

基于事件先验的高效视觉理解视觉语言模型

Haotong Qin¹, Cheng Hu^{2*}, and Michele Magno¹

¹ Center for Project-Based Learning D-ITET, ETH Zürich, Switzerland
{ haotong.qin,michele.magno } @pbl.ee.ethz.ch

² State Key Laboratory of Industrial Control Technology, Zhejiang University, China
22032081@zju.edu.cn

Abstract. 基于大语言模型 (LLM) 的视觉-语言模型 (VLMs) 大大扩展了视觉理解能力的边界。然而, 其高计算需求阻碍了在资源受限的边缘设备上的部署。效率的关键缺失源于 VLM 需要处理密集且冗余的视觉信息。视觉输入包含与文本语义无关的重要区域, 使相关计算在推理时无效。本文介绍了一种新的基于事件先验的视觉-语言模型, 称为 EP-VLM。其核心贡献是一种利用动态事件视觉导出的运动先验来增强 VLM 效率的新机制。受人类视觉认知的启发, EP-VLM 首先使用事件数据来引导 RGB 视觉输入的逐块稀疏化, 逐步将 VLM 计算集中在视觉输入的显著区域。随后, 我们构建了一种位置保留的标记策略, 用于 VLM 架构中的视觉编码器。该策略在准确保持视觉输入的位置信息的同时, 处理事件导向的非结构化稀疏视觉输入。实验结果表明, EP-VLM 在保持与 Qwen2-VL 系列基线模型几乎无损的准确性的同时, 实现了显著的效率提升。例如, 相对于原始 Qwen2-VL-2B, EP-VLM 在 RealWorldQA 数据集上保留了原始准确性的 98 %, 并节省了 50 % FLOPs。此项工作展示了基于事件的视觉先验在提高 VLM 推理效率方面的潜力, 为在边缘创建更高效和可部署的 VLM 铺平了道路。

1 介绍

视觉-语言模型 (VLMs) [4, 38, 25, 33, 14] 通过统一视觉和语言模态彻底改变了视觉理解。这些模型通常基于大型语言模型 (LLMs) [40, 29, 13, 19, 24] 的强大推理和语言理解能力, 将其扩展到通过大规模预训练来解释和处理视觉信息。借助复杂的 LLM 骨干, VLMs 在包括图像描述 [4]、视觉问答 [3]、目标检测 [21] 和多模态内容生成 [15] 在内的广泛应用中展现出了非凡能力。它们在自动驾驶 [5]、机器人学 [35]、医疗成像 [22] 和互动 AI 助手 [12] 等领域得到应用。这一成功也推动了对其在边缘设备上的部署探索, 如设备上的视觉搜索、实时监控分析和便携设备上的人机交互。

然而, 在资源受限的边缘环境中部署 VLMs, 这些 VLMs 本身内含集成的 LLM 的计算负荷, 面临显著挑战。这些挑战包括高延迟、有限的部署可行性和过度能量消耗。现有的主流 VLM 框架, 例如 Qwen-VL [4] 和 LLaVA [25], 表现出可观的资源需求。这部分是由于其底层 LLM 的大量参数。例如, 大规模的 VLM 可能拥有数百亿的参数, 而像 LLaVA-1.5-7B 这样的小变体仍然依赖于数十亿参数的 LLM。这些模型通常需要大量的 GPU 内存和计算能力, 推理时经常超过 10GB, 每生成一个 token 则需要数百个 GFLOP。相较之下, 主流移

* Corresponding Author

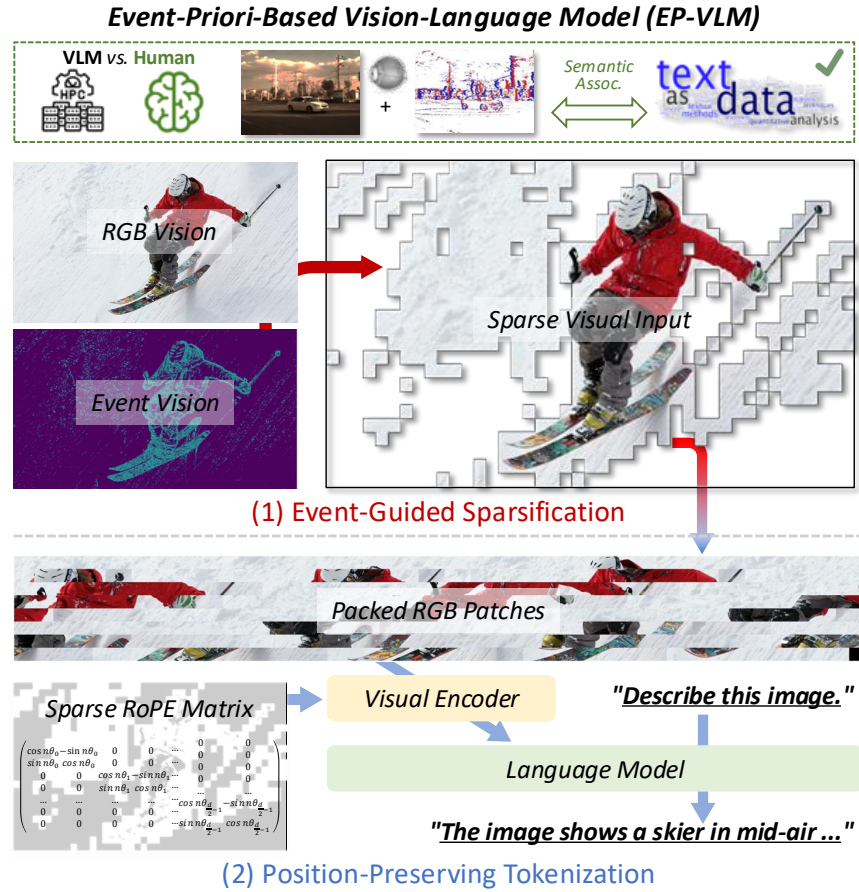


Fig. 1. EP-VLM 概述。

动设备通常具有有限的 RAM，明显较低的计算能力，以及严格的电池限制。尽管针对视觉前端和 LLM 后端的压缩方法，如剪枝 [27, 36]、量化 [32, 16, 18, 8] 和知识蒸馏 [39]，已被探索，但它们往往在复杂视觉任务上牺牲准确性。这些方法通常无法在准确性和效率之间达到帕累托最优。因此，迫切需要计算上高效的 VLM 机制，以实现可持续的边缘视觉理解。

现有的视觉-语言模型 (VLMs) 通过在推理过程中直接处理编码的输入信息，即 RGB 图像 (或视频帧) 和文本，实现视觉-语言理解。然而，由于视觉信息稀疏地分布在密集的像素阵列中，前者会不断将大量语义无关的背景信息引入到 VLM 的计算推理中。视觉输入的处理发生在视觉编码器和 LLM 骨干网络中：前者将输入编码为令牌，而后者计算和管理视觉和文本令牌。因此，视觉相关的计算构成了 VLM 总体计算负荷的重要部分。根据架构和输入分辨率，直接归因于视觉输入的计算甚至可以在图像描述生成的过程中占到总计算成本的 90% 以上。因此，效率的浪费在语义无关的视觉信息上仍然是基于 VLM 的视觉内容理解中的一个关键且常被忽视的瓶颈。

为了解决这个问题, 本文提出了一种基于事件的视觉-语言模型, 即 EP-VLM, 以实现高效的视觉理解 (图 1)。与现有的视觉-语言模型 (VLM) 依赖于昂贵的内部计算来实现跨模态理解不同, 人类利用来自生物视觉线索的运动信息作为先验知识来协助大脑实现高效的视觉认知 [6]。受到这种生物机制的启发, 我们提出了一种利用从动态事件视觉中获取的运动先验来稀疏化视觉信息的新方法, 从而在保持准确性的同时提高视觉理解的效率。首先, 我们引入与 RGB 图像采样在时间上对齐的事件视觉数据, 并将其作为先验知识来指导 RGB 视觉输入的逐片稀疏化处理。这意味着我们只处理事件数据指示为显著或变化的视觉区域。随后, 我们调整 VLM 架构中的视觉编码器以处理这种事件引导的非结构化稀疏视觉输入, 确保位置编码的正确传播和匹配, 以实现稀疏视觉信息的准确理解, 同时节省计算量。

我们的初步实验表明, 所提出的 EP-VLM 在效率提升方面相较于其基线模型 Qwen2-VL 取得了显著的改进。例如, 在 RealWorldQA 数据集上, EP-VLM 保持了基线响应准确率的 98 %, 同时其计算 FLOPs 和 MACs 减少了约 50 %。同时, 我们的定性研究表明, EP-VLM 中的事件视觉先验有效地保留了视觉输入的语义内容和空间关系。因此, EP-VLM 利用事件视觉显著提高了 VLM 的效率, 为在边缘设备上部署更强大且更易部署的 VLM 铺平了道路。

2 相关工作

视觉-语言模型。视觉-语言模型 (VLMs) 已经成为人工智能中的一种主导范式, 在弥合视觉感知与自然语言理解的差距方面取得了显著成功。最近一些强大的多模态 MLLMs, 如 LLaVA [25]、GPT-4o [19]、Gemini [37]、Claude-3.5 [2] 和 Qwen-VL [4, 38, 40], 通过视觉编码器扩展 LLMs, 使它们能够基于视觉输入执行复杂的推理和对话。这些模型通常通过投影层或适配器连接预训练的视觉编码器 (例如, ViT [10]) 与预训练的 LLM, 然后在各种视觉-语言任务上进行微调。尽管这些大型 VLMs 具备令人印象深刻的能力, 但其显著的计算和内存占用推动了对高效 VLMs 的研究。常见的方法包括知识蒸馏 [39], 即较小的学生模型从较大的教师 VLM 学习; 量化 [41, 23, 17], 降低模型权重和激活的精度; 剪枝 [27], 从模型中移除冗余的权重或结构组件; 以及设计更轻量级的架构或高效的注意机制 [43]。然而, 许多这些高效的 VLM 设计仍然难以在复杂、开放式的视觉理解任务上实现效率与性能的最佳平衡, 尤其是在与较大的模型相比时。

事件相机 [11], 也称为神经形态或动态视觉传感器 (DVS), 代表了与传统基于帧的相机的范式转变。事件相机不是以固定速率捕获图像, 而是异步记录像素级亮度变化, 生成具有微秒级时间分辨率、高动态范围 (HDR) 和低功耗的稀疏“事件”流。每个事件通常编码其空间坐标 (x, y)、时间戳 (t) 以及极性 (p , 指示亮度增加或减少)。这种独特的数据格式在捕捉快速运动、减少静态场景中的数据冗余以及在复杂光照条件下操作方面提供了固有优势。事件数据的独特性质导致了专门的处理技术和模型的发展。早期应用集中于高速跟踪 [7]、手势识别 [1] 和同步定位与地图 (SLAM) [31] 等任务, 通常采用生物启发的尖峰神经网络 (SNNs) [9] 或量身定制的深度学习架构, 如特定于事件的卷积神经网络 (CNNs) [28] 和图神经网络 (GNNs) [34] 来处理稀疏、异步事件流。最近, 对将事件数据与大规模感知和推理系统集成的兴趣日益增长, 包括大型语言模型。例如, EventGPT [26] 是将 LLM 与事件流理解相结合的

开创性尝试，使预训练的 LLM 能够理解基于事件的场景。Yu 等人探讨了基于 LLM 的纯零样本事件识别 [42]。

3 方法

3.1 预备知识

本节概述了基于多模态大型语言模型 (LLMs) 的视觉语言模型 (VLMs) 的基本架构，并建立相关的符号表示。这样的 VLMs 通常整合了一个预训练的 LLM 和一个专门的视觉处理模块。这种协同组合旨在有效利用每个组件的独特能力：视觉模型用于解释图像，而 LLM 用于理解文本指令并在两种模态之间执行复杂推理。为了阐明这些模型的操作流程，我们将在处理图像和文本输入时参考 Qwen2-VL 架构，这是此类 VLM 的一个当代理论例子。

视觉处理管道的初始阶段涉及对输入图像进行标准化。一个任意的输入图像首先被重塑为一个预定义的分辨率，记作 $\mathbb{R}^{H \times W \times C}$ ，其中 H 、 W 和 C 分别代表高度、宽度和颜色通道的数量。此标准化格式确保了与后续预训练视觉编码器的兼容性。为了与可能处理视频输入（帧序列）的架构保持一致，图像输入通常具有一个显式的时间维度 T 。对于静态图像，这可以被概念化为长度为 $T = 1$ 的序列，或者如果下游架构需要一个固定长度的时间输入，图像特征可能会被复制 T 次。这导致一个视觉输入张量 $\mathbf{X}_v \in \mathbb{R}^{H \times W \times C \times T}$ 。在调整大小之后，图像被分解为一系列不重叠的补丁。这种基于补丁的表示是一种常用策略，特别是针对基于 Transformer 的架构以实现序列处理。具体而言，具有维度 $H \times W$ 的图像被分成 $N_v = \left\lfloor \frac{H}{p} \right\rfloor \times \left\lfloor \frac{W}{p} \right\rfloor$ 个独立的补丁，其中每个补丁由 $p \times p$ 个像素组成。

处理后的视觉输入，通常是从 $\mathbf{X}_v$ 派生的补丁序列，然后被输入到一个预训练的视觉编码器中，记为 $g(\cdot)$ 。在 Qwen2-VL 模型中，这个角色由视觉转换器 (ViT) 完成。一个关键组成部分是引入位置嵌入，以使 ViT 能够理解这些补丁的空间排列。Qwen2-VL 使用一种二维旋转位置嵌入 (RoPE)，即 $\mathbf{P}_v^{2D}$ ，它应用于 ViT 内的补丁嵌入。这一机制使得模型能够保留和利用关于不同图像区域相对位置的信息。视觉编码器的输出是一组丰富的视觉特征，即 $\mathbf{Z}_v$ ：

$$\mathbf{Z}_v = g\left(\text{Patches}(\mathbf{X}_v), \mathbf{P}_v^{2D}\right). \quad (1)$$

二维 RoPE $\mathbf{P}_v^{2D}$ 通过应用旋转矩阵来注入空间意识，这些矩阵由补丁坐标 (i, j) 参数化。对于一个 d 维嵌入向量，旋转矩阵在不同维度对上操作。更具体地说，坐标为 (i, j) 的补丁的旋转矩阵 $\mathbf{R}_{(i,j)}^{2D}$ 被构造成一个块对角矩阵。这个矩阵由 $d/4$ 个相同的 4×4 块的直接和 ($\oplus$) 组成，其中每个块基于 i 对其前两个分量进行旋转，并基于 j 对其后两个分量进行旋转：

$$\mathbf{R}_{(i,j)}^{2D} = \bigoplus_{m=1}^{d/4} \begin{bmatrix} \cos(i\theta_m) & -\sin(i\theta_m) & 0 & 0 \\ \sin(i\theta_m) & \cos(i\theta_m) & 0 & 0 \\ 0 & 0 & \cos(j\theta_m) & -\sin(j\theta_m) \\ 0 & 0 & \sin(j\theta_m) & \cos(j\theta_m) \end{bmatrix}, \quad \text{where } \theta_m = 10000^{-2m/d}. \quad (2)$$

这种 RoPE 设计由于其旋转等变性固有地保留了相对位置关系，同时在计算上也很高效。在视觉编码器提取特征后，生成的视觉特征 $\mathbf{Z}_v$ 由多层感知器 (MLP)

处理。这个 MLP 的主要功能是将视觉特征投影到一个新的嵌入空间，使其与 LLM 的输入要求兼容。这个投影通常使用投影矩阵 $\mathbf{W}$ 所定义的线性变换来实现，常常之后还会有非线性激活函数。这一阶段还可以用于减少视觉标记的维度或整合信息。原文表明，目的是通过将相邻 $M_v \times M_v$ 组的视觉标记信息（如果 $\mathbf{Z}_v$ 保持可以映射到这些组的网格状结构）压缩成一个更紧凑的表示来减少视觉输入的有效标记大小。合并大小 M_v 因此指的是这种信息整合的概念性分组。这个投影阶段的输出是一个视觉嵌入标记序列， $\mathbf{H}_v$ ：

$$\mathbf{H}_v = \text{MLP}_{\mathbf{W}}(\mathbf{Z}_v), \quad (3)$$

，其中 $\text{MLP}_{\mathbf{W}}(\cdot)$ 表示 MLP 运算，突出地包含投影矩阵 $\mathbf{W}$ 。为了简化，如果仅考虑线性投影方面，这可以近似为 $\mathbf{H}_v \approx \mathbf{W}\mathbf{Z}_v$ （假设 $\mathbf{Z}_v$ 适于进行矩阵乘法）。处理过的视觉嵌入令牌序列 $\mathbf{H}_v$ 随后准备与文本信息集成。同时输入文本查询 $\mathbf{X}_q$ 通过 LLM 的输入嵌入层（通常涉及标记化和嵌入查找）转换为语言嵌入令牌序列 $\mathbf{H}_q$ 。这两组嵌入 $\mathbf{H}_v$ 和 $\mathbf{H}_q$ 连接以形成一个统一的多模态序列。然后将该组合序列输入到预训练的 LLM 主干网络中，表示为 $f(\cdot)$ 。LLM 处理这种融合表示以进行跨模态推理并生成最终的文本答案 $\mathbf{X}_a$ ：

$$\mathbf{X}_a = f(\text{concat}(\mathbf{H}_v, \mathbf{H}_q)). \quad (4)$$

。因此，LLM 被赋予理解交织的视觉和语言信息，以产生一个在语境上相关且连贯的回应。

描述的 VLM 架构关键依赖于两个主要阶段来连接和解释视觉与语言数据：(1) 视觉编码和投影管道，它将原始视觉输入从密集的 RGB 像素表示抽象为更压缩的嵌入令牌序列。(2) 计算需求高的 LLM 骨干负责理解这些组合输入令牌并生成所需输出。然而，一个显著的挑战来自许多 VLM 采用的标准逐块视觉编码管道。观察到这种方法常常在视觉输入流中引入显著的冗余。自然图像通常包含大量语义稀疏或缺乏与当前任务相关信息的区域。因此，视觉编码器和 LLM 骨干在处理来自这些无信息或冗余图像区域的视觉令牌时耗费了相当大的计算资源。

3.2 视觉输入的事件引导稀疏化

受到人类视觉系统效率的启发，这项工作引入了基于事件的数据，作为传统 RGB 图像的补充模态。我们的目标是通过稀疏化语义内容较低的区域来预处理视觉输入，从而减少冗余。

标准 RGB 图像提供了场景的密集、丰富多彩的表现。相比之下，由动态视觉传感器（DVS）捕获的事件数据记录了亮度的异步局部变化。这一固有特性使得事件数据在空间上是稀疏的，并且比其密集的 RGB 对应数据紧凑得多。因此，我们为每个 RGB 帧结合了相应的事件视觉数据，以作为视觉先验，特别是运动的指示。从直观上看，表现出运动的地点与关联的文本查询具有更大的相关性概率。相反，静态背景区域通常特征在于信息含量较低及最小的运动，一般在视觉输入中不那么重要或可以更紧凑地表示。利用这一原理，我们在视觉编码阶段之前使用事件先验信息来稀疏化 RGB 输入。这种有针对性的稀疏化旨在提高随后的视觉处理任务的计算效率。

具体方法如下：首先，对事件数据进行处理，使其在时间和空间上与相应的 RGB 图像对齐。事件数据点在一个定义的时间窗口内积累形成一个二维图，然

后调整大小以匹配 RGB 输入 $\mathbf{X}_v$ 的尺寸 ($W \times H$)，产生事件基础的表示 $\mathbf{E}_v \in \mathbb{R}^{W \times H}$ 。随后，将 $\mathbf{E}_v$ 划分为不重叠的块，保持与 RGB 图像采用的块策略一致。设 p 为这些方块的边长。对于每个块，我们计算其 ℓ_1 范数以量化运动强度。这样就形成了矩阵 $\mathbf{S}_v^E \in \mathbb{R}^{\lfloor H/p \rfloor \times \lfloor W/p \rfloor}$ ，其中每个元素 (u, v) 的计算如下：

$$\mathbf{S}_{v,uv}^E = \sum_{(x,y) \in \text{Patch}_{uv}(\mathbf{E}_v)} |\mathbf{E}_v(x,y)| \quad (5)$$

，其中 $\text{Patch}_{uv}(\mathbf{E}_v)$ 指的是 $\mathbf{E}_v$ 中的第 u 行第 v 列的图像块， $\mathbf{E}_v(x,y)$ 是该块内像素坐标 (x,y) 处的值。在 $\mathbf{S}_v^E$ 中更高的值表示在相应块中运动发生的更高频率。

一个事件优先的视觉掩码 $\mathbf{M}_v^E \in \{0,1\}^{\lfloor H/p \rfloor \times \lfloor W/p \rfloor}$ 随后通过指定的分位数阈值 $\tau \in [0,1]$ 得出：

$$\mathbf{M}_{v,uv}^E = \mathbb{1}_{(\mathbf{S}_{v,uv}^E \geq Q_{1-\tau}(\mathbf{S}_v^E))} \quad (6)$$

，其中 $Q_{1-\tau}(\mathbf{S}_v^E)$ 表示 $(1-\tau)$ 分位数的所有运动强度值在 $\mathbf{S}_v^E$ 中的位置， $\mathbb{1}_{(\cdot)}$ 是指示函数。 $\mathbf{M}_v^E$ 中的一个值为 1 表示相应区域的运动强度在所有区域中排名在前 τ 的分数（例如，如果 $\tau = 0.5$ ，即前 50%），因此将其标记为保留。

这种方法的有效性通过使用事件优先的视觉掩码对 RGB 图像进行掩蔽来证明，其中 $\tau = 0.5$ （即保留运动最大的 50% 的补丁）。如图 1 所示，即使根据此掩码遮挡了 RGB 图像的一半补丁，图像的整体语义内容仍然大致保持完整。可以通过利用基于事件的数据作为显著性的动态指导，有效减轻冗余，而不会大幅丢失关键的语义信息，从而为更高效的视觉理解系统铺平道路。

3.3 用于视觉编码器的保位分词

在获得事件优先的视觉掩码 $\mathbf{M}_v^E$ （如前一部分所述）之后，直接在视觉语言模型（VLMs）中应用它以提高计算效率会遇到某些挑战。这个复杂性产生的原因是，因为有效的视觉理解关键依赖于保持输入图像的固有空间结构。由事件数据捕获的运动模式在各种视觉输入中本质上是可变和动态的，导致掩码稀疏且不规则。视觉编码器通常需要根据视觉输入的完整尺寸计算的位置嵌入，以编码每个小块或标记的空间位置。简单的直接连接（打包）只有未被掩盖的小块的方法会破坏它们原本的空间关系，从而扭曲编码器可用的位置信息。相反，为保持固定输入结构而保留被掩盖的小块（例如，作为零向量），则会抵消从稀疏化中获得的计算节省。

为了解决这些挑战，我们提出了一种针对非结构化稀疏视觉输入的保持位置的推理策略。这一策略首先根据视觉掩码 $\mathbf{M}_v^E$ 提供的指导，从原始稠密输入补丁序列 $\mathbf{X}_v = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N\}$ 中有选择地保留视觉补丁。与 $\mathbf{M}_v^E$ 中零值条目对应的补丁被丢弃。这个压缩过程可以形式化地表示为：

$$\tilde{\mathbf{X}}_v = \text{Pack}(\mathbf{X}_v, \mathbf{M}_v^E) \quad (7)$$

，其中 $\text{Pack}(\cdot, \cdot)$ 表示从 $\mathbf{X}_v$ 中选择和连接与所对应的掩码值在 $\mathbf{M}_v^E$ 中为 1 的补丁的函数。张量 $\tilde{\mathbf{X}}_v$ 是结果的打包（稀疏）视觉输入序列，包含 $N' < N$ 个补丁。

关键是，为了保持稀疏输入 $\tilde{\mathbf{X}}_v$ 的位置信息完整性，首先根据原始的、密集的输入维度计算旋转位置嵌入（RoPE），记为 $\mathbf{R}^{2D}$ ，从而得到针对所有 N 原始

贴片位置的一组位置嵌入 $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N\}$ 。然后使用相同的视觉掩码 $\mathbf{M}_v^E$ 选择性地打包这些全分辨率的位置嵌入：

$$\tilde{\mathbf{R}}^{2D} = \text{Pack}(\mathbf{R}^{2D}, \mathbf{M}_v^E) \quad (8)$$

生成的打包 RoPE 嵌入 $\tilde{\mathbf{R}}^{2D}$ (一系列 N' 嵌入) 直接对应于打包的视觉输入序列 $\tilde{\mathbf{X}}_v$ 中的贴片。这确保了每个保留的贴片 $\tilde{\mathbf{p}}_j \in \tilde{\mathbf{X}}_v$ 都与其正确的原始位置编码 $\tilde{\mathbf{r}}_j \in \tilde{\mathbf{R}}^{2D}$ 相关联。

压缩的视觉输入张量 $\tilde{\mathbf{X}}_v$ 及其对齐的压缩 RoPE 嵌入 $\tilde{\mathbf{R}}^{2D}$ 随后由视觉编码器处理，通常是一个 ViT。标准 Transformer 层中的操作可以抽象如下。如果 $\text{Patches}(\mathbf{X}_v)$ 表示从原始图像或其对应的完整补丁序列 $\mathbf{X}_v$ 中派生的所有补丁标记序列，而 $\mathbf{P}_v^{2D}$ 表示这些补丁的完整位置嵌入集合（例如，RoPE $\mathbf{R}^{2D}$ ），则基本的处理步骤涉及一个函数 g 。这个函数 g 通常包括自注意机制，其中位置信息 $\mathbf{P}_v^{2D}$ 与补丁特征相结合：

$$\mathbf{Z}_v = g\left(\text{Patches}(\mathbf{X}_v), \mathbf{P}_v^{2D}\right) \quad (9)$$

。随后，表示 $\mathbf{Z}_v$ 进一步被一个多层感知机 (MLP) 利用权重 $\mathbf{W}$ 进行变换，这是 Transformer 块的另一个标准组成部分：

$$\mathbf{H}_v = \text{MLP}_{\mathbf{W}}(\mathbf{Z}_v) \quad (10)$$

。在我们提出的高效推理方法中，这些操作 g 和 $\text{MLP}_{\mathbf{W}}$ 有效地在稀疏化输入上执行。这意味着 g 的输入对应于 $\text{Patches}(\mathbf{X}_v)$ 成为我们的压缩序列 $\tilde{\mathbf{X}}_v$ ，而位置输入 $\mathbf{P}_v^{2D}$ 成为压缩的 RoPE 嵌入 $\tilde{\mathbf{R}}^{2D}$ 。这种方法确保了计算主要用于选定的显著标记，同时仍然利用了 $\tilde{\mathbf{R}}^{2D}$ 中固有的准确保留的位置信息。因此，生成的张量 $\mathbf{H}_v$ 包含精炼的视觉特征，这些特征是从稀疏的活动标记集中高效构建的。

这种位置保留推理机制使视觉编码器能够处理显著减少的标记数量 (N' 而不是 N)，从而提高计算效率，同时保持对视觉内容的准确空间理解。将这种策略应用于打包的视觉输入，显著提高了 ViT 组件和随后 LLM 阶段的推理效率。这种改进是实现整体高效 VLM 运行的关键因素，尤其是在资源受限环境或对延迟敏感的应用中。

4 实验

本节对所提出的 EP-VLM 框架进行了全面评估。量化结果展示了该框架对模型准确性和计算效率的影响，而定性案例研究则阐明了事件引导视觉稀疏化的好处。

4.1 量化结果

我们在 RealWorldQA 基准上评估了 Qwen2-VL-2B/7B [4] 及其 EP-VLM 变体，使用了基于事件优先的稀疏性 τ ，稀疏度为 0.3、0.5 和 0.7。如表 1 所示，效

Table 1. Qwen2-VL 在 RealWorldQA 基准上的性能比较。

Model	Sparsity (τ)	Params (B)	Acc. (%)	FLOPs (T)	MACs (T)
Qwen2-VL-2B 0		2.21	62.9	14.7	7.4
EP-VLM	0.3	2.21	62.7 (-0.3)	10.3 -29.9%	5.2 -29.7%
EP-VLM	0.5	2.21	61.4 (-2.3)	7.4 -49.7%	3.7 -50.0%
EP-VLM	0.7	2.21	59.1 (-6.0)	4.5 -69.4%	2.2 -70.3%
Qwen2-VL-7B 0		8.29	70.1	31.1	15.5
EP-VLM	0.3	8.29	67.3 (-2.8)	24.7 -20.1%	12.3 -20.6%
EP-VLM	0.5	8.29	67.2 (-2.9)	17.9 -42.4%	8.9 -42.6%
EP-VLM	0.7	8.29	64.8 (-5.3)	11.1 -64.3%	5.5 -64.5%

**Fig. 2.** 情况 1: 海滩场景描述。红色 代表主体 (人/动物) 的描述, 蓝色 代表背景的描述。

率指标 [30] 包括浮点运算 (FLOPs)、乘累加运算 (MACs) 和参数数量, 精度反映了视觉问答能力。

表格 1 中的结果揭示了三个关键见解。首先, 对于每种类型的模型, 参数数量在各个配置中保持稳定, 证实 EP-VLM 作为一种不需要架构修改的输入调节机制。其次, 增加稀疏性会引发准确性与效率的权衡, 50 % 稀疏性只显示了 1.5 % 的准确性下降 (从 62.9 % 下降到 61.4 %), 同时处理了一半的视觉标记。第三, 理论上的 FLOPs/MACs 增加显然源于当前实现中稀疏数据索引的开销, 而由于 LLM 主体中标记处理的减少, 实际推理延迟和能量消耗降低。这种 50 % 稀疏配置显示了接近帕累托最佳的性能, 验证了我们的假设, 即事件数据为语义视觉信息提供了有效的先验。

为了展示 EP-VLM 利用事件视觉实现高效和准确视觉理解的能力, 我们研究了从广泛使用的基准中提取的定性案例研究。第一个案例展示了在由事件数据引导的 50 个 % 稀疏的复杂海滩场景中的演示。第二个案例对比了树叶中小鸟的标准 ImageNet 照片与由 EP-VLM 处理的对应的 n-ImageNet [20] 事件表示。这些例子共同展示了动态视觉先验如何能够选择性地突出显著内容, 同时抑制冗余的背景信息。



Fig. 3. 案例 2：鸟类场景的描述。

首先看 Fig. 2 中的海滩场景，当通过稀疏的事件驱动输入进行指导时，EP-VLM 在语义精确性和关系推理上表现出显著的改进。在 RGB 演示中（左侧），模型正确识别了主要角色，即一个女人和一只狗，以及广泛的海岸线背景、波浪和发光的天空。然而，描述仍然有些泛泛且结构松散，对对象之间的关系强调有限。相反，事件驱动的稀疏输入（右侧）使 EP-VLM 能够关注重要的运动线索和边缘动能，生成更丰富、更有结构的标题。模型不仅识别出坐在沙子上的女人和以击掌姿势出现的狗，还准确编码了“女人在狗右边”的空间关系。此外，EP-VLM 捕捉到细致入微的配件细节（例如，女人的格子衬衫和手表），并将这些与狗的鞍具和牵绳在情境上相关联，反映了更深的场景理解。通过利用异步事件流，模型过滤掉背景噪声，如波浪和天空照明，并优先考虑动态元素，从而生成简明而全面的实体、属性及其空间互动的描述。

同样地，在图 3 的鸟类场景中，密集的 RGB 和稀疏的事件引导之间的对比同样鲜明。在标准的 ImageNet 输入下，EP-VLM 生成了一个正确但相对平淡的字幕，即“在高大的绿草中间的一只小雀形目鸟”，仅仅注意到了基本的羽毛对比和栖息地。然而，当提供 n-ImageNet 事件数据时，模型的描述变得更加生动和富有关系细微差别：它强调了鸟的栖息动作，其喙向附近草叶倾斜的方向，以及光线在其橙棕色羽毛上的交错的效果。事件驱动的表达强调微小的动作，如头部倾斜和翼部调整，使 EP-VLM 能够推断行为背景（“栖息在潮湿地带草丛中央的草地中”）。这个案例强调了 EP-VLM 将稀疏的时间线索转化为精确的空间和语义关系的能力，反映了人类视觉系统优先考虑运动以解释场景的特点。

本文提出了 EP-VLM，一种新颖的基于事件先验的视觉语言模型，在保持准确性的同时显著提高了计算效率。受人类视觉认知的启发，EP-VLM 利用事件数据中的运动先验来动态稀疏化 RGB 输入，将计算集中在动态视觉传感器识别的语义显著区域。关键的是，我们的保持位置的标记化策略使视觉编码器能够处理这种非结构化、稀疏输入，同时通过打包的旋转位置嵌入保持准确的空间关系。在 Qwen2-VL 基线上进行的实验表明，EP-VLM 在 RealWorldQA 基准上的浮点运算量减少了 50%，而准确率保持在 98%，验证了事件引导的稀疏化能够有效消除冗余的视觉计算而不影响理解。这项工作将基于事件的先验确立为一种强大的范式，为资源受限环境中的高效多模态智能开辟了新的途径。

References

1. Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., et al.: A low power, fully event-based gesture recognition system. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7243–7252 (2017)
2. Anthropic: Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet> (2024), <https://www.anthropic.com/news/claude-3-5-sonnet>
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Baumann, N., Hu, C., Sivasothilingam, P., Qin, H., Xie, L., Magno, M., Benini, L.: Enhancing autonomous driving systems with on-board deployed large language models. arXiv preprint arXiv:2504.11514 (2025)
6. Berry, M.J., Brivanlou, I.H., Jordan, T.A., Meister, M.: Anticipation of moving stimuli by the retina. *Nature* **398**(6725), 334–338 (1999)
7. Chamorro Hernández, W.O., Andrade-Cetto, J., Solà Ortega, J.: High-speed event camera tracking. In: Proceedings of the The 31st British Machine Vision Virtual Conference. pp. 1–12 (2020)
8. Chen, H., Lv, C., Ding, L., Qin, H., Zhou, X., Ding, Y., Liu, X., Zhang, M., Guo, J., Liu, X., et al.: Db-llm: Accurate dual-binarization for efficient llms. arXiv preprint arXiv:2402.11960 (2024)
9. Cordone, L., Miramond, B., Thierion, P.: Object detection with spiking neural networks on automotive event data. In: 2022 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**(1), 154–180 (2020)
12. Guan, Y., Wang, D., Chu, Z., Wang, S., Ni, F., Song, R., Li, L., Gu, J., Zhuang, C.: Intelligent virtual assistants with llm-based process automation. arXiv preprint arXiv:2312.06677 (2023)
13. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
14. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
15. He, Y., Liu, Z., Chen, J., Tian, Z., Liu, H., Chi, X., Liu, R., Yuan, R., Xing, Y., Wang, W., et al.: Llms meet multimodal generation and editing: A survey. arXiv preprint arXiv:2405.19334 (2024)
16. Huang, W., Liu, Y., Qin, H., Li, Y., Zhang, S., Liu, X., Magno, M., Qi, X.: Billm: Pushing the limit of post-training quantization for llms. arXiv preprint arXiv:2402.04291 (2024)

17. Huang, W., Qin, H., Liu, Y., Li, Y., Liu, X., Benini, L., Magno, M., Qi, X.: Slim-llm: Saliency-driven mixed-precision quantization for large language models. arXiv preprint arXiv:2405.14917 (2024)
18. Huang, W., Zheng, X., Ma, X., Qin, H., Lv, C., Chen, H., Luo, J., Qi, X., Liu, X., Magno, M.: An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence* **2**(1), 36 (2024)
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. Kim, J., Bae, J., Park, G., Zhang, D., Kim, Y.M.: N-imagenet: Towards robust, fine-grained object recognition with event cameras. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 2146–2156 (2021)
21. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639 (2022)
22. Li, J., Guan, Z., Wang, J., Cheung, C.Y., Zheng, Y., Lim, L.L., Lim, C.C., Ruamvi-boonsuk, P., Raman, R., Corsino, L., et al.: Integrated image-based deep learning and language models for primary diabetes care. *Nature medicine* **30**(10), 2886–2896 (2024)
23. Li, Z., Yan, X., Zhang, T., Qin, H., Xie, D., Tian, J., Kong, L., Zhang, Y., Yang, X., et al.: Arb-llm: Alternating refined binarizations for large language models. arXiv preprint arXiv:2410.03129 (2024)
24. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
26. Liu, S., Li, J., Zhao, G., Zhang, Y., Meng, X., Yu, F.R., Ji, X., Li, M.: Event-gpt: Event stream understanding with multimodal large language models. arXiv preprint arXiv:2412.00832 (2024)
27. Ma, X., Fang, G., Wang, X.: Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems* **36**, 21702–21720 (2023)
28. Messikommer, N., Gehrig, D., Loquercio, A., Scaramuzza, D.: Event-based asynchronous sparse convolutional networks. In: *European Conference on Computer Vision*. pp. 415–431. Springer (2020)
29. Meta, A.: The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on 4(7), 2025 (2025)
30. MrYxJ: calcflops: a flops and params calculate tool for neural networks. <https://github.com/MrYxJ/calculate-flops.pytorch> (2024), <https://github.com/MrYxJ/calculate-flops.pytorch>
31. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International journal of robotics research* **36**(2), 142–149 (2017)
32. Qin, H., Ma, X., Zheng, X., Li, X., Zhang, Y., Liu, S., Luo, J., Liu, X., Magno, M.: Accurate lora-finetuning quantization of llms via information retention. arXiv preprint arXiv:2402.05445 (2024)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

- natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
34. Schaefer, S., Gehrig, D., Scaramuzza, D.: Aegnn: Asynchronous event-based graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12371–12381 (2022)
 35. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2998–3009 (2023)
 36. Sun, M., Liu, Z., Bair, A., Kolter, J.Z.: A simple and effective pruning approach for large language models. arXiv preprint arXiv:2306.11695 (2023)
 37. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
 38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
 39. Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., Zhou, T.: A survey on knowledge distillation of large language models. arXiv preprint arXiv:2402.13116 (2024)
 40. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
 41. Yang, G., He, C., Guo, J., Wu, J., Ding, Y., Liu, A., Qin, H., Ji, P., Liu, X.: Llm-cbench: Benchmarking large language model compression for efficient deployment. arXiv preprint arXiv:2410.21352 (2024)
 42. Yu, Z., Qu, Q., Chen, X., Wang, C.: Can large language models grasp event signals? exploring pure zero-shot event-based recognition. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
 43. Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., Liu, X., Wu, J., Huang, L.: Tinyllava: A framework of small-scale large multimodal models. arXiv preprint arXiv:2402.14289 (2024)