arXiv:2602.01522v1 [cs.LG] 2 Feb 2026

何时秩-1就已足够？面向参数高效微调的几何引导初始化

Haoran Zhao [1] **Soyeon Caren Han** [1] **Eduard Hovy** [1]

摘要

参数高效微调（PEFT）是适配多模态大语言模型的标准方法，但极低秩设置——尤其是秩1 LoRA——往往并不稳定。我们表明，这种不稳定性并非仅源于容量有限：在秩-1范式下，优化对更新方向高度敏感。具体而言，预训练的视觉和文本特征形成不匹配的各向异性区域，从而产生一个占主导地位的“间隙”方向；该方向如同一个平移分量，并在秩-1约束下不成比例地引导早期梯度。通过分析预训练表示，我们识别出一条主导早期梯度流的模态间隙轴，而随机秩一初始化不太可能与之对齐，进而导致梯度微弱和训练崩溃。我们提出**Gap-Init**，一种几何感知的初始化方法，它将秩1 LoRA的方向与由小型校准集估计得到的模态间隙向量对齐，同时保持初始LoRA更新为零。在多个视觉语言任务和骨干网络上，Gap-Init能够持续稳定秩-1训练，并可达到或超越强秩-8基线的表现。我们的结果表明，在极端低秩极限下，初始对齐的重要性可以不亚于秩本身。

1. 引言

大型多模态语言模型（多模态大语言模型）在图像描述和视觉问答等视觉语言任务上取得了出色的表现（Alayrac 等人，2022；Li等人，2022；2023）。然而，将这些模型适配到下游任务仍需耗费大量计算资源，因此催生了 LoRA 等参数高效微调方法（Houlsby 等人，2019；Hu 等人，2022b）。通过将更新限制在低秩适配器中，参数高效微调能够降低训练成本，同时通常保持模型性能。尽管低秩设置取得了成功，但在实际应用中，低秩设置，尤其是秩1适配，往往不够稳定，这通常被归因于表达能力有限或初始化匹配不佳（Hu 等人，2022b；Dettmers等人，2023；Meng 等人，2024）。因此，提高秩通常被视为提升稳定性和准确率的主要补救方法。

在本研究中，我们表明，在现代多模态学习场景中，仅凭容量不足以完全解释秩-1极限下观察到的不稳定性。在多个广泛使用的多模态大语言模型架构中，我们发现，秩-1优化可能会受到更新方向是否在激活空间中对几何显著的跨模态轴具有非平凡投影的过度影响。在秩-1约束下，随机初始化的方向很可能在这类轴上的投影微乎其微，从而削弱早期梯度信号并导致优化失败。这一观察与高维表示的已知性质一致：在实践中，随机方向往往近乎正交于固定的语义轴（Ethayarajh，2019；Mu & Viswanath，2018a；Vershynin，2018年）。

重要的是，我们并未声称单一秩一方向能够捕获所有任务相关变化。相反，我们的证据表明，在极低秩范围内，许多实际场景中的早期优化可能主要由与少数几何显著轴的对齐所主导，即使这些轴只能解释一小部分表征方差。基于这一洞见，我们提出**Gap-Init**，这是一种简单的免训练初始化策略：利用一个小型校准集估计跨模态差距方向，并使秩一更新与之对齐，同时确保 LoRA 更新在初始化时为零。

在多个多模态基准上，Gap-Init 提升了秩1 LoRA 的稳定性和有效性；在若干设置中，其性能接近秩更高的基线，而可训练参数则大幅减少。总体而言，我们的结果表明，在极低秩极限下，从几何角度看，初始化和表征几何可能是决定可训练性的首要因素，这补充了关于 PEFT 的基于容量的观点。

我们的贡献总结如下：

[1]墨尔本大学计算与信息系统学院，澳大利亚墨尔本。联系邮箱：Soyeon Caren Han <caren.han@unimelb.edu.au>。

预印本。*2026*年*2*月*3*日。

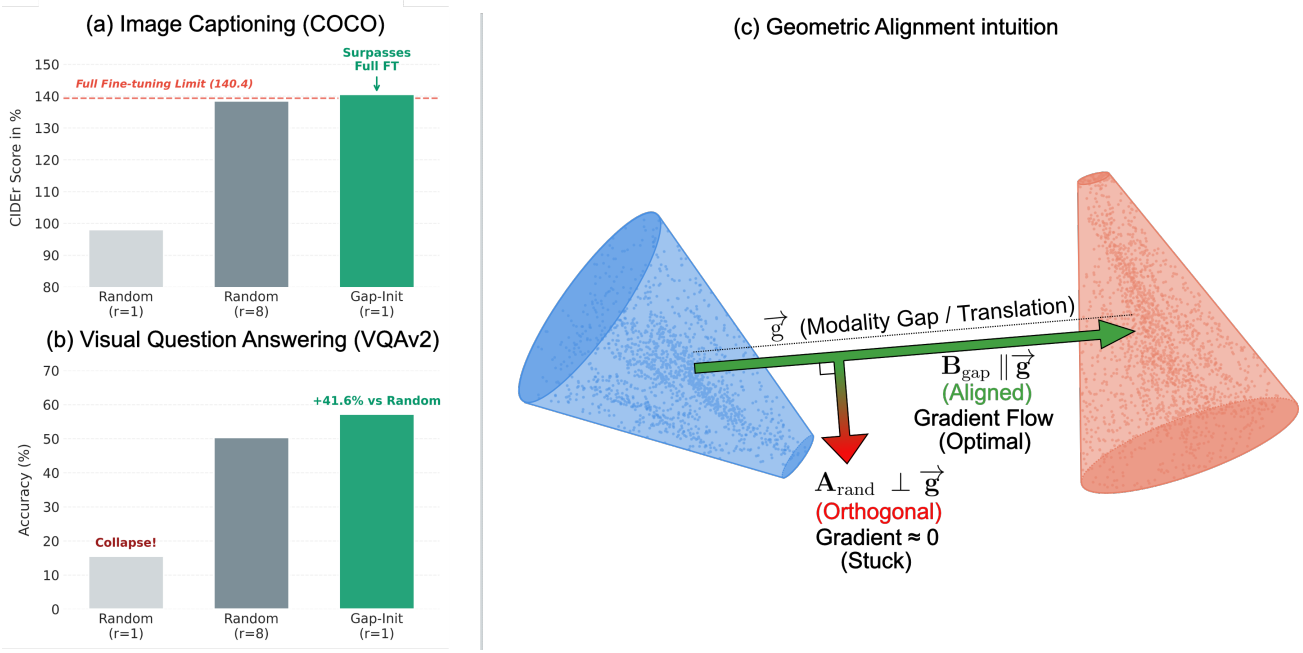


图 *1.* **Gap-Init 通过几何感知初始化，在多模态任务和不同训练设置下稳定秩1适配。(a) 图像描述（COCO）**。在此设置中，秩-1 Gap-Init 的表现可媲美甚至略优于强大的秩-8基线，同时大幅减少可训练参数。**(b) 视觉问答（VQAv2）**。与随机初始化相比，Gap-Init 提升了秩-1准确率，并实现稳定收敛。**(c) 几何直觉**。视觉和文本嵌入形成各向异性锥；二者的不匹配包含一个类平移分量，该分量对秩为 1 时的优化尤为重要。随机秩一初始化通常在这一轴上的投影微乎其微，因而抑制了有益的梯度流。Gap-Init 使更新方向与估计的差距向量对齐，从第一步起并在整个训练过程中改善优化动态，且适用于不同任务。

- **秩-1 不稳定性的几何视角：**我们发现，秩-1 参数高效微调中存在一种系统性且内在的优化障碍：随机低秩更新可能几乎与显著的跨模态轴正交，从而抑制早期梯度流并导致训练不稳定。

- **Gap-Init：**我们提出一种轻量级的校准式初始化方法，使秩1 LoRA 的方向与估计的模态间隙向量对齐，同时在初始化时保留预训练函数。

- **跨任务/骨干网络的实证证据：**我们表明，Gap-Init 能够在多个多模态基准上实现稳定且有效的秩1适配，并且在采用相同标准训练协议时，其表现通常可与更高秩基线相媲美。

2. 相关工作

参数高效微调（PEFT）。LoRA 在冻结预训练权重的同时学习低秩更新，大幅减少可训练参数，并取得有竞争力的性能（Hu 等人，2022b）。QLoRA 将低秩适配器与 4 位量化相结合，以实现内存高效的微调（ Dettmers 等人，2023）。除标准LoRA外，近期方法还通过利用预训练权重中的结构来改进低秩适配——DoRA 将权重分解为幅度和方向分量，并在分解后的形式中应用低秩更新（Liu 等人，2024）；PiSSA 则以基于奇异值分解的谱初始化取代标准初始化，该初始化源自预训练权重矩阵（Meng 等人，2024）。这些方法主要在权重空间中运作，并能在中等秩下改善优化和泛化。这些方法的一个共同假设是，无需明确参考数据诱导的激活模式，仅凭预训练权重的几何结构便可推断出与任务相关的适配方向。

视觉语言模型的参数高效微调。 适配器和LoRA式调优广泛用于多模态迁移。VL-Adapter通过注入轻量级适配器，在仅更新少量参数的同时，使视觉语言任务上的性能接近全量微调（Sung等人，2022）。BLIP-2和LLaVA等大规模多模态大语言模型在适配过程中通常依赖适配器/LoRA模块（Li等人，2023；Liu 等人，2023）。然而，大多数以往的多模态参数高效微调研究聚焦于中等秩（例如， $r \geq 8$），并未专门分析在极端秩-1范式下标准LoRA初始化的不稳定性；在这种情况下，实际优化会对初始化几何十分敏感。

表 *1.* **秩1适配：权重空间与激活空间。** 具有代表性的参数高效微调改进方法（PiSSA（Meng 等人，2024）、DoRA（Liu 等人，2024）聚焦于权重空间参数化，而 Gap-Init 则考虑在实际秩1适配中根据数据估计的激活空间几何。

方法	几何	数据	秩1 重点
PiSSA	权重	No	No
DoRA	权重	No	No
Gap-Init	激活	Yes	Yes

表征几何与模态差距。 众所周知，神经嵌入具有很强的各向异性：早期分析表明，去除主导成分可以改善各向同性和下游性能（Mu &Viswanath，2018b），而上下文表征则表现出强烈的各向异性以及随层变化的几何特性（Etha-yarajh，2019）。在多模态编码器中，Liang 等人 (2022) 对模态差距进行了形式化描述：图像和文本嵌入在共享空间中占据彼此分离的区域。

我们的方法定位。Gap-Init 与权重空间参数高效微调改进方法（例如 DoRA、PiSSA）互为补充：我们不通过对预训练权重进行对角化来选择子空间，而是利用数据诱导的激活空间几何，通过从小规模校准集中估计跨模态差距方向，并用它来初始化秩1适配。这一视角在优化受限于单一更新方向时尤其 relevant。

3. 模态对齐的几何结构

我们刻画了预训练多模态大语言模型中一种简单的几何结构，这种结构对于极端低秩适配具有实际意义。经验上，（i）视觉和文本隐藏状态具有高度各向异性，并集中在窄锥中；（ii）它们的跨模态不匹配包含显著的均值偏移（类似平移）成分。在秩为 1 的约束下，优化会高度依赖更新方向是否在该间隙轴上具有非平凡投影；在高维空间中，随机秩1方向很可能几乎与其正交，从而削弱有效梯度信号。Gap-Init 利用这一几何结构，将秩1 LoRA 方向初始化为与估计得到的差距向量对齐。

3.1. 随机秩一初始化为何效果不佳：近似正交集中现象

令 $u = g/\|g\|$ 表示归一化的模态差距方向。标准LoRA初始化会诱导出一个随机秩-1方向 b（除缩放因子外），该方向在 $\mathbb{S}^{d-1}$ 上近似均匀分布。在高维空间中，随机向量几乎必然与任意固定方向近乎正交，且这一结论以高概率成立。以下命题将这一随机方向的集中现象形式化。

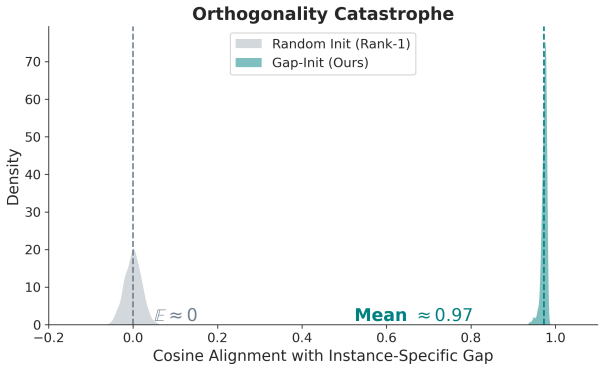


图 *2.* **初始化时的近似正交性。** 随机秩1 LoRA 方向集中在 $\cos(u, b) \approx 0$ 附近，而 Gap-Init 则通过构造使方向与估计得到的间隙轴对齐。

命题 3.1（随机方向的集中现象）。 若 b 在 $\mathbb{S}^{d-1}$ 上均匀随机分布，且 u 为固定方向，则

$$\mathbb{E}[\langle u, b\rangle] = 0, \qquad \mathbb{E}[\langle u, b\rangle^2] = \frac{1}{d},$$

且对于任意 $\varepsilon > 0$,

$$\mathbb{P}(|\langle u, b\rangle| > \varepsilon) \le 2\exp\left(-\frac{(d-1)\varepsilon^2}{2}\right).$$

对于 BLIP-2 ($d = 4096$)，这意味着在高维下，$|\langle u,\ b\rangle|$ 通常处于 $1/\sqrt{d}$ 的量级。

图2 展示了初始对齐程度的经验分布：随机秩一方向集中在余弦值 $\approx$ 0附近，而 Gap-Init 则实现了较高的对齐度（例如，$\cos \approx 0.97$ 在我们的设置中）。因此，在随机秩一初始化下，更新中沿间隙轴的分量（也就是早期梯度）可能会大幅衰减。

3.2. 模态差距中的优化相关平移分量

如果跨模态对齐需要协调许多彼此独立的方向，那么秩1适配在实践中就不太可能提供可靠的优化信号。相反，逐样本间隙向量的奇异值谱（图3）表明存在结构化失配：(i) 主导方向捕获了间隙变化中不可忽略的一部分（例如，~16% 在我们的分析中），且 (ii) 后续分量的贡献逐渐减小。

这促使我们使用跨样本共享的类平移分量来建模这种失配：

$$T(f_{\text{vision}}(x)) \approx f_{\text{vision}}(x) + \lambda(x)\, g,$$

其中，g 是全局间隙方向，$\lambda(x)$是依赖于样本的标量。虽然多模态对齐本质上并非一维，但在秩1约束下，早期优化期间的梯度信号可能主要由与这一显著方向的对齐所主导

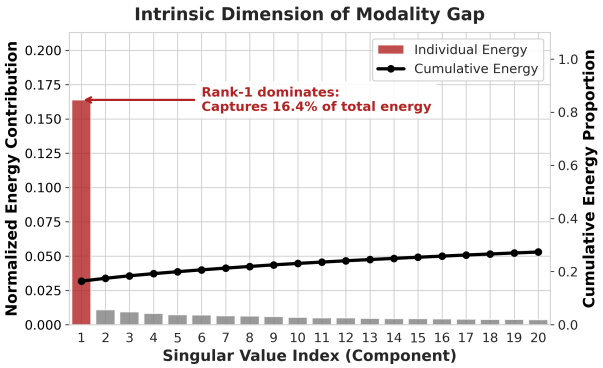


图 *3.* **实例特定间隙向量的奇异值谱**。 奇异值谱显示出一个显著的主导方向，而更高秩分量的贡献则逐渐减小。

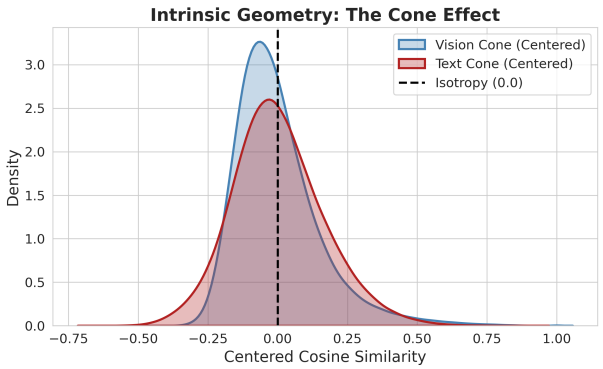


图 *4.* **内在几何（锥效应）**。 中心化后，视觉和文本嵌入仍高度集中在各自的均值方向附近，表明每种模态内部都存在各向异性。

。Gap-Init 通过在初始化时将秩1 LoRA 子空间与 g 对齐，将这一观察付诸实践。

附录A中通过高斯平移模型进行的形式化分析及合成验证，B附录中则进一步讨论了何时会出现类似平移的结构。

3.3. 锥假说：模态簇由均值主导

关于嵌入各向异性的既有研究（Mu & Viswanath, 2018b；Ethayarajh, 2019）以及多模态对比模型（Liang 等人, 2022）表明，高维表示通常会集中在狭窄的角区域（锥）内。我们通过测量中心化后的视觉/文本嵌入与各自均值向量之间的余弦相似度，确认了 BLIP-2 中也存在这一现象。

如图所示 4，两种模态都呈现出紧密的集中趋势，这与均值主导的几何特征一致。这促使我们将跨模态不匹配建模为一个占主导地位的类平移方向，而模态均值之差可作为一种简单的估计量：

$$g = \mu_{\text{text}} - \mu_{\text{vision}}.$$

接下来，我们将研究在此设置下，当适配子空间被限制为秩-1 时，与这一间隙方向对齐会如何影响优化。

4. 方法

受第3节几何分析的启发，我们提出了 **Gap-Init**，一种几何感知初始化策略，通过显式地将 LoRA 更新方向与模态差距对齐，来弥补秩1适配的不足。关键在于，我们的方法具有层特异性：在每个 *Transformer* 层的隐藏状态空间内估计模态差距，并使用相应的差距向量初始化该层的 LoRA 参数。这一设计确保了严格的维度和坐标一致性，并避免依赖一个跨异构层的全局方向。

我们的方法由三个部分组成：（1）重新审视秩1低秩适配的失效模式，（2）估计对齐表示空间中的平移主导的模态差距，以及（3）初始化 LoRA 参数，使其与这一差距对齐。我们进一步通过间隙引导的层选择策略，验证该差距的诊断价值。

4.1. 预备知识：秩-1 适配及其局限性

令 $W_0 \in \mathbb{R}^{d\times k}$ 表示 Transformer 层中冻结的预训练权重矩阵。LoRA（Hu 等人, 2022a）将可训练更新参数化为

$$\Delta W = BA,$$

其中 $B \in \mathbb{R}^{d\times r}$ 和 $A \in \mathbb{R}^{r\times k}$ 满足 $r \ll \min(d,\ k)$。前向计算变为

$$h = (W_0 + \Delta W)x = W_0 x + BAx. \tag{1}$$

按照标准做法， B 初始化为零， A从高斯分布中初始化。

对于中等秩（$r \geq 8$），优化器可以探索多个方向，并逐渐使更新子空间与任务相关几何结构对齐。然而，当秩为 1 时，可训练子空间会坍缩为一条完全由初始化决定的直线。如第 3 节所示，在高维空间（$d \approx 10^3$）中，随机初始化的方向几乎必然与模态差距方向正交。因此，沿有用方向的梯度分量会被抑制，其抑制因子为 $O(d^{-1/2})$，从而导致不稳定优化或完全坍缩——我们将这一现象称为正交性灾变。

4.2. 估计平移主导的模态差距

第 3 节表明，视觉与文本表示之间的主导差异表现为各自表示锥之间的平移。我们使用一个小型无标注校准集 $\mathcal{D}_{\text{cal}}$和冻结的预训练模型来估计这一平移。校准数据不需要精确监督或精确对齐；弱配对图文示例足以恢复表示空间中的主导平移方向。

对齐表示空间。 首先，使用视觉语言对齐模块（例如，Q-Former 后接线性投影）将视觉特征映射到语言模型的语义空间中。因此，视觉和文本表示都位于语言模型的同一隐藏状态空间中。

逐层差距估计。 对于每个 Transformer 层 ℓ，以及每个校准样本（$x^{(i)}$ 图像$x^{(i)}$ 文本），我们提取：

$$z_v^{(i,\ell)} = h_v^{(\ell)}(x_{\text{img}}^{(i)}), \qquad z_t^{(i,\ell)} = h_t^{(\ell)}(x_{\text{txt}}^{(i)}),$$

其中， $h_v^{(\ell)}$和 $h_t^{(\ell)}$ 分别表示第 ℓ 层对齐后的视觉和文本隐藏状态。

我们将第 ℓ 层的样本投影差距定义为

$$g^{(i,\ell)} = z_t^{(i,\ell)} - z_v^{(i,\ell)}.$$

SPOT-Gap。为减少噪声并提升鲁棒性，我们可以选择使用局部邻域一致性（SPOT-Gap；详情见附录C）对这些向量进行稳定化处理。最终的逐层模态差距计算如下：

$$g^{(\ell)} = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_i \tilde{g}^{(i,\ell)}.$$

该过程只需在校准集上进行一次前向传播，计算开销几乎可以忽略不计。

4.3. Gap-Init：几何感知初始化

Gap-Init 对 LoRA 参数进行初始化，使每一层的更新子空间与相应的模态差距对齐。值得注意的是，差距向量和 LoRA 参数定义在同一逐层隐藏空间中，从而确保维度和坐标一致性。

对于秩为 1 的情况，LoRA 矩阵相应地简化为向量。在层 ℓ，我们初始化

$$B^{(\ell)} \leftarrow \frac{g^{(\ell)}}{\|g^{(\ell)}\|_2}, \qquad A^{(\ell)} \leftarrow 0.$$

算法1 Gap-Init：用于秩-1 适配 1：输入：包含视觉编码器 f_V、文本编码器 f_T 和投影模块 ϕ 的预训练模型；校准集 $\mathcal{D}_{\text{cal}}$；LoRA 秩 r。2：输出：初始化后的 LoRA 矩阵（B, A）。3：步骤 1：提取对齐嵌入 4：对于每个样本（x_{img}， x_{txt}） $\in \mathcal{D}_{\text{cal}}$ 执行 5：$z_v \leftarrow \phi$（f_V（x_{img}））6：$z_t \leftarrow f_T$（x_{txt}）7：$g^{(i)} \leftarrow z_t - z_v$（样本投影差距）8：结束循环 9：步骤 2：估计逐层模态差距（SPOT-Gap）10：对于每一层 ℓ执行 11：对 $\{g^{(i,\ell)}\}$应用 SPOT-Gap 平滑 12：$g^{(\ell)} \leftarrow \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_i \tilde{g}^{(i,\ell)}$13：结束循环 14：步骤 3：初始化 LoRA 矩阵 15：按所需形状初始化 B 和 A 16：设置 $B^{(\ell)}$的第一列： $B^{(\ell)}$：$_1 \leftarrow g^{(\ell)}/\|g^{(\ell)}\|_2$17：如果 $r > 1$ 则 18：随机初始化 B 的其余列 19：结束条件 20： $A \leftarrow 0$（在初始化时保持 $\Delta W = 0$）21：返回（B， A）

设置 $A^{(\ell)} = 0$ 可确保在初始化时满足 $\Delta W^{(\ell)} = 0$，从而保留预训练模型的行为。

尽管前向传播保持不变，但反向传播的动态过程发生了根本变化。相对于 $A^{(\ell)}$ 的梯度变为

$$\frac{\partial \mathcal{L}}{\partial A^{(\ell)}} = \left(B^{(\ell)}\right)^\top \frac{\partial \mathcal{L}}{\partial h^{(\ell)}} x^\top,$$

这会将梯度直接投影到模态差距方向本身。相比之下，标准 LoRA 必须先通过优化发现这一方向，而这在秩-1 约束下效果不佳。

对于更高的秩（$r > 1$），我们将 $B^{(\ell)}$的第一列与 $g^{(\ell)}$ 对齐，并随机初始化其余列。

值得注意的是，Gap-Init 不会引入新的参数、目标函数或训练阶段——它只改变现有秩-1 适配器的初始方向。

4.4. 差距引导的层选择（GG-Saf e)

尽管即使采用朴素的层选择，Gap-Init 也表现出色，我们仍进一步考察模态差距的诊断价值，并利用它选择适配层。

我们计算一个逐层 *SPOT-Gap* 分数 $S_{\text{gap}}^{(l)}$，用于量化第 l 层视觉与文本隐藏表示之间的差异。该分数衡量该层中平均样本投影差距的幅度（详见附录 C）。

我们采用两项约束：

1. **安全性：** 我们排除视觉编码器的底层，因为它们编码低级的模态特定特征，应保持不变。

2. **诊断：** 在其余层中，我们选择值最高的 top-k 层。

在第 5 节的实证结果中，随机初始化下，GG-Safe 的表现显著优于朴素选择，这验证了 SPOT-Gap 度量在诊断方面的相关性。不过，所提出的 Gap-Init 足够稳健，即使实际应用中不使用 GG-Safe，其优势依然存在。

5. 实验

我们在图像描述（COCO）、视觉问答（VQAv2、OK-VQA）、零样本迁移（Flickr30k）和幻觉鲁棒性（POPE）任务上评估 Gap-Init，所有分数均以百分比形式报告。

5.1. 实验设置

模型。 所有实验均采用 BLIP-2 OPT-2.7B（Li等人, 2023），因为它提供了一种简洁且广泛采用的架构，将视觉编码器与语言模型解耦，因此特别适合对跨模态表示进行受控分析。

训练与评估。 模型在 COCO 图像描述（Karpathy 划分）和 VQAv2 上进行微调，并在 Flickr30k 和 OK-VQA 上评估零样本泛化能力，同时在 POPE 上评估幻觉鲁棒性。

基线。 我们比较以下方法：（1）采用随机初始化的标准 *LoRA*，（2）*PiSSA*（Meng 等人, 2024），一种基于SVD的权重初始化方法；以及（3）秩-*8 LoRA*和全层*LoRA*，作为容量上限。

协议。 我们的主要关注点是秩-1 范式，它代表了本研究中参数效率的实际极限。除非另有说明，否则我们采用朴素的层选择（顶部 6 层），并冻结 Q-Former。

5.2. COCO 图像描述结果

在 COCO 图像描述任务中，秩-1 标准LoRA表现出严重的优化不稳定性，CIDEr分数仅为 98.08（表 2）。相比之下，Gap-Init能够稳定训练，并取得 140.59 的 CIDEr分数，与

表 *2.* **COCO 图像描述（Karpathy 划分）。** 指标以百分比表示。在相同训练协议下，秩为 1 的 Gap-Init 所取得的性能可与更强的秩为 8 的基线相媲美，甚至超过后者。

方法	秩/配置	CIDEr	BLEU-4
标准LoRA	1 / 朴素	98.08	24.48
PiSSA	1 / 朴素	130.42	39.07
标准LoRA	8 / 朴素	138.49	40.63
全层LoRA	1 / 全部	140.36	41.65
Gap-Init（我们的方法）	**1 / 朴素**	**140.59**	**41.87**

表 *3.* **零样本迁移至 Flickr30k。** 指标以百分比表示。在相同训练协议和评估设置下，秩1的 Gap-Init 在所有秩-1方法中取得了最强的零样本迁移性能。

方法	Rank	CIDEr	Δ 与基线相比
标准LoRA	1	63.00	–
标准LoRA	8	78.35	+15.35
PiSSA	1	76.00	+13.00
Gap-Init（我们的方法）	**1**	**79.60**	**+16.60**

在该设置下的秩-8 LoRA相比仍具竞争力，同时使用的可训练参数减少了 $8\times$，且在 COCO 图像描述任务上可媲美全层LoRA。这表明，对齐更新方向可能足以恢复多模态对齐所需的最小子空间。

5.3. 向 Flickr30k 进行零样本迁移

在向 Flickr30k 进行零样本迁移时，Gap-Init 达到 79.60 CIDEr，优于其他秩-1 方法；在相同训练协议下，其性能与秩-8 LoRA 相当（表 3），且评估采用了完全相同的实验条件。尽管 PiSSA 在域内表现具有竞争力，但其迁移性能较低（76.00 CIDEr），这表明以权重为中心的初始化可能与预训练权重几何结构的关联更为紧密，而数据依赖型对齐则能更好地支持领域无关迁移。

5.4. 鲁棒性与推理：视觉问答和 POPE

Gap-Init 显著提升了秩-1 VQAv2 的性能（15。58 → 57. 23），并实现了稳定的收敛（表4）。就 VQA 准确率而言，在我们的训练协议下，PiSSA 获得了更高的分数，这表明权重空间谱初始化能够有效提升域内任务性能。相比之下，Gap-Init 获得了更高的 POPE F1（92.4 对 86.9），反映出其在针对幻觉的评估中具有更好的忠实性。这些结果表明，激活空间间隙对齐在秩 1 条件下能够带来互补优势，尤其体现在稳定性和鲁棒性方面-

表4. **视觉问答性能与幻觉鲁棒性（POPE）**。指标以百分比报告。Gap-Init稳定了秩-1训练，在保持较高POPE忠实度的同时取得了出色的视觉问答性能。

方法	秩	VQA2	OKVQA	POPE F1
标准LoRA	1	15.58	2.93	50.1
标准LoRA	8	50.39	17.02	**94.8**
PiSSA	1	**63.35**	**26.56**	86.9
Gap-Init（我们的方法）	**1**	57.23	21.03	92.4

相关指标方面，相较于权重空间初始化方法也是如此。我们强调，这些比较聚焦于极低秩范围（秩 1），在这一范围内，优化稳定性成为主导因素。

6. 消融与分析

6.1. 初始化与层选择

我们比较了秩1适配设置下层选择和初始化策略的影响。如表5所示，采用朴素层选择并随机初始化时，基线性能为 98.08 CIDEr。在此设置下，采用 GG-Safe 进行层选择，同时保持随机初始化，性能小幅提升至 102.15 CIDEr。

将 GG-Safe 层选择与 Gap-Init 结合后，性能大幅提升至 140.06 CIDEr。进一步在朴素层选择下应用 Gap-Init，性能略高，达到**140.59CIDEr**。两种 Gap-Init 设置之间的差距很小，这表明一旦初始化方向得到恰当对齐，适配层的具体选择就不那么重要了。

这些结果表明，虽然 GG-Safe 层选择有助于秩1适配，但性能提升主要源于对初始化方向的纠正。在极低秩约束下，优化主要受初始方向的几何对齐程度限制，而非适配层的具体选择。

关于初始化方向鲁棒性的进一步分析，包括校准集大小、数据组成和方向性噪声扰动的影响，见附录 D和附录E。

6.2. 与更强的参数高效微调基线进行比较

我们进一步按照 DoRA 基线的默认推荐配置，将 Gap-Init（秩 1）与其进行比较。

表 6 比较了所提出的 Gap-Init 与不同秩下的 DoRA 基线。秩为 1 的 DoRA 达到 140.51 CIDEr 和 41.55 BLEU-4，而增加秩达到 8 并不会带来额外收益。

值得注意的是，仅使用秩 1 的 Gap-Init 就取得了最佳整体性能，达到 140.59 CIDEr 和 41.81 BLEU-4，超过了两种 DoRA 配置。这些结果表明，单纯增加秩并不能保证性能提升，恰当的初始化可能比适配容量更为关键，因此 Gap-Init 是现有参数高效微调基线的一种参数高效替代方案。

表5。**秩1适配下的初始化与层选择比较**。指标以百分比表示。结果表明，虽然有指导的层选择能带来 modest 的收益，但性能提升主要源于对初始化方向的纠正。

方法	CIDEr
朴素选择 + 随机初始化	98.08
GG-Safe 选择 +随机初始化	102.15
GG-Safe 选择 +Gap-Init	140.06
朴素选择 + Gap-Init	**140.59**

表 6. **与采用低秩适配的 DoRA 进行比较**。 指标以百分比表示。在相同训练协议下，尽管采用了更低的适配秩，秩为 1 的 Gap-Init 在所评估的图像描述指标上仍然保持了与 DoRA 相当的竞争力。

方法	CIDEr	BLEU-4
DoRA (r=1)	140.51	41.55
DoRA (r=8)	139.55	40.92
Gap-Init (r=1)	**140.59**	**41.81**

6.3. 多种子评估

为严格评估秩-1适配的优化稳定性，我们使用五个不同的随机种子，对比 Gap-Init 与标准LoRA。统计摘要见表7。

表 7中的结果揭示了标准低秩适配的一个关键弱点。在随机初始化下，标准LoRA（$r = 1$）表现出严重的优化不稳定性，CIDEr得分的标准差高达 7.10，充分证明了这一点。尽管某些随机种子能够取得具有竞争力的性能，但其他种子则会遭遇灾难性崩溃，使训练过程实际上变成了一场碰运气。

与之形成鲜明对比的是，Gap-Init 展现出卓越的稳健性。通过使更新方向与模态差距对齐，我们的方法取得了始终如一的高平均性能 140.57，同时将标准差降低了近 5 倍。方差的大幅降低证实，Gap-Init 消除了高维随机初始化固有的“正交性风险”，确保秩1适配无论随机种子如何，都能可靠地收敛到高性能解。

表 7. **多随机种子稳定性分析**（$r = 1$）。我们报告 5 个随机种子的均值和标准差。标准LoRA表现出较高的方差和不稳定性，而 Gap-Init 则确保了稳健收敛，并显著降低了方差。

方法	CIDEr	BLEU-4
标准LoRA（$r = 1$）	135.37 ± 7.10	40.54 ± 1.28
Gap-Init ($r = 1$)	**140.57 ± 1.44**	**41.52 ± 0.73**

表 8。 **秩-1 适配下的跨骨干网络泛化**。 指标以百分比报告。与随机初始化相比，Gap-Init 在不同架构和规模的视觉-语言骨干网络上始终能提升性能。

方法	CIDEr	BLEU-4	METEOR
Qwen2-VL-7B-Instruct			
秩=1 ，随机	143.67	42.55	31.89
秩=1、Gap-Init	**144.11**	**42.97**	**31.95**
Gemma3-4B			
秩=1 ，随机	90.35	24.95	24.28
秩=1，Gap-Init	**95.24**	**27.27**	**26.33**

6.4. 视觉-语言骨干网络之间的跨模型泛化

我们评估了 Gap-Init 在不同架构和规模的视觉–语言骨干网络上的泛化能力，其中包括 Qwen2-VL-7B-Instruct (Wang 等人，2024) 和 Gemma3-4B(DeepMind，2025)，两者的设计存在显著差异。结果汇总见表8。

在大型骨干网络 Qwen2-VL-7B 上，秩-1 适配已表现出较强的竞争力，而 Gap-Init 在各项指标上均带来稳定提升。在较小的骨干网络 Gemma3-4B 上，受约束优化条件下 Gap-Init 的优势更加明显。在训练预算有限时，Gap-Init 的表现显著优于秩-1 和秩-8 随机初始化，这表明，将低秩更新与主要的跨模态差异对齐，能够提高优化效率。更多采用扩展训练预算的跨骨干网络结果见附录F。

6.5. 研究结果总结

在各项消融研究中，Gap-Init 在多种实验条件下均展现出有效性、鲁棒性和参数效率方面的优势。首先，它在紧凑的校准方案下实现了最佳性能，凸显出较高的数据效率。其次，对校准组成和噪声注入的研究揭示了一种微妙的几何权衡：尽管模态差距严格具有域特异性（分布外校准失效便证明了这一点），但引入兼容多样性——例如混合域数据或受控噪声——能够发挥有益的正则化作用。

第三，多种子评估表明，在秩1适配下，Gap-Init 的性能上限高于 LoRA。

此外，在极低秩约束下，Gap-Init 的表现与 DoRA 等更强的参数高效微调基线相当，甚至更胜一筹。这表明，在实验条件相同的情况下，恰当的初始化可能比提高适配秩更为关键。具体而言，尽管随着训练时间延长，性能差距会缩小，但与间隙方向显式对齐仍能提供更优的初始化，从而加速收敛并提升性能。

综上，这些结果有力地证明，所提出的基于初始化的适配策略能够在不同模型和训练机制下实现有效、鲁棒且参数高效的低秩适配。

关于校准鲁棒性、噪声敏感性和跨骨干网络泛化的全面消融研究见附录 D、E和 F。

7. 结论

本研究揭示了多模态模型中秩-1 参数高效微调失效背后的一个根本性几何瓶颈。通过分析表示几何，我们发现视觉和文本特征形成了各向异性锥，其差异包含一个以平移为主导的分量，该分量会对秩-1 约束下的优化产生不成比例的影响。这种结构意味着，标准随机初始化很可能使秩一更新方向与该分量错位，进而导致梯度流减弱和优化不稳定。

我们提出**Gap-Init**，一种几何感知初始化策略，它以简单而有效的方式将 LoRA 更新子空间与估计的模态差距方向对齐。Gap-Init 仅需一个小型且与领域对齐的校准集，无需额外参数，却能持续稳定秩-1 训练，并在图像描述、视觉问答、零样本迁移和幻觉鲁棒性方面达到或略微超过强劲的秩-8 基线。综合我们的分析与实验表明，在极端低秩极限下，方向的重要性可以不亚于秩。尽管我们关注的是秩-1 适配，该分析也为极度受限的适配机制揭示了一条更广泛的原则。

与 DoRA 或 PiSSA 等权重空间改进不同，Gap-Init 完全在激活空间中运行，并专为病态的秩-1 范式而设计；在该范式中，优化受方向限制，而非容量限制。这些发现凸显了数据与激活空间几何在参数高效微调中的重要性，并表明理解表征对齐或可为多模态大语言模型解锁更高效的适配策略。

在秩为 1 时，瓶颈从秩转向方向。

影响声明

本研究探讨多模态表示的几何属性，并提出一种更稳定、参数效率更高的方法，用于适配大型视觉–语言模型。由于我们的方法降低了微调的计算成本，因此可能让资源有限的研究团队也能更广泛地使用多模态适配技术。更高效的微调也可能减少训练大型模型所带来的能耗。

与此同时，扩展和适配多模态模型会引发重要的伦理考量。效率的提升并不能消除预训练模型固有的风险，例如偏见关联、幻觉，或在分布偏移下的错误推理。Gap-Init聚焦于低秩适配器的初始化，并不直接应对这些更广泛的安全挑战。因此，在真实场景中部署经过适配的模型时应当谨慎，尤其是在涉及高风险决策或敏感内容的环境中。

我们的分析凸显了模态对齐的几何结构，但并未改变模型所基于的底层数据。因此，训练数据中存在的任何社会偏见在适配后仍可能持续存在。我们鼓励从业者采用适当的审计方法，并从公平性、鲁棒性以及潜在的下游危害等方面评估模型。

总体而言，我们希望这项工作有助于加深对多模态表示学习的理解，同时也强调对经过适配的大规模模型进行负责任评估与部署的必要性。

参考文献

Alayrac, J.-B., Donahue, J., Luc, P., 等。Flamingo：一种用于少样本学习的视觉语言模型。神经信息处理系统进展，35，2022。

DeepMind。Gemma 3 技术报告。*arXiv* 预印本 *arXiv:2503.19786*，2025。

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA：高效微调量化大语言模型。载于神经信息处理系统进展（*NeurIPS*），2023。网址https://arxiv.org/abs/2305.14314。

Ethayarajh, K. 上下文化词表示的上下文性如何？载于*2019*年自然语言处理经验方法会议论文集（*EMNLP*），第668–677页，2019。

Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. 面向自然语言处理的参数高效迁移学习。发表于国际机器学习会议，2019。

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA：大语言模型的低秩适配。发表于国际学习表征会议（*ICLR*），2022a。Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA：大语言模型的低秩适配。发表于国际学习表征会议（*ICLR*），2022b。网址：https://openreview.net/forum?id=nZeVKeeFYf9。Li, J., Li, D., Xiong, C., and Hoi, S. BLIP：用于统一视觉语言理解与生成的语言-图像预训练引导式自举。发表于国际机器学习会议，2022。

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2：通过冻结的图像编码器和大语言模型引导语言-图像预训练。载于国际机器学习会议（*ICML*），2023。

Liang, W., Zhang, Y., Kwon, Y., Yeung, S. 和 Zou, J. 注意差距：理解多模态对比表征学习中的模态差距。载于神经信息处理系统进展 *(NeurIPS)*，2022 年。网址https://openreview.net/forum?id=S7Evzt9uit3。

Liu, H., Li, C., Wu, Q. 和 Lee, Y. J. 视觉指令微调。载于神经信息处理系统进展 *(NeurIPS)*，2023 年。

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.- C. F., Cheng, K.-T. 和 Chen, M.-H. DoRA：权重分解低秩适配。载于 第 *41* 届国际机器学习会议论文集 *(ICML)*，《机器学习研究论文集》，2024 年。网址 https://proceedings.mlr.press/v235/liu24bn.html。

Meng, F., Wang, Z., 和 Zhang, M. PiSSA：大语言模型的主奇异值与奇异向量适配。载于神经信息处理系统进展（*NeurIPS*），2024。网址：https://arxiv.org/abs/2404.02948。

Mu, J. 和 Viswanath, P. All-but-the-top：词表示的简单有效后处理。载于 国际学习表征会议，2018a。

Mu, J. 与 Viswanath, P. All-but-the-top：简单有效的词表示后处理。载于国际学习表征会议（*ICLR*），2018b。URL https://arxiv.org/abs/1702.01417。

Sung, Y.-L.、Cho, J. 与 Bansal, M. VL-Adapter：面向视觉与语言任务的参数高效迁移学习。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集（*CVPR*），第 5217–5227 页，2022。doi: 10.1109/CVPR52688.2022.00516。

Vershynin, R. 高维概率：数据科学应用导论。剑桥大学出版社，2018年。

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., 等。《Qwen2-vl：提升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。

A. 秩一优化坍塌的高斯玩具模型

本附录对正文中讨论的秩一优化坍塌的几何解释提供形式化分析。我们引入一个最简高斯平移模型，并证明在高维情况下，随机初始化的秩1 LoRA更新会遭受严重的梯度抑制，而对齐感知初始化则可以避免这一问题。

A.1. 高斯平移模型。

我们将视觉隐藏状态和文本隐藏状态建模为 $\mathbb{R}^d$中的两个各向同性高斯簇:

$$x_v \sim \mathcal{N}(\mu_v, \sigma^2 I_d), \qquad x_t \sim \mathcal{N}(\mu_v + \vec{g}, \sigma^2 I_d), \tag{2}$$

其中 $\vec{g}$表示模态间隙平移向量。由于球形协方差， $\vec{g}$是空间中唯一的非各向同性结构。

A.2. 秩一更新与最优方向

一个秩1 LoRA ada pter $\Delta W = ba^\top$ 会诱导更新 $x_v \mapsto x_v + \Delta W x_v$，从而导致预期的对齐损 s

$$\mathcal{L} = \mathbb{E}\|x_v + \Delta W x_v - x_t\|^2. \tag{3}$$

对高斯分布取期望可将优化问题简化为

$$\min_{b,\,\theta} \|\theta b - \vec{g}\|^2,$$

这意味着最优方向满足

$$b^* \parallel \vec{g}. \tag{4}$$

这表明任何成功的秩一更新都必须恢复模态差距方向。

A.3. 随机初始化下的梯度抑制。

标准 LoRA 在单位球面上进行均匀随机初始化 b： $b \sim$ Unif($\mathbb{S}^{d-1}$)。有用梯度分量与其在 $\vec{g}$上的投影成正比:

$$\nabla\mathcal{L} \propto \langle b, \vec{g}\rangle. \tag{5}$$

对于 $\mathbb{R}^d$中的随机方向，平方余弦相似度 $\langle b,\ \hat{}g\rangle^2$服从 Beta($_{1\,2}$， $^{d-1}_{2}$) 分布，因此

$$\mathbb{E}[\langle b, \hat{g}\rangle^2] = \frac{1}{d}, \qquad |\langle b, \hat{g}\rangle| = \mathcal{O}(d^{-1/2}). \tag{6}$$

A.4. 优化坍缩定理

定理 A.1 （优化坍缩）。在高斯平移模型下，随机初始化的秩*1 LoRA*更新会使有用梯度分量 $\mathcal{O}$（$d^{-1/2}$）受到抑制 $\langle b, \vec{g}\rangle$。随着维度增加，优化器会进入一个平坦区域，无法可靠地恢复模态差距方向。初始化 $b \propto \vec{g}$ 可恢复 $\mathcal{O}$（1）梯度流并避免这种坍缩。

A.5. 合成验证。

我们使用合成高斯数据验证预测的缩放行为。随机秩1方向呈现出预期的$\mathcal{O}$（$d^{-1/2}$）衰减，而间隙对齐初始化在不同维度下保持信号恒定（图5）。这证实，观测到的秩1坍缩源于高维几何，而非表征能力不足。

B. 为何平移主导会出现在多模态对齐中

本附录从机理层面探讨并佐证这一经验观察：现代视觉-语言模型中的模态差距通常主要由类平移分量主导。尽管多模态失配本质上并非低维的，但当代训练目标和架构的若干特性会使失配偏向于形成一致的全局偏移。

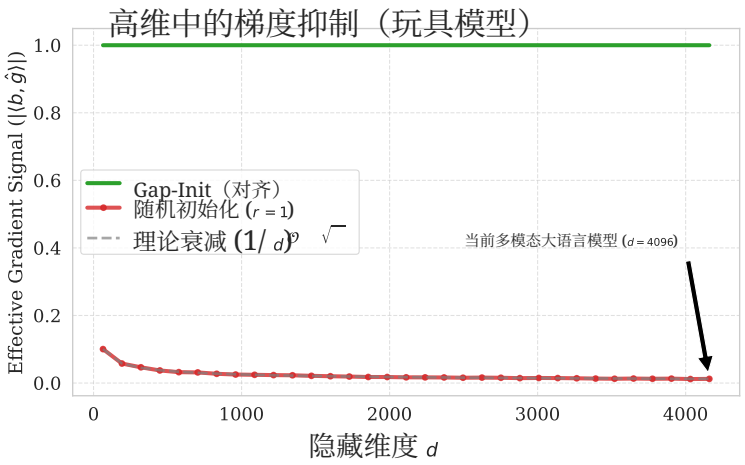


图 *5.* **高维中的梯度抑制（高斯玩具模型）**。我们计算维度为 $d \in [64, 4096]$ 时秩1 LoRA 方向的有效梯度幅值 $|\langle b, \hat{g} \rangle|$。随机初始化呈现出预测的 $\mathcal{O}(d^{-1/2})$ 衰减（红色），这表明沿模态差距方向的几乎所有梯度信号都会在高维情况下消失。相比之下，GAP-INIT（绿色）保持了与维度无关的 $\mathcal{O}(1)$ 信号。这验证了秩1 LoRA 中的优化坍缩是高维几何造成的，而非秩不足所致。

B.1. 对比目标促使锥体形成。

对比预训练目标函数（例如 CLIP、BLIP-2）通过拉近匹配对、推远不匹配对来优化表征。这一过程会在嵌入空间中引发强烈的全局各向异性（Mu & Viswanath，2018b；Ethayarajh，2019），使表征坍缩为模态特定锥。一旦这种锥状结构形成，两种模态之间的主要差异便可由其质心间的位移很好地近似，从而自然产生类平移间隙。

B.2. 异构编码器架构引入系统性偏差。

视觉编码器和语言模型在架构上各不相同，并且采用不同的归一化和缩放约定进行训练。这些系统性差异（例如残差路径统计、分词效应）会体现为嵌入均值的整体偏移。由于这些偏差在不同实例间保持一致，它们会累积形成稳定的全局平移向量，而非因实例而异的失真。

B.3. 多模态桥接模块近似呈线性运行。

许多视觉语言模型依赖浅层多模态桥接模块（例如 BLIP-2 中的 Q-Former），将视觉表征投影到语言嵌入空间。相较于冻结的语言模型，这些模块运行于近似线性的区间。因此，它们需要纠正的主要失配也可以通过线性、类似平移的变换来很好地描述。

B.4. 下游对齐只需粗略调整位置。

视觉问答和图像描述等下游任务，主要要求模态之间具备语义兼容性，而非对表征进行精细的几何扭曲。因此，有效对齐往往只需将视觉特征移向语言流形，而无需重塑其内部结构。平移能够捕捉这种粗略的一阶校正。

B.5. 总结

综合来看，强烈的嵌入各向异性、系统性的架构偏移，以及近似线性的多模态桥接，共同使模态差距呈现出以平移为主导的结构。这直观地解释了为何与平移方向对齐的秩一更新，尽管底层

对齐问题具有更高维度。

C. SPOT-Gap度量的完整定义

我们给出*SPOT-Gap* 度量的完整公式，该度量用于第4.4 节，量化逐层模态错位。SPOT-Gap 是**语义保持最优传输差距**的缩写。这是一种轻量、无需训练的诊断方法，灵感源自最优传输原理：我们不求解完整的最优传输方案，而是估计在保持局部结构的同时，将视觉特征传输到文本流形所需的语义位移。SPOT-Gap 由三个部分组成：（1）样本投影语义差距；（2）结构保持几何差异；（3）邻域一致性平滑。

令 $H_v^{(l)} = \{h_{v,i}^{(l)}\}$ 和 $H_t^{(l)} = \{h_{t,i}^{(l)}\}$ 分别表示从校准集 $\mathcal{D}_{\text{cal}}$中提取的第 l层视觉和文本隐藏状态。所有向量均属于 $\mathbb{R}^d$。

C.1. 样本投影语义差距

对于每个配对样本（i），我们考虑将视觉特征对齐到其文本对应项所需的语义位移：

$$g_{\text{sem},i}^{(l)} = h_{t,i}^{(l)} - h_{v,i}^{(l)}.$$

幅度

$$s_{\text{sem},i}^{(l)} = \|g_{\text{sem},i}^{(l)}\|_2$$

衡量在样本层面匹配两个模态的语义传输成本。这是在对角最优传输耦合下，将 $h_{v\,i}^{(l)}$,传输至 $h_{t\,i}^{(l)}$,的成本的一阶近似。

C.2. 结构保持几何差异

为了保持局部语义结构这一关键的最优传输概念，我们引入一个差异项，用于衡量不同模态之间的局部几何差异。对于每个特征 $h_{v\,i}^{(l)}$,，令 $\mathcal{N}_k(h_{v\,i}^{(l)},)$ 表示其在 $H_v^{(l)}$中的 k 个最近邻。我们计算其局部离散度：

$$d_{v,i}^{(l)} = \frac{1}{k} \sum_{j \in \mathcal{N}_k(h_{v,i}^{(l)})} \|h_{v,i}^{(l)} - h_{v,j}^{(l)}\|_2.$$

同理，从 $H_t^{(l)}$ 计算 $d_{t\,i}^{(l)}$,。

几何差异定义为

$$s_{\text{geom},i}^{(l)} = |d_{t,i}^{(l)} - d_{v,i}^{(l)}|.$$

这刻画了需要多少结　构扭曲，才能在局部将视觉邻域“传输”到文本邻域 s.

C.3. 邻域一致性平滑

为降低噪声并加强结构保持，我们在每个样本邻域上对组合后的语义分数与几何分数进行平滑：

$$\bar{s}_i^{(l)} = \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} \left(\alpha_{\text{sem}}\, s_{\text{sem},j}^{(l)} + \alpha_{\text{geom}}\, s_{\text{geom},j}^{(l)} \right),$$

其中 α_{sem} 与 α_{geom} 为固定权重。该步骤在不求解完整传输映射的情况下，近似最优传输的全局平滑约束。

C.4. 最终 SPOT-Gap 分数

逐层 SPOT-Gap 分数是平均平滑后的差异：

$$S_{\text{gap}}^{(l)} = \frac{1}{N} \sum_{i=1}^{N} \bar{s}_i^{(l)}.$$

算法 2 SPOT-Gap 评分过程（层 l）

输入： 视觉隐状态 $H_v^{(l)} = \{h_{v,i}^{(l)}\}$，文本隐状态 $H_t^{(l)} = \{h_{t,i}^{(l)}\}$、邻域大小 k、权重 α_{sem}、α_{geom}。**输出：** **SPOT-Gap 分数** $S_{\text{gap}}^{(l)}$。**步骤 1：计算样本投影语义差距** **对于每个样本** i **执行** $g_{\text{sem},i} \leftarrow h_{t,i}^{(l)} - h_{v,i}^{(l)}$，$s_{\text{sem},i} \leftarrow \|g_{\text{sem},i}\|_2$。**结束循环。** **步骤 2：计算几何差异** **对于每个样本** i **执行** **识别** k **最近邻** $\mathcal{N}_k(h_{v,i}^{(l)})$ **于** $H_v^{(l)}$：$d_{v,i} \leftarrow \frac{1}{k} \sum_{j \in \mathcal{N}_k} \|h_{v,i}^{(l)} - h_{v,j}^{(l)}\|_2$。**识别** k **最近邻** $\mathcal{N}_k(h_{t,i}^{(l)})$ **于** $H_t^{(l)}$：$d_{t,i} \leftarrow \frac{1}{k} \sum_{j \in \mathcal{N}_k} \|h_{t,i}^{(l)} - h_{t,j}^{(l)}\|_2$。$s_{\text{geom},i} \leftarrow |d_{t,i} - d_{v,i}|$。**结束循环。** **步骤 3：邻域一致性平滑** **对于每个样本** i **执行** **在配对样本之间识别** k **最近邻** $\mathcal{N}_k(i)$：$\bar{s}_i \leftarrow \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k} (\alpha_{\text{sem}} s_{\text{sem},j} + \alpha_{\text{geom}} s_{\text{geom},j})$。**结束循环。** **步骤 4：最终层分数：** $S_{\text{gap}}^{(l)} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \bar{s}_i$。**返回** $S_{\text{gap}}^{(l)}$

解释。 较高的 $S_{\text{gap}}^{(l)}$ 表明，层 l 在模态之间表现出较大的语义位移和结构不一致——这正是使某一层难以对齐的特征。因此，SPOT-Gap 作为一种保持语义、受最优传输启发的诊断方法，用于识别存在严重多模态错位的层。

用法。 在第4.4节中，我们使用 SPOT-Gap 来验证我们的几何分析，并指导可选的层选择。尽管 Gap-Init 即使在朴素选择下也表现强劲，SPOT-Gap 仍使我们能够探究内部模态结构并确认对齐敏感层。

D. 校准集大小与数据组成

D.1. 校准集大小敏感性

为评估估计的模态间隙向量 g的鲁棒性，我们改变校准集大小 $|\mathcal{D}_{\text{cal}}| \in \{16, 64, 256, 1024\}$，并使用 CIDEr 和 BLEU-4 报告在 COCO 图像描述上的结果。

如表9所示，Gap-Init 的有效性取决于能否获得对模态差距具有统计代表性的估计。在校准集极小（$|\mathcal{D}_{\text{cal}}| = 16$）时，由于估计方向的方差较高，性能下降，表明对实例特定噪声的平均不足。

随着校准集大小增加，性能迅速提升，并在 $|\mathcal{D}_{\text{cal}}| = 256$ 处达到明显峰值。进一步将大小增加到 1024 时，收益逐渐减小，表明模态差距对应的是一种低秩全局几何特征，无需大规模采样。除非另有说明，我们采用 $|\mathcal{D}_{\text{cal}}| = 256$ 作为默认配置。

表 *9.* **校准集大小对秩-1适配的影响** $|\mathcal{D}_{\text{cal}}|$。 性能在中等校准集大小时达到峰值，这表明无需穷举采样即可稳健地估计全局平移方向。

$\lvert D_{\text{cal}}\rvert$	CIDEr	BLEU-4
16	109.99	32.36
64	137.68	40.47
256	**142.05**	**42.30**
1024	140.03	41.36

表*10.* **校准数据来源对秩1适配的影响**。 纯粹的分布外校准会降低性能，而混合兼容域则比域内基线略有提升。

校准来源	CIDEr	BLEU-4
Flickr30k（分布外）	108.85	31.50
COCO（领域内）	140.56	41.44
混合（50% COCO + 50% Flickr）	**141.91**	**41.95**

D.2. 校准数据构成：领域特异性与多样性

上一小节研究了校准数据规模，接下来我们将考察校准数据分布的影响。我们固定$|\mathcal{D}_{\text{cal}}| = 256$，并使用以下数据评估 Gap-Init：（i）纯领域内数据（COCO）、（ii）纯分布外数据（Flickr30k）以及（iii）混合校准集（50% COCO，50% Flickr30k）。

表 10 中的结果凸显了模态间隙估计的两个关键特性。

首先，模态差距具有领域特异性。与域内校准相比，使用纯分布外数据（Flickr30k）进行校准会导致性能大幅下降，这表明平移方向会随数据分布而变化。

其次，引入兼容多样性可以起到温和的正则化作用。混合校准集的表现略优于纯域内基线。由于 Flickr30k 和 COCO 共享相似的自然图像语义流形，将两者混合不会扰乱主要的平移方向，反而会引入额外的变异性，从而使估计更加稳定。这表明，有效的秩-1初始化位于一个与目标领域高度对齐、同时又足够宽泛以提升鲁棒性的区域。

E. 对间隙方向噪声的鲁棒性

E.1. 向差距向量注入噪声

为了评估 Gap-Init 对方向扰动的敏感性，我们构造带噪声的间隙向量

$$g' = \text{normalize}(g + \epsilon \cdot \eta),$$

其中 $\eta \sim \mathcal{N}(0, I)$ 和 $\epsilon \in \{0.0, 0.2, 0.6, 1.0\}$。结果报告于表 11。

如表 11 所示，性能对噪声注入呈现非单调响应。在低噪声水平（$\epsilon = 0.2$）下，性能略有下降，这表明估计的间隙方向在局部已经良好对齐，较小的扰动可能有害。

在中等噪声水平（$\epsilon = 0.6$）下，性能恢复并略有提升，BLEU-4 达到峰值。这一行为与随机正则化效应一致，即从估计方向略微偏离中心进行初始化，有助于探索更宽的吸引域。

相比之下，当噪声淹没信号（$\epsilon = 1.0$）时，性能会急剧下降。这种情况可作为阴性对照：一旦方向信息在很大程度上被破坏，秩1适配便会失效，这证实了 Gap-Init 极其依赖保留平移方向中编码的语义对齐。

表*11.* **噪声注入的影响**（ϵ）。指标以百分比表示。尽管该方法对较大偏差（$\epsilon = 1.0$）较为敏感，但适度的噪声（$\epsilon = 0.6$）能取得最高的 BLEU-4 分数。

噪声水平（ϵ）	CIDEr	BLEU-4
0.0	141.84	42.55
0.2	140.04	41.62
0.6	**142.00**	**43.43**
1.0	133.45	39.62

表*12.* 跨骨干网络泛化能力在低秩适配下的 Gap-Init。指标以百分比表示。Gemma3-4B 的结果分别展示了单轮和五轮训练预算下的结果。

方法	CIDEr	BLEU-4	METEOR
Qwen2-VL-7B-Instruct			
秩=8，随机	142.57	41.72	32.04
秩=1，随机	143.67	42.55	31.89
秩=1，Gap-Init	**144.11**	**42.97**	**31.95**
Gemma3-4B（1 轮）			
秩=8，随机	89.62	22.67	23.86
秩=1，随机	90.35	24.95	24.28
秩=1, Gap-Init	**95.24**	**27.27**	**26.33**
Gemma3-4B（5 轮）			
秩=8，随机	94.82	27.35	27.91
秩=1，随机	95.43	28.12	28.36
秩=1，Gap-Init	**96.17**	**28.89**	**28.74**

F. 额外的跨骨干网络结果

本附录提供了更多结果，展示 Gap-Init 在具有不同架构、规模和训练预算的视觉-语言骨干网络上的泛化能力。这些实验补充了主要结果，说明在模型容量和优化机制各异的情况下，显式地与模态差距方向对齐会产生怎样的表现。

F.1. 视觉-语言骨干网络间的跨模型泛化

我们在两个具有代表性的骨干网络上评估 Gap-Init：Qwen2-VL-7B-Instruct，这是一种容量充足的大规模视觉-语言模型；以及 Gemma3-4B，这是一种规模较小、受更严格的优化条件限制的模型。结果见表12。

在大规模骨干网络 Qwen2-VL-7B 上，采用随机初始化的秩-1适配，其表现已可与更高秩的基线相媲美，表明当模型容量充足时，性能对秩的敏感性会降低。尽管如此，在秩为 1 时应用 Gap-Init，仍能在所有评估指标上持续提升性能，并在所比较的设置中取得最佳总体结果。

在规模较小的骨干网络 Gemma3-4B 上，Gap-Init 的优势更为显著。在单轮训练预算下，采用随机初始化的秩-1适配仅带来有限的提升，而 Gap-Init 则在所有指标上均取得显著改善，其中 CIDEr分数大幅提升。这凸显了在优化资源有限时，使低秩更新与主要跨模态差异相匹配的重要性。

重要的是，随着训练预算增加，这一优势依然存在。经过五轮训练，随机初始化的低秩更新能够部分恢复有效方向，从而缩小性能差距。然而，Gap-Init 仍优于秩-1和秩-8随机基线，这表明，与模态差距方向显式对齐能够提供更优的初始化，加快收敛并提升最终性能。

表 *13.* LoRA 配置详情。

参数	值
秩（r）	1（主设置），8（基线对比）
缩放因子 (α)	2（秩 1）、16（秩 8）
LoRA 丢弃率	0.05
初始化（基线）	随机高斯
初始化（Gap-Init）	$B^{(\ell)} \leftarrow g^{(\ell)}/\|g^{(\ell)}\|_2, \quad A^{(\ell)} \leftarrow 0$

表*14*。 训练超参数。

参数	值
优化器	AdamW
学习率	2×10^{-4}
学习率调度器	余弦衰减
批量大小（每个设备）	32
梯度累积步数	2
有效批量大小	64
训练轮数	5
精度	FP16 / BF16（取决于硬件）
随机种子	{42, 123, 2024, 999, 7}

G. 实现细节

本附录提供完整的实现细节，以确保所有涉及 Gap-Init 及相关基线的实验均可复现。

G.1. 模型架构与适配范围

除非另有说明，我们在以 OPT-2.7B 语言模型为骨干网络的 BLIP-2 上进行实验。语言模型的隐藏维度为 $d = 2560$。视觉特征通过冻结的 Q-Former 和后续的线性投影层对齐到语言模型空间。所有模态差距计算和 LoRA 适配均在对齐后的语言模型隐藏状态空间中进行。

适配的模块。LoRA 适配器被插入语言模型以及（可选的）视觉编码器的线性层中。具体而言，我们针对以下子模块:

- 注意力层：q 投影、k 投影、v 投影、o 投影。
- MLP 层：fc1、fc2

除非另有说明，所有 Gap-Init 实验中 Q-Former 均被冻结。

G.2. LoRA 配置

表13 汇总了全文所使用的 LoRA 超参数。

对于秩-$r > 1$， $B^{(\ell)}$ 中只有第一列与模态差距对齐；其余列均为随机初始化。

G.3. 训练超参数

表 14 报告了图像描述和视觉问答任务所使用的训练超参数。

所有实验均为 con 采用相同的训练计划评估不同初始化方法，以确保比较公平 on.

G.4. 模态间隙估计的校准流程

模态差距向量 使用一个小型、无标签校准集 $\mathcal{D}_{\text{cal}}$ 进行估计。除非另有说明，校准样本取自 COCO Train2017 划分（域内）。

校准规模。 我们使用 256 个图文对，这可以在几乎不增加计算开销的同时，稳定地估计差距。关于校准规模的消融实验见第 5节。

流程。 校准过程如下：

1. 从 $\mathcal{D}_{\text{cal}}$ 中随机抽取 $(x_{\text{img}},\ x_{\text{txt}})$ 对。2. 使用冻结的预训练模型进行前向传播。3. 对于每个 Transformer 层 ℓ，提取：

   - 对齐后的视觉隐状态 $h_{(\ell)v}$，
   - 文本隐藏状态 $h_{(\ell)\,t}$。

4. 计算样本投影差距 $g^{(i,\ell)} = h_t^{(\ell)} - h_v^{(\ell)}$。

5. 应用 SPOT-Gap 平滑（附录 C），并对样本取平均以得到 $g^{(\ell)}$。

由此得到的间隙向量会被存储并用于初始化；在此阶段不会计算梯度。

G.5. 层选择策略

我们考虑两种层选择策略：

- **朴素选择：** 顶部的k Transformer 层。
- **间隙引导选择（GG-Safe）：** 按 SPOT-Gap 分数排名靠前的k 层，同时为确保安全性排除早期视觉编码器层。

除非另有说明，否则 Gap-Init 使用朴素选择；GG-Safe 仅用于验证模态差距的诊断价值。

G.6. 全量微调说明

我们在本工作中并未对骨干模型进行全量微调。图中凡提及“全量微调上限”之处，均表示根据先前已报告结果得出的近似上界，并非由我们训练的实验基线。这一点将在修订稿中予以明确说明。