 92.92 | 88.89 | 63.36 | 94.81 | 84.92 |
| s2m-add | 50.40 | 47.49 | 47.63 | 29.02 | 43.64↓13.29 | 83.22 | 67.67 | 75.45↓8.25 | 73.94 | 79.30 | 87.84 | 84.63 | 22.35 | 92.79 | 73.48↓11.44 |
| s2m-mul | 48.04 | 47.85 | 39.12 | 29.60 | 41.15↓15.78 | 82.30 | 63.00 | 72.65↓11.05 | 70.46 | 77.68 | 84.96 | 80.38 | 21.12 | 90.08 | 70.78↓14.14 |
| cl-t-c | 10.12 | 12.81 | 8.12 | 4.38 | 8.86↓48.07 | 29.66 | 23.82 | 26.74↓56.96 | 3.79 | 11.21 | 21.97 | 11.20 | 2.44 | 18.30 | 11.49↓73.43 |
| cl-t-m | 9.38 | 10.36 | 6.42 | 4.63 | 7.70↓49.23 | 35.86 | 32.84 | 34.35↓49.35 | 3.56 | 21.05 | 35.14 | 12.97 | 1.45 | 24.43 | 16.43↓68.49 |
| cl-s-c | 8.58 | 16.36 | 5.41 | 5.20 | 8.89↓48.04 | 28.62 | 23.40 | 26.01↓57.69 | 5.46 | 11.14 | 23.14 | 11.35 | 1.54 | 21.57 | 12.37↓72.55 |
| cl-s-m | 10.84 | 9.96 | 6.62 | 2.80 | 7.56↓49.37 | 34.06 | 29.69 | 31.88↓51.82 | 3.93 | 22.24 | 33.74 | 13.14 | 1.42 | 23.76 | 16.37↓68.55 |
| c-sem | 16.30 | 13.79 | 8.34 | 8.42 | 11.71↓45.22 | 49.08 | 35.41 | 42.25↓41.45 | 7.40 | 29.57 | 32.02 | 15.54 | 5.76 | 35.26 | 20.93↓63.99 |
| 多图像 | 29.32 | 39.47 | 30.32 | 16.76 | 28.97↓27.96 | 75.23 | 52.91 | 64.07↓19.63 | 72.68 | 70.18 | 85.20 | 63.27 | 23.75 | 81.21 | 66.05↓18.87 |
| **ColParse** | **62.40** | **59.73** | **68.43** | **57.90** | **62.12**↓5.19 | **84.30** | **83.94** | **84.12**↑0.42 | **87.12** | **83.94** | **92.84** | **89.62** | **62.53** | **94.97** | **85.17**↑0.25 |
| UniME-V2-2B | 9.50 | 15.78 | 15.51 | 19.03 | 14.96 | 54.24 | 77.98 | 66.11 | 61.19 | 65.48 | 76.76 | 59.65 | 45.25 | 77.50 | 64.31 |
| s2m-add | 30.82 | 38.23 | 23.31 | 20.59 | 28.24↑13.28 | 58.58 | 50.31 | 54.45↓11.66 | 45.94 | 55.05 | 67.28 | 55.13 | 12.43 | 72.62 | 51.41↓12.90 |
| s2m-mul | 29.67 | 34.92 | 21.62 | 19.02 | 26.31↑11.35 | 56.92 | 48.58 | 52.75↓13.36 | 45.21 | 55.60 | 66.84 | 53.32 | 12.49 | 71.71 | 50.86↓13.45 |
| cl-t-c | 24.01 | 15.49 | 14.30 | 12.37 | 16.54↑1.58 | 46.27 | 38.93 | 42.60↓23.51 | 18.00 | 24.42 | 29.87 | 16.38 | 4.20 | 33.42 | 21.05↓43.26 |
| cl-t-m | 21.67 | 10.06 | 13.15 | 9.43 | 13.58↓1.38 | 42.79 | 35.18 | 38.99↓27.12 | 17.18 | 27.33 | 42.30 | 14.99 | 5.25 | 44.29 | 25.22↓39.09 |
| cl-s-c | 23.56 | 15.34 | 12.00 | 12.62 | 15.88↑0.92 | 48.96 | 39.41 | 44.19↓21.92 | 19.80 | 24.93 | 32.56 | 21.55 | 4.33 | 38.90 | 23.68↓40.63 |
| cl-s-m | 22.75 | 13.02 | 12.48 | 10.46 | 14.68↓0.28 | 43.51 | 37.60 | 40.56↓25.55 | 17.20 | 29.53 | 42.92 | 16.67 | 5.00 | 47.36 | 26.45↓37.86 |

续见下页

表3 –接上页

| 方法 | ViDoRe-V2 | | | | | ViDoSeek | | | VisRAG | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bio-L | Eco-R | ESG-H | ESG-M | Avg. | Doc | Page | Avg. | arXiv | 图表 | InfoV | MP-Doc | Plot | 幻灯片 | Avg. |
| c-sem | 35.10 | 27.35 | 21.50 | 28.81 | 28.19↑13.23 | 64.50 | 60.58 | 62.54↓3.57 | 32.65 | 51.23 | 57.19 | 40.32 | 14.91 | 64.66 | 43.49↓20.82 |
| 多图像 | 31.08 | 38.51 | 16.09 | 18.02 | 25.93↑10.97 | 56.06 | 43.06 | 49.56↓16.55 | 46.47 | 46.80 | 65.91 | 42.99 | 11.82 | 63.81 | 46.30↓18.01 |
| ColParse | **44.52** | **45.86** | **48.15** | **51.48** | **47.50**↑32.54 | **80.50** | **79.46** | **79.98**↑13.87 | **58.51** | **63.17** | **74.52** | **66.58** | **35.57** | **80.21** | **63.09**↓1.22 |
| UniME-V2-7B | 26.77 | 23.69 | 24.68 | 31.17 | 26.58 | 78.25 | 82.25 | 80.25 | 60.60 | 79.43 | 80.61 | 64.94 | 45.35 | 82.17 | 68.85 |
| s2m-add | 41.25 | **54.38** | 36.91 | 24.88 | 39.36↑12.78 | 71.49 | 64.32 | 67.91↓12.34 | 51.03 | 64.77 | 75.82 | 67.21 | 15.38 | 81.12 | 59.22↓9.63 |
| s2m-mul | 41.49 | 53.84 | 35.49 | 23.31 | 38.53↑11.95 | 71.69 | 64.24 | 67.97↓12.28 | 50.89 | 65.45 | 76.45 | 67.54 | 15.32 | 81.68 | 59.56↓9.29 |
| cl-t-c | 22.76 | 16.17 | 8.68 | 14.14 | 15.44↓11.14 | 49.38 | 45.31 | 47.35↓32.90 | 19.44 | 24.63 | 35.22 | 17.75 | 4.44 | 36.13 | 22.94↓45.91 |
| cl-t-m | 27.93 | 14.34 | 19.03 | 16.99 | 19.57↓7.01 | 55.94 | 52.84 | 54.39↓25.86 | 27.05 | 38.47 | 62.09 | 20.81 | 7.64 | 55.56 | 35.27↓33.58 |
| cl-s-c | 24.75 | 19.62 | 12.41 | 11.69 | 17.12↓9.46 | 52.10 | 48.38 | 50.24↓30.01 | 20.58 | 21.04 | 34.39 | 23.49 | 5.53 | 35.42 | 23.41↓45.44 |
| cl-s-m | 30.78 | 20.40 | 18.14 | 15.62 | 21.24↓5.34 | 56.69 | 53.63 | 55.16↓25.09 | 28.73 | 37.84 | 62.65 | 23.01 | 6.78 | 57.14 | 36.03↓32.82 |
| c-sem | 37.96 | 29.51 | 27.33 | 21.95 | 29.19↑2.61 | 71.01 | 68.82 | 69.92↓10.33 | 51.40 | 65.80 | 72.53 | 54.42 | 21.98 | 76.81 | 57.16↓11.69 |
| 多图像 | 35.99 | 45.82 | 24.74 | 14.71 | 30.32↑3.74 | 65.89 | 53.21 | 59.55↓20.70 | 51.79 | 62.63 | 73.67 | 48.06 | 17.25 | 73.00 | 54.40↓14.45 |
| ColParse | **54.95** | 50.07 | **54.92** | **50.14** | **52.52**↑25.94 | **81.16** | **83.32** | **82.24**↑1.99 | **61.90** | **77.80** | **78.41** | **71.89** | **44.43** | **84.68** | **69.85**↑1.00 |
| B3-2B | 38.41 | 31.80 | 45.23 | 45.10 | 40.14 | 78.56 | 74.87 | 76.72 | 51.75 | 66.86 | 70.43 | 45.73 | 36.69 | 77.81 | 58.21 |
| s2m-add | 32.14 | **40.97** | 22.62 | 19.74 | 28.87↓11.27 | 67.40 | 56.61 | 62.01↓14.71 | 43.01 | 51.27 | 68.01 | 50.92 | 13.40 | 73.78 | 50.07↓8.14 |
| s2m-mul | 30.87 | 39.31 | 21.36 | 19.69 | 27.81↓12.33 | 66.73 | 56.12 | 61.43↓15.29 | 42.74 | 49.75 | **68.21** | 51.27 | 13.31 | 73.79 | 49.85↓8.36 |
| cl-t-c | 11.90 | 13.53 | 9.61 | 5.70 | 10.19↓29.95 | 44.40 | 41.39 | 42.90↓33.82 | 14.13 | 37.31 | 41.30 | 17.72 | 3.32 | 44.50 | 26.38↓31.83 |
| cl-t-m | 15.16 | 14.20 | 13.12 | 6.61 | 12.27↓27.87 | 44.92 | 44.57 | 44.75↓31.97 | 14.12 | 32.08 | 42.61 | 19.91 | 1.06 | 43.03 | 25.47↓32.74 |
| cl-s-c | 16.03 | 20.23 | 10.10 | 8.92 | 13.82↓26.32 | 48.54 | 47.73 | 48.14↓28.58 | 13.09 | 36.05 | 44.30 | 24.60 | 2.76 | 48.77 | 28.26↓29.95 |
| cl-s-m | 16.43 | 16.68 | 14.29 | 9.21 | 14.15↓25.99 | 48.08 | 47.08 | 47.58↓29.14 | 13.51 | 35.91 | 44.17 | 25.47 | 1.20 | 45.45 | 27.62↓30.59 |
| c-sem | 25.26 | 25.23 | 18.89 | 18.29 | 21.92↓18.22 | 60.79 | 59.97 | 60.38↓16.34 | 32.30 | 54.90 | 61.44 | 39.77 | 10.80 | 61.17 | 43.40↓14.81 |
| 多图像 | 21.83 | 23.38 | 7.29 | 11.37 | 15.97↓24.17 | 52.38 | 37.41 | 44.90↓31.82 | 38.52 | 41.26 | 53.67 | 24.33 | 7.97 | 37.56 | 33.89↓24.32 |
| ColParse | **45.03** | 39.23 | **48.20** | **49.30** | **45.44**↑5.30 | **79.98** | **80.79** | **80.39**↑3.67 | **51.00** | **62.94** | 67.06 | **53.86** | **32.02** | **80.01** | **57.82**↓0.39 |
| B3-7B | 47.29 | 44.81 | 50.84 | 48.05 | 47.75 | 82.07 | 82.26 | 82.17 | 65.83 | 76.77 | 84.54 | 68.55 | 52.86 | 85.75 | 72.38 |
| s2m-add | 44.47 | 48.61 | 35.86 | 28.82 | 39.44↓8.31 | 75.93 | 65.24 | 70.59↓11.58 | 56.95 | 69.80 | 78.61 | 69.19 | 19.74 | 85.71 | 63.33↓9.05 |
| s2m-mul | 44.79 | **49.78** | 35.45 | 28.68 | 39.68↓8.07 | 76.17 | 64.83 | 70.50↓11.67 | 57.17 | 70.52 | 79.24 | 70.67 | 20.17 | 85.50 | 63.88↓8.50 |
| cl-t-c | 18.50 | 22.50 | 20.46 | 15.85 | 19.33↓28.42 | 66.30 | 63.26 | 64.78↓17.39 | 18.07 | 43.88 | 52.94 | 23.57 | 5.85 | 58.35 | 33.78↓38.60 |
| cl-t-m | 22.98 | 22.71 | 22.21 | 14.16 | 20.52↓27.23 | 67.41 | 67.43 | 67.42↓14.75 | 23.92 | 43.10 | 62.54 | 28.35 | 7.43 | 61.50 | 37.81↓34.57 |
| cl-s-c | 20.76 | 26.27 | 28.71 | 22.95 | 24.67↓23.08 | 68.09 | 65.59 | 66.84↓15.33 | 16.71 | 47.07 | 53.44 | 31.71 | 7.56 | 60.01 | 36.08↓36.30 |
| cl-s-m | 22.17 | 26.43 | 26.15 | 19.85 | 23.65↓24.10 | 67.85 | 64.05 | 65.95↓16.22 | 19.23 | 43.79 | 53.45 | 35.24 | 8.50 | 61.19 | 36.90↓35.48 |
| c-sem | 32.31 | 30.23 | 33.27 | 18.99 | 28.70↓19.05 | 75.30 | 71.83 | 73.57↓8.60 | 46.32 | 67.80 | 75.28 | 54.93 | 24.73 | 80.26 | 58.22↓14.16 |
| 多图像 | 26.48 | 33.14 | 11.25 | 11.91 | 20.70↓27.05 | 64.84 | 47.83 | 56.34↓25.83 | 51.50 | 46.37 | 63.64 | 34.80 | 11.94 | 44.09 | 42.06↓30.32 |
| ColParse | **53.72** | 49.50 | **52.40** | **50.15** | **51.44**↑3.69 | **83.15** | **83.60** | **83.38**↑1.21 | **65.97** | **75.92** | **80.19** | **73.87** | **48.05** | **86.13** | **71.69**↓0.69 |

## C.5. 变体研究

### C.5.1. 算法工作流程

**算法4** 提出了一个统一的离线索引框架，用于处理这三种变体。在共享的布局解析和双流编码（阶段 1–2）之后，阶段 3 会根据指定的变体类型有所不同： 单转多 保留原始子图像向量；类型聚类通过求平均值，按语义内容类型聚合向量；全局向量纳入将完整文档的全局向量添加到局部集合中。

**算法4** 面向 `ColParse` 变体的集成离线索引 输入：文档图像 $d \in \mathbb{R}^{H\times W\times 3}$；文档解析器模型 $\Psi_{\text{parse}}$；单向量编码器 $\Phi_{\text{enc}}$ ： $\mathbb{R}^{H'\times W'\times 3} \to \mathbb{R}^{D}$；模式 $M \in$ {s2m、s2m-t-c、s2m-g-i}输出：多向量表示$\mathrm{D}_{\text{variant}}$/* 阶段1：布局感知文档解析 */$[\{b_j, c_j\}]_{j=1}^{k} \leftarrow \Psi_{\text{parse}}(d)$； // 获取 $k$ 边界框和内容类型$\mathcal{S}_d \leftarrow \emptyset$ 对于 $j \leftarrow 1$ 到 $k$ 执行$s_j \leftarrow$ 裁剪$(d, b_j)$； // 为每个布局组件提取子图像 $\mathcal{S}_d \leftarrow \mathcal{S}_d \cup \{s_j\}$结束 /* 阶段2：区域编码 */ $\mathrm{D}_{\text{local}} \leftarrow \emptyset$ 对于每个子图像 $s_j \in \mathcal{S}_d$ 执行 $\mathrm{v}^{(j)}$ 局部$\leftarrow \Phi_{\text{enc}}(s_j)$； // 独立区域编码 $\mathrm{D}_{\text{local}} \leftarrow \mathrm{D}_{\text{local}} \cup \{\mathrm{v}^{(j)}$ 局部$\}$结束 /* 阶段3：特定变体的表示构建 */ 如果 $M =$ s2m则$\mathrm{D}_{\text{variant}} \leftarrow \mathrm{D}_{\text{local}}$； // 标准布局分解集合 否则如果 $M =$ s2m-t-c则$\mathrm{D}_{\text{variant}} \leftarrow \emptyset$ $\mathcal{T} \leftarrow$ 唯一$(\{c_1, \ldots, c_k\})$； // 识别唯一内容类型 对于每种类型 $t \in \mathcal{T}$执行 $\mathrm{v}_{\text{avg}}^{(t)} \leftarrow$ 均值$(\{\mathrm{v}^{(j)}$ 局部 $\mid c_j = t\})$； // 按类型聚类并取平均 $\mathrm{D}_{\text{variant}} \leftarrow \mathrm{D}_{\text{variant}} \cup \{\mathrm{v}_{\text{avg}}^{(t)}\}$结束 结束 否则如果 $M =$ s2m-g-i则$\mathrm{v}_{\text{global}} \leftarrow \Phi_{\text{enc}}(d)$； // 编码原始页面以提供上下文 $\mathrm{D}_{\text{variant}} \leftarrow \mathrm{D}_{\text{local}} \cup \{\mathrm{v}_{\text{global}}\}$； // 将全局向量追加到集合 结束 返回$\mathrm{D}_{\text{variant}}$

### C.5.2. 更多分析

由于正文篇幅有限，我们将 `ColParse` 与其变体之间的性能比较雷达图留在图9中。

**引入全局页面级上下文对于消除孤立布局组件中的语义歧义不可或缺**。表4中的定量结果表4表明，即使通过简单纳入全局上下文（`s2m-g-i`），性能也会大幅超越仅使用局部信息的单转多基线，使VLM2Vec-V1-2B在ViDoRe-V1上的得分从34.39提升至49.93。图5突出显示，在需要整体理解的基准测试中，这一差距最为显著；表格或图表等局部子图像往往缺少页面其他位置提供的必要上下文标题。我们推测，全局向量充当“语义锚点”，提供文档的整体主题，这对于后期交互机制准确地将特定查询词元与相关子区域对齐至关重要。

**保持布局组件各自的空间与语义完整性，优于启发式类型级聚类**。表4中的性能趋势表4以及图5中的对比柱状图`s2m-t-c`变体通常会导致性能下降，相较于标准的单转多方法尤为如此，例如VLM2Vec-V1-2B在ViDoRe-V1基准测试上的得分下降了1.56分。图9图9中的雷达图也呈现出相同趋势：类型聚类变体的性能覆盖范围始终最窄。这表明，空间局部性是视觉文档中至关重要的语义载体；将多个不同组件合并为同一类型的平均表示，会使模型失去足够的细粒度分辨能力，无法让`MaxSim`算子区分同一类型中具体的相关区域与无关区域。




图*9.* 在十种主流单向量多模态检索模型上， `ColParse` 及其变体在五个 VDR 基准测试上的性能对比（以 nDCG@5 评估）。受篇幅限制，详细结果记录请参见表 4和表 5。

表4. 在 MMLongBench 和 ViDoRe-V1基准上进行的消融研究。对于每个模型区块，我们将每列中表现最佳的方法加粗（基础结果除外）。平均分及其相对于基准模型的相对增益（↑/↓）如表所示。

| 方法 | MMLongBench | | ViDoRe-V1 | |
|---|---|---|---|---|
| | 文档 | Avg. | arXiv | Avg. |
| VLM2Vec-V1-2B | 25.62 26.23 | 25.93 | 17.80 13.98 39.41 9.18 36.32 10.56 16.39 15.96 23.56 24.11 | 20.73 |
| 单转多 | 25.19 | | 38.78 9.98 | |
| s2m-t-c | 25.37 18.43 | 21.90↓4.03 | 34.95 15.04 55.09 15.31 40.07 9.78 32.83 43.07 44.84 37.31 | 32.83↑12.10 |
| s2m-g-i | 31.57 | 29.06↑3.13 | **48.84** 60.83 | |
| ColParse | **34.31** | | 47.66 | |
| VLM2Vec-V1-7B | 23.85 37.63 | 30.74 | 28.07 17.93 44.47 2.06 16.78 5.86 17.93 25.04 28.90 14.59 | 20.16 |
| 单转多 | 35.75 | | 53.54 | |
| s2m-t-c | 35.62 28.64 | 32.13↑1.39 | 50.56 24.08 59.42 35.90 64.84 18.29 65.39 58.49 72.13 63.75 | 51.29↑31.13 |
| s2m-g-i | 37.95 | | **62.29** 64.14 | |
| ColParse | **43.34** | | 60.47 | |
| VLM2Vec-V2-2B | 48.55 50.34 | 49.45 | 78.98 38.51 82.21 64.57 87.64 44.68 85.06 82.99 89.89 87.08 | 74.16 |
| 单转多 | 45.89 | | 69.32 | 70.99↓3.17 |
| s2m-t-c | 44.99 40.97 | 42.98↓6.47 | 67.44 34.25 74.79 48.36 71.32 43.73 82.22 81.18 84.78 76.42 | 66.45↓7.71 |
| s2m-g-i | 48.03 | | 79.83 87.46 | 77.44↑3.28 |
| ColParse | **49.49** | | **80.17** | **78.41**↑4.25 |
| LamRA-Ret | 19.78 13.24 | 16.51 | 29.31 19.56 63.00 15.83 51.44 7.70 21.10 29.81 37.18 31.95 | 30.69 |

下页继续

表4 –接上一页

| 方法 | MMLongBench | | ViDoRe-V1 | |
|---|---|---|---|---|
| | 文档 | Avg. | arXiv | Avg. |
| 单转多 | 33.04 | | 6.06 13.60 | 23.45↓7.24 |
| s2m-t-c | 31.74 17.49 | 24.62↑8.11 | 6.13 10.90 35.44 16.81 28.56 15.77 15.45 27.59 24.45 19.65 | 20.08↓10.61 |
| s2m-g-i | **33.16** | **25.82**↑9.31 | 6.06 15.79 29.03 | 25.15↓5.54 |
| ColParse | 30.74 | | **17.27** 26.29 | **30.10**↓0.59 |
| GME-2B | 52.07 53.14 | 52.61 | 82.59 56.46 88.97 89.72 93.20 70.33 98.49 92.15 98.15 95.65 | 86.57 |
| 单转多 | 50.56 | | 72.46 89.86 | 77.05↓9.52 |
| s2m-t-c | 49.51 44.73 | 47.12↓5.49 | 70.90 37.34 77.92 67.59 79.85 59.12 86.13 86.87 87.23 83.35 | 73.63↓12.94 |
| s2m-g-i | 51.81 | | 80.89 90.49 | 81.76↓4.81 |
| ColParse | **53.06** | | **82.39** | **86.10**↓0.47 |
| GME-7B | 54.01 55.80 | 54.91 | 87.59 56.05 91.96 94.25 93.72 76.26 99.63 95.45 99.63 99.06 | 89.36 |
| 单转多 | 53.55 | | 75.57 | 81.71↓7.65 |
| s2m-t-c | 52.97 48.05 | 50.51↓4.40 | 75.21 44.11 82.69 78.44 86.17 66.14 91.53 89.86 91.38 88.60 | 79.41↓9.95 |
| s2m-g-i | 54.32 | | 84.29 93.70 | 85.44↓3.92 |
| ColParse | **54.96** | | **87.35** | **89.56**↑0.20 |
| UniME-V2-2B | 18.52 40.10 | 29.31 | 36.52 12.43 42.41 14.09 51.11 7.39 20.23 32.96 24.21 19.25 | 26.06 |
| 单转多 | 38.47 | | 52.64 | |
| s2m-t-c | 35.13 30.33 | 32.73↑3.42 | 48.29 21.24 62.26 43.83 58.58 27.37 70.42 66.86 74.85 67.57 | 54.13↑28.07 |
| s2m-g-i | 41.90 | | **64.23** 83.62 | |
| ColParse | **44.22** | | 62.39 **85.84** | |
| UniME-V2-7B | 33.19 45.72 | 39.46 | 63.23 24.91 65.25 11.16 41.54 14.18 41.89 40.56 57.44 42.78 | 40.29 |
| 单转多 | 39.83 | | 57.08 87.49 | |
| s2m-t-c | 38.66 36.94 | 37.80↓1.66 | 54.98 25.88 65.87 56.63 70.76 41.65 80.00 77.76 86.67 77.88 | 63.81↑23.52 |
| s2m-g-i | 41.35 | | 64.50 87.90 | |
| ColParse | **45.90** | | **64.78** | |
| B3-2B | 37.10 32.07 | 34.59 | 57.00 29.38 68.09 48.31 71.55 18.09 74.13 64.64 75.44 63.13 | 56.98 |
| 单转多 | 36.47 | | 46.63 66.85 | 55.25↓1.73 |
| s2m-t-c | 35.15 28.33 | 31.74↓2.85 | 43.88 18.92 54.65 41.90 54.30 26.86 67.36 61.92 72.05 60.66 | 50.25↓6.73 |
| s2m-g-i | 38.87 | | 56.06 67.17 | 60.82↑3.84 |
| ColParse | **42.06** | | **56.47** | **62.74**↑5.76 |
| B3-7B | 46.09 45.10 | 45.60 | 68.95 43.38 79.86 66.56 84.12 37.06 81.01 81.25 88.57 81.30 | 71.21 |
| 单转多 | 44.43 | | 58.94 | 70.04↓1.17 |
| s2m-t-c | 43.32 39.84 | 41.58↓4.02 | 56.41 29.10 68.99 53.11 70.75 44.35 83.69 77.98 84.69 78.56 | 64.76↓6.45 |
| s2m-g-i | 45.62 | | 67.23 | 73.47↑2.26 |
| ColParse | **49.11** | | **67.68** | **74.41**↑3.20 |

表 5. 在 ViDoRe-V2、ViDoSeek 和 VisRAG 基准测试上的消融研究。对于每个模型区块，我们将表现最佳的方法加粗（每列中，基础结果除外）。平均分以相对增益（↑/↓）的形式展示，并与基准模型进行比较。

| 方法 | ViDoRe-V2 | | | | | ViDoSeek | | | VisRAG | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bio-L | Eco-R | ESG-H | ESG-M | Avg. | Doc | Page | Avg. | arXiv | 图表 | InfoV | MP-Doc | Plot | 幻灯片 | Avg. |
| VLM2Vec-V1-2B | 6.88 | 14.15 | 12.25 | 20.54 | 13.46 | 56.40 | 67.73 | 62.07 | 41.68 | 58.21 | 70.79 | 42.74 | 23.83 | 74.07 | 51.89 |
| 单转多 | 17.15 | 18.19 | 10.24 | 7.44 | 13.26↓0.20 | 50.51 | 41.52 | 46.02↓16.05 | 30.17 | 42.91 | 61.28 | 37.60 | 6.80 | 65.76 | 40.75↓11.14 |
| s2m-t-c | 13.46 | 18.31 | 11.58 | 8.85 | 13.05↓0.41 | 50.22 | 40.75 | 45.49↓16.58 | 27.87 | 41.64 | 59.37 | 34.82 | 6.94 | 62.76 | 38.90↓12.99 |
| s2m-g-i | 28.49 | 26.29 | 29.18 | **39.15** | 30.78↑17.32 | 73.38 | 68.14 | 70.76↑8.69 | **40.39** | 52.54 | **70.43** | 47.82 | 8.96 | 76.40 | 49.42↓2.47 |
| ColParse | **30.33** | **29.55** | **33.21** | 38.33 | **32.86**↑19.40 | **75.23** | **70.19** | **72.71**↑10.64 | 38.18 | **60.09** | 69.44 | **48.29** | **18.83** | **76.95** | **51.96**↑0.07 |
| VLM2Vec-V1-7B | 4.93 | 13.74 | 6.82 | 11.27 | 9.19 | 54.26 | 77.39 | 65.83 | 52.58 | 69.83 | 71.43 | 52.86 | 34.24 | 73.22 | 59.03 |
| 单转多 | 34.67 | 37.55 | 33.77 | 26.91 | 33.23↑24.04 | 66.88 | 58.17 | 62.53↓3.30 | 45.06 | 50.30 | 63.65 | 50.23 | 10.83 | 74.00 | 49.01↓10.02 |
| s2m-t-c | 31.17 | 41.18 | 26.15 | 21.07 | 29.89↑20.70 | 64.53 | 54.75 | 59.64↓6.19 | 43.98 | 52.75 | 61.47 | 45.75 | 10.49 | 71.33 | 47.63↓11.40 |
| s2m-g-i | 41.08 | 37.99 | 40.11 | 36.54 | 38.93↑29.74 | 75.46 | 75.28 | 75.37↑9.54 | 53.97 | 64.14 | 65.37 | 55.41 | 27.98 | 77.07 | 57.32↓1.71 |
| ColParse | **42.63** | **42.89** | **50.55** | **42.86** | **44.73**↑35.54 | **78.34** | **78.61** | **78.48**↑12.65 | **54.43** | **70.30** | **69.27** | **58.49** | **33.46** | **77.98** | **60.66**↑1.63 |
| VLM2Vec-V2-2B | 44.45 | 45.77 | 48.77 | 46.98 | 46.49 | 80.88 | 83.68 | 82.28 | 77.38 | 82.30 | 86.27 | 71.60 | 66.96 | 92.04 | 79.43 |
| 单转多 | 42.12 | 51.08 | 41.33 | 24.84 | 39.84↓6.65 | 73.81 | 67.71 | 70.76↓11.52 | 65.36 | 69.18 | 79.34 | 73.19 | 19.59 | 86.77 | 65.57↓13.86 |
| s2m-t-c | 41.81 | 51.83 | 37.07 | 25.78 | 39.12↓7.37 | 73.69 | 66.16 | 69.93↓12.35 | 64.18 | 66.49 | 77.61 | 64.32 | 19.58 | 85.70 | 62.98↓16.45 |
| s2m-g-i | 44.34 | 51.99 | 40.57 | 34.53 | 42.86↓3.63 | 79.23 | 81.01 | 80.12↓2.16 | 77.03 | 74.83 | 82.98 | **78.17** | 54.13 | 90.93 | 76.35↓3.08 |
| ColParse | **50.06** | **53.76** | **57.41** | **46.40** | **51.91**↑5.42 | **80.94** | **83.87** | **82.41**↑0.13 | **77.18** | **78.05** | **84.37** | 78.07 | **58.74** | **91.95** | **78.06**↓1.37 |
| LamRA-Ret | 10.75 | 9.65 | 6.32 | 11.18 | 9.48 | 60.17 | 28.81 | 44.49 | 11.17 | 63.50 | 59.78 | 33.57 | 29.42 | 57.59 | 42.51 |
| 单转多 | 11.11 | 26.32 | 20.54 | 23.76 | 20.43↑10.95 | 53.77 | 30.82 | 42.30↓2.19 | 1.94 | 25.59 | 30.56 | 27.49 | 3.95 | 44.28 | 22.30↓20.21 |
| s2m-t-c | 9.36 | 15.57 | 14.49 | 17.11 | 14.13↑4.65 | 49.02 | 29.38 | 39.20↓5.29 | 2.03 | 22.46 | 31.39 | 21.10 | 3.96 | 42.90 | 20.64↓21.87 |
| s2m-g-i | 11.11 | **26.56** | 18.78 | 21.76 | 19.55↑10.07 | 53.74 | 31.77 | 42.76↓1.73 | 1.91 | 28.55 | 31.15 | 32.12 | 25.09 | 44.22 | 27.17↓15.34 |
| ColParse | **15.81** | 17.65 | **23.75** | **25.48** | **20.67**↑11.19 | **58.55** | **39.99** | **49.27**↑4.78 | **5.91** | **60.76** | **54.91** | **38.77** | **25.27** | **62.34** | **41.33**↓1.18 |
| GME-2B | 54.25 | 50.65 | 59.44 | 49.15 | 53.37 | 81.44 | 79.62 | 80.53 | 81.37 | 81.70 | 91.31 | 85.03 | 63.81 | 93.60 | 82.80 |
| 单转多 | 47.25 | 43.18 | 42.80 | 35.67 | 42.23↓11.14 | 82.42 | 64.31 | 73.37↓7.16 | 69.46 | 75.14 | 82.71 | 77.65 | 19.43 | 87.64 | 68.67↓14.13 |
| s2m-t-c | 47.87 | 50.70 | 39.72 | 31.03 | 42.33↓11.04 | 82.06 | 63.82 | 72.94↓7.59 | 68.24 | 74.56 | 81.25 | 71.08 | 20.06 | 87.98 | 67.20↓15.60 |
| s2m-g-i | 49.19 | 43.73 | 47.90 | 40.90 | 45.43↓7.94 | 83.46 | 72.22 | 77.84↓2.69 | 78.82 | 77.90 | 85.79 | 81.37 | 51.04 | 90.55 | 77.58↓5.22 |
| ColParse | **55.14** | **52.08** | **56.24** | **51.36** | **53.71**↑0.34 | **83.93** | **80.16** | **82.05**↑1.52 | **81.65** | **83.87** | **91.06** | **85.89** | **63.17** | **93.76** | **83.23**↑0.43 |
| GME-7B | 53.66 | 54.34 | 65.38 | 54.32 | 56.93 | 83.21 | 84.18 | 83.70 | 87.20 | 82.32 | 92.92 | 88.89 | 63.36 | 94.81 | 84.92 |
| 单转多 | 44.39 | 42.94 | 50.47 | 35.10 | 43.23↓13.70 | 83.51 | 66.19 | 74.85↓8.85 | 75.24 | 77.08 | 84.57 | 82.41 | 20.95 | 90.67 | 71.82↓13.10 |
| s2m-t-c | 47.29 | 51.24 | 46.31 | 30.35 | 43.80↓13.13 | 83.03 | 65.61 | 74.32↓9.38 | 74.40 | 76.84 | 83.60 | 77.58 | 21.35 | 90.80 | 70.76↓14.16 |
| s2m-g-i | 45.99 | 42.98 | 56.63 | 39.79 | 46.35↓10.58 | 84.04 | 73.61 | 78.83↓4.87 | 84.45 | 80.35 | 88.21 | 87.03 | 53.24 | 92.27 | 80.93↓3.99 |
| ColParse | **62.40** | **59.73** | **68.43** | **57.90** | **62.12**↑5.19 | **84.30** | **83.94** | **84.12**↑0.42 | **87.12** | **83.94** | **92.84** | **89.62** | **62.53** | **94.97** | **85.17**↑0.25 |
| UniME-V2-2B | 9.50 | 15.78 | 15.51 | 19.03 | 14.96 | 54.24 | 77.98 | 66.11 | 61.19 | 65.48 | 76.76 | 59.65 | 45.25 | 77.50 | 64.31 |
| 单转多 | 37.44 | 47.15 | 27.88 | 26.68 | 34.79↑19.83 | 67.36 | 59.21 | 63.29↓2.82 | 47.81 | 54.17 | 71.61 | 61.52 | 9.44 | 76.72 | 53.55↓10.76 |
| s2m-t-c | 35.32 | 42.69 | 28.26 | 25.47 | 32.94↑17.98 | 63.55 | 55.59 | 59.57↓6.54 | 44.16 | 48.21 | 67.30 | 52.26 | 9.72 | 72.55 | 49.03↓15.28 |
| s2m-g-i | **44.64** | **48.14** | 46.83 | 44.53 | 46.04↑31.08 | 78.62 | 77.90 | 78.26↑12.15 | 60.30 | 60.76 | **77.30** | **68.15** | 30.63 | **82.23** | **63.23**↓1.08 |
| ColParse | 44.52 | 45.86 | **48.15** | **51.48** | **47.50**↑32.54 | **80.50** | **79.46** | **79.98**↑13.87 | **58.51** | **63.17** | 74.52 | 66.58 | **35.57** | 80.21 | 63.09↓1.22 |
| UniME-V2-7B | 26.77 | 23.69 | 24.68 | 31.17 | 26.58 | 78.25 | 82.25 | 80.25 | 60.60 | 79.43 | 80.61 | 64.94 | 45.35 | 82.17 | 68.85 |
| 单转多 | 45.06 | 53.38 | 44.29 | 27.79 | 42.63↑16.05 | 72.30 | 67.77 | 70.04↓10.21 | 52.99 | 64.35 | 74.06 | 69.25 | 16.48 | 82.14 | 59.88↓8.97 |
| s2m-t-c | 43.60 | **58.23** | 42.39 | 28.00 | 43.06↑16.48 | 70.92 | 65.12 | 68.02↓12.23 | 52.28 | 59.47 | 69.08 | 62.67 | 16.85 | 80.61 | 56.83↓12.02 |
| s2m-g-i | 50.51 | 53.70 | 46.85 | 37.13 | 47.05↑20.47 | 77.78 | 80.58 | 79.18↓1.07 | 60.99 | 68.25 | 76.65 | **72.59** | 34.78 | **85.05** | 66.39↓2.46 |
| ColParse | **54.95** | 50.07 | **54.92** | **50.14** | **52.52**↑25.94 | **81.16** | **83.32** | **82.24**↑1.99 | **61.90** | **77.80** | **78.41** | 71.89 | **44.43** | 84.68 | **69.85**↑1.00 |
| B3-2B | 38.41 | 31.80 | 45.23 | 45.10 | 40.14 | 78.56 | 74.87 | 76.72 | 51.75 | 66.86 | 70.43 | 45.73 | 36.69 | 77.81 | 58.21 |
| 单转多 | 36.98 | **45.67** | 23.09 | 18.57 | 31.08↓9.06 | 67.88 | 60.70 | 64.29↓12.43 | 42.15 | 56.31 | 60.69 | 51.81 | 11.43 | 74.38 | 49.46↓8.75 |
| s2m-t-c | 34.49 | 41.79 | 18.12 | 20.80 | 28.80↓11.34 | 65.79 | 58.75 | 62.27↓14.45 | 39.49 | 53.55 | 56.26 | 43.83 | 11.33 | 71.09 | 45.93↓12.28 |
| s2m-g-i | 40.64 | 44.71 | 33.32 | 38.76 | 39.36↓0.78 | 75.41 | 77.59 | 76.50↓0.22 | **51.45** | **64.51** | 62.71 | **55.60** | 30.12 | 78.04 | 57.07↓1.14 |
| ColParse | **45.03** | 39.23 | **48.20** | **49.30** | **45.44**↑5.30 | **79.98** | **80.79** | **80.39**↑3.67 | 51.00 | 62.94 | **67.06** | 53.86 | **32.02** | **80.01** | **57.82**↓0.39 |
| B3-7B | 47.29 | 44.81 | 50.84 | 48.05 | 47.75 | 82.07 | 82.26 | 82.17 | 65.83 | 76.77 | 84.54 | 68.55 | 52.86 | 85.75 | 72.38 |
| 单转多 | 45.33 | 52.80 | 39.17 | 30.33 | 41.91↓5.84 | 77.57 | 66.96 | 72.27↓9.90 | 54.91 | 67.44 | 72.97 | 68.97 | 17.68 | 82.51 | 60.75↓11.63 |
| s2m-t-c | 42.29 | 50.60 | 34.12 | 25.95 | 38.24↓9.51 | 76.48 | 66.12 | 71.30↓10.87 | 52.95 | 61.93 | 71.13 | 62.39 | 18.19 | 81.25 | 57.97↓14.41 |
| s2m-g-i | 49.68 | **53.53** | 48.27 | 38.41 | 47.47↓0.28 | 81.14 | 79.17 | 80.16↓2.01 | 64.55 | 71.83 | 77.07 | 72.23 | 39.18 | 84.84 | 68.28↓4.10 |
| ColParse | **53.72** | 49.50 | **52.40** | **50.15** | **51.44**↑3.69 | **83.15** | **83.60** | **83.38**↑1.21 | **65.97** | **75.92** | **80.19** | **73.87** | **48.05** | **86.13** | **71.69**↓0.69 |

## C.6. 超参数研究

### C.6.1. 平衡因子的影响

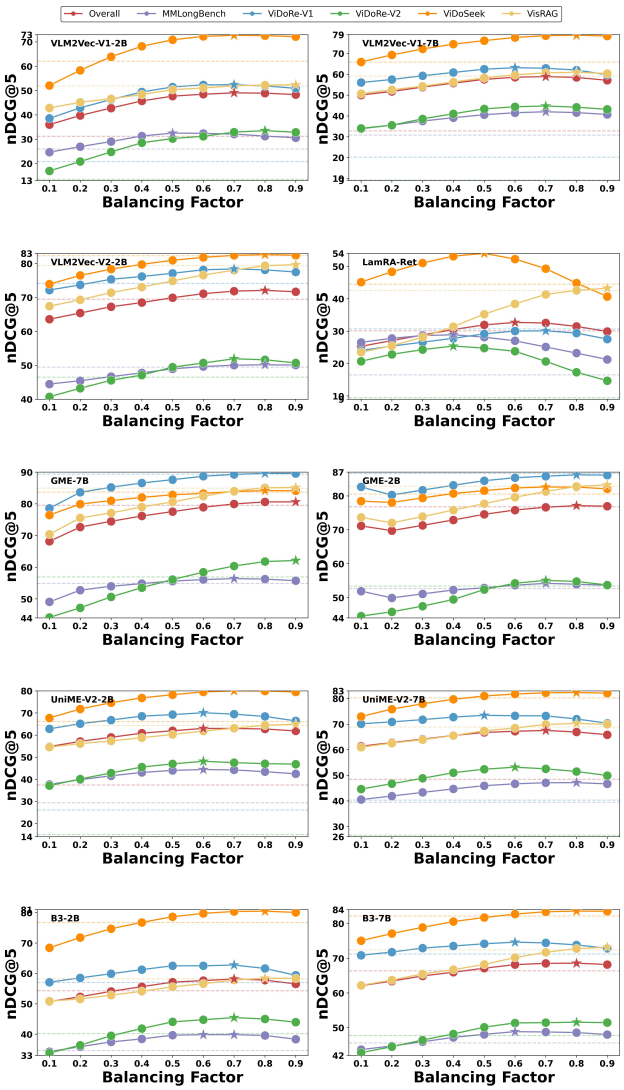



图*10.* 使用 `ColParse` 在不同平衡因子下的模型级性能对比。虚线表示基准结果；星形点表示性能最佳的平衡因子。

### C.6.2. 文档解析模型的影响

评估结果表明，与现有的专用视觉语言模型相比，MinerU2.5 在效率与准确率之间实现了更出色的平衡。表 6 量化了在 A100 (80G) 硬件上的推理效率，使用 Tokens/sec 衡量生成速度，并以页/秒评估端到端吞吐量。结果显示，尽管 0.9B 的 MinerU2-VLM 保持着最高处理速度，MinerU2.5 以 2422 tokens/s 和 2.25 页/秒位居第二，两项指标均显著优于诸如 3B 参数基线方法

表 *6.* MinerU2.5 的推理效率比较。MinerU2.5 与基线方法的结果均在 A100（80G）机器上测试。每列中的最佳结果和次佳结果分别以**粗体** 和下划线标示。

| 模型 | 参数 | 词元/秒 | 页/秒 |
|---|---|---|---|
| MinerU2-VLM | 0.9B | **3089** | **2.84** |
| dots.ocr | 3.0B | 311 | 0.28 |
| MonkeyOCR-pro-3B | 3.7B | 520 | 0.47 |
| MonkeyOCR-pro-1.2B | 1.9B | 589 | 0.53 |
| Nanonets-OCR-s | 3.7B | 605 | 0.55 |
| MinerU2.5 | 1.2B | <u>2422</u> | <u>2.25</u> |

dots.ocr 和 MonkeyOCR-pro。与此同时，表表 7在 OmniDocBench 上对多个类别的解析准确率进行了基准测试，采用综合总体得分、文本和阅读顺序的编辑距离，以及针对公式和表格的结构相似性指标（CDM 和 TEDS）。MinerU2.5 在全部六项指标上均达到最先进水平，总体得分高达 90.67，且在文本和布局识别方面的错误率最低。MonkeyOCR-pro-3B 和 dots.ocr 在结构与文本任务中交替位居第二，但 MinerU2.5 是唯一在文档解析质量的所有评估维度中始终领先的方法。

### C.6.3. 效率分析

表格8总结了五个VDR 基准测试中24个数据集上每篇文档的平均解析向量数量。

表7. OmniDocBench 上的文档解析性能（Ouyang 等人，2025年），涵盖多项任务。最佳和次佳结果分别以**粗体**和下划线标示。

| 方法 | 段落。 | 总体↑ | 文本$^{Edit}$↓ | 公式$^{CDM}$↑ | 表格 TEDS↑ | 表格 TEDS-S↑ | 阅读顺序$^{Edit}$↓ |
|---|---|---|---|---|---|---|---|
| MinerU2-VLM | 0.9B | 85.56 | 0.078 | 80.95 | 83.54 | 87.66 | 0.086 |
| dots.ocr | 3B | 88.41 | 0.048 | 83.22 | 86.78 | 90.62 | 0.053 |
| MonkeyOCR-pro-3B | 3.7B | 88.85 | 0.075 | 87.25 | 86.78 | 90.63 | 0.128 |
| MonkeyOCR-pro-1.2B | 1.9B | 86.96 | 0.084 | 85.02 | 84.24 | 89.02 | 0.130 |
| Nanonets-OCR-s | 3.7B | 85.59 | 0.093 | 85.90 | 80.14 | 85.57 | 0.108 |
| **MinerU2.5** | 1.2B | **90.67** | **0.047** | **88.46** | **88.22** | **92.38** | **0.044** |

表 8. 五个 VDR 基准测试中 24个数据集上每篇文档平均解析向量数量的汇总。这些数值表示由文档解析器（MinerU2.5）生成的布局感知子图像嵌入（$k$）的数量。

| 基准测试 | 数据集 | 平均值 #向量 ($k$) |
|---|---|---|
| MMLongBench | MMLongBench-doc | 6.04 |
| | MMLongBench-page | 6.04 |
| ViDoSeek | ViDoSeek-doc | 5.60 |
| | ViDoSeek-page | 5.77 |
| VisRAG | VisRAG ArxivQA | 1.97 |
| | VisRAG ChartQA | 2.98 |
| | VisRAG InfoVQA | 4.20 |
| | VisRAG MP-DocVQA | 5.82 |
| | VisRAG PlotQA | 2.06 |
| | VisRAG SlideVQA | 4.66 |
| ViDoRe-v2 | ViDoRe 生物医学讲座 v2 | 3.88 |
| | ViDoRe 经济学报告 v2 | 5.89 |
| | ViDoRe 人工标注的 ESG 报告 v2 | 6.92 |
| | ViDoRe 多语言 ESG 报告 v2 | 6.91 |
| ViDoRe-V1 | ViDoRe arxiviva | 1.99 |
| | ViDoRe docvqa | 5.64 |
| | ViDoRe infovqa | 4.52 |
| | ViDoRe shiftproject | 8.97 |
| | ViDoRe syntheticDocQA 人工智能 | 5.71 |
| | ViDoRe syntheticDocQA 能源 | 5.07 |
| | ViDoRe syntheticDocQA 政府报告 | 5.98 |
| | ViDoRe syntheticDocQA 医疗保健行业 | 6.11 |
| | ViDoRe tabfquad | 2.10 |
| | ViDoRe tatdqa | 8.58 |