arXiv:2512.15885v1 [cs.CV] 17 Dec 2025

超越文字：多模态大语言模型的自监督视觉学习

Davide Caffagni[1] Sara Sarto[1] MarcellaCornia[1] Lorenzo Baraldi[1]Pier Luigi Dovesi[2]
Shaghayegh Roohi[2] MarkGranroth-Wilding[2] Rita Cucchiara[11]摩德纳-雷焦艾米利亚大学 [2]
AMDSilo AI[1] {name.surname}@unimore.it [2] {name.surname}@amd.com

摘要

多模态大语言模型（*MLLM*）近年来在连接视觉与语言方面展现出令人瞩目的能力，但其在基础视觉推理任务上的表现仍然有限。这一局限可归因于 *MLLM* 主要通过文本描述学习视觉理解，而文本描述作为监督信号，具有主观性且本质上并不完整。此外，与大规模纯文本预训练相比，多模态指令微调的规模较小，导致 *MLLM* 过度拟合语言先验，同时忽略视觉细节。为解决这些问题，我们提出 *JARVIS*，这是一种受 *JEPA* 启发、用于提升 *MLLM* 自监督视觉能力的框架。具体而言，我们将 *I-JEPA* 学习范式整合到 *MLLM* 训练的标准视觉-语言对齐流程中。我们的方法利用冻结的视觉基础模型作为上下文编码器和目标编码器，同时训练预测器（由大语言模型的早期层实现），使其无需完全依赖语言监督，即可从图像中学习结构和语义规律。在标准多模态大语言模型基准测试上进行的大量实验表明，*JARVIS* 能够持续提升不同大语言模型系列在以视觉为中心的基准测试上的表现，且不会降低多模态推理能力。我们的源代码已公开，地址为：*https://github.com/aimagelab/JARVIS*。

1. 引言

大型语言模型（LLM）[10,54, 58] 的迅速成功表明，这类模型越来越需要处理文本以外的各种模态并进行推理。这一需求推动了多模态大语言模型（MLLM）[11], 的兴起，它们将不同的输入模态转换到大语言模型的同一嵌入空间中，从而使其能够理解 [3, 30, 64], 甚至生成 [51], 其他模态，尤其是

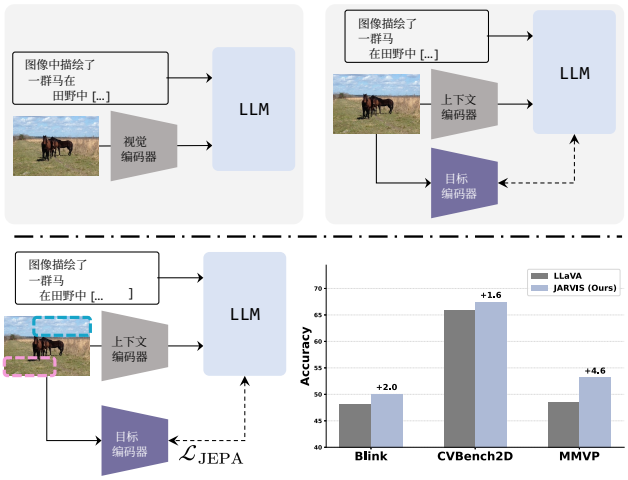


图1. LLaVA（左上）、将选定层的输出与目标编码器的输出对齐的基线方法（右上），以及通过掩码预测损失对齐输出的 JARVIS（左下）的比较。我们还报告了 JARVIS 和 LLaVA 在三个视觉基准测试上的结果（右下）。

图像。尽管取得了令人瞩目的进展，自视觉指令微调首次提出以来，设计多模态大语言模型的基本方案一直没有改变。视觉指令微调最初由 LLaVA 提出。LLaVA 展示了轻量级投影器如何通过将视觉编码器生成的图像表征与大语言模型的文本嵌入空间对齐，来连接视觉与文本模态。经过这种对齐后，投影后的视觉特征便能被大语言模型有效解读，使其能够对图像进行推理，并生成以视觉内容为条件的文本。

尽管该流程已被证明在广泛的任务中极为有效，但当前的多模态大语言模型在一些出人意料地简单的视觉推理场景中仍表现出明显的局限性，例如确认物体是否存在、对其进行计数、理解它们之间的空间关系，或估计它们的相对距离 [18, 56, 57, 63]。当前多模态大语言模型在这些视觉任务中的低水平表现，凸显了其视觉感知方面的严重不足。我们认为，

这一缺陷源于多模态大语言模型仅通过文本描述来学习观察图像。在 LLaVA 提出的对齐阶段，多模态大语言模型会接收到一幅图像，其学习目标是生成该图像的标题。直观地说，如果多模态大语言模型能够描述一幅图像，那么它就应该已经看见 了它。然而，图像标题本质上具有主观性 [13, 48]：它们反映了标注人员认为 相关的内容，往往会省略从其他角度来看可能至关重要的细节。此外，在实践中不可能假设能够获取一幅图像的所有可能描述。因此，图像本身包含的信息天然比其任意子集的文本描述更加丰富、全面。与此同时，与在海量文本语料库上进行的无监督预训练相比，多模态训练的规模相对有限，因此多模态大语言模型在对图像进行推理时往往过度依赖语言先验，从而忽略视觉细节 [8, 16, 61, 67]。

有鉴于此，在这项工作中，我们主张利用图像中固有的自监督信号来训练多模态大语言模型，使其能够发现并理解视觉输入，而不局限于文本描述。具体而言，基于联合嵌入预测架构（JEPAs）在自监督学习方面取得的可喜成果 [27]，，我们提出将 I-JEPA [4]，——一种在无标注图像上训练的 JEPA 模型——整合到经典的视觉指令微调流程中。

具体来说，给定一张图像，我们首先对预训练的上下文视觉编码器提取的嵌入中的连续块进行掩码。随后，将剩余嵌入传递给大语言模型，并像 LLaVA 一样训练该模型根据被掩码的图像生成图像字幕。与此同时，我们还要求大语言模型根据可见嵌入恢复缺失的嵌入。重要的是，这些目标嵌入的监督信号由额外的目标视觉编码器提供，该编码器只能访问图像，因此不会受到图像字幕中的语言条件影响。这种学习范式以预测编码理论为依据 [46]，，该理论认为，信号中相互兼容部分的表征（即可见上下文与被掩码目标嵌入）应当能够相互预测。这一设计使大语言模型能够学习图像中已有的模式和结构，而这些模式和结构往往无法仅凭文本描述捕捉到。我们将此方法命名为 JARVIS，其概览如图1所示。

总而言之，本工作强调了将自监督视觉学习融入多模态大语言模型的重要性及其益处。在使用相同训练数据和架构的情况下，我们的方法 JARVIS 在标准多模态大语言模型基准测试上优于原始的 LLaVA 训练范式，其中在 Cambrian评测套件的视觉中心任务上提升最为显著 [56]。我们的实验表明，这些改进在不同的大语言模型系列中均保持一致，并且不会牺牲通用认知性能。

2. 相关工作

多模态大语言模型（多模态大语言模型）。 多模态大语言模型（多模态大语言模型） [3, 30, 51, 64]已经成为大语言模型向文本以外其他模态自然扩展的产物。根据所关注的额外模态，大多数多模态大语言模型都配备专用的单模态编码器和适配器，其作用是将编码器生成的单模态嵌入转换到大语言模型嵌入空间中，而后者充当推理骨干网络 [11]。就图像感知而言，包括专有的、最先进的多模态大语言模型在内的大多数模型 [2, 6, 52]，，都可以追溯到 LLaVA 提出的两阶段训练方案 [36, 37]，，该方案以现成的大语言模型为基础。在第一训练阶段，大语言模型保持冻结，仅训练图像到文本适配器（例如，线性投影器或多层感知机投影器），利用图像描述数据，最大化给定图像时大语言模型生成描述的概率。随后，视觉指令微调阶段通过让解冻后的大语言模型在存在图像的情况下遵循用户指令来完成训练。在本工作中，我们遵循 LLaVA 框架，因为它提供了一种可轻松应用于多个大语言模型的训练方案，并且完全基于开源数据。然而，我们对第一训练阶段进行了干预，加入了一种不涉及语言监督的学习信号。

面向多模态大语言模型的视觉中心策略。 增强多模态大语言模型的视觉感知一直是一项长期挑战。一些方法聚焦于视觉编码器，设计图像到文本投影器，在输入大语言模型之前汇集多层激活 [12, 33]。另一类研究则将为大多数多模态大语言模型提供支持的 CLIP视觉编码器，与不同的视觉专家模型相结合，以提升其在视觉领域的表现 [25, 35, 56, 57]。关于语言骨干网络，针对大语言模型如何处理视觉词元的研究发现，参与处理视觉词元的注意力头数量有限 [8]，，而且可以利用这些注意力头来提升定位等视觉任务的表现 [24]。

以图像重建形式进行的自监督视觉学习也得到了探索：训练期间仅将外部视觉编码器用作教师模型，并在推理阶段将其弃用。ROSS [60]将多模态大语言模型与去噪网络一同训练，以恢复由视觉分词器生成的视觉嵌入。与我们工作更为接近的是 VIRAL [65]，，它将大语言模型的中间激活与外部视觉基础模型产生的中间激活对齐。

与相关工作相比，据我们所知，我们首次将类 I-JEPA监督 [4]作为额外的学习目标，纳入多模态大语言模型的训练方案中。具体而言，除了对齐中间激活之外，我们还要求多模态大语言模型恢复图像的掩码区域的表征。值得注意的是，我们在LLaVA预训练阶段便应用了这一方法，此时大语言模型仍在学习如何处理和理解视觉词元。这

样一来，我们影响了大语言模型感知图像的方式，超越了纯粹的语言监督。

联合嵌入预测架构。 JEPA [27]是一种新兴的自监督学习框架，其基础原理是：相容输入信号的潜在表示应当能够相互预测 [46]。根据应用领域的不同 [1, 21, 28, 47, 55], , JEPA 模型识别这些相容输入对的方式也各不相同。例如，I-JEPA [4]训练上下文编码器和目标编码器，使得预测器网络能够根据同一图像中相邻部分的上下文表示，推断出图像裁剪块的目标表征。随后，V-JEPA [5, 7] 将这一理念拓展到视频领域，利用由掩码图像块构成的时空块来为预测器提供条件，使其根据目标编码器输出这些图像块未经掩码的表示。

通常，JEPA 会联合学习上下文网络和目标网络。然而，近期研究表明，即使使用冻结的目标编码器，JEPA 也能高效地学习到强大的表示 [31]。在这项工作中，我们将 I-JEPA 用作自监督目标函数，从图像中学习，而不局限于其文本描述。大多数基于 JEPA 的方法专注于学习编码器网络，并在推理时舍弃预测器网络；与之不同的是，我们将视觉基础模型 [43, 44]用作冻结的上下文编码器和目标编码器，并让大语言模型充当我们这个 JEPA 增强模型中的预测器。

3. 提出的方法

3.1. 背景

在这项工作中，我们旨在提升多模态大语言模型的视觉感知能力。除预训练大语言模型 $\mathcal{G}$ 外，为一般性起见，多模态大语言模型的架构还包括：(i) 用于处理图像的预训练视觉编码器 $\mathcal{F}_v$；以及 (ii) 可训练的投影器 proj，用于将 $\mathcal{F}_v$ 的输出嵌入空间与 $\mathcal{G}$ 的输入嵌入空间对齐。形式上，给定图像 I 会被转换为由 N 个 d 维视觉嵌入构成的序列，随后作为输入词元提供给 $\mathcal{G}$，如下所示：

$$\begin{aligned} I &= \text{proj}(\mathcal{F}_v(\text{I})) \\ &= \{\mathbf{v}_i \in \text{R}^d\}_{i=1\ldots\text{N}}\}. \end{aligned}$$

遵循 LLaVA [36]，推广的方法，我们使用下一词元预测（NTP）目标函数训练 $\mathcal{G}$，具体而言，通过最小化在图像 I 条件下生成文本 $\mathbf{x}$ 的负对数似然：

$$\mathcal{L}_{\text{NTP}} = -\sum_i \log P\left(\mathbf{x}_i | I, \mathbf{x}_{1,\ldots,i-1}; \mathcal{G}, \text{proj}\right). \quad (1)$$

训练过程分为两个阶段。在第一阶段（称为 对齐）中，$\mathbf{x}$ 表示
I 的描述，且只有 proj 经过训练，以对齐视觉和

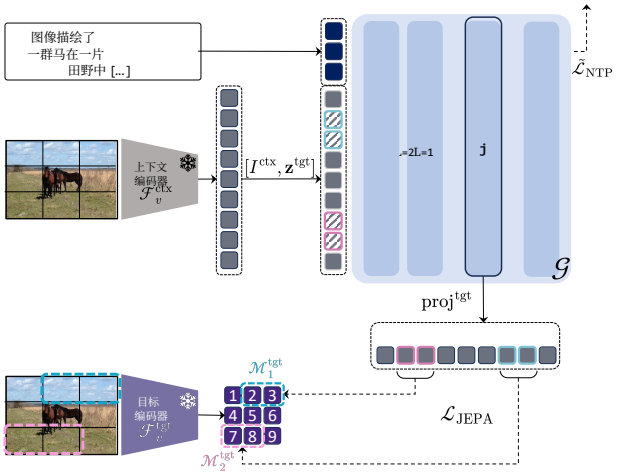


图 2。我们的 **JARVIS** 方法概览。我们利用单个上下文块，通过掩码预测损失预测多个目标块的表示，使大语言模型的预测嵌入与目标编码器的输出对齐。

文本表示。在第二阶段，即视觉指令微调中，$\mathbf{x}$ 表示围绕图像 I 展开的多轮视觉对话，在此过程中，语言模型 $\mathcal{G}$ 和投影器 proj 都经过训练。

3.2. 将 JEPA 引入多模态大语言模型

通过在对齐阶段最小化公式1，多模态大语言模型学会用自然语言描述图像。直观地说，在这项任务中取得成功意味着模型已经具备观察 并理解图像的能力。然而，仅依靠文本监督本质上存在局限，因为图像说明无法完整捕捉嵌入图像中的丰富视觉信息。为了解决这一局限，我们提出引入一种纯视觉的自监督学习目标函数。具体而言，受 JEPA [27]模型近期进展的启发，我们将 I-JEPA [4] 公式引入 LLaVA [36, 37]的对齐阶段。

重新审视 I-JEPA 方法。 I-JEPA 的目标是学习一种视觉嵌入函数，使得可以根据图像可见部分的表征，预测图像缺失部分的潜在表征，即目标；而图像可见部分的表征称为即上下文。给定一幅图像 I，I-JEPA 首先采样一组整数 $\mathcal{M}^{\text{ctx}} = \{i \subseteq 1, \ldots, \text{N}\}$，每个整数对应 I 中一个图像块的索引，这些图像块共同构成可见上下文。同样，k 组不同的图块索引$\mathcal{M}^{\text{tgt}} = \{\mathcal{M}_j^{\text{tgt}} \subseteq 1, \ldots, \text{N}\}_{j=1,\ldots,k}$ 也会被采样，用作预测目标，因此会从 I 中屏蔽掉。如图2（左下）所示，这些集合对应于由连续图像块组成的区域，其宽高比和尺寸（因此也包括 $\mathcal{M}^{\text{ctx}}$ 和$\mathcal{M}^{\text{tgt}}$的基数）均从预定义范围内采样。需要特别指出的是，目标块之间可以重叠，但上下文与目标之间不得有交集。

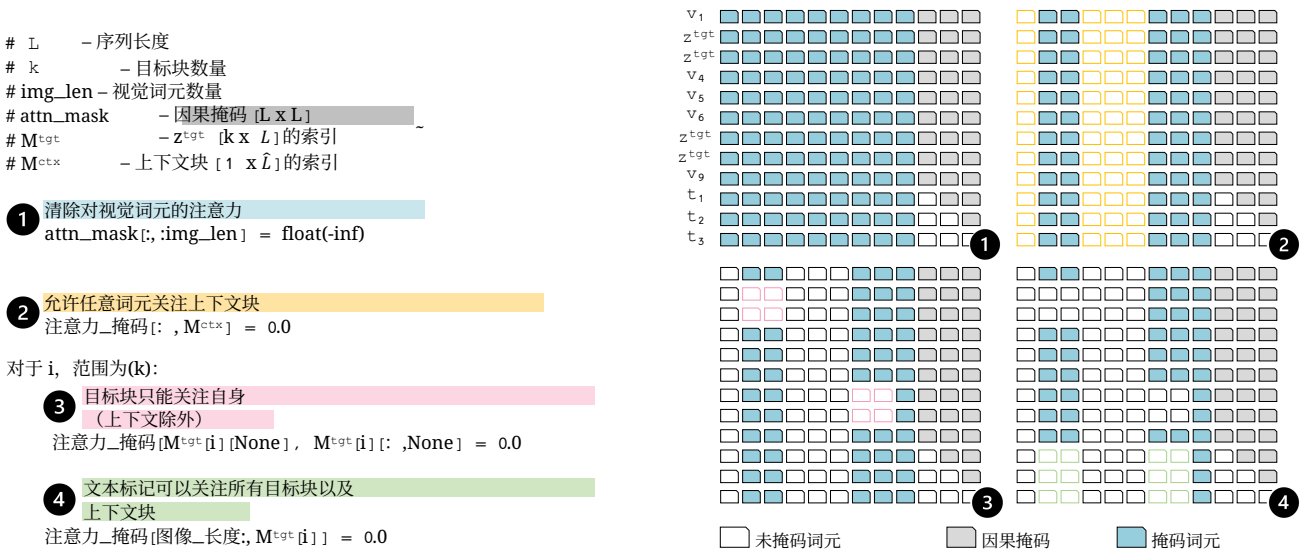


图3. 注意力掩码实现的可视化。**左侧：**概述计算注意力掩码所用连续步骤的伪代码。**右侧：**图示展示了每一步的效果，并突出显示掩码如何改变词元注意力。

与 I-JEPA 不同，我们不会在像素级屏蔽输入图像，而是直接屏蔽冻结视觉编码器 $\mathcal{F}_v$ 提取的视觉嵌入；为清晰起见，下文将其称为 $\mathcal{F}_v^{\text{ctx}}$。此外，还会使用另一个冻结视觉编码器 $\mathcal{F}_v^{\text{tgt}}$，从目标块集合中提取目标嵌入 I^{tgt}。形式上，上下文视觉嵌入集和目标视觉嵌入集定义如下：

$$I^{\text{ctx}} = \{\mathbf{v}_i \in \text{proj}(\mathcal{F}_v^{\text{ctx}}(\text{I})) \text{ s.t. } i \in \mathcal{M}^{\text{ctx}}\} \tag{2}$$

$$I^{\text{tgt}} = \{\tilde{\mathbf{v}}_i \in \mathcal{F}_v^{\text{tgt}}(\text{I}) \text{ s.t. } i \in \mathcal{M}^{\text{tgt}}\}. \tag{3}$$

根据 I-JEPA，训练预测器网络以预测 I^{tgt} from I^{ctx} 以及潜变量 $\mathbf{z}^{\text{tgt}}$，后者指示掩码目标图像块的位置。在实践中，$\mathbf{z}^{\text{tgt}}$被实现为可学习嵌入$\mathbf{z} \in \mathrm{R}^d$ 与每个掩码图像块的位置编码 $\phi(\cdot) \in \mathrm{R}^d$ 相加：

$$\mathbf{z}^{\text{tgt}} = \{\mathbf{z} + \phi(i), \forall i \in \mathcal{M}^{\text{tgt}}\}. \tag{4}$$

逐层目标预测。在我们的模型中，预测器网络对应于大语言模型 $\mathcal{G}$，因此它能够学习恢复掩码图像块的嵌入，从而发现图像的潜在模式和结构。具体而言，近期研究 [8, 61, 67] 表明，多模态大语言模型在其早期到中间层会将注意力集中于视觉词元，因此我们从 $\mathcal{G}$ 的 Transformer 层堆栈中选取一个浅层子集来实现预测器。令 $\mathcal{G}_j(\cdot)_i$ 表示所选层在 j 第 i个词元位置处的激活。现在，我们可以将待最小化的 JEPA 损失函数定义为预测器在目标位置的输出与参考目标之间的距离度量 I^{tgt}。形式上，可表示为：

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{|\mathcal{M}^{\text{tgt}}|} \sum_{i \in \mathcal{M}^{\text{tgt}}} d\left(\text{proj}^{\text{tgt}}(\mathcal{G}_j([I^{\text{ctx}}, \mathbf{z}^{\text{tgt}}]))_i, I_i^{\text{tgt}}\right), \tag{5}$$

其中，$d(\cdot)$ 是距离函数，proj^{tgt} 是一个前馈网络，用于将 $\mathcal{G}_j(\cdot)$ 的维度与 $\mathcal{F}_v^{\text{tgt}}$ 的维度匹配，而 $[\cdot]$ 是沿词元轴进行的拼接。

与此同时，我们仍然训练多模态大语言模型以生成图像字幕 $\mathbf{x}$，同时可以访问可见的上文 I^{ctx} 和潜变量 $\mathbf{z}^{\text{tgt}}$。值得注意的是，得益于 $\mathcal{L}_{\text{JEPA}}$，经过逐层处理，$\mathbf{z}^{\text{tgt}}$ 的表示会逐渐收集图像缺失部分的信息，从而降低为掩码图像生成字幕的难度。因此，公式1 中的下一词元预测目标函数修改如下：

$$\tilde{\mathcal{L}}_{\text{NTP}} = -\sum_i \log P\left(\mathbf{x}_i | [I^{\text{ctx}}, \mathbf{z}^{\text{tgt}}], \mathbf{x}_{1,\ldots,i-1}; \mathcal{G}, \text{proj}\right). \tag{6}$$

通过注意力掩码实现高效计算。整体训-˜练目标由以下各项之和给出：$\mathcal{L}_{\text{NTP}}$ 和 $\mathcal{L}_{\text{JEPA}}$。请注意，这两个损失在一次前向传播中计算完成，无需对大语言模型进行任何额外计算$\mathcal{G}$。这是通过调整 $\mathcal{G}$ 每个注意力层中应用的注意力掩码实现的，具体如下（如图 3所示）。首先，由于上下文和目标视觉嵌入可以交错排列在视觉输入序列$[I^{\text{ctx}}, \mathbf{z}^{\text{tgt}}]$中，我们允许它们之间进行双向注意力，但有以下限制：（i）根据 I-JEPA，上下文嵌入 I^{ctx} 不能关注目标 I^{tgt}；（ii）来自不同块的目标嵌入不能相互关注，但可以关注 I^{ctx}。接下来，由于图像

字幕词元 $\mathbf{x}$ 遵循$\mathcal{G}$的因果注意力机制，并且位于输入序列中的视觉嵌入之后，因此视觉上下文 I^{ctx} 和潜变量 $\mathbf{z}^{\text{tgt}}$ 不会受到图像字幕$\mathbf{x}$的影响，所以在预测 I^{tgt}时不存在文本监督。

平衡视觉与文本监督。在优化步数相同的情况下，仅使用掩码图像训练多模态大语言模型，会产生较弱的图像-文本对齐。此外，这会导致对齐阶段与后续的视觉指令微调和推理阶段之间出现不一致，因为在后续阶段，模型可以访问完整图像。为了解决这一问题，我们提议以概率 $\mathcal{L}_{\text{JEPA}}$ 跳过 λ的计算，从而基于未掩码图像计算原始的下一词元预测损失（参见公式1）。这相当于将 I^{ctx} 扩展至整个图像，因此无需插入潜变量$\mathbf{z}^{\text{tgt}}$，因为没有需要重建的掩码图像块。

遵循 LLaVA 的做法，在完成对齐训练后，我们进入视觉指令微调阶段。在此设置中，不再使用目标视觉编码器 $\mathcal{F}_v^{\text{tgt}}$和投影器 proj^{tgt}，模型则使用未掩码图像进行下一词元预测目标函数训练。

4. 实验

实现与训练细节。 我们严格遵循 LLaVA-1.5 [37]配方训练 JARVIS、基线及其他方法，并且只进行实现所提出方法所必需的更改。除非另有说明，所有模型均采用 CLIP ViT- L/14@336 [44] 作为（上下文）视觉编码器 $\mathcal{F}_v^{\text{ctx}}$，图文投影器 proj 则由一个两层 MLP网络实现。在包含 558k 个图像-字幕对的对齐阶段，我们冻结大语言模型 $\mathcal{G}$，仅训练 proj，以及 JARVIS 的目标投影器 proj^{tgt}。相反，在使用 LLaVA-1.5 发布的数据集中的 665k 个样本进行视觉指令微调时， $\mathcal{G}$ 会解除冻结。不同的是，两个阶段始终冻结视觉编码器。所有模型的其他训练超参数均保持一致，详见补充材料。

为了选择上下文 $\mathcal{M}^{\text{ctx}}$ 和目标 $\mathcal{M}^{\text{tgt}}$（参见公式2）视觉嵌入的索引，我们采用 I-JEPA 提出的精确分块掩码策略 [4]。对于 $\mathcal{M}^{\text{ctx}}$，我们采样图块索引以获取单个视觉嵌入块，该块对应于输入图像的矩形裁剪，覆盖图片面积的 85% 至 100%。另一方面，我们采样 $k = 4$ 组索引，对应于尺寸小于 $\mathcal{M}^{\text{ctx}}$的裁剪，其并集构成目标嵌入 $\mathcal{M}^{\text{tgt}}$ 的索引，并将其从 $\mathcal{M}^{\text{ctx}}$ 中移除，以避免产生简单的预测。

为了实现预测器网络并计算 $\mathcal{L}_{\text{JEPA}}$，我们采用 $\mathcal{G}$ 的前四分之一层，即我们将 j （参见公式5）设为各特定大语言模型深度四分之一处对应层的索引。投影 $^{\text{tgt}}$ 的架构遵循 投影 [37]，的架构，即内部采用带有 GELU 非线性激活函数的两层多层感知机 [20] 。对于目标视觉编码器 $\mathcal{F}_v^{\text{tgt}}$，除非另有说明，我们使用 DINOv2-L/14 [43]，，它以提供高质量空间特征而著称。在对齐训练期间，所有实验中我们均将 λ 设为 0.2，这意味着我们有 $\mathcal{L}_{\text{JEPA}}$ 20% 的概率跳过优化步骤。最后，我们采用负余弦相似度作为公式 d（·；·）中预测视觉嵌入与目标视觉嵌入之间的距离函数5。

评测基准。评测时，我们采用 Cambrian评测套件 [56]，，这是一个综合性基准测试，包含 16 项任务，涵盖四个类别：通用（4 项）、知识（4 项）、光学字符识别（3 项）和视觉中心（5 项）。在视觉中心类别中，Cambrian 包含 RealWorldQA [63]，，用于评估基于视觉输入的常识推理能力；MMVP [57]，，用于考察模型在九类问题上的视觉感知能力；Blink [18] 和 CVBench2D [34, 68]，，侧重于有关空间关系的问题；以及 CVBench3D [9]，，用于评估模型从相机视角判断物体相对深度以及物体间相对距离的能力。对于每项任务，我们报告使用官方评测工具包计算的准确率。通用、知识和光学字符识别类别的详细信息，以及各项任务的实验结果明细，均见补充材料。

4.1. 消融研究

我们首先介绍一组消融研究，旨在了解关键的架构选择和目标函数选择如何影响 JARVIS 的行为。在这些实验中，我们采用 Vicuna-7B 作为基础大语言模型。

改变大语言模型层 $\mathcal{L}_{\text{JEPA}}$。将大语言模型应用于语言任务时，通常会提取其最后一层的激活值用于词元生成。然而，在计算 $\mathcal{L}_{\text{JEPA}}$ （参见公式5）时，JARVIS 也会利用大语言模型在潜在空间中预测图像缺失的部分。鉴于最后一层应对生成连贯语言输出所需的句法和语法结构更为敏感，我们研究这一学习目标是否更适合在中间层上计算。表1（上半部分）报告了在不同层计算 $\mathcal{L}_{\text{JEPA}}$ 时的性能。请注意，$\mathcal{L}_{\text{NTP}}$ （参见公式6）始终在最后一层计算。作为基线，我们还在第一行列出了按照 LLaVA [37]训练时 Vicuna-7B 的评估结果。

在通用、知识和光学字符识别任务上的平均准确率保持稳定的同时，使用中间层而非最终层激活值时，视觉任务上的表现却存在显著差异。例如，视觉中心准确率提升了

表1. 消融研究结果。所有实验均采用 Vicuna-7B 作为基础大语言模型，使用 CLIP ViT-L@336 作为上下文视觉编码器，并使用 DINOv2-L/14 作为目标视觉编码器。

	通用	知识	OCR	视觉中心					
	Avg	Avg	Avg	RealWorldQA					
LLaVA [37]	65.8	41.4	36.2	54.9	31.3	46.8	57.6	63.7	50.9
更改大语言模型层的影响									
$j = 4$（早期）	**65.1**	41.1	36.1	55.9	28.7	45.8	57.7	64.1	50.4
$j = 8$（1/4 深度）	**65.1**	41.6	**36.2**	55.6	**30.7**	**48.3**	60.4	**63.8**	**51.7**
$j = 16$（中期）	64.5	41.7	**36.2**	54.6	28.0	45.3	**60.8**	63.0	50.3
$j = 24$（3/4 深度）	65.0	**42.3**	**36.2**	**57.0**	26.7	46.5	59.4	61.5	50.2
$j = 31$（接近最终层）	64.9	41.9	35.8	55.7	27.3	45.5	56.6	60.9	49.2
$j = 32$（最终）	64.4	41.7	35.9	**57.0**	27.3	45.4	55.5	61.7	49.4
投影器缩放的影响									
采用线性投影	64.8	**42.1**	35.9	54.9	26.0	45.3	58.3	63.7	49.6
采用 MLP 投影	**65.1**	41.6	**36.2**	**55.6**	**30.7**	**48.3**	**60.4**	**63.8**	**51.7**
更改 $\mathcal{L}_{JEPA}$ 中距离函数的影响									
采用平滑 L1 距离（I-JEPA [4]）	64.7	**41.9**	35.8	**56.2**	24.7	45.5	58.1	61.1	49.1
使用余弦距离	**65.1**	41.6	**36.2**	55.6	**30.7**	**48.3**	**60.4**	**63.8**	**51.7**

个百分点： +1.0从第 31 层下降到第 24层时。在中间层中，我们发现位于深度四分之一处的那一层（即， $j = 8$ 对于 Vicuna-7B 而言，即第 32 层，占总深度的四分之一）能够实现最佳的视觉性能；与最终层相比，MMVP 和 CVBench2D 分别有显著提升，提升幅度为+3.4 和 +4.9 个百分点。这些发现支持了我们先前的观察结果，也与近期研究 [8, 61, 67], 一致，后者表明，多模态大语言模型在中间层更加关注视觉词元。

扩展目标投影器。 接下来，我们消融分析扩展目标投影器 投影^{tgt} 从简单的线性层扩展为两层多层感知机对性能的影响，结果见表1（中）。 投影^{tgt} 用于转换大语言模型第 j 层的中间激活，即， $\mathcal{G}_j(\cdot)$ ，使其维度与目标嵌入 I^{tgt} 保持一致。目标嵌入由 $\mathcal{F}_v^{tgt}$生成，用于计算 $\mathcal{L}_{\text{JEPA}}$。使用线性投影不仅无法提升视觉性能，甚至会导致性能低于基线。例如，在 Vision-Centric Blink 基准测试中，性能从 46.8 降至45.3。相比之下，采用多层感知机投影器可提升性能，在视觉中心任务上提高 +1.7 个百分点。这一发现表明，线性变换不足以将大语言模型的中间嵌入投影到视觉编码器的输出空间。

更改 $\mathcal{L}_{\text{JEPA}}$ 的距离函数。针对 $\mathcal{L}_{\text{JEPA}}$ 进行优化，等价于最小化预测视觉嵌入与目标视觉嵌入之间的距离。默认情况下，我们选择负余弦相似度（即余弦距离）作为距离函数。因此，我们要求大语言模型猜测目标嵌入的方向。相反地，我们探索了切换到平滑 L1 距离度量的效果，正如原始 I-JEPA 公式 [4], 所采用的方法一样；该方法不仅要求匹配预测目标的方向，还要求匹配其大小，以便将其最小化。具体而言，当预测嵌入与目标嵌入之间的（逐元素）绝对距离小于 1 时，平滑 L1 距离呈二次下降；否则呈线性上升。根据表1（底部），将余弦距离切换为平滑 L1 距离会导致视觉任务上的 +2.6 平均准确率下降若干个百分点。关键在于，使用平滑 L1 距离最小化 $\mathcal{L}_{\text{JEPA}}$ 会严重削弱相对于基线的通用与视觉中心性能。

在本文余下部分，我们将展示基于消融实验结果所得见解开展的实验。总而言之，我们利用大语言模型前四分之一的层来实现 I-JEPA 所需的预测器网络，并搭配两层多层感知机投影器，使预测视觉嵌入与目标视觉嵌入的维度相匹配，再通过负余弦相似度衡量二者的距离。

4.2. 主要实验结果

与相关方法和基线的比较。我们分析了应用于两个大语言模型 Vicuna-7B 的多模态大语言模型的不同训练方案 [14]（上），以及 Qwen2-7B [54]（下）的结果，并汇总于表2中。除标准版 LLaVA [37]（第一行）外，其他所有方法均采用 DINOv2-L/14 [43]作为目标视觉编码器实现。具体而言，我们加入了 VIRAL [65]（第二行），这是一种用于对齐大语言模型与基础视觉编码器视觉表征的新方法。VIRAL 在 LLaVA 的视觉指令微调阶段加入额外目标函数，对齐大语言模型中间层的激活与外部目标视觉编码器的输出，从而对该阶段进行正则化。

此外，我们还实现了另一个基线（第三行），其架构与 JARVIS 完全相同，但未采用 I- 的掩码预测目标函数

表 2。比较 与包括原始 LLaVA [37] 模型、VIRAL [65], 和我们的基础版 es.

	LLM	通用 Avg	知识 Avg	OCR Avg	视觉中心 RealWorldQA					
LLaVA [37]	Vicuna-7B	**65.8**	41.4	36.2	54.9	**31.3**	46.8	57.6	63.7	50.9
VIRAL [65]		64.6	**42.6**	36.0	54.8	30.0	42.9	55.7	60.0	48.7
我们的模型（无掩码）		65.0	41.3	**36.3**	**55.7**	28.0	42.7	57.7	60.8	49.0
JARVIS（我们的模型）		65.1	41.6	36.2	55.6	30.7	**48.3**	**60.4**	**63.8**	**51.7**
LLaVA [37]	Qwen2-7B	72.0	46.7	36.9	**58.0**	36.7	**52.7**	62.5	66.8	55.3
VIRAL [65]		70.9	47.2	36.1	57.9	36.0	49.3	62.0	62.3	53.5
我们的无掩码版本		**72.4**	**48.1**	**37.0**	56.2	35.3	49.6	62.8	67.6	54.3
JARVIS（我们的方法）		71.9	47.7	36.3	55.8	**38.0**	49.3	**63.9**	**73.0**	**56.0**



图4. 使用 Qwen2-7B 的多模态大语言模型三种训练方法的定性比较 [54]。我们展示了 CVBench2D 的样例 [56]（计数和相对位置，第1和第2列）、CVBench3D 的样例 [56]（深度和距离，第3列），以及 Blink 的样例 [18]（最后一列）。

JEPA [4]。相反，我们训练它只需将大语言模型深度四分之一处那一层的激活与目标视觉编码器对齐。虽然这一方法在概念上与 VIRAL 类似，但与 JARVIS 在视觉指令微调期间应用额外视觉目标不同，它是在对齐阶段应用的。该基线使我们能够衡量 $\mathcal{L}_{\text{JEPA}}$ 对我们训练方法的影响。换句话说，这项比较回答了这样一个问题：要释放大语言模型的视觉感知能力，预测图像缺失部分是否比预测外部视觉编码器生成的完整图像表示更适合作为无监督学习信号？根据表2，答案是肯定的，因为 VIRAL 和不使用掩码的基线模型都无法提升原始 LLaVA 在视觉中心基准上的表现。值得注意的是，这一结论对两种大语言模型都成立。

相反，采用 Vicuna-7B 的 JARVIS 在大多数视觉任务上都超越了标准 LLaVA 模型，平均提升了 +0.8 个百分点。切换到 Qwen2-7B 后，JARVIS 相比 LLaVA 的平均优势为 +0.7 个百分点，并且在具有挑战性的 CVBench3D 上显著提升了 +6.2个百分点，同时在通用、知识和光学字符识别任务上保持了具有竞争力或更高的得分。图 4展示了一些定性结果，其中我们比较了 LLaVA、VIRAL 和 JARVIS 在包括 CVBench2D、CVBench3D 和 Blink 在内的多项视觉任务上的表现。

跨不同大语言模型的泛化。 表3 全面概述了 JARVIS 在不同大语言模型系列中的表现[1]。从规模更紧凑的语言模型开始，JARVIS 在各项视觉任务上始终展现出改进效果。对于 Gemma2-2B [53] 模型，JARVIS 使视觉中心准确率平均提升了 +0.8 个百分点，将其性能从50。0提升至50。8。对于 LLaMA-3.2-3B [19] 模型，JARVIS 实现了更为显著的 +1.4个百分点提升，视觉中心平均值从 52。1 提升至53。5。这些结果证明，JARVIS 能够有效增强即使是资源受限程度更高的多模态大语言模型的基础视觉理解能力。

扩大大语言模型规模后，Vicuna-7B [14] 与 JARVIS 配合，在除 MMVP 之外的所有视觉任务上均胜过 LLaVA，并在具有挑战性的 CVBench2D基准测试上取得显著 +2.8 提升。使用 Ministral-8B（一种 Mistral [23],变体）时，我们注意到，JARVIS 在 MMVP、Blink 和 CVBench3D 上表现更优，在视觉中心任务上的平均准确率提升了 +1.6个百分点。值得注意的是，JARVIS 在通用任务上也超过了 LLaVA，得分从 65.5 提高到 67.2（+1.7），在光学字符识别上则从 33.0 提高到 34.0，同时在知识方面保持了相同的准确率

[1]有关这些实验所用配置的更多详情，请参阅补充材料。

表3. JARVIS 在多个大语言模型系列上的评估表现。

	LLM	视觉编码器	通用知识	光学字符识别		视觉中心					
			Avg	Avg	Avg	RealWorldQA					
LLaVA [37]	Gemma2-2B [53]	CLIP ViT-L	**66.5**	**42.0**	**33.9**	53.3	44.3	30.7	**57.3**	64.7	50.0
JARVIS（我们的方法）			66.1	41.9	33.4	**54.2**	**45.3**	**31.3**	**57.3**	**65.7**	**50.8**
LLaVA [37]	LLaMA-3.2-3B [19]	CLIP ViT-L	**67.4**	**44.7**	**34.7**	56.3	46.4	**32.0**	**63.7**	62.3	52.1
JARVIS（我们的方法）			67.0	44.3	34.3	**58.7**	**49.2**	**32.0**	63.6	**64.1**	**53.5**
LLaVA [37]	Vicuna-7B [14]	CLIP ViT-L	**65.8**	41.4	**36.2**	54.9	**31.3**	46.8	57.6	63.7	50.9
JARVIS（我们的方法）			65.1	**41.6**	**36.2**	**55.6**	30.7	**48.3**	**60.4**	**63.8**	**51.7**
LLaVA [37]	Ministral-8B	CLIP ViT-L	65.5	**42.2**	33.0	**59.2**	30.0	44.8	**61.4**	62.1	51.5
JARVIS（我们的方法）			**67.2**	**42.2**	**34.0**	58.6	**34.7**	**49.2**	57.6	**65.5**	**53.1**
LLaVA [37]	Qwen2-7B [54]	CLIP ViT-L	**72.0**	46.7	**36.9**	**58.0**	36.7	**52.7**	62.5	66.8	55.3
JARVIS（我们的方法）			71.9	**47.7**	36.3	55.8	**38.0**	49.3	**63.9**	**73.0**	**56.0**
LLaVA [37]	Qwen2-7B [54]	SigLIP2 ViT-L	72.4	47.4	39.0	57.5	48.5	48.0	65.8	**68.1**	57.6
JARVIS（我们的方法）			**74.0**	**48.2**	**43.8**	**59.6**	**53.1**	**50.0**	**67.4**	66.8	**59.4**

与 LLaVA 相比。如前文所述，在 Qwen2-7B 上对 JARVIS 和 LLaVA 进行比较，进一步证实了这一趋势：JARVIS 在 CVBench2D 和 CVBench3D 等视觉感知任务上表现出色，同时在其他任务类别上的表现相当或有所提升。

总体而言，JARVIS 与 Qwen2-7B 搭配时，在所考察的大语言模型中表现最佳。基于这一点，我们还测试了这一组合在扩展上下文视觉编码器时的表现，将编码器从 CLIP ViT-L/13@336 [44] 切换为更先进的 SigLIP2-So400M L/14@384 [59]。在这种设置下，JARVIS 在全部四个类别上的表现均优于 LLaVA，在通用类别上领先 LLaVA $+1.6$ 分，在知识类别上领先 $+0.8$ 分，在光学字符识别类别上领先 $+4.8$ 分，在 $+1.8$ 视觉中心任务上领先。结果表明，与原始视觉指令微调 [37]方案相比，JARVIS 从更强大的视觉编码器中获益更多。

4.3. 扩展目标编码器

采用 JARVIS 的最佳配置，即 将 Qwen2-7B 与作为上下文视觉编码器的 SigLIP2 配对，我们评估了改变和扩展目标编码器的影响。表5 报告了 JARVIS 在三个以视觉为中心的数据集（Blink、CVBench2D 和 MMVP）上的性能，以及所有视觉任务的平均表现。我们考察了与语言监督一同训练的视觉编码器，例如 CLIP [44]和 SigLIP 2 [59], , 以及来自 DINOv2 [43] 和 DINOv3 [49]系列的纯无监督模型，其规模从基础版到巨型版（1B）不等。所有视觉编码器均以分辨率 $[384 \times 384]$运行。我们首先比较 CLIP、SigLIP2 和 DINOv2，发现 DINOv2 在所有数据集上始终取得最佳表现，这表明语言监督编码器提取的密集特征不如 DINO 提取的特征。受此启发，我们进一步探索了多种 DINO 变体，包括基础版、大型版、巨型版和超大型版模型。总体而言，大型版或更先进的编码器往往能带来更高的性能，证明了扩展 JARVIS 中目标编码器的益处。

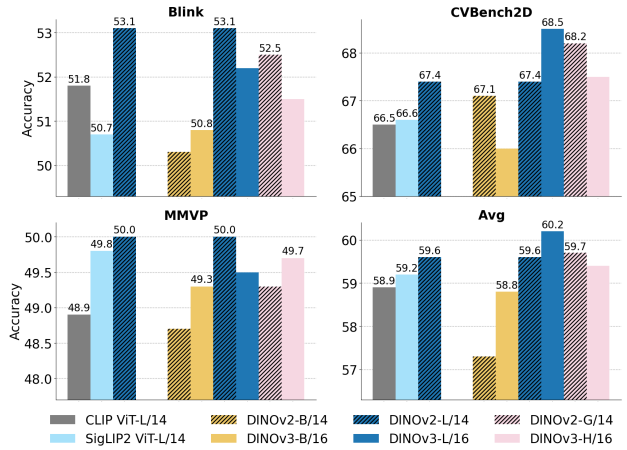


图5. 扩大目标视觉编码器规模对以视觉为中心的基准测试的影响。

5. 结论

本文介绍了 JARVIS，这是一个旨在提升多模态大语言模型视觉感知能力的框架，其集成了受 I-JEPA 启发的自监督学习目标。JARVIS 的核心洞见在于，多模态大语言模型可以通过在潜在空间中预测图像缺失部分，获得更稳健、更精细的视觉表征，突破仅依赖文本描述所产生的监督所固有的局限。在 LLaVA 的对齐阶段引入 I-JEPA 的预测目标后，JARVIS 引导大语言模型捕捉视觉世界内在的结构与语义规律，从而加深对视觉内容的理解。针对多种大语言模型系列开展的大量实验表明，这种方法在广泛的以视觉为中心的基准测试上持续取得显著提升，凸显了自监督视觉学习能够使多模态大语言模型有效地超越文字进行观察，并更好地推理复杂的视觉信息。

致谢

我们感谢 ISCRA 计划授予的 CINECA 资助，以及 EuroHPC 超级计算机 LUMI 提供的高性能计算资源。本研究获得了欧盟地平线计划项目“ELLIOT——面向任意数据流稳健泛化的欧洲大型开放式多模态基础模型”（编号：101214398）、EuroHPC联合事业项目“MINERVA”（资助协议编号：101182737），以及由欧盟—下一代欧盟资助的 PNRR 项目“ITSERR”（CUPB53C22001770006）的支持。

参考文献

[1] Mohamed Abdelfattah 和 Alexandre Alahi。《S-JEPA：用于骨骼动作识别的联合嵌入预测架构》。发表于 ECCV，2024 年。3[2] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。《GPT-4 技术报告》。arXiv 预印本 arXiv:2303.08774，2023 年。2[3] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katherine Millican、Malcolm Reynolds 等。《Flamingo：用于少样本学习的视觉语言模型》。发表于 NeurIPS，2022 年。1，2[4] Mahmoud Assran、Quentin Duval、Ishan Misra、Piotr Bojanowski、Pascal Vincent、Michael Rabbat、Yann LeCun 和 Nicolas Ballas。《利用图像进行自监督学习：联合嵌入预测架构》。发表于 CVPR，2023 年。2，3，5，6，7，12，14[5] Mido Assran、Adrien Bardes、David Fan、Quentin Garrido、Russell Howes、Matthew Muckley、Ammar Rizvi、Claire Roberts、Koustuv Sinha、Artem Zholus 等。《V-JEPA 2：自监督视频模型实现理解、预测与规划》。arXiv 预印本 arXiv:2506.09985，2025 年。3[6] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。《Qwen2.5-VL 技术报告》。arXiv 预印本 arXiv:2502.13923，2025 年。2[7] Adrien Bardes、Quentin Garrido、Jean Ponce、Xinlei Chen、Michael Rabbat、Yann LeCun、Mido Assran 和 Nicolas Ballas。《重新审视特征预测：从视频中学习视觉表征》。TMLR，2024 年。3[8] Jing Bi、Junjia Guo、Yunlong Tang、Lianggong Bruce Wen、Zhang Liu、Bingjie Wang 和 Chenliang Xu。《揭示语言模型中的视觉感知：一种注意力头分析方法》。发表于 CVPR，2025 年。2，4，6[9] Garrick Brazil、Abhinav Kumar、Julian Straub、Nikhila Ravi、Justin Johnson 和 Georgia Gkioxari。《Omni3D：用于野外三维目标检测的大型基准测试与模型》。发表于 CVPR，2023 年。5[10] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。载于 NeurIPS，2020。1[11] Davide Caffagni、Federico Cocchi、Luca Barsellotti、Nicholas Moratelli、Sara Sarto、Lorenzo Baraldi、Lorenzo Baraldi、Marcella Cornia 和 Rita Cucchiara。多模态大语言模型的变革：一项综述。载于 ACL Findings，2024。1, 2[12] Gongwei Chen、Leyang Shen、Rui Shao、Xiang Deng 和 Liqiang Nie。LION：以双层视觉知识赋能多模态大语言模型。载于 CVPR，2024。2[13] Kanzhi Cheng、Wenpo Song、Jiaxin Fan、Zheng Ma、Qiushi Sun、Fangzhi Xu、Chenyang Yan、Nuo Chen、Jianbing Zhang 和 Jiajun Chen。CapArena：大语言模型时代的详细图像描述基准测试与分析。arXiv 预印本 arXiv:2503.12329，2025。2[14] Wei-Lin Chiang、Zhuohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E. Gonzalez、Ion Stoica 和 Eric P. Xing。Vicuna：一款开源聊天机器人，以 90%* 的 ChatGPT 质量令 GPT-4 印象深刻，2023。6, 7, 8, 12, 13, 14, 15[15] Federico Cocchi、Nicholas Moratelli、Davide Caffagni、Sara Sarto、Lorenzo Baraldi、Marcella Cornia 和 Rita Cucchiara。LLaVA-MORE：用于增强视觉指令微调的大语言模型与视觉骨干网络比较研究。载于 ICCV Workshops，2025。1[16] Ailin Deng、Tri Cao、Zhirui Chen 和 Bryan Hooi。文字还是视觉：视觉语言模型会盲目相信文本吗？载于 CVPR，2025。2[17]Chaoyou Fu、Peixian Chen、Yunhang Shen、Yulei Qin、Mengdan Zhang、Xu Lin、Jinrui Yang、Xiawu Zheng、Ke Li、Xing Sun 等。MME：多模态大语言模型综合评测基准。arXiv 预印本 arXiv:2306.13394，2023。14[18] Xingyu Fu、Yushi Hu、Bangzheng Li、Yu Feng、Haoyu Wang、Xudong Lin、Dan Roth、Noah A Smith、Wei-Chiu Ma 和 Ranjay Krishna。BLINK：多模态大语言模型看得见，却无法感知。载于 ECCV，2024。1, 5, 7[19] Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。Llama 3 模型家族。arXiv 预印本 arXiv:2407.21783，2024。7, 8, 12[20] D Hendrycks。高斯误差线性单元（GELU）。arXiv 预印本 arXiv:1606.08415，2016。5[21] Hai Huang、Yann LeCun 和 Randall Balestriero。LLM-JEPA：大语言模型邂逅联合嵌入预测架构。arXiv 预印本 arXiv:2509 14252，2025 3 . . [22] Drew A Hudson 和 Christopher D Manning。GQA：用于现实世界视觉推理和组合式问答的新数据集。载于 CVPR，2019。14[23] Albert Q. Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Florian Bressand、Gianna Lengyel、Guillaume Lam- ple、Lucile Saulnier、L´elio Renard Lavaud、Marie-Anne Lachaux、Pierre Stock、Teven Le Scao、Thibaut Lavril,

Thomas Wang、Timoth´ee Lacroix 和 William El Sayed。Mistral 7b。arXiv 预印本 arXiv:2310.06825，2023。7[24] Seil Kang、Jinyeong Kim、Junhyeok Kim 和 Seong Jae Hwang。你的大型视觉语言模型只需少数注意力头即可实现视觉定位。载于 CVPR，2025。2[25] O˘guzhan Fatih Kar、Alessio Tonioni、Petra Poklukar、Achin Kulshrestha、Amir Zamir 和 Federico Tombari。BRAVE：拓展视觉语言模型的视觉编码。载于 ECCV，2024。2[26] Aniruddha Kembhavi、Mike Salvato、Eric Kolve、Minjoon Seo、Hannaneh Hajishirzi 和 Ali Farhadi。一张示意图胜过十二张图像。载于 ECCV，2016。14[27] Yann LeCun。迈向自主机器智能之路。2022。2, 3[28] Hongyang Lei、Xiaolong Cheng、Qi Qin、Dan Wang、Huazhen Huang、Qingqing Gu、Yetao Wu 和 Luo Ji。M3-JEPA：基于联合嵌入预测架构、通过多门控混合专家实现多模态对齐。载于 ICML，2025。3[29] Bohao Li、Rui Wang、Guangzhi Wang、Yuying Ge、Yixiao Ge 和 Ying Shan。SEED-Bench：以生成式理解为多模态大语言模型进行基准测试。arXiv 预印本 arXiv:2307.16125，2023。14[30] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型引导语言-图像预训练。载于 ICML，2023。1, 2[31] Xianhang Li、Chen Huang、Chun-Liang Li、Eran Malach、Josh Susskind、Vimal Thilak 和 Etai Littwin。重新思考 JEPA：借助冻结教师模型实现计算高效的视频自监督学习。arXiv 预印本 arXiv:2509.24317，2025。3[32] Ji Lin、Hongxu Yin、Wei Ping、Pavlo Molchanov、Moham-mad Shoeybi 和 Song Han。VILA：视觉语言模型的预训练。载于 CVPR，2024。1[33] Junyan Lin、Haoran Chen、Yue Fan、Yingqi Fan、Xin Jin、Hui Su、Jinlan Fu 和 Xiaoyu Shen。多模态大语言模型中的多层视觉特征融合：方法、分析与最佳实践。载于 CVPR，2025。2[34] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于 ECCV，2014。5[35] Ziyi Lin、Chris Liu、Renrui Zhang、Peng Gao、Longtian Qiu、Han Xiao、Han Qiu、Chen Lin、Wenqi Shao、Keqin Chen 等。SPHINX：面向多模态大语言模型的权重、任务与视觉嵌入联合混合。arXiv 预印本 arXiv:2311.07575，2023。2[36] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。载于 NeurIPS，2023。1, 2, 3[37] Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。通过视觉指令微调改进基线。载于 CVPR，2024。1, 2, 3, 5, 6, 7, 8, 12, 13, 14, 15, 16[38] Yuan Liu、Haodong Duan、Yuanhan Zhang、Bo Li、Songyang Zhang、Wangbo Zhao、Yike Yuan、Jiaqi Wang、Conghui He、Ziwei Liu 等。MMBench：你的多模态模型是全能选手吗？载于 ECCV，2024。14

[39] Ilya Loshchilov 和 Frank Hutter。《解耦权重衰减正则化》。发表于 ICLR，2019 年。12[40] Pan Lu、Swaroop Mishra、Tanglin Xia、Liang Qiu、Kai-Wei Chang、Song-Chun Zhu、Oyvind Tafjord、Peter Clark 和 Ashwin Kalyan。《学会解释：通过思维链进行多模态推理以实现科学问答》。发表于 NeurIPS，2022 年。14[41] Pan Lu 、 Hritik Bansal 、 Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao。《Mathvista：评估基础模型在视觉场景中的数学推理能力》。发表于 ICLR，2024 年。14[42] Ahmed Masry、Xuan Long Do、Jia Qing Tan、Shafiq Joty 和 Enamul Hoque。《Chartqa：一种通过视觉和逻辑推理回答图表相关问题的基准测试》。发表于 ACL，2022 年。15[43] Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。《DINOv2：无需监督，学习稳健的视觉特征》。TMLR，第 1–31 页，2024 年。3、5、6、8、14[44]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《从自然语言监督中学习可迁移视觉模型》。发表于 ICML，2021 年。3、5、8[45] Samyam Rajbhandari、Jeff Rasley、Olatunji Ruwase 和 Yuxiong He。《ZeRO：面向万亿参数模型训练的内存优化》。发表于 IEEE SC20，2020 年。12[46] Rajesh PN Rao 和 Dana H Ballard。《视觉皮层中的预测编码：对某些经典外感受野效应的功能性解释》。《自然·神经科学》，2：79–87，1999 年。2、3[47]Ayumu Saito、Prachi Kudeshia 和 Jiju Poovvancheri。《Point-JEPA：一种用于点云自监督学习的联合嵌入预测架构》。发表于 WACV，2025 年。3[48] Sara Sarto、Marcella Cornia 和 Rita Cucchiara。《多模态大语言模型时代的图像描述评估：挑战与未来展望》。发表于 IJCAI，2025 年。2[49] Oriane Sim´eoni、Huy V Vo、Maximilian Seitzer、Federico Baldassarre、Maxime Oquab、Cijo Jose、Vasil Khalidov、Marc Szafraniec、Seungeun Yi、Micha¨el Ramamonjisoa 等。《DINOv3》。arXiv 预印本 arXiv:2508.10104，2025 年。8[50] Amanpreet Singh、Vivek Natarajan、Meet Shah、Yu Jiang、Xinlei Chen、Dhruv Batra、Devi Parikh 和 Marcus Rohrbach。《迈向能够阅读的视觉问答模型》。发表于 CVPR，2019 年。15[51] Chameleon 团队。《Chameleon：混合模态早期融合基础模型》。arXiv 预印本 arXiv:2405.09818，2024 年。1、2[52]Gemini 团队、Rohan Anil、Sebastian Borgeaud、Jean-Baptiste Alayrac、Jiahui Yu、Radu Soricut、Johan Schalkwyk、Andrew M Dai、Anja Hauth、Katie Millican 等。《Gemini：一系列能力卓越的多模态模型》。arXiv 预印本 arXiv:2312.11805，2023 年。2

[53]Gemma 团队、Morgane Riviere、Shreya Pathak、Pier Giuseppe Sessa、Cassidy Hardin、Surya Bhupati-raju、L´eonard Hussenot、Thomas Mesnard、Bobak Shahriari、Alexandre Ram´e 等。《Gemma 2：以实用规模改进开放语言模型》。arXiv 预印本 arXiv:2408.00118，2024 年。7、8、12[54] Qwen 团队等。《Qwen2 技术报告》。arXiv 预印本 arXiv:2407.10671，2024 年。1、6、7、8、12、13、14、15、16[55] Hugo Thimonier、Jos´e Lucas De Melo Costa、Fabrice Popineau、Arpad Rimmel 和 Bich-Liˆen DOAN。《T-JEPA：无需数据增强的表格数据自监督学习》。发表于 ICLR，2025 年。3[56] Peter Tong、Ellis Brown、Penghao Wu、Sanghyun Woo、Adithya Jairam Vedagiri IYER、Sai Charitha Akula、Shusheng Yang、Jihan Yang、Manoj Middepogu、Ziteng Wang 等。《Cambrian-1：对多模态大语言模型进行全面开放的视觉中心探索》。发表于 NeurIPS，2024 年。1、2、5、7、14、15、16[57] Shengbang Tong、Zhuang Liu、Yuexiang Zhai、Yi Ma、Yann LeCun 和 Saining Xie。《闭目塞听？探索多模态大语言模型的视觉缺陷》。发表于 CVPR，2024 年。1、2、5、16[58]Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar 等。《LLaMA：开放且高效的基础语言模型》。arXiv 预印本 arXiv:2302.13971，2023 年。1[59] Michael Tschannen、Alexey Gritsenko、Xiao Wang、Muham-mad Ferjad Naeem、Ibrahim Alabdulmohsin、Nikhil Parthasarathy、Talfan Evans、Lucas Beyer、Ye Xia、Basil Mustafa 等。《SigLIP 2：具备增强语义理解、定位能力和密集特征的多语言视觉语言编码器》。arXiv 预印本 arXiv:2502 14786，2025 年 8。 .[60] Haochen Wang、Anlin Zheng、Yucheng Zhao、Tiancai Wang、Zheng Ge、Xiangyu Zhang 和 Zhaoxiang Zhang。《重建式视觉指令微调》。发表于 ICLR，2025 年。2[61] Sudong Wang、Yunjian Zhang、Yao Zhu、Jianing Li、Zizhe Wang、Yanwei Liu 和 Xiangyang Ji。《迈向理解大型视觉语言模型中的知识如何演变》。发表于 CVPR，2025 年。2、4、6[62] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、R´emi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。《Transformers：自然语言处理的最新进展》。发表于 EMNLP，2020 年。12[63] xAI。《Grok》，2024 年。1、5[64] Jiabo Ye、Haiyang Xu、Haowei Liu、Anwen Hu、Ming Yan、Qi Qian、Ji Zhang、Fei Huang 和 Jingren Zhou。《mPLUG-Owl3：迈向多模态大语言模型中的长图像序列理解》。发表于 ICLR，2025 年。1、2[65] Heeji Yoon、Jaewoo Jung、Junwan Kim、Hyungyu Choi、Heeseong Shin、Sangbeom Lim、Honggyu An、Chaehyun Kim、Jisang Han、Donghyun Kim 等。《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。2，6，7，14，15，16[66] Xiang Yue、Yuansheng Ni、Kai Zhang、Tianyu Zheng、Ruoqi Liu、Ge Zhang、Samuel Stevens、Dongfu Jiang、Weiming Ren、Yuxuan Sun 等。《MMMU：面向专家级通用人工智能的大规模多学科多模态理解与推理基准》。发表于*CVPR*，2024 年。14[67] Zhi Zhang、Srishti Yadav、Fengze Han 和 Ekaterina Shutova。《多模态大语言模型中的跨模态信息流》。发表于 *CVPR*，2025 年。2，4，6[68] Bolei Zhou、Hang Zhao、Xavier Puig、Tete Xiao、Sanja Fidler、Adela Barriuso 和 Antonio Torralba。《通过 ADE20K 数据集理解场景语义》。*IJCV*，127(3):302–321，2019 年。5

超越文字：多模态大语言模型的自监督视觉学习

补充材料

A. 其他实现细节

训练细节。 对于所有实验，我们遵循 LLaVA-1 提出的训练超参数。5 [37], 相关细节见表 4。从计算角度来看，JARVIS 仅需在对齐阶段对目标视觉编码器进行额外的前向传播；与大语言模型的前向传播相比，这一过程相对轻量，并且无需计算任何梯度。

大语言模型细节。 我们使用可在 Hugging Face 的 Transformers 上免费获取的开源大语言模型的指令微调版本进行实验。表5中给出了每个大语言模型的确切引用。

掩码细节 - $\mathcal{M}^{\text{tgt}}$。 I-JEPA 的目标 [4] 是根据可见的上下文图像块 $\mathcal{M}^{\text{ctx}}$，预测 k 图像块中的目标块的潜在表征 $\{\mathcal{M}_j^{\text{tgt}}\}_{j=1,\ .....k}$。实际上，这些块被实现为整数集合，其中集合中的每一项对应于输入图像中一个图像块的索引。

对于每个样本，JARVIS 都会预测四个可能彼此重叠的目标块的表示（即，$k=4$）。在 I- JEPA 中，预测器会针对每个目标块执行一次专门的前向传播，每次前向传播的输入都是上下文嵌入与对应于掩码图像块的位置编码的拼接。由于 JARVIS 中的预测器是大语言模型，为了高效实现，我们将上下文嵌入和 k 目标块（通过可学习嵌入加位置编码实现）打包到同一输入中，即，$[I^{\text{ctx}}, \mathbf{z}^{\text{tgt}}]$ 式5中的内容，从而使大语言模型只需执行一次前向传播。为复现 I-JEPA 对各目标块的隔离，我们修改注意力掩码，使属于不同目标块的词元无法相互关注（见图3）。请注意，若目标块存在重叠，块间仍可能发生信息泄漏；不过，这不会影响性能，表6 （上部和中部）证明了这一点。每个目标块的尺度 $\mathcal{M}_j^{\text{tgt}}$，即掩码图像块数量占图像块总数的比例，在每个批次中均匀采样，范围为$(0.15,\ 0.20)$；宽高比的范围为$(0.75,\ 1.5)$。

掩码细节 - $\mathcal{M}^{\text{ctx}}$。 同样，可见上下文$\mathcal{M}^{\text{ctx}}$ 是由一组宽高比在 (0.75,1.5) 范围内的图像块构成的单个块，但覆盖了图像中更大的区域。实际上，尺度是在 (0.85,1.0) 范围内均匀采样的。为避免出现简单的预测，我们从 $\mathcal{M}^{\text{ctx}}$ 中移除任何对应图像块出现在任一 k 目标块中的索引。图 6 展示了将输入图像划分为上下文（即，黑色虚线 边框）和目标块（即，彩色点线边框）的示例 6。需要强调的是，实际操作中掩码并非在像素级进行，因为上下文视觉编码器 $\mathcal{F}_v^{\text{ctx}}$仍会处理完整的未掩码图像。相反，掩码作用于编码器输出的嵌入，因此只有与$\mathcal{M}^{\text{ctx}}$ (即， I^{ctx}) 对应的视觉嵌入会输入大语言模型，而与目标图像块对应的嵌入则会被潜变量 $\mathbf{z}^{\text{tgt}}$替代。

表 4. 第一训练阶段（即，对齐）和第二阶段（即，视觉指令微调）的训练超参数。我们采用与 LLaVA 1.5 [37] 相同的取值。

超参数	训练	
	阶段 1	阶段 2
批次大小	256	256
学习率	1e-3	2e-5
学习率调度	余弦衰减	
学习率预热比例	0.03	
权重衰减	0	
训练轮数	1	
优化器	AdamW [39]	
DeepSpeed 阶段 [45]	2	3

表5. 本文使用的各个大语言模型与其在 Hugging Face 上页面的对应关系 [62]。

LLM	Hugging Face 页面
Gemma2-2B [53]	google/gemma-2-2b-it
LLaMA-3.2-3B [19]	meta-llama/Llama-3.2-3B-Instruct
Vicuna-7B [14]	lmsys/vicuna-7b-v1.5
Ministral-8B	mistralai/Ministral-8B-Instruct-2410
Qwen2-7B [54]	Qwen/Qwen2-7B-Instruct

B. 补充实验结果

B.1. 补充消融研究

验证掩码策略。表6（上）比较了训练期间目标区域之间存在重叠与否时的性能。当允许目标之间重叠 $\{\mathcal{M}_j^{\text{tgt}}\}$时（如 I-JEPA [4] 所示，见图6），模型在大多数基准测试中取得了略高的分数（例如， +1.6视觉中心平均值从 50.1 提高到51.7）。这表明，有限的重叠可以起到温和的正则化作用，促使模型更好地利用空间上下文并提升跨区域一致性。相反，强制目标互不重叠（无重叠）会略微降低下游性能，这可能是因为学习到的视觉表征多样性有所下降，且空间对应关系较弱。

在表6 （中），我们进一步分析不同注意力配置的影响——

表6. 额外消融研究结果。所有实验均采用 Vicuna-7B 作为基础大语言模型，使用 CLIP ViT-L@336 作为上下文视觉编码器，DINOv2-L/14 作为目标视觉编码器。

	通用	知识	OCR	视觉中心					
	Avg	Avg	Avg	RealWorldQA					
LLaVA [37]	65.8	41.4	36.2	54.9	31.3	46.8	57.6	63.7	50.9
防止目标重叠									
无重叠	**65.2**	**41.7**	35.9	55.3	28.7	46.8	58.7	61.1	50.1
有重叠	65.1	41.6	**36.2**	**55.6**	**30.7**	**48.3**	**60.4**	**63.8**	**51.7**
改变注意力模式的效果									
$\text{attn}_{\text{目标}\to\text{目标}}$	64.7	41.4	35.7	**56.5**	29.3	45.0	59.9	61.8	50.5
$\text{attn}_{\text{目标}\nrightarrow\text{文本}}$	65.0	**42.2**	35.7	54.9	25.3	46.2	57.3	62.5	49.2
$\text{attn}_{\text{目标}\nrightarrow\text{目标，目标}\to\text{文本}}$	**65.1**	41.6	**36.2**	55.6	**30.7**	**48.3**	**60.4**	**63.8**	**51.7**
更改 λ的影响									
$\lambda = 0.0$	63.8	**41.9**	35.6	**58.3**	28.7	47.4	57.1	60.9	50.5
$\lambda = 0.2$	**65.1**	41.6	**36.2**	55.6	30.7	**48.3**	**60.4**	**63.8**	**51.7**
$\lambda = 0.5$	64.6	**41.9**	36.0	57.1	**32.0**	45.8	59.0	62.3	51.2

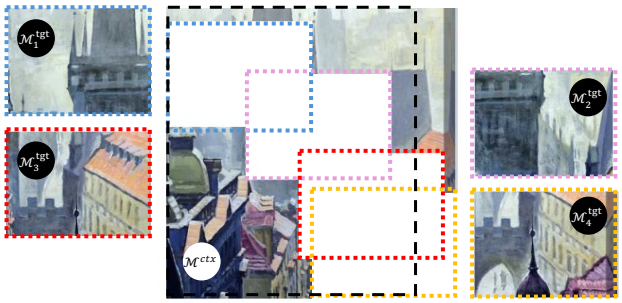


图 6. I-JEPA 的掩码策略示例。首先，按照给定的尺度和宽高比采样四个图像块区域 $\{\mathcal{M}_j^{\text{tgt}}\}_{j=1,\,2,\,3,\,4}$（彩色虚线框）。这些区域将由 $\mathcal{F}_v^{\text{tgt}}$处理，并作为 $\mathcal{L}_{\text{JEPA}}$的预测目标。同时，采样一个较大的图像块区域 $\mathcal{M}^{\text{ctx}}$（黑色虚线框），并移除与目标块重叠的图像块。剩余图像块由$\mathcal{F}_v^{\text{ctx}}$转换为视觉嵌入，然后作为条件信号传递给大语言模型，以预测被掩码的目标。

训练期间视觉词元和文本词元之间的注意力配置。我们首先讨论为何要阻止来自不同目标块的词元彼此关注（$\text{attn}^{\text{tgt}\nrightarrow\text{tgt}}$）。我们最初选择这一策略是为了复现 I-JEPA：其中，目标块通过不同的前向传播进行预测，因此彼此之间不存在交互。当我们允许来自不同块的目标彼此关注时，即 $\text{attn}^{\text{tgt}\to\text{tgt}}$，通用、知识和光学字符识别类别上的性能保持稳定，但在视觉中心类别上，性能降至 LLaVA 基础版水平以下。

接下来，我们通过禁用目标词元对文本标记的注意力（即，$\text{attn}^{\text{tgt}\nrightarrow\text{txt}}$）进行实验，从而阻止大语言模型在生成图像描述时查看其对图像缺失部分的预测。这一改动实际上将.....的计算分离出来

$\mathcal{L}_{\text{NTP}}$ 与 $\mathcal{L}_{\text{JEPA}}$，并导致视觉感知能力严重下降，视觉中心平均分从 51。7 降至 49。2。总体而言，我们提出的配置即，$\text{attn}^{\text{tgt}\nrightarrow\text{tgt},\ \text{tgt}\to\text{txt}}$，在阻断不同目标块之间注意力的同时，允许文本关注预测目标，在通用、光学字符识别和视觉中心任务上取得了最佳成绩。

.....的影响 $\mathcal{L}_{\text{JEPA}}$ 丢弃率。在表6 （底部）中，我们通过消融研究探究了更改 λ 与 $\mathcal{L}_{\text{JEPA}}$ 损失相关的概率所带来的影响。正如正文中所述，这种丢弃机制旨在平衡掩码图像建模（通过 $\mathcal{L}_{\text{JEPA}}$）的目标，同时维持稳健的图文对齐，以供后续指令微调。我们假设，较小但非零的丢弃率（λ）能让模型定期接受整幅图像监督，从而提升整体图文对齐效果和下游性能。为验证这一点，我们通过改变概率 $\lambda \in \{0.0, 0.2, 0.5\}$，同时保持其他训练参数不变，进行了消融研究。当 $\lambda = 0.0$，$\mathcal{L}_{\text{JEPA}}$ 始终处于启用状态。这导致大多数基准测试上的性能下降，说明持续的掩码监督会损害高层次的图像文本理解。使用 $\lambda = 0.5$时，我们在某些任务上取得了略好的结果，例如在 CVBench2D 上，得分从57.1 提升至59.0。然而，整体表现最佳的是 $\lambda = 0.2$。JARVIS 最终采用了这一数值，表明动态平衡视觉（掩码）监督与文本（整幅图像下一词元预测）监督能够带来益处。

关于 Qwen2-7B 的消融研究。我们重新进行主论文中表1 的消融研究，这次将 Vicuna-7B [14] 替换为 Qwen2-7B [54] 作为 JARVIS 的底层大语言模型。首先，我们讨论计算 j 的层 $\mathcal{L}_{\text{JEPA}}$ 的选择（表 6，上）。与 Vicuna-7B 类似，选择 j 后半部分的层

表 7. 以 Qwen2-7B 作为基础大语言模型、CLIP ViT-L@336 作为上下文视觉编码器、DINOv2-L/14 作为目标视觉编码器的消融研究。

	通用	知识	OCR	视觉中心					
	Avg	Avg	Avg	RealWorldQA					
LLaVA [37]	72.0	46.7	36.9	58.0	36.7	52.7	62.5	66.8	55.3
更改大语言模型层的影响									
$j = 4$（早期）	**72.3**	47.4	36.9	**57.5**	**39.3**	47.7	63.7	69.2	55.5
$j = 7$（1/4 深度）	71.9	**47.7**	36.3	55.8	38.0	49.3	**63.9**	**73.0**	**56.0**
$j = 14$（中间）	72.1	**47.7**	36.5	56.5	34.7	48.9	63.0	69.4	54.5
$j = 21$（3/4 深度）	72.1	**47.7**	36.7	56.9	**39.3**	48.9	62.9	66.5	54.9
$j = 27$（接近最终层）	72.2	**47.7**	**37.1**	56.5	33.3	48.9	62.4	66.3	53.5
$j = 28$（最终）	72.2	47.6	36.4	56.2	37.3	**50.2**	62.5	67.3	54.7
投影器缩放的影响									
使用线性投影	**72.0**	**48.1**	**36.6**	**55.8**	**38.7**	48.8	61.6	66.8	54.3
使用 MLP 投影	71.9	47.7	36.3	**55.8**	38.0	**49.3**	**63.9**	**73.0**	**56.0**
更改 $\mathcal{L}_{JEPA}$ 中距离函数的影响									
使用平滑 L1 距离（I-JEPA [4]）	71.7	47.6	**36.4**	**57.0**	**40.0**	**49.8**	62.0	66.3	55.0
使用余弦距离	**71.9**	**47.7**	36.3	55.8	38.0	49.3	**63.9**	**73.0**	**56.0**

无法有效提升相对于 LLaVA 的视觉中心性能。对于 Qwen2-7B，将视觉感知提升到最佳效果的层位于深度的四分之一处，这进一步证实了我们的方法选择可以推广到 Vicuna-7B 之外。接下来，我们验证将 MLP网络作为单一线性投影替换为用于实现投影[tgt]的影响。结果（表 6，下）证实，采用非线性算子将第 j层 Qwen2 的输出转换到 DINOv2 [43], 的目标嵌入空间，相较于线性投影器，在视觉任务上平均提升了 +1.7个百分点。最后，我们将平滑 L1 距离作为 $\mathcal{L}_{\text{JEPA}}$ （表 7，下）的距离度量。虽然就平均视觉中心准确率而言，余弦距离仍是最优选择，但我们注意到，相较于平滑 L1 距离的提升幅度在使用 Vicuna-7B 作为大语言模型时为 +2.6 个百分点，而在 Qwen2-7B 上则降低至1。0个百分点。这表明，更先进的大语言模型可以更容易地与视觉基础模型对齐，因此对它们而言，同时预测视觉嵌入的幅度和方向是可行的。

B.2. CVBench 上的详细结果

在表格 8中，我们详细列出了 CVBench [56]各项任务的性能表现。无论使用 Vicuna-7B 还是 Qwen2-7B 大语言模型，JARVIS 的表现始终优于 LLaVA 基线及其他方法。使用 Vicuna-7B 时，JARVIS 在 CVBench2D 和 CVBench3D 上均取得最高总分，并且在关系^{2D} 任务上提升尤为显著，得分达到 64.6，而 LLaVA 的得分为 59.8 分。使用更强大的 Qwen2-7B 模型时，性能优势更加明显，JARVIS 在所有报告的任务上均取得最佳结果。具体而言，它将 CVBench3D 的总分提升至

表8. CVBench [56]各单项任务的性能细分

	CVBench2D			CVBench3D		
	计数 ^{2D}	关系 ^{2D}	整体深度	距离 ^{3D}	总体 ^{3D}	
Vicuna-7B [14]						
LLaVA [37]	57.0	59.8	57.6	71.8	55.5	63.7
VIRAL [65]	53.2	60.2	55.7	64.2	**55.8**	60.0
我们的方法（无掩码）	56.2	60.8	57.7	69.5	52.0	60.8
JARVIS（我们的方法）	**57.9**	**64.6**	**60.4**	**73.4**	54.2	**63.8**
Qwen2-7B [54]						
LLaVA [37]	52.9	76.0	62.5	76.3	57.3	66.8
VIRAL [65]	53.2	74.2	62.0	74.0	50.5	62.3
我们的模型（无掩码）	53.2	76.5	62.8	78.3	56.8	67.6
JARVIS（我们的模型）	**53.8**	**78.2**	**63.9**	**81.7**	**64.3**	**73.0**

73.0，比 LLaVA 高出 +6.2 分，表明其在三维深度（即估计物体与相机之间的距离）和距离理解（即估计物体之间的相对距离）方面均取得了显著进步。这些结果彰显了我们的方法在细粒度视觉中心评估中的有效性。

B.3. 多模态大语言模型任务的完整结果 s

与前面报告的其他实验不同，表9聚焦于通用、知识和光学字符识别类别，并明确列出 Cambrian 评估基准中包含的各个数据集 [56]。具体而言，与原论文相同，我们将数据集分为三个主要类别，而非视觉中心类别：

• **通用**。其中包括 MME [17], GQA [22], MMBench（MMB） [38], 和 SEED-Bench（SEED） [29]。这些数据集共同通过量化、颜色识别和多领域视觉理解等任务评估感知能力与场景理解能力。

•**知识**。它包含 ScienceQA（SQA） [40],MMMU [66], MathVISTA [41], 和 AI2D [26]。这一

表9。结果 来自 Cambrian 基准测试的通用、知识、光学字符识别和以视觉为中心的基准测试结果 [5 6].

	通用				知识				OCR			视觉中心
	MME	GQA	MMB	SEED	SQA	MMMU	MathVISTA	AI2D	ChartQA	OCRBench	TextVQA	Avg
Vicuna-7B [14] & CLIP ViT-L												
LLaVA [37]	**1507.4**	**62.6**	**62.1**	**66.8**	**69.6**	**35.1**	5.7	55.2	**17.8**	**32.6**	**58.4**	50.9
JARVIS（我们的模型）	1489.3	62.4	61.0	**66.8**	69.2	33.8	**7.9**	**55.7**	17.6	**32.6**	58.2	**51.7**
Qwen2-7B [54] & CLIP ViT-L												
LLaVA [37]	1603.7	63.0	**73.3**	70.3	74.8	39.4	7.1	**65.7**	**19.1**	**32.8**	**58.7**	55.3
JARVIS（我们的方法）	**1585.5**	**63.3**	**73.3**	**70.4**	**76.3**	**41.8**	**7.9**	64.8	18.9	31.6	58.6	**56.0**
Qwen2-7B [54] 和 SigLIP2 ViT-L												
LLaVA [37]	1575.5	63.9	73.7	71.7	75.2	**43.2**	6.9	64.2	17.8	37.1	62.2	57.6
JARVIS（我们的方法）	**1612.2**	**64.9**	**75.4**	**73.5**	**76.5**	42.0	**8.5**	**66.0**	**25.8**	**40.5**	**65.1**	**59.4**

组用于衡量事实知识和学科特定推理能力，通过科学、数学和示意图理解方面的测试，评估模型所需的文本与视觉知识。

• **光学字符识别**。它涵盖 ChartQA [42], OCRBench [37], 和 TextVQA [50]。这些基准测试侧重于识别图像中的嵌入文本和数值信息并进行推理，评估光学字符识别的准确性以及以文本为基础的视觉推理能力。

作为本文的一项核心主张，JARVIS 在增强视觉推理能力的同时，在语言和知识相关任务上保持了有竞争力的表现。值得注意的是，使用 Vicuna-7B 时，JARVIS 在通用基准测试和知识任务上的表现与 LLaVA 相当，同时保持了较强的光学字符识别准确率。例如，尽管基线在 GQA 上仅以 $+0.2$ 略微领先，但我们的模型在 OCRBench 上的表现与其完全持平。扩展至 Qwen2-7B 后，无论采用 CLIP ViT-L 还是 SigLIP2 ViT-L 作为视觉编码器，JARVIS 在知识密集型任务上的表现均持续提升（例如，采用 CLIP 时在 MMMU 上提升 $+2.4$，在 MathVISTA 上提升 $+0.8$），同时在通用任务和光学字符识别任务上保持与基线持平，甚至超越基线。值得注意的是，使用 SigLIP2 时，TextVQA 上的表现从 62.2 提升至 65.1

C. 其他定性结果

在本节中，我们展示了更多定性示例，以说明我们的 JARVIS 模型在 Cambrian 基准测试四大类别中的表现。图7对比了 LLaVA [37], VIRAL [65],与我们的方法。第一行展示通用类别中的样例，涵盖广泛的视觉识别和常识查询。第二行聚焦知识密集型示例，在这些示例中，模型必须依赖事实知识或特定领域的视觉定位。第三行评估光学字符识别能力，涉及图表、文本密集型图像或数值推理。最后三行则展示视觉中心案例。总体而言，JARVIS 展现出强大的视觉定位和几何推理能力，能够正确处理反射率差异、深度排序、计数和遮挡等问题，而以往的多模态大语言模型经常在这些方面表现不佳。此外，我们的方法在其他类别中也始终优于基线和竞争对手，展现出更可靠的识别能力，以及更准确的图表和图形解读能力。这些定性示例凸显了我们的方法在各种视觉推理场景中的稳健性和通用性。

D. 局限性与影响

在这项工作中，我们通过定量和定性分析证明，JARVIS 借助自监督视觉学习增强了多模态大语言模型的视觉感知。然而，正如图中报告的失败案例所表明的那样，要实现人类水平的空间推理，我们还有很多工作要做。8例如，从第一个案例可以看出，所有多模态大语言模型都认为蜘蛛有八条腿，但实际上它只有六条。此外，我们通过实验发现，表现最佳的层即大语言模型深度的四分之一；我们在该层之上计算 $\mathcal{L}_{\text{JEPA}}$。尽管实验表明，这一选择在不同模型间具有良好的迁移性，我们仍希望未来的工作能为自监督视觉学习中最佳层的选择带来有原则的理论标准。

问：这幅画的标题是《牧羊人的朝拜》吗？



LLaVA [37]: 是 ✗VIRAL [65]: 是 ✗JARVIS（我们的）：否 ✓

问：红色边界框内的人叫 Nick Porrazzo 吗？



LLaVA [37]: 是 ✗VIRAL [65]: 是 ✗JARVIS（我们的）：否 ✓

问：谁穿着外套？



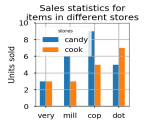
LLaVA [37]: 男人 ✓VIRAL [65]：女孩 ✗ JARVIS（我们的）：男人 ✓

问：铁丝网围栏外面有什么？



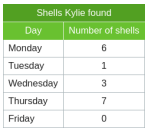
LLaVA [37]: C）一条小溪 ✗VIRAL [65]: B）地平线 ✗ JARVIS（我们的）：A）另一片田野 ✓

问：至少在一家商店中，售出少于 5 件的商品有多少件？



LLaVA [37]: 10 ✗VIRAL [65]: 10 ✗JARVIS（我们的）：2 ✓

问：星期四到星期五期间，贝壳数量的变化率是多少（每天）？



Shells Kylie found	
Day	Number of shells
Monday	6
Tuesday	1
Wednesday	3
Thursday	7
Friday	0

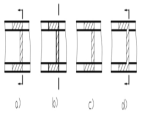
LLaVA [37]: -3 ✗VIRAL [65]: -3 ✗JARVIS（我们的）：-7 ✓

问：建筑的内部墙壁和穹顶覆盖着什么？



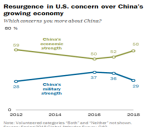
LLaVA [37]: C）马赛克 ✗VIRAL [65]: C）马赛克 ✗JARVIS（我们的）：B）壁画 ✓

问：发动机上有多少个驱动轮？



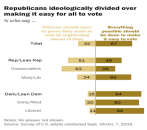
LLaVA [37]: D ✗VIRAL [65]: C ✗JARVIS（我们的方法）：B ✓

问：蓝色图表中的最小值是多少？



LLaVA [37]: 35 ✗VIRAL [65]: 25 ✗JARVIS（我们的方法）：28 ✓

问：有多少百分比的人支持让投票尽可能简单？



LLaVA [37]: 68 ✗VIRAL [65]: 76 ✗JARVIS（我们的方法）：67 ✓

问：哪个银行在举办这场杯赛？



LLaVA [37]: 富国银行 ✗VIRAL [65]: 富国银行 ✗ JARVIS（我们的）：嘉信理财 ✓

问：紧凑空间有多大？



LLaVA [37]: 2299.5 ✗VIRAL [65]:22995 ✗ JARVIS（我们的）：229.95 ✓

问：能看到多少名行人？



LLaVA [37]:1 ✗ VIRAL [65]:1 ✗ JARVIS（我们的）：2 ✓

问：水在哪里？



LLaVA [37]: B) 在你身后✗VIRAL [65]: C) 在右侧✗JARVIS（我们的方法）：A) 正前方✓

问题：这张图像中能看到鲨鱼的腹部吗？



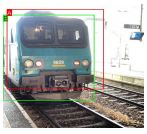
LLaVA [37]: B) 否✗VIRAL [65]: B) 否✗ JARVIS（我们的方法）：A) 是✓

问题：有曲奇饼叠放在其他曲奇饼上吗？



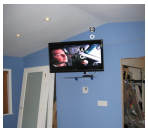
LLaVA [37]: B) 否✗VIRAL [65]: B) 否✗ JARVIS（我们的方法）：A) 是✓

问题：哪个边界框能更准确地定位并框住火车？



LLaVA [37]: A）框 A✗VIRAL [65]: A）框 A✗JARVIS（我们的方法）：B）框 B ✓

问：考虑 A 点和 B 点的表面颜色。哪一个更暗？



LLaVA [37]: C）颜色相同✗VIRAL [65]: C）颜色相同 ✗JARVIS（我们的方法）：A）A 更暗✓

问：图像中 A 点和 B 点被圈出。哪一个距离相机更近？



LLaVA [37]: B）B 更近 ✗VIRAL [65]: B）B 更近 ✗ JARVIS（我们的方法）：A）A 更近 ✓

问：图像中 A 点和 B 点被圈出。哪一个距离相机更近？



LLaVA [37]: A）A 更近 ✗VIRAL [65]: A）A 更近 ✗ JARVIS（我们的模型）：B）B 更近 ✓

问：壁灯（红框）相对于图片位于哪里？



LLaVA [37]: B）右侧 ✗VIRAL [65]: B）右侧 ✗JARVIS（我们的模型）：A）左侧✓

问：人相对于飞盘位于哪里？



LLaVA [37]: A）上方 ✗VIRAL [65]: A）上方 ✗JARVIS（我们的模型）：B）下方✓

问：毛巾（红框）和台灯（蓝框）哪个离相机更近？



LLaVA [37]: B)台灯 ✗VIRAL [65]: B)台灯 ✗JARVIS（我们的方法）：A) 毛巾 ✓

问：汽车（红框）旁边的是自行车（蓝框）还是公共汽车（绿框）？



LLaVA [37]: B)公共汽车 ✗VIRAL [65]: B)公共汽车 ✗JARVIS（我们的方法）：A) 自行车 ✓

图 7. 使用 Qwen2-7B 对多模态大语言模型的三种训练方法进行定性比较 [54]。我们展示了 Cambrian 所有类别的样本 [56]：通用（第一行）、知识（第二行）、光学字符识别（第三行）和视觉中心（最后三行）。

问：以下说法是否正确：蜘蛛有 8 条腿？



LLaVA [37]: A) 是✗VIRAL [65]: A) 是✗ JARVIS（我们的方法）：A) 是✗

问：复活节彩蛋上的装饰是平的还是凸起的？



LLaVA [37]: B) 凸起✗VIRAL [65]: B) 凸起✗J ARVIS（我们的）：B) 凸起✗

问：红色方框中的盆栽相对于马位于-置于何处？



LLaVA [37]: A) 左侧 ✗VIRAL [65]: A) 左侧 ✗JARVIS（我们的）：A) 左侧 ✗

问：哪个离相机更近，罐-（红色方框）还是垃圾桶（蓝色方框）？



LLaVA [37]: B) 垃圾桶 ✗VIRAL [65]: B) 垃圾桶 ✗ JARVIS（我们的方法）：B) 垃圾桶 ✗

图 8. 使用 Qwen2-7B 执行视觉中心任务时的一些失败案例 [54]。我们展示了来自 MMVP 的样本 [57]（第 1、2 列）、CVBench2D [56]（第 3 列）和 CVBench3D [56]（第 4 列）。