LightVLA

The Better You Learn, The Smarter You Prune: Towards Efficient Vision-language-action Models via Differentiable Token Pruning

Titong Jiang[1,2]*, Xuefeng Jiang[1,3]*, Yuan Ma[1]†, Xin Wen[1], Bailin Li[1],
Kun Zhan[1], Peng Jia[1], Yahui Liu[2], Sheng Sun[3], and Xianpeng Lang[1]‡

> 原论文地址：The Better You Learn, The Smarter You Prune: Towards Efficient Vision-language-action Models via Differentiable Token Pruning

题目：学得越好，剪枝越精：通过可微分令牌剪枝实现高效视觉 - 语言 - 动作模型

蒋 titong[1・2]、*蒋雪峰*[1・3]、马源[1†]、文鑫[1]、李柏霖[1]、詹坤[1]、贾鹏[1]、刘亚辉[2]、孙胜[3]、郎咸鹏[1‡]

摘要：本文提出 LightVLA—— 一种简单且高效的可微分令牌剪枝框架，适用于视觉 - 语言 - 动作（Vision-Language-Action, VLA）模型。尽管 VLA 模型在执行现实世界机器人任务时展现出了出色的能力，但其在资源受限平台上的部署常受限于大规模视觉令牌上繁重的基于注意力的计算，这成为了关键瓶颈。LightVLA 通过**自适应、以性能为导向的视觉令牌剪枝**解决这一挑战：它生成动态查询以评估视觉令牌的重要性，并采用 Gumbel softmax 实现可微分令牌选择。通过微调，LightVLA 学会保留信息最丰富的视觉令牌，同时剪去对任务执行无贡献的令牌，从而在提升效率的同时改善性能。值得注意的是，LightVLA 无需启发式 “超参魔法值”，且不引入额外可训练参数，因此可与主流推理框架兼容。实验结果表明，在 LIBERO 基准测试的各类任务中，LightVLA 的性能优于不同的 VLA 模型及现有令牌剪枝方法，在大幅降低计算开销的同时实现了更高的成功率。具体而言，LightVLA 将浮点运算次数（FLOPs）降低 59.1%，延迟降低 38.2%，同时任务成功率提升 2.6%。此外，本文还研究了基于可学习查询的令牌剪枝方法 LightVLA*—— 该方法引入额外可训练参数，同样取得了令人满意的性能。本文研究表明，当 VLA 模型追求最优性能时，LightVLA 能从 “以性能为导向” 的角度自主学习剪枝令牌。据我们所知，LightVLA 是首个将自适应视觉令牌剪枝应用于 VLA 任务、并以 “效率与性能兼顾” 为目标的工作，为构建更高效、更强大且更实用的实时机器人系统迈出了重要一步。项目网站：https://liauto-research.github.io/LightVLA。

1 引言

从大规模工业操作到个人医疗健康与休闲活动，机器人技术正重塑人类生活的几乎每一个方面。近年来，随着人工智能（AI）被引入机器人领域，机器人技术迎来了最新的技术飞跃 ——**具身智能（embodied intelligence）** 的兴起，而这一飞跃的核心驱动力正是视觉 - 语言 - 动作（VLA）模型的出现。VLA 模型可视为一类大型视觉 - 语言模型（Vision-Language Models, VLMs），其能直接将视觉信息与语言指令转化为可执行的动作策略。借助从大型语言模型（LLMs）继承的通用知识与推理能力，VLA 模型在应对复杂的机器人推理、规划与操作任务方面展现出了变革性潜力 [1]-[8]。

遗憾的是，VLA 模型的成功伴随着高昂的计算复杂度。典型的 VLA 模型通常包含参数规模达数十亿的 LLM，其基于注意力的高昂计算成本与较大的前向传播延迟，使其难以在边缘设备（如家用机器

人 [9]、自动驾驶车辆 [10],[11]）等资源受限系统上实现实时应用。因此，VLA 模型的加速技术对于提升其效率与实用性至关重要 [12]。

现有研究已探索了多种针对 VLM 与 VLA 模型的加速方法，包括模型量化、层跳过 [12]、令牌剪枝 [13]-[20] 及轻量级模型设计 [21]-[23]。在这些方法中，**视觉令牌剪枝**尤为值得关注 —— 因为 VLA 模型的输入令牌中，绝大多数是视觉令牌。由于视觉模态本身具有稀疏性 [14]，许多视觉令牌仅携带少量信息或冗余信息，这为 VLA 模型带来了显著却不必要的计算负担。同时，尽管视觉令牌剪枝在 VLM 领域已得到较广泛研究 [13]-[15]，但近年研究 [24]-[26] 表明，这些方法迁移到 VLA 模型时性能表现不佳：原因在于 VLM 关注全局语义，而特定机器人任务更依赖局部语义。因此，面向 VLA 模型的视觉令牌剪枝具有巨大潜力，但相关探索仍较为有限。

学界普遍认为，VLA 模型加速过程中存在 “效率与泛化性能” 的权衡关系。因此，以往的视觉令牌剪枝方法往往优先追求效率，而容忍一定程度的性能损失。例如，EfficientVLA [12] 首先将保留令牌数量设为超参数，再提出多种方法以最小化令牌减少导致的性能下降。然而，本文认为 “效率与性能并非本质矛盾”。需注意的是，视觉输入的稀疏性不仅导致计算效率低下，还会通过引入噪声、分散注意力而损害性能。基于此，本文提出：通过消除视觉输入的稀疏性，可同时优化效率与性能，打破 VLA 模型中 “效率 - 性能” 的权衡困境。

具体而言，本文研究了 VLA 模型中视觉令牌的内在稀疏性，并提出**LightVLA**—— 一种以性能为导向的可微分视觉令牌剪枝框架，用于实现高效 VLA 模型。为评估视觉令牌对任务执行的重要性，LightVLA 通过视觉令牌与任务指令令牌间的交叉注意力生成动态查询；随后，每个查询采用 Gumbel-softmax [27] 技术以可微分方式选择有用令牌。基于微调范式，本文以 OpenVLA-OFT [7] 为基础模型，训练 LightVLA 区分并仅保留对整体性能有贡献的信息性视觉令牌。如图 1 所示，在 LIBERO 基准测试中，LightVLA 以显著更少的视觉令牌实现了当前最优性能。与 OpenVLA-OFT 相比，LightVLA 不仅将总 FLOPs 降低 59.1%，还将任务成功率提升 2.6%—— 这充分证明 “效率与性能可作为兼顾目标同时实现”。

为进一步填补 VLA 模型视觉令牌剪枝领域的空白，本文在 “讨论” 章节（第 5 节）提出 LightVLA* —— 该方法是 LightVLA 的高效变体，通过引入 “可学习查询” 作为额外可训练参数，引导模型选择信息性令牌。综上，本文的贡献总结如下：

1. 实证表明，VLA 模型的性能与效率可同时优化；
2. 提出 LightVLA—— 一种以性能为导向的可微分视觉令牌剪枝框架，适用于 VLA 模型；
3. 在 LIBERO 基准测试上的全面实验表明，与基础模型及其他现有模型相比，LightVLA 实现了当前最优的性能与效率；
4. 为进一步填补 VLA 模型令牌剪枝领域的空白，提出 LightVLA*—— 一种基于可学习查询的令牌剪枝初步探索，该方法同样提升了性能与效率。

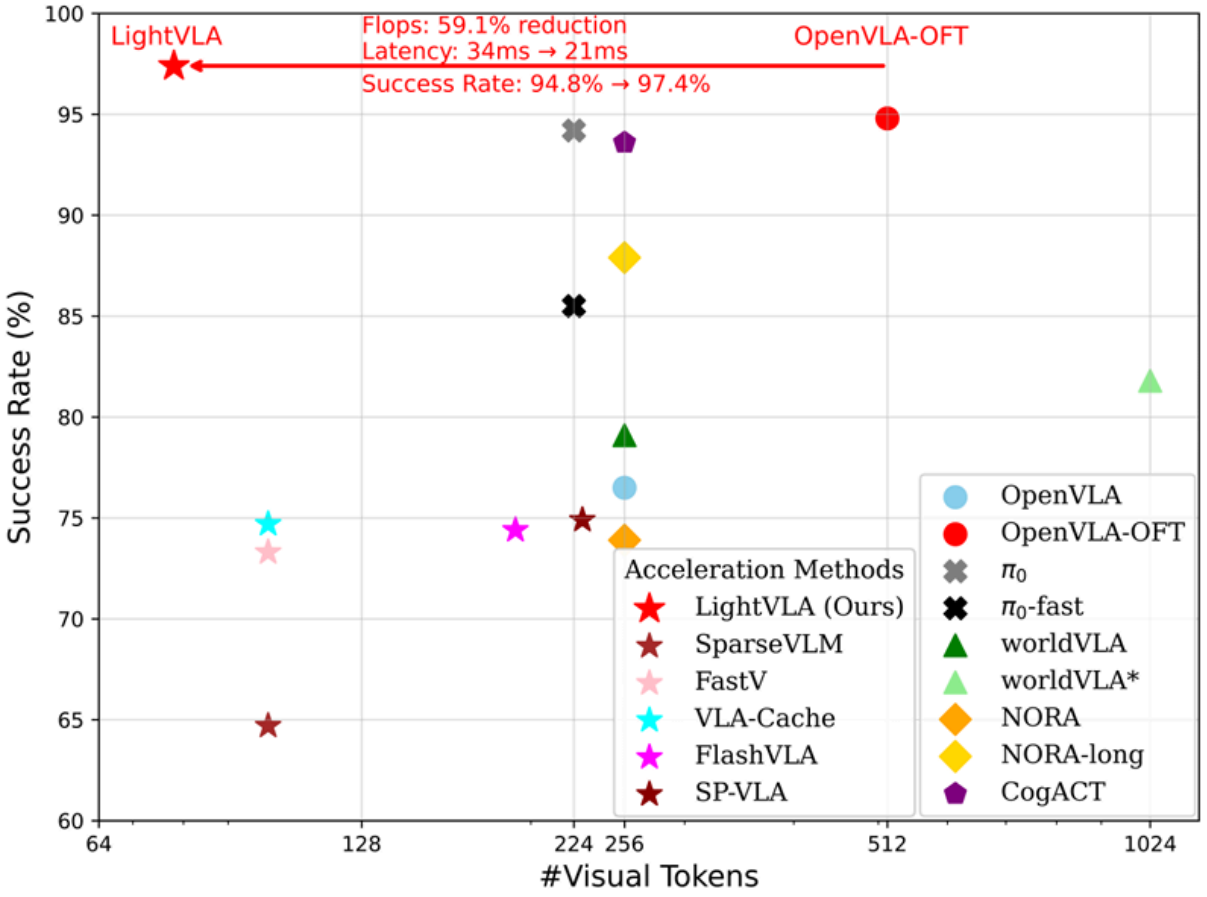

Fig. 1: LightVLA achieves better performance than common VLA models and acceleration methods with fewer visual tokens, yielding efficient computation and lower latency.

2 相关工作

2.1 视觉 - 语言模型（VLM）

视觉 - 语言模型（VLMs）融合视觉与语言模态，将 LLMs 的推理能力扩展到视觉输入处理领域。这种融合通常通过 “将图像编码为数百个与文本令牌对齐的视觉令牌” 实现 [28]。典型的 VLM 包括 LLaVa [28]、BLIP-2 [29]、InternVL [30] 及 Qwen-VL [31]，其参数规模通常在 70 亿至 700 亿之间。尽管 VLMs 具备上述优势，但其并非为 “直接生成任务特定机器人动作策略” 而设计 —— 这一需求催生了 VLA 模型的出现与发展。

2.2 视觉 - 语言 - 动作模型（VLA）

VLA 模型 [1]-[4] 将 VLMs 扩展至具身智能领域，能为复杂机器人任务（如操作任务）生成可行的动作策略，填补了 “感知与动作” 间的鸿沟。与 VLMs 类似，典型的 VLA 模型（如 OpenVLA [2]、π_0[3]、CogACT [6]）通常将图像块 token 化为数百个视觉令牌，再将其与任务令牌拼接，最终生成动作（既可为离散令牌 [2],[21],[32]，也可为连续值 [6],[7]）。

早期 VLA 模型（如 OpenVLA）生成离散动作令牌，通过自回归方式进一步生成动作策略；而近年 VLA 模型（如 CogACT、OpenVLA-OFT）则致力于生成连续动作令牌。此外，“动作分块技术（action chunking）”[7],[21],[22]（旨在预测动作策略序列）也被证明能有效提升整体性能。然而，VLA 模型数十亿的参数规模与高昂的推理成本，使其难以部署于低延迟实时机器人任务中。

为优化计算开销与延迟，现有工作多致力于设计轻量级 VLA 模型，如 TinyVLA [23]、SmolVLA [22]、NORA [21]。除模型架构外，令牌剪枝为 “通过减少输入令牌优化效率” 提供了可行方向，但该方向在 VLA 研究中仍未得到充分探索。

2.3 视觉令牌剪枝

令牌剪枝已成功应用于多种神经网络架构，从原始 Transformer、视觉 Transformer（ViT）[33]-[35] 到 LLMs、VLMs 等现代大规模模型 [13]-[20]。现有方法通常预先定义一个超参数，用于确定保留视觉令牌的固定数量 —— 这需要大量实证探索以选择该超参数的最优值。

在 VLA 模型场景中，为 VLM 优化的视觉令牌剪枝方法迁移到 VLA 时往往性能不佳 [24]-[26]。而专门针对 VLA 的现有工作 [12],[24]-[26] 虽探索了 “以注意力分数为指导的无训练视觉令牌剪枝”，但仍依赖固定的令牌剪枝比例。这种剪枝比例存在两大局限：（1）引入强归纳偏置，可能限制模型对不同视觉输入与任务的适应性；（2）通常不可避免地导致性能下降。

与现有方法不同，本文提出的可微分视觉令牌剪枝框架 LightVLA，能为每个任务场景动态选择必要令牌，在保持显著计算效率的同时提升性能。另一方面，vLLM [36]、SGLang [37] 等主流推理导向平台虽针对高吞吐量生成进行了优化，但推理过程中不暴露中间注意力分数 —— 这使得传统 “基于注意力分数的令牌选择方法”[12],[14] 无法适用。而 LightVLA 无需依赖 LLM 内部的注意力分数，可与这些平台良好兼容，从而便于实际部署。

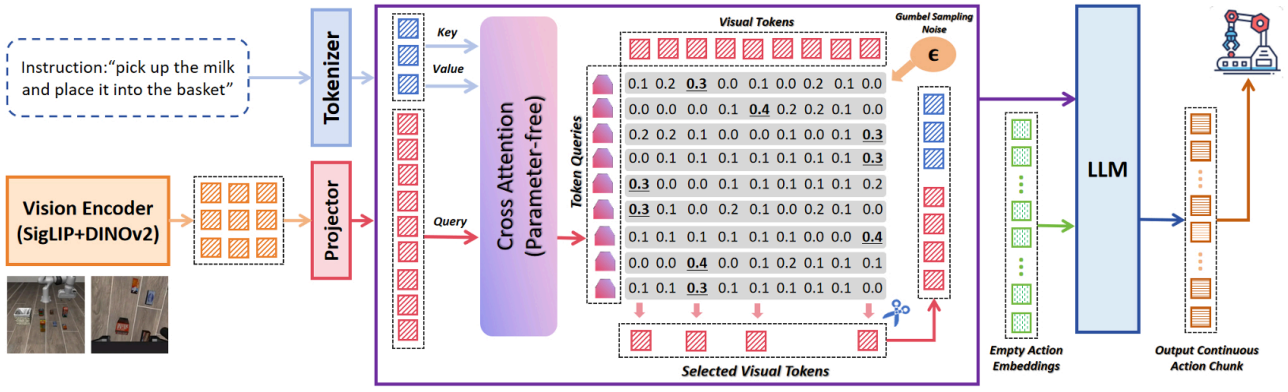

Fig. 2: Illustration of the proposed LightVLA framework. Gray regions indicate the use of Gumbel-softmax for differentiable token selection.

3 方法

本文引入 LightVLA—— 一种用于 VLA 加速的可微分视觉令牌剪枝框架。与现有研究不同，本文的策略完全以**性能为导向**：即唯一优化目标是提升性能。LightVLA 使模型能自适应保留或丢弃令牌，在此过程中，模型自主学习仅保留有用令牌以最大化性能，进而提升效率。此外，LightVLA 无需额外参数、超参数或辅助损失，是一种可兼容大多数 VLA 模型的通用框架。

3.1 问题定义

典型的 VLA 模型可分解为三个组件：视觉编码器（含投影函数 f_v ）、LLM 骨干网络 f_ϕ 、动作头（或解令牌器）f_a 。具体流程如下：

1. 视觉编码器将输入图像 X_I 编码为 L_v 个初始视觉令牌 $H_{v'} = f_v(X_I) \in \mathbb{R}^{L_v \times D'}$ ；
2. 初始视觉令牌经投影得到 $H_v = f_v(X_I) \in \mathbb{R}^{L_v \times D}$ ，并与语言令牌 $H_l \in \mathbb{R}^{L_l \times D}$ 拼接后输入 LLM 骨干网络；

- 其中，D' 为初始视觉令牌嵌入维度，D为文本令牌嵌入维度；

3. 动作头 f_a 最终将 LLM 骨干网络输出的隐藏状态转化为面向机器人任务的动作策略。

LLM 输出最终隐藏状态（即输出令牌）$H = f_\phi(H_v, H_l)$ 。计算瓶颈主要存在于 LLM 骨干网络 f_ϕ 的解码器层 —— 该层需对所有视觉令牌 H_v 与文本令牌 H_l 执行大量基于注意力的计算。最终，动作头将 LLM 的隐藏状态H转化为连续动作策略 $A = f_a(H)$ 。

由于多数情况下 $L_v \gg L_l$ ，减少计算开销的有效方法是引入高效视觉令牌剪枝器 f_p 。视觉令牌剪枝器的目标是确定待保留的剪枝后视觉令牌集 $H'_v = f_p(H_v) \subseteq H_v$ ，以在不损害 VLA 模型性能的前提下降低计算成本。需注意的是，本文保留 [CLS] 令牌（因其承载全局视觉信息 [38]），仅对块级（patch-level）视觉令牌进行剪枝。

3.2 LightVLA 框架

现有研究 [13]-[20] 中的视觉令牌剪枝器通常将视觉令牌数量减少至预定义值 L'_v 。这种方法虽简单有效，但可能因信息丢失导致性能下降 —— 尤其当任务与场景复杂、L'_v 个令牌无法承载足够信息时。因此，VLA 模型亟需根据输入动态确定 H'_v，以进一步最小化信息丢失。

本文提出的 LightVLA 是一种**基于查询的新型视觉令牌剪枝策略**，能自适应地从 H_v 中区分信息性视觉令牌。如图 2 所示，LightVLA 采用一系列 L_v 个令牌查询 $Q = \{q_1, q_2, \cdots, q_{L_v}\}$ ，每个查询负责从所有令牌中选择一个有用的视觉令牌（即 $h_k = q_i(H_v)$ ）。所有查询选择的视觉令牌共同构成剪枝后集合 $H'_v = \{h_k \mid \exists q_i, h_k = q_i(H_v)\}$ 。

- 极端情况下，若每个查询选择唯一令牌，则所有令牌均被保留，即 $H'_v = H_v$ ；
- 若多个查询选择同一令牌，则重复令牌不会被重复保留，即 $H'_v \subset H_v$ 。

剪枝过程分为三个步骤：查询生成、令牌评分、令牌选择。

3.2.1 查询生成

现有 VLM 研究中，常用 “可学习嵌入作为查询”[13],[29]，但这种方法会为模型引入额外参数，使其在 VLA 模型常部署的资源受限平台上可行性降低。因此，本文提出一种**无参数查询生成方法**，以提升兼容性。

本文观察到：视觉令牌的有用性可通过其视觉信息与语言指令的交互体现。例如，当指令为 “拿起牛奶并放入篮子” 时，VLA 模型应更关注图像中的两个关键语义对象（牛奶、篮子），而非其他信息性较低的对象或背景。

基于此，查询可通过视觉令牌与语言令牌间的交叉注意力生成，公式如下：

$$Q = softmax\left(\frac{H_v H_l^T}{\sqrt{D}}\right) H_l,$$

其中，$Q \in \mathbb{R}^{L_v \times D}$ 。需注意的是，与传统注意力设计不同，为简化流程，查询生成过程中不包含权重矩阵或偏置矩阵。

3.2.2 令牌评分

在该步骤中，每个查询通过令牌评分独立评估所有令牌的有用性，公式如下：

$$S = \frac{QH_v^T}{\sqrt{D}} \quad (2)$$

其中，$S \in \mathbb{R}^{L_v \times L_v}$ 为评分矩阵，$s_{i,j}$ 表示第 i 个查询为第 j 个令牌分配的分数。

3.2.3 令牌选择

为确定剪枝后的令牌集，每个查询选择分数最高的令牌，公式如下：

$$H_v' = \{h_k \mid k = argmax_j(s_{i,j}), j = 1, 2, \cdots, L_v\} \quad (3)$$

然而，训练过程中需注意：argmax操作不具备可微性。现有研究 [18] 提出的解决方案是在令牌分数上引入辅助损失，但辅助损失不仅会使训练流程复杂化，还可能导致性能下降与不同优化目标间的梯度冲突 —— 且令牌分数的真值难以定义。

为解决这一问题，本文采用 Gumbel-softmax 采样技术 [27]，使argmax操作具备可微性。该技术可使离散令牌采样过程在反向传播中保持可微，从而引导 VLA 模型学习选择信息最丰富的视觉令牌。具体而言，本文将评分矩阵 $S \in \mathbb{R}^{L_v \times L_v}$ 转化为指示矩阵 $I \in \mathbb{R}^{L_v \times L_v}$，相关公式如下：

$$S' = S + \epsilon \quad (4)$$

$$S_{soft} = softmax_j\left(S'\right) \quad (5)$$

$$S_{hard} = one\text{-}hot\left(argmax_j\left(S'\right)\right) \quad (6)$$

$$I = S_{hard} + S_{soft} - S_{soft}^{SG} \quad (7)$$

其中：

- S' 为注入 Gumbel 采样噪声 $\epsilon \in U(0, \alpha)$ 后的评分矩阵；
- S_{soft} 与 S_{hard} 分别为软评分与硬评分；
- $one\text{-}hot$ 表示独热编码函数；
- SG表示停止梯度（stop gradient）操作。

需注意的是，与原始 Gumbel-softmax 操作（$\epsilon \in U(0,1)$）不同，本文通过**衰减噪声上界** α，随训练进程逐步降低采样噪声强度。第 4.4 节的消融实验验证了该设计的有效性：训练初期，该设计鼓励模型探索更多样的令牌选择方案；训练后期，有助于模型稳定收敛。

通过指示矩阵 I，可得到剪枝后的令牌集：

$$H_v' = IH_v^T \quad (8)$$

由于 I 是指示矩阵，H_v' 仅包含查询选择的令牌。此外，I 的梯度等于 S' 的梯度，因此可通过梯度下降以端到端方式优化查询。值得注意的是，推理阶段无需 Gumbel 噪声，直接通过argmax操作选择查询对应的视觉令牌。

此外，本文观察到 LLM 骨干网络严重依赖视觉令牌的位置 ID 以理解空间关系 —— 因此，令牌选择过程中需保留位置 ID。

4 实验

4.1 实验设置

4.1.1 数据集

本文在 LIBERO 基准测试 [39] 上评估 LightVLA—— 该基准测试基于 Franka Emika Panda 机械臂仿真环境，提供的演示数据包含相机图像、机器人状态、任务标注及末端执行器增量位姿动作。本文采用四个任务套件：LIBERO-Spatial（空间任务）、LIBERO-Object（对象任务）、LIBERO-Goal（目标任务）、LIBERO-Long（长程任务），每个套件包含 10 个任务，每个任务有 500 条专家演示数据。

为评估策略在不同空间布局、对象选择、任务目标及长程规划任务中的泛化能力，本文对每个任务进行 50 次测试（每个套件共 500 次测试），并报告每个任务套件的成功率（%）。

4.1.2 基准模型

为进行对比，本文将 LightVLA 与两类基准模型比较：一类是不同架构的 VLA 模型，另一类是对现有 VLA 模型应用令牌剪枝的方法，具体如下：

- **不同架构的 VLA 模型**：OpenVLA [2]、π_0系列 [3],[4]、NORA 系列 [21]、SmolVLA [22]、OpenVLA-OFT [7]、CogACT [6]、WorldVLA（256 个视觉令牌）、WorldVLA*（1024 个视觉令牌）[32]；
- **令牌剪枝方法**：FlashVLA[24]、SP-VLA[25]、VLA-cache[26]、FastV[14]、SparseVLM[15]；
  - 注：SparseVLM 与 FastV 最初为 VLM 设计，本文参考 [26] 将其应用于 OpenVLA 骨干网络。

4.1.3 实现细节

所有实验在 8 张 Nvidia® H20 GPU 上进行。本文以开源的 OpenVLA-OFT [7] 为基础模型，该模型包含双分支视觉编码器（DINOv2 [40] 与 SigLIP [41]），并以 LLaMA-2-7B [42] 为语言模型骨干网络。

微调过程中，采用秩为 32 的 LoRA [43] 技术，对整个模型（包括视觉编码器、LLM 骨干网络、动作头）进行微调。基础模型初始化采用 HuggingFace 上开源的 OpenVLA-OFT checkpoint，总微调步数为 40,000 步；学习率从 5e-4 衰减至 5e-5（衰减节点为 30,000 步）。单设备批次大小为 8，全局批次大小为 64。

4.2 主实验

表 1 展示了 LightVLA 与基于不同基础模型架构的基准模型在 LIBERO 基准测试上的对比结果（CogACT 结果来自 [44]）。结果表明，LightVLA 在所有任务中均实现最优性能 —— 尤其显著优于其基础模型 OpenVLA-OFT。此外，OpenVLA-OFT 需消耗 512 个视觉令牌，而 LightVLA 平均仅保留 78 个视觉令牌，这表明大多数视觉令牌对整体性能无贡献。该结果不仅验证了视觉模态的稀疏性，更证明 “性能与效率可作为兼顾目标同时优化”。

表 1：LIBERO 基准测试实验结果（注：TP 表示令牌剪枝（Token Pruning）；AR、FM、PD 分别表示不同动作解码（生成）范式：自回归（Auto-Regressive）、流匹配（Flow Matching）、并行解码（Parallel Decoding）；LightVLA 的 “保留视觉令牌数” 标注于成功率后，格式为 “平均值 ± 标准差”；* 表示本文复现结果，因硬件差异与原论文 [7] 略有不同。）

方法	规模	TP	解码方式	骨干网络	空间任务成功率（%）	对象任务成功率（%）	目标任务成功率（%）	长 成 （
OpenVLA [2]	7B		AR	PrismaticVLM	84.7	88.4	79.2	53
SparseVLM [15]	7B	✓	AR	PrismaticVLM	79.8	67.0	72.6	39
FastV [14]	7B	✓	AR	PrismaticVLM	83.4	84.0	74.2	51
VLA-Cache [26]	7B	✓	AR	PrismaticVLM	83.8	85.8	76.4	52
FlashVLA [24]	7B	✓	AR	PrismaticVLM	84.2	86.4	75.4	51
SP-VLA [25]	7B	✓	AR	PrismaticVLM	75.4	85.6	84.4	54
WorldVLA [32]	7B		AR	Chameleon	85.6	89.0	82.6	59
WorldVLA* [32]	7B		AR	Chameleon	87.6	96.2	83.4	60
NORA [21]	3B		AR	Qwen-VL	85.6	87.8	77.0	45
SmolVLA [22]	2.25B		FM	SmolVLM	93.0	94.0	91.0	77
CogACT [6]	7B		FM	PrismaticVLM	97.2	98.0	90.2	88
π_0 [3]	3.3B		FM	PaliGemma	96.8	98.8	95.8	85
π_0-fast [4]	3.3B		FM	PaliGemma	96.4	96.8	88.6	60
NORA-Long [21]	3B		PD	Qwen-VL	92.2	95.4	89.4	74
OpenVLA-OFT* [7]	7B		PD	PrismaticVLM	97.6	94.2	95.2	92
LightVLA（本文）	7B	✓	PD	PrismaticVLM	98.4（90±15）	98.4（78±11）	98.2（64±10）	94 （

4.3 计算效率分析

表 2 对比了 LightVLA 与其他 VLA 加速方法的加速效果与性能。结果表明，与其他强基准模型相比，LightVLA 在实现最高平均成功率的同时，大幅降低了延迟与计算需求。与 OpenVLA-OFT 相比，LightVLA 不仅将 FLOPs 降低 59.1%、延迟降低 38.2%，还将成功率提升 2.6%。值得注意的是，在表 2 所有 VLA 加速方法中，LightVLA 是唯一实现性能提升的方法。该发现证明：在追求最优性能的过程中，通过消除视觉稀疏性，VLA 模型的效率也可得到优化。

表 2：LightVLA 与其他方法在 LIBERO 基准测试上的加速效果与性能对比（注：表格报告视觉令牌数、GPU 型号、计算成本（TFLOPs）、端到端延迟、任务平均成功率。）

方法	视觉令牌数	GPU 型号	计算成本（TFLOPs）	延迟（ms）	平均成功率（%）
OpenVLA [2]	256	A100	-	-	76.5
SparseVLM [15]	100（最大）	RTX 4090	1.4	83	64.7
FastV [14]	100（最大）	RTX 4090	1.9	53	73.3
VLA-Cache [26]	100（最大）	RTX 4090	1.4	32	74.7
FlashVLA [24]	192	H100	0.7	55	73.7
SP-VLA [25]	229（平均）	A100	3.1	-	74.9
OpenVLA-OFT [7]	512	H20	8.8	34	94.8
LightVLA（本文）	78（平均）	H20	3.6	21	97.4

4.4 消融实验

4.4.1 噪声因子调度的影响

本文提出 “随训练进程逐步降低采样噪声强度”，以实现更多样的令牌选择方案。为验证该设计的有效性，本文将 LightVLA 与两个变体对比：（1）无采样噪声的 LightVLA；（2）恒定采样噪声的 LightVLA（即不衰减噪声）。

表 3 结果表明，LightVLA 性能优于两个变体。对变体令牌剪枝方案的深入分析揭示了噪声因子调度的作用机制：与 LightVLA 相比，“无采样噪声” 变体保留的视觉令牌更少，在对象任务、目标任务等语义密集场景中，常导致关键语义信息丢失；而采样噪声的引入通过鼓励更多样的令牌剪枝选择，缓解了这一问题。此外，“恒定采样噪声” 变体表明：若噪声强度不衰减，模型难以学习剪枝令牌，导致保留令牌数量显著增加。

表 3：噪声因子调度对 LightVLA 的影响

变体	空间任务成功率（%）	对象任务成功率（%）	目标任务成功率（%）	长程任务成功率（%）	平均值（%）	平均保留令牌数
LightVLA（本文）	98.4	98.4	98.2	94.6	97.4	78
无噪声 LightVLA	98.8	97.6	97.2	94.2	97.0	72
无调度 LightVLA	99.4	97.8	96.0	94.8	97.0	112

4.4.2 保留令牌对性能的影响

LightVLA 的核心优势之一是 “自适应区分有用令牌”。为验证这一能力，本文通过以下两种方式操纵 LightVLA 的令牌剪枝方案：

1. **补充随机令牌**：LightVLA 保留k个令牌后，向剪枝集补充k个随机令牌，最终向 LLM 输入2k个令牌 —— 验证 LightVLA 是否遗漏有用令牌；
2. **丢弃部分令牌**：LightVLA 保留k个令牌后，从剪枝集中随机丢弃 10% 的令牌，最终向 LLM 输入(0.9k)个令牌 —— 验证 LightVLA 是否保留无用令牌。

表 4 结果表明，对令牌剪枝方案的任何操纵均会导致性能下降，这证明 LightVLA 能精准保留有用令牌并丢弃无用令牌。

表 4：保留令牌对性能的影响

模型	令牌数	空间任务成功率（%）	对象任务成功率（%）	目标任务成功率（%）	长程任务成功率（%）	平均值（%）
LightVLA（本文）	k	98.4	98.4	98.2	94.6	97.4
LightVLA（补充k个）	2k	98.0	97.6	97.8	93.8	96.8
LightVLA（丢弃 10%）	(0.9k)	98.2	97.8	97.2	93.0	96.6

4.5 定性可视化

为更直观展示令牌剪枝过程，本文选取一个任务片段作为演示，展示 VLA 模型执行操作任务时的令牌选择动态。选取的关键帧涵盖操作任务的关键阶段（对象交互、任务完成）。如图 3 所示，保留的令牌集中于关键关注对象（摩卡壶、炉灶、机械臂本身），而大多数背景令牌被剪去。此外，对比 Frame#175 与 Frame#275 可发现，LightVLA 能根据需求灵活调整保留或剪去的令牌数量 —— 这进一步验证了其自适应令牌剪枝的有效性。

（图 3 说明：LIBERO-Long 任务示例 ——“将两个摩卡壶放在炉灶上”。每个帧包含 4 张图像：左上为第三人称视角相机图像；右上为腕部相机图像；左下为 “剪枝后令牌掩码标注的第三人称视角图像”；右下为 “剪枝后令牌掩码标注的腕部相机图像”。）

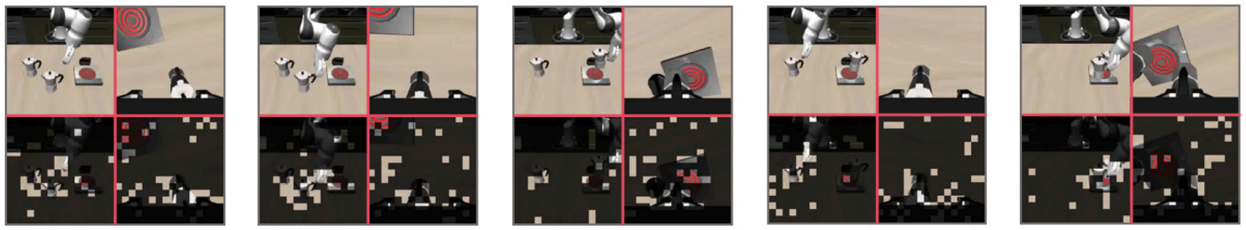

Fig. 3: An example of LIBERO-Long task: ’Put both moka pots on the stove’. Each frame consists of 4 images. Upper left: The 3rd person view camera. Upper right: The wrist camera. Lower left: The 3rd person view camera with pruned tokens masked. Lower right: The wrist camera with pruned tokens masked.

5 讨论

5.1 基于可学习查询的令牌剪枝

由于令牌剪枝（尤其是 “训练感知型令牌剪枝”）在 VLA 研究中仍较少被探索，除前文介绍的 “无参数令牌剪枝方法 LightVLA” 外，本文还实现了 LightVLA*—— 该方法采用 “可学习查询”（含额外可训练参数）[13],[29] 选择信息性视觉令牌。本文考虑将可学习查询应用于两个不同位置，分别基于 “仅视觉特征” 或 “视觉 - 语言联合特征” 进行剪枝。具体而言，为滤除冗余视觉令牌，本文引入 N_q 个压缩查询作为令牌查询头，引导模型从 L_v 个视觉令牌中选择有用令牌。

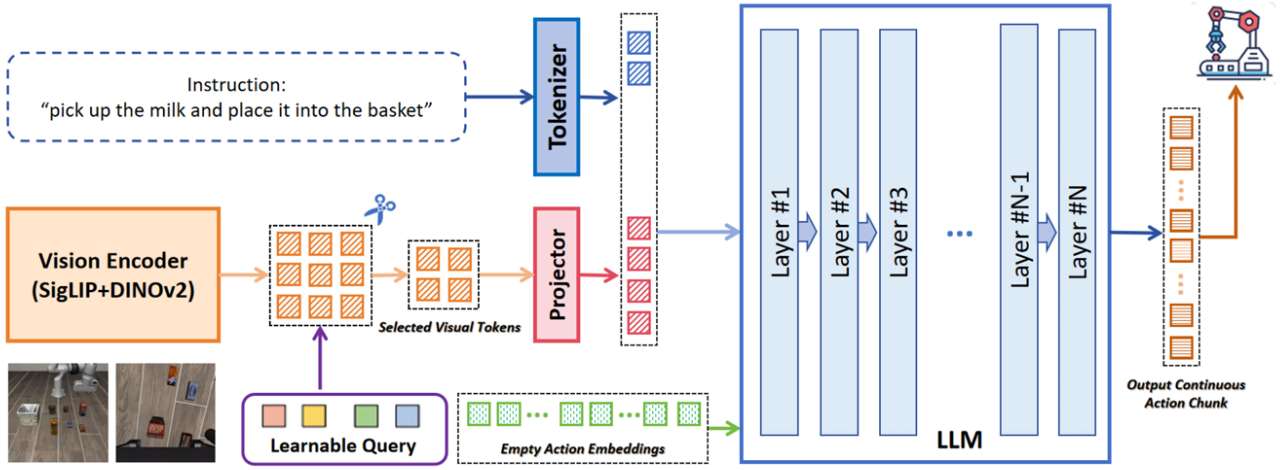

Fig. 4: Illustration of LightVLA* when pruning visual tokens at the vision encoder with the learnable query.

5.1.1 在视觉编码器端进行选择

N_q 个压缩查询与所有视觉令牌 H_v 交互，选择性提取关键视觉信息以生成剪枝后的视觉令牌。与无参数 LightVLA 的公式（1）不同，本文在视觉编码器后引入可学习查询 $Q^* \in \mathbb{R}^{N_q \times D'}$（如图 4 所示）。

需注意的是，视觉令牌首先在通道维度拼接，再执行令牌剪枝。令牌评分计算如下：

$$S^* = \frac{LN(Q^*) \cdot LN(H_v^T)}{\sqrt{D'}},$$

其中：

- D' 为投影层前的视觉令牌维度；
- LN表示层归一化（Layer Normalization），用于稳定训练过程；
- $S^* \in \mathbb{R}^{N_q \times L_v}$ 为评分矩阵。

本文采用 RMSNorm [45] 实现高效层归一化。基于 S^*，每个查询选择分数最高的视觉令牌；且通过 Gumbel-softmax 操作保证训练过程的可微性 —— 这与 LightVLA 一致。LightVLA *引入的额外参数包括 “可学习查询 $Q^$ ” 及 “层归一化映射参数”*，且可与 vLLM、SGLang 等高效推理导向平台兼容。

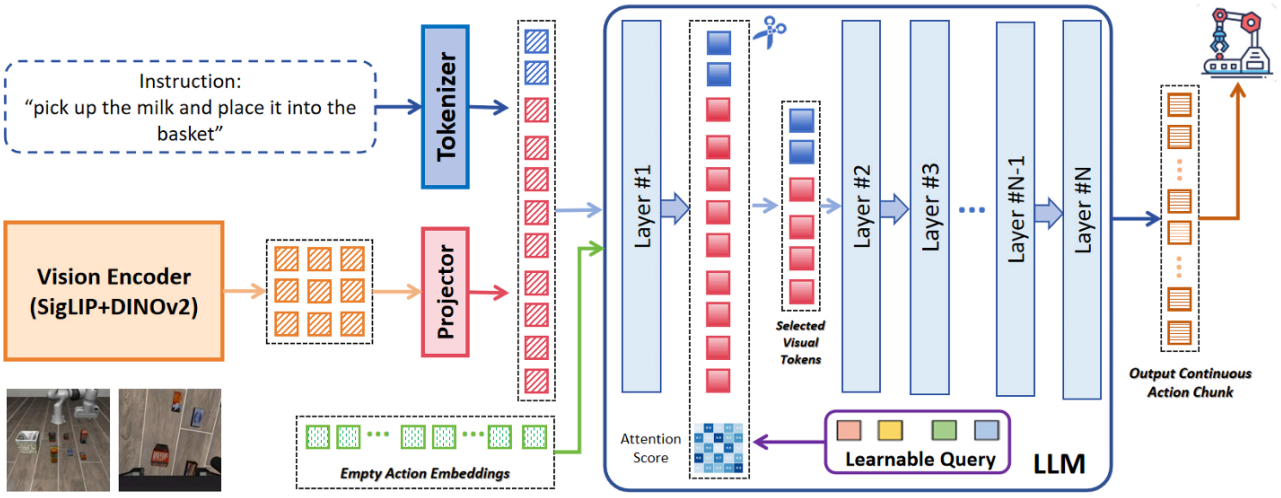

Fig. 5: Illustration of LightVLA* when pruning visual tokens at the first decoder layer of LLM with the learnable query.

5.1.2 在 LLM 早期层进行选择

不同于在图像编码器端操作，压缩查询 Q_v 可在 LLM 中与 “视觉令牌 H_v ” 及 “早期层文本令牌 H_l ” 交互（如图 5 所示）—— 这使模型能利用任务级文本的语义信息引导令牌剪枝过程。

其核心思路与 “在视觉编码器端选择” 类似，但引入 “文本令牌对视觉令牌的注意力分数”，公式如下：

$$S^{\dagger} = \frac{LN(Q^{\dagger}) \cdot LN(H_v^T) + \zeta \cdot attn}{\sqrt{D}} \quad (10)$$

其中：

- D为投影后视觉令牌维度；
- attn表示 “视觉令牌对文本令牌的注意力分数”；
- $Q^{\dagger} \in \mathbb{R}^{N_q \times D}$ 为可学习查询；
- ζ 为 “交叉注意力权重的可学习权衡系数”，初始值设为 1.0；
- 注意力分数由 LLM 对应的解码器层输出。

本文选择在 LLM 早期层（第 1-3 层）而非深层（第 4-32 层）剪枝冗余视觉令牌 —— 原因在于：深层已形成丰富的跨模态表示，视觉与文本特征深度融合且语义纠缠；而在早期层剪枝视觉令牌，能更有效降低计算量（因每个解码器层均需执行大量二次注意力计算）。需注意的是，若在解码器层剪枝视觉令牌，LightVLA * 需输出注意力分数 —— 这与前文提到的部分推理导向平台不兼容。

（图 4 说明：LightVLA_在视觉编码器端采用可学习查询进行视觉令牌剪枝的示意图。）

（图 5 说明：LightVLA_在 LLM 第一个解码器层采用可学习查询进行视觉令牌剪枝的示意图。）

5.1.3 分析

在 LightVLA * 实验中，本文初始化 $N_q = 128$ 个查询 —— 首先保留 25% 的视觉令牌（因 LightVLA 实验表明，自适应保留 10%-20% 的视觉令牌已足够）。表 5 展示了实验结果。

结果表明：LightVLA*系列与 LightVLA 的性能均优于表 1 中的对应基准模型；其中，“在 LLM 第一个解码器层进行剪枝的 LightVLA”* 在复杂长程任务套件 LIBERO-Long 上实现最优性能。此外，还可观察到：在 LLM 深层解码器层剪枝视觉令牌时，平均性能略有下降。

表 5：基于可学习令牌查询的实验结果（注：SR 表示成功率（Success Rate）。）

方法	空间任务 SR（%）	对象任务 SR（%）	目标任务 SR（%）	长程任务 SR（%）	平均值（%）	平均保留令牌数
LightVLA（本文，无参数）	98.4	98.4	98.2	94.6	97.4	78
LightVLA*（视觉编码器端）	98.2	98.6	97.8	95.0	97.4	128
LightVLA*（LLM 第 1 层）	98.6	98.8	98.0	95.2	97.7	128
LightVLA*（LLM 第 3 层）	98.0	98.2	97.6	94.8	97.2	128

5.2 与专家混合模型（MoE）的对比

LightVLA 与专家混合（Mixture of Expert, MoE）技术 [46] 的核心思路存在相似性：两者均通过从整体集合中选择 “密集子集元素” 优化前向传播效率。但两者的目标差异导致了行为上的本质不同：

- LightVLA 的目标是 “在提升效率的同时最大化性能”，因此其令牌选择聚焦于 “仅保留信息性令牌”；
- MoE 的目标是 “将特定任务分解为子任务，由不同专家处理”，为平衡专家间的知识与工作量，其选择会在专家间均匀分布，无特定聚焦。

综上，LightVLA 与 MoE 在目标与行为上均存在本质差异，是两种截然不同的技术。

6 结论

本文研究了视觉 - 语言 - 动作（VLA）模型的内在视觉冗余问题，提出无额外参数的视觉令牌剪枝框架 LightVLA。该框架通过 “基于查询的可微分令牌剪枝过程”，自适应选择信息性视觉令牌；在 LIBERO 基准测试上，LightVLA 实现了当前最优性能，同时显著优化了计算效率。此外，本文还提出含可学习查询（额外可训练参数）的框架 LightVLA*，其性能同样优于对应基准模型。

关于未来工作，本文计划从两方面展开：

1. 进一步研究端到端 VLM 或面向自动驾驶的 VLA 模型中的视觉冗余问题，优化整体效率与延迟，以促进实际部署；
2. 探索 “高效令牌剪枝” 在 “面向复杂空间智能任务的 VLM/VLA 模型” 中的应用，推动其在消费级设备（如家用机器人）上的广泛应用。

文档中 Gumbel softmax 的详细解释

在《LightVLA.pdf》中，Gumbel softmax 是实现 “可微分视觉令牌选择” 的核心技术，其核心目标是解决令牌剪枝过程中**argmax 操作不可微分**的关键问题，使离散令牌采样过程能在模型反向传播阶

段保持可微性，从而支持端到端训练，最终引导 VLA 模型精准选择对任务有用的视觉令牌（）。以下结合文档内容，从技术定位、实现流程、文档改进及应用价值四方面展开详细解释：

一、技术定位：解决 “离散选择不可微分” 的核心工具

在 LightVLA 的令牌剪枝流程中，模型需要通过 “选择分数最高的视觉令牌” 确定剪枝后集合（公式 3：$H_v' = \{h_k \mid k = argmax_j(s_{i,j})\}$ ）。但传统的**argmax 操作是离散的、不可微分的**—— 若直接使用，模型无法通过反向传播优化 “令牌选择逻辑”（如生成动态查询的参数），导致剪枝策略无法与任务性能目标联动。

Gumbel softmax 的核心价值的在于：通过 “引入可控噪声 + 软 / 硬分数结合” 的方式，为离散的令牌选择过程赋予可微性，使模型能在训练阶段通过梯度下降端到端优化查询与令牌选择策略，同时在推理阶段保留 “硬选择” 的准确性。

二、文档中 Gumbel softmax 的具体实现流程

LightVLA 将 Gumbel softmax 应用于令牌剪枝的 “令牌选择” 阶段，具体分为 4 步，对应文档中的公式（4）-（8），核心是将 “离散的令牌选择” 转化为 “可微分的连续计算”：

1. 注入 Gumbel 采样噪声（公式 4）

首先对令牌评分矩阵S（ $S \in \mathbb{R}^{L_v \times L_v}$ ，元素 $s_{i,j}$ 表示 “第i个查询对第j个视觉令牌的评分”）注入**均匀分布的 Gumbel 噪声** ϵ ，得到带噪声的评分矩阵 S' ： $S' = S + \epsilon$ 其中，$\epsilon \in U(0, \alpha)$ （U表示均匀分布，α 为噪声上界）。注入噪声的目的是为了 “软化” 离散的分数排序，为后续软选择提供基础。

2. 计算 “软评分” 与 “硬评分”（公式 5-6）

- **软评分（S_{soft}）**：对带噪声的评分矩阵 S' 沿 “令牌维度”（j维度）计算 softmax，得到连续的概率分布，模拟 “软选择” 过程： $S_{soft} = softmax_j(S')$ 此时 S_{soft} 的元素为 0-1 之间的概率值，具备可微性，但无法直接对应 “确定的令牌选择结果”。
- **硬评分（S_{hard}）**：对 S' 执行传统 argmax 操作后，通过 one-hot 编码转化为离散的指示向量，模拟 “硬选择” 过程（即最终推理时的实际令牌选择逻辑）： $S_{hard} = one\text{-}hot\left(argmax_j(S')\right)$ 此时 S_{hard} 仅有 “选中令牌” 对应位置为 1，其余为 0，能准确表示选择结果，但不可微分。

3. 构建可微分的指示矩阵I（公式 7）

为同时保留 “软评分的可微性” 与 “硬评分的准确性”，LightVLA 通过 “直通估计（Straight-Through Estimator）” 思想，将 S_{hard} 与 S_{soft} 结合，构建最终的指示矩阵I：

$$I = S_{hard} + S_{soft} - S_{soft}^{SG}$$

其中，S_{soft}^{SG} 表示对 S_{soft} 执行 “停止梯度（Stop Gradient）” 操作 —— 即前向传播时使用 S_{hard} 的离散结果（保证选择准确性）（*前向传播过程中没有梯度，后两项直接相互抵消了，只剩下S_hard*），反向传播时使用 S_{soft} 的梯度（保证可微性）（*反向传播过程中只有 S_soft有梯度，借助它的梯度可以去更新最初的 S，进而 S_hard 也会被更新，进而在前向传播时就会有新的 I ，就会有新的选择*）。最终 I 既能准确指示 “选中的令牌”，又能支持梯度回传，实现 “离散选择 + 连续优化” 的统一。

4. 生成剪枝后令牌集（公式 8）

通过指示矩阵 I 与原始视觉令牌 H_v 的矩阵乘法，得到剪枝后的令牌集 H_v'：$H_v' = IH_v^T$ 由于 I 仅在“选中令牌” 位置有非零值，H_v' 自然过滤掉无用令牌；同时，因 I 的梯度来自 S_{soft}，模型可通过梯度下降优化 “查询生成逻辑”，使后续选择更精准。

三、文档对 Gumbel softmax 的关键改进：噪声上界动态衰减

原始 Gumbel softmax 通常固定噪声分布为 $\epsilon \in U(0,1)$（噪声上界 $\alpha = 1$），但 LightVLA 对此做了针对性改进：**随训练进程逐步降低噪声上界 α。**

该改进的核心目的是平衡 “训练探索性” 与 “后期稳定性”：

- **训练早期**：α 较大，噪声强度高 —— 鼓励模型探索更多样的令牌选择方案，避免陷入 “局部最优剪枝策略”（如过度剪枝关键语义令牌）；（*经过归一化处理后，S‘ 元素的值都是-1到1的小数，有些元素可能很接近于0，但是如果噪声比较大，接近于0的元素就可能因为噪声的影响值变得比较大，也就是说，所有元素的之都可能变大，而在softmax下，元素值如果都较大的话，结果会比较平均，就不会轻易的把某些元素的值设为零，就不会大量丢弃令牌，就不会陷入局部最优*）
- **训练后期**：α 逐步减小至接近 0，噪声强度低 —— 使模型的令牌选择策略趋于稳定，最终收敛到 “性能最优的剪枝方案”。

文档通过消融实验验证了该改进的有效性：对比 “无噪声”“恒定噪声” 变体，采用 “动态衰减噪声” 的 LightVLA 在所有任务套件中均实现更高成功率，且保留的令牌数更合理。

四、在 LightVLA 中的应用价值

Gumbel softmax 是 LightVLA 实现 “性能与效率双赢” 的关键支撑，其价值体现在三方面：

1. **突破 “效率 - 性能权衡”**：通过可微分训练，模型能自主学习 “保留有用令牌、剪去无用令牌”，既降低计算开销（减少 FLOPs 与延迟），又避免因信息丢失导致的性能下降 —— 最终实现 59.1% FLOPs 降低、38.2% 延迟降低，同时提升 2.6% 成功率；
2. **支持端到端优化**：无需引入额外辅助损失（传统方法常通过辅助损失优化不可微操作），简化训练流程，且避免 “辅助损失与主任务损失冲突” 的问题；
3. **兼容推理框架**：推理阶段可直接移除 Gumbel 噪声，采用 “argmax 硬选择”，无需依赖 LLM 内部的注意力分数，能与 vLLM、SGLang 等主流推理平台兼容，便于实际部署。