arXiv:2611.32439v1 [cs.LG] 24 Nov 2026

看见答案，学会嵌入：从后见之明中学习通用多模态嵌入

Yoonki Cho YoungjuNa Taegeun Yang Jinhwan Seo Sung-Eui Yoon 韩国科学技术院

https://yoonkicho.github.io/HEME

摘要

通用多模态嵌入（*UME*）模型将异构多模态输入映射到共享嵌入空间中，在该空间中，成对相似度支持多种下游任务。近期的推理驱动 *UME* 方法利用中间推理生成额外的嵌入上下文，但通常会在推理时带来额外的计算开销。本文探索将匹配候选项（即配对目标）作为替代来源的上下文，并探讨模型能否在训练期间“看见答案”，从而在推理时无需答案也能“学会嵌入”。然而，直接以这一特权上下文为条件，可能引入仅在候选项中存在、而查询本身无法提供的信息，或导致候选项条件化的查询表示向候选项嵌入坍缩。为此，我们提出后见增强多模态嵌入（***HEME***），从这类候选项上下文中学习以查询为基础的事后表示。*HEME* 首先利用事后注意力掩码、不匹配候选条件化和查询词元对齐训练事后教师，以限制对候选项的过度依赖。随后冻结教师模型，并将其事后表示和相似度分布蒸馏到独立学生模型中。在推理时，学生模型无需候选项条件化或事后教师，即可独立编码查询和候选项。在 *MMEB-V2* 以及 *MMEB-V3* 的智能体和文本任务上的实验表明，*HEME* 有效，并且相比显式推理驱动方法能够实现更高效的推理。

1. 引言

通用多模态嵌入（UME）模型旨在将文本、图像、视频和视觉文档等多种模态及其组合表示在一个共享的嵌入空间中，以支持各种下游任务 [11, 36, 70]。除了对齐这些异质输入之外，一个核心挑战是确定针对特定任务，输入的哪些方面应在其嵌入表示中得到强调。

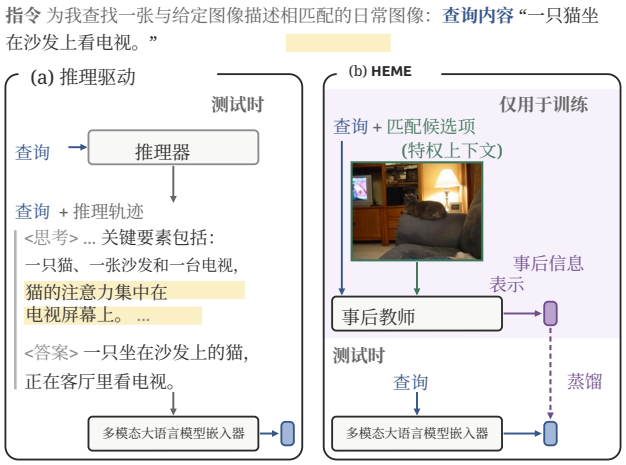

图 1. **测试时的推理与训练期间的事后信息**。(a) 推理驱动的 UME 方法在测试时通过中间推理生成额外上下文。(b) HEME 则在训练期间使用匹配候选项作为特权上下文：事后教师生成候选项条件化的事后信息表示，并将其蒸馏到独立嵌入器中。测试时，HEME 独立编码查询和候选项，无需事后教师或额外推理步骤。

多模态大语言模型（MLLM） [31, 51] 为实现这一目标奠定了坚实基础，能够结合任务指令来理解多模态内容 [21, 22, 30]。然而，即便有了此类指令，标准嵌入目标函数也只监督应由哪个候选项与查询匹配，却不会直接指出哪些查询线索 使该候选项变得相关。因此，当匹配取决于需要复杂理解的细粒度属性或关系时，模型可能难以应对 [6, 32]。为解决这一问题，近期的推理驱动 UME 方法会利用中间推理为嵌入生成额外上下文，但往往需要在推理阶段进行额外计算 [26, 52, 56]。

在本文中，我们探索一种训练期间可用的替代来源：匹配候选项 （即，

与查询配对的真实目标）。观察该目标有助于揭示查询中哪些细节与匹配相关。例如，在图1中，给定图像描述“猫在沙发上看电视”，匹配图像有助于识别图像描述中的相关细节，包括猫与电视之间的关系。因此，在训练期间，我们允许多模态大语言模型嵌入器将查询与其匹配候选项联合处理，同时在推理时仍对查询和候选项进行独立编码。我们将匹配候选项视为训练时的特权上下文 [47, 71], ，从而获得候选项条件化查询嵌入，并将其称为事后信息表示。这种表示为学习如何单独编码查询提供了由目标信息指导的监督信号。

尽管匹配候选项提供了有用的上下文，但从中学习有效的事后表示并非易事 [5, 62]。直接访问候选项可能会使事后表示过度受到候选项特有信息的影响，而这些信息无法仅从查询中获得 [12]。它甚至可能向候选项嵌入坍缩，使所得监督在很大程度上与标准的查询—候选项匹配重复。因此，关键挑战在于利用这种特权上下文，同时使事后表示立足于查询，并在没有候选项条件化的情况下具备可迁移性。

为此，我们提出了事后增强多模态嵌入（**HEME**），这是一种两阶段框架：首先训练事后教师，然后将其监督蒸馏到独立嵌入器中。在第一阶段，事后教师利用候选项上下文对查询编码进行条件化，同时限制对候选项特有信息的依赖。具体而言，事后注意力掩码仅允许候选项信息通过查询内容词元影响嵌入，从而阻断候选项到读出层的直接访问。我们还在不匹配候选项上对事后通路进行条件化，避免教师简单地跟随所提供的候选项。此外，我们进一步对齐事后通路与仅查询通路之间的中间查询词元表示，促使候选项条件化表示保持接近其仅查询对应表示。随后，我们冻结事后教师，并通过匹配教师的事后嵌入和相似度分布来训练独立的学生嵌入器。在推理时，学生分别对查询和候选项进行编码，无需访问匹配候选项或事后教师。实验结果表明，在单次推理下，HEME 在 20亿和70亿规模上均取得了可比模型中最高的整体 MMEB-V2 得分，同时在 MMEB-V3 智能体和文本任务上也保持竞争力，且无需针对这些任务使用专门的训练数据。

我们的主要贡献可总结如下：

- 我们为通用多模态嵌入引入事后监督，将匹配的候选项作为训练时特权上下文，同时在推理时保留独立编码。
- 我们提出了 HEME，通过学习以查询为基础的事后表示，并将其监督信号蒸馏到独立嵌入器中。
- 我们表明，HEME 在 MMEB-V2 上实现了整体更优的性能，并在 MMEB-V3 的智能体和文本任务上取得了具有竞争力的结果；这些结论得到了对所提出框架的大量分析的支持。

2. 相关工作

通用多模态嵌入（UME）。基于嵌入的检索已从对比训练的双编码器 [23, 41] 发展到能够处理多种检索任务的指令感知检索器 [30, 33, 53, 67, 70]。随着基于多模态大语言模型的嵌入器兴起，这一方向进一步拓展为通用多模态嵌入，即由单一的指令条件模型支持异构输入和广泛的下游任务 [11,19, 21, 22, 28, 36, 45]。近来，推理驱动方法通过显式推理轨迹 [6, 20, 26, 56], 、潜在推理 [13, 57],或自适应推理 [52, 68]丰富了嵌入构建过程。不同于这些推理时推理方法，HEME 仅将匹配的候选项作为训练时特权上下文，用于优化查询表示，同时在推理时保留独立编码，无需额外的推理步骤。

后见之明学习。 利用仅在训练期间可用的信号进行特权信息学习 [15,47], ，而广义蒸馏则将这一原则扩展到教师-学生学习 [35]。近年来，这一范式在大语言模型自蒸馏中重新受到关注：教师以特权上下文为条件，对必须在没有该上下文的情况下运行的学生进行监督 [18, 40, 43, 71]。对于检索，先前的工作同样利用了目标侧信息，通常通过教师或后验模型，引入参考输出或图像 [38, 54], 相关文档 [4, 58], 以及未来交互 [44]。然而，特权条件可能使教师监督偏向特权内容 [12], ，其优势也取决于能否有效迁移到学生模型 [5, 62]。为减轻这种偏差，HEME 学习以候选项为条件、同时保持以查询为基础的事后信息表示，并将其监督信号迁移到最终嵌入器。

嵌入模型蒸馏。知识蒸馏 [14] 广泛用于检索，通过将交叉编码器的相关性分数迁移到独立编码的检索器 [16, 37, 42], ，近期的 UME 模型 [28, 29]也采用了这一方法。除了分数级监督，表征蒸馏还会对齐教师和学生嵌入或它们的关系结构 [9, 25, 39, 48, 61, 66]。与此同时，非对称检索方法会在编码器之间进行蒸馏

，这些编码器的容量各不相同，通常使用更强大的编码器进行离线索引，并以此监督轻量级查询编码器 [1, 8, 24, 34]。而 HEME 蒸馏的是信息不对称，而非容量差距：教师和学生共享相同的架构，但只有教师会将匹配候选项作为特权上下文接收。

3. 方法

HEME 学习一种通用多模态嵌入器，在训练期间利用特权候选项上下文，同时在推理时保留独立编码。我们首先阐述基于多模态大语言模型的嵌入（第3.1节），然后介绍事后表示以及候选项条件化带来的挑战（第3.2节）。接着，我们描述 HEME 的两个阶段（图 2）：在候选项依赖约束下训练事后教师（第 3.3节），以及将其事后监督蒸馏到一个无需候选项上下文即可运行的独立嵌入器中（第3.4节）。

3.1. 预备知识

基于多模态大语言模型的嵌入。 每个输入查询或候选项都由多模态内容（例如文本、图像、视频、视觉文档）以及可选的任务指令构成。遵循近期的通用嵌入器 [22, 36], ，我们使用因果仅解码器多模态大语言模型，并将指定的读出标记 [Embed] 置于每个输入序列的末尾。我们提取其最后一层隐藏状态，并应用 ℓ_2 归一化以获得嵌入向量。令 f_θ 表示所得的嵌入器，令 $\mathbf{q} = f_\theta(Q)$ 和 $\mathbf{c} = f_\theta(C)$ 分别表示查询序列 Q和候选项序列 C的嵌入。由于两个嵌入均已进行单位归一化，其内积 $\mathbf{q}^\top \mathbf{c}$ 等于它们的余弦相似度，并作为相关性分数。查询和候选项均独立编码，因此可以预先计算候选项嵌入并离线建立索引。

对比目标。 给定一个包含 N 个匹配查询—候选项对 $\{(Q_i, C_i)\}_{i=1}^N$ 的批次，我们采用标准的 InfoNCE 损失 [46]。对于查询嵌入$\mathbf{q}_i$，我们定义

$$\mathcal{L}_{\text{NCE}} = \ell(\mathbf{q}_i) = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{c}_i/\tau)}{\sum_{j=1}^N \exp(\mathbf{q}_i^\top \mathbf{c}_j/\tau)}, \quad (1)$$

其中 τ 表示温度。为简化记法，所有目标函数均针对单个查询 i 定义，并在整个批次上取平均。下文中，我们通过将其他查询侧表示代入$\mathbf{q}_i$，在式（1）中复用$(\cdot)$。

3.2. 事后表示

虽然对比目标 $\mathcal{L}_{\text{NCE}}$ 会监督哪个候选项与查询匹配，但它并未明确指出哪些查询线索促成了这种匹配。为了在训练期间提供此类信号，我们通过让查询关注其匹配的候选项，构造一个候选项条件化查询表示。具体而言，我们在因果注意力下将候选项序列C_i 前置于查询序列 Q_i，并提取查询末尾 [嵌入]标记处的表示：

$$\mathbf{h}_i = f_\theta([C_i; Q_i]), \quad (2)$$

我们将其称为事后表示。与仅查询嵌入 $\mathbf{q}_i$不同，$\mathbf{h}_i$ 可以在训练时利用匹配的候选项作为特权上下文 [47], ，同时仍作为查询侧嵌入，与候选项嵌入进行评分。

然而，直接进行候选项条件化可能会使事后表示过度依赖候选项特有的信息。首先，在标准因果注意力下，末尾的 [嵌入] 词元可以直接关注候选项词元，从而使候选信息主导$\mathbf{h}_i$。如果 $\mathbf{h}_i$ 趋于坍缩为候选项嵌入，那么除对比目标已提供的正样本对信号外，它所提供的依赖查询的监督就十分有限。其次，事后表示可能会编码仅凭查询无法获取的候选项特有细节，因此这类监督难以迁移到仅查询嵌入器。为此，我们寻求一种以查询为基础的 事后表示，使其既能受益于候选项上下文，又能在不进行候选项条件化的情况下迁移。

3.3. 事后教师训练

为应对这些挑战，我们通过三个组成部分训练事后教师：事后信息掩码和不匹配的候选项条件化以防止复制，以及查询词元对齐以确保表示以查询为基础。

事后信息通路。 事后教师是一个单独的多模态大语言模型f_θ，使用共享权重执行候选项、查询和事后信息通路。候选项通路和查询通路分别计算 $\mathbf{c}_i = f_\theta(C_i)$和 $\mathbf{q}_i = f_\theta(Q_i)$。与此同时，事后信息通路根据下文所述的事后信息掩码计算 $\mathbf{h}_i$，如公式（2）所示；下文中，$\mathbf{h}_i$ 表示这一经过掩码处理的表示。由于因果注意力使候选项前缀的计算独立于后续查询，我们复用候选项通路中已分离的键值缓存，仅处理查询序列。我们使用相同的对比目标 $\mathbf{h}_i$ 进行训练 $\ell(\mathbf{h}_i)$，该目标作用于候选嵌入 $\{\mathbf{c}_j\}_{j=1}^N$。

事后信息掩码。 为防止候选项直接访问嵌入读出，事后信息掩码会选择性地限制对候选前缀的注意力。只有查询内容词元，记作 $\mathcal{T}_i$，可以关注候选标记；其他所有查询侧词元，包括任务指令和读出 [嵌入]词元，都会被阻断，无法访问候选前缀（图 2a 中的事后信息掩码）。这同时阻断了直接的候选项到嵌入 [嵌入] 路径，以及通过指令标记形成的潜在捷径。因此，候选项上下文只能通过调节查询内容词元的表示来影响$\mathbf{h}_i$。

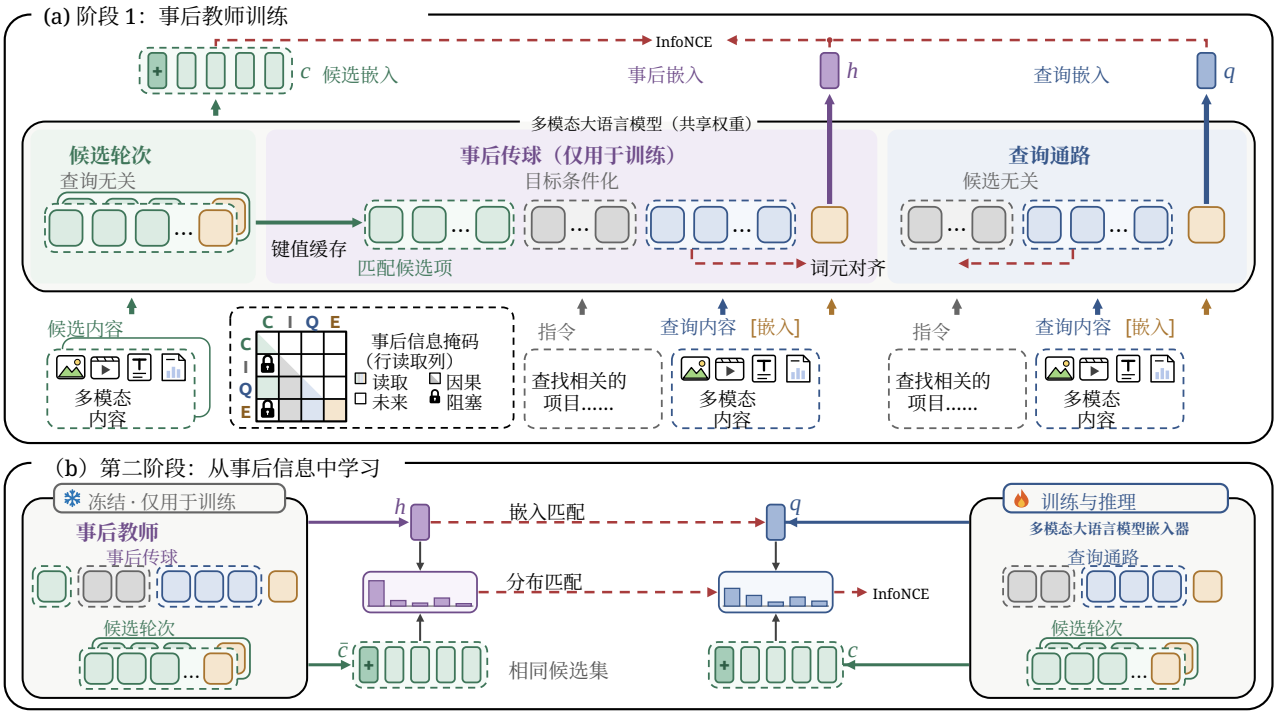

图2. **HEME 概览。(a) 事后教师训练。**候选项通路和查询通路分别独立编码各自的输入，而事后通路则通过复用候选项键值缓存，使查询以其匹配候选项为条件。事后信息掩码确保候选项上下文只能通过查询内容词元影响嵌入。结合不匹配候选条件化和查询词元对齐，这有助于使事后表示始终以查询为依据。为清晰起见，省略了不匹配候选项的事后通路。**(b) 从事后信息中学习。**冻结的事后教师通过嵌入和相似度分布匹配来指导独立学生模型，同时学生也接受标准 InfoNCE 训练。分布匹配会在相同的批内候选项上对齐教师与学生的相似度分布，并使用各自的候选项嵌入（$\bar{\mathbf{c}}$ 和 $\mathbf{c}$）。推理时仅使用独立学生嵌入器。

不匹配候选项。 即使采用事后信息掩码，查询内容词元仍然可以将候选项信息传递给 [嵌入]，，从而使$\mathbf{h}_i$仅仅跟随所提供的候选项。为抑制这种行为，我们还会使用不匹配候选项对查询进行条件化 C_k（一个批内负样本，其与 $k \neq i$），在相同掩码下得到$\tilde{\mathbf{h}}_i = f_\theta([C_k; Q_i])$。尽管条件输入是不匹配的候选项 C_k，但$\mathbf{h}_i$仍经过训练，以检索匹配候选项的嵌入$\mathbf{c}_i$：

$$\mathcal{L}_{\mathrm{H}} = \ell(\mathbf{h}_i) + \ell(\tilde{\mathbf{h}}_i). \tag{3}$$

由于 C_k 是 Q_i的负样本，若遵循不匹配候选项，就会给错误的候选项较高分数，而这会被 $\ell(\mathbf{h}_i)$直接惩罚。

词元对齐。 为进一步促使事后信息表示始终以查询为基础，我们借鉴以往关于中间表示对齐的研究，对事后信息通路与查询通路之间的中间查询词元表示进行对齐 [27, 63, 64]。具体而言，对于查询内容词元 $\mathcal{T}_i$，我们最小化

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left[1 - \cos\left(\mathbf{z}^{\mathrm{H},m}_{i,t}, \mathbf{z}^{\mathrm{Q},m}_{i,t}\right)\right], \tag{4}$$

其中$\mathbf{z}^{\mathrm{H}}_{i,t}{}^{m}$和$\mathbf{z}^{\mathrm{Q}}_{i,t}{}^{m}$分别是事后传球（使用匹配候选项）和查询通路中，查询z t在第 m 层的词元隐藏状态。该对齐直接正则化了候选项条件表示与仅查询表示之间针对相同查询词元的差异。

教师目标函数。 事后教师通过标准对比目标、后见之明目标和词元对齐目标进行训练：

$$\mathcal{L}_{\mathrm{teacher}} = \mathcal{L}_{\mathrm{NCE}} + \mathcal{L}_{\mathrm{H}} + \mathcal{L}_{\mathrm{align}}. \tag{5}$$

我们停止对复用的候选项键值缓存和嵌入进行梯度传播，因此后见之明目标函数仅通过查询侧计算进行反向传播。

3.4. 从事后信息中学习

事后教师可以利用匹配候选项生成目标信息查询表示，但这种特权上下文在推理时无法获得。因此，我们

将教师的事后监督迁移到一个独立的学生嵌入器中，该嵌入器独立编码查询和候选项（图2b）。为了在蒸馏过程中提供固定的监督信号，我们冻结训练好的教师，并训练一个独立的学生 f_ϕ，其初始化参数与教师一样，均来自同一个预训练多模态大语言模型。在这里，$\mathbf{q}_i = f_\phi(Q_i)$ 和 $\mathbf{c}_i = f_\phi(C_i)$ 表示学生的嵌入，而冻结的教师则提供事后表示 $\mathbf{h}_i$、仅查询嵌入 $\bar{\mathbf{q}}_i = f_\theta(Q_i)$ 以及候选项嵌入 $\bar{\mathbf{c}}_i = f_\theta(C_i)$。学生通过三个目标函数进行训练：嵌入匹配、分布匹配，以及标准的 InfoNCE。

嵌入匹配。 我们首先通过将学生查询嵌入 $\mathbf{q}_i$ 与冻结教师表示 $\mathbf{h}_i$ 进行匹配，来迁移事后信息表示：

$$\mathcal{L}_{\text{emb}} = 1 - \cos(\mathbf{q}_i, \mathbf{h}_i) = 1 - \mathbf{q}_i^\top \mathbf{h}_i. \tag{6}$$

由于冻结教师是在促使 $\mathbf{h}_i$ 保持以查询为基础的约束下训练得到的，$\mathbf{h}_i$ 为学生的查询表示提供了由目标信息引导的训练信号。

分布匹配。 嵌入匹配虽然逐点对齐单个表示，但不会显式监督查询与不同候选项之间的相对相似度。因此，我们在同一个批内候选集上，匹配教师和学生所诱导的相似度分布。具体而言，设 $\mathbf{p}_i^{\text{T}}$ 和 $\mathbf{p}_i^{\text{S}}$ 表示所有 N 批内候选项上的概率分布，其中 $p_{i,j}^{\text{T}} \propto \exp(\mathbf{h}_i^\top \bar{\mathbf{c}}_j / \tau_{\text{d}})$ 和 $p_{i,j}^{\text{S}} \propto \exp(\mathbf{q}_i^\top \mathbf{c}_j / \tau_{\text{d}})$，其中 τ_{d} 为蒸馏温度。我们最小化教师分布与学生分布之间的 Kullback–Leibler（KL）散度：

$$\mathcal{L}_{\text{dist}} = D_{\text{KL}}\big(\mathbf{p}_i^{\text{T}} \,\|\, \mathbf{p}_i^{\text{S}}\big) = \sum_{j=1}^{N} p_{i,j}^{\text{T}} \log \frac{p_{i,j}^{\text{T}}}{p_{i,j}^{\text{S}}}. \tag{7}$$

该目标将事后教师在候选项集合上的相似度结构迁移过来，提供了与嵌入匹配互补的关系监督。

学生目标函数。 学生将对比损失 $\mathcal{L}_{\text{NCE}} = \ell(\mathbf{q}_i)$（针对自身的候选项嵌入 $\{\mathbf{c}_j\}_{j=1}^N$ 计算）与以下两个蒸馏目标函数结合：

$$\mathcal{L}_{\text{student}} = \alpha \mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{emb}} + \mathcal{L}_{\text{dist}}, \tag{8}$$

其中，α 控制对比目标与蒸馏目标之间的平衡，正如先前的检索蒸馏 [49,50]中所采用的那样。对比目标保留了标准的嵌入训练，而嵌入匹配和分布匹配则分别提供逐点监督和关系监督。由于教师与学生共享相同的架构和预训练初始化，HEME 蒸馏的是一种<人类可读>信息不对称，而非容量差距：教师可以特权访问匹配候选项，而学生仅对查询进行编码。在推理时，仅使用独立的学生嵌入器分别对查询和候选项进行编码。

4. 实验

4.1. 实验设置

基准与指标。 我们在 MMEB- V2 [36], 上评估 HEME。该多模态嵌入基准包含 78 个数据集，涵盖图像（36 个）、视频（18 个）和视觉文档（24 个）任务。按照标准协议，我们报告图像和视频任务的 Hit@1，以及视觉文档任务的 NDCG@5。此外，我们还在 MMEB-V3 [17], 的 47 个智能体任务和 53 个文本任务上进行评估，并分别报告 Hit@1 和 NDCG@5。由于 Qwen2-VL 不支持音频输入，我们排除 11 个音频任务。各类别得分为该类别内各数据集得分的平均值，总体得分则为每个基准中所有已评估数据集得分的平均值。完整的基准详情和任务级结果见附录B和C。

实现细节。 我们基于 Qwen2-VL-20亿和 Qwen2-VL-7B [51]构建 HEME。我们的训练混合数据包括来自 MMEB- train [22], 的基于图像的视觉语言任务、来自 LLaVA- Hound [69], 的视频语言指令，以及来自 Vi- DoRe [10] 和 VisRAG [65]的视觉文档检索任务。教师模型和学生模型均从同一个预训练 MLLM 初始化，并分别在 NVIDIA H200 GPU 上以 512 的全局批大小训练一个轮次。我们使用 AdamW，学习率为 1×10^{-5}（7B 模型为 5×10^{-6}），权重衰减为 0.01，先进行 100 个线性预热步骤，再采用余弦衰减。我们冻结视觉编码器，并对多模态投影器和语言解码器进行完全微调。我们将对比温度设为 $\tau = 0.02$，将蒸馏温度设为 $\tau_{\text{d}} = 0.07$，并将平衡权重设为 $\alpha = 0.2$。我们在 28 层语言解码器的第 $m = 21$ 层应用查询令牌对齐。对于每个查询，我们从批内负样本中均匀采样一个不匹配候选项。所有消融研究均使用 20亿骨干网络，并在 MMEB-V2 上进行。

4.2. 主要结果

基线。 对于 MMEB-V2，我们将 HEME 与基于相同 Qwen2-VL 骨干网络、规模为 2B 和 7B 的模型进行比较。这些模型包括标准 UME 基线（GME [70], VLM2Vec-V 2 [36]）以及采用显式生成（UME-R1 [26], TTE$_s$ [6]、RIME [56]）、潜在推理（PLUME [13], LaME [57], VaME [55]）和自适应推理（MMEmb-R1 [52]）的推理驱动模型。我们还将采用更大骨干网络或外部训练数据 [3, 19, 28, 72] 的模型作为单独的参考进行报告。MMEB-V2 的基线分数取自原论文和基准测试报告 [36]。对于 MMEB-V3，我们采用基准测试 [17]报告的基线分数。

表 1。 **MMEB-V 上的结果2 [36]**。 HEME 在 ∼2B 和 ∼7B 两个对比组中均取得最高总体得分；采用更大骨干网络或外部训练数据的模型将单独列出作为参考。对于图像和视频任务，我们报告 Hit@1；对于视觉文档任务，我们报告 NDCG@5。Avg 和 All 分别表示每个类别内的数据集平均值和总体平均值。CLS：分类，QA：问答，RET：检索，GD：定位，MRET：时刻检索，VDR：ViDoRe，VR：VisRAG，OOD：分布外。

模型	图像					视频					VisDoc					All
	CLS	QA	RET	GD	Avg	CLS	QA	RET	MRET	平均值	VDRv1	VDRv2	VR	OOD	Avg	
# 数据集	10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
参考模型（更大的骨干网络 / 外部训练数据）																
RzenEmbed-v2-7B [19]	70.6	71.7	78.5	92.1	75.9	58.8	63.5	51.0	45.5	55.7	89.7	60.7	88.7	44.4	77.1	71.6
Qwen3-VL-Embed-8B [28]	74.2	81.1	80.2	92.3	80.1	78.4	71.0	58.7	56.1	67.1	87.2	69.9	88.7	73.3	82.4	77.8
DME-9B [3]	74.5	80.9	78.2	94.6	79.8	87.7	71.0	61.0	58.5	70.8	87.6	57.8	94.5	73.5	82.0	78.4
WeMM-Embedding-9B [72]	76.2	82.2	81.7	95.6	81.9	87.4	77.7	67.4	58.5	74.3	90.5	65.1	90.9	71.8	83.3	80.6
∼2B 模型规模																
GME [70]	54.4	29.9	66.9	55.5	51.9	34.9	42.0	25.6	32.4	33.9	86.1	54.0	82.5	43.1	72.7	54.1
VLM2Vec-V2 [36]	62.9	56.3	69.5	77.3	64.9	39.3	34.3	28.8	38.5	34.9	75.5	44.9	79.4	39.4	65.4	58.0
UME-R1 [26]	64.8	62.8	67.6	77.2	66.6	44.3	51.2	32.9	39.7	42.2	72.4	46.2	79.2	37.2	63.9	60.1
PLUME [13]	66.5	59.2	67.6	79.7	66.3	45.0	52.3	33.5	46.7	44.1	72.1	49.8	78.1	57.4	67.5	61.6
TTE_s [6]	67.9	66.6	70.2	84.1	70.1	47.3	49.1	33.2	32.1	41.3	77.5	53.2	83.2	41.1	68.8	63.1
RIME [56]	67.9	64.4	69.8	82.1	69.1	48.0	52.1	33.6	39.2	43.7	76.4	51.4	81.7	63.9	71.4	64.1
LaME [57]	67.6	66.2	70.5	81.2	69.3	45.6	52.3	36.3	43.5	44.5	78.2	48.4	82.8	55.4	72.1	64.4
MMEmb-R1 [52]	64.5	68.1	70.0	88.9	70.0	56.3	52.8	47.6	42.5	50.6	74.1	56.0	78.9	48.3	68.0	65.0
VaME [55]	67.6	66.6	70.0	80.3	69.5	46.2	54.8	38.2	45.5	46.3	79.4	55.2	85.6	64.1	74.4	65.7
HEME（我们的方法）	66.3	66.5	71.4	81.6	69.7	49.4	53.2	37.1	39.2	45.3	84.9	55.3	87.3	65.7	77.3	66.4
∼7B 模型规模																
GME [70]	57.7	34.7	71.2	59.3	56.0	37.4	50.4	28.4	38.2	38.6	89.4	55.6	85.0	44.4	75.2	57.8
UME-R1 [26]	67.1	69.2	71.9	84.9	71.3	48.6	60.7	38.2	39.3	47.5	75.7	50.5	83.7	37.6	67.1	64.5
TTE_s [6]	69.7	72.4	74.0	90.6	74.2	49.1	60.6	36.4	37.2	46.8	84.1	62.7	91.9	47.6	76.4	68.6
RIME [56]	70.3	71.7	73.2	86.3	73.4	52.6	62.0	38.4	41.6	49.4	80.9	55.6	85.8	66.9	75.6	68.6
LaME [57]	70.0	71.0	73.1	85.8	73.0	51.0	64.8	40.3	44.6	50.8	81.5	58.2	85.0	68.2	75.9	68.8
MMEmb-R1 [52]	68.1	71.8	77.7	87.1	74.4	60.2	56.0	51.2	41.1	53.3	79.6	65.3	84.7	58.0	74.9	69.7
VaME [55]	71.3	70.3	73.7	84.1	73.2	49.9	58.6	42.0	49.3	50.0	83.2	59.0	89.4	67.9	78.2	69.4
HEME（我们的方法）	69.5	71.8	74.9	87.0	73.9	53.2	60.2	40.7	41.5	49.7	88.1	60.9	88.7	69.3	80.6	70.4

表2。**推理效率**。 延迟是在单张 NVIDIA H100、批大小为 1 的设置下，使用 MMEB-V2 样本测得的。Steps 表示平均顺序推理步骤数；†LaME 在单次前向传播中处理八个潜在标记。

方法	推理	步骤	延迟（毫秒）	MMEB-V2
UME-R1 [26]	显式	408	10,903	60.1
RIME [56]	显式	153	4,145	64.1
PLUME [13]	潜在	8	295	61.6
LaME [57]	潜在	0†	51	64.4
HEME（我们的方法）	–	0	49	66.4

MMEB-V2。桌子 1 报告了 MMEB-V2 上的结果。HEME 在 ∼2B 和 ∼7B 两组中均取得最高总体得分，分别达到 66.4 和 70.4，并且比各组中最强的基线高出 0.7 分。HEME 在两组中的视觉文档性能也均为最佳，平均分分别为 77.3 和 80.6。在无需额外测试时推理的情况下，这些结果仍通过单次推理实现。

推理效率。 表 2 比较了 HEME 与推理驱动的基线在∼2 70亿模型上的推理延迟。我们使用从每个 MMEB-数据集中采样的 10 个匹配的查询–候选项对，在单张 NVIDIA H100 上以批大小 1 测量每个输入的平均延迟。

表 3。 **MMEB-V 的智能体和文本任务结果3 [17]**。HEME 在 70亿规模上取得了最高的智能体得分，并在基于 Qwen2-VL 的模型中取得了最佳总体得分。我们报告智能体任务的 Hit@1 和文本任务的 NDCG@5；All 表示全部 100 个数据集的平均值。† 模型使用了额外的纯文本训练数据。各子任务的结果见附录C。

模型	骨干网络	智能体	文本	All
# 数据集		47	53	100
VLM2Vec-V2 [36]	Qwen2-VL-2B	28.7	24.5	26.5
E5-Omni† [2]	Qwen2.5-Omni-3B	36.9	26.7	31.5
Omni-Embed-Nemotron† [60]	Qwen2.5-Omni-3B	36.5	39.2	37.9
Qwen3-VL-Embedding† [28]	Qwen3-VL-2B	39.3	39.2	39.2
HEME（我们的方法）	Qwen2-VL-2B	36.9	34.4	⋮
VLM2Vec [22]	Qwen2-VL-7B	19.7	22.2	21.0
LCO-Embedding-Omni† [59]	Qwen2.5-Omni-7B	27.8	32.4	30.2
E5-Omni† [2]	Qwen2.5-Omni-7B	36.7	26.9	31.5
GME† [70]	Qwen2-VL-7B	35.6	37.1	36.4
Qwen3-VL-Embedding† [28]	Qwen3-VL-8B	38.4	42.5	40.6
HEME-7B（我们的方法）	Qwen2-VL-7B	39.8	37.7	⋮

V2 数据集。与需要显式生成或迭代式潜在推理的方法不同，HEME 无需额外推理步骤，只需一次前向传播即可生成最终嵌入。HEME 耗时 49 ms，与单次传播的潜在基线 LaME（51 ms）相当；相比 PLUME、RIME 和 UME-R1，速度分别快约 6×、85×和 223× 。在延迟几乎相同的情况下，HEME

表 4。**事后教师设计消融**。阶段 1 约束对教师诊断和学生性能的影响。$\text{Copy}_{\text{rand}}$ / $\text{Copy}_{\text{hard}}$ （%）：将给定的不匹配候选项排在首位的比例。不匹配候选项要么是随机的批内负样本，要么是使用 InfoNCE 基线挖掘的困难负样本。$\cos(\mathbf{h}, \bar{\mathbf{q}})$：事后信息嵌入与仅查询嵌入之间的相似度。学生全部：阶段 2 后的总体 MMEB-V2 得分。

教师设计	教师			学生
	复制$_{\text{rand}}$(↓)	复制$_{\text{hard}}$ (↓)	余弦$(\mathbf{h}, \bar{\mathbf{q}})$	All
InfoNCE（无教师）	–	–	–	63.7
HEME	0.4	6.3	0.93	66.4
无 事后信息掩码	3.2	13.4	0.89	65.8
无 词元对齐	2.8	9.2	0.79	65.4
无 不匹配的候选项	67.6	89.5	0.72	63.1
无 全部约束	97.6	99.2	0.59	62.7

取得了比 LaME 更高的 MMEB-V2 分数（66.4 对 64.4）。高效编码对于大规模嵌入系统尤为重要，因为可能需要处理数百万个项目。这一点对 MMEB-V3 尤其重要，其百万规模的语料库可能导致基于推理的编码计算开销很大 [7]。

MMEB-V3。 我们进一步在 MMEB-V3 的智能体和文本任务上评估 HEME（表 3）。HEME-7B 在表格中取得了最高的智能体得分（39.8）。在 20亿和 70亿规模下，HEME 在基于 Qwen2-VL 的基线模型中也都取得了最高的总体得分，分别达到 35.6 和 38.7。从 20亿扩展到 70亿后，智能体和文本性能均有所提升。值得注意的是，这些结果是在没有专门的智能体训练数据或纯文本训练数据的情况下取得的，同时总体表现仍与使用额外文本到文本检索数据训练的模型具有竞争力。

4.3. 消融研究

事后教师设计。 我们对用于训练事后教师的约束进行消融，并评估教师行为和蒸馏学生模型的性能（表4）。我们衡量教师将给定的不匹配候选项排在首位的频率（复制率）、事后嵌入与仅查询嵌入之间的相似度，以及从每个教师蒸馏出的学生模型得分。我们在学生训练期间，采用与事后监督相同的批内设置进行教师诊断。为减少评估开销，我们使用 200 个训练步骤（∼102k 个查询），困难不匹配项取自基线模型批内负样本中排名前 8 的项。所有教师变体均在相同批次上进行评估；复制率根据批内候选项集合计算，而事后表示与查询之间的相似度则使用匹配候选项的事后表示计算。

移除事后信息掩码会使困难负样本的复制率从 6.3% 增加到 13.4%，而移除词元对齐则会使事后信息与查询之间的相似度从 0.93 降至 0.79。相应的学生得分分别从 66.4 降至 65.8 和 65.4。在没有不匹配候选条件化的情况下，即使保留掩码和词元对齐，复制率仍会升至 67.6% 和 89.5%，学生得分也会降至 63.1，低于 63.7 的 InfoNCE 基线。因此，仅靠掩码和词元对齐并不能阻止复制，因为候选项信息仍可通过查询内容词元传递**到**查询表示中；不匹配候选条件化对于抑制这一行为起着关键作用。移除全部约束后，两种复制率都会进一步升至 97% 以上，学生得分则降至 62.7。这些结果表明，当教师利用候选项上下文、同时不让其主导查询表示时，事后信息监督是有效的。

表5。**事后迁移消融**。 第二阶段目标对下游性能的影响。所有变体均使用相同的事后教师，并保留标准 InfoNCE目标。

方法	图像	视频	VisDoc	All
InfoNCE	68.2	42.4	73.0	63.7
$+\mathcal{L}_{\text{emb}}$	69.4	43.0	76.8	65.6
$+\mathcal{L}_{\text{dist}}$	69.5	44.3	76.9	65.9
$+\mathcal{L}_{\text{emb}}+\mathcal{L}_{\text{dist}}$ (**HEME**)	69.7	45.3	77.3	66.4

表6。**蒸馏目标消融实验**。用于第二阶段的教师和目标表示的影响。$\mathbf{q}$：仅使用 InfoNCE 的教师的查询嵌入；$\bar{\mathbf{q}}$：事后教师仅查询嵌入；$\mathbf{h}$：事后嵌入。 ×：直接评估。

教师	目标	图像	视频	VisDoc	All
仅使用 InfoNCE	×	68.2	42.4	73.0	63.7
	$\mathbf{q}$	67.6	42.0	72.3	63.1
事后信息（阶段 1）	×	68.7	43.2	75.8	65.0
	$\bar{\mathbf{q}}$	68.8	43.4	76.4	65.3
	$\mathbf{h}$	69.7	45.3	77.3	66.4

事后迁移。 表 5考察了教师的事后监督如何传递给学生。嵌入匹配和分布匹配分别将 InfoNCE 基线从 63.7 提升至 65.6 和 65.9。二者在图像和视觉文档任务上的得分接近，而 $\mathcal{L}_{\text{dist}}$ 在视频得分上更高（44.3 对 43.0 ）。结合这两个目标函数可将总体得分提升至 66.4，在图像、视频和视觉文档任务上均优于单独使用任一目标函数，这表明这两种监督形式并非完全冗余。

蒸馏目标。 表6 考察学生增益究竟来自蒸馏本身、更强的教师，还是特定的事后目标。对仅使用 InfoNCE 的教师的查询嵌入进行蒸馏，并未优于 InfoNCE 基线，得分为 63.1，而基线为 63.7。事后教师的查询通路已经强于基线（65.0 对 63.7），对其仅查询表示进行蒸馏后 $\bar{\mathbf{q}}$ 得分达到 65.3。在保持教师不变的情况下，对其事后表示进行蒸馏



图3. **查询锚定的后见信息及其向推理的迁移。** (a) 对于目标 C^+ 和一个具有挑战性的近似未命中项 C^-，事后传球在用于区分两个候选项的词上显示出更大的查询词元偏移；在推理时，HEME在所示示例中表现出更强的目标–近似未命中项分离。(b) 对于回答两个不同问题的单个页面 C^+，以同一答案为条件会产生问题特定的后见排序；仅根据问题，HEME表现出相应的检索行为。

h 则得分达到 66.4，相较于$^-$**q**进一步提升了 1.1 分。这表明，候选项条件化的后见目标在教师的仅查询表示之外提供了额外监督。

事后行为。图 3 考察候选项上下文如何改变后见表示，以及相应行为是否会出现在仅查询学生模型中。在图 3（a）中，我们固定查询，并比较查询词元在以目标或由教师的查询通路从前10个困难负例中选出的近似未命中候选项为条件时的隐藏状态。在所示示例中，最大变化出现在区分匹配候选项与近似未命中候选项的词上。这些区别涵盖不同层次的视觉语义，包括局部对象属性（衬衫颜色）、场景上下文（海与草地），以及全局图像外观（彩色与黑白）。接下来，我们通过测量查询在匹配候选项与同一近似未命中候选项之间的区分度，评估这些区别是否会传递到学生模型。在推理时，在三个示例中，HEME 的标准化目标项–近似未命中项间隔均大于 InfoNCE 基线；其中，间隔按每个模型的候选项相似度的标准差进行归一化；以目标为条件的事后传球产生的间隔甚至更大（斜线柱）。除这些示例外，在来自 COCO、VisualNews、CIRR 和 VisDial 的超过1.8万个近似未命中对中，HEME 有79.7%的比例取得了大于基线的标准化间隔。

图3（b）通过固定匹配页面并改变问题，对VisRAG页面检索提供了补充分析。后见排序不同于仅页面本身的最近邻，表明它们并非仅由候选项的嵌入邻域决定。此外，以同一页面为条件会为两个问题产生不同的排序：与可持续发展相关的页面更受环境问题青睐，而关于社区项目和慈善活动的页面更受慈善问题青睐。仅根据问题，HEME表现出与后见排序相似的查询特定检索偏好。这些分析表明，事后教师以依赖于查询的方式使用候选项上下文，而独立的学生则表现出相关行为。

5. 结论

我们提出了 HEME，这是一种利用匹配候选项作为训练时特权上下文的框架，用于学习通用多模态嵌入，同时在推理时保留独立编码。HEME 通过控制对特权上下文的依赖来学习以查询为基础的事后信息表示，

并将其蒸馏到独立的嵌入器中。在 MMEB-V2 和 MMEB-V3 上进行的实验表明，HEME 在各种任务中均具有效果，同时推理效率高且无需额外的推理步骤。我们的分析表明，控制这种依赖至关重要，而事后目标能够在仅基于查询的教师表示之外提供额外监督。

大语言模型使用声明

我们使用了 ChatGPT 和 Codex（OpenAI）以及 Claude（Anthropic）来协助进行稿件编辑、代码实现与调试，以及实验设计。所有人工智能辅助生成的内容均经过作者审阅和核验；作者对科学论断、引用、代码、实验结果以及最终稿件承担全部责任。

致谢

本工作部分得到了大韩民国政府（科学技术信息通信部）资助的“先进 GPU 利用支持计划”的支持。

参考文献

[1] Mateusz Budnik 和 Yannis Avrithis。用于知识迁移的非对称度量学习。载于 CVPR，2021年。3[2] Haonan Chen、Sicheng Gao、Radu Timofte、Tetsuya Sakai 和 Zhicheng Dou。E5-Omni：面向全模态嵌入的显式跨模态对齐。载于 ACL Findings，2026年。6、17[3] Haonan Chen、Chu Li、Zhicheng Wang、Yuanwei Liu、Yuanjiang Wang、Shaohua Jiang 和 Zhicheng Dou。抖音多模态嵌入模型技术报告。arXiv 预印本，arXiv:2608.02148，2026年。5、6[4] Teng Chen、Sheng Xu、Feixiang Guo、Xiaoyu Wang、Qingqing Gu、Hongyan Li 和 Luo Ji。高效的基于推理过程的检索：基于 JEPA 从生成式重排序器进行在线策略蒸馏。载于 ICMR，2026年。2[5] Sanjiban Choudhury 和 Paloma Sodhi。胜过你的教师：从特权人工智能反馈中学习的大语言模型智能体。载于 ICLR，2025年。2[6]Xuanming Cui、Jianpeng Cheng、Hong-You Chen、Satya Narayan Shukla、Abhijeet Awasthi、Xichen Pan、Chaitanya Ahuja、Shlok Mishra、Taipeng Tian、Qi Guo、Ser-Nam Lim、Aashu Singh 和 Xiangjun Fan。先思考后嵌入：生成式上下文改善多模态嵌入。载于 ICLR，2026年。1、2、5、6、15[7] Xuanming Cui、Shlok Kumar Mishra、Wentao Bao、Aashu Singh、Zihao Wang、Xiangjun Fan、Jun Xiao、Ser-Nam Lim 和 Jianpeng Cheng。MoEMB：利用高效混合专家模型扩展通用多模态嵌入。arXiv 预印本，arXiv:2609.08663，2026年。7[8] Yongchao Du、Min Wang、Wengang Zhou、Shuping Hui 和 Houqiang Li。基于 Image2Sentence 的非对称零样本组合图像检索。载于 ICLR，2024年。3

[9] Zhiyuan Fang、Jianfeng Wang、Lijuan Wang、Lei Zhang、Yezhou Yang 和 Zicheng Liu。《SEED：用于视觉表征的自监督蒸馏》。发表于 ICLR，2021年。2[10] Manuel Faysse、Hugues Sibille、Tony Wu、Bilel O mrani、Gautier Viaud、C´eline Hudelot 和 Pierre Colombo。《Col-Pali：使用视觉语言模型进行高效文档检索》。发表于 ICLR，2025年。5, 20[11]Tiancheng Gu、Kaicheng Yang、Ziyong Feng、Xingjun Wang、Yanzhao Zhang、Dingkun Long、Yingda Chen、Weidong Cai 和 Jiankang Deng。《打破模态壁垒：借助多模态大语言模型实现通用嵌入学习》。发表于 ACM MM，2025年。1, 2[12] Sarthak Harne、Chinmay Karkar、Yash Pandya、Ahmed Awadallah 和 Akshay Nambi。《特权但有偏：PI条件教师模型如何破坏自蒸馏》。arXiv 预印本 arXiv:2608.04794，2026年。2[13] Chenwei He、Xiangzhao Hao、Tianyu Yang、Yuxiang Ma、Yuheng Jia、Lingxiang Wu、Chaoyang Zhao、Haiyun Guo 和 Jinqiao Wang。《PLUME：基于潜在推理的通用多模态嵌入》。发表于 ACM MM，2026年。2, 5, 6[14] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。《神经网络中的知识蒸馏》。arXiv 预印本 arXiv:1503.02531，2015年。2[15] Judy Hoffman、S aurabh Gupta 和 Trevor Darrell。《利用模态幻觉进行侧信息学习》。发表于 CVPR，2016年。2[16] Sebastian Hofst¨ atter、Sophia Althammer、Michael Schr¨oder、Mete Sertkan 和 Allan Hanbury。《通过跨架构知识蒸馏改进高效神经排序模型》。arXiv 预印本 arXiv:2010.02666，2020年。2[17] Haohang Huang、Xuan Lu、Mingyi Su、Xuan Zhang、Ziyan Jiang、Ping Nie、Kai Zou、Tomas Pfister、Wenhu Chen、Wei Zhang、Xiaoyu Shen 和 Rui Meng。《MMEB-V3：衡量全模态嵌入模型的性能差距》。发表于 COLM，2026年。5, 6, 17[18] Jonas H¨ubotter、Frederike L¨ ubeck、Lejs Deen Behric、Anton Baumann、Marco Bagatella、Daniel Marta、Ido Hakimi、Idan Shenfeld、Thomas Kleine Buening、Carlos Guestrin 和 Andreas Krause。《通过自蒸馏进行强化学习》。发表于 ICML，2026年。2[19] Weijian Jian、Yajun Zhang、Dawei Liang、Chunyu Xie、Yixiao He、Dawei Leng 和 Yuhui Yin。《RzenEmbed：迈向全面的多模态检索》。arXiv 预印本 arXiv:2510.27350，2025年。2, 5, 6[20] Haonan Jiang、Yuji Wang、Yongjie Zhu、Xin Lu、Wenyu Qin、Meng Wang、Pengfei Wan 和 Yansong Tang。《Embed-RL：面向推理驱动的多模态嵌入的强化学习》。发表于 ECCV，2026年。2[21] Ting Jiang、Minghui Song、Zihan Zhang、Haizhen Huang、Weiwei Deng、Feng Sun、Qi Zhang、Deqing Wang 和 Fuzhen Zhuang。《E5-V：借助多模态大语言模型实现通用嵌入》。arXiv 预印本 arXiv:2407.12580，2024年。1, 2[22] Ziyan Jiang、Rui Meng、Xinyi Yang、Semih Yavuz、Yingbo Zhou 和 Wenhu Chen。《VLM2Vec：训练视觉语言模型以完成大规模多模态嵌入任务》。发表于 ICLR，2025年。1, 2, 3, 5, 6, 17, 20

[23]Vladimir Karpukhin、Barlas O˘guz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wentau Yih。《用于开放域问答的密集段落检索》。发表于 EMNLP，2020年。2[24] Jaeyoon Kim、Yoonki Cho 和 Sung-Eui Yoon。《通过非对称查询处理实现测试时高效的视觉地点识别》。发表于 AAAI，2026年。3[25] Seungyeon Kim、Ankit Singh Rawat、Manzil Zaheer、Sadeep Jayasumana、Veeranjaneyulu Sadhanala、Wittawat Jitkrittum、Aditya Krishna Menon、Rob Fergus 和 Sanjiv Kumar。《EmbedDistill：用于信息检索的几何知识蒸馏》。arXiv 预印本 arXiv:2301.12005，2023年。2[26] Zhibin Lan、Liqiang Niu、Fandong Meng、Jie Zhou 和 Jinsong Su。《UME-R1：探索推理驱动的生成式多模态嵌入》。发表于 ICLR，2026年。1, 2, 5, 6, 15[27] Fuhao Li、Wenxuan Song、Han Zhao、Jingbo Wang、Pengxiang Ding、Donglin Wang、Long Zeng 和 Haoang Li。《空间强制：视觉-语言-动作模型的隐式空间表征对齐》。发表于 ICLR，2026年。4[28] Mingxin Li、Yanzhao Zhang、Dingkun Long、Keqin Chen、Sibo Song、Shuai Bai、Zhibo Yang、Pengjun Xie、An Yang、Dayiheng Liu、Jingren Zhou 和 Junyang Lin。《Qwen3-VL-Embedding 和 Qwen3-VL-Reranker：最先进的多模态检索与排序统一框架》。arXiv 预印本 arXiv:2601.04720，2026年。2, 5, 6, 17[29] Xiaojie Li、Chu Li、Shi-Zhe Chen 和 Xi Chen。《U-MARVEL：借助多模态大语言模型，通过嵌入学习揭示通用多模态检索的关键因素》。发表于 ICLR，2026年。2[30] Sheng-Chieh Lin、Chankyu Lee、Mohammad Shoeybi、Jimmy Lin、Bryan Catanzaro 和 Wei Ping。《MM-Embed：借助多模态大语言模型实现通用多模态检索》。发表于 ICLR，2025年。1, 2[31] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。《视觉指令微调》。发表于 NeurIPS，2023年。1[32] Qianying Liu、Xiao Liang、Zhiqiang Zhang、Yibo Chen、Xu Tang、Zhongfei Qing、Fengfan Zhou、Yao Hu 和 Paul Henderson。《ReMatch：通过匹配增强多模态检索的表征》。发表于 CVPR，2026年。1[33] Yikun Liu、Yajie Zhang、Jiayin Cai、Xiaolong Jiang、Yao Hu、Jiangchao Yao、Yanfeng Wang 和 Weidi Xie。《LamRA：将大型多模态模型用作先进的检索助手》。发表于 CVPR，2025年。2[34] Zhuchenyang Liu、Yao Zhang 和 Yu Xiao。《NanoVDR：将20亿参数的视觉-语言检索器蒸馏为7000万参数的纯文本编码器，用于视觉文档检索》。发表于 EMNLP，2026年。3[35] David Lopez-Paz、L´eon Bottou、Bernhard Sch¨olkopf 和 Vladimir Vapnik。《统一蒸馏与特权信息》。发表于 ICLR，2016年。2[36] Rui Meng、Ziyan Jiang、Ye Liu、Mingyi Su、Xinyi Yang、Yuepeng Fu、Can Qin、Raghuveer Thirukovalluru、Xuan Zhang、Zeyuan Chen、Ran Xu、Caiming Xiong、Yingbo Zhou、Wenhu Chen 和 Semih Yavuz。《VLM2Vec-V2：推进视频、图像和视觉文档的多模态嵌入》。发表于 TMLR，2026年。1, 2, 3, 5, 6, 14, 15, 17[37] Antoine Miech、Jean-Baptiste Alayrac、Ivan Laptev、Josef Sivic 和 Andrew Zisserman。《思考，快与慢：高效……》借助 Transformer 实现高效的文本到视觉检索。在 CVPR，2021年。2[38] Ashwin Paranjape、Omar Khattab、Christopher Potts、Matei Zaharia 和 Christopher D. Manning。事后信息：通过后验引导训练检索器，以提升开放式生成效果。在 ICLR，2022年。2[39] Wonpyo Park、Dongju Kim、Yan Lu 和 Minsu Cho。关系知识蒸馏。在 CVPR，2019年。2[40] Emiliano Penaloza、Dheeraj Vattikonda、Nicolas Gontier、Alexandre Lacoste、Laurent Charlin 和 Massimo Caccia。面向语言模型的特权信息蒸馏。在 ICML，2026年。2[41]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移视觉模型。在 ICML，2021年。2[42] Ruiyang Ren、Yingqi Qu、Jing Liu、Wayne Xin Zhao、Qiao-qiao She、Hua Wu、Haifeng Wang 和 Ji-Rong Wen。RocketQAv2：一种用于密集段落检索和段落重排序的联合训练方法。在 EMNLP，2021年。2[43] Idan Shenfeld、Mehul Damani、Jonas H¨ubotter 和 Pulkit Agrawal。自蒸馏助力持续学习。在 ICML，2026年。2[44] Jiakai Tang、Yang Zhang、See-Kiong Ng、Xu Chen、Wen Chen、Jian Wu 和 Han Zhu。向未来学习：用于序列推荐的特权自蒸馏。arXiv 预印本 arXiv:2607.27055，2026年。2[45] Raghuveer Thirukovalluru、Rui Meng、Ye Liu、Karthikeyan K、Mingyi Su、Ping Nie、Semih Yavuz、Yingbo Zhou、Wenhu Chen 和 Bhuwan Dhingra。通过智能批次挖掘打破对比学习的批次屏障（B3）。在 NeurIPS，2025年。2[46] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。arXiv 预印本 arXiv:1807.03748，2018年。3[47] Vladimir Vapnik 和 Akshay Vashist。一种新的学习范式：利用特权信息进行学习。《神经网络》，2009年。2, 3[48] Henrique Schechter Vera、Sahil Dua、Biao Zhang、Daniel Salz、Ryan Mullins、Sindhu Raghuram Panyam、Sara Smoot、Iftekhar Naim、Joe Zou、Feiyang Chen 等。EmbeddingGemma：强大且轻量的文本表征。arXiv 预印本 arXiv:2509.20354，2025年。2[49] Liang Wang、Nan Yang、Xiaolong Huang、Binxing Jiao、Linjun Yang、Daxin Jiang、Rangan Majumder 和 Furu Wei。通过弱监督对比式预训练获得文本嵌入。arXiv 预印本 arXiv:2212.03533，2022年。5[50] Liang Wang、Nan Yang、Xiaolong Huang、Binxing Jiao、Linjun Yang、Daxin Jiang、Rangan Majumder 和 Furu Wei。SimLM：利用表征瓶颈进行密集段落检索预训练。在 ACL，2023年。5[51] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhihao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-VL：增强视觉语言模型在任意分辨率下对世界的感知能力。arXiv 预印本 arXiv:2409.12191，2024年。1, 5, 20

[52] Yuchi Wang、Dingkang Yang、Haiyang Yu、Weikang Bian、Jiefeng Long、Xiao Liang、Chao Feng 和 Hongsheng Li。MMEmb-R1：通过配对感知选择和自适应控制增强推理能力的多模态嵌入。在 EMNLP，2026年。1, 2, 5, 6[53] Cong Wei、Yang Chen、Haonan Chen、Hexiang Hu、Ge Zhang、Jie Fu、Alan Ritter 和 Wenhu Chen。UniIR：通用多模态信息检索器的训练与基准测试。在 ECCV，2024年。2[54] Haokun Wen、Xian Zhang、Xuemeng Song、Yinwei Wei 和 Liqiang Nie。目标引导的组合图像检索。在 ACM MM，2023年。2[55] Peixi Wu、Mingzhou Jiang、Feipeng Ma、Biao Yang、Yun-hao Zhou、Wei Yuan、Bosong Chai、Huizu Lin、Jie Chen、Zhangchi Hu、Fan Yang、Wenwu Ou、Hebei Li 和 Xiaoyan Sun。VaME：探索用于多模态嵌入的变分潜在推理。arXiv 预印本 arXiv:2609.33402，2026年。5, 6[56] Peixi Wu、Ke Mei、Feipeng Ma、Bosong Chai、Zhibin Lan、Chenxi Zhao、Shannan Yan、Jie Chen、Zhangchi Hu、Yan-song Peng 等。超越思维链：将重写作为生成式多模态嵌入的通用接口。在 ACM MM，2026年。1, 2, 5, 6, 15[57] Peixi Wu、Biao Yang、Feipeng Ma、Bosong Chai、Bo Lin、Wei Yuan、Fan Yang、Tingting Gao、Hebei Li 和 Xiaoyan Sun。LaME：通过信息瓶颈学习在潜在空间中思考，以实现多模态嵌入。在 EMNLP Findings，2026年。2, 5, 6[58] Zehua Xia、Yuyang Wu、Yiyun Xia 和 Cam-Tu Nguyen。用于多跳密集检索的动量后验正则化。在 COLING，2025年。2[59] Chenghao Xiao、Hou Pong Chan、Hao Zhang、Weiwen Xu、Mahani Aljunied 和 Yu Rong。扩展以语言为中心的全模态表征学习。在 NeurIPS，2025年。6, 17[60] Mengyao Xu、Wenfei Zhou、Yauhen Babakhin、Gabriel Moreira、Ronay Ak、Radek Osmulski、Bo Liu、Even Oldridge 和 Benedikt Schifferer。Omni-Embed-Nemotron：面向文本、图像、音频和视频的统一多模态检索模型。arXiv 预印本 arXiv:2510.03458，2025年。6, 17[61] Chuanguang Yang、Zhulin An、Libo Huang、Junyu Bi、Xinqiang Yu、Han Yang、Boyu Diao 和 Yongjun Xu。CLIP-KD：关于 CLIP 模型蒸馏的实证研究。在 CVPR，2024年。2[62] Shuo Yang、Sujay Sanghavi、Holakou Rahmanian、Jan Bakus 和 S. V. N. Vishwanathan。探索学习排序中的特权特征蒸馏。在 NeurIPS，2022年。2[63] Heeji Yoon、Jaewoo Jung、Junwan Kim、Hyungyu Choi、Heeseong Shin、Sangbeom Lim、Honggyu An、Chaehyun Kim、Jisang Han、Donghyun Kim、Chanho Eom、Sunghwan Hong 和 Seungryong Kim。面向多模态大语言模型的视觉表征对齐。arXiv 预印本 arXiv:2509.07979，2025年。4[64] Sihyun Yu、Sangkyung Kwak、Huiwon Jang、Jongheon Jeong、Jonathan Huang、Jinwoo Shin 和 Saining Xie。面向生成的表征对齐：训练扩散变换器比你想象的更容易。在 ICLR，2025年。4[65] Shi Yu、Chaoyue Tang、Bokai Xu、Junbo Cui、Junhao Ran、Yukun Yan、Zhenghao Liu、Shuo Wang、Xu Han、Zhiyuan Liu 和 Maosong Sun。VisRAG：面向多模态文档的基于视觉的检索增强生成。发表于 ICLR，2025年。5, 20[66] Dun Zhang、Jiacheng Li、Ziyang Zeng 和 Fulong Wang。Jasper 和 Stella：最先进的嵌入模型的蒸馏。arXiv 预印本 arXiv:2412.19048，2024年。2[67] Kai Zhang、Yi Luan、Hexiang Hu、Kenton Lee、Siyuan Qiao、Wenhu Chen、Yu Su 和 Ming-Wei Chang。MagicLens：通过开放式指令进行自监督图像检索。发表于 ICML，2024年。2[68] Longxiang Zhang、Weilong Dai、Guanghao Zhang、Hao Jiang 和 Pipei Huang。按需思考：采用双LoRA架构的自适应推理驱动的多模态嵌入。arXiv 预印本 arXiv:2605.14448，2026年。2[69] Ruohong Zhang、Liangke Gui、Zhiqing Sun、Yihao Feng、Keyang Xu、Yuanhan Zhang、Di Fu、Chunyuan Li、Alexander G. Hauptmann、Yonatan Bisk 和 Yiming Yang。基于语言模型奖励的视频大型多模态模型直接偏好优化。发表于 NAACL，2025年。5, 20[70] Xin Zhang、Yanzhao Zhang、Wen Xie、Mingxin Li、Ziqi Dai、Dingkun Long、Pengjun Xie、Meishan Zhang、Wenjie Li 和 Min Zhang。打通模态：借助多模态大语言模型改进通用多模态检索。发表于 CVPR，2025年。1, 2, 5, 6, 15, 17[71] Siyan Zhao、Zhihui Xie、Mengchen Liu、Jing Huang、Guan Pang、Feiyu Chen 和 Aditya Grover。自蒸馏推理器：大型语言模型的在线策略自蒸馏。发表于 ICML，2026年。2[72] Junjie Zhou、Ke Mei、Lei Li、Tianyi Wang、Fengyun Rao 和 Jing Lyu。WeMM-Embedding：微信多模态嵌入技术报告。arXiv 预印本 arXiv:2608 24053，2026年 5, 6。

A. 补充定性分析

我们补充提供了图像（图S1）、视觉文档（图S2）和视频（图S3）检索的其他定性示例。

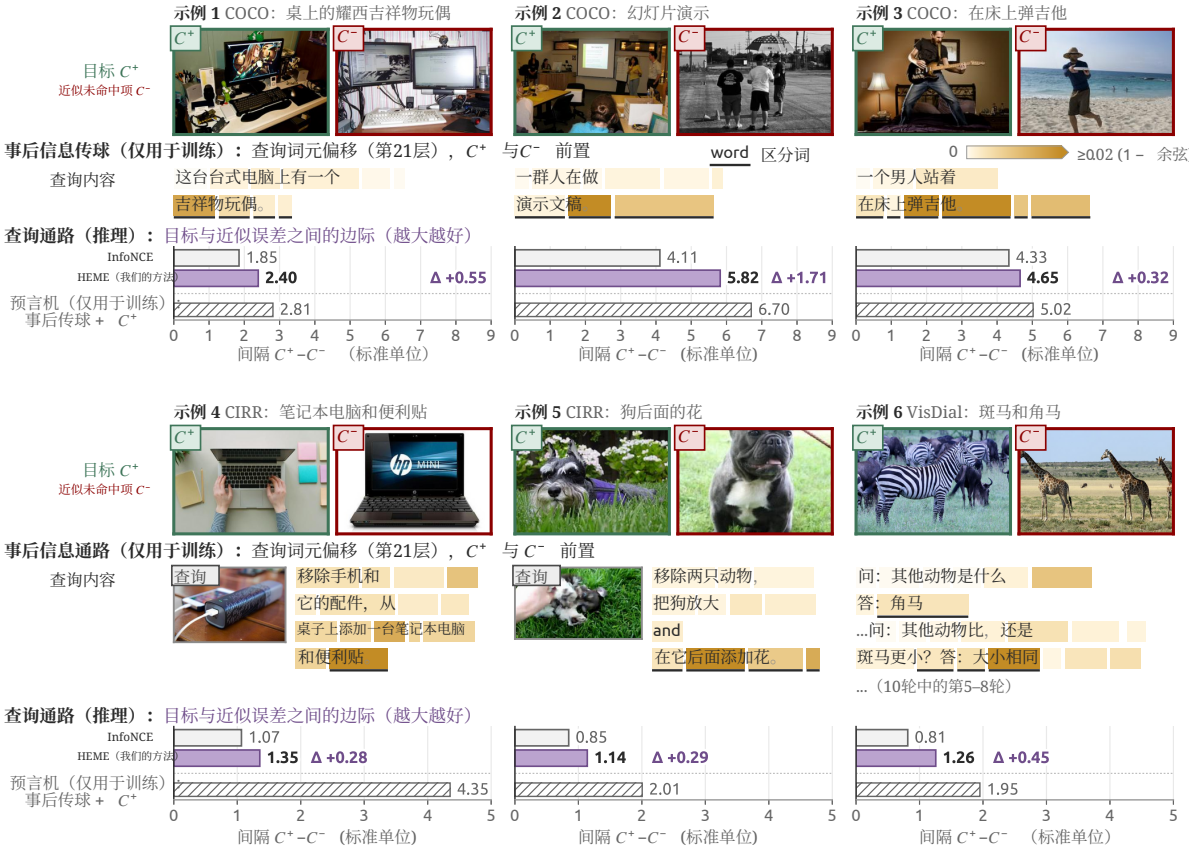

图 S1. **图像任务中事后信息的定性分析**。阴影部分显示以 C^+ 和 C^-为条件时的查询词元偏移。柱形表示经各模型候选项得分标准差归一化的检索间隔。



图 S2. **视觉文档检索中的问题条件化事后回顾**。三个问题共用匹配页面 C^+。对于每个问题，我们展示事后传球、HEME 和 InfoNCE 排名前三的页面；为保持清晰，省略 C^+。HEME 和 InfoNCE 仅使用问题。彩色边框标出也出现在事后信息排名前三中的页面。

视频检索。 图 S3将相同的词元偏移分析扩展到 MSVD、VATEX、YouCook2 和 DiDeMo 上的文本到视频检索。



图 S3。**事后信息在文本到视频检索中的定性分析。** 阴影表示在以 C^+ 和 C^- 为条件时的查询词元偏移。每一对展示来自 C^+ 和 C^- 的帧，按时间顺序排列。下划线标记目标特定短语，三角形标记词元偏移最大的词。

B. MMEB-V2 评估

基准与指标。我们遵循 MMEB-V2 评估协议 [36]。我们分别对查询和候选项进行编码，并根据候选项与查询的相似度对候选项排序。每个数据集在其类别平均值和总体得分中所占权重相同。表S1 按模态和任务列出了 78 个数据集。对于 36 个图像数据集和 18 个视频数据集，我们报告 Hit@1；对于 24 个视觉文档（VisDoc）数据集，我们报告 NDCG@5。对于分类、问答和视觉定位，候选项分别为类别标签、答案和图像区域。

表 S1. **按模态和任务划分的 MMEB-V2 数据集**。缩写 t2i 和 i2t 分别表示文本到图像检索和图像到文本检索。

模态	元任务	#	数据集
图像	分类（CLS）	10	ImageNet-1K、N24News、HatefulMemes、VOC2007、SUN397、Place365、ImageNet-A、ImageNet-R、ObjectNet、Country211
	问答（QA）	10	OK-VQA、A-OKVQA、DocVQA、InfographicsVQA、ChartQA、Visual7W、ScienceQA、VizWiz、GQA、TextVQA
	检索（RET）	12	VisDial、CIRR、VisualNews（文本到图像）、VisualNews（图像到文本）、MSCOCO（文本到图像）、MSCOCO（图像到文本）、NIGHTS、WebQA、FashionIQ、Wiki-SS-NQ、OVEN、EDIS
	定位（定位）	4	MSCOCO、RefCOCO、RefCOCO-Matching、Visual7W-Pointing
视频	分类（CLS）	5	K700、SSv2、HMDB51、UCF101、Breakfast
	问答（QA）	5	MVBench、Video-MME、NExT-QA、EgoSchema、ActivityNetQA
	检索（RET）	5	MSR-VTT、MSVD、DiDeMo、VATEX、YouCook2
	时刻检索（MRET）	3	QVHighlights、Charades-STA、MomentSeeker
VisDoc	ViDoRe v1（VDRv1）	10	ArxivQA、DocVQA、InfoVQA、TabFQuAD、TAT-DQA、ShiftProject、SyntheticDocQA（人工智能，能源、政府报告、医疗保健行业）
	ViDoRe v2（VDRv2）	4	ESG报告（人工标注），以及多语言生物医学讲座、经济学报告、和ESG报告
	VisRAG（VR）	6	ArxivQA、ChartQA、MP-DocVQA、SlideVQA、InfoVQA、PlotQA
	域外（OOD）	4	ViDoSeek（页面、文档），MMLongBench-Doc（页面、文档）
总计		78	

超参数分析。对于 Qwen2-VL-2B，我们每次改变一个参数（表S2）。改变对比损失权重 α 或蒸馏温度 τ_d时，我们固定教师模型。当改变词元对齐层 m时，我们会重新训练教师模型和学生模型。默认设置（$\alpha = 0.2$， $\tau_d = 0.07$， $m = 21$）在每次参数扫描中都取得了最高总体得分。

表 S2。**MMEB-V2 上的超参数分析**。阴影表示默认设置。粗体表示每次参数扫描中各列的最高得分。

（a）平衡权重 α

α	图像	视频	VisDoc	All
0.1	69.4	**46.0**	76.4	66.1
0.2	**69.7**	45.3	**77.3**	**66.4**
1.0	68.7	43.5	76.7	65.3

(b) 蒸馏温度 τ_d

τ_d	图像	视频	VisDoc	All
0.05	69.2	44.3	77.0	65.9
0.07	**69.7**	**45.3**	**77.3**	**66.4**
0.20	68.9	43.7	76.7	65.5

(c) 对齐层 m

m	图像	视频	VisDoc	All
14	68.8	42.4	74.9	64.6
21	69.7	**45.3**	**77.3**	**66.4**
27	**69.9**	43.5	76.8	65.9

额外的 InfoNCE 训练。 在第二轮次进行 InfoNCE 训练，使总体得分提高了 0.2 分（63。7 → 63.9；表 S3）。HEME 达到 66.4，在视频（+4.0）和 VisDoc（+3.4）上的提升最大，相较于两轮次基线模型而言。

表 S3。**在 MMEB-V2 上进行额外的 InfoNCE 训练**。所有模型均使用 Qwen2-VL-2B。InfoNCE 训练一或两个轮次；HEME 分别对教师和学生各训练一个轮次。粗体标记每列中的最高分。

方法	图像	视频	VisDoc	All
InfoNCE（1 轮）	68.2	42.4	73.0	63.7
InfoNCE（2 轮）	68.5	41.3	73.9	63.9
HEME（我们的方法）	**69.7**	**45.3**	**77.3**	**66.4**

B.1. 按数据集划分的结果

我们报告标准嵌入和显式推理基线在任务级别的 MMEB-V2 分数及类别平均值。VLM2Vec-V2 和 GME 的分数 [70] 来自基准测试报告 [36]。对于 UME-R1 [26], TTE$_s$ [6]和 RIME, [56], 我们采用其各自论文中的分数。

图像任务。 表格S4 将 36 个数据集划分为分类、问答、检索和定位。HEME-2B 和 HEME-7B 的平均 Hit@1 分数分别为 69.7 和 73.9。

表 S4。**MMEB-V2 图像结果（Hit@1）**。第一行给出了涵盖全部 78 项数据集的整体 MMEB-V2 得分。粗体表示每行中的最高得分。

数据集	GME		VLM2Vec-V2	UME-R1		TTE$_s$		RIME		HEME	
	2B	7B	2B	2B	7B	2B	7B	2B	7B	2B	7B
平均值 – 全部（78项任务）	54.1	57.8	58.0	60.1	64.5	63.1	68.6	64.1	68.6	66.4	**70.4**
平均值 – 图像（36项任务）	51.9	56.0	64.9	66.6	71.3	70.1	**74.2**	69.1	73.4	69.7	73.9
分类（10）	54.4	57.7	62.9	64.8	67.1	67.9	69.7	67.9	**70.3**	66.3	69.5
问答（10）	29.9	34.7	56.3	62.8	69.2	66.6	**72.4**	64.4	71.7	66.5	71.8
RET（12）	66.9	71.2	69.5	67.6	71.9	70.2	74.0	69.8	73.2	71.4	**74.9**
定位（4）	55.5	59.3	77.3	77.2	84.9	84.1	**90.6**	82.1	86.3	81.6	87.0
ImageNet-1K	58.3	64.6	80.8	75.3	80.4	83.3	**84.3**	81.2	80.9	81.0	83.0
N24News	50.1	50.5	72.9	81.1	82.3	78.6	**83.1**	80.0	82.7	78.3	82.8
HatefulMemes	52.5	53.6	56.3	75.2	**79.0**	64.0	67.4	68.4	76.2	64.2	71.7
VOC2007	75.9	80.3	85.0	80.0	90.8	86.3	86.6	90.4	**91.0**	86.2	89.0
SUN397	67.3	69.5	71.0	79.4	80.3	77.5	78.9	80.1	80.6	80.7	**81.5**
Place365	35.8	39.1	35.9	42.6	**46.8**	45.7	44.6	45.3	45.5	44.1	46.0
ImageNet-A	28.8	41.2	47.4	50.4	53.9	50.9	**60.4**	52.1	57.4	50.9	54.5
ImageNet-R	78.6	83.9	89.3	88.7	90.1	89.7	90.5	89.9	89.9	90.9	**91.2**
ObjectNet	70.6	69.0	65.2	52.0	42.3	**74.1**	72.6	66.1	72.7	64.6	68.6
Country211	26.5	24.8	25.2	23.4	25.0	28.5	**29.0**	25.5	26.4	22.4	26.7
OK-VQA	29.9	33.2	51.5	62.4	71.7	68.4	74.7	65.8	74.1	68.5	**74.8**
A-OKVQA	18.6	21.0	43.6	51.1	58.7	57.1	**66.1**	56.4	61.8	58.6	62.9
DocVQA	29.8	41.4	90.1	92.2	93.8	94.2	**95.6**	93.4	94.4	94.3	95.0
InfographicsVQA	11.6	20.3	58.8	67.7	**79.2**	65.6	77.5	63.6	79.1	69.7	78.3
ChartQA	13.4	17.8	47.4	64.9	75.1	57.5	70.9	61.7	**77.4**	57.0	68.7
Visual7W	16.2	22.2	52.9	54.1	55.2	54.1	57.9	54.8	54.9	54.6	**61.2**
ScienceQA	27.3	28.0	38.2	42.7	53.7	50.7	**60.0**	45.7	59.0	50.0	59.1
VizWiz	37.0	39.0	43.3	46.8	51.6	55.1	53.8	47.7	**55.3**	52.7	55.1
GQA	75.1	76.9	64.9	67.3	69.3	77.0	**80.9**	73.1	73.6	75.3	75.6
TextVQA	39.7	46.8	72.2	78.6	83.5	86.2	87.0	81.3	87.0	84.5	**87.5**
VisDial	48.1	60.8	82.7	76.6	80.7	81.2	84.4	80.7	82.6	83.4	**85.9**
CIRR	44.2	54.9	57.5	53.7	55.3	59.4	**65.1**	57.0	60.0	55.7	63.1
VisualNews 文本到图像	74.7	79.7	74.5	71.7	76.8	72.8	78.5	72.4	79.8	75.4	**80.7**
VisualNews 图像到文本	78.3	**83.6**	78.2	74.2	82.0	76.5	81.3	78.2	83.5	78.8	**83.6**
MSCOCO 文本到图像	68.1	71.2	75.3	75.1	78.3	75.2	77.9	76.1	77.8	77.3	**79.5**
MSCOCO 图像到文本	63.1	57.7	71.4	68.9	71.4	71.1	73.1	70.0	72.6	72.5	**75.3**
NIGHTS	67.0	67.6	68.6	67.2	68.1	**70.8**	69.8	67.9	68.6	69.1	69.3
WebQA	88.8	91.4	90.6	90.0	90.9	90.4	90.8	90.7	90.7	91.3	**91.5**
FashionIQ	32.9	**37.8**	19.5	17.1	23.4	26.3	29.7	19.8	26.0	21.5	27.4
Wiki-SS-NQ	73.9	**78.2**	66.9	62.0	72.5	64.2	70.5	68.9	76.3	72.2	77.1
OVEN	72.3	**75.1**	64.3	66.9	71.4	67.6	72.7	67.5	68.6	68.8	71.2
EDIS	91.8	**96.0**	84.1	88.0	92.0	87.0	93.9	88.9	91.6	90.3	93.7
MSCOCO	28.6	31.4	67.1	69.5	72.7	67.7	74.1	69.0	72.0	68.6	**74.2**
RefCOCO	55.9	60.9	87.1	83.3	91.4	91.4	**97.7**	88.9	91.7	89.3	93.5
RefCOCO-Matching	73.3	78.4	85.8	84.4	91.1	95.0	**96.3**	90.1	93.7	90.5	93.9
Visual7W-Pointing	64.1	66.5	69.2	71.5	84.2	82.5	**94.3**	80.5	87.9	78.0	86.5

视频任务。 表格 S5 报告了分类、问答、检索和时刻检索的 Hit@1。在这 18 个数据集上，HEME-2B 的平均分为 45.3，HEME-7B 的平均分为 49.7。

表 S5 . **MMEB-V2 视频结果（Hit@1）**。每行最高分以粗体标出 .

数据集	GME		VLM2Vec-V2	UME-R1		TTE_s		RIME		HEME	
	2B	7B	2B	2B	7B	2B	7B	2B	7B	2B	7B
平均值 – 视频（18项任务）	33.9	38.6	34.9	42.2	47.5	41.3	46.8	43.7	49.4	45.3	**49.7**
分类（5）	34.9	37.4	39.3	44.3	48.6	47.3	49.1	48.0	52.6	49.4	**53.2**
问答（5）	42.0	50.4	34.3	51.2	60.7	49.1	60.6	52.1	**62.0**	53.2	60.2
RET（5）	25.6	28.4	28.8	32.9	38.2	33.2	36.4	33.6	38.4	37.1	**40.7**
MRET（3）	32.4	38.2	38.5	39.7	39.3	32.1	37.2	39.2	**41.6**	39.2	41.5
K700	35.2	39.7	38.0	35.8	42.8	49.6	**55.0**	47.7	**55.0**	52.4	54.9
SSv2	29.9	30.6	42.8	44.1	50.4	50.4	44.9	48.5	55.1	49.6	**55.5**
HMDB51	43.4	47.9	40.9	54.4	58.3	52.5	51.7	56.7	**58.9**	54.6	54.3
UCF101	52.4	54.7	60.0	67.2	70.0	58.3	64.2	66.4	68.5	69.5	**70.3**
Breakfast	13.6	14.3	14.8	20.1	21.5	25.4	29.7	20.6	25.4	20.8	**30.9**
MVBench	37.5	46.6	33.7	49.9	58.2	48.5	59.5	49.1	**59.9**	46.9	53.5
Video-MME	34.3	39.2	30.7	41.7	47.3	45.8	**53.1**	41.6	49.6	42.4	48.4
NExT-QA	39.5	53.6	20.9	59.9	69.6	53.8	**70.1**	58.9	69.7	59.6	69.2
EgoSchema	40.8	46.8	34.0	45.4	52.4	36.4	**55.6**	40.4	**55.6**	41.0	53.4
ActivityNetQA	58.0	65.6	52.3	57.8	76.0	60.8	64.6	70.5	75.1	76.2	**76.5**
DiDeMo	22.0	26.4	30.4	32.4	40.0	33.5	34.9	33.8	38.4	37.5	**42.7**
MSR-VTT	27.3	31.8	28.3	34.3	38.9	34.8	37.6	36.4	41.5	39.9	**42.3**
MSVD	47.6	49.7	48.1	55.4	60.8	56.5	58.5	56.6	59.4	60.9	**63.3**
VATEX	23.0	24.9	26.5	29.9	32.6	25.6	31.0	28.1	32.7	32.1	**36.2**
YouCook2	7.9	9.1	10.6	12.7	18.5	15.8	**19.9**	13.3	**19.9**	15.1	19.0
QVHighlights	43.6	**59.5**	49.4	57.5	54.9	38.9	51.0	55.1	56.9	47.8	53.1
Charades-STA	14.9	14.0	20.2	20.4	21.9	19.5	18.9	19.4	21.6	22.4	**23.1**
MomentSeeker	34.8	37.4	40.8	41.2	41.1	37.7	41.5	43.2	46.2	47.4	**48.4**

视觉文档任务。 表S6 按数据集报告 NDCG@5。域外组包含 ViDoSeek 和 MMLongBench-Doc 分别在页面级和文档级的结果。

表 S6。**MMEB-V2 视觉文档结果（NDCG@5）**。粗体表示每行中的最高分 .

数据集	GME		VLM2Vec-V2	UME-R1		TTE_s		RIME		HEME	
	2B	7B	2B	2B	7B	2B	7B	2B	7B	2B	7B
平均值 – VisDoc（24 个任务）	72.7	75.2	65.4	63.9	67.1	68.8	76.4	71.4	75.6	77.3	**80.6**
ViDoRe-V1 (10)	86.1	**89.4**	75.5	72.4	75.7	77.5	84.1	76.4	80.9	84.9	88.1
ViDoRe-V2 (4)	54.0	55.6	44.9	46.2	50.5	53.2	**62.7**	51.4	55.6	55.3	60.9
VisRAG (6)	82.5	85.0	79.4	79.2	83.7	83.2	**91.9**	81.7	85.8	87.3	88.7
分布外 (4)	43.1	44.4	39.4	37.2	37.6	41.1	47.6	63.9	66.9	65.7	**69.3**
ViDoRe arxivqa	82.8	86.9	80.6	73.9	73.6	80.7	84.6	82.1	84.1	89.7	**91.4**
ViDoRe docvqa	53.1	57.5	44.9	37.9	41.1	44.5	46.0	45.4	46.9	57.5	**61.0**
ViDoRe infovqa	90.2	**91.6**	83.7	76.2	80.8	84.8	88.7	81.2	85.4	89.9	91.3
ViDoRe tabfquad	93.3	94.6	89.2	86.1	90.2	88.4	94.7	85.5	**95.3**	90.7	94.1
ViDoRe tatdqa	69.9	74.1	43.8	40.6	46.7	50.4	59.4	48.9	52.6	69.8	**75.4**
ViDoRe shiftproject	89.5	**96.8**	60.8	66.8	65.0	65.2	81.6	69.9	73.7	77.7	85.3
ViDoRe 人工智能	97.5	**99.6**	88.5	85.9	89.5	91.9	98.1	89.8	96.1	96.8	96.9
ViDoRe 能源	91.9	**95.3**	86.5	83.3	85.7	88.7	93.5	86.2	88.5	89.6	95.1
ViDoRe 政府报告	94.6	**98.8**	85.0	82.6	89.8	86.9	96.7	86.4	91.7	92.5	92.6
ViDoRe 医疗保健行业	98.7	**99.3**	92.2	90.8	94.3	92.8	97.9	88.3	94.9	94.3	98.4
ViDoRe ESG 报告人工标注版 v2	61.0	63.4	45.6	50.2	50.4	59.0	**69.4**	56.5	60.2	60.2	65.0
ViDoRe 生物医学讲座多语言版 v2	54.0	49.5	44.3	46.2	50.7	52.0	**60.8**	47.6	51.5	53.1	59.9
ViDoRe 经济学报告多语言版 v2	50.2	54.2	43.0	45.7	57.8	49.8	60.4	47.0	59.2	53.4	**62.0**
ViDoRe ESG 报告多语言版 v2	50.7	55.4	46.6	42.7	43.2	52.1	**60.3**	54.3	51.6	54.5	56.6
VisRAG ArxivQA	82.0	87.4	76.9	74.3	80.5	78.5	**94.5**	79.4	84.0	85.7	89.1
VisRAG ChartQA	79.9	81.9	84.4	86.0	85.0	84.4	**91.2**	83.3	85.4	89.8	87.5
VisRAG MP-DocVQA	84.4	89.2	71.8	75.6	83.4	79.2	90.1	82.3	87.4	88.1	**90.5**
VisRAG SlideVQA	93.4	94.5	91.5	87.1	91.5	92.3	**95.6**	91.0	94.1	94.5	**95.6**
VisRAG InfoVQA	91.4	93.5	85.7	84.4	89.2	87.2	93.0	82.4	91.8	90.9	**93.9**
VisRAG PlotQA	64.1	63.4	66.1	68.0	72.7	77.5	**86.9**	71.6	72.1	74.6	75.4
ViDoSeek-page	21.6	23.2	21.9	21.2	21.3	22.6	35.0	80.7	**85.6**	79.1	82.6
ViDoSeek-doc	83.6	83.9	80.2	75.9	75.3	82.0	**84.4**	79.7	80.9	82.6	83.9
MMLongBench-page	15.8	16.2	11.9	11.9	12.3	12.9	20.7	48.3	52.4	50.1	**55.9**
MMLongBench-doc	51.4	54.3	43.7	39.7	41.3	47.0	50.4	46.9	48.8	50.9	**54.7**

C. MMEB-V3 评估

基准测试与指标。 我们遵循 MMEB-V3 评估协议 [17]，并对各数据集的分数取平均值。表S7列出了用于评估的 47 个智能体任务和 53 个文本任务。对于每项任务，每个查询都会对完整的候选语料库进行排序。输入长度超过 Qwen2-VL 的 32,768 词元上下文长度时，会从文本末尾截断，同时保留读出标记。我们报告智能体任务的 Hit@1 和文本任务的 NDCG@5。这些基准数据集不包含在我们的 HEME 微调混合数据集中。由于 Qwen2-VL 不接受音频输入，我们省略了 11 项音频任务。

表 S7。**MMEB-V3 智能体和文本任务。** 括号内给出了每个来源的任务数量。

类别	子类别	来源	#
智能体	工具检索	Tool-REX（35）	35
	图形用户界面检索	GAE-Bench：GUIAct（4）、Mind2Web（4）	8
	记忆检索	KnowMeBench、REALTALK、PeerQA、DeepPlanning	4
Text	推理密集型检索 (RR)	BRIGHT (12)、R2MED (8)	20
	指令遵循检索 (IF)	FollowIR (3)、InfoSearch (6)	9
	长上下文检索 (LC)	LongEmbed (6)	6
	多条件检索 (MC)	MultiConIR（5）	5
	通用检索（GR）	NanoBEIR（13）	13
总计			100

子类别结果。在表S8中的 100 项任务中，HEME-2B 的平均值为 35.6，HEME-7B 的平均值为 38.7。基线得分及其子类别平均值来自基准测试报告 [17]。对于该报告中在音频上进行评估的模型，我们仅使用智能体和文本任务重新计算总体平均值。

表 S8。**MMEB-V3 按子类别划分的结果。**智能体得分采用 Hit@1，文本得分采用 NDCG@5。最后一列为全部 100 项任务的平均值。†模型使用了额外的纯文本训练数据。

模型	骨干网络	智能体				Text						All
		Tool	图形用户界面记忆		Avg	RR	IF	LC	MC	GR	Avg	
# 任务		35	8	4	47	20	9	6	5	13	53	100
VLM2Vec-V2 [36]	Qwen2-VL-2B	27.6	36.2	23.3	28.7	7.8	29.2	11.5	50.3	41.2	24.5	26.5
E5-Omni† [2]	Qwen2.5-Omni-3B	37.7	35.6	32.3	36.9	11.5	44.4	8.3	60.7	34.4	26.7	31.5
Omni-Embed-Nemotron† [60]	Qwen2.5-Omni-3B	38.1	32.0	32.2	36.5	17.2	42.8	50.0	69.7	56.9	39.2	37.9
Qwen3-VL-Embed† [28]	Qwen3-VL-2B	42.6	30.4	28.4	39.3	16.6	40.6	53.1	61.2	58.0	39.2	39.2
HEME（我们的方案）	Qwen2-VL-2B	37.6	39.5	24.8	36.9	11.7	39.0	48.8	59.0	50.1	34.4	35.6
VLM2Vec-Qwen2VL [22]	Qwen2-VL-7B	19.8	21.4	15.9	19.7	7.2	28.1	5.9	40.9	41.7	22.2	21.0
LCO-Embed-Omni† [59]	Qwen2.5-Omni-7B	29.0	25.0	23.0	27.8	13.6	52.0	12.6	61.4	52.1	32.4	30.2
E5-Omni† [2]	Qwen2.5-Omni-70亿	37.4	38.0	27.3	36.7	11.1	42.1	9.7	56.9	35.9	26.9	31.5
GME† [70]	Qwen2-VL-70亿	39.0	30.0	17.1	35.6	12.5	52.4	17.8	59.0	62.5	37.1	36.4
Qwen3-VL-Embed† [28]	Qwen3-VL-8B	41.3	33.5	22.8	38.4	18.2	44.8	58.0	61.2	61.2	42.5	40.6
HEME（我们的模型）	Qwen2-VL-7B	40.4	43.1	27.9	39.8	15.7	42.7	51.5	57.6	53.9	37.7	38.7

C.1. 按任务划分的结果

智能体任务。 桌子S9 列出了各项工具、图形用户界面和记忆检索任务。HEME-2B 和 HEME-7B 在这些任务上的 Hit@1 平均值分别为 36.9 和 39.8。

桌子 S9。 **MMEB-V3 智能体检索结果（Hit@1）。** 第一行同时包含智能体和文本任务。粗体表示每行中的最高分 .

Task	VLM2Vec -V2	VLM2Vec -Qwen2VL	E5-Omni		全能-嵌入 -Nemotron	LCO-嵌入 -全能	GME	Qwen3-VL -嵌入		HEME	
	2B	7B	3B	7B	3B	7B	7B	2B	8B	2B	7B
平均值 – 全部（100 项任务）	26.5	21.0	31.5	31.5	37.9	30.2	36.4	39.2	**40.6**	35.6	38.7
平均值 – 智能体（47 项任务）	28.7	19.7	36.9	36.7	36.5	27.8	35.6	39.3	38.4	36.9	**39.8**
工具（35）	27.6	19.8	37.7	37.4	38.1	29.0	39.0	**42.6**	41.3	37.6	40.4
图形用户界面（8）	36.2	21.4	35.6	38.0	32.0	25.0	30.0	30.4	33.5	39.5	**43.1**
记忆 (4)	23.3	15.9	**32.3**	27.3	32.2	23.0	17.1	28.4	22.8	24.8	27.9
工具-craft 数学代数	65.0	57.1	75.7	81.4	**85.0**	63.9	84.6	84.6	83.9	74.3	80.0
工具-craft-tabmwp	14.4	10.9	19.0	29.9	29.3	25.3	27.6	**36.8**	34.5	19.0	23.0
工具-craft-vqa	43.5	23.0	47.0	50.5	**65.5**	34.5	60.5	52.0	54.5	46.0	51.0
工具-gorilla-huggingface	20.2	18.6	25.8	19.2	26.6	20.2	24.6	26.4	**28.4**	25.0	27.4
Tool-gorilla-pytorch	4.7	0.0	**20.9**	14.0	11.6	16.3	14.0	11.6	9.3	9.3	14.0
Tool-gorilla-tensor	7.3	1.8	16.4	12.7	20.0	5.5	**21.8**	16.4	16.4	0.0	3.6
Tool-toolink	40.4	36.4	68.4	**70.6**	67.0	65.6	66.4	67.8	67.6	58.1	55.9
Tool-apibank	38.6	1.0	53.5	47.5	45.5	7.9	52.5	**62.4**	58.4	50.5	50.5
Tool-apigen	46.5	49.3	59.7	61.2	**64.0**	50.2	54.0	55.4	55.5	55.0	51.8
Tool-mnms	30.3	27.3	**54.5**	51.5	42.4	27.3	30.3	36.4	30.3	33.3	**54.5**
Tool-reversechain	51.5	22.5	**64.0**	56.5	50.0	18.5	54.5	61.0	57.5	56.5	59.5
Tool-rotbench	1.8	1.1	5.1	4.2	6.9	2.5	**7.3**	3.3	4.2	3.1	5.5
Tool-t-eval-dialog	22.0	26.0	16.0	24.0	14.0	20.0	**34.0**	24.0	20.0	18.0	30.0
Tool-t-eval-step	10.0	**38.0**	26.0	12.0	12.0	22.0	30.0	16.0	12.0	24.0	26.0
Tool-taskbench-daily	52.5	45.0	**90.0**	75.0	67.5	62.5	72.5	65.0	62.5	52.5	85.0
Tool-toolace	56.4	33.3	67.1	**71.4**	67.0	51.7	69.0	65.7	62.3	60.9	67.0
Tool-toolbench	24.8	6.0	29.2	28.9	13.3	16.9	17.4	**40.6**	36.0	28.6	35.5
Tool-toolemu	10.5	0.0	10.5	5.3	**34.2**	7.9	13.2	31.6	26.3	31.6	26.3
Tool-tooleyes	15.8	4.2	**24.2**	12.6	10.5	11.6	20.0	19.0	20.0	18.9	23.2
Tool-toollens	8.9	1.3	5.1	10.2	1.0	2.9	1.3	8.0	7.0	9.9	**12.1**
Tool-ultratool	4.6	0.4	44.0	43.8	43.6	41.2	40.4	**50.6**	41.2	30.2	42.2
Tool-autotools-food	4.6	9.1	9.1	18.2	13.6	4.5	18.2	50.0	50.0	**54.5**	40.9
Tool-autotools-music	0.0	0.0	0.0	0.0	3.1	0.0	0.0	12.5	37.5	**59.4**	28.1
Tool-autotools-weather	0.0	0.0	0.0	9.1	0.0	0.0	0.0	18.2	9.1	**45.5**	9.1
Tool-restgpt-spotify	12.5	27.5	20.0	30.0	25.0	0.0	20.0	60.0	**67.5**	7.5	20.0
Tool-restgpt-tmdb	0.0	0.0	9.3	13.0	16.7	9.3	11.1	37.0	**38.9**	13.0	11.1
Tool-appbench	65.6	65.6	87.5	84.4	87.5	59.4	81.2	71.9	84.4	**90.6**	**90.6**
Tool-gpt4tools	46.9	28.1	37.5	50.0	75.0	81.2	81.2	**84.4**	68.8	50.0	71.9
Tool-gta	0.0	14.3	28.6	14.3	**42.9**	28.6	35.7	14.3	21.4	14.3	21.4
Tool-taskbench-huggingface	56.5	4.3	47.8	56.5	52.2	52.2	56.5	**60.9**	39.1	47.8	47.8
Tool-taskbench-多媒体	65.0	32.5	**92.5**	82.5	77.5	60.0	87.5	70.0	62.5	72.5	80.0
Tool-metatool	44.0	32.0	46.0	51.0	49.0	48.5	**54.0**	52.5	51.5	52.5	51.0
Tool-tool-be-honest	46.0	47.1	54.0	52.0	54.3	35.4	**55.4**	51.1	49.1	43.7	50.3
Tool-toolalpaca	51.1	26.6	58.5	60.6	50.0	60.6	59.6	62.8	**67.0**	55.3	60.6
Tool-toolbench-sam	4.1	1.0	5.6	6.6	8.1	2.5	7.6	**11.7**	**11.7**	5.6	6.1
GAE-guiact-q2s	25.9	7.7	29.9	32.3	26.1	24.1	31.3	19.8	29.5	31.8	**37.2**
GAE-guiact-q2t	52.6	42.6	50.9	55.5	39.8	42.8	31.3	48.4	46.3	52.6	**57.4**
GAE-guiact-s2s	35.8	33.2	36.5	36.4	35.9	17.9	32.7	25.5	30.2	**44.6**	43.1
GAE-guiact-t2s	24.1	8.6	21.9	23.0	20.7	13.8	17.1	25.0	24.9	33.0	**33.6**
GAE-mind2web-q2s	43.0	12.2	43.0	44.9	41.7	33.0	42.1	27.1	37.7	44.5	**54.5**
GAE-mind2web-q2t	47.8	32.4	40.3	50.2	33.6	25.3	32.4	44.3	37.2	46.6	**51.4**
GAE-mind2web-s2s	37.1	32.4	38.3	37.7	37.1	25.9	37.1	30.2	35.2	**38.6**	**38.6**
GAE-mind2web-t2s	23.4	1.9	23.7	23.7	21.5	17.1	15.9	23.1	27.4	24.0	**28.7**
KnowMeBench	32.6	22.3	**45.1**	38.4	44.4	35.8	19.0	41.5	35.3	35.6	40.4
REALTALK	18.4	11.8	27.8	27.5	32.0	28.1	19.6	**35.4**	30.5	26.7	29.6
PeerQA	21.3	11.0	**22.8**	19.9	18.4	15.4	14.7	17.7	16.2	16.9	19.9
DeepPlanning	20.8	18.3	33.3	23.3	**34.2**	12.5	15.0	19.2	9.2	20.0	21.7

文本任务。 表格 S10 报告了 53 个文本检索任务的 NDCG@5。HEME-2B 和 HEME-7B 的平均值分别为 34.4 和 37.7。推理密集型组包括 12 个 BRIGHT 任务和 8 个 R2MED 任务。

表 S10。**MMEB-V3 文本结果（NDCG@5）**。任务按检索子类别分组。粗体表示每行中的最高分。

Task	VLM2Vec -V2	VLM2Vec -Qwen2VL	E5-Omni		全能-嵌入 -Nemotron	LCO-Embed -全能	GME	Qwen3-VL -嵌入		HEME	
	2B	7B	3B	7B	3B	7B	7B	2B	8B	2B	7B
平均值 – 文本（53 项任务）	24.5	22.2	26.7	26.9	39.2	32.4	37.1	39.2	**42.5**	34.4	37.7
推理密集型检索 (20)	7.8	7.2	11.5	11.1	17.2	13.6	12.5	16.6	**18.2**	11.7	15.7
指令遵循检索 (9)	29.2	28.1	44.4	42.1	42.8	52.0	**52.4**	40.6	44.8	39.0	42.7
长上下文检索 (6)	11.5	5.9	8.3	9.7	50.0	12.6	17.8	53.1	**58.0**	48.8	51.5
多条件检索 (5)	50.3	40.9	60.7	56.9	**69.7**	61.4	59.0	61.2	61.2	59.0	57.6
通用检索 (13)	41.2	41.7	34.4	35.9	56.9	52.1	**62.5**	58.0	61.2	50.1	53.9
BRIGHT 奥赛数学	1.2	2.2	3.7	3.9	3.3	**8.2**	5.0	2.4	1.1	2.5	4.0
BRIGHT 生物学	12.3	9.4	2.0	1.6	13.6	11.1	3.0	14.7	**17.1**	7.7	11.8
BRIGHT 地球科学	11.4	0.7	7.7	6.1	26.4	13.6	12.8	26.2	**28.9**	14.5	15.6
BRIGHT 经济学	3.9	2.8	4.5	3.8	15.4	10.3	7.0	15.9	**16.7**	8.7	11.0
BRIGHT 力扣	8.5	4.7	17.9	**20.3**	18.5	5.6	17.3	13.7	17.3	13.4	15.6
BRIGHT 小马	1.5	**9.5**	2.4	1.3	2.6	0.5	2.1	2.5	9.1	3.4	4.2
BRIGHT 心理学	4.4	3.0	5.6	4.2	18.1	8.5	11.0	**18.6**	12.0	9.0	15.5
BRIGHT 机器人学	4.8	1.5	7.9	7.0	12.1	2.8	5.7	11.3	**12.5**	7.9	6.6
BRIGHT Stack Overflow	2.1	0.0	6.8	6.0	9.6	1.6	6.1	**14.5**	12.0	4.9	7.5
BRIGHT 可持续生活	2.6	3.6	4.7	4.9	6.6	10.9	9.7	9.7	**14.2**	6.2	9.0
BRIGHT TheoremQA 问题	9.4	15.1	12.3	**16.9**	13.9	9.9	12.8	10.6	12.8	11.9	14.2
BRIGHT TheoremQA 定理	4.7	3.2	11.2	11.6	**23.0**	8.6	17.0	12.8	21.2	5.9	12.0
R2MED 生物信息学	19.4	15.1	17.5	33.7	36.9	14.0	33.2	35.8	**42.2**	23.0	34.6
R2MED 生物学	12.3	9.3	2.0	1.6	13.9	11.0	3.2	14.7	**17.2**	7.6	11.9
R2MED IIYi-临床	8.5	2.3	15.5	21.9	15.7	4.2	18.5	21.3	**26.3**	16.2	18.3
R2MED MedQA-Diag	21.8	16.4	8.4	6.8	6.9	6.7	28.9	33.6	**34.4**	7.6	12.6
R2MED MedXpertQA-Exam	5.6	6.0	5.3	6.9	5.3	6.5	5.3	11.5	**17.4**	4.9	10.9
R2MED 医学科学	4.7	3.5	25.8	20.0	27.3	**33.1**	9.3	10.1	13.1	27.0	29.2
R2MED PMC-临床	14.0	13.5	28.8	25.7	21.7	8.7	32.2	21.4	**37.1**	22.4	31.3
R2MED PMC-Treatment	28.3	21.4	27.6	34.9	14.2	13.9	**42.0**	30.4	37.4	30.5	39.0
FollowIR core17	21.3	21.0	28.6	26.9	30.1	**45.3**	34.8	34.5	34.3	41.1	44.3
FollowIR news21	4.6	9.7	25.3	24.9	31.6	26.7	13.3	21.8	20.4	40.9	**48.1**
FollowIR robust04	24.7	21.1	32.1	27.4	33.5	40.2	30.6	39.5	**44.0**	37.6	39.6
InfoSearch Audience-v1	17.2	25.0	29.1	34.2	27.4	48.6	**61.8**	31.3	34.5	17.5	20.4
InfoSearch Clarity-v1	43.6	34.5	62.3	43.8	34.5	50.5	57.9	**62.8**	55.4	36.7	28.5
InfoSearch Format-v1	20.4	10.2	49.9	49.2	**52.5**	47.3	47.5	38.2	46.2	29.7	42.0
InfoSearch Language-v1	49.2	45.3	67.9	66.5	71.3	68.9	**77.5**	56.5	67.6	57.2	67.1
InfoSearch Length-v1	43.9	40.0	56.4	58.6	56.9	73.3	**75.4**	44.2	55.9	44.9	47.2
InfoSearch Source-v1	37.6	46.1	47.8	46.9	47.7	67.0	**72.5**	36.8	44.5	45.7	47.2
LongEmbed 2wikimqa	13.7	6.0	1.4	1.2	71.1	2.5	14.7	83.8	**89.0**	54.0	50.7
LongEmbed narrativeqa	3.6	1.3	1.2	1.1	32.7	1.8	6.8	52.9	**60.9**	45.5	55.6
LongEmbed needle	7.6	3.2	14.9	21.7	24.3	17.1	21.2	15.5	21.2	25.9	**27.3**
LongEmbed 口令	11.3	12.7	27.7	31.4	35.9	31.8	26.7	32.8	**38.2**	31.0	34.1
LongEmbed QMSum	11.9	4.7	1.5	1.5	40.2	8.4	9.8	37.5	42.3	44.0	**46.2**
LongEmbed SummScreen FD	20.9	7.5	3.1	1.3	96.0	13.8	27.4	95.9	**96.3**	92.4	95.0
MultiConIR 书籍	47.1	35.0	52.1	52.5	**64.0**	56.1	56.8	62.1	60.9	55.8	55.4
MultiConIR 法律文档	52.4	38.0	57.7	54.7	**68.8**	59.8	56.7	57.2	57.7	56.7	60.8
MultiConIR 医学案例	43.9	35.8	57.7	51.6	**65.7**	59.4	52.8	60.0	61.4	55.0	52.7
MultiConIR 电影	48.7	42.0	66.3	61.9	**74.7**	65.2	64.6	64.9	62.9	66.5	60.7
MultiConIR 人物	59.2	53.4	69.8	64.0	**75.1**	66.2	64.1	61.8	62.9	60.8	58.2
NanoBEIR NanoArguAna	36.1	32.4	53.3	49.9	42.8	25.7	**54.3**	49.4	45.0	41.9	49.7
NanoBEIR NanoClimateFEVER	14.4	14.2	2.2	1.3	23.6	25.0	31.9	29.5	**33.5**	21.6	25.0
NanoBEIR NanoDBPedia	48.4	51.7	38.2	36.2	53.8	**65.8**	60.5	60.6	61.1	55.0	59.2
NanoBEIR NanoFEVER	52.3	55.9	18.6	19.0	75.4	80.0	87.1	89.2	**89.8**	77.4	76.4
NanoBEIR NanoFiQA2018	21.0	20.1	15.4	24.0	54.4	43.2	**58.8**	43.7	53.2	37.4	39.3
NanoBEIR NanoHotpotQA	50.0	53.6	73.0	56.9	**92.0**	62.2	84.5	73.8	77.4	62.3	63.8
NanoBEIR NanoMSMARCO	38.6	28.1	47.0	44.4	53.4	45.7	**65.5**	53.6	64.9	46.1	48.0
NanoBEIR NanoNFCorpus	29.5	31.8	10.8	13.2	31.8	34.8	**39.9**	39.5	**39.9**	28.2	34.5
NanoBEIR NanoNQ	28.2	34.5	28.8	23.9	64.5	58.5	65.9	56.0	**66.6**	53.2	60.1
NanoBEIR NanoQuoraRetrieval	92.8	93.4	92.9	87.5	91.5	90.5	95.2	94.9	96.0	92.4	**96.1**
NanoBEIR NanoSCIDOCS	24.3	30.1	7.2	33.1	32.7	33.0	**44.1**	35.7	37.5	29.5	34.5
NanoBEIR NanoSciFact	58.4	61.7	33.5	43.4	73.8	66.9	73.4	77.5	**78.5**	63.6	68.5
NanoBEIR NanoTouche2020	41.6	34.2	25.7	33.6	50.3	46.3	**52.0**	50.1	**52.0**	42.8	45.4

D. 其他实现细节

事后教师、学生和 InfoNCE 基线采用相同的训练混合数据、采样、预处理、优化以及查询/候选项格式化。

训练数据和预处理。 表 S11 列出了 MMEB-train、 [22], ViDoRe [10], VisRAG [65],和 LLaVA-Hound [69]的采样权重。每一步，我们以概率 s 抽取一个来源 $w_s / \sum_{s'} w_{s'}$，并根据该来源构建全局批次。我们将四个分类数据集汇集起来，以减少一个批次内重复类别标签导致的假负例。图像保留其原始宽高比，每张图像使用 4 至 1,280 个视觉词元（28×28 每个词元对应的像素数）。对于每个视频，我们均匀采样 8 帧，总像素预算为 1,228,800。

表 S11。**训练来源和采样权重。** †汇集的分类来源包括 N24News、HatefulMemes、VOC2007 和 SUN397。

模态	来源	权重	模态	来源	权重
图像	ImageNet-1K	1.0	图像	VisDial	1.0
图像	分类†	1.0	图像	CIRR	0.5
图像	OK-VQA	0.5	图像	VisualNews t2i	1.0
图像	A-OKVQA	0.5	图像	VisualNews i2t	1.0
图像	DocVQA	1.0	图像	MSCOCO 文本到图像	1.0
图像	InfographicsVQA	0.5	图像	MSCOCO 图像到文本	1.0
图像	ChartQA	0.5	图像	NIGHTS	0.5
图像	Visual7W	1.0	图像	WebQA	0.5
图像	MSCOCO 定位	1.0			
VisDoc	ViDoRe	10.0	VisDoc	VisRAG	12.0
视频	LLaVA-Hound，视频→图像描述	5.0	视频	LLaVA-Hound，图像描述→视频	5.0
视频	LLaVA-Hound，视频问答	5.0			

优化。 我们从同一个预训练 Qwen2-VL 主干网络初始化教师和学生 [51]，并分别以全局批大小 512 训练一轮。我们冻结视觉编码器，并更新多模态投影器和语言解码器。我们使用 AdamW，权重衰减为 0.01；对于 20亿和 70亿模型，学习率分别为 10^{-5} 和 5×10^{-6}。学习率在线性预热 100 步后采用余弦调度。对于 HEME-2B，教师和学生阶段分别约需 110 和 90 个 H200 GPU 小时；对于 HEME-7B，分别约需 310 和 190 个 GPU 小时。

输入格式化与读出。 我们使用 Qwen2-VL 聊天模板将每个输入格式化为单轮用户对话，并省略系统提示词。最后的换行符充当 [嵌入] 位置。我们将其输入嵌入替换为一个从换行符嵌入初始化的可学习向量，并对其最后一层隐藏状态进行 ℓ_2归一化。在事后传球中，注意力掩码会阻止 [嵌入]以及查询内容之后的模板标记关注候选前缀。

事后信息实现。 我们为拼接序列 $[C_i; Q_i]$计算 M-RoPE [51] 位置索引。缓存的候选键保留其前缀位置，而查询位置则从前缀末尾继续。对于 $\mathcal{L}_{\text{align}}$，我们将每个查询内容词元与另一遍中的相同词元配对，并通过两遍传播对齐梯度。